Dipl.-Ing. Selver Softic, BSc

Knowledge Discovery through Mining and Profiling from Semi-Structured Sparse Text Artifacts Using Semantic Technologies and Linked Data for Education and Research

PhD Thesis

Graz University of Technology

Institute of Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Supervisor: Priv.-Doz. Dipl.-Ing. Dr. techn. Martin Ebner

Graz, March 2017

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …… ……..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…… ……..
date (signature)

For my dear beloved parents who always believed in me and thought me that love, knowledge and honesty are the most valuable qualities a man can reach. I would only add that sharing knowledge makes it even more valuable. For my dear wife who always supported me during all these years. Thank you for your patience and love. For my beloved new born son who filled my life with pride and love. This should be an inspiration for you that everything in a life is reachable with dedication and hard work. Many thanks to my supervisor and friend Dr. Martin Ebner whose advices has been always inspiration, comfort and support in moments where I doubted my scientific path. I also want to thank Prof. Erik Mannens from IDLab at Ghent University for having always understanding for my requests and for giving me a chance to work with his extraordinary team. To my friends and colleagues Behnam Taraghi and Laurens De Vocht with whom I always could discuss and improve my work. I would never be able to achieve this without you guys. Going this way was half hard thanks to your friendship and your cooperation. Finally, many thanks to all my master and bachelor students. Without their excellent works this thesis could not be approved. I also want to thank to all the colleagues who I met in labs and conferences as well to my mentor regarding semantics Dr. Michael Hausenblas. Thank you Dr. Mike.

Abstract

This thesis describes the contribution made by author to the field of Knowledge Discovery in semi-structured text fragments such as tweets or system logs from online learning system PLE (Personal Learning Environment) at Graz University of Technology for the field of Research and Education. The thesis elaborated on the applications area related use cases such as Research 2.0 and Visual (Learning) Analytics. The primary addressed scientific fields are Semantic (Data) Modeling, Data Profiling and Data Mining of tacit information hidden within the semi-structured text fragments. Based upon Semantic Modeling of such data and its interlinking with reliable Linked Data sources is shown that mining, profiling and search approaches contribute Knowledge Discovery, in particular, in finding new relevant resources, persons, events for researchers and in tracking the learners and learning objects as well as related actions within the PLE at Graz University of Technology.

Kurzfassung

Diese Arbeit beschreibt den Beitrag des Autors auf dem Gebiet der Wissensentdeckung in semi-strukturierten Textfragmenten wie Tweets oder Systemprotokollen aus dem Online-Lernsystem PLE (Personal Learning Environment) der Technischen Universität Graz für die Bereiche der Forschung und Lehre. Für die genannte Anwendungsbereiche wurden Anwendungsfälle wie Research 2.0 und Visual (Learning) Analytics untersucht. Primär adressierte wissenschaftliche Felder waren Semantic (Data) Modeling, Data Profiling und Data Mining von impliziten Informationen, die in den semi-strukturierten Textfragmenten verborgen sind. Basierend auf semantischer Modellierung der semi-strukturierten Textfragmente und ihrer Vernetzung mit zuverlässigen Linked Data Quellen wurde gezeigt, dass die Mining-, Profiling- und Suchansätze dazu beitragen können, effiziente Wissensentdeckung zu betreiben insbesondere bei der Suche nach neuen relevanten Ressourcen, Personen, und Veranstaltungen für Forscher sowie bei der Verfolgung der Lernenden und Lernobjekten und damit zusammenhängenden Handlungen innerhalb des PLE an der Technischen Universität Graz.

Contents

List of Figures

1 Introduction

This chapter introduces the main research topics of this work. It begins with an general overview of current situation and narrows down the focus to microblogs and system logs as representatives of very common form of short semi-structured data and its potential for knowledge discovery. It also elaborates on the motivation of Web technologies and trends like the Web of Data, which is a main challenge of this research effort together with user groups of scholars targeted with this work. Finally, the chapter rounds up with the definition of research questions to be answered in subsequent chapters. The closing subsection gives a short overview over the following chapters.

1.1 General Overview

Currently, most of the content on the Web is user generated. Microblogs in the Web 2.0, as short form of blogging, gained strong popularity and importance in recent years. Microblogging platforms like Twitter[1], Tumblr[2], or Friendfeed[3] as many others[4] attract daily many users with different social, cultural and educational backgrounds. While posting users share their emotions, opinions or commonly useful information. The same situations is also with background data generated by machines. Millions of systems are creating logs and reports for monitoring purposes. Within their line they hide valuable information about system, users and activities. This information reformulated in appropriate form can easily contribute

[1] http://www.twitter.com, last access: 2017–05–29
[2] http://www.tumblr.com, last access: 2017–05–29
[3] http://http://friendfeed.com/, last access: 2017–05–29
[4] https://tinyurl.com/h482bkh, last access: 2017–05–29

better visibility and system design which has direct influence to the users. Together, users and machines which produce content and interact with each other (useruser, usermachine, machinemachine), they form the Social Web which is the essence of internet known nowadays.
Many information sources of public interest remain captured behind the "Walled Garden"[5]. Combining information resources over its walls leads to a high degree of mismatches between vocabulary and data structure of the different sources because of lack of structured and standardized data exchange models and protocols which could serve different technology platforms. Valuable data produced by the masses and clusters of machines remains in so-called “data silos” bound to a specific platform or somewhere within databases without proprietary interfaces, though public but barely reused. The access to this data is associated with specialized application interfaces (API’s) which require a high degree of technical knowledge to retrieve the data in a desirable form.
The Web of Data also known as "Semantic Web" extends the Web by functionality by bringing structure and giving well-defined meaning to the content and it enables humans and machines to work together using controlled vocabularies.

1.2 Problem Statement

Sparse[6] semi-structured contextually bound text fragments like tweets or system logs suit well for automated context analysis and semantically enrichment. Feature extraction or named entity recognition on text fragments performs well and works more precisely on mid-sized and small text artifacts as shown in Bacchelli et al. (2011). The possibility of monitoring such content, not only for humans but also for machines, contributes to creation of more intelligent user interfaces, better common information awareness, and to more technically profound agent, search and recommendation systems.

[5] https://en.wikipedia.org/wiki/Closed_platform, last access: 2017–05–29
[6] "sparse" in my thesis means short or small in terms of information content

Current research on microblogs focused on analysis of their content and social aspects around the content. On the other hand system logs are completely neglected. No approach tackled the problem of using this information to transform it in form more understandable to machines for better exploration and acquisition of additional knowledge implicitly existent in their content for specific groups of users.
A particular group of users with such case specific needs are researchers. Resources for research are not always easy to explore, and rarely come with strong support for identifying, linking and selecting those that can be of interest for further investigation. Personalized adaptation of the Web to the needs of researchers is the main vision of Research 2.0. The other beneficial group from analyzing the system logs would be learning environments which nowadays need to be more reflective towards the users and teachers using them. Efforts addressing these potential in research are Learning and Visual Analytics and Educational Data Mining. Bundling the insights of addressed to find a common approach to prepare and mine data represents the mission of current work. This thesis wants to list the possibilities whether an approach with semantic modeling of data and enhancing it with Linked Data knowledge bases would contribute to better solutions for these fields.

1.3 Research Questions

As main objective this work aims at answering following questions based upon observations and assumptions elaborated in "Problem Statement":

- RQ1: Do sparse semi-structured text fragments (as tweets and user logs) contain information useful for better exploration of research related and learning resources?
- RQ2: How such content can be semantically modeled and explored with machines using semantic (web) technologies and Linked Data?
- RQ3: How such modeled data can be used to profile researchers on Twitter and explore related resourced on the Web or in case of system logs for reflections and improvements of learning systems?

1.4 Hypotheses

Based upon research questions following hypotheses will be tested:

- Hyp1: Semi-structured text fragments (as tweets and user logs) contain information useful for better profiling, search and exploration of research repositories and for better reflection of learning activities, as well they build solid base for improvement recommendations in TEL (Technology Enhanced Learning).
- Hyp2: Useful information from semi-structured texts (tweets and user logs) can be extracted, semantically modeled and retrieved with machines to serve fields like Learning Analytics, Semantic Search and User (Researcher, Learner) Profiling.
- Hyp3: Using short semi-structured text fragments (as tweets and user logs) semantically modeled and connected with domains of interests can contribute better profiling, search and exploration of researchers and related content as well as to better reflection and improvements in learning systems.

1.5 Methodology

My thesis applies the principle for a good experimental study presented by Robert (Bob) Taylor in the Workshop on "Introduction on Research in Experimental Computer Science" held in Palo Alto, California in 1991. The principle says: "you should build what you design and use what you build, as only through the extensive use of an artifact, you truly understand the implications of your work" Liskov (1992).

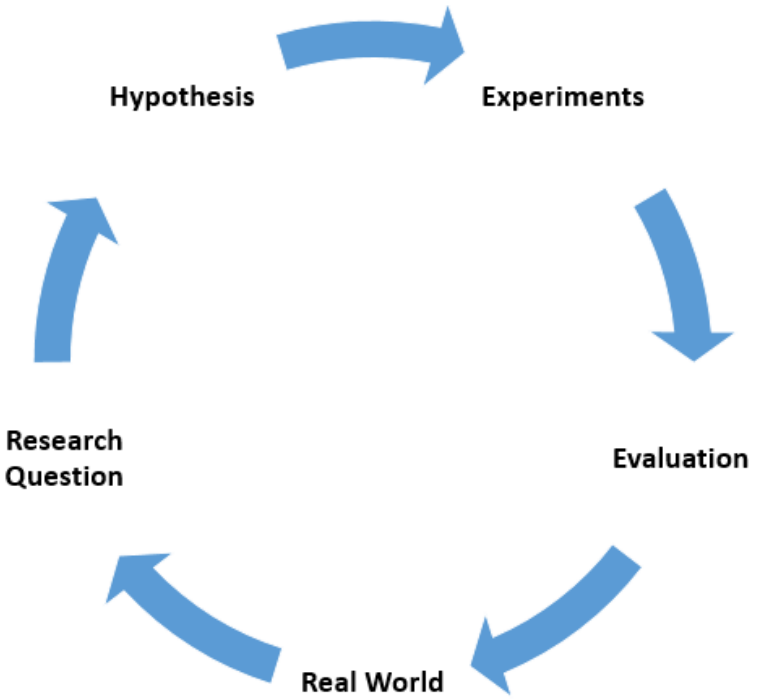


Figure 1.1: Research methodology of this thesis.

The methodology in my thesis follows the principles of "Experimental Science". The Experimental Science is based upon a process that uses experiments formulated as "tests or prototypes under controlled conditions to examine the validity of a hypothesis, or determine the efficacy of something previously untried"[7]. The used research methodology applies iterative

[7] http://www.thefreedictionary.com/Experimental+science, last access: 2017–05–

process shown in Figure 1.1. This iterative process starts with a real world problem, which provides a set of premises supporting the formulation of the initial research questions (here it is a research questions RQ1). Through the iteration of initial research question, with deeper insights to the occurring problems, challenges and circumstances, additional research questions emerged (research questions RQ2 and RQ3). In this sense, the iteration continues until the evaluation step is successful in terms of determination of answer giving insights and results.

1.6 Main Publications

This section introduces a list of main papers used for completing my thesis. The position of the author (myself) is outlined in bold text. To each paper there is a short description of the topic, as well as short statement to author's (my) contribution. Further, it is mentioned which research question the paper addresses and at which chapter the text from the paper was used. Introduced list of publications is chronologically ordered starting by 2010 and ending in 2017.

1. **Paper A: Softic, S.**, Ebner, M., Mühlburger, H., Altmann, T., and Taraghi, B. (2010). @twitter Mining #Microblogs Using #Semantic Technologies. In 6th Workshop on Semantic Web Applications and Perspectives (SWAP 2010). (pp. 1-12).
   Topics: Semantic Data Modeling, Twitter Mining, Data Profiling, Data Mining.
   Author's contribution: Main contribution
   Relates to research questions: RQ1, RQ2, RQ3
   Used in chapters: 2, 4, 4.3
2. **Paper B:** Ebner, M., Altmann, T., and **Softic, S.** (2011). @twitter Analysis of #edmedia10– is the #informationstream Usable for the #mass. Form@re [Elektronische Ressource], (74), 1-11. .
   Topics: Data Profiling, Data Mining, Twitter Mining
   Author's contribution: Co-author, conception of methodology and experiment

29

Relates to research questions: RQ1, RQ2
Used in chapters: 2, 3, 10.1

3. **Paper C:** De Vocht, L., **Softic, S.**, Ebner, M., and Mühlburger, H. (2011). Semantically driven Social Data Aggregation Interfaces for Research 2.0. In 11th International Conference on Knowledge Management and Knowledge Technologies. (pp. 43:1-43:10). New York, NY, USA: ACM.
Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Data Mining and Data Profiling, Mesh-up Interfaces
Author's contribution: Co-author on methodology, and contributor on implementation and evaluation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 5, 5.1
4. **Paper D: Softic, S.** (2012). Towards Identifying Collaborative Learning Groups using Social Media. International journal of emerging technologies in learning (Elektronische Ressource), 7(S2;2012), 15-21.
Topics: Twitter Mining, Technology Enhanced Learning, Data Profiling
Author's contribution: Main contributor
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 6, 6.1
5. **Paper E:** Thonhauser, P., **Softic, S.**, and Ebner, M. (2012). Thought Bubbles: a Conceptual Prototype for a Twitter based Recommender System for Research 2.0. In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies. (pp. 1-4). New York: ACM. 10.1145/2362456.2362496
Topics: Twitter Mining, Recommender Systems, Data Profiling, Research 2.0
Author's contribution: Co-author on methodology, and contributor on conception of implementation and evaluation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 3, 3.2
6. **Paper F: Softic, S.**, Taraghi, B., Ebner, M., De Vocht, L., Mannens, E., and Van De Walle, R. (2013). Monitoring Learning Activities in PLE Using Semantic Modelling of Learner Behaviour. In Human Factors in Computing and Informatics. (Lecture Notes in Computer Science ed., Vol. 7946, pp. 74-90). (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer. 10.1007/978-3-642-39062-3_5
Topics: Semantic Data Modeling, PLE, Visual Analytics, Learning

Analytics, Technology Enhanced Learning
Author's contribution: Main author, contributor on concept of the paper, implementer of the data related and analytic parts
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 6, 7.1

7. **Paper G:** Taraghi, B., **Softic, S.**, Ebner, M., and De Vocht, L. (2013)Learning Activities in Personal Learning Environment. In Proceedings of the 25th World Conference on Educational Media and Technology. (pp. 2466-2475).
   Topics: User Logs, PLE, Visual Analytics, Learning Analytics, Technology Enhanced Learning
   Author's contribution: Co-author of Semantic Web / mining related content, contributor on concept of Learning Analytics idea
   Relates to research questions: RQ1, RQ2
   Used in chapters: 2, 6, 7.1
8. **Paper H: Softic, S.**, Taraghi, B., and De Vocht, L. (2013). Activities and Trends Analytics in a Widget based PLE using Semantic Technologies. In Proceedings of the 5th International Conference on Computer Supported Education (CSEDU) 2013. (pp. 199-203). Porto, Portugal: SCITEPRESS.
   Topics: User Logs, PLE, Visual Analytics, Learning Analytics,Technology Enhanced Learning
   Author's contribution: Co-author of Semantic Web / mining related content, contributor on concept of Learning Analytics idea
   Relates to research questions: RQ1, RQ2
   Used in chapters: 2, 6, 7.1
9. **Paper I:** De Vocht, L., Mannens, E., Van de Walle, R., **Softic, S.**, and Ebner, M. (2013). A Search Interface for Researchers to Explore Affinities in a Linked Data Knowledge Base 21-24. In Proceedings of the ISWC 2013 Posters and Demonstrations Track a track within the 12th International Semantic Web Conference (ISWC 2013). (Vol. 1035, pp. 21-24).
   Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Data Mining, Data Profiling, Explorative Semantic Search
   Author's contribution: Co-author on methodology and implementation (architecture and user interface)
   Relates to research questions: RQ1, RQ2, RQ3

Used in chapters: 2, 5, 5.5

10. **Paper J**: **Softic, S.**, Ebner, M., Vocht, L. D., Mannens, E., and de Walle, R. V. (2013). A Framework Concept for Profiling Researchers on Twitter Using the Web of Data. In WEBIST 2013 - Proceedings of the 9th International Conference on Web Information Systems and Technologies, Aachen, Germany, 8-10 May, 2013, pages 447–452.
    Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Data Mining, Data Profiling
    Author's contribution: Co-author on methodology and implementation (architecture and user interface)
    Relates to research questions: RQ1, RQ2, RQ3
    Used in chapters: 2, 4, 4.1, 4.3
11. **Paper K**: **Softic, S.**, De Vocht, L., Mannens, E., Van de Walle, R., and Ebner, M. (2014). Finding and Exploring Commonalities between Researchers Using the ResXplorer. In P. Zaphiris, and A. Ioannou (Eds.), Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Collaboration. (1 ed., pp. 486-494). (Lecture Notes in Computer Science, Volume 8523). Springer International Publishing.
    Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Explorative Semantic Search
    Author's contribution: Main author on methodology and implementation (architecture and user interface), main contributor evaluation
    Relates to Research questions: RQ1, RQ2, RQ3
    Used in chapters: 2, 5, 5.4
12. **Paper L:** De Vocht, L., **Softic, S**., Mannens, E., Ebner, M. and Van de Walle, R. (2014), Aligning Web Collaboration Tools with Research Data for Scholars. in Proceedings of the first workshop on Big Scholarly Data (BigScholar)., pp. 1203-1210, International World Wide Web Conference, Seoul, Democratic People's Republic of Korea, 7-11 April.
    Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Data Mining, Data Profiling
    Author's contribution: Co-author on methodology and implementation and on design of evaluation
    Relates to research questions: RQ1, RQ2, RQ3
    Used in chapters: 2, 5, 5.2.2
13. **Paper M**: **Softic, S.**, De Vocht, L., Taraghi, D., Ebner, M., Mannens,

E., and Van de Walle, R. (2014). Leveraging learning analytics in a personal learning environment using linked data. BULLETIN OF THE IEEE TECHNICAL COMMITTEE ON LEARNING TECHNOLOGY, 16(4), 10–13.
Topics: User Logs, PLE, Visual Analytics, Learning Analytics, Education, Technology Enhanced Learning
Author's contribution: Main author, main contributor on concept of Learning Analytics idea and exploitation results.
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 8, 8.1

14. **Paper N:** De Vocht, L., **Softic, S.**, Dimou, A., Verborgh, R., Mannens, E., Ebner, M., and Van de Walle, R. (2015). Visualizing Collaborations and Online Social Interactions at Scientific Conferences for Scholarly Networking. in WWW 2015 Companion., pp. 1053-1054, International World Wide Web Conference, Florenz, Italy, 18-22 May.
Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Linked Data, Data Mining, Data Profiling, Visualization
Author's contribution: Co-author on methodology and implementation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 5, 5.3
15. **Paper O: Softic, S.**, Taraghi, B., Ebner, M., De Vocht, L., Mannens, E., and Van de Walle, R. (2015). Mining and Visualizing Usage of Educational Systems Using Linked Data. In Immersive Education. (1 ed., Vol. 486, pp. 17-26). (Communications in Computer and Information Science). New York, Berlin, Heidelberg: Springer.
Topics: User Logs, PLE, Visual Analytics, Learning Analytics, Technology Enhanced Learning
Author's contribution: Main author, main contributor on concept of Learning Analytics idea and exploitation results
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 7
16. **Paper P:** Salkic, S., **Softic, S.**, Taraghi, B., and Ebner, M. (2015). Linked Data Driven Visual Analytics for Tracking Learners in a PLE. In DeLFI 2015 - Die 13. E-Learning Fachtagung Informatik. (1 ed., pp. 329-331). Bonn: Köllen Druck + Verlag GmbH.
Topics: User Logs, PLE, Visual Analytics, Learning Analytics,Technology

Enhanced Learning
Author's contribution: Co-author, contributor on concept of Learning / Visual Analytics idea and exploitation results
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 8, 8.2

17. **Paper Q: Softic, S.**, De Vocht, L., Mannens, E., Ebner, M., and Van de Walle, R. (2015). COLINDA: modeling, representing and using scientific events in the web of data. 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2015), Proceedings (Vol. 1363, pp. 12–23, CEUR-WS), Co-located with the 12th Extended Semantic Web Conference (ESWC 2015), Portoroz, Slovenia.
Topics: Semantic Data Modeling, Twitter Mining, Research 2.0, Data Mining, Data Profiling
Author's contribution: Main author of the text, main contributor on methodology and implementation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 4, 4.2
18. **Paper R:** De Vocht, L., **Selver, S.**, Verborgh, R., Mannens, E., Ebner, M., and Van de Walle, R. (2015). Benchmarking the Effectiveness of Associating Chains of Links for Exploratory Semantic Search. In 4th International Workshop on Intelligent Exploration of Semantic Data (IESD 2015)., pp. 1-15, 14th International Semantic Web Conference (ISWC 2015), Bethlehem, USA, United States, 12 October.
Topics: Explorative Semantic Search, Benchmarking, Research 2.0, Data Mining, Data Profiling
Author's contribution: Co-author of the text, contributor on concept, methodology and evaluation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 5
19. **Paper S:** De Vocht, L., **Softic, S.**, Verborgh, R., Mannens, E., Ebner, M., and Van de Walle (2016). ResXplorer: Revealing Relations between Resources for Researchers in the Web of Data. Computer Science and Information Systems, ISSN: 1820-0214 (Print) 2406-1018 (Online), ComSIS Consortium, Vol. 14, No. 1, pp. 25–50.
Topics: Explorative Semantic Search, Benchmarking, Research 2.0, Data Mining, Data Profiling

Author's contribution: Main co-author of the text, main contributor on concept, methodology and evaluation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 5, 5.5, 5.3

20. **Paper T:** De Vocht, L., **Softic, S.**, Verborgh, R., Mannens, E., Ebner, M. and Van de Walle, R. (2017), Social Semantic Search: A Case Study on Web 2.0 for Science, International Journal On Semantic Web and Information Systems, Vol. 13 No. 3. (in print).
Topics: Explorative Semantic Search, Research 2.0, Data Mining, Data Profiling
Author's contribution: Main co-author of the text, main contributor on concept, methodology and evaluation
Relates to research questions: RQ1, RQ2, RQ3
Used in chapters: 2, 5, 5.5, 5.3

Thesis concept overview with relation of papers to specific field is shown in 1.2.

In order to attach the position of the particular contribution regarding the use cases and related area of research author created visualizations with the corresponding information in chapter 9.

1.7 Thesis Outline

The rest of the thesis is organized into introductory part which formalizes the motivation and research questions followed by chapter which describes related work 2 and state of the art on achievements related to the defined questions. This introductory part is followed by two blocks of chapters. The first block including chapters 3, 4 and 5 describes potentials, semantic data modeling and experiments of knowledge discovery and data mining in sparse text fragments in form of tweets for the purposes of researchers and the contribution to research. The second block of chapters 6, 7 and 8 describes potentials, semantic data modeling and experiments of knowledge discovery and data mining in sparse text fragments in form of user logs for the purposes of monitoring education processes in the Personal Learning Environment (PLE) at Graz University of Technology. Chapter describing

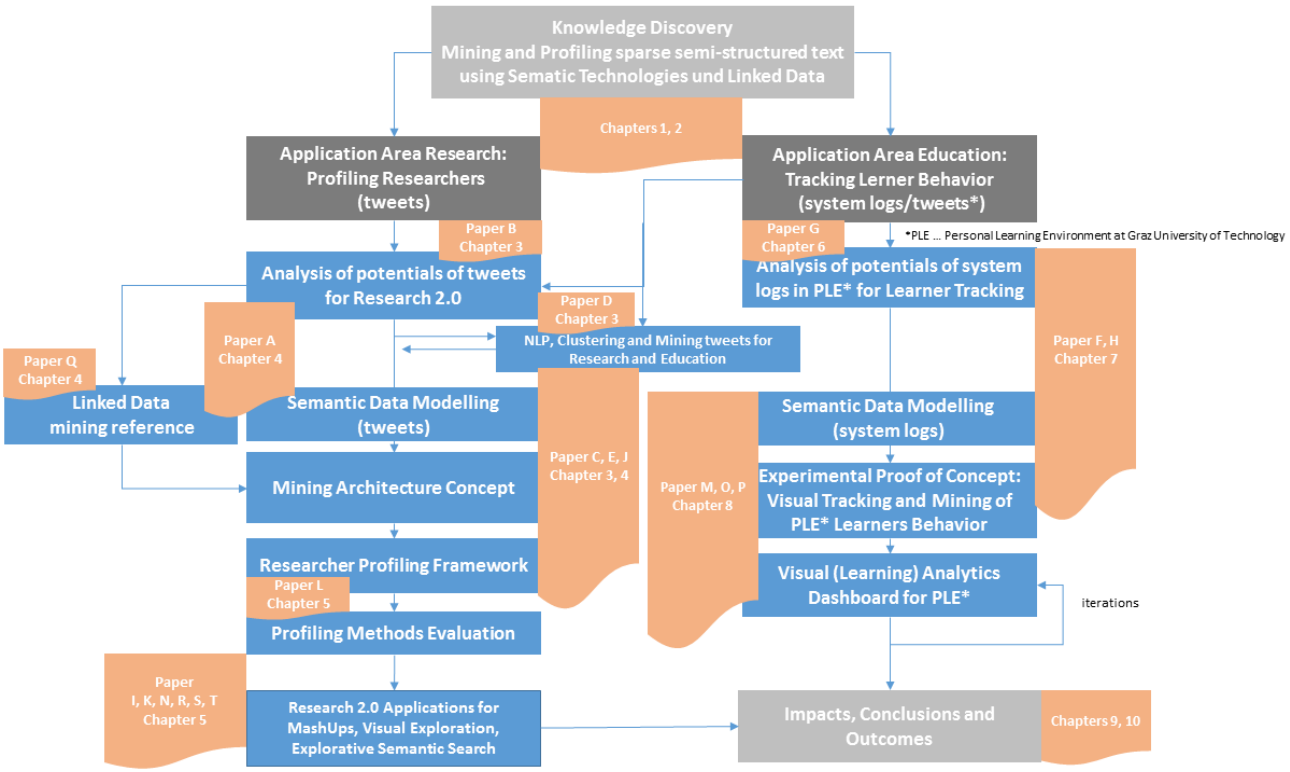


Figure 1.2: Thesis concept overview with relation of papers to specific field.

experiments in both chapter blocks reflect the direct effort trying to give an answer to the research questions stated in 1.3. This main part of the thesis is followed by the discussion 9 and retrospective how presented experiments apply to single research questions and fields. Finally, in chapter 10 conclusions are drawn and some outlook on future work is delivered. Each section in methodology part contains a subsection with "statement to own contribution" that explains how the author of this thesis was involved into preparation an publishing of particular part of presented work.

2 Related Work

Chapter 2 offers an overview over related work on areas relevant for this thesis.

2.1 Knowledge Discovery and Data Mining

Knowledge Discovery as term was firstly introduced during a KDD conference workshop in 1989 by Fayyad et al. (1996). The group of authors consider the knowledge discovery as data driven discovery process with well defined phases and steps where as output a certain knowledge is produced. According to them finding useful patterns and matching in data has different names as information harvesting, information discovery, knowledge extraction and many other. Most popular and widely accepted by statisticians, data analysts and in database field is Data Mining, which is according to Fayyad et al. (1996) considered as one of the steps in knowledge discovery where specific algorithms for extraction of patterns from data is done. Data mining as such contributes to number of fields, including retail sales, bio-informatics, stock prediction and counter-terrorism among many others. The information system society, especially in German speaking area uses the term Data Mining as equivalent to Knowledge Discovery. Knowledge discovery process in figure 2.6 presented by Fayyad et al. (1996) beside data mining contains "additional steps, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining". These steps according to Fayyad et al. (1996) "are essential to ensure that useful knowledge is derived from the data". Fayyad et al. (1996) emphasize that "blind application of data-mining methods can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns".

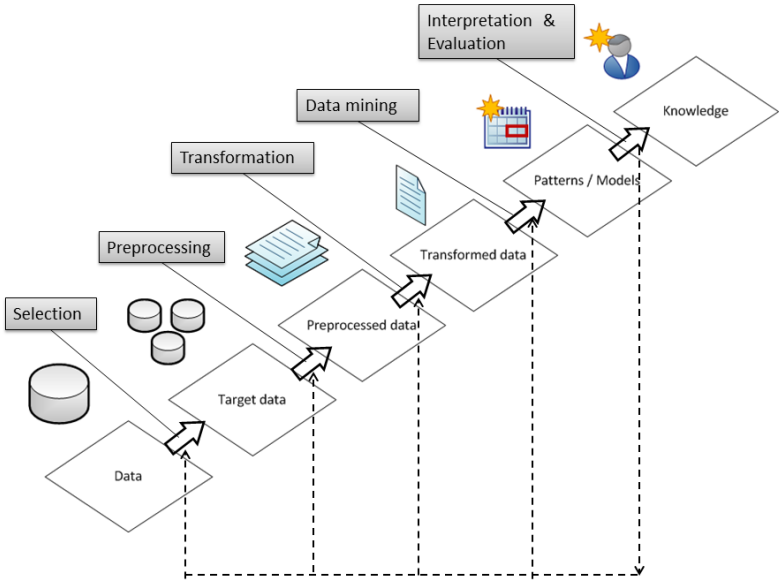


Figure 2.1: Knowledge Discovery Pipeline as described in Fayyad et al. (1996).

2.2 Educational Data Mining

In their reviews and survey on the state of the art of overview Romero and Ventura (2010, 2007), define Educational data mining (EDM) as interdisciplinary research area dealing with application of data mining and other methods to make useful discoveries from data with educational settings. Hereby EDM uses computational approaches to tackle the instrumentation of insights gathered by analysis of the educational data. According to Baker et al. (2010) methods applied in Educational data mining could be divided into following categories: Prediction,Clustering, Relationship Mining, Discovery with Models, and Distillation of data for human judgment. The first three categories are more acquainted to regular data mining while the last two are typical for Educational Data Mining. In comparison to traditional educational research methods Educational Data Mining offers alternatives such as laboratory experiments, in-vivo experiments, and design research. Through the years of continuous research four areas of application within Educational Data Mining gained special importance. The first one is im-

provement of student models. Hereby the researcher try to provide models using the detailed information on student characteristics such as motivation, knowledge, meta cognition and attitudes. Second area of application is aiming at discovery, adaptation and improvement of knowledge structure of the domain. Methodology used in this area resides on building predictive models. Third area of application resides on relationship mining and focuses on studying the pedagogical support provided by learning software. One of the methods used here is for instance learning decomposition introduced by Beck and Mostow (2008): fitting exponential learning curves to performance data, relating student success to the amount of each type of pedagogical support a student has received. The fourth area of application is tightly bound to scientific discovery about learning and learners and can also address problems from previous three areas. However, all of these areas have discovery with models in common which is a key method for scientific discovery in Educational Data Mining.

2.3 Research 2.0

Research 2.0 aims to adapt the Web 2.0 to the needs of researchers. Research 2.0 comprises interacting with information published on Social Media, and on line collaboration platforms (and other Web 2.0 tools). These tools and services, according to the specifications of Research 2.0, postulated by Parra Chico and Duval (2010); Ullmann et al. (2010), are considered as Mash-Ups, API's, publishing feeds, search and discovery services and specially designed interfaces based on social profiles. Data from such platforms in form of posts, thread, tags and user information is easily transferable into semantic form, since widely used and accepted vocabularies for these domain exist. Weaving microblogs into the Web of Data is interesting from the perspective of researcher centric semantic search and profiling. Weller et al. (2011) showed that Twitter, as exemplary microblog, can help resolving scientific citations. This thesis along with related publications introduces methods for knowledge discovery and prototypes as set of tools and services which researcher can use to discover related resources, such as persons, publications or events. Those methods and prototypes are introduced in

De Vocht et al. (2011); Thonhauser et al. (2012); Softic et al. (2013a); De Vocht et al. (2013b,c, 2014b, 2015, 2016, 2017).

2.4 Visual Analytics

According to Thomas and Cook (2005); Wong and Thomas (2004) the term "Visual Analytics" is defined as "the science of analytical reasoning facilitated by interactive visual interfaces". It is considered as combination of fields of information visualization and scientific visualization. The "Analytical Reasoning" introduced by in Thomas and Cook (2005) as such aims at applying human judgments to draw conclusions from a combination of evidence and assumptions. It offers a conceptual base for application of Visual Analytics in several cases as e.g. threat analysis, prevention, and responsiveness. Beside analytical reasoning "Visual Analytics" contributes to many other areas like data transformations and representations for computation and visualization or analytic reporting and technology transition as (see Keim et al. (2010)). Visual Analytic's research agenda unites several scientific and technical communities: computer science, information visualization, cognitive and perceptual sciences, interactive design, graphic design, and social sciences. Figure 2.2 shows the model of "Visual Analytics" process introduced by Keim. According to this model the Visual Analytics process includes key factors as: data, model, user and visualization. As result of interaction of these factors and with additional parameter refinement the users are able to gain knowledge. The key actions are taken in modeling of data and interactions and through the visualization and user as well through parameter adjustment, enabling in this way drawing of conclusions and insights supported by presented information.

2.4.1 Visualizing Relations between Researchers

Besides tracking, modeling and analyzing science related data from the Web and social media, very important approach for studying collaborations between researchers represents their visualization. Kraker et al. (2014) analyzed "the adequacy and applicability of readership statistics recorded

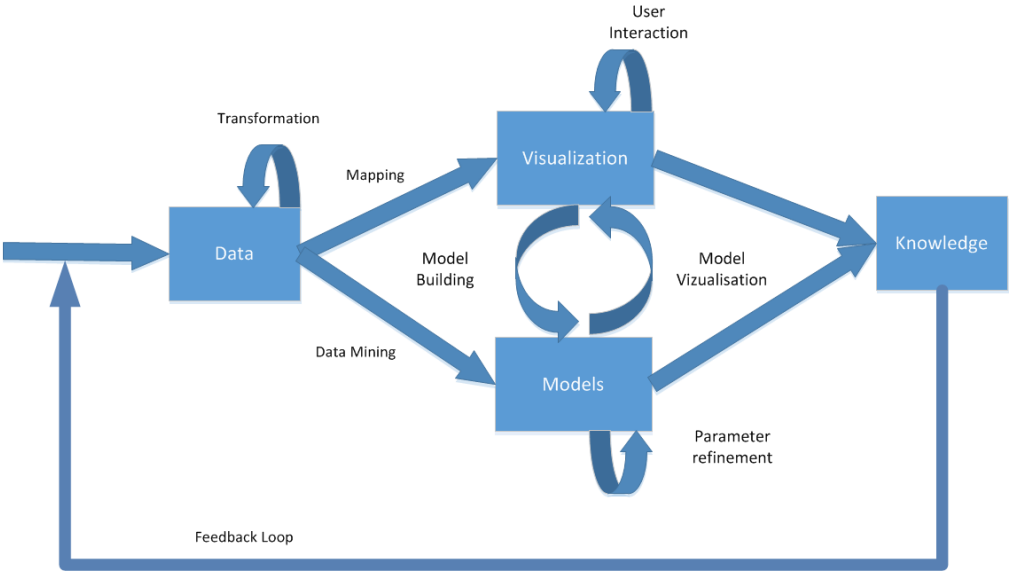


Figure 2.2: Visual Analytics concept defined by Keim et al. (2010).

in social reference management systems for creating knowledge domain visualizations". In this case study, authors used information from Mendeley[1] and Web of Science[2]. They found that visual representation allows alternative insights on underlying scientific data and found it reasonable for future works to involve more person related data. The quality however according to them depends on the setup of source data and precision of the methods used.

[1] http://www.mendeley.com/, last access: 2017–05–29
[2] http://wokinfo.com/, last access: 2017–05–29

2.5 Relevant Semi-Structured Sparse Texts for Education and Reasearch on the Web

2.5.1 Academic Social Platforms in the Web

As the number of Web 2.0 users and platforms increased, Social Media arrived. Currently, there is a number of research related publicly accessible social platforms like Mendeley[3], Researchr[4], ResearchGate[5] etc. Those platforms contain information about researchers and their publications as well as about their researcher networks. These platforms also track the value and popularity of their contributions and try to reflect the impact of their work. For researcher it is also a possibility to promote their work in the public. The content on these platforms is edited and acquired by researchers and partly by system itself by the analysis of data and external scientific web sources like publication archives. Some of these platforms like Mendeley provides an API[6] to encourage the developers to implement creative applications which use scientific materials in the form of mash ups.

2.5.2 Relevance of Twitter as Valuable Data Source

Although the beginning of first serious microblogs dates back couple of years ago their leverage on the web grows rapidly. This observation was also reported by Zhao and Rosson (2009); Boyd et al. (2010). Most significant among them is Twitter, which induced a new culture of communication as reported also by McFedries (2007); Java et al. (2007). In 2013, Twitter generated in average 500 million Tweets a day with 100 million active users daily[7]. Java et al. (2007) defined four main user behaviors why people are using Twitter - for daily chats, for conversation, for sharing information and

[3] http://www.mendeley.com/en/1/1/, last access: 2017–05–29
[4] http://researchr.org/, last access: 2017–05–29
[5] http://www.researchgate.net/, last access: 2017–05–29
[6] http://dev.mendeley.com/, last access: 2017–05–29
[7] https://tinyurl.com/mbkv9t6, last access: 2017–05–29

for reporting news. Researchers especially appreciate this development. For instance, studies on the use of microblogs like Twitter[8] conducted by Ebner et al. (2010a, 2011) within the science community has shown that researchers are using Twitter to discuss and asynchronously communicate on topics during conferences and in their everyday work. These findings are also supported by prior works by Reinhardt et al. (2009a); Letierce et al. (2010a). A survey of the use of Twitter for scientific purposes by Letierce et al. (2010a) has shown that Twitter is not only a communication medium but also reliable source of data for scientific analysis, profiling tasks and trends detection, outlined also through Tao et al. (2011); Mathioudakis and Koudas (2010); Softic et al. (2010); Bakshy et al. (2011). Twitter hashtags have a strong influence on the structuring of communication within Twitter as well as for community building (see also Laniado and Mika (2010); Bakshy et al. (2011)).

Grabeeter

At Graz University of Technology a tool called Grabeeter[9] has been implemented for storing and caching social data from Twitter. Grabeeter introduced by Mühlburger et al. (2010) is an application that allows you to search tweets of single Twitter users online and offline. In contrast to the Twitter API, Grabeeter provides all stored tweets and makes no restriction over time. The database of Grabeeter tool includes the tweets from more than 3000 users from mostly educational and research area. This tool developed by a E-Learning lab at Graz University of Technology[10] simply grabs the user timeline via the regular Twitter API[11] and stores the data. Therefore potentially every person or institution that owns a Twitter account can grab own Tweets using the Grabeeter. Tweets are stored in the Grabeeter database and on the file system as Apache Lucene[12] index. In order to ensure an efficient search tweets must be indexed. These tweets can be searched by web interface or by a JavaFX based client. Alternatively,

[8] http://www.twitter.com, last access: 2017–05–29
[9] http://grabeeter.tugraz.at/, last access: 2017–05–29
[10] elearning.tugraz.at/, last access: 2017–05–29
[11] https://dev.twitter.com/, last access: 2017–05–29
[12] https://lucene.apache.org/core/, last access: 2017–05–29

Grabeeter offers a rudimentary REST API with export possibility of timeline to XML (eXtensible Markup Language), JSON (JavaScript Object Notation) format. Grabeeter serves primarily as tweet storage. In contrast to Twitter API which allows the insights on only last 300 tweets, Grabeeter provides all stored tweets and makes no restriction over time. At the moment when experiment were conducted, Grabeeter database contained approximately 5 million tweets, which makes it a very reliable source. Grabeeter was shut down in 2013. However, the database stayed available internally as experimental source of information accessible for researchers at Graz University of Technology.

2.5.3 Digital Publication Archives

Most research publications at least in life and medical like PubMed[13], education of technical sciences e.g. DBLP (Digital Bibliography & Library Project)[14] are available via the Web. Many digital libraries and scientific online journals offer access to their content. Usually a paid membership is needed to get full access to all articles, but most of the educational institutions can afford this kind of service. At the same time a growing number of "Open Journals" offer free access to all published works. Most prominent archives in this area are Directory Of Open Access Journals (DOAJ)[15] as well as Online Journals[16]. Crawling pre-processing and bringing such content into contextual form brings benefits for the researchers. Same as in the case of other researcher platforms a set of publicly available APIs[17] is availabe for mining the additional information for the researchers.

[13] http://www.ncbi.nlm.nih.gov/pubmed, last access: 2017–05–29
[14] http://dblp.uni-trier.de/, last access: 2017–05–29
[15] http://www.doaj.org/, last access: 2017–05–29
[16] http://online-journals.org/, last access: 2017–05–29
[17] http://libguides.mit.edu/apis, last access: 2017–05–29

2.5.4 PLE at Graz University of Technology

The main idea of PLE at Graz University of Technology is to integrate existing university services and resources with services and resources from the World Wide Web in one platform and in a personalized way (see Taraghi et al. (2010); Ebner and Taraghi (2010)). The TU Graz PLE contains widgets (see Taraghi et al. (2009, 2010); Ebner and Taraghi (2010)) that represent the resources and services integrated from the World Wide Web. Web today provides lots of different services; each can be used as supplement for teaching and learning. The PLE has been redesigned in 2011, using metaphors such as apps and spaces for a better learner-centered application and higher attractiveness. This was explicitly pointed in Ebner et al. (2010b); Taraghi et al. (2012). In order to enhance PLE in general and to improve the usability as well as usefulness of each individual widget a tracking module was implemented (see also Taraghi et al. (2011)).

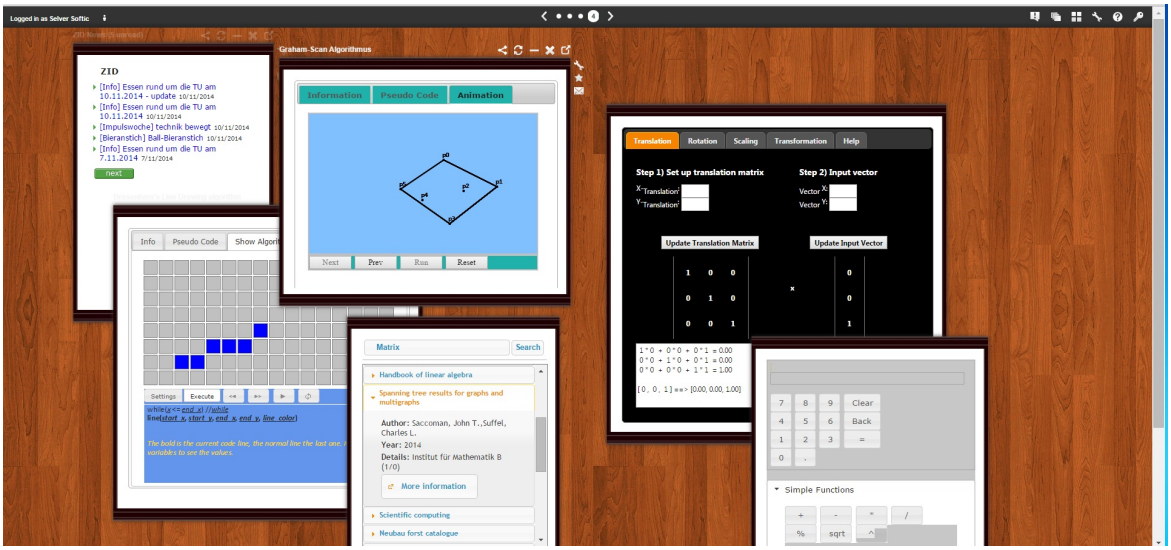


Figure 2.3: PLE at Graz University of Technology.

2.6 Web of Data and Semantic Technologies

The Semantic Web envisioned by Tim Berners-Lee in Berners-Lee et al. (2001) extends the Web with functionality by bringing structure and giving well-defined meaning to the content and it enables humans and machines to work together using controlled vocabularies. By machines are meant software agents which carry out sophisticated tasks such as intelligent search. Semantic Web Community provides a set of widely used schema (also known as vocabularies) useful to cover the description of user generated content and user profiles attached to them.

2.6.1 The architecture of the Semantic Web

Before understanding the essence of Semantic Web it is necessary to deal with its structure. The architecture of Semantic Web is called Semantic Web Stack[18] depicted in Figure 2.4. The Semantic Web Stack illustrates general hierarchy of necessary languages, standards and technologies involved into the construction of Web of Data vision. Hereby each layer uses the capabilities of underlying layer and provides some capabilities to the layers above. The idea of Semantic Web Stack was initially introduced in by Sir Tim Berners-Lee in Berners-Lee (1996); Berners-Lee et al. (2001) and refined considering reasoning and description logic by Ian Horrocks and a group of researches around him in Horrocks et al. (2005).

2.6.2 RDF

RDF[19] (Resource Description Framework) is a markup language for presenting and interchange of information and resources on the World Wide Web. Resource Description Framework (RDF) serves as foundation for processing meta data with semantics in general and extends the linking structure of the Web. It provides interoperability between applications that exchange information on the Web. In this way, machines can understand

[18] https://tinyurl.com/hdlpozp, last access: 2017–05–29
[19] http://www.w3.org/RDF/, last access: 2017–05–29

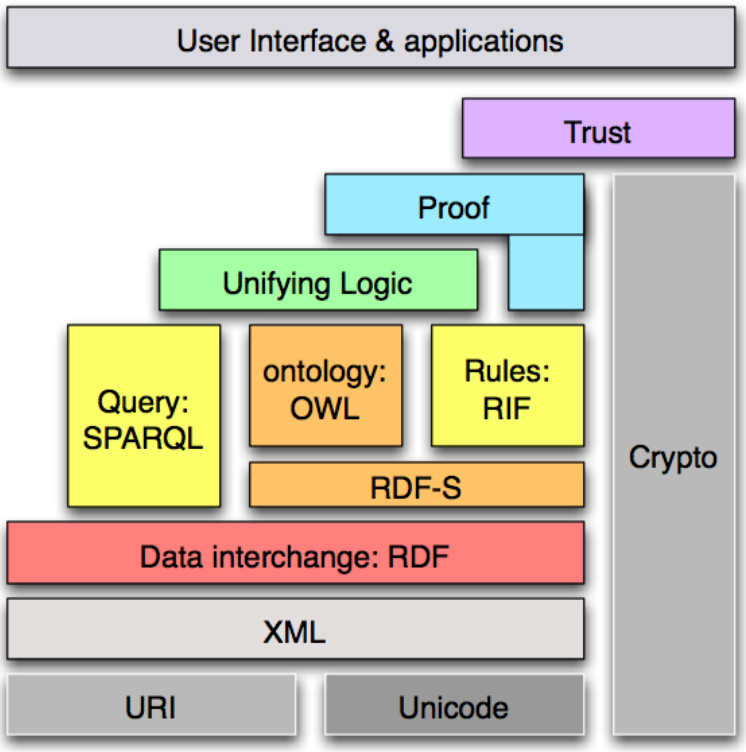


Figure 2.4: Semantic Web Stack as adapted from Horrocks et al. (2005).

each other and no classical web services are needed. With other words RDF emphasizes facilities to enable automated processing of Web resources. The RDF can be also imagined as a directed, labeled graph data format for representing information in the Web. Areas of appliances are manifold. For example, it can be used in resource discovery to provide better search engine capabilities, in cataloging for describing the content and content relationships available at a particular Web site. Further, it can be useful for digital libraries, in social networks and on-line communities e.g to resolve the social relations of community members, by intelligent software agents to facilitate knowledge sharing and exchange, in content rating, in describing collections of pages that represent a single logical 'document' or digital artifact, for representation of personal information, describing intellectual property rights of Web pages, expressing and privacy preferences of a user

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://memyselfandi.com/selvers> dc:publisher <http://memyselfandi.com/selvers/about>;
                                  dc:title "About Selver Softic" .
```

Listing 2.1: Sample RDF.

as well as the privacy policies of a Web site etc. In the future, RDF with digital signatures is proposed to be the key technology for building the "Web of Trust" for e-commerce, collaboration and administration tasks, and other comparable applications. The numerous examples presented altogether show that the RDF can be considered as the base data model of the Web of Data with precisely defined formal semantics.
The base of the RDF model is built upon 'triples'. A simple RDF triple consist from Subject, Predicate or Property and the Object. The Subject must be always a valid URI (Uniform Resource Identifier), the Predicate is a property of distant vocabulary and Object can be either a Literal or a valid URI as depicted in listing 2.1. Also so called blank nodes (e.g. "_bNode") are allowed to be used for Subjects or Objects instead of URIs. Description of the triples is usually done in specific notations like: RDF/XML a XML based representation, N3 notation (designed by Sir Tim Berners-Lee) and also very often used Turtle notation. N3 or Turtle are shorthand non-XML serializations designed with the aim of easier human-readability while RDF/XML was designated for machines. There is also a graph based representation of triples recommended by the RDF developer group at W3C that should contribute that humans can read and understand the triples more easier.

In the listing 2.1 two triples containing literal and URI as object are presented in order to demonstrate the advantages of the N3 notation regards the usage as a format that is easier to read for humans. Listing 2.2 above shows the same example represented in RDF/XML form. As already mentioned before, RDF/XML format was designed primary for so called Semantic Web agents.

Looking at the graph in Figure 2.5 or at the N3 Notation in listing 2.1 it is much easier to recognize the meaning of the two triples and relations between them described in this context than by just using the RDF/XML representation in listing 2.2. Those two triples say that resource

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://memyselfandi.com/selvers">
    <dc:publisher>http://memyselfandi.com/selvers/about</dc:publisher>
    <dc:title>About Selver Softic</dc:title>
  </rdf:Description>
</rdf:RDF>
```

Listing 2.2: Sample RDF in RDF/XML form.

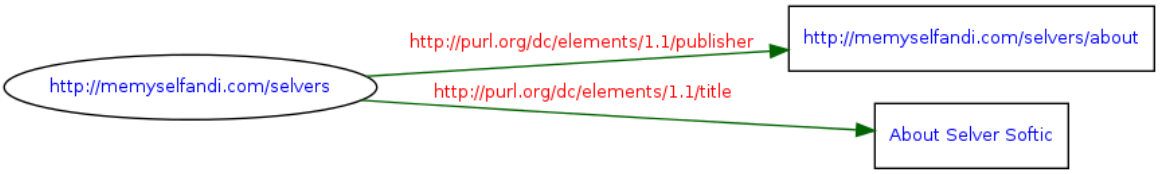


Figure 2.5: RDF graph.

"http://memyselfandi.com/selvers" was published by somebody having a on-line profile at "http://memyselfandi.com/selvers/about" and that it carries the title "About Selver Softic". Presented examples above outlines the essence of triple based data instance generation in the world of Semantic Web, which is necessary for elementary understanding of the concept of Semantic Web Architecture.

2.6.3 OWL

OWL (Web Ontology Language)[20] is a language for defining ontologies. Ontology describes knowledge in a explicit way using concepts and simple logical relations between them. OWL unites actually a family of knowledge representation languages and can be also considered as extension of RDF and RDF Schema which allows to clearly specify restrictions considering specific knowledge domain. Using these more specified restrictions OWL increases the decidability in the realm of a certain knowledge base. It can be used by humans or software e.g. intelligent agents, to share information about certain objects, occurrences and the like. There can be for example an

[20] http://www.w3.org/TR/owl-features/, last access: 2017–05–29

ontology about cars, books, products, services, finances, etc. Primary assignment of OWL is processing instead of displaying the information[21]. With other words, OWL opens a gate for adoption of artificial intelligence into the concept of Semantic Web. Machines or humans should be able to get an answer on a search quest considering a certain knowledge domain. Hereby OWL uses so called Open World Assumption in contrast to technologies like SQL or Prolog with Closed World Assumption premises. Under Open World Assumption a statement can not be interpreted automatically as false if it was not proven true using the current knowledge. Only certain thing according the Open World Assumption in this case is that on the specific statement no further conclusions can be drawn.

Historical Backgrounds

Ontologies as such have a long history dating far before the modern computer science.They have roots originally in philosophy. At the beginning of nineties of last century, a number of research efforts were made on exploration of the idea how the representation of knowledge for artificial intelligence can contribute to the development of the World Wide Web. As base of this research served HTML based language SHOE and XML based XOL (later known as OIL). Today's OWL is revised version of DAML+OIL web language. DAML+OIL was developed by a special group named "US/UK ad hoc JointWorking Group on Agent Markup Languages" founded by US Defense Research Agency (DARPA) and EU's IST funding project. In the year 2001 the W3C Consortium started the so called "Web Ontology Working Group". First concepts, synopsis, reference and abstract syntax resulted then already in the middle of 2002. Finally, on February 10th, 2004 OWL became an W3C (World Wide Web Consortium) recommendation. The name OWL originate from switching the first and second place of letters from Web Ontology Language acronym. This was originally suggested by Tim Finin in his E-Mail to W3C Web Ontology Mailing list in December 2001[22].

[21] http://www.w3.org/TR/owl-semantics/, last access: 2017–05–29
[22] https://tinyurl.com/zqxg36a, last access: 2017–05–29

```
owl:class
owl:oneOf
owl:unionOf
owl:intersectionOf
```

Listing 2.3: Some class related constructs.

OWL, Language Elements and Constructs

An OWL ontology can be considered as set of individuals and properties. Using the "properties" allows the expression of relation between the "individuals". A set of axioms places constraints on sets of so called individuals (also called classes) and the types of relationships permitted between them. Using the axioms an additional implicit semantics is obtained. This can be provided by inferring from the explicitly provided data. OWL provides three different languages each designed for specific area of use:

- OWL Lite
- OWL DL
- OWL Full

OWL Lite serves those users who need only a classification hierarchy and some simple constraints. It basically supports cardinality but permits only cardinality values of 0 and 1. OWL has also a lower formal complexity then OWL DL and requires simpler tools. OWL DL delivers maximum expressiveness while retaining computational completeness and decidability. OWL DL guarantees for all computable conclusions that all computations will terminate in a finite time. Although OWL language constructs are included in the bundle, they can be used only under certain restrictions. The name OWL DL was given due to its correspondence with description logic, the logic that provided formal foundation of OWL. OWL Full beside maximum expressiveness allows also the syntactic freedom of RDF. Nevertheless, in this case no computational guarantees are offered. OWL Full implements the treatment of a class either as a collection of individuals or as an single individual on its own. It also allows an ontology an enhancement of meaning of the pre-defined (RDF or OWL) vocabulary.

Some class-related constructs: Constructs presented in the listing 2.3 are commonly used for definitions or as axioms, boolean combinations of class

```
owl:Restriction
owl:allValuesFrom
owl:someValuesFrom
```

Listing 2.4: Some properties related constructs.

```
owl:sameAs
```

Listing 2.5: Some instance related constructs.

expressions. Some properties-related constructs: In the listing 2.4 property-related constructs foreseen for restrictions are depicted. OWL contains also instance-related constructs: Listing 2.5 refers to very often used instance-related construct expressing the equality matters regarding instances.

2.6.4 RDFS

RDF-S, RDFS or RDF Schema like RDF is a language for description of vocabularies on the Web. It can be considered as semantic extension of RDF. RDF itself knows only about properties as attributes of resources that additionally can represent relations between them and offers no possibility for their description and definition neither for definitions of relations among them. This restriction of RDF motivated the creation of RDF Schema. The RDF Schema provides classes and properties that allow description of classes, properties and other resources. In this way describing groups of related resources and the relationships between these resources is made possible. The vocabulary descriptions of RDF Schema are written in RDF. This enables the resources to determine characteristics of other resources, such as the domains and ranges of their properties. The first version was published by W3C in April 1998. Finally, on February 10th 2004 RDFS became W3C recommendation[23]. In listing 2.6 we can see how easily a class of type "Person" can be defined in RDFS containing a special subclass of type "Student".

[23] http://www.w3.org/TR/2004/REC-rdf-schema-20040210/, last access: 2017–05–29

```
@prefix myns: <#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

myns:Person a rdfs:Class;
 rdfs:comment "Class describing general person";
 rdfs:subClassOf rdf:Resource .

myns:Student a rdfs:Class;
 rdfs:comment "Class describing a special person subclass student";
 rdfs:subClassOf myns:Person .
```

Listing 2.6: Example in RDFS.

2.6.5 SPARQL

SPARQL (pronounced as "sparkle")[24] is an RDF query language and protocol. Its name stands for Simple Protocol and RDF Query Language. It is standardized by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium. Initially it was released as a Candidate Recommendation in April 2006, and on 15th January 2008, SPARQL became an official W3C Recommendation. SPARQL allows that a query can consist from triple patterns, conjunctions, disjunctions, and optional patterns. Currentl,y it is a very common practice to expose RDF data instances using so called SPARQL Endpoints. Those are web interfaces allowing remote firing of queries on underlying RDF graphs using adequate connectors.

SPARQL has four query forms in its specification[25]. These query forms use pattern matching to form result sets or RDF graphs. The query forms are:

- SELECT Returns all, or a subset of, the variables bound in a query pattern match.
- CONSTRUCT Returns an RDF graph constructed by substituting variables in a set of triple templates.
- ASK Returns a boolean indicating whether a query pattern matches or not.
- DESCRIBE Returns an RDF graph that describes the resources found.

The query depicted in listing 4.5 retrieves names and optionally age of all team players and the names of teams as well they play for assumed their

[24] http://www.w3.org/TR/rdf-sparql-query/, last access: 2017–05–29
[25] http://www.w3.org/TR/rdf-sparql-query/#QueryForms, last access: 2017–05–29

```
PREFIX player: <http://example.com/playerOntology#>
PREFIX team: <http://example.com/teamOntology#>
SELECT ?name ?team ?age
WHERE {
?x player:name ?name.
OPTIONAL{player:age ?age}
?y team:name ?team.
?x player:playsFor ?y.
?y team:wasFoundedIn '1954'^xsd:int.
}
```

Listing 2.7: Sample SPARQL query.

team is founded in the year 1954. As can been seen already this simple example demonstrates the complexity and the power that SPARQL obtains. SPARQL became today mostly used query language for RDF graphs (also recommended by W3C).

2.6.6 R2RML

R2RML[26] is a W3C recommendation since September 27th 2012 and represents a language for customized mappings of content from relational databases into RDF. Hereby,each mapping is individual and depends on database schema and target vocabulary. R2RML refers logical tables such as: base table, view or SQL query to obtain the input data for mapping. Each logical tables is mapped using the so called **'triples map'**. A 'triples map' represents a rule that maps each row in logical table. This 'rule' contains two main parts: a **'subject map'** and **'multiple predicate-objects maps'**. A simple R2RML mapping is shown in listing 2.8 where a simple logical table STUDENT (see 2.1) with columns STDNO (student number), STDNAME (student name) is mapped via 'triples map' into RDF shown in listing 2.9.

STUDENT	
STDNO	STDNAME
1234	SOFTIC

Table 2.1: Sample logical table STUDENT.

[26] https://www.w3.org/TR/r2rml/, last access: 2017–05–29

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix ex: <http://example.com/ns#>.

<#TriplesMap1>
    rr:logicalTable [ rr:tableName "STUDENT" ];
    rr:subjectMap [
        rr:template "http://data.example.com/student/{STDNO}";
        rr:class ex:Student;
    ];
    rr:predicateObjectMap [
        rr:predicate ex:name;
        rr:objectMap [ rr:column "STDNAME" ];
    ].
```

Listing 2.8: Sample R2RMLmapping for the STUDENT logic table.

```
<http://data.example.com/student/1234> rdf:type ex:Student.
<http://data.example.com/student/1234> ex:name "SOFTIC".
```

Listing 2.9: Triples generated through R2RML mapping for the STUDENT logic table.

2.6.7 RML

The RDF Mapping language (RML)[27] is a generic mapping language defined to express customized mapping rules from heterogeneous data structures and serializations to the RDF data model. According to Dimou et al. (2014b) RML is meant as a superset of the W3C standardized mapping language R2RML. It is aiming to extend its applicability to a broader set of input sources such as CSV (Comma Separated Values), JSON and XML.

2.6.8 Widely used Ontologies and Vocabularies

This subsection introduces two most relevant vocabularies for describing user related web content and scientific events.

FOAF

Main objective of Friend of a Friend (FOAF)[28] project is creation of a machine-readable content on the Web describing people, their activities,

[27] http://rml.io/, last access: 2017–05–29

[28] http://www.foaf-project.org/, last access: 2017–05–29

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

:me a foaf:Person;
foaf:homepage <http://selsofti.org/>;
foaf:img </images/me.jpg>;
foaf:mbox_sha1sum "241021fb0e6289f92815fc210f9e9137262c252e";
foaf:name "Selver Softic" .
```

Listing 2.10: Sample FOAF data.

relations and the links between them. It is a simple vocabulary for sharing and transferring of information like contacts, photos, calendars, links to web blogs, etc Brickley and Miller (2004). FOAF enables the users to merge and re-use this sources on-line. FOAF project exists since year 2000.

SIOC

Semantically-Interlinked On-line Communities or SIOC[29] (pronounced "shock" as old Celtic word for ice) represent a framework consisting of a core ontology and a number of tools that enable connecting online community sites and diverse forms of internet-based discussions. On-line communities like message boards, web blogs, discussion forums and the like contain valuable information but this information is isolated as separate "data island" in so called "walled-garden". Nevertheless, this information could be more valuable if it were interconnected. SIOC approach to online communities description allows to interlink these "islands", and enables the extraction of richer information from various discussion web sources. The vocabulary of SIOC Core Ontology provides the main concepts and properties required to describe information from on-line communities (e.g., message boards, wikis, web blogs, etc.) on the Semantic Web (see Berrueta et al. (2007)). Online community sites can provide in this way information about their structure and contents to the outside world making this information readable and understandable for machines. This information can be used by tools that understand SIOC data to suggest related information from other community sites what offers the possibility of structured and interlinked overview over distributed conversations across blogs, forums and mailing lists. It can be

[29] http://sioc-project.org/, last access: 2017–05–29

```
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://example.org/posts/post?id=1> a :Post;
dcterms:created "2006-09-07T09:33:30Z";
dcterms:title "Sample SIOC post";
 :content "This is a sample SIOC post demo.";
 :has_container <http://example.org/posts/>;
 :has_creator <http://selversoftic.com/contact/>;
 :has_reply <http://example.org/posts/post?id=23>;
 :topic [
 rdfs:label "Semantic Web" ],
  [rdfs:label "SIOC" ],
  <http://example.org/posts/category/SIOC/>,
  <http://example.org/posts/category/semantic-web/>
  ] .
```

Listing 2.11: Sample of SIOC representation of a post.

also used as enhanced export/import format, with access to either the entire content or summaries. Besides this aspect it offers the implicitly ability of publishing and subscribing to decentralized discussion channels and communities. SIOC is a highly sophisticated vocabulary. It is usually often combined with FOAF and Dublin Core vocabulary offering very efficient description of on line communities. Such approaches are introduced by Bojars et al. (2008a,b) in the past. Exposing data in SIOC offers structure and enhanced semantics with explicit flexibility considering manipulation of data. Since this thesis is aiming at analysis of user generated content SIOC seems to be the right choice for structuring the data for further processing. The SIOC project was started by John G. Breslin and Uldis Bojars at DERI, NUI (National University of Ireland), Galway in year 2004. SIOC became a W3C Member Submission in 2007.

The listing 2.11 represents a sample RDFised blog post in a already introduced N3 notation using the SIOC vocabulary combined with concepts from Dublin Core consisting from post date, post creator information, post title, content and categorization description using the topic keywords.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@base <http://example.org/ns/> .

<A> rdf:type skos:Concept ;
  skos:prefLabel "love"@en ;
  skos:inScheme <MyScheme> .

<B> rdf:type skos:Concept ;
  skos:prefLabel "adoration"@en ;
  skos:inScheme <AnotherScheme> .

<A> owl:sameAs <B> .

ex3:milkBySourceAnimal rdf:type skos:Concept;
   skos:prefLabel "milk by source animal"@en;
   skos:broader ex3:milk;
   skos:narrower ex3:cowMilk;
   skos:narrower ex3:goatMilk;
   skos:narrower ex3:buffaloMilk.
```

Listing 2.12: Sample SKOS representation.

SKOS

SKOS[30] stands for Simple Knowledge Organisation System and it is widely spread vocabulary that was developed to support the organization of knowledge in Semantic Web in form of RDF. The classes and properties of this vocabulary enable construction of concept systems as thesauri, indexes and taxonomies in semantic form which can be exchanged and passed between the computer applications in an interoperable way.

SWRC

The "Semantic Web for Research Communities" (SWRC) ontology Sure et al. (2005), is used to represent knowledge about researchers and research communities. It models key entities relevant for typical research communities and the relations between them. Top level concepts used by the SWRC ontology are: the Person, Publication, Event, Organization, Topic and Project concepts. Additionally the ontology offers a possibility to annotate each of the concepts with labels which is especially useful for text- and pattern-based matching in the mining process.

[30] https://www.w3.org/2004/02/skos/, last access: 2017–05–29

Dublin Core

Dublin Core is a bibliographic data format. It is a set of small vocabularies. They are used to describe web and physical resources such as web pages, online documents, books and other electronic media. Dublin Core as such was initiated by the Dublin Core Metadata Initiative (DCMI) an open forum for metadata standards founded 1994 by Metadata Workshop at World Wide Web conference held in Chicago. The name "Dublin" refers to Dublin, Ohio, USA where the schema was originally introduced. The schema was introduced by Online Computer Library Center (OCLC), a library consortium based in Dublin, and the National Center for Supercomputing Applications (NCSA). Original Dublin Core Metadata Set includes 15 elements. They have been extended over the time by DCMI[31] .

2.7 Linked Data

The foundational data layer of the Semantic Web is considered to be Linked Data[32]. Linked Data consists from Resource Description Framework (RDF)[33] graph based representation of data instances retrievable via the SPARQL[34] W3C standard. The idea of Linked Data Berners-Lee (2006) and growing activity in this field accumulated into LOD Cloud (Link Open Data Cloud). Currently the Linked Open Data (LOD) Cloud[35] has reached a respectable size[36,37]. LOD Cloud has billions of triples and includes meanwhile reliable semantic data sets like e.g. DBPedia Auer et al. (2008), which is semantic version of Wikipedia as reasonable mapping source as many others. DBPedia offers SPARQL endpoints and lookup services. Beside DBPedia there are some other verified Linked Data sets with endpoints in the LOD

[31] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms, last access: 2017–05–29
[32] http://linkeddata.org/, last access: 2017–05–29
[33] http://www.w3.org/RDF/, last access: 2017–05–29
[34] http://www.w3.org/TR/rdf-sparql-query/, last access: 2017–05–29
[35] http://lod-cloud.net/, last access: 2017–05–29
[36] http://lod-cloud.net/state/, last access: 2017–05–29
[37] http://stats.lod2.eu/, last access: 2017–05–29

Cloud e.g . for identifying people (FOAF.O-Sphere), locations Geonames[38], scientific events COLINDA[39] etc. Linked data services like these offer a solid base for enrichment of unstructured content as shown in De Vocht et al. (2011). Linking semantic sources using simple principles described in Berners-Lee (2006); Bizer and Cyganiak (2006); Bizer et al. (2008, 2009) turns the web into large database not only available for human but also to intelligent agents. Bringing Twitter data into this infrastructure would increase the relevance of content and offer more profound information on social aspects of contributed content. Related work on this topic has been published by Tao et al. (2011).

2.7.1 Tools for Publishing and Interlinking the Linked Data

Currently, there are several tools for publishing and interlinking data: RDF Refine[40], KARMA[41] and Silk Framework[42]. RDF Refine is a a Google Refine extension for exporting RDF, which can reconcile and interlink data within the application. Further, reconciliation against SPARQL endpoints and RDF dumps as well as search within the Web for related RDF datasets is supported. RDF Refine has an export function to RDF with a GUI (Graphic User Interface) for defining the shape of the RDF graph including support for own vocabularies or existing ones. During the shaping of the RDF Graph auto complete function for property and class names resolution is provided. KARMA[43] introduced by Knoblock et al. (2011) is a tool for information integration provided by the Information Sciences Institute at the University of Southern California, which enables users to integrate data from different sources: databases, spreadsheets, delimited text files, XML (eXtensible Markup Language), JSON (JavaScript Object Notation), KML (Keyhole Markup Language), and Web APIs (Application Programmable Interfaces). The tool supports a user to integrate the information by modeling it according to the target ontology of his choice. The whole interaction can

[38] http://geonames.org, last access: 2017–05–29
[39] http://colinda.org, last access: 2017–05–29
[40] http://refine.deri.ie/, last access: 2017–05–29
[41] http://www.isi.edu/integration/karma/, last access: 2017–05–29
[42] http://silkframework.org/, last access: 2017–05–29
[43] http://www.isi.edu/integration/karma/, last access: 2017–05–29

be done in a user interface. KARMA recognizes further mappings within the process and makes proposals for a model that ties the classes together. Data from KARMA can be transferred to various formats for adjustment. Once the model is fixed and data is integrated, it can be published to a RDF data store or database. The survey made by Wölger et al. (2011) from STI Innsbruck on data interlinking methods claims that none of the currently existent and evaluated frameworks and tools can be used as general purpose interlinking tool. One of the reasons for this claim, lies according to the authors, on the trade-off which must be made regarding granularity and flexibility. In their work published in Volz et al. (2009) the authors of Silk introduce an approach which closely fits to this demand. Silk is a framework developed by the Free University of Berlin for creating interlinkage between domain specific RDF data sets. The Silk framework uses a declarative language to search for the links within the mapping data sets. It aims at discovering and creating *owl:sameAs* links using the domain specific rules and parameters. In this way publishers can integrate the discovered connections between the links into their export in order to enable the data consumer. The Silk framework uses an XML based Silk-LSL language to specify the rules for linking the resources. The Silk framework has been successfully used to resolve connections between the drugs repositories for the application of life sciences as well as to enrich a sparse data with additional meta data. Publishing quality Linked Data requires adequate mappings tools and approaches to handle the issues of data cleaning, reconciliation, and data transformation. D2R Server[44] introduced by Bizer and Cyganiak (2006) is a tool for publishing the content from relational databases to RDF. Hereby the D2R Server uses a customizable D2RQ mapping language[45] to weave the database content into this format, and it allows the RDF data to be browsed and searched. This on-the-fly translation allows the publishing of RDF from large live databases and eliminates the need for replicating the data into a dedicated RDF triple store. It also supports content negotiation, SPARQL endpoint, and CLOB (Character Large Objects) / BLOB (Binary Large Objects) dumps as well. TheDataTank[46] Vander Sande et al. (2012) is a RESTful (REST - Representational state transfer) data management system.

[44] http://d2rq.org/d2r-serve, last access: 2017–05–29
[45] http://d2rq.org/d2rq-language, last access: 2017–05–29
[46] http://thedatatank.com/, last access: 2017–05–29

It was initially built to become a datahub for any organisation that wanted to publish Open Data. But it also can serve as an adapter for formats like CSV (comma separated values) / XLS (Microsoft Spreadsheets) / XML / JSON / SQL files to supply a RESTful interface using generic resources in *tdt/core* and publish 5 star data. This software is managed by OFKN Belgium. It has a tool to convert structured data to linked data and expose it as a service using a mapping file.

2.7.2 Linked Data for Research

The efforts to make sharing scientific resources a reality occupied researchers in science and educational informational systems for a long time. The outcome of such quests lead to an increasing variety of heterogeneous technologies, schemas, repositories and query mechanisms. This trend brings with it a constant growing amount of publicly available Linked Data including scientific repositories. Within the research community commercial digital libraries like Association for Computer Machinery Digital Library[47] and Springer[48] started to publish their archives in the LOD Cloud providing, in this special case, more than tens of millions of triples. Parallel to the commercial scientific content providers some academic institutions as well as the most famous public libraries, such as Library of Congress[49], British National Library[50] and Bibliothèque Nationale de France[51], provided their public Linked Data. Besides the initiative of big digital and national libraries, the efforts made by the scientific community like bootstrapping the eScience assets from the Open Archives Initiative - Object Reuse and Exchange (OAI-ORE) project[52] into the Web of Data are worth mentioning. According to the LOD Cloud stats[53] publication repositories are the most numerous. Around 10% of the overall distribution of triples comes from the research publication repositories and publications are the source of around 30% of the overall

[47] http://acm.rkbexplorer.com/, last access: 2017–05–29
[48] http://lod.springer.com, last access: 2017–05–29
[49] http://id.loc.gov, last access: 2017–05–29
[50] http://bnb.data.bl.uk, last access: 2017–05–29
[51] http://data.bnf.fr, last access: 2017–05–29
[52] https://www.openarchives.org/ore/, last access: 2017–05–29
[53] http://stats.lod2.eu/, last access: 2017–05–29

links distribution[54].Information present within the LOD Cloud offers a solid base of re-usable information to weave the Web and adapt information for researchers and scientists.

2.7.3 Interfaces for Research based on Linked Data

A review on related search interfaces for science despite the huge amount of published Linked Data especially publications meta data leads us to a few working solutions worth mentioning. One of them is *RKB Explorer*[55] introduced by Glaser et al. (2008) which is a visual browser originated from the ReSIST[56] network of excellence. It includes many sources of scientific data. The visual browsing interface is based on categorized pre-selection which focuses on people, organizations, publications and courses and materials. The view always focuses on the selected category which makes the context based browsing less flexible but focused. Within the visualization RKB Explorer evaluates relationships of the first degree. In comparison to RKB Explorer approach presented in my thesis is rather user and search centric than concept and context centric. In the interface introduced as my common work with Ghent University user profile affects the pre-selection of search results. Users can configure the search context by executing searches for resources or by expanding one or more resources. Another advanced research related effort is *Faceted DBLP search*[57] which resulted from two European research projects KnowledgeWeb[58] and ViKEF[59] and maintained by L3S Research Center, Leibniz University of Hannover. The search approach in this case resides on DBLP++ data set which enhances DBLP and on additional keywords and abstracts available on the public web pages. It integrates facets on time, venues, publications years and authors and delivers the results in various formats like BibTeX, regular web pages, as DOI - Document Object Identifiers or in RDF format. Faceted DBLP offers a good flexibility on filtering and narrowing down the results as well

[54] http://lod-cloud.net/state/, last access: 2017–05–29
[55] http://www.rkbexplorer.com, last access: 2017–05–29
[56] http://www.resist-noe.org/, last access: 2017–05–29
[57] http://dblp.l3s.de/, last access: 2017–05–29
[58] http://knowledgeweb.semanticweb.org/, last access: 2017–05–29
[59] http://www.vikef.net/, last access: 2017–05–29

implements basic syntactic query expansion based upon single word and whole phrase, but still in an anonymous way. Retrieval is done by classical search engines and result selection is done by ranking without any possible relation to the user profile. BibBase[60] introduced by Xin et al. (2013) has an interface to leverage the personal publications into the Web of Data and integrates the retrieval of author publications from Mendeley[61], DBLP[62] and Zotero[63].

2.7.4 Linked Data for Education

Research on interoperability of technology-enhanced learning (TEL) repositories throughout the last decade led to a fragmented landscape of competing approaches, such as metadata schemas and interface mechanisms. However, so far web-scale integration of resources is not facilitated, mainly due to the lack of existence of shared principles, datasets and schemas. On the other hand, the Linked Data approach has emerged as the de-facto standard for sharing data on the Web and offers a large potential to solve interoperability issues in the field of TEL. This achievement relies on establishing principles that support sharing of large datasets on the Web together using the defined technology stack. Technology stack is simple and based upon URIs, RDF, and SPARQL. The initiative has lead to success and widespread Liked Data efforts which accumulated in huge amounts of public data such as DBPedia, WordNet RDF or similar. The Linked Data movement also supports the exposure of large amounts of reusable data and resources into Linked Data Cloud, a network of interlinked Linked Data sets[64]. The nature of data involved is ranging from domain specific expert knowledge up to data about cultural heritage like e.g. the Europeana dataset. Recently, the notion about these approaches is getting more adopted and accepted by education institutions. Within this realm, Linked Data technologies are being used to expose public information regarding: course

[60] http://bibbase.org, last access: 2017–05–29
[61] http://www.mendeley.com/, last access: 2017–05–29
[62] http://www.informatik.uni-trier.de/~ley/db/, last access: 2017–05–29
[63] http://zotero.org, last access: 2017–05–29
[64] http://richard.cyganiak.de/2007/10/lod/, last access: 2017–05–29

offering, educational resources and facilities. This lead to the creation of an sub initiative named "Web of Educational Data" including institutions such as the Open University (UK) or the National Research Council (Italy), as well as Linked Data about publicly available educational resources, such as the mEducator – Linked Educational Resources (see also Dietze et al. (2012, 2013)).

2.7.5 Semantic Modeling in Semantic Search

Many different concepts and definitions for semantic search have been delivered by research so far (see Guha et al. (2003); Kiryakov et al. (2003); Zhang et al. (2005); Chu-Carroll et al. (2006)). The understanding of Semantic Search in the scope of information retrieval (IR) presented by Castells et al. (2007) differs in many aspects from the one in the Semantic Web community introduced by Tran et al. (2011). However, common to all Semantic Search approaches is the use of a Semantic Model which includes resources, query and a matching framework. An overview on related work in this area can be found in Uren et al. (2007).

2.8 Social Networks and Streams

Every fourth citizen of the world dealt in some way with social streams[65]. Most of the users are from Facebook[66] followed by micro blogging platform like Twitter[67] or by image sharing platform Instagram[68]. Usually such platforms once you log in contain a streams of your own posts an posts of other users from your network who follow you or are followed by you. Twitter differs from conventional social networks through limit to 140 characters (recently some changes happened in this field) for each message (so-called tweets) and through own user defined structure of messages through hashtags, mentions and re-tweets to mention the most important

[65] https://tinyurl.com/pp69opo, last access: 2017–05–29
[66] https://www.facebook.com/, last access: 2017–05–29
[67] http://www.twitter.com, last access: 2017–05–29
[68] http://www.instagram.com, last access: 2017–05–29

of them. In a special way tweets represent semi-structured text fragments which became a matter of analytic efforts in many different ways. Strohmaier and Wagner introduced 2010 in Wagner and Strohmaier (2010) a tripartite model for formalizing tweets called "Tweetonomies" consisting from users, resources and messages which offers a very interesting base for research on nature and structure of tweets and their relatedness.

2.8.1 Twitter and Microblogging

According to Tempelton (2008): "Microblogging is a small-scale form of blogging consisting from short succinct messages used by both consumers and businesses to share news, post status updates and carry on conversations". This definition is extended by 'Techterms.com'[69] as follows: "The most common microblogging platform is Twitter, which allows you to post updates of 140 characters or less. These updates, called tweets, may include hashtags, mentions (links to other Twitter users), or links to online resources, such as webpages, images, or videos. When you microblog using Twitter, your updates are seen by all users who have chosen to 'follow' you". Further the same website reports: "... the Microblogging on Facebook is more flexible than on Twitter, since you can post longer updates and include media directly in your posts. You can also share content with other users, similar to Twitter's "retweet" feature. Though Facebook makes it easy to post quick updates, its focus is more towards social networking than microblogging. Therefore, Twitter remains the most popular microblogging platform. While Facebook and Twitter dominate the microblogging scene, there are several other options available. One popular service is Tumblr, a website (owned by Yahoo!) that was designed specifically for microblogging. Tumblr allows you to easily insert photos, videos, quotes, and links into your posts and includes a 'reblog' feature for sharing other users' posts. Another service is Google+[70], which is similar to Facebook, and allows you to post updates that can be seen by the public or specific user within Google+ circles. Instagram (owned by Facebook) is a microblogging platform designed for

[69] http://techterms.com/definition/, last access: 2017–05–29
[70] https://plus.google.com/, last access: 2017–05–29

sharing images, while Vine[71] allows you to share short videos."

2.8.2 Semantic Relatedness and Metrics for Tweets

Semantic Relatedness and Metrics aims at defining methods and rules to extract as much as possible useful information and to identify in which context the information is related and linked to each other Milikic et al. (2011) in tweets. The premise used that a term is always defined through its context. However dynamically changing of context re-define the meaning of the terms. Milikic et al. (2011) developed so called "Normalized Micropost Distance" to track the changes of context of tweets over time. This approach tackles the context changing issue however what is still missing in this approach is detailed analysis of content of the tweets and terms significance as well the analysis of the significance of user and pre-strucured parts of tweet text like hashtags.

2.8.3 Trend Detection in Tweets

Trend detection represents a widely spread discipline in analysis of tweets. The idea behind the trend detection does not only aims the detection of trending words but also pre-processing, filtering and summarizing of context information relevant for different interest groups such as music fans, car fans, voters etc. Kraker et al. (2011) introduced a trend detection system defined through: specified taxonomy of keywords, specific list of users and their combination. For this puprose they used Part Of Speech (POS) tagging[72]. Laniado and Mika (2010) also introduced an approach for trend detection in 2010 in. In order to sort out the hashtags which are strong identifiers they defined in their work four main attributes for rating of them. Those are: frequency of users who used this tag, specificity of hashtags versus the meaning of hashtag word used in normal context, consistency of term used in hashtag and stability of hashtag term for a specific topic

[71] https://vine.co/, last access: 2017–05–29
[72] https://tinyurl.com/j4tyfet, last access: 2017–05–29

over time. Based upon those attributes they defined a vector space model to detect trends.

2.8.4 User Categorization and Profiling

Very relevant option in the realm of this work will be also the task of user categorization and profiling of Twitter users and users in general. Horn et al. (2010) used supervised classification in Twitter to tackle this challenge by implementing an application which follows the path of supervised learning. Such approach did not solve the hurdle of big amount of noisy data. Another try was done by Choudhury and Breslin (2011) with sport related tweets. The authors tried to isolate and identify reliable classifiers to identify users on twitter who tweet about certain sport event. Efforts like these are found necessary and valuable since Twitter as electronic "word of mouth" includes also customizable user-related content regarding commercial products and services. Knowing about them and the users related to them may bring competitive advantages for those who are able to govern this process as correctly observed by Jansen et al. (2009). Profiling and categorization reminds on battle at two fronts. On the way to the perfect user profile, while using the technology to re-fine and define the useful user context the research always struggles with the danger of loss of important and valuable information.

2.8.5 Semantic Modeling of Social Web Content

A formal definition of basic "Semantic Model" can be found in Tran et al. (2011). Data formatted in RDF and aligned to some context with concepts in RDFS or/and OWL is considered as semantic data. Currently only a limited number of works describe semantic modeling of data from social platforms. In Rowe (2009) authors applied semantic modeling to different social platforms in common contexts and evaluated the potentials of reasoning on such an infrastructure. According to the authors even a small amount of data yields good results with simple reasoning and delivers very precise matches. In Passant et al. (2010a,b) improved mapping social profiles with

related content, such as via interlinking the content tags. Semantic modeling for Twitter data has been applied by works introduced in my thesis Softic et al. (2010); De Vocht et al. (2011, 2014b).The have been identified as good resolvers for the retrieval of information and a solid interlinking base for the Linked Data Cloud. Similar use of semantic modeling of Twitter users was introduced on service level by Tao et al. (2011) and confirmed the benefits of this approach. These findings have been extended by the work on the "Researcher Affinity Browser" introduced in De Vocht et al. (2011), as a prototype of Research 2.0 mash-ups based upon a personal semantic model from Twitter connected with the Linked Data set COLINDA, allowing researchers to find and identify colleagues with the same or similar affinities and to track scientific events they visited. SemanticTweet[73] and Twitter-Based User Modelling Service (TUMS) presented in Tao et al. (2011) provide infrastructure to store the profiles data from Twitter in form of RDF (Resource Description Framework) graphs useful for further analysis. in this context FOAF (Friend of a Friend) vocabulary describes semantically users and relation between them according to Miller and Brickley (2010). Very commonly used vocabulary for description of posts is SIOC (Semantically Interlinked Online Communities) was introduced by Breslin et al. (2005, 2006a). For tag binding the Modular Unified Tagging Ontology (MUTO)[74] created and introduced by Lohmann et al. (2011) combines the best approaches from earlier efforts on defining a tag ontology.

2.8.6 Semantic Web, Semantics and Micro Blogs

Recent works on using Semantic Web technologies to structure and describe the data from microblogs, like in the case of SMOB (Semantic Micro Blogging) project introduced by Passant et al. (2008, 2009, 2010a,b) and some other scientific approaches introduced by Softic et al. (2010); Tao et al. (2011), outlines the potential of semantically based approaches and their usage for retrieving focused views on information stored within micro blogs. Combined with semantics, microblog content as current research shows presented in Softic et al. (2010); Stankovic et al. (2010); Rowe and Stankovic

[73] https://github.com/sflinter/semantictweet, last access: 2017–05–29
[74] http://muto.socialtagging.org/core, last access: 2017–05–29

(2010); Ebner et al. (2011) is a useful source of information. It is useful for instance for resolving scientific citations accordint to findings in Weller et al. (2011). Further, hashtags from tweets are trustful resolver to link entities to micro blog posts when combined with Linked Data. This observation is confirmed by Laniado and Mika (2010); Thonhauser et al. (2012). Extracting semantic entities and events from sports tweets has also been successfully done by Choudhury and Breslin (2011) with a very high precision using classical NER (Named Entity Recognition) and SVMs (Support Vector Machines). SMOB introduced by Passant et al. (2009, 2010a,b) is one of the open source micro blogging tools which also provides a SPARQL API for data querying. Beside SMOB there is also Smesher [75], a micro blogging client for twitter which allows storing tweets in form of RDF triples at local storage. It provides a SPARQL API as well.

2.8.7 Semantic Modelling of Tweets

Semantic Web Community provides a set of widely used schema (a.k.a. vocabularies) useful to cover the description of micro blog posts and user profiles attached to them. FOAF (Friend of a Friend) previously introduced in this chapter vocabulary describes semantically user and relation between them and was initially introduced by Miller and Brickley (2010). SemanticTweet[76] and Twitter-Based User Modelling Service (TUMS) a work done by Tao et al. (2011) provide infrastructure to store the profiles data from Twitter in form of RDF (Resource Description Framework) graphs useful for further analysis. Very commonly used vocabulary for description of posts is SIOC (Semantically Interlinked Online Communities) introduced by Breslin et al. (2005, 2006a). For tag binding the Modular Unified Tagging Ontology (MUTO)[77] published byLohmann et al. (2011) combines the best approaches from earlier efforts on defining a tag ontology.

[75] https://tinyurl.com/hdylvyr, last access: 2017–05–29
[76] https://github.com/sflinter/semantictweet, last access: 2017–05–29
[77] http://muto.socialtagging.org/core, last access: 2017–05–29

2.9 Learning Analytics and Importance of Reflection of Learning Activity

The current learning analytics research community defines according to Santos et al. (2012) learning analytics as the analysis of communication logs (see also Rosen et al. (2011); Bakharia and Dawson (2011)), learning resources (found in Niemann et al. (2011)), learning management system logs as well existing learning designs (see contributions: Lockyer and Dawson (2011); Richards and DeVries (2011)) and the activity outside of the learning management systems (referred by Pardo and Kloos (2011); Blikstein (2011)). The result of this analysis improves the creation of predictive models as introduced by Sharkey (2011); Fancsali (2011), recommendations as shown in Verbert et al. (2011); Drachsler et al. (2010) and reflection as described in Verbert et al. (2012). Learning Analytics resides on algorithms, formulas, methods, and concepts that translate data into meaningful information. Modeling, structuring and processing the collected data derived from e.g. user behaviour tracking plays a decisive role for the evaluation. Different works outlined the importance of tracking activity data in Learning Management Systems (for references see in Santos et al. (2012); Verbert et al. (2011, 2012); Rosen et al. (2011); Mazza and Milani (2005)). Nevertheless, none of these works addressed the issue of intelligently structuring learner data in context and processing it to provide a flexible interface that ensures maximum benefit from collected information.

2.9.1 Semantic Modeling of Learner Activities

The Semantic Web standards like RDF[78] and SPARQL[79] enable data for standardized interchange and to be queried as graphs. Data schema is usually projected on specific knowledge domain using adequate ontologies. This approach has been fairly successful used to generate correct interpretation of web tables reporter by Mulwad et al. (2010), to advance the learning process as introduced in Jeremić et al. (2012); Prinsloo et al.

[78] http://www.w3.org/RDF, last access: 2017–05–29
[79] http://www.w3.org/TR/rdf-sparql-query/, last access: 2017–05–29

(2012) as well to support the controlled knowledge generation in E-learning environments as described in Softic et al. (2009). Exploratory graphics introduced in Kirchberg et al. (2011) show that the sum of (web) user data on the access paths and the linkage of the resources within an environment (site) at a particular time window gives sufficient insight at what constitutes relevance; important properties and linkages between data resources. This potential was also recognized by recent research in *IntelLEO Project*[80]. The *IntelLEO* project delivered an ontology framework where *Activities Ontology*[81] is used to model learning activities and events related to them. In the same framework the Learning Context Ontology[82] offers formalization of learning context as general learning situation. Due to the relatedness to the problem that is addressed by this work these ontologies have been used to model the context of analytic data collected from PLE (Personal Learning Environment) logs.

2.10 Recommendation Systems

Recommendation systems have a wide area of appliance. They are especially successful in e-commerce. Most recently, they are also used in e-learning tasks for recommending relevant resources (e.g. papers, books to the learners (students) as reported in Thai-Nghe et al. (2010). The idea of recommendation systems dates from 90s of the last century. In 1992 PARC Tapestry System was introduced by Goldberg et al. (1992). This was a first prototype of collaborative filtering system where system was supposed to assist users in finding interesting content. It was the first time that explicit data was combined with behavioral data at the same data storage. Later on 1994, Resnick et al. (1994) introduced the GroupLens project which allowed an easy use of recommendation systems by providing an automated mechanism for generation of collaborative filters in the sphere of news articles. Boosted by the burst of internet bubble[83] in year 2000. The research on

[80] http://intelleo.eu, last access: 2017–05–29

[81] http://www.intelleo.eu/ontologies/activities/spec/, last access: 2017–05–29

[82] http://www.intelleo.eu/ontologies/learning-context/spec/, last access: 2017–05–29

[83] https://tinyurl.com/znre2gb, last access: 2017–05–29

recommendation systems in the period up to year 2006 developed rapidly. Big e-commerce companies like Amazon integrated recommendation systems into their platform gaining huge advantages through their use. These efforts were reported in Linden et al. (2003). Netflix a web movie streaming platform started also an competition for collaborative filtering recommendation algorithms to predict user ratings on movies which ended 2009 [84]. Nowadays, in a lot of segments of our life in so-called digital assistance systems we have have integrated recommendation system like in e.g. car navigation, movie streaming platforms, web shops etc. The main objective of recommendation systems is to elicit and identify personal preferences of users (customers) of an (web) information system and to offer for them identical or similar alternatives as option e.g. product items in a web shop. Recommendation systems use a number of different technologies. Some of those approaches are described in Rajaraman et al. (2014). Differing on approach how they try to solve the problem of recommendation we differ different groups of recommendation systems. Today we distinguish two main techniques by recommendation systems: collaborative filtering and content based recommendation. Additionally some literature also distinguishes knowledge-based and hybrid versions of recommendation.

2.10.1 Collaborative Filtering

Collaborative Filtering approach represents best researched type of recommendation. It has been subject of research for last 20 years. Many experts already used this approach to recommend users in several environments including social content as reported by Solskinnsbakk and Gulla (2011) and Personal Learning Environments as introduced by Mödritscher (2010). This kind of recommendation deals with similarities between two different users. In comparison to other recommendation forms in the case of collaborative filtering the advantage of the system is that it does not need to know anything about semantic of recommended items. This kind of systems evaluates the opinion of users (see also Schafer et al. (2007)).Tthere are several examples of usages of tags in combination with this method. Hereby, the tags are treated as addition to the user-item matrix and latter as additional

[84] http://www.netflixprize.com/, last access: 2017–05–29

semantics. However, this approach brings also some disadvantages like the cold-start problem. This means that the system has no reference values for similarity comparison for new items in system such as e.g. products new to the commercial platform. There are also issues of sparsity and scalability. For instance, huge systems like Facebook or Last.fm[85] and the like have millions of users. Calculating recommendations for all users and items represents a challenge. The problem of sparsity points to the fact that big selling platforms like Amazon.com sell a big amount of items, not bought by all users, which means that there is a lack of ratings for all of the products because not each user rates the product. A collaborative filtering approach produces always two type of results: a comparable numeric quantity of user rating related to a specific item and a list of items that are recommended. Depending on concept we differ user based and item based nearest neighbor recommendation.

User based recommendation

User based recommendation works on detection and application of user similarities to the recommendation hereby the most similar users and their item sets are recommended to the target user or group. One of the issues this approach tries to solve is also to predict which items a user might like in the future. The methods used vary from conventional methods like nearest-neighbor approach presented in Chen et al. (2010), or clustering introduced in Gong (2010) up to self defined methods like in Chen et al. (2012) where authors propose a method of making tweet recommendations based on collaborative ranking based on capture personal interests and their context including: "tweet topic level factors, user social relation factors and explicit features such as authority of the publisher, and quality of the tweet". This recommendation method works quite well and it has been applied in variety of use cases. However, it faces the problem of scalability with increasing number of users e.g. in big e-commerce systems. This approach in recommendation systems is also called *memory based*, because all of the data needed for prediction needs to be in memory.

[85] http://www.last.fm/, last access: 2017–05–29

Item based recommendation

Item based similarity uses pre-processing and pre-computation to prevent the problem with the scalability and to offer real-time predictions even in a huge systems like the one of Amazon. For this purpose, item similarity matrices have been calculated in ahead to lower and simplify the amount of possible items and to enable fast recommendation actions that scale. Hereby, the operations are done at runtime. Predictions are calculated respectively the number of nearest neighbors equal to the number of rating a user produced (for details see Linden et al. (2003)). Within this process several methods and techniques have been used to decrease the complexity and to filter the most important items. Maybe the most efficient one is where the users through choices determinate the scale of rates. In general, there is no universal solution to set and define the scale of efficiency at the beginning. It is always an evolving process. Choice and procedure of testing and test data sets are key success factors in this matter. This is what all recommendation systems have in common. Lack of appropriate useful data is the biggest obstacle in the implementation of recommendation systems. This obstacle was especially reported for learning systems by Drachsler, 2010 Drachsler et al. (2010). Item based nearest neighbor methods are considered as *model based* because they reside on models which are pre-computed offline and then used live to calculate prediction. An overview of such methods will be presented in following subsections.

2.10.2 Model based methods

Current state of the art literature distinguishes several approaches for model based methods. The most efficient one is the matrix factorization introduced by the winners of the Netflix Prize (see Bell and Koren (2007); Koren (2009)). Another widely used approaches are association rule mining introduced by Romero and Ventura (2007, 2010) and probabilistic recommendation introduced by Jannach et al. (2011).

Matrix Factorization

Matrix Factorization is model based recommendation approach which resides on derivation of latent factors from rating patterns. Those are calculated by approximation of a bigger matrix **X**, representing partially observed rating matrix, by two smaller matrices **W** and **H**. According to method introduced in Thai-Nghe et al. (2010) the matrix **W** contains latent factors of the user u_k, while **H** contains i_k items with feature vectors. The formula presented in equation 2.1 serves for calculation of rating that a user u gave to an item i.

$$\sum_{K}^{k=1} = w_{u,k} H_i k = \left(WH^T\right)_u, i \tag{2.1}$$

The success of matrix factorization is tightly bound to the efficient elimination of noisy data and to a proper training of the system. According to Jannach et al. (2011) mostly used approaches to get rid of the noise in data are stochastic methods.

2.10.3 Similarity measures

Similarity measures are used for both **memory based** and for **model based** approaches by recommendation systems to detect the similarity between the users or items. There are generally several similarity methods. Most famous for sure are Cosine and Adjusted Cosine Similarity and Pearson Correlation Coefficient.

Cosine Similarity is mostly used in item based approached and it it established most accurate similarity in this field. The usage of this similarity measure excels up to Informational Retrieval and Text Mining. The similarity between two vectors of items $\vec{a}$ and $\vec{b}$ is defined as follows in equation 2.2:

$$sim\left(\vec{a}, \vec{b}\right) = \frac{\vec{a} \cdot \vec{b}}{[\vec{a}] * \left[\vec{b}\right]} \tag{2.2}$$

Adjusted Cosine Similarity is an adjustment or extension of regular Cosine Similarity where average user ratings are subtracted additionally from the ratings. The equation 2.3 represents this similarity measure. The U represents the set of users:

$$sim\left(\vec{a},\vec{b}\right) = \frac{\sum u \in U \left(r_u, a - \bar{r_u}\right)\left(r_u, b - \bar{r_u}\right)}{\sqrt{\sum u \in U \left(r_u, a - \bar{r_u}\right)^2 \left(r_u, b - \bar{r_u}\right)^2}} \tag{2.3}$$

In both cases the prediction are calculated as weighted sums of user ratings for items.

Pearson Similarity Coefficient represents the counterpart to Cosine Similarity for the case of user based nearest neighbor recommendation. In other words it calculates the similarity for users instead of items. The equation 2.4 represents this similarity measure. The I represents the set of items:

$$sim\left(\vec{a},\vec{b}\right) = \frac{\sum i \in I \left(r_i, a - \bar{r_i}\right)\left(r_i, b - \bar{r_i}\right)}{\sqrt{\sum i \in I \left(r_i, a - \bar{r_i}\right)^2 \left(r_i, b - \bar{r_i}\right)^2}} \tag{2.4}$$

Usually, all user item sets are compared to each other by using the formula and than based on the conditions of how nearest neighbor is determinated a prediction of the items for a certain user can be calculated.

2.10.4 Content-based Recommendation

Content-based Recommendation in comparison to collaborative filtering needs the awareness of the system users and their items to produce recommendations. The most important task is to distinguish between irrelevant and relevant items that might be of interest for a certain user. This means that not all of the items and related information needs to be stored. Content-based recommendation systems may be used in a variety of domains e.g for recommending web pages, news articles, restaurants, television programs, and items for sale in E-commerce platforms. This kind of recommendation systems has been primary developed for Semantic Recommender systems by analyzing news feeds and documents and checking of their semantic

relatedness. Best way of representation for many information domains is by semi-structured data including some attributes with a set of restricted values and some free-text fields. A common practice in dealing with free text fields is to convert it to a structured form. Main objective of content-based recommendation is to find relevant tags and text fragments from unstructured content of documents, posts or online reviews Pazzani and Billsus (2007). Social tags as such are also representatives of semi-structured texts in blogging platforms. As such they may be used as starting point for content-based recommendation. Very common techniques applied in content-based recommendation will be described in following sub-sections.

Vector Space Models (VSM)

According to Laniado and Mika (2010), Vector Space Models are often used in the field of Information Retrieval as representation of documents where each dimension corresponds with a term in the collection and each value measures the weight of this term for the given document. The procedure is usually as follows: firstly the content of the document is encoded by keywords (tags, hashtags) by using TF-IDF (Term Frequency - Inverse Document Frequency) method. The TF represents the measure of frequency of a certain term in a document (see Jannach et al. (2011)). The definition of TF of an keyword i in a docment j is represented in equation 2.5. Term $freq(i,j)$ represent the absolute frequency of i in j. The latter term $OtherKeywords(i,j)$ in numerator of the fraction denotes the set of other keywords appearing in the document j. In order to know the TF we also need to know the $max(freq(k,j)$ where $k \in OtherKeywords(i,j)$.

$$TF\,(i,j) = \frac{freq\,(i,j)}{max\,(freq\,(k,j) : k \in OtherKeywords(i,j))} \tag{2.5}$$

The IDF is used as measure for weighting the relevance of keywords determined by their occurrence in other documents. The idea behind the IDF is that the keywords, that appear in many documents are less important then the keywords occurring in very few of them. The definition of IDF is shown

in equation 2.6. The term N denotes the overall number of the documents and the term $n(i)$ are the documents in which term i appears.

$$IDF\,(i) = log_2 \left(\frac{N}{(n(i))} \right) \tag{2.6}$$

Combined from those two definitions we get the final definition of $TF - IDF(i,j)$ (see equation 2.7).

$$TFIDF\,(i,j) = TF\,(i,j) * IDF\,(i) \tag{2.7}$$

In this way we get an final vector space model consisting from weighted vectors for keyword i in document j. Nonetheless, weighting keywords with non-context awareness produces inaccurate recommendations especially in complex documents and texts with e.g. scientific context because of the very narrow definition of terms used in them.

k-nearest Neighbor (kNN)

This technique is much alike the ones for the collaborative approaches because it analyzes primary the user profiles and their likes and dislikes which are then tracked and stored. It also uses Cosine Similarity already described in section 2.10.3 to determinate the similarity between two documents. For each new added document the technique findes k-nearest set of documents for a certain user based upon his profile and rating history. This kind of approach has been already successfully used for recommendation in Folksonomy Systems (reported by Gemmell et al. (2009)) and by automatic annotation of documents with concepts extracted from social data (introduced in Solskinnsbakk and Gulla (2011)). The advantage of kNN method is that it is easy to implement and that it is highly adaptive respectively new documents.

Classification

Classification as such can be also applied for the content-based recommendations. This technique as such allows applying of different machine learning

methods even on a sparse semi-structured text fragments as introduced in Horn (2010); Horn et al. (2010). Applying machine learning methods as for instance Supervised Learning opens new challenges and problems as choice of adequate representative training set. Also a very common obstacle in this context represents the determination of relevance of new text (document) arrivals as relevant or irrelevant. This process is called "labeling" and requires ad-hoc classification of new documents (texts) which is sometimes very inaccurate due to less semantic precision of classification features, sparsity of input or poor coverage of classification features. Also a very common issue is dealing with the problem of outliers: documents with unexpected input texts.

Application of Bayesian Classifiers

Bayes Classifiers especially Naïve Bayes as presented in Chen et al. (2009); Frank and Bouckaert (2006) are very often used for the tasks of text classification. The approach is based on Bayes Theorem as presented in equation 2.8.

$$P\left(A|B\right) = \left(\frac{P\left(B|A\right) \cdot P\left(A\right)}{P\left(B\right)}\right) \tag{2.8}$$

In equation 2.8 $P(A|B)$ represents the probability relationship between A and B. $P(B|A)$ is the probability on B assumed that A happened. $P(A)$ and $P(B)$ are so-called *a priori* probabilities expressed as fraction of possible options A and B. In Naïve Bayes each document is treated as collection of words without paying attention to their order and arrangement. In equation 2.9 taken from Frank and Bouckaert (2006) the Bayesian formula is applied to describe probability relationship that class value c fits a test document d. Words w occur $n_w d$ times in d. $P(c)$ and $P(d)$ are *a priori* probabilities for class c and document d.

$$P\left(c|d\right) = \left(\frac{P\left(c\right) \prod w \in d \cdot P\left(w|c\right)_w^n, d}{P\left(d\right)}\right) \tag{2.9}$$

Seth et al. (2010) used the Naïve Bayes to develop a subjective credibility model for participatory media like Twitter. The authors defined a set of rules and credibility metrics for identifying a message as useful for specific user. This is done through calculation of credibility of a certain document (message) for a certain user profile. This model was developed to improve and as extension to existing collaborative filtering recommendation system. Together they represent a hybrid form of recommendation (see also section 2.10.6).

2.10.5 Knowledge-based Recommendation

Knowledge-based recommendation relies exclusively on user ratings and demographic information. This approach is very suitable for very infrequent occurrences and situations where a recommendation should be chosen very carefully. This means in case of making recommendation for such decision a long time span of records should be considered. The whole action is focused on personalization towards single user. Generally, there are two different types of this kind of recommendation system distinguished: constrain-based and case-based. The first one relies on pre-defined user-given set of constraints that fulfill user's requirements. The second type is focused on retrieval of similar items using the similarity measures.

2.10.6 Hybrid Recommendations

Hybrid approach by recommendation systems combines the latter three approaches in order to overcome their weaknesses and to use their strengths to achieve better recommendation. Jannach et al. (2011) presented some of the mostly used algorithms and methods to combine the named approaches and also presented in their work a table for categorization of input data requirements of recommendation algorithms which should serve the easier choice of recommendation algorithm for specific cases. Almost every known algorithm has segments of hybrid approach like e.g. Netflix Prize winner approach introduced by Bell and Koren (2007); Koren (2009). Also, a very interesting approach of a tag cloud based recommendation system that uses

social relations to recommend user generated content has been introduced by De Pessemier et al. (2009). Another very useful survey paper on presenting current state of the art approaches with instruction table how to chose the proper algorithm for individual recommendation solution was presented by Adomavicius and Tuzhilin (2005).

2.11 Classification of Semi-structured Text Artifacts and Part of Speech Tagging

Successful knowledge discovery through mining an profiling of information from semi-structured sparse text artifacts relies as in the case of longer texts on efficient approaches for classification of features describing the context of information needed to fulfill this task. This section introduces some of the state of the art methods in this field.

2.11.1 Natural Language Processing (NLP)

As introduced in the sections 2.10.3 and 2.10.4 methods like Cosine Similarity or TF-IDF (Term Frequency–Inverse Document Frequency) do not include any semantics into their processing. Therefore, it is also important to involve methods and techniques which process words as part of text by their meaning in order to get an insight into context of implicit information hidden in text fragments. In this subsection two main aspects of implementing text contextualization will be described: NLP Pipeline and POS (Part of Speech) Tagging. Both of them are part of a broader concept called Natural Language Processing or abbreviated NLP.

Part of Speech (POS) Tagging

POS Tagging is a technique which tags each single word in the sentence assigning in this way a meaning for each of them respectively the information context. In this way it is possible for machines to detect the semantic

relatedness between the words in different context. POS Tagging is just one part of the NLP Pipeline and it is implemented either through supervised or unsupervised learning approach. The difference lies at the ability to tag ad-hoc unknown class of the word. Gimpel et al. (2011a) developed for instance a specific POS Tagger for Twitter. Similar efforts has been also done by Ritter et al. (2011a). Both achievements are relevant for the matter of this work. This technique leans on standardized text corpora and it is language dependent which limits their usage only to supported languages. There is a series of institution who do their research in the field of computer linguistic and offer their text corpora as reference data sets. Depending on their content they can be applied for certain domains with different success rate.

NLP Pipeline

In his book Russell (2011) introduced a typical NLP Pipeline as follows:

- End of sequence detection: breaking the text into meaningful sentences.
- Tokenization: making each word in a sentence a token.
- POS Tagging: assigning POS information to each token.
- Chunking: analysing tagged tokens and detecting logical concepts in them.
- Extraction: analysing chunks and tagging chungs as named entities which can be e.g. events, persons, locations etc.

This is just a formal definition of NLP Pipeline. Of course the real-world implementation slightly differs from the presented form in terms of implementation of parts which are necessary. This also depends from the language, size and form of the text that should be processed. Information gained through NLP Pipeline is usually than proceeded to further processing e.g. for calculating similarities.

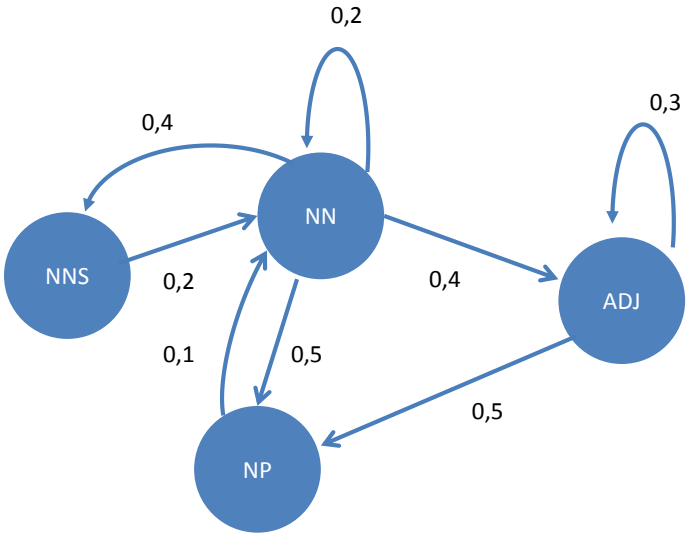


Figure 2.6: Example HMM with four different kinds of POS tags (NN,NP,NNS,ADJ) with transition probabilities from one state to another

2.11.2 Hidden Markov Models

Tweets and similar sources of sparse text often represent collection with single text artifacts having time dependent and context related dynamically changing meaning. Very commonly used approach dealing with prediction of changing circumstances are Hidden Markov Models (HMMs). HMMs describe the probability for tacit information. In the context of knowledge discovery in texts HMMs are applied often as POS taggers (see Goldwater and Griffiths (2005)). The usage of HMMs within POS Taggers helps the identification of certain words in relation to the sentence. Hereby each POS tag represents a single state in a HMM.

2.11.3 Support Vector Machines

Suport Vector Machine (SVM) is supervised learning technique used for the purposed of classification and regression analysis of an given feature vector space. SVM models are built upon training data and strongly depend on it. The built up model predicts the target values of test data, given test data attributes Hsu et al. (2003). In this procedure *hyperplanes* separate sets of objects in classes, aiming to produce the best possible separation degree near the class border. The distance between the first occurring object of a class and hyperplane is called *margin*. SVM tries to find hyperplanes, with as many margins as possible. In order to achieve the classification in two dimensional space not linear objects need to be mapped into a higher dimensional space. This is done by mapping the feature vectors. SVM use several kinds of so called *kernel functions* which allow to calculate functions, hyperplanes in this case, in higher dimensional spaces. According to Hsu et al. (2003) most applies functions in this procedure are: linear, polynomial, radial basis functions and sigmoid.

Since SVM are binary classifiers and in case of POS tagging more than two classes have to be classified. Therefore in case this method is applied it is usual to engage the K classifiers. Those classifiers are created to separate a class from all other classifiers. Authors in Nakagawa et al. (2001) used SVM to successfully predict POS tags using following features: POS context (word next to unknown), word context (lexical forms of two conterminous words) and substrings (prefixes and suffixes of unknown word - up to character length of 4). Kudo and Matsumoto (2001) found that SVMs are especially useful for chucking approaches. The method is slightly differs from the POS approach through kernel function used and adaptations of hyperplanes.

2.11.4 Clustering

Clustering represents a very common technique in unsupervised learning[86]. According to Witten et al. (2011) concept of clustering relies on method

[86] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/, last access: 2017–05–29

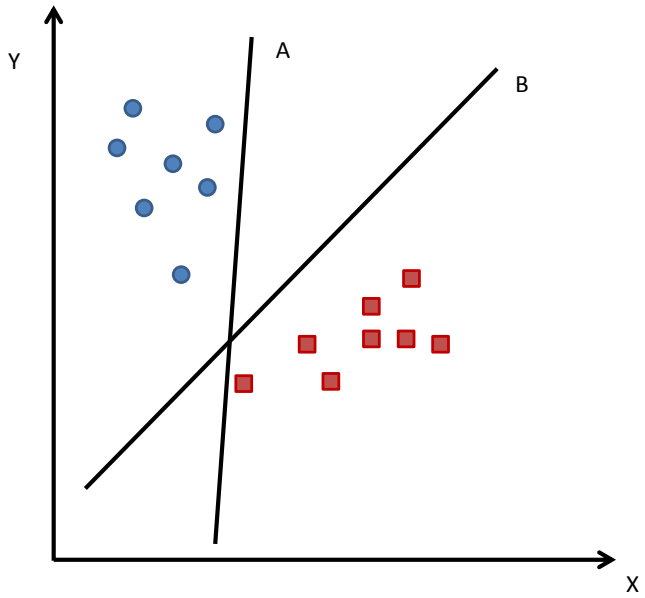


Figure 2.7: Sample of linearly separable problem through SVM with two possible hyperplanes A,B

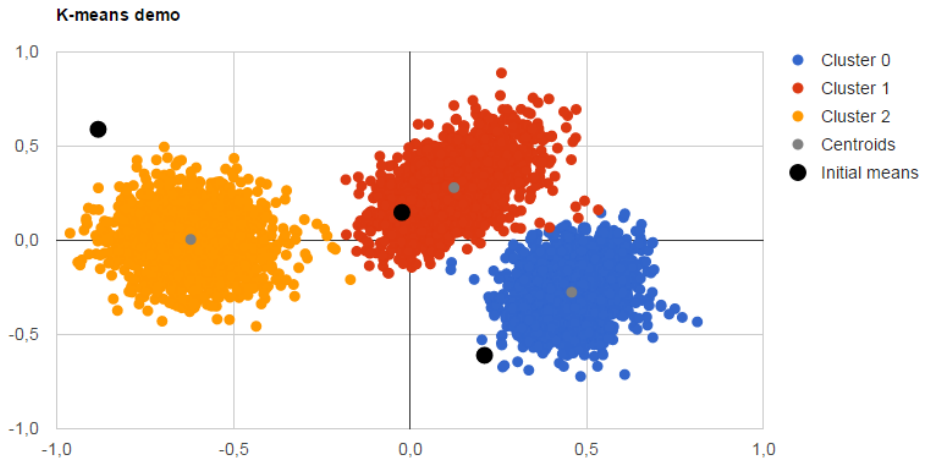


Figure 2.8: Sample of k-means clustering for k=3

of grouping similar items to clusters based on their stronger mutual resemblance regarding one special criteria. Most important component of clustering are distance measures. Figure 2.8 shows an example of terminated clustering process simulated in an online demo for k-means clustering[87].

Generally there are several ways to formulate the clustering criteria. Authors in Fung (2001) offer also an overview of common clustering algorithms. Based upon the method there are:

- Exclusive Clustering - one item can be assigned to exactly one cluster
- Overlapping Clustering - one item can be assigned to many clusters
- Probabilistic Clustering - degree of affiliation to single cluster is defined through the probability function e.g. k-means
- Hierarchical Clustering - based upon union of two "nearest" clusters

Usually a clustering algorithms consists from four steps:

- Assumption - creation of initial clusters and starting points
- Initialization - checking the initial affiliation through e.g. probabilistic functions

[87] http://syskall.com/kmeans.js/, last access: 2017–05–29

- Iteration - recalculation of the centroid of cluster and re-affiliation of items
- Termination - done when certain criteria is fulfilled

Application of clustering is manifold and reaches many domains like: marketing, psychology, libraries, city planning, insurance, documents classification etc. In the realm of this work as machine driven complement test method to semantic modeling of information from semi-structured sparse texts algorithms as k-means and hierarchical clustering as most commonly used will be applied in different use cases.

3 Potentials for Knowledge Discovery in Online Research Communities

This chapter introduces experiments, findings and concepts published as separate scientific works listed in 1.6 which are trying to unveil the potentials of Twitter as platform and tweets as exemplary sparse semi-structured text fragments as source for knowledge discovery and data mining relevant for researchers as target group. The text from the listed publications has been used also partly to describe the methodology and concept, implementation, experiment conduction, preliminary findings and conclusion. Description of these experiments should outline the eligibility of research questions specified in 1.3. This chapter addresses in particular research questions: RQ1 and RQ2.

3.1 Tracking Researchers on Twitter Using the Conference Hashtags

3.1.1 Statement to Own Contribution

This section with all subsection includes partly or at whole the text and figures from Softic et al. (2010); Ebner et al. (2011) which were written by myself in cooperation with my supervisor and my colleagues. My contribution beside writing the articles was developing of the concept of the testing use case to track the users at specific research events as conferences and design and conduction of analysis with twitterStat, formerly a.k.a. STAT

Tool. The implementation of the tool was done by the one co-authors, more precisely by Mr. Thomas Altmann in the realm of his bachelor thesis.

3.1.2 Why is Twitter Interesting as Source of Tacit Information?

The phenomenon of Web 2.0 has brought the concept of user generated content, later on social networks and as a part of it micro blogging. Especially micro blogging platforms as Twitter gained strong importance in recent years. Twitter is daily generating hundreds million of Tweets [1] and reached already than billion search queries a day mark[2]. According to statistics[3] from recent years, Twitter has over 300 million active users. A growing number of people that are linked via acquaintances and online social networks such as Twitter allows indirect access to a huge amount of ideas. These ideas are contained in a massive human information flow (see Jansen et al. (2009)). Users provide many relevant data for specific purposes, as shown in many studies before such as Reinhardt et al. (2009b); Java et al. (2007); Rowe and Stankovic (2010); De Vocht et al. (2011).

3.1.3 Usage and Form of the Tweets

Studies on the use of microblogs like Twitter[4] conducted by Ebner et al. (2010a, 2011) and by Reinhardt et al. (2009a); Letierce et al. (2010a) within the science community has shown that researchers are using Twitter to discuss and asynchronously communicate on topics during conferences and in their everyday work. Short form of tweets posted by microbloggers offers also a solid base for automated content processing and analysis. As reported by Reinhardt et al. (2009a) conference related Twitter-streams based upon a hashtag search reflect the ongoing occurrences within the actual event. Twitter info-walls placed at the conference location also support

[1] https://tinyurl.com/kvc8oqd, last access: 2017–05–29
[2] https://tinyurl.com/znbwc69, last access: 2017–05–29
[3] https://tinyurl.com/pm7txe9, last access: 2017–05–29
[4] http://www.twitter.com, last access: 2017–05–29

the conference administration, communication and discussion between the scientific tracks and sessions. Tweets as such have some significant parts like hashtags, retweets or mentions (denoted with leading '#', RT or '@'). It is aimed to show whether there is a possibility to get significant information from a pool of tweets with the right mining reference in form of reliable linked data repository. A survey of the use of Twitter for scientific purposes conducted by Letierce et al. (2010a) showed that Twitter is not only a communication medium, but also a reliable source of data for scientific analysis and profiling tasks and trends detection. The same findings are confirmed also by Mathioudakis and Koudas (2010); Softic et al. (2010); Tao et al. (2011).

3.1.4 Making use of Tweet's Semi-structured Form to Track Research Events

For exploring the potentials of tweets as meaningful source for mining and knowledge discovery, the team at E-Learning Lab at Graz University of Technology developed adequate a set of analytical tools[5] for Twitter. Those tools enabled basic capturing the text of tweets and user profile information and allowed an simple automatic analysis of tweet content based upon statistical indicators. Especially interesting in the given context is so called *twitterStat (Semantic Tweeter Analysis Tool)* (semantic in this context is related to the linguistic attachment). The twitterStat is still in development at this moment, but the first beta-version is already online[6]. However future efforts including this thesis are aiming to create an analysis system that will be able to answer simple questions about people and their actions and interactions based upon their information from Twitter for researcher as target person. A snapshot of the current status of STAT is depicted in 3.1.

Testing the potential of tweets (here treated as semi-structured text fragments) and Twitter as such was done in experimental way using a specific use case. For this purpose tweets of a big e-learning conference are examined. It is aimed to show whether there is a possibility to get significant

[5] http://twitter.learninglab.tugraz.at/, last access: 2017–05–29
[6] http://twitter.learninglab.tugraz.at/stat/, last access: 2017–05–29

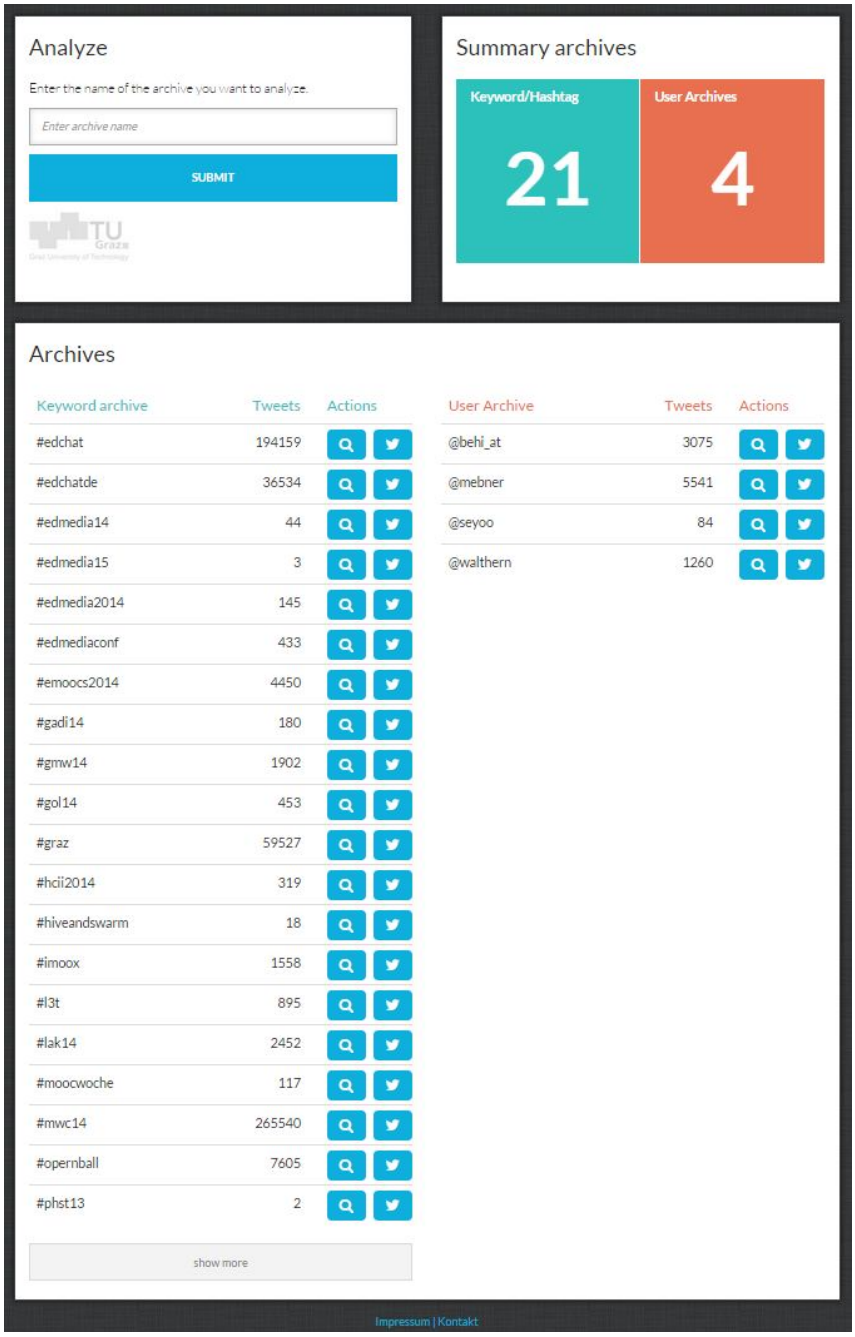


Figure 3.1: twitterStat (fomerly STAT) tool for analysis of tweets.

information from a pool of postings or not. For this experiment the ED-MEDIA 2014 conference has been used. ED-MEDIA is a well established international conference on "Educational Multimedia, Hyper- media & Telecommunication" and started in 1993 as follow-up after 6 years of International Conferences on Computers and Learning (ICCAL). The main purpose as stated on their web page is to serve as a multidisciplinary forum for the discussion and exchange of information on the research, development, and applications on all topics related to multimedia, hypermedia and telecommunication/distance education. Nowadays it is certainly one of the largest international conferences on these topics. Over 1000 participants every year attend numerous sessions and workshops for couple of days. Two recent publications Khan et al. (2009); Ochoa et al. (2009) pointed out the huge amount of contributions, the relationship of authors, the key players and lots of more trends. An initial test analysis produced results andswers to questions such as: Which persons (@) were using the hash tag #edmedia14 and how often? Which hash tags (#) were used with #edmedia14 and how often?

3.1.5 Discussion and Conclusions on First Results

First results have shown that keyword extraction can be taken as basis for further investigations and treatment of data and that it is possible to get meaningful outcomes, for instance filter relationships between words used in order to find important content or users themselves. Of course the interpretation of such analysis is limited and will never replace personal participation, but it gave a short and general overview. This first analysis can be seen as starting point for further semantic analysis for example to interlink other resources automatically. In this way simple information froma tweet can be enhanced to become more reliable.

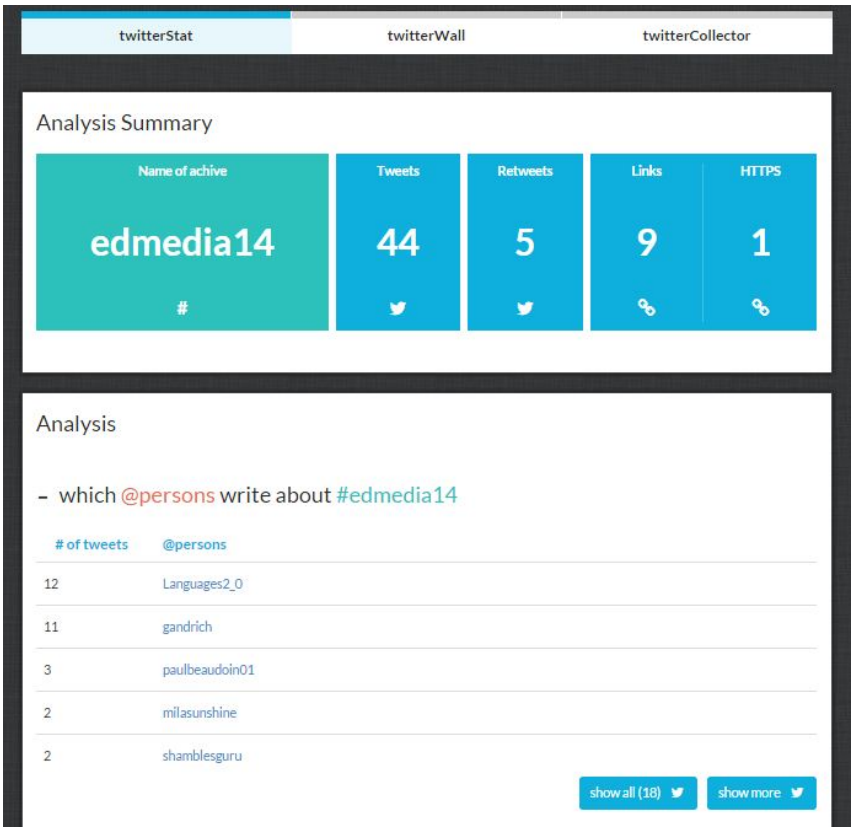


Figure 3.2: twitterStat (fomerly STAT) tool for analysis of tweets detecting persons on #edmedia14 hashtag stream.

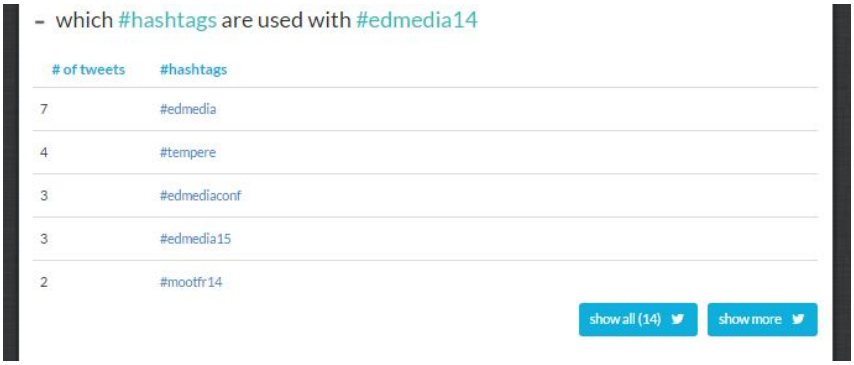


Figure 3.3: twitterStat (fomerly STAT) tool for analysis of tweets detects popular hashtags #edmedia14 hashtag stream.

3.2 Clustering of Interest Groups for Recommendation of Researchers on Twitter

3.2.1 Statement to Own Contribution

The text in following section origins from Thonhauser et al. (2012) and from Master Thesis of Mr. Patrick Thonhauser who I co-supervised with my mentor Dr. Ebner. The overall idea, the use case and technical concept in this work as well description of the system including evaluation methodology and evaluation itself was done by myself together as part of supervision of the master thesis of Mr. Thonhauser. Implementation of the prototype was done by Mr. Thonhauser as part of his master Thesis. This section aims to pinpoint the potentials of applying unsupervised algorithms on sparse semi-structured texts as tweets and Twitter user profiles in order to support profiling and recommendation of research related contacts. It is concept of a novel approach for finding new interesting users and information for a specific Twitter account. Work presented here is not intended to serve as a detailed description of a semantic recommender system for research 2.0, but rather as a brief overview of a proof of concept application, which's main task is the classification and recommendation of Twitter users based on their profile description and their tweets. Also preliminary results of extensive categorization task are presented at the end.

3.2.2 Twitter and Its Users

Every millions of users are communicating via Twitter, exchanging the latest news and discussing millions of diverse topics. Everybody as Twitter user who is interested in a specific person, specific tweet topic, has the ability to retrieve the information by exploring the tweeted resources. Twitter has become one of the most popular applications for the dissemination of information stated authors in Kraker et al. (2011) and it is therefore an ideal candidate to serve as the main source for mining data concerning users and provided information of scientific interest. The interesting questions from researcher's standpoint is how to make use of the information contained

within millions of tweets and what to extract from those 140 character of a single tweet. How much useful information is hidden within and how can we separate useful information from noise? What about the user researcher user profiles in general. Can this information be used for finding persons (other users in our case researchers) having similar affiliations based only upon the content they are posting? As pre-study for targeted deeper semantic analysis and categorization of tweet content a set of conventional clustering algorithms (hierarchical clustering and k-means) have been tested to detect the group of users with similar interests from researcher domain and to check whether later aimed profiling can be applied once the tweets are made retrievable through semantic form.

3.2.3 Concept and Use Case: Thought Bubbles

Twitter users follow other users for several reasons (e.g. because of similar fields of interest). This does not implies that the connection between similarly interested Twitter users have to be necessary mutual. In Twitter which is respectively user relations not implicitly bidirectional, an individual user does not have to know his followers or to communicate with them to engage the creation of relation. Usually, the follower is interested and involved with similar topics, to the followed user. Therefore, it is most likely that other followers of the same user have similar connections, that may be of certain interest to that particular user. A single user in Twitter is active and interested usually in several kinds of "topic based bubbles". Hereby, users interested on such "micro community" do not necessarily know all participants of such a "bubble". The users within user's specific bubble, might be of interest to each other. Figure 3.4 shows an example of a network graph, which reveals the sphere of activity within diverse "Thought Bubbles". Users marked with a star (*) are potentially highly interesting for this particular user (centered in figure 3.4). These users belong to the same topic specific bubble, as illustrated here, to the "Science Bubble". This implies that following a specific user within a certain field of interest increases the probability of finding further relevant users who are also engaged in such specific field. The missing bi-directionality of certain user connections, indicates interest based relationships. This assumption reflects the basic

concept of "Thought Bubbles". Such circumstances offer the possibility of recommending people and information, which is contained within a bubble and not yet explored by a specific Twitter user.

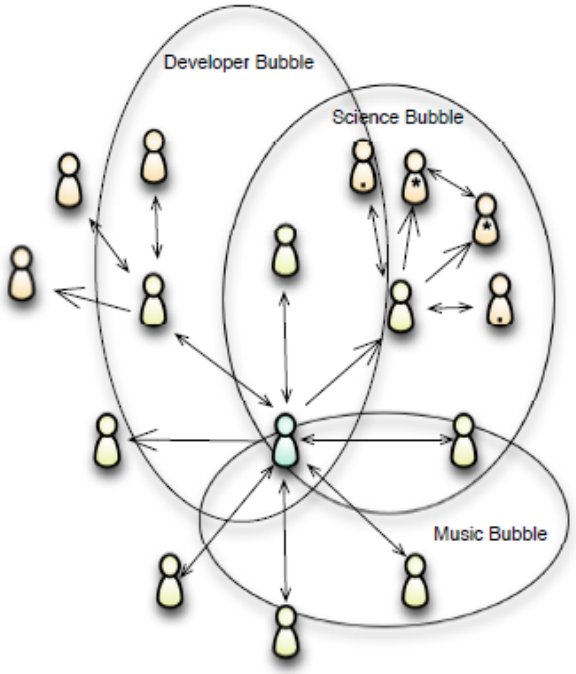


Figure 3.4: An example of how a user can be placed in Twitter network graph. As presented in Thonhauser et al. (2012)

3.2.4 System Design, Methodology and Implementation

The conceptual realization of specific "Thought Bubbles" can be split into several sub modules.

Finding Relevant Users

The first sub module separates less useful users or so called "noisy" users, from those that spread interesting news, personal thoughts and facts. To simplify this process a pool of people who are connected to ones Twitter account has to be defined. This connection exists because either a user is following other users or because other users are following the user self. This pool of people will be further denoted as the so called "inner circle". Separating the inner circle of people by filtering useful information provided by those people helps to reveal further potential matching interests, which are hidden in the so called "outer circle". However, the "outer circle" of people represents the connection to every person acting within ones inner circle. Subsequently, a second cycle of filtering is performed to efficiently narrow down and identify the people of potential interest. Horn Horn (2010) uses in his master thesis Support Vector Machines (SVMs) for this challenging classification task. SVMs are a commonly used technique for text classification and are recommended by many researchers like Rios and Zha (2004), Hsu et al. (2003) or Nakagawa et al. (2001). By applying this method, a potentially interesting set of users would remain for further selection. Also thinkable is the usage of a POS-tagger and a chunker in advance. This is helpful by checking whether a Twitter account belongs to a person or an entity as company, organization or the like. Achieving clearly arranged set of Twitter accounts worth exploring in depth also includes eliminating duplicates within this set and eliminating the accounts that are already followed.

Tackling Categorization of Users

Granular categorization of users represents the most challenging part. Categorization of active users is done by using the "Thought Bubble" service. At the very beginning, a set of appropriate categories that cover all relevant user interests need to be defined. For example, such categories would be e.g. developing, science, teaching, etc. The annotation of words in user's Tweets is done by applying Natural Language Processing (NLP) techniques, as done before in Ritter et al. (2011b). Classifying tweets, in comparison to the regular classification of text artifacts, is a very specific task because of

the length of tweets (limited by 140 characters), the often changing context in which a term is used and the frequent usage of uncommon vocabulary in tweets. The elimination of irrelevant words like conjunctions, stop words or prepositions can be realized by tagging in Tweets (using Part-Of-Speech tagging). For instance, Gimpel et al. (2011b) already developed a POS-tagger especially suited to Twitter. Summarizing the results of all categorized user's Tweets, leads to classification of a user expressed in percent. As referred previously, SVMs can be used for such a task as applied by Nakagawa et al. (2001). There are also several other methods, for accomplishing classification such as Bayesian approaches as it has been done by Goldwater and Griffiths (2005). Future research and evaluation in this field will hopefully offer the answers which of the mentioned methods is the best one for the categorization of Twitter users. One example for evaluation of such approaches is presented by Choudhury and Breslin (2011).

User Related Indicators for Recommendation

In addition to measuring of the similarity of "Thought Bubble" attributes, with regard to the affiliation of a user into a category, several other indicators for determining the significance of a user's recommendation are used to sharpen the matching accuracy:

1. **Tweet Frequency** is the amount of Tweets a Twitter user is firing within a defined period of time.
2. The **Follower ratio**. The more followers a user has, the more influence or credibility one might posses. When a user has very few followers, but is following a huge amount of other users, might hint towards the user being a Blast Follower[7].
3. The **Amount of Retweets** a users Tweets have, indicates the magnitude a user's reputation has.
4. Clients will have the possibility of **Rating Recommended Users or Tweets**. By comparing these with potential recommendations for a "Thought Bubble", similarities between them will influence the users overall rating score within a bubble.

[7] http://www.makeuseof.com/dir/blastfollow-mass-follow-twitter-users/, last access: 2017–05–29

Presented indicators are considered as parameters for fine-tuning of the selection of recommended tweets and Twitter users. However, the main task regarding applying these indicators, is to find an appropriate weighting scheme respectively thresholds and significance of each single indicator. Same indicators could be used to detect influential and relevant recommendation with other methods like e.g. neural networks in supervised and unsupervised version. The problem however remains also the same: finding right weighting and thresholds.

Recommendation

Recommendation decisions are met by calculating ratings for each potentially relevant user, based on their category classification and the additional indicators, introduced in 3.2.4. Subsequently, category classification of an active service user is compared to the classified categories of potentially interesting users. In advance, all additional indicators have different weights, what influences the ranking of the user in the final recommendation list.

3.2.5 Results and Application Setup

First experimental prototype (proof-of-concept) of "Thought Bubbles" was implemented in the realm of the master thesis of Patrick Thonhause Thonhauser et al. (2012). Overall architecture of this implementation is shown in figure 3.5. Django[8] was used as Web framework and the Natural Language Toolkit (NLTK)[9] for classification and word processing tasks. Data storage is handled by using SQLite[10] and Twitter related requests are handled by the Python Twitter framework[11]. Presented implementation was not made with focus on offering the best performance. It was primary the proof of concept. Re-implementing it at large scale I would definitely re-consider the usage of SQLite and maybe use some more performant relational or NoSQL database. The same applies to the used processing framework. It would

[8] https://www.djangoproject.com/, last access: 2017–05–29
[9] http://www.nltk.org/, last access: 2017–05–29
[10] http://www.sqlite.org/, last access: 2017–05–29
[11] http://code.google.com/p/python-twitter/, last access: 2017–05–29

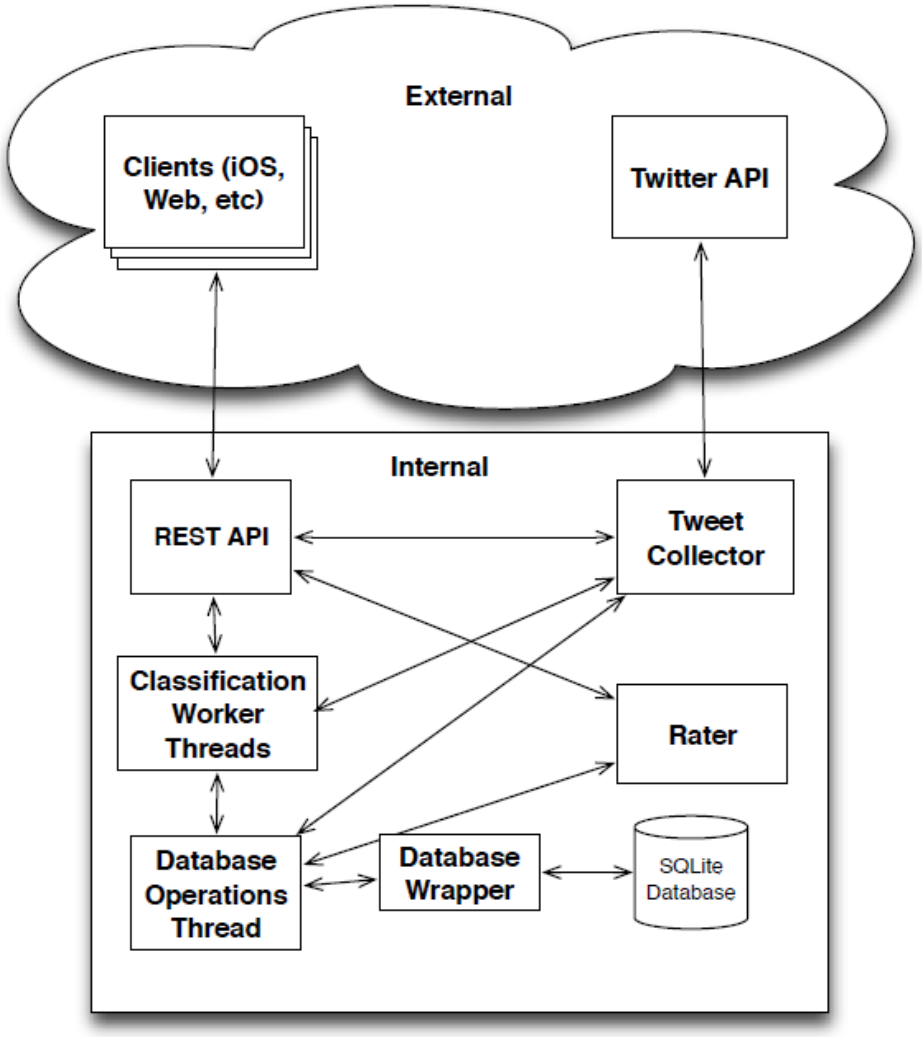


Figure 3.5: Thought Bubble infrastructure. Adapted from Thonhauser et al. (2012).

be replaced by some strong Big Data processing engine such e.g. Apache

Spark[12].

Proof of Concept Implementation Setup

The classification is done by filtering hashtags within tweets of users and by applying the POS tagging and chunking to the last 200 Tweets fo each Twitter user's timeline. For POS tagging was used NLTKs Trigram Tagger[13], trained with Conll-2000 training data introduced by Tjong Kim Sang and Buchholz (2000). Similar approach was applied in similar use case described in Ritter et al. (2011c). After applying POS tagging, sentences are brought into the form of so called chunk trees (see Abney (1992)). Feature vectors are compiled by iterating through the chunk trees and searching for detected phrases, names and nouns. To emphasize the influence of hashtags, their single occurrence has been counted twice within a vector. Additionally, in order to reduce the weight of words that are not useful for categorization, the words that occur most frequently in the English language (the 200 most used English words) are removed from the vectors. This task is performed for all Twitter users within a potential "Thought Bubble". Afterward, the results have been compared by applying cosine similarity. In this way rating of similarity of tweeted content was implemented. This similarity is measured through comparison of the word frequency counts of words and phrases, which were classified as relevant by the preliminary steps (POS tagging, chunking and phrase, noun and name filtering).

First test results

Conduction of a first test run included all steps mentioned in 3.2.5. All tweets of test users have been cached previously to ensure that all observed accounts are in the exact same actuality during testing and that occasional tweeting of some of them would not affect the results for the chosen sample period. Exactly 49 Twitter accounts were compared to Twitter account @mebner belonging Martin Ebner. Within the test set of users, 21 Twitter

[12] https://spark.apache.org/, last access: 2017–05–29

[13] http://nltk.googlecode.com/svn/trunk/doc/howto/tag.html, last access: 2017–05–29

accounts of people and their students who work in the same or similar fields as @mebner were added to measure the reliability of the system. The rest of the Twitter accounts for this test run have been chosen randomly. The number of test users does not allow general conclusions regarding the system classification abilities however it is big enough to test the targeted use case and overall usefulness of the implementation. Figure 3.6 visualizes the results of a first test run, based on Martin Ebner's (@mebner) Twitter account. The best scoring users were achieved whether by students or

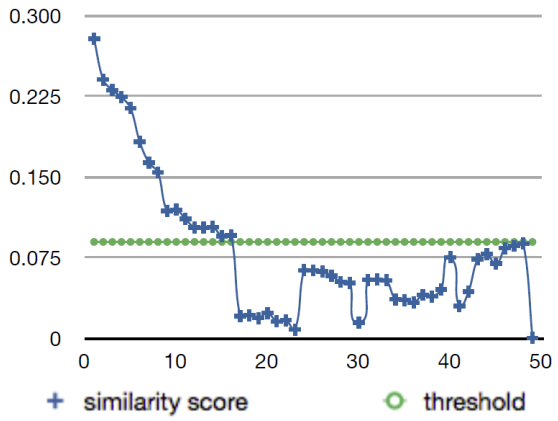


Figure 3.6: Test run with 50 Twitter users including @mebner account. First 21 data points represent the hand picked users. Adapted from Thonhauser et al. (2012).

researchers, whose profile description is similar to @mebners. The highest scoring account was already followed by @mebner. Not a single random pick scored more than slightly above 0.09, but still lower than 0.1. The highest scoring from non-researcher data set was achieved by a tech blogger's Twitter account, which could indeed be of potential interest to a professor at a university of technology. Five out of 21 manually added researchers and students scored lower than expected. The reason for this occurrence lies most probably on the amount of useful tweets. Possibly they used the Twitter account for profession unrelated postings. By applying more indicators (features) as discussed in section 3.2.4, it is valid to expect that the error rate will be reduced to a acceptable level. Nonetheless, the 0.1 mark seems to be a good threshold for deciding, whether a Twitter account should still be considered for further analysis. At least in the case of

@mebner. Similar to the first test run, further test runs were conducted for every member of the manually picked users. Figure 3.7 visualizes all found optimal thresholds, which would enable the categorization of an account to reach a similar accuracy to @mebners test run. By observation

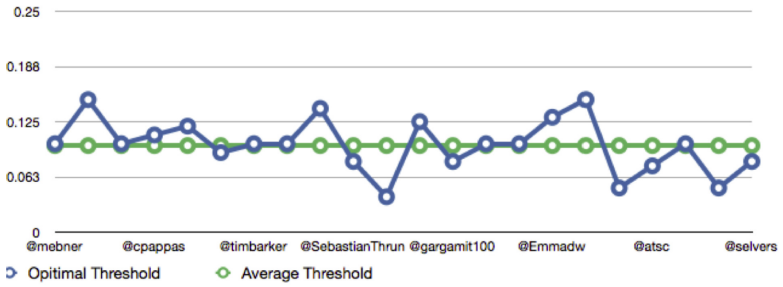


Figure 3.7: Thresholds of the 22 hand picked users including @mebner. Adapted from Thonhauser et al. (2012).

of each set of results for the tested users, thresholds were parametrized. These thresholds were set to reach a minimum 75% limit, where at least three quarters of the hand picked users were categorized as potentially interesting. The 75% rate of correct classification is motivated by the results of @mebners Twitter account. The average threshold of 0.098 has been calculated by summing of all specific thresholds and their division by the count of tested users. Although the calculated average threshold of 0.098 is very close to the presumed 0.1 of @mebners case, the statistical spreading of the specific thresholds are up to 50% and more. Therefore, the assumption that the usage of a threshold isn't the best choice for pre-elimination is valid, because the amount of accounts for further processing may vary too much. Applying a simple k-nearest neighbor algorithm would be more appropriate for limiting the number of potential recommendations in advance. All top "n" picks within a test set, are now part of "Thought Bubbles" of target user. Within this set of potentially interesting users, category specific bubbles can be extracted and then recommended as a topic based subset of users. In advance to this, "Thought Bubbles" for the target user of the service, will be available via the REST API as visualized in figure 3.5. The bubbles will be delivered as JSON[14] objects where the client then has the responsibility of deciding how those recommendations are presented to the selected user.

[14] http://www.json.org/, last access: 2017–05–29

3.2.6 Discussion, Conclusion and Outlook

The results achieved in the categorization within the additional 21 test runs are varying. This can be explained by taking into account that different user use different words, have different language proficiency and phrases and have different interests beside their professions. Nonetheless, the usage of re-tweets in the set of tweets that where POS tagged and chunked, lowered the scores significantly within the set of accounts, which should at least score close to a specific threshold. As a result of that, it is recommended for future test runs to exclude usage of re-tweets in classification task. In the first place, as a big advantage of presented approach in comparison to similar approaches like in De Vocht et al. (2011), the concept of Thought Bubbles isn't limited to the movement of a specified community like *Research 2.0*. It can be used in any kind of topic related community because of its generic approach. The fact that people are classified, basically on the content of their tweets and not only on hashtags, mentions or already existing connections, leads to new and so far undiscovered personalized recommendations of users with similar interests. Hereby, a thing to consider is that all recommendations are always based on the context of the latest "n" tweets of a user (where length of "n" depends on technical limitation of the system), and therefore they change over time dependently on content and number of tweets used. Nevertheless, first insights made in this work show to what extent is Twitter useful for discovering useful and interesting information. Considering that recommendations depend on the content of Tweets, it raises the challenge to find metrics and techniques that enable us to filter as much as possible noisy content and detect significant facts within a dynamically changing context. The classification of user profiles represents a pre-work for user-related social recommendation in social networks isn't just a Twitter related topic, it can also be used for similar applications as Facebook, Google+ and the like. Presented methods can help to establish connections between people with similar interests, particularly scientific interests or expertise which is from high significance as mentioned already in Stankovic et al. (2010). Thought Bubble service as implemented in this experiment, gives the opportunity to Twitter users to access other peoples knowledge just by tweeting about what they do. This is implicitly not only an alternative way for finding new and interesting people, but

rather a way of creating a personal subset of people, who might be able to answer your questions or influence own work. With other words, this is also a step towards a personalized and focused stream of information for everyone. Based on the presented findings, future development may also include the answer to the question of whether the huge amount of noise can be eliminated in a satisfactory amount of computation time in order to provide a competitive system.

4 Semantic Modeling and Mining Approach for Online Researcher Profiles

Based upon findings in chapter 3 the author made the decision to focus more to the domain specific view on the Data. Although the results in section 3.2 brought some very promising results and insights the main intention behind the research questions follows the idea of practical use of the standardized technologies without massive pre-considerations on mining infrastructure and without large scale tuning of algorithms. However, very useful insight from the referred section is the efficient use of similarity measures and Natural language Processing (NLP) methods that can be re-used in further experiments.
This chapter presents the methodology and architecture which includes semantic modeling of data, SPARQL based retrieval and interlinking with useful sources from Linked Data Cloud (section 4.3. The main source of data is Twitter with use cases focused on Profiling, Data Mining and Discovery and Semantic Search for researchers. As reliable and scientifically approved mining reference Conference Linked Data (COLINDA)[1] will be introduced in a separate section (see section 4.2). Due to specificity of data source (Twitter) and the time when I came up with the idea for implementing mining architecture and publishing COLINDA promising concepts like R2RML and RML introduced in related work subsections 2.6.6 and 2.6.7 for ETL (Extract, Transform, Load) work flow have not been presented yet to the scientific community. In particular, this chapter addresses research questions: RQ1, RQ2 and RQ3.

[1] http://colinda.org, last access: 2017–05–29

4.1 Why Modeling and Querying Tweets and Twitter User Profiles?

Although Twitter already has an API with advanced search functionality, retrieved data lacks of practical usability. Bringing these results into a structured form with appropriate domain description using wide accepted vocabularies for a specific knowledge domain would increase the relevance of information retrieved through mining and exploration of such content. Second disadvantage of Twitter API is that results were in the moment of experiment conduction restricted to the last 3200 tweets. The intention behind the following approach is to create an architecture adaptable to the flow of data produced by the users. This effort requires well thought mining infrastructure with reliable mining sources.

4.2 Modeling Scientific Events with Semantic Vocabularies and Possibilities of their Exploitation

4.2.1 Statement to Own Contribution

Texts and ideas for this subsection are from the workshop paper about COLINDA (see Softic et al. (2015b)). All work presented in the realization of COLINDA is my own work. My colleagues mentioned as co-authors especially Mr. De Vocht contributed to integration and implementation of use cases where COLINDA was used as as mining source for scientific events.

4.2.2 Introduction and Motivation

COLINDA[2] contains information about scientific events worldwide (including location and proceedings references), published as Linked Data. The data contained in COLINDA is extracted and accumulated from the data dumps of WikiCfP , which are published yearly and freely available on request for research[3] purposes, and from data gathered via JSON interface from Eventseer. WikiCfP and Eventseer are two very popular online scientific event archives. WikiCfP contains calls for paper for about approximately 30.000 conferences and has approximately 100.000 registered users. Eventseer contains according the latest information[4] calls for around 21000 events and serves more then 1 million users. Also included are the Twitter[5] feeds of both sites integrating on the fly arrival of upcoming scientific events using the Twitter API[6] to receive the data from Twitter profiles of Wiki CfP and Eventseer. Currently COLINDA includes data about more than 15000 conferences. Event instances are enriched through interlinking with information from Linked Data proceedings repositories DBLP (L3S)[7] and Semantic Web Dog Food[8] as well by location information from Geonames and DBPedia. Primary intention of COLINDA was to provide hashtag based identification system for scientific events in Twitter in the manner of the "5-star" quality Open Data[9]. Researchers are using very often hashtags, while they are discussing on Twitter. Specially during scientific events, they are using hashtags as abbreviated reference to the event they are attending (see Reinhardt et al. (2009a)). E.g. WWW (World Wide Web) 2013 is often referred as "www13" or "www2013". DBLP (L3S) Linked Dataset and Semantic Web Dog Food also use this kind of notation to reference the event of conference proceedings; e.g. for 'www2013' at DBLP(L3S)[10] and e.g. for

[2] Available at: http://colinda.org/, see also http://datahub.io/dataset/colinda, last access: 2017–05–29

[3] http://www.wikicfp.com/cfp/data.jsp, last access: 2017–05–29

[4] http://eventseer.net/data/, last access: 2017–05–29

[5] http://www.twitter.com/, last access: 2017–05–29

[6] http://dev.twitter.com, last access: 2017–05–29

[7] http://datahub.io/dataset/l3s-dblp, last access: 2017–05–29

[8] http://datahub.io/dataset/semantic-web-dog-food, last access: 2017–05–29

[9] http://5stardata.info/, last access: 2017–05–29

[10] http://dblp.l3s.de/d2r/page/publications/conf/WWW/2013, last access: 2017–05–29

'iswc2012' at Sematic Web Doog Food[11]. The overall idea of COLINDA is to serve as mining reference for creation of semantically driven microblog data Mesh-ups for Research 2.0 and as interlinking hub for other science relevant sources from the LOD cloud as well as to enhance and support exploratory search for researchers. Efforts made in this field using COLINDA will be introduced in detail in chapter 5.

4.2.3 Extraction, Modeling, Creation and Publishing of Linked Scientific Events

COLINDA data covers generally three domains: The first domain originates from WikiCfP and Eventseer and describes the **Conference** as basic scientific event with a start date, location, description, label and link to the event web page. Second domain is the **Location** of the event with geographic parameters resolved using the GeoNames[12] and DBPedia[13] data set in interlinking process. Each location contains reference to the city, country and coordinates of the location. Further, as extension and third domain we have **Proceedings** of the conference represented by the links from DBLP (L3S) or Semantic Web Dog Food.

Linked Scientific Events Creation Process

The data creation process comprises the following steps:

- Extraction - extraction and pre-processing of data sources (Subsection 4.2.3)
- Modeling of Events using SWRC Ontology - concept coverage (Subsection 4.2.3)
- Triplification - creating RDF data triples (Subsection 4.2.3)
- Interlinking - connection to other Linked Data sets (Subsection 4.2.4)

[11] http://data.semanticweb.org/conference/iswc/2012/, last access: 2017–05–29
[12] http://www.geonames.org, last access: 2017–05–29
[13] http://dbpedia.org, last access: 2017–05–29

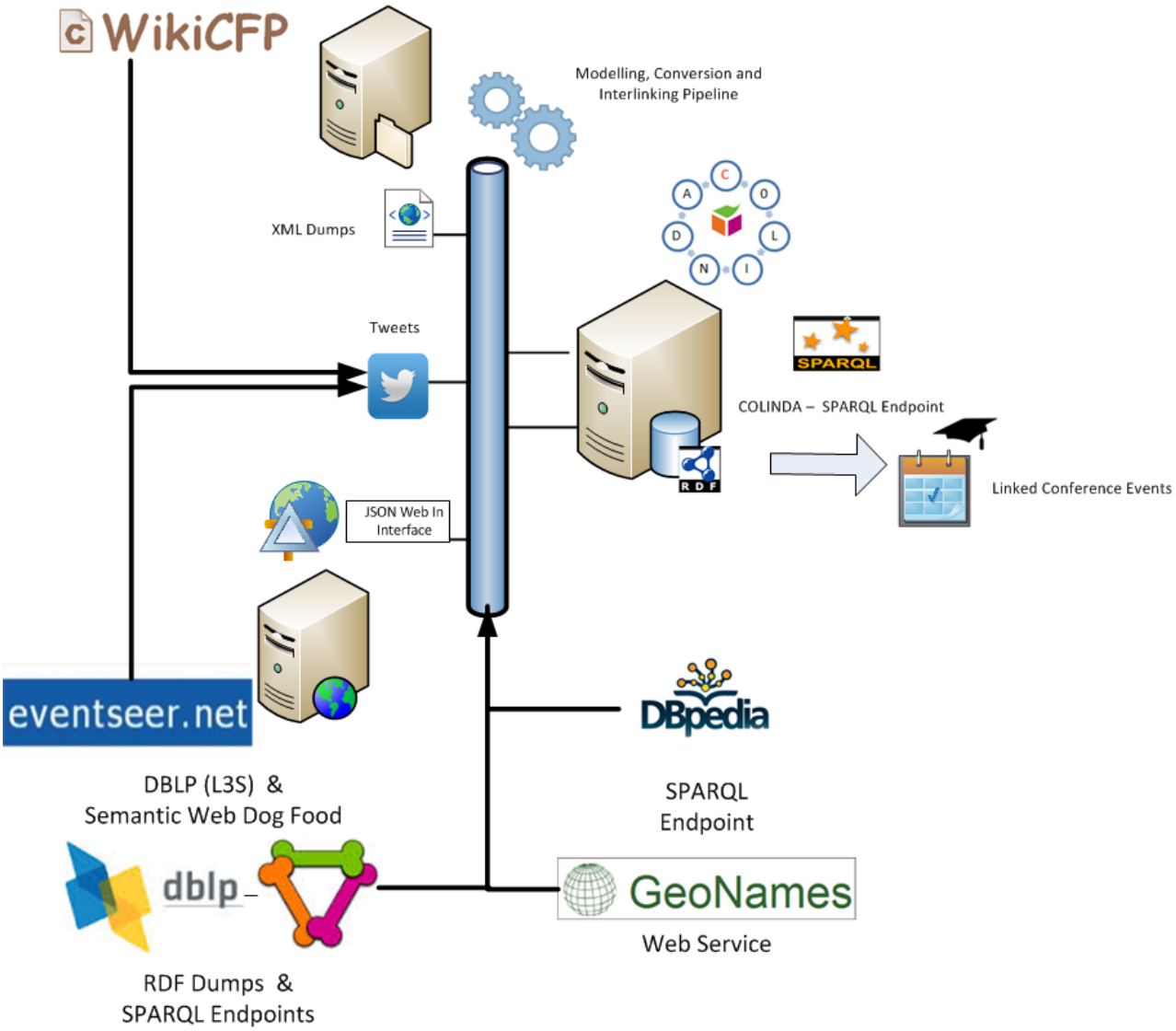


Figure 4.1: Creation process of linked scientific events.

Data Extraction

COLINDA is constructed from variously structured sources. Therefore a minimal set of properties has been defined that describe the minimum useful **Conference** concept for a single RDF instance. During extraction,

all properties from sources are being mapped to defined normalized set in order to harmonize the federated data. The **Location** and **Proceedings** concepts related to conference events as such are considered as optional enrichment which will be treated in the interlinking process. This decision was made having in mind that all conference descriptions do not explicitly include the venue information. The quality of source data depends on the users that provide the information. Thus such data sources implicitly exclude assumption of completeness. Table 4.1 represents the minimal set of properties a **Conference** and **Location** instance should include. The extraction process includes steps of either pre-processing of XML dumps from WikiCfP or JSON from Tweets and Eventseer into the temporary tables of values formatted as Comma Separated Value (CSV). During the pre-processing cycle data fields like e.g. date or labels are being normalized to achieve uniform representation, and to provide easier processable input for triplification step which converts the extracted values from temporary tables into RDF formatted instances of Linked Data.

Table 4.1: Harmonized COLINDA - minimal properties set. Entries denoted with * are optional.

Concept	*Property*
Conference	*label* *title* *description* *date** *link** *location**
Proceedings	*proceedings**
Location	*placename* *city* *country* *longitude* *latitude*

Modeling Scientific Events in the Web of Data

Basic representation of scientific events was well elaborated in previous research work about the SWRC ontology introduced by Sure et al. (2005). This practice has been already approved and adapted by the implementation

of Linked Data proceedings repositories DBLP (L3S) and Semantic Web Dog Food. Same vocabulary was used for COLINDA following the good practice of re-using existing vocabularies before defining own. Minimal field set defined in table 4.1 for RDF instance generation matches well the range of SWRC concepts. Therefore, the SWRC Ontology[14] and basic RDFS Schema[15] have been chosen as established vocabularies to describe **Conference** instances. The same approach was applied for **Location** concept; needed set of geographical features to describe conference venues is well covered by elements from GeoNames[16] and Basic Geo (WGS84) Vocabulary[17]. Complete model with interlinked properties (proceeding and location) can be seen in figure 4.2, where a single complete and interlinked instance of a conference (WWW2013) is depicted. Matching between features and the vocabulary properties is shown in table 4.2.

Table 4.2: COLINDA concept to ontology model mapping (note: geonames - GeoNames Ontology, geo - W3C GEO Vocabulary, swrc - SWRC Ontology). Entries denoted with * are optional.

Concept/Property	*RDF Class/Property*
Conference	swrc:Conference
label	rdfs:label
title	swrc:eventTitle
description	swrc:description
*date**	swrc:startDate
*link**	owl:sameAs
*location reference**	swrc:location
*location reference**	dcterms:spatial
Proceedings*	rdfs:seeAlso
Location*	geo:SpatialThing
*placename**	geonames:P
*city**	geonames:name
*country**	geonames:countryName
*longitude**	geo:long
*latitude**	geo:lat

[14] http://ontoware.org/swrc/, last access: 2017–05–29
[15] http://www.w3.org/TR/rdf-schema/, last access: 2017–05–29
[16] http://www.geonames.org/ontology/, last access: 2017–05–29
[17] http://www.w3.org/2003/01/geo/wgs84_pos#, last access: 2017–05–29

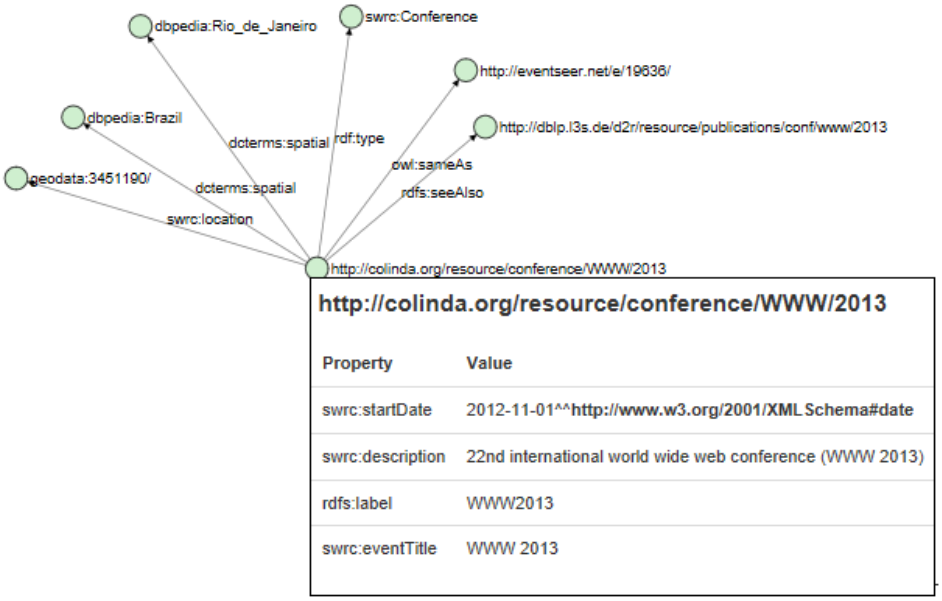


Figure 4.2: Sample interlinked **Conference** RDF instance of WWW 2013 generated by Visual RDF.

Triplification - Creation of RDF Instances of Scientific Events

The triplification[18] process uses as input temporary data tables in Comma Separated Values (CSV)[19] like format generated in extraction and pre-processing step. Input generated in this way represents tabular set of values compatible with properties from table 4.1. The temporary data table contains pre-processed and normalized data, which means the content was checked and treated for errors, misspellings and missing data. This input is then read line by line and conference instance is generated as single RDF graph using the vocabulary properties defined in table 4.2. Each conference instance is accessible via REST (Representational State Transfer) call as described in subsection 4.2.4. To make them accessible by SPARQL endpoint,

[18]Under 'triplification' we understand 'triple-wise' creation of Linked Data instances as RDF graphs.

[19]http://www.ietf.org/rfc/rfc4180.txt, last access: 2017–05–29

background batch process loads the conference instances into the ARC2[20] RDF triple store running on the server.

4.2.4 Interlinking to Other Interesting Sources

In order to provide 5-star data and led by the design issues described in Berners-Lee (2006), we used ***swrc:location*** as interlinking property in order to interlink the location data with GeoNames. The interlinking process uses GeoNames query service to resolve geographical information and retrieve coordinates. Although usually ***owl:sameAs*** is used to interlink to other data set we used this property to resolve the connection to the conference web page and since ***swrc:location*** seems regarding the GeoNames to be more appropriate choice. How this connection looks like can be seen in the sample depicted in figure 4.2 as well as online[21],[22]. Further, the dumps of DBPedia and Semantic Web Dog Food have been used to enhance the instances with DBPedia location info using the ***dcterms:spatial*** property and for interlinking the proceedings from DBLP (L3S) ans Semantic Web Dog Food we match the conference's ***rdfs:label*** to the corresponding labels in those data sets via SPARQL queries. In matching case a link is established with correlating results using the ***rdfs:seeAlso*** property.

URI Design and Public Accessibility

Access to instances of COLINDA is possible via URIs with following pattern:

- http://colinda.org/resource/conference/{label}/{year}

[20] https://github.com/semsol/arc2/, last access: 2017–05–29

[21] http://www.colinda.org/resource/conference/WWW/2013?format=html, last access: 2017–05–29

[22] http://graves.cl/visualRDF/?url=www.colinda.org/resource/conference/WWW/2013, last access: 2017–05–29

All responses from COLINDA are formatted as RDF/XML fragment. Other supported formats are: HTML, Text, N3, NTRIPLES format[23]. Alternative access offers the SPARQL[24] endpoint. Current endpoint supports up to 250000 result triples per query and delivers results in different formats like: JSON, RDF/XML, XML, TSV (Tab Separated Values) etc. How to query the endpoint is shown by simple example in listings 4.1. Results from the query return the COLINDA link, city, country and the geo-location of WWW 2013 conference.

```
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?x ?city ?country ?long ?lat
{
 ?x rdfs:label "WWW2013";
    swrc:location ?loc.
    OPTIONAL
    {
     ?loc gn:name ?city;
          gn:countryName ?country;
          geo:lat ?lat;
          geo:long ?long.
    }
}
```

Listing 4.1: Sample SPARQL query for retrieval of conference (geo) location. Adapted from Softic et al. (2015b).

Actuality of Data

COLINDA is kept up-to-date by a cron job which grabs the newest event announcements over the Twitter API for accounts of WikiCfP and Eventseer. The cron job parses, creates, interlinks and synchs new events. Each tweet also includes information about the call page link which allows retrieval of the extended information about events via web (WikiCfP) or available JSON (Eventseer) interface during the update task. The automated cron job collects and interlinks the data. The update of triple store is done manually, by scripts. Also manual (semi-automatic) updates are ran as soon as the fresh dumps from both sites are available.

[23] e.g. http://www.colinda.org/resource/conference/WWW/2013?format=html, last access: 2017–05–29

[24] http://colinda.org/endpoint.php, last access: 2017–05–29

4.2.5 Public Availability and Influence

Recently, a dump of COLINDA was made available as Linked Data Fragments[25] in order offer easier and faster mining. Linked Data Fragments use SPARQL query patterns and the infrastructure can be run local without huge technology requirements. One additional advantage of Link Data Fragments approach is that it provides a lightweight infrastructure with very good performance for mining COLINDA. COLINDA RDF data dumps are also accessible via the CKAN Registry[26] of LOD Cloud. COLINDA was also used as reference data set for interlinking task in Semantic Publishing Challenges 2015-2016 [27],[28] co-located at European Semantic Web Conference what additionally underlines its contribution to the scientific community around Linked Open Data.

4.3 Semantic Modeling and Mining Researchers Profiles from Twitter

4.3.1 Statement to Own Contribution

Concepts and insights from this section are extended version of previously published work in Softic et al. (2010) which is partly extended in De Vocht et al. (2011); Softic et al. (2013a). Overall idea of semantic modeling and mining of researcher profiles from Twitter as well as test implementation and evaluation of the experimental mining architecture is done solely by author of this thesis.

[25] http://data.linkeddatafragments.org/colinda#dataset, last access: 2017–05–29
[26] http://datahub.io/dataset/colinda, last access: 2017–05–29
[27] https://github.com/ceurws/lod/wiki/SemPub2015, last access: 2017–05–29
[28] https://github.com/ceurws/lod/wiki/SemPub2016, last access: 2017–05–29

4.3.2 Introduction

Microblogs in the Web 2.0, as short form of blogging, gained strong popularity and importance in recent years. Microblogging platforms like Twitter attract daily many users with different social, cultural and educational backgrounds. While tweeting users share their emotions, opinions or commonly useful information. The possibility of monitoring such content, not only for humans but also for machines, contributes to creation of more intelligent user interfaces, better common information awareness, and to more technically profound agent, search and recommendation systems. This part of the chapter aims to answer the questions: whether it is possible to use semantic technologies and Linked Data to mine useful tacit information from microblogs, in our case in particular Twitter, as well to describe an architecture which serves such purpose. To approve the presented approach an experimental setup of such architecture has been implemented and tested. Finally the achieved results have been analyzed for future research in this field.
Following subsections are organized as follows: First, a short review on related work has been done. Then, the holistic architecture for mining Twitter based upon semantic technologies and natural language processing techniques is introduced. Architecture building blocks have been used to explain stepwise the method of information processing, operational components, and interfaces of the proposed mining architecture. Based on a real world use case identified by research community, a proof of concept is presented that was implemented for mining scientific events visited by the Twitter user describing them as researchers. The tweets and Twitter user information the experimental prototype used was archived in experimental local storage developed by Graz University of Technology called Grabeeter (see details in 2.5.2. Finally this part of the thesis is closed up with discussion of the results and future steps.

4.3.3 Twitter Usage at Scientific Conferences

Although the beginning of first serious micro blogs dates back couple of years ago their leverage on the web grows rapidly Zhao and Rosson (2009); Boyd et al. (2010). Most significant among them is Twitter, which induced a new culture of communicationMcFedries (2007); Java et al. (2007). In 2013 Twitter generated in average 500 million Tweets a day with 100 million active users daily[29]. Java et al. (2007) defined four main user behaviors why people are using Twitter - for daily chats, for conversation, for sharing information and for reporting news. Usage of Twitter at conferences helps to increase reports, statements, and announcements as well as supports fast conversation between participants. Nowadays very often so-called Twitter-streams done by hash tag search nearby projection of an ongoing presentation (see Reinhardt et al. (2009b)) or placed at any other location at the conference support the conference administration, organization, discussions or knowledge exchange. As reported in already in related work chapter micro blogging and as such also Twitter became a valuable service reported by different publications (see Reinhardt et al. (2009b); Ebner et al. (2010a); Priem and Costello (2010); Letierce et al. (2010b)).

4.3.4 Modeling Context

Semantic Web Community provides a set of widely used schema (a.k.a vocabularies) useful to cover the description of micro blog posts and user profiles attached to them. FOAF (Friend of a Friend) vocabulary describes semantically user and relation between them (see Miller and Brickley (2010) for details). SemanticTweet[30] and Twitter-Based User Modelling Service (TUMS) introduced by Tao et al. (2011) provide infrastructure to store the profiles data from Twitter in form of RDF (Resource Description Framework) graphs useful for further analysis. Very commonly used vocabulary for description of posts is SIOC (Semantically Interlinked Online Communities) Breslin et al. (2005, 2006a). For tag binding the Modular Unified Tagging

[29] http://tinyurl.com/mbkv9t6, last access: 2017–05–29
[30] https://github.com/sflinter/semantictweet, last access: 2017–05–29

Ontology (MUTO)[31] introduced by Lohmann et al. (2011) combines the best approaches from earlier efforts on defining a tag ontology.

4.3.5 Mining Architecture

Mining architecture approach introduced in this subsection makes use of semantic technologies (RDF, SPARQL), Web of Data and simple Natural Language Processing methods like similarity measures and regular expressions.
First a generic structure of the architecture is introduced, and then, based upon modular construction, a stepwise method of information processing is explained together with operational components and in between results, and interfaces of the architecture. The main objective of presented experiment is to provide an intelligent data source useful for data mining, information and context profiling or further use in recommendation and search systems.

Proposed architecture consists of four main modules:

1. **Harvesting module** which is dependent on data source (here Twitter)
2. Module for converting the retrieved data into semantically described triples (**Triplifier module**)
3. **Interlinking module** that aligns the generated semantic links to the Linked Open Data sets.
4. **Querying interface** - e.g. consumer API, or SPARQL endpoint. interface

Figure 4.3 shows the architecture in general (applied on Twitter). Following subsections describe the building modules more into detail.

Harvesting Module

Harvesting module operates the pre-step before conversion of data into semantic form of RDF triples, by gathering the data from the native microblog platform. Hereby, the module uses commonly provided REST API

[31] http://muto.socialtagging.org/core, last access: 2017–05–29

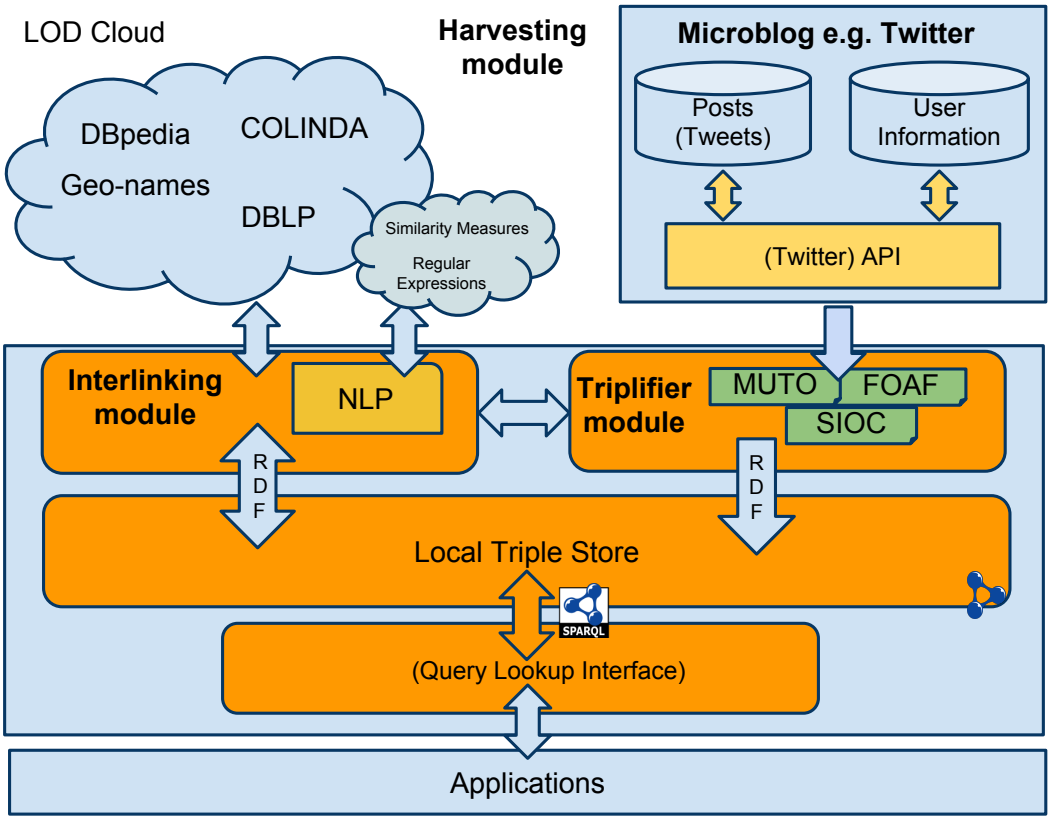


Figure 4.3: Architecture for mining microblogs (applied on Twitter use case). as from Softic et al. (2010)

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix sioc_t: <http://rdfs.org/sioc/types#> .
@prefix dcterms: <http://purl.org/dc/terms/#> .

<https://twitter.com/laurens_d_v/status/334401818572492801>
    rdf:type  sioct:MicroblogPost ;
    sioc:content
    "My presentation at #www2013 #ldow2013 in Rio
    about the engine behind #mmlab's Everything is Connected
    on @slideshare
    http://www.slideshare.net/laurensdv/discovering-meaningful-connections-between-resources-in-the-web-
        of-data" ;
    sioc:has_creator  <https://twitter.com/laurens_d_v/> ;
    foaf:maker  <https://twitter.com/laurens_d_v/> ;
    dcterms:created  "2014-05-14" ;
    sioc:topic <#www2013>

<http://twitter.com/laurens_d_v/>
    rdf:type foaf:Person ;
    foaf:name  "Laurens De Vocht" ;
    foaf:knows <http://twitter.com/selvers>
    foaf:knows <http://twitter.com/mebner>
    foaf:depiction <https://pbs.twimg.com/profile_images/1131586475/sshot-picsmal.png> .
```

Listing 4.2: Sample corresponding metadata from sample tweet as RDF triples in N3 notation.

with restricted interfaces (in this case Twitter API) for this purpose. The data set gained from Twitter API is usually in JSON or XML format. This module also parses and formats the data before transformation in semantic from. Collection of micro posts (tweets) by **Harvesting module** happens user wise. After retrieval and reformatting data **Harvester module** passes it further to the **Triplifier module**.

Triplifier Module

Transformation of microblog posts (tweets) and user information, into RDF triples **Triplifier module** uses following vocabularies: SIOC to describe the post structure, FOAF to express the person information. This method of conversion was also approved in Passant et al. (2008); Softic et al. (2010); De Vocht et al. (2011). Exemplary result of "triplification" (conversion of tweet information into RDF triples) process is shown in Figure 4.2. This listing describes shows a single tweet along with information about his creator.

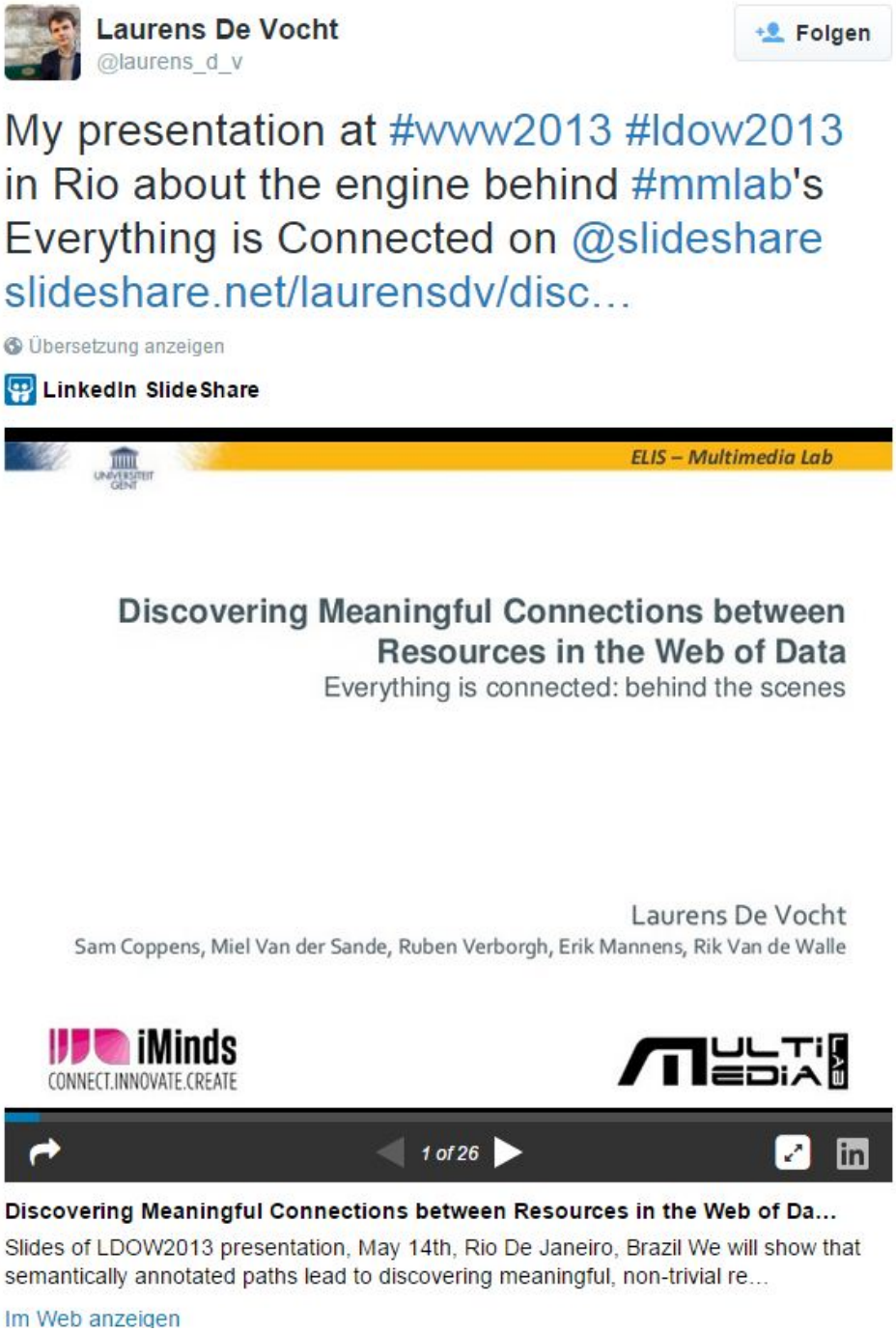


Figure 4.4: Sample tweet demonstrating event reference.

```
@prefix muto: <http://purl.org/muto/core#> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<#www2013>
    rdf:type muto:AutoTag ;
    muto:taggedResource <https://twitter.com/laurens_d_v/status/334401818572492801> ;
    muto:hasCreator <http://www.twitter.com/laurens_d_v> ;
    muto:tagLabel "www2013" ;
    muto:tagMeaning <http://www.colinda.org/resource/conference/WWW/2013> .
```

Listing 4.3: Aligning tags to tweets and interlinking meaning to tags.

Interlinking Module

Interlinking module queries the local RDF store and retrieves the text of posts. NLP technologies like stammers or regular expressions extract in interlinking process hashtags from tweet body. Module uses the extracted hashtags to lookup for the fields of content in DBPedia, COLINDA, Geonames graphs or similar knowledge bases from Linked Open Data Cloud. In case of match and module creates 'owl:sameAs' links with the corresponding property. During the mapping and interlinking process similarity measures on character and token level check whether or not a link and its description is an appropriate match for interlinking and for ambiguity resolution. Listing 4.3 shows discovered match for hashtags from sample listing before. MUTO instances bind hashtags from Twitter with entities in Linked Data repositories.

Tagged resource offers additional useful information linked to other Linked Open Data bases as shown on example in listing 4.4.

The main intention of interlinking is to enhance the value of searchable content as well to offer verification of search results for a specific domain search. For instance as the ontologies own concept as classes and relation are defined as properties these can be used for faceted or exploratory search or as single categories in a recommendation system.

```
@prefix swrc: <http://swrc.ontoware.org/ontology#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://colinda.org/resource/conference/WWW/2013>
  a swrc:Conference ;
  rdfs:label "WWW2013" ;
  swrc:location <http://sws.geonames.org/3451190/> ;
  swrc:eventTitle "WWW 2013" ;
  rdfs:seeAlso <http://dblp.l3s.de/d2r/resource/publications/conf/www/2013> ;
  owl:sameAs <http://eventseer.net/e/19636/> ;
  dc:spatial <http://dbpedia.org/resource/Brazil>, <http://dbpedia.org/resource/Rio_de_Janeiro> ;
  swrc:startDate "2012-11-01"^^xsd:date ;
  swrc:description "22nd international world wide web conference (WWW 2013)" .
```

Listing 4.4: Information behind the tag meaning.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/> .
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
PREFIX muto: <http://purl.org/muto/core#> .

SELECT ?name WHERE
{
 ?x rdf:type muto:AutoTag ;
    muto:tagLabel "www2013" ;
    muto:hasCreator ?creator .

 ?creator  foaf:name ?name .
}
```

Listing 4.5: SPARQL Query: Tracking all persons who tagged conference www2013.

Querying and Lookup Interfaces

In case that the architecture should be offered as service to machines (intelligent agents, search interfaces and recommendation structures) a simple exposure of SPARQL endpoint is sufficient. SPARQL endpoints offer beside RDF also JSON and XML retrieval formats. All these measures deliver possibility to explore the implicit information of structured tweets using simple text search or specified selective filtering of desired information. Data collected in this way offers also additional insight on related content which is explicitly named inside the text of micro blogs. To point out how data from this architecture can be retrieved in a selective manner we use the listing 4.5 that pictures in a best way with a simple sample of a SPARQL query which retrieves all person names who tagged in their tweets the World Wide Web 2013 conference.

4.3.6 Proof of Concept Experiment

In order to demonstrate the functionality of described architecture a focused use case identified by recent research on usage of Twitter in scientific community has been chosen for the proof of the concept. Proposed use case was derived from how the researcher use the Twitter to communicate with each other and express opinions and impressions about certain scientific events they are attending. The question which the experiment is trying to answer is whether is it possible to use the proposed architecture to determinate which conferences the researcher has visited based upon users Twitter profile and tweets, how accurate is the identification. Further we want to answer which conferences, that he missed or could visit, and which persons could be suggested to the researcher by the infrastructure? The latter is thinkable as extension to results retrieved by query in listing 4.5 where all conferences visited by corresponding users could be selected and offered as potential point of interest.

Use Case

Researchers are focused in their work on specific areas and conferences that deal with achievements related to them. We can make an assumption that their profiles are also characterized by mentions about conferences they are tweeting about. This trend is especially very strong in technical researcher communities according to Ebner et al. (2011). Researchers tweet about what they have noticed at such events as well what they have remarked as interesting regarding their own interest. This finding in our case should serve as testing ground for the reliability of proposed infrastructure for knowledge mining and context enhancement.

Methodology and Data Set

Out of base of more then 4000 users and more than one million tweets in local twitter data storage from database 52 different user profiles have been selected by queries and regular expression knowing that they contain hash-tags having the form that corresponds the conferences tags (a sequence of

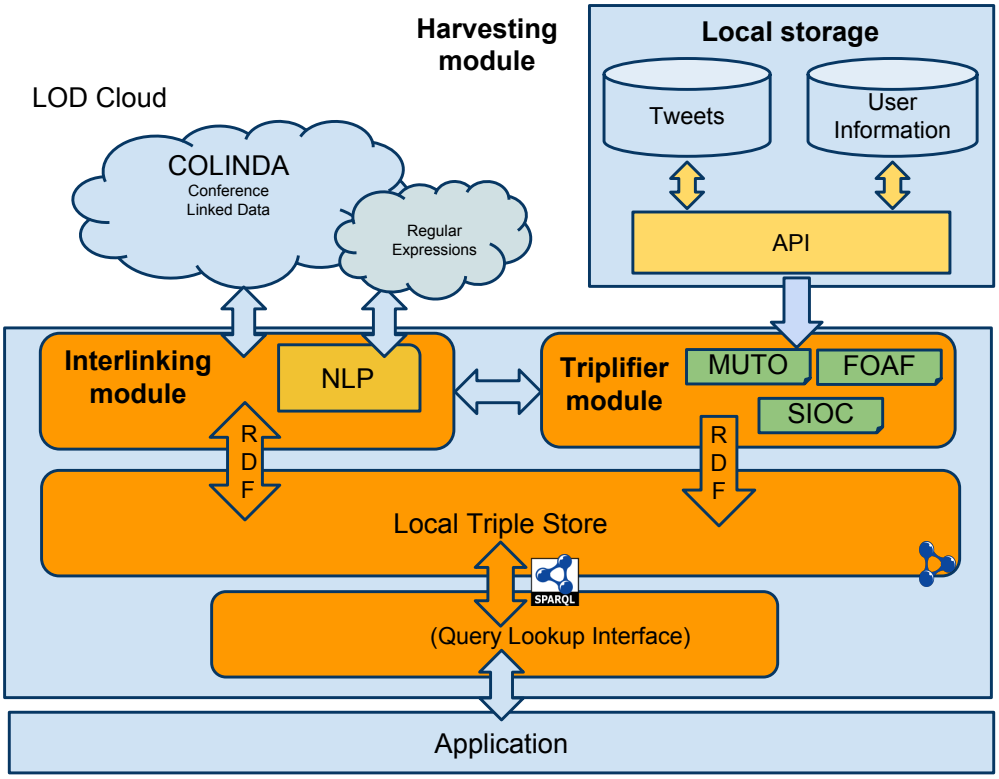


Figure 4.5: Experimental Architecture derived from Figure 1 using COLINDA Linked Data set.

chars followed by a number e.g.WWW2013, WWW13) and that their description of profile contains the word 'researcher'. We archived picked tweets having potential conference hashtags user-wise and user profile information into an experimental local store for our proof of concept implementation. Detection Linked Data set used in this experiment is COLINDA Softic et al. (2015b). COLINDA Linked Data set contained at the moment of experiment information about 15000 mainly technical conferences from 2006 up to 2015. Our experimental architecture (concept see Figure 4.5) extracted personal information and hashtags and modeled this data into semantic form as described in "Triplifier Module" in section 4.3.5. For each hash tag that was extracted from tweet body the interlinking module starts a run which uses regular expression to pre-filter conference-like hashtags, and COLINDA SPARQL endpoint as query interface for detection of conferences corresponding to the hashtags.

Measures

Definitions represented in equation bellow express precision, recall and F-measure as combination of *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN) results. The results are interpreted as follows: each detected and verified conference hashtag (by the correctness of link from COLINDA data set) is treated as true positive. Each correctly undetected hashtag is true negative. Each correctly undetected event hashtag due to algorithm or data set lacks is treated as false negative. Falsely identified links which do not correspond to the detected event are false positives. All event hashtags have been extracted user wise that correspond the pattern #+[a-z]+[0-9] or in other words char sequence followed by numbers sequence e.g. WWW2015 or WWW15.

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{4.2}$$

$$F - Measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.3}$$

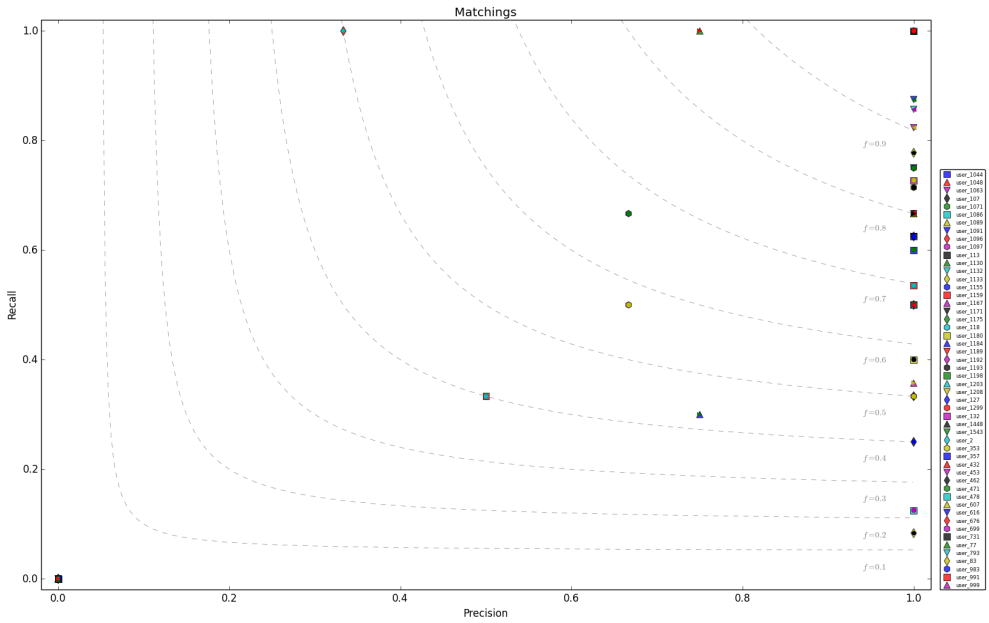


Figure 4.6: Precision-Recall diagram with F-measure.

Evaluation of Results

Whole evaluation is done user wise for the group of 52 user previously selected and identified as potential candidates having hashtags corresponding the reference to a scientific event. For each user, potential conference hashtags have been extracted and semantic representation of them has been created. Those then were matched through interlinking module with COLINDA. After this step for each user the connected tags have been verified and not de-referenced tags have been checked manually once again. A user with the set of a pre-filtered potential conference hashtags as shown in listing 4.4 represents a matter of a single query. Based upon how the de-refencing, respectively to the total number of hashtags, has been user-wise calculated the precision, recall and F-measures values.

Figure 4.6 represents the results user-wise for the 52 selected test users and all hashtags found in their tweets. As we can observe the overall precision

with couple of outliers lies between 40% and 100% while by the majority of of users (approx. 75% of them) the precision lies between 60% and 100%. Very promising is the fact that 65% of the users tested reached a precision of 100%. The precision in this case is an indicator for the method which verifies the approach of the architecture at least for the given test. Regarding recall which is an indicator for the detection process and quality of mining source (in this case it is COLINDA) we could observe following results: 88% of all tests perform an recall between 30% and 100%. Around 48% of all test user accounts achieved a recall above and equal to 60%. This makes COLINDA a very solid base for data mining, the same applies for the detection process. Of course there is a plenty of place for improvements through enhancement of COLINDA but also in the detection process of interlinking by tuning the used regular expressions. Overall F-measure ranges for the majority from 40% up to 100%. This group represent 73% of all tested users. Around 55% of all tested users belong to the group having F-measure higher or equal 60%. All outliers mentioned are caused by falsely identification of hashtags as conference hashtags without being it and by the lack of mining source for some existing conferences which are not within COLINDA.

4.4 Conclusions on Findings in this Chapter

At first glance, proposed approach of mining architecture performs quite precise. The precision for 75% of test users lies between 60% and 100%, 65% reached the 100% precision. Also the recall as an indicator for the detection process and quality of mining source COLINDA was very promising and lies between 30% and 100%. Around 48% of all test user accounts achieved a recall above and equal to 60%. Altogether, the first evaluation results indicate the possibility of exploitation of proposed approach as integral part of interactive Research 2.0 related applications, frameworks and interfaces as well as part of exploratory search systems for researchers. This is exactly what the following chapter 5 tries to evaluate through series of implemented prototypes based partly upon findings from this chapter.

5 Exploitation of Proposed Approaches for Research

This chapters represents the practical implementations of concepts presented in chapters 3 and 4 and their integration into more complex systems like interface mesh-ups, semantic search systems and advanced semantically driven user interfaces for researchers that support and enforce exploratory search overall. Research questions addressed with this chapters are: RQ1, RQ2 and RQ3.

5.1 Research 2.0 Mesh-up Application and Framework to Match Affinities between Researchers

5.1.1 Statement to Own Contribution

Ideas, texts, solutions and results introduced in following subsections originate from De Vocht et al. (2011) written and published by Mr. Laurens De Vocht and myself as co-author. The publication presents the results of Mr. Devocht's master thesis which I co-supervised together with my mentor Dr. Martin Ebner from Graz University of Technology and Prof. Erik Duval from University of Leuven. My own contribution to the the work were in developing the overall idea of use case, architecture, modeling of semantic data, design of user interface and its evaluation and implementation of the proposed solution as adviser. I also co-authored introduction, related

work and discussion of this publication as well did the proof-reading. Implementation work on user interface and web service as well conduction of evaluation was completely done by Mr. De Vocht as part of his master thesis.

5.1.2 Introduction and Goal Definition

This section introduces a framework to address an important issue in the context of the ongoing adoption of the "Web 2.0" in science and research, often referred to as "Science 2.0" or "Research 2.0". A growing number of people are linked via acquaintances and online social networks such as Twitter allows indirect access to a huge amount of ideas. These ideas are contained in a massive human information flow (see Jansen et al. (2009)). That users of these networks produce relevant data is being shown in many studies Reinhardt et al. (2009b); Java et al. (2007); Rowe and Stankovic (2010). The problem however lies in discovering and verifying such a stream of semi-structured text fragments. Another related problem is locating an expert that could provide an answer to a very specific research question. Solution presented here is using semantic technologies (RDF, SPARQL), common vocabularies (SIOC, FOAF, SWRC) and Linked Data (DBpedia, GeoNames, COLINDA) introduced by Auer et al. (2008); Berners-Lee (2006); Bizer et al. (2007); Softic et al. (2015b) to extract and mine the data about scientific events out of context of micro blogs, in particular Twitter, as proposed in Softic et al. (2010). Within this process the presented approach aims at identifying researchers and information related to them based on entities of time, place and topic. The framework provides an API that allows quick access to the information that is analyzed by the system. As a proof-of-concept an implementation and evaluation of a researcher profiling use case has been done. It involves the development of a framework that focuses on the proposition of researches based on topics and conferences they have in common. A demonstration application: "Researcher Affinity Browser" shows how the API supports developers to build rich internet applications (mesh-ups) for Research 2.0. This application also introduces the concept 'affinity' that exposes the implicit proximity between entities and users based on the content users produced. The usability of a demonstration

application and the usefulness of the framework itself are investigated with an explicit evaluation questionnaire. This user feedback led to important conclusions about successful achievements and opportunities to further improve this effort.

5.1.3 Use Case on Researcher Profiling

One of the most visible trends on the Internet is the emergence of "Social Web" sites. Social interactions with people who share affinities with you can support progress in research and learning as Mejas (2005) reported in his work. Scientists and researchers are interested in very specific topics; this is best verified by the conferences they are attending. Another trend is that many of them blog and tweet about these events, especially in communication and technical research communities (see Ebner et al. (2010a); Reinhardt et al. (2009a)). This creates huge opportunities for profiling. The attendees tweet about what they notice, what they notice as interesting for their own projects (see Zhao and Rosson (2009)). What if we could connect these users using this information? We could call an application that does just that "Researcher profiling". This approach comes from the concept that the data produced in social networks can have true value if properly annotated and interlinked. This can be done by choosing community approved ontologies and Linked Open Data resources. A second requirement is to create a suitable interface in which this information can be explored. The user interface and the data quality merged together will determine the user satisfaction for this system. Assumed that researchers want to find either: interesting events, to which many people in their field of interest are going to; people, based on matching interests or events; new challenges such as companies, organizations, topics which are related to events and people this scientist is interested in. This application is also “Researcher Profiling” as shown in Figure 5.1.

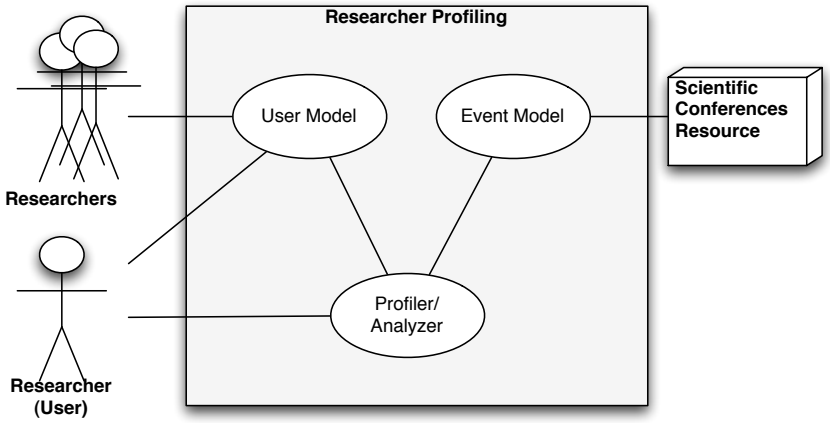


Figure 5.1: Researcher Profiling use case diagram. Adopted from De Vocht et al. (2011).

5.1.4 Data Source and Semantic Technologies for Profiling

At Graz University of Technology a tool called Grabeeter was implemented for storing and caching social data from Twitter. Grabeeter developed by Mühlburger et al. (2010) is an application that allows you to search tweets of single Twitter users online and offline. In contrast to the Twitter API, Grabeeter provides all stored tweets and makes no restriction over time. Details on Grabeeter are described in section 2.5.2. The Semantic Web Technology stack (see also 2.6.1) is well defined and applying frameworks such as SIOC (Semantically Interlinked Online Communities) introduced by Breslin et al. (2005) and FOAF (Friend-Of-A-Friend) introduced by Brickley and Miller (2004) can lead to a an interlinked and semantically rich knowledge source according to Bojars et al. (2008a). This approach has been experimentally approved for Twitter by findings also described in section 4.3. The knowledge source intended to be realized in this experimental implementation will be built with user profiles and the content they produce on various social networks as a basis (primarily Twitter but also some other like Mendeley[1] is planed for future efforts). To achieve this goal a certain process has to be run through which is basically realized in three steps. The first step is referred to as 'triplification' or 'rdfization', Data is extracted and annotated with the help of domain vocabularies and

[1] https://www.mendeley.com/, last access: 2017–05–29

ontologies. The triples that result are then being stored and made accessible as Linked Data in the second step. The third and final step in the process is the publication of the data URIs in various RDF formats or exposure of data as a query-able SPARQL endpoint initially presented by Bizer et al. (2007); Oren et al. (2008). Overall idea of presented use case is aiming to provide a scientific architecture paradigm for building semantic applications that rely on social data. Furthermore this section describes the architecture for the framework that builds on semantic social data layer. It aims to gain more knowledge and mine usable data out of the social context of micro blogs as Twitter Reinhardt et al. (2009a); Java et al. (2007); Letierce et al. (2010a); Boyd et al. (2010); Honeycutt and Herring (2009). In order to verify the data mined using "hashtags" and simple regular expressions the framework runs verification queries against conference knowledge represented through COLINDA introduced in previous chapters (see 4.2 and 4.3 and in Softic et al. (2010, 2015b)). COLINDA contains data from WikiCfP and Eventseer described with the SWRC vocabulary (see Sure et al. (2005)). Those are two currently very popular online Web 2.0 sites containing data about calls for papers, locations and topics of conferences as well as some other meta-data that can be used for the identification and verification tasks within the profiling and mining process. Additional enhancement and verification of data is planned to be done using DBPedia. Since we are getting information from a social data source where scientific significance is essential, we can call this analysis "Researcher Profiling". The results of the social data 'triplification' are still not proven accessible to researchers without the need for an expert with extended information mining skills. Researcher profiling could help in the understanding of scientific relevance and importance to other researchers specific needs; this is because the extent of social network data is massive and individual researchers are only likely to be interested in specific parts of the overall knowledge on the basis of their area of specialization. Online social connections can be built around common entities that the users link to. At the same time it will create new opportunities to co-relate existing Research 2.0 integration efforts and applications (see Bojars et al. (2008a)).

5.1.5 Framework

Following subsection describes the principles and structures of the proposed and realized mining framework.

Architecture

Figure 5.2 presents the global system architecture. First data from social networks will be aggregated and archived or cached. This data is annotated and stored as triples in a RDF (triple) store (graph based database for RDF data) and exposed for querying. Thus the data becomes available as Linked Data. To improve the quality of the Linked Data, the stored triples are interlinked with the Linked Open Data Cloud (sources like COLINDA or DBPedia). To make an abstraction of the complex linked data graph structure, the analysis of the data is done by using SPARQL queries tries to make the data available as scientific relevant information in common forms like tabular data (CSV) or in JSON or XML format. In this way top-level knowledge discovery applications can benefit from framework outputs. The implementation of the framework as a proof-of-concept was limited to what is marked green (b) in Figure 5.2. The Grabeeter aggregates data from Twitter. This data is used in **Researcher Profiling Framework**, which consists out of two important parts: the **Semantic Profiling Network** and the **Researcher Profiling API**. The Semantic Profiling Network is the collection of annotated and interlinked data that comes from Grabeeter. The Researcher Profiling API allows analyzing and retrieving targeted information and retrieving it for further use.

Implementation

The main intention behind framework is to support a researcher profiling application that meets the requirements of the use cases presented in 5.1.3. The mining infrastructure of the framework is based on my research work conducted at Graz University of Technology and published in Softic et al. (2010). It is also described in 4.3. The mining architecture design consists

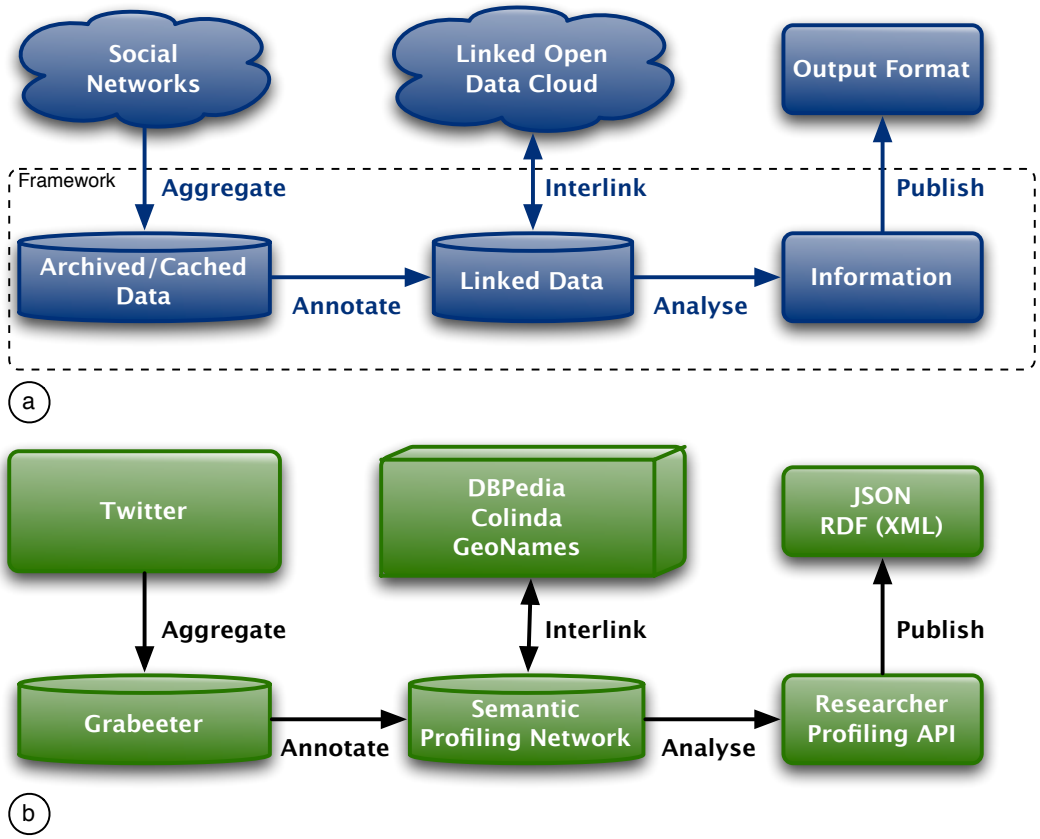


Figure 5.2: The general solution (a) and the implemented solution (b), as in De Vocht et al. (2011).

out of three modules: a data extraction module, an interlinking module and an analysis module.

- **Extraction / Harvesting Module**: Extracts data from various resources and annotates it using relevant ontologies for that specific data context.
- **Interlinking Module**: Enhances the annotated data (triples) and creates a SPARQL endpoint for it. It is responsible for requesting more data if needed for certain information query. It parses high level requests and translates them to SPARQL queries. The results are then being returned to "Analysis Module".
- **Analysis / Querying Module**: In this module user needs expressed as API requests are interpreted and translated into SPARQL queries which are further processed by the "Interlinking Module". The results are then combined, formatted and returned to user of the upper level application. This module also contains some metrics to rank and evaluate the returned results.

In addition to mining infrastructure a programming interface to this framework is provided. The extraction module loads triples into the RDF store on user request ("add me to the system") or in a periodical script that keeps the data up-to-data. The interlinking module communicates with the RDF store using SPARQL Queries. This structure between the modules and the RDF store is depicted in the design diagram in figure 5.3. The extraction module collects data of a person from Grabeeter. This data is requested directly from the Grabeeter database using MySQL queries. It parses the user profile and if a profile does not exist in Grabeeter then the user request is queued on a list to be analyzed in the future. Data for that user account will be available at later time. It annotates the data using relevant entities from ontologies. The result of the extraction is a collection of annotated data in the form of triples, which are finally stored in a RDF Store. The Interlinking module accesses the stored triples created in the "Extraction Module" and provides it on demand to upper level operational layer of the API. Frankly, it is impossible to create a generic framework that supports all thinkable data contexts, but the intention behind the framework ist to cover a broad range of data contexts for the pruposes od Research 2.0. For now the focus is on three data contexts:

- **User:** Social microblogs, more precisely annotated data from Twitter

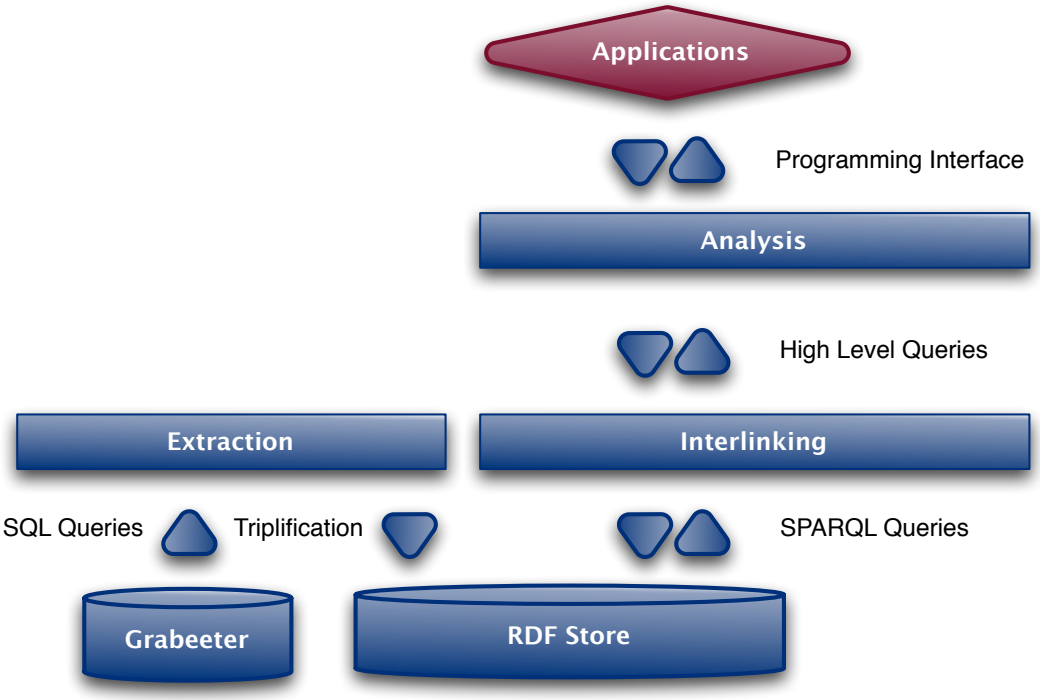


Figure 5.3: The scientific profiling network design., adopted from De Vocht et al. (2011).

```
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX gn:   <http://www.geonames.org/ontology#>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo:  <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:  <http://www.w3.org/2002/07/owl#>

SELECT * {
 ?x rdfs:label "I-KNOW2010";
 OPTIONAL
 {
 ?x swrc:description ?des;
    swrc:keywords ?key;
    swrc:startDate ?st;
    swrc:endDate ?end;
    swrc:location ?loc;
 }
}
```

Figure 5.4: Querying conference data., adopted from De Vocht et al. (2011).

users (SIOC, FOAF, DublinCore) necessary for user profiling.

- **Domain:** Annotated data of scientific conferences (COLINDA) to enable the framework to recognize and link to conferences and scientific events.
- **General:** OpenCalais, Linked Data, CommonTag Ontology (with links to DBPedia, GeoNames) to give a meaning to hashtags from a user. At the moment, the number of tags that can be linked to DBPedia is mostly attached to user profile. However, for most of the users there are almost no references in DBPedia.

Extending the framework with more domain knowledge could quickly increase the number of applications which this framework could support. In current state the "Researcher Profiling Framework" supports most thinkable "Research 2.0" use cases. They are very similar to the two use cases presented in this section. It is about discovering new resources created by researchers and of course researchers and scientific events themselves.

For example "Conferences" are interlinked through following process: first a SPARQL query to the COLINDA repository retrieves properties by the conference 'code' (used in the user's hashtag) like depicted in Figure 5.4. Additionally if existent also location data is fetched using SPARQL queries

```
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX gn:   <http://www.geonames.org/ontology#>
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo:  <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:  <http://www.w3.org/2002/07/owl#>

SELECT * {
 ?x rdfs:label "I-KNOW2010";
    swrc:location ?loc.
 OPTIONAL
 {
 ?loc gn:name ?city;
   gn:countryName ?country;
   geo:lat ?lat;
   geo:long ?long .
 }
}
```

Figure 5.5: Querying conference location, adopted from De Vocht et al. (2011).

```
INSERT INTO <root>{ <root/tags/LREC2008> rdf:type swrc:Conference }
INSERT INTO <root>{ <root/tags/LREC2008> ctag:means ?s }
INSERT INTO <root>{ <root/tags/LREC2008> swrc:location ?y }
```

Listing 5.1: Adding meaning and location reference. Adopted from De Vocht et al. (2011).

like in Figure 5.5.

The results are then interlinked to the tags by storing the properties as additional triples in the RDF Store.

The "Analysis Module" consists from several PHP classes that perform algorithms on the Linked Data graph. This graph is created during extraction and interlinking phase. The algorithms basically translate the high level API requests into chains of SPARQL queries.
The higher level API and web services don't have to care about the underlying semantic graph structure of the data. The functionality to find users and events is based on several query parameters such as the user or event itself, date and location. The ranking is determined according to a basic metric: the number of entities that two resources have in common (see Shinavier (2010)). An example translated high level query that returns a ranking of users by common tags with the given user as argument (Figure 5.6). A

```
SELECT DISTINCT ?z COUNT(?l) as ?tags WHERE
{
   ?t sioc:has_creator <user_uri> .
   ?t sioc:tagged ?y .
   ?y ctag:label ?l .
   ?tw sioc:tagged ?y .
   ?tw sioc:has_creator ?z
}
GROUP BY ?z
ORDER BY DESC(?tags)
```

Figure 5.6: Similarity query for ranking by common tags, adopted from De Vocht et al. (2011).

similar function also exists for entities, mentions and friends.

API as REST Web Service

The web service on the top of the Researcher Profiling Framework is made available as an API (see Table 5.1). The calls can be done by REST HTTP GET calls that return a JSON object. There are four main API calls that can be made, each with their specific parameters. Application developers can ask the profile for a specific person. Furthermore, users can find out relevant events or persons given the name of a specific person. There is also a possibility to discover the most popular mentions, events and friends. There is an API call to register a new user and it is possible to retrieve details about an event such as the URL and the users referring to that event.

API function Parameter	User Profile Allowed Values	profile Action
user	<screenname>	Returns the User Profile as JSON
uri	user_id=<uid>	Returns the User Profile as JSON
tweet	NONE	Same as above, but includes the latest tweet
API function	Event Details	event
Parameter	Allowed Values	Action
code	<eventcode>	Returns the Event Details as JSON
users	NONE	Same as above, but includes the users mentioning the eventcode
API function	Discovery	discovery
Parameter	Allowed Values	Action
find	persons|events|popular_friends|popular_mentions|popular_events	Returns the most frequent entities according to given find value
user	if(persons|events) <screenname> else NONE	Same as above, but returns only persons for the given user
API function	Register	register
Parameter	Allowed Values	Action
user	<screenname>	Returns whether the user registration is successful or not
Other API functions Parameter	 Allowed Values	 Action
All Screens	alluserNames	Returns a list of all user screennames
All Uris	allusers	Returns a list of all user uris.

Table 5.1: Web Service API Overview Table as published in De Vocht et al. (2011).

5.1.6 Evaluation and Results

Having researcher profiling use case in mind user interface was implemented as a web application and named "Researcher Affinity Browser". Hereby an important user feedback on the usability has been taken into account. The test users from the target group, of course scientific researchers, tested the usefulness (as user satisfaction and search quality) of the framework using an explicit evaluation questionnaire. Since only a few people know about the existence of this application, the evaluation relied on explicit user feedback about the usefulness of the framework and the relevance of the personalized results.

Approach

Researchers with a Twitter account who registered in Grabeeter were invited to try out and take part in the evaluation of the "Researcher Affinity Browser". It was explained that the evaluation was about one of the first web applications to expose affinities between Twitter users and the likely the first to be built on top of a "Semantic Profiling Framework". It is intended for researchers who are using Twitter to report about their research, their interests or the conferences they are attending or tracking. After registration they had to wait a few hours before their data was analyzed and suggestions could be made for them. As soon as that was done, they could start exploring people. Demo video of the "Researcher Affinity Browser" is available online [2].Figure 5.7 shows a screenshot of the application. The left column shows the different affinity facets that can be explored. The center view displays the results in a grid, an affinity plot and a map. Details about a person and their different affinities are displayed in the bottom zone. This application is used for the evaluation as follows: in a first phase the users tested an interactive wire-frame prototype that led to the current user interface design.

The "Researcher Affinity Browser" Application retrieves a list of relevant users using the Researcher profiling API. The results are a current snapshot, not static data. Every time users produce new content on social networks,

[2] https://www.youtube.com/watch?v=A25DrP3Mv8w, last access: 2017–05–29

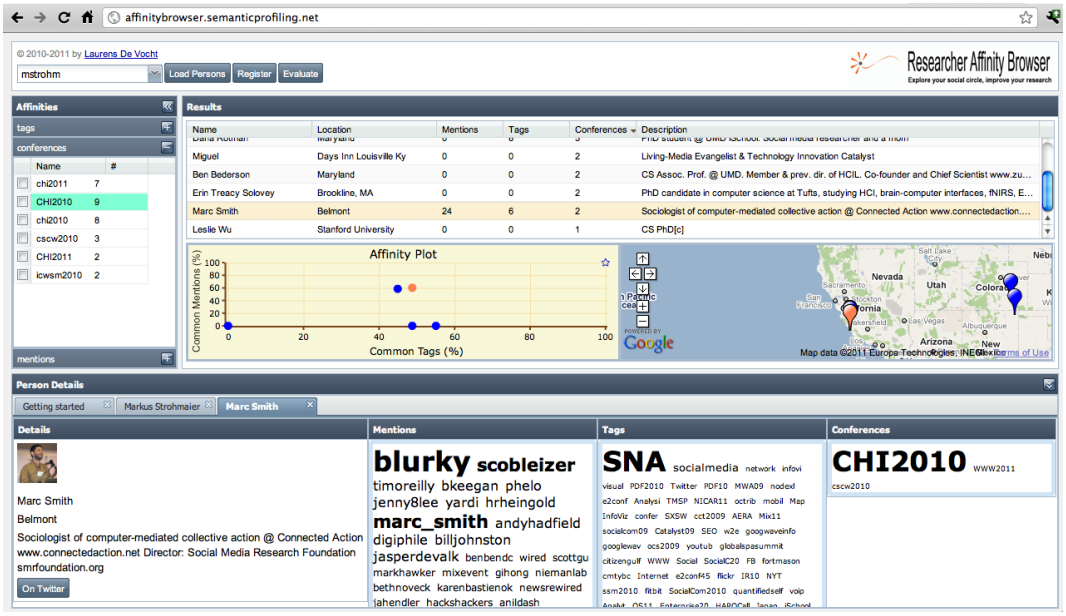


Figure 5.7: Screenshot of the Researcher Affinity Browser, adopted from De Vocht et al. (2011)

the analyzed data evolves with it. In our project this means that when people are using different hashtags over time, our system will suggest other persons to them. The relevancy is measured according to the number of common entities (thus affinities) that are shared with the user. The different affinity facets are displayed on the left. Users can explore in the demo version three types of affinities: conferences, tags and mentions. Activating a certain affinity narrows down and filters the list of matching persons. Users can explore their matches in several ways. Firstly, there is the result table that displays detailed information about each person and how many affinities are shared. Secondly, there is a map view and an affinity plot synchronized with the result table. The purpose of the map is to get a better impression of where the affiliations of the found persons lie. The affinity plot visualizes in a quick overview how 'good' the affinity with the user is. One dimension shows the mentions, the other dimension shows the tags. The more to the top right a person's dot is plotted, the more affinity there is with the user. Thirdly and finally, users can double click on any person in the result list to get a tab that displays a profile with more information

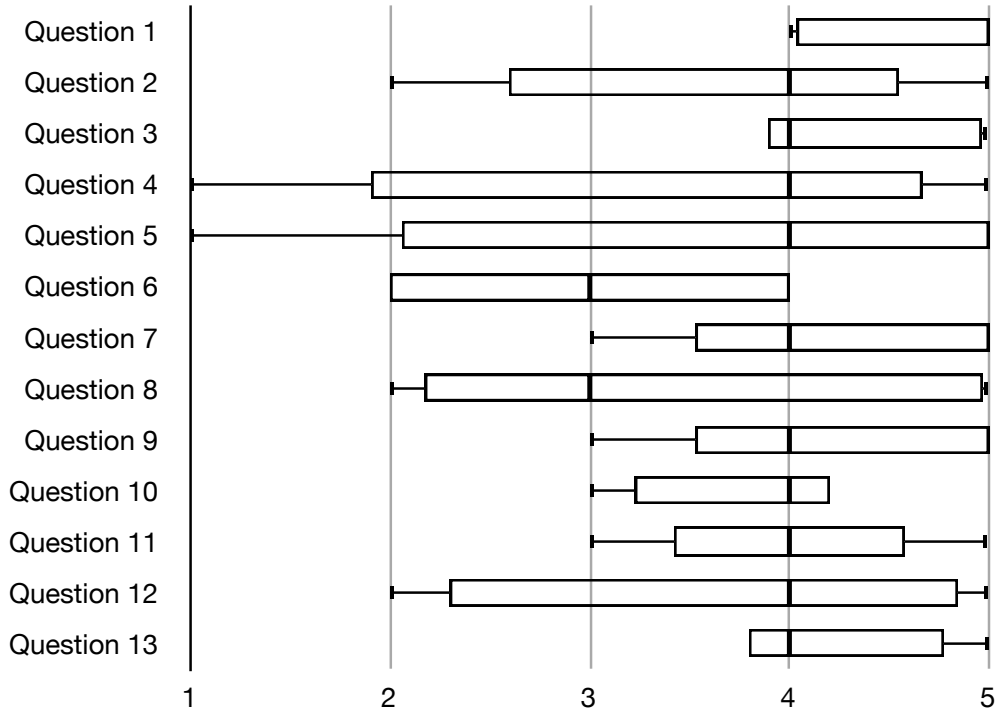


Figure 5.8: Usefulness Questionnaire evaluation results box plot, adopted from De Vocht et al. (2011)

and allows them to get more insight into certain affinities of that person. They can also click on links to get in touch with that person. For example if the profile of someone is extracted from Twitter, a link to the Twitter profile will be displayed.

Results

Figure 5.8 shows the evaluation questionnaire results. The questions are being referred to with their number between brackets.

The most important results are:

- All users agreed and half of the users even strongly that the use of the concept affinity is a great benefit in this context (1).
- Four out of seven users judged that just a few resulting persons is relevant while the other three judged that definitely more than half of the persons presented to them were relevant. Almost all users agreed, but not strongly that there were enough relevant users whom they considered contacting (11).
- All users concurred , and 2/7 strongly that the system makes data from Twitter more useful to find relevant researchers (13). 1/3rd agreed strongly and half of the users agreed that the additional information displayed was relevant to what they wanted to find (9).
- The users are carefully positive – no strong agreements no disagreements – about the reliability of the results: whether the displayed details about a person correspond with what they talk about and do in practice or not (10).
- There is no agreement or disagreement, the answers among the users lie across the entire possible spectrum about the fact if the system shows clearly the affinities between people (2), displays information more confusing and distracting (8), never does anything unexpected (6); if the filtering works fast enough (4, 5) and if the results presented can change daily is obvious enough (12).
- The tendency is to a strong agreement among users that they understand why the persons are plotted and displayed on a map and how the convention between the different views works (3).

In general the application was perceived as a great effort, which already shows that data from social networks can be the source of useful information for researchers. The information, shown in the "Researcher Affinity Browser" application and provided by the web service, exposes affinities and makes great use of Twitter hashtags and mentions. The current identification of scientific conferences and the possibility to easily identify other entities demonstrates the added value of using Linked Data instead of a more traditional relational database approach. Test users confirm this by agreeing on the fact that different types of affinities allow a relevant perspective on persons. The true power of the framework and applications however can only become visible and verifiable after more resources, besides Twitter and COLINDA are linked. Users found that some persons seemed to be relevant

at first sight, but there was not enough detailed information available about those persons to be sure of that.

5.1.7 Conclusion and Discussion on Achieved Results

The "Semantic Profiling Network" has a service that presents for each user a list of suggested entities. Those entities are now limited to persons and conferences. A Researcher Profiling use case was implemented as a web application to test the usefulness of the framework. This application, the "Researcher Affinity Browser", introduced the concept affinity. The use of this concept is not new in the broader field of information science: it has been used before to express relations between abstract objects in software engineering and was reported initially by Pintado (1995) or to rank products based on social data (see Li et al. (2011)). Therefore, it is not entirely surprising that the users all agreed on the benefit of using this concept in a social data context where the information displayed wants to suggest entities and show the proximity between them. How the users evaluate the relevance of their results can differ strongly. Some users find that almost all or at least more than half of them are relevant, this is probably because they acknowledge the fact that when someone has enough entities in common, therefore they are of interest. Other users use a more strict approach and believe that persons they already know are actually not relevant. On top of that they might look beyond the common entities and take the time to actually check those persons out, for example by reading their posts ans tweets. Before they judge someone as relevant they check them actually out. This can not be done of course for all the persons in their results. If at least one person attracted the attention of the user that shows the idea that during their testing they have encountered interesting persons and then they reported those that they quickly checked out to confirm as relevant. Actually since all users reported at least a few relevant persons in the short time of evaluation it proves that the application driven by the framework deliver clear and concise results. All users concurred this, they answered positive on the question whether the system makes data from Twitter more useful to find relevant users. The users also found enough relevant users they wanted to contact. Not all users are convinced by this

last statement, this is probably due to the fact that it is not possible to really find out more about those persons because only Twitter data is used. There are some warnings and areas that could be improved. There is a careful agreement about the reliability of the results: it is not possible to verify if the displayed details correspond with what the persons really do in practice. Some users even explicitly commented on that: more detailed information about each person is missing. Though most users found that the additional displayed information relevant. Furthermore the user satisfaction about how the different views are presented and synchronized varies greatly. This is probably a matter of preference and it depends on what the user expects of an affinity browser. This leads to assumption that some users think of it as an overview of their social circle with new suggestions. Others probably think of it as an expert finding system in which they can dig deeply to discover who shares important affinities with them. To assert those assumption. Two scenarios are needed to make test with implicit feedback from user actions. In this way it will be possible to identify and rate the scenarios that score the best for different types of users. It is not obvious that the results can change daily. This could of course be implicitly assumed since new users are added daily and the tags people are using evolve. However the application and the framework do not really emphasize this and a timestamp of the analysis for example could help or a history overview could be kept and referred to. This paper presented the proof-of-concept for using Linked Data to enhance unstructured semi-structured sparse tweets. More specifically it is an effort to integrate microblogging data from Twitter, combined with scientific conference data from COLINDA. To realize the proof-of-concept it was necessary to set-up a "Semantic Profiling Framework"' as an attempt to make this integrated data available with an interface for programmers and demonstration application. The extension to other linked data resources or other social networks touched briefly to show that the true value of this approach lies indeed in the quick expandability of the used data sets. The approach presented in this paper aims at gaining more knowledge and getting usable data out of social context of micro blogs with a framework driven methodology based upon Semantic Web standards and tools. Introducing the interesting aspects about microblogs, we tried to answer how far they correspond with ideas from other research areas like Science 2.0, Research 2.0, Semantic Web or Linked Data and to outline the importance and relevance of such or similar efforts by examples and

arguments from current research and with examples from current work. Presented work is based on state of the art technologies and brings in a novel approach of usage and dissemination of knowledge accumulated in social data silos like Twitter using the semantic tools and techniques as Semantic Data Modeling and Mining for the domains of appliance like Research 2.0 and Science 2.0. To the best of author's knowledge there were no current known research effort at the time the experiment was conducted that handle the problematic of connecting scientist using tweets, Semantic Web and Linked Data. The more resources, the more types of entities can be interlinked to improve the verifiability of the results. The framework can easily be enriched with additional RDF resources, a new handle in the "Interlinking Module" suffices. Some more effort has to be done to add data from another source that is not yet available as RDF. In that case it is necessary to write an additional Model class for the Extraction module and a handle in the Annotator class that includes data from that module by annotating it appropriately. This process is completely comparable to the extraction of Twitter data presented in this thesis. On the high level, new functionality can easily be added by proper translation into SPARQL queries. As more different data models and resources become available it might be of interest to extend the API as such. Again the same approach can be used for the discovery and presentation of persons and scientific events. Although it is not part of the scope of this project but an entity recognizer such as Open Calais on the entire microblog text could increase the available social data, instead of only focusing on hashtags and mentions. This would however decrease the explicitness, introduce a dependency on the quality of the used entity recognition engine and break the fact that users explicitly mark certain words as important because they want to take part in a conversation or be listed somewhere as the literature has shown. The choice to use only hashtags and mentions is one of the most important limitations. It also interesting to just mention briefly that not all tags are equally important. Algorithms or social methods to rank the importance of tags could improve the quality of the results. They can be seen for example as weighted links in the semantic network. Currently all the links are considered equally important and only the frequency determine the proximity of entities. Together with publishing COLINDA to the LOD cloud here presented efforts want to provide a platform to mesh-up location, user profiles and conference data in the way that is accessible for

humans and machines and to tackle the questions like: which scientists in Twitter considering their profiles fit to me? Which conferences they visited recently and they probably want to visit in the future? Predictions and generating forecast reports about scientist that will participate at some specific conferences and about upcoming conferences that match the own research focus is also an issue that might be the matter of future steps. Further, linking the scientists automatically to sub communities based on their interests can be a thinkable extension of proposed context.

5.2 Alignment of Researcher Profiles with Research Relevant Web Resources and Collaboration Tools

5.2.1 Statement to Own Contribution

Following subsections originate from De Vocht et al. (2014b) and represent the results of research conducted in cooperation with IDLab (former Data Science/Multimedia Lab) at Ghent University. Together with Mr. Laurens De Vocht I elaborated and evaluated the concept of profiling Twitter users (see also Softic et al. (2010, 2013a); De Vocht et al. (2011, 2012)). I also delivered the COLINDA[3] (COnfernce LInked Data) as my own contribution (see also section 4.2 or Softic et al. (2015b)) to this paper and as a base for alignment of researcher Twitter profiles to the conference data. Practical implementation more precisely the algorithm of alignment between profiles and COLINDA and was mainly implemented by Mr. De Vocht. The main objective was to test our commonly developed semantic profiling model for researchers where Twitter accounts of researches are aligned with information about the conferences they visited and works they published, using the information from open digital archives which offer also linked data resources such as DLBP (Digital Bibliography & Library Project)[4] and

[3] http://www.colinda.org/, last access: 2017–05–29
[4] http://dblp.uni-trier.de/, last access: 2017–05–29

COLINDA as main source for conference matching. Additionally the corresponding researchers profile from Mendeley[5], were used as enhancement of the testing profiles. Synthesized profiles have been compared against a golden sample and published the results at the BigScholar 2014 workshop at the WWW 2014 conference. Work described in this section uses and develops further the concepts of researcher profiles and researcher profiling introduced in De Vocht et al. (2012, 2011) and also delivers the back-end administration and Social Linked Data enhancement of the knowledge base that is used in Softic et al. (2013a, 2015b). As consequence of our contribution to the scientific community around the workshop where the paper related to this section was published Mr. De Vocht and myself, we were invited to be the members of the program committee and reviewer at the subsequent workshops BigScholar[6] 2015 and BigScholar 2016[7].

5.2.2 Introduction to the Research Topic

Resources for research are not always easy to explore, and rarely come with strong support for identifying, linking and selecting those that can be of interest to researchers. In this section research on a model that uses state-of-the-art semantic technologies to interlink structured research data and data from Web collaboration tools, Social Media and Linked Open Data will be described as reflection of common research introduced in De Vocht et al. (2014b). This model was used to build a platform that connects researchers, using their profiles as a starting point to explore novel and relevant content for their research. Researchers can easily adapt to evolving trends by synchronizing new Social Media accounts or collaboration tools and integrate then with new data sets. The approach is evaluated by a scenario of personalized exploration of research repositories where real world scholar profiles have been analyzed and compared to a reference profile. All findings from this experiment should serve as starting point for personalization of exploratory search solution for researchers that will be presented in later sections: 5.3, 5.4 and 5.5.

[5] http://www.mendeley.com/, last access: 2017–05–29
[6] http://thealphalab.org/bigscholar/2015/, last access: 2017–05–29
[7] http://thealphalab.org/bigscholar/2016/, last access: 2017–05–29

5.2.3 Motivation for Semantic Modeling of Research Data

Publication repositories and online journals all have search engines to help scholars find interesting resources. However, these approaches are often ineffective, mostly because scholars:
(i) only look-up resources based, at best, on their topics or keywords, not taking into account the specific context and the scholar's profile; (ii) are restricted to resources from a single origin. Of course, aggregations exist that index resources from multiple sources.
The challenge is therefore in matching research needs and contexts to opportunities from multiple, heterogeneous sources. In other words, we should make the most of the wealth of resources for research through relating and matching their scholar profile with the online available resources, publications and other scholar's profiles. Usually researchers need a paid membership to get full access to journals' articles, the library 'paywall'. At the same time a growing number of "Open Journals" offer free online access to all their published works. Most prominent archives in this area are Directory of Open Access Journals[8] as well as Online Journals[9]. Many of these bibliographic archives provide APIs or are already published as Linked Data. Big national libraries followed this example. According to the Linked Open Data (LOD) Cloud stats[10] publication repositories are abundant[11]. Currently the Linked Open Data (LOD) Cloud[12] has reached a respectablesize[13,14]. Around 10% of the overall distribution of triples comes from the research publication repositories. Publications are the source of around 30% of the overall links[15]. Researchers have embraced internet technologies in ways that broaden the scope of their research work beyond college walls and in ways reaching beyond data silos forced by libraries. Microblogging platforms such as Twitter can be a useful way to expand their community even further by following others and sharing research

[8] http://www.doaj.org/, last access: 2017–05–29
[9] http://online-journals.org/, last access: 2017–05–29
[10] http://stats.lod2.eu/, last access: 2017–05–29
[11] http://lod-cloud.net/state/, last access: 2017–05–29
[12] http://lod-cloud.net/, last access: 2017–05–29
[13] http://lod-cloud.net/state/, last access: 2017–05–29
[14] http://stats.lod2.eu/, last access: 2017–05–29
[15] http://lod-cloud.net/state/, last access: 2017–05–29

interests. Following subsection describe the alignment model for researcher developed by myself and my colleague Laurens De Vocht by explaining (i) which vocabularies have been used; (ii) the data sets selected for the implementation; (iii) the custom developed system for dynamic alignment of resources of social media, collaboration tools and selected datasets; and (iv) evaluation of the alignment and measure as well how well it can interlink conferences, publications and authors with researcher user profiles.

5.2.4 Datasets

The datasets used in experimental implementation, combine existing Linked Open Data sets: DBpedia[16], DBLP[17] and GeoNames[18] interlinked with research oriented datasets such as COLINDA[19] and a Social Linked Data set containing information about conferences and social profiles of the researchers from Twitter and Mendeley and the data they generated.
Weaving such sources into the Web of Data is also interesting from the scholars perspective. Twitter, as exemplary Social Media microblogging platform, can help resolving scientific citations as Weller et al. (2011) reported in their work.
The approach uses Twitter[20] data to profile scientists. Besides Twitter as a profiling source the implementation used Mendeley[21], a popular example of a research publication and citation sharing tool, for linking with scientific resources. This source was used to access the publications, tags and profile information of registered authors and link them with the author's social profiles. Table 5.2 highlights the statistics of the used datasets (M = millions, G = gigabytes).

[16] http://dbpedia.org, last access: 2017–05–29
[17] http://dblp.l3s.de, last access: 2017–05–29
[18] http://www.geonames.org/ontology/, last access: 2017–05–29
[19] http://colinda.org, last access: 2017–05–29
[20] http://www.twitter.com/, last access: 2017–05–29
[21] http://www.mendeley.com/, last access: 2017–05–29

Dataset	Size (G)	#Triples/Rows	#Instances	#Literals
DBpedia	38	332M	27.1M	161M
DBLP (L3S)	12	95.2M	13.1M	17.5M
COLINDA	0.15	0.143M	0.016M	0.070M
Social LD*	0.06	0.041M	0.007M	0.015M

Table 5.2: Linked Data used within the search experiments.* Average per user profile. Adoped from De Vocht et al. (2014b).

5.2.5 Vocabularies

Already approved vocabularies (provided by scientific community and current research efforts and reported through Passant et al. (2010b); Tao et al. (2011); Softic et al. (2010); De Vocht et al. (2011)) have been used to annotate social media content as Linked Data. For semantic modeling was specifically applied: Friend of A Friend (FOAF)[22], Semantically Interlinked Online Communities (SIOC)[23], Semantic Web for Research Communities Ontology (SWRC)[24], and the Dublin Core[25]. FOAF describes the user profiles, their social relations and resources. A combination of SIOC with FOAF and the Dublin Core was used for creating model instances of web entries like blogs, microblogs, mailing list entries and forum posts as well as other entries from collaboration tools as previously introduces by Passant et al. (2010b); Breslin et al. (2006b). The SWRC ontology implemented by Sure et al. (2005) was used to describe the academic resources and events with corresponding meta data in order to be compliant with research related Linked Data sets (COLINDA and DBLP). The Modular Unified Tagging Ontology (MUTO)[26] introduced by Lohmann et al. (2011) was used for tag binding as it combines the best approaches from earlier efforts on defining a tag ontology. MUTO instances bind hashtags from Twitter with entities in a user's context.

[22] http://xmlns.com/foaf/spec/, last access: 2017–05–29
[23] http://rdfs.org/sioc/spec/, last access: 2017–05–29
[24] http://ontoware.org/swrc/, last access: 2017–05–29
[25] http://dublincore.org/documents/dcmi-terms/, last access: 2017–05–29
[26] http://muto.socialtagging.org/core, last access: 2017–05–29

5.2.6 Alignment of Researcher Profiles

For the purpose of alignment of researcher profiles three components have been implemented: a *Profiler* which extracts the timeline and followers of the researcher's account and annotates them using the FOAF and SIOC vocabularies, an *Interlinker* which aligns various sources from DBLP (pubications), Geonames and DBpedia (venues) and COLINDA (scientific events) and an *Extractor* which generates the semantically modeled data. They represent a pipeline which produces semantically modeled research profiles aligned to relevant scientific events, publications and other researchers. A sample result of such process is listed in 5.2.

5.2.7 Evaluation

The main intention behind the experiment is testing the aspects of proposed semantic data model and its implementation for making research data available through the interlinking of multiple data sources. The aligning of multiple data sources should improve the quality of the presented content. Test users noted this as an important criterion for improvement during earlier iterations e.g. see subsection 5.1.7 or conclusions in De Vocht et al. (2011). Achieving this allows researchers a more refined and personalized access to heterogeneous sources for data they may find useful. To measure the quality of the linking three parameters have been observed and evaluated: precision, sensitivity and accuracy of the linking applied to four types of resources: authors, friends, publications and hashtags. Each of these measures is a combination of *true positives* (TP), *false positives* (FP), *true negatives* (TN) and *false negatives* (FN) (see Powers (2011)).

$$precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{5.2}$$

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.3}$$

Following text will first describe the scenario to which the interlinked resources contributed, secondly it will describe the user profiles part of this scenario and finally present and discuss the measured results.

5.2.8 Scenario: Personalization

Users start by logging in with their Twitter accounts. Preferably researchers would authorize a social user account in which they often interact with the scientific community. After authorizing their accounts, users get access to a panel where they: manage configured repositories (see figure 5.9); browse a list to connect and disconnect account of available social media or collaboration tools and synchronize data from these accounts with the configured research data repositories.

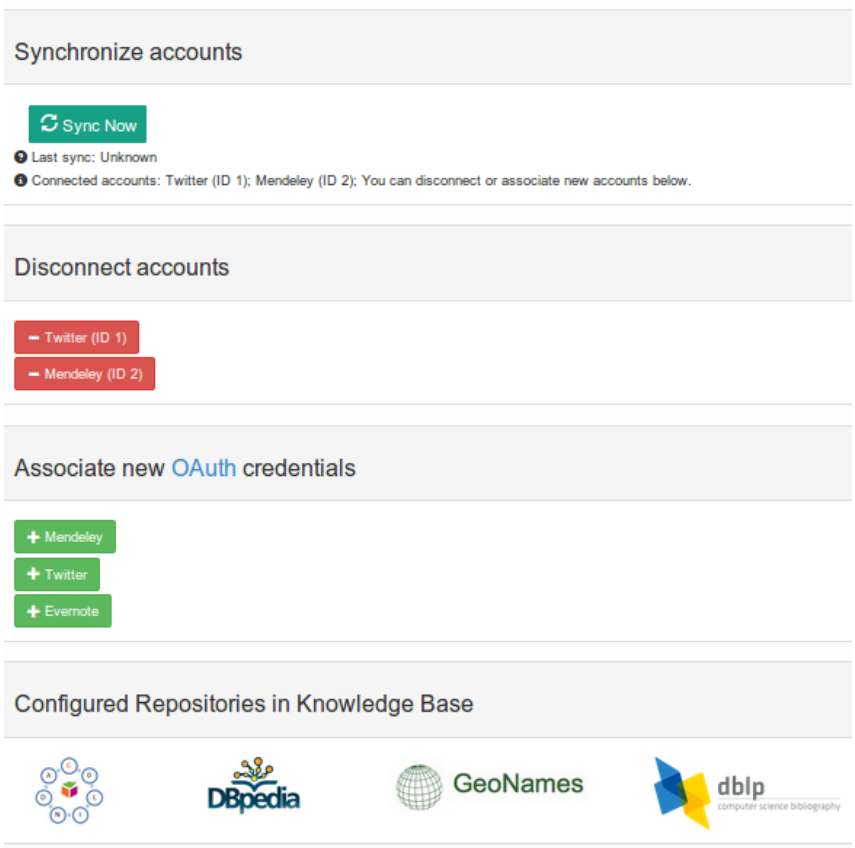
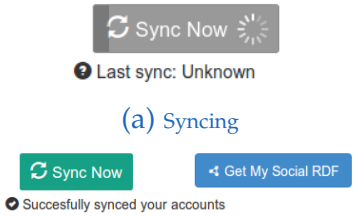


Figure 5.9: How researchers can manage and configure desired resources by creation of profile. Adopted from initial version of De Vocht et al. (2014b).

They can then synchronize the latest social data (from Twitter) with the newest version of their public personal library (on Mendeley) and link it to

the configured research data repositories. This synchronization happens on the client side. After synchronization users can download their profiles in RDF which is automatically posted back to the server (see figure 5.10).



(a) Syncing

(b) Succesfully synced

Figure 5.10: Client-side synchronization of resources: after synchronization users can download their profiles' RDF. Adopted from initial version of De Vocht et al. (2014b).

The goal of the scenario is to expose affinities, otherwise hidden proximities or likings for specific resources, of the synchronized user profiles. It shows context-relevant relations for scholars based on common affinities using hashtags, mentions, people or conferences. The nature of implemented model enables the creation of personalized context as a starting point for further exploration of content made available through the interlinking process described here. The synchronization as pre-setup should enable researchers to explore content closely related to their interests more effectively if there is a sufficient number of accurate and precise links available.

User Profiles

Each researcher (scholar) profile contains a Mendeley library and a Twitter feed. The libraries contain their bookmarked citations and publications, and the Twitter feed contains recent tweets of the researcher and the users followed by researcher. Three different types of researcher profiles fitting the scenario have been compared:

1. An 'intense scholar profile' which uses all the tools efficiently and with a dense community of scholarly related people. This profile has been constructed as 'Golden Profile' (*GP*). It is the only profile which we customly created for use a reference. The others are live profiles

belonging to real users. It has a Mendeley containing publications only from the *Proceedings of the Linked Data on the Web Workshop* (2008-2012). The Twitter profile was created by adding the organizing committee of this workshop series and adding all Twitter recommended profiles to follow mentioning 'Linked Data' or 'Semantic Web' in their description.

2. Two 'typical scholar profiles' using these tools, but the Twitter account is not exclusively used for sharing academic resources for tweeting about conferences. One has a fairly large personal library (*UP1*) while the other has a small personal library (*UP2*). Both libraries contain a variety of publications, not all of these publications are indexed in DBLP.
3. A 'basic profile', making only use of Twitter, and this use is not limited to academic purposes either (*UP3*).

Characteristics for each of the profiles are listed in table 5.3.

Characteristics	GP	UP1	UP2	UP3
Mendeley				
Articles in Personal Library	65	100	33	N/A
Twitter				
Following	30	245	258	N/A
Authors Following	21	35	140	N/A
Hashtags	21	26	18	22
Conference Hashtags	9	5	3	1

Table 5.3: An overview of the contents of each profile. Adopted from initial version of De Vocht et al. (2014b).

Listing 5.2 shows an interlinked article, person and tag. We see that the article's authors are recognized in DBLP as well as the identifier of the article. An *owl:sameAs* connects the person representation with a link to the reference of the social account with the author profile. The example tag shown displays the *muto:tagMeans* property to link the conference hashtag with the URI of the conference on COLINDA.

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix muto: <http://purl.org/muto/core#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix swrc: <http://swrc.ontoware.org/ontology#> .

<http://resxplorer.org/articles/6018551401> a swrc:Article ;
    rdfs:label "4th Linked Data on the Web Workshop ( LDOW2011 )" ;
    dc:creator <http://dblp.l3s.de/d2r/resource/authors/Christian_Bizer>,
        "Christian Bizer" ;
    dc:identifier "10.1145/1963192.1963323", "6018551401" ;
    dc:source <http://www.mendeley.com/c/6018551401/p/27542461/bizer-2011-4th-linked-data-on-the-web-
        workshop--ldow2011-/> ;
    owl:sameAs <http://dblp.l3s.de/d2r/resource/publications/conf/www/BizerHBH11> .

<http://resxplorer.org/people/timberners_lee> a foaf:Person ;
    rdfs:label "Tim Berners-Lee" ;
    dc:identifier "timberners_lee" ;
    owl:sameAs <http://dblp.l3s.de/d2r/resource/authors/Tim_Berners-Lee> ;
    foaf:account <http://resxplorer.org/accounts/timberners_lee> ;
    foaf:name "Tim Berners-Lee" .

<http://resxplorer.org/tags/www2010> a <http://rdfs.org/sioc/types#Tag> ;
    rdfs:label "www2010" ;
    dc:description "The World Wide Web Conference 2010, Raleigh, NC" ;
    muto:tagLabel "www2010" ;
    muto:tagMeans <http://colinda.org/resource/conference/WWW/2010> .
```

Listing 5.2: Excerpt from interlinked data of the GP as published in De Vocht et al. (2014b).

Tags

For tag-entity linking the accuracy was measured by fraction conference tags and the sensitivity by the precision. In all cases it is clear that the GP delivers the best output (higher score is better). We also see in Figure 5.11 that UP1 has a slightly higher accuracy. UP1 also has the largest Mendeley library and used the most conference tags.

While all three UP's have a much lower sensitivity than the GP, they have a considerably high precision, as shown in figure 5.12. The sensitivity for UP1 is better than UP2 for the same level of precision, this is due to the fact that UP1 has a slightly higher fraction of conference tags. Conference tags are better recognized than other tags, not surprising as the model is optimized for it.

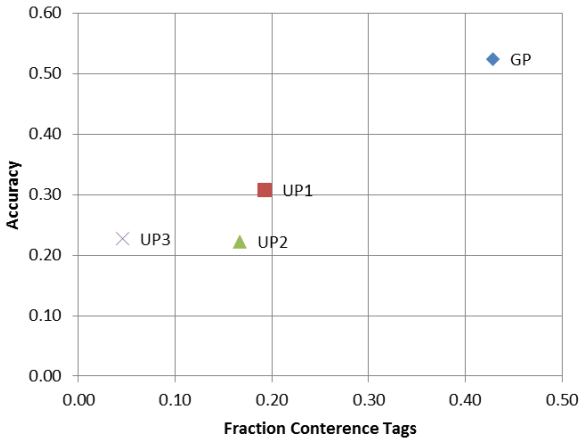


Figure 5.11: Accuracy by fraction of tags which represent conferences: shows that higher fraction of conferences leads to better accuracy as published in De Vocht et al. (2014b).

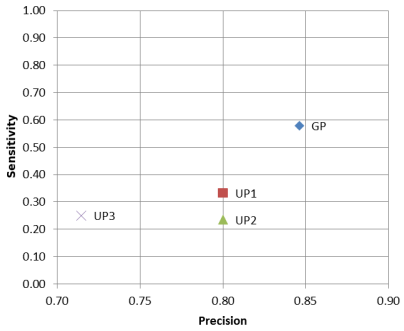


Figure 5.12: Sensitivity by precision: the precision and sensitivity of entity matching of the tags for the GP is as expected the highest. As published in De Vocht et al. (2014b).

Articles and Authors

When interlinking articles and authors, the version of the article and author in the personal library of the user with the version available in DBLP has been considered. Obviously, except for the GP, not all publications are available in DBLP, so there are no TN in that case. In all these cases there are no FP, so precision is equal to 1. This is good and expected, as the links for articles and authors are based on the schema matching of the vocabularies rather than recurrences off the strings as is the case with the tags.
Figure 5.13 shows a relative high precision for authors in the UP1 and UP2

case compared to GP, spread is just above 20%. The spread with the article links is twice as high, GP's library consisted of publications all in DBLP and was centered around the same community. UP1 and UP2 also have articles in their library not available in DBLP. Table 5.4 indicates that linking of

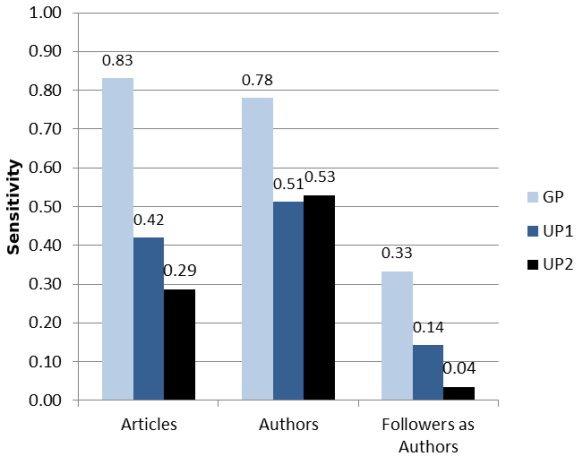


Figure 5.13: Sensitivity by type of resource linked. The sensitivity of linking entities for the GP is as expected the highest in all cases. As published in De Vocht et al. (2014b).

followed users as authors has a bad sensitivity. This is because the personal library of the user which used to identify the link of the social profile of each other with their publications is limited by the scope of each user's library. So it only contains a fraction of the available authors in DBLP. This is however normalized in the accuracy score, which takes into account the TN as well.

User	Sensitivity	Accuracy
GP	0.33	0.53
UP1	0.14	0.88
UP2	0.04	0.48

Table 5.4: Sensitivity and accuracy for linking followed users as authors: a high difference, especially for UP1 and UP2, because many of the followed users are not researchers or are unrelated. As published in De Vocht et al. (2014b).

5.2.9 Retrospective to Existing Work

Studies on the use of microblog platforms like Twitter within scientific communities [27] Ebner et al. (2011) have shown that researchers (scholars) use Twitter to discuss and asynchronously communicate on topics during conferences and in their everyday work (see Reinhardt et al. (2009a)). A survey of the use of Twitter for scientific purposes conducted by Letierce et al. (2010a) showed that Twitter is not only a communication medium, but also a reliable source of data for scientific analysis and profiling tasks (for examples refer Softic et al. (2010); Tao et al. (2011)). In Laniado and Mika (2010) reported that Twitter users adopted hashtags to create threads of communication around a certain topic. Hashtags can be suitable to link entities from microblog posts when combined with Linked Data as findings in Laniado and Mika (2010); Thonhauser et al. (2012) show. In our earlier work on this subject presented in previous sections and chapters, an interface has been built and presented in De Vocht et al. (2011) to allow scholars to browse their affinities such as interpersonal shared commonalities. The efforts to make sharing scientific resources a reality occupied researchers in science and educational informational systems for a long time. The outcome of such quests lead to an increasing variety of heterogeneous technologies, schema, repositories and query mechanisms. This trend brings with it a constant growing amount of publicly available Linked Data including scientific repositories. Within the research community commercial digital libraries like Association for Computer Machinery) Digital Library[28] started to publish their archives in the LOD Cloud providing, in this special case, more than 12 million triples. Parallel to the commercial scientific content providers some academic institutions as well as the most famous public libraries, such as Library of Congress[29], British National Library[30] and Bibliothèque Nationale de France[31] provided their public Linked Data. Besides the initiative of big digital and national libraries, the efforts made by the scientific community like bootstrapping the eScience assets from the

[27] http://www.twitter.com, last access: 2017–05–29
[28] http://acm.rkbexplorer.com/, last access: 2017–05–29
[29] http://id.loc.gov, last access: 2017–05–29
[30] http://bnb.data.bl.uk, last access: 2017–05–29
[31] http://data.bnf.fr, last access: 2017–05–29

Open Archives Initiative - Object Reuse and Exchange (OAI-ORE) project[32] into the Web of Data are worth mentioning. Currently only a limited number of works describe semantic modeling of data from social platforms. In Rowe (2009) authors applied semantic modeling to different social platforms in common contexts and evaluated the potentials of reasoning on such an infrastructure. According to the authors even a small amount of data yields good results with simple reasoning and delivers very precise matches. Passant et al. (2010b) improved mapping social profiles with related content, such as via interlinking the content tags. Semantic modeling for Twitter data has been applied by Softic et al. (2010) identifying hashtags as good resolvers for the retrieval of information and a solid interlinking base for the Linked Data Cloud. Similar use of semantic modeling of Twitter users was introduced on service level by Tao et al. (2011) and confirmed the benefits of previous approaches. These findings have been extended by the work on the "Researcher Affinity Browser" we introduced in previous section of the thesis (see De Vocht et al. (2011)). It is a prototype of Research 2.0 mesh-up based upon a personal semantic model from Twitter connected with the Linked Data set COLINDA, allowing researchers to find and identify colleagues with the same or similar affinities and to track scientific events they visited.

5.2.10 Conclusions and Future Work

Based upon previous research introduced in previous chapters this section presented a new approach for dynamic alignment of research data from social media and collaboration tools with Linked Open Data for scholars and researchers. The implemented experimental prototype was able to match resources from researchers based on their personal library and contributions on social media. This achievement is essential for the effective realization of a tool to facilitate the personalized exploration of heterogeneous data sources containing both research data and social data. Both providers of research data, through opening up their data to a broader audience, and scholars, through actively using collaboration tools and social media, will benefit.

[32] https://www.openarchives.org/ore/, last access: 2017–05–29

Preliminary results indicate sensitivity, precision and accuracy when linking tags, authors and articles to conferences. Conference tags are better recognized than other tags, this is not surprising because the experimental prototype had a optimized model for this task. The prototype never obtained false positives when interlinking authors and articles. By interlinking followed users on Twitter as authors, tested prototype encountered a high amount of negatives. All found links of users as authors were correct but there is room for reducing false negatives.

Future research will focus on how to determine the efficiency of the model and a user evaluation of the environment involved. The environment needs enough incentives for the users to remain synchronized. Further goal is to improve the accuracy of the interlinking by processing the contributed links that weren't immediately recognized. Special focus is set to interlinking the user's personal library with the libraries of other users. This will allow links to be made to social and research data beyond a single user's scope. This should lead to more fine-grained details facilitating researchers to obtain a more sophisticated selection and linking of contributed resources based on previous assessments and explored links.

5.3 Visualizing Relations between Researchers based on Semantically Modeled Researcher Profiles

5.3.1 Statement to Own Contribution

Content of this section originates from the De Vocht et al. (2015). This poster was presented at SAVE-SD Workshop held at the WWW 2015 Conference and extends the former findings published in Softic et al. (2010); De Vocht et al. (2011, 2012, 2013b,c); Softic et al. (2013a); De Vocht et al. (2014b); Softic et al. (2015b). As my contribution I conceptualized together with the main author Mr. De Vocht and other co-authors the use case and experimental prototype of visualization interface and participated in the evaluation of

the presented prototype. My most important contribution was the well-approved approach of semantic modeling as background concept of the implementation for the paper that I presented in my prior work as well in previous chapters and section here along with the interlinking social media data of researchers and relevant Linked Data sources as COLINDA and DBLP.

5.3.2 Motivation and Outline

The various ways of interacting with social media, web collaboration tools, co-authorship and citation networks for scientific and research purposes remain distinct. In this contribution, a solution is proposed to align such information. More particularly, an exploratory visualization of research networks was developed. The result is a scholar centered, multi-perspective view of conferences and people based on their collaborations and online interactions. Also an early stage measurement of the relevance and user acceptance of this type of interactive visualization has been done. Preliminary results indicate a high precision both for recognized people and conferences. The majority in a group of test-users responded positively to a set of statements about the acceptance.

5.3.3 Introduction and Retrospective to Previous Efforts

Social media used by researchers resulted in the emergence of alternative scientific networks beyond the traditional co-authorship and citation networks. However, the various ways of scientific interaction, including these with collaboration tools (e.g. Mendeley[33], ResearchGate[34] and social media (e.g. Twitter [35]) are reflected, but remain distinct from the scholar networks formed in the frame of their publications. Co-authorship, citation and social media based networks are rarely associated, let alone combined in a single

[33] http://www.mendeley.com, last access: 2017–05–29
[34] http://www.researchgate.net, last access: 2017–05–29
[35] http://www.twitter.com, last access: 2017–05–29

visual interface. Social media captures an aspect of conferences that proceedings do not, they reflect the "talk" and networking that goes on during and in-between presentations.

Researchers, as other Twitter users, tend to adopt hashtags to create threads of communication around a certain topic, e.g. *#SemWeb* or *#savesd15*. When used appropriately, searching for these hashtags returns messages that belong to the same conversation (even if they do not contain the same keywords). Results are promising concerning the compliance between Twitter hashtags and URIs, and detecting concepts and entities valuable to be treated as new identifiers (see Softic et al. (2010); Laniado and Mika (2010); De Vocht et al. (2011, 2012); Thonhauser et al. (2012); Softic et al. (2013a); De Vocht et al. (2014b)). Applying semantic modeling for Twitter data led to identifying hashtags as good resolvers for the retrieval of information and a solid interlinking base with the rest of the Linked Data Cloud (see Softic et al. (2010); De Vocht et al. (2011); Softic et al. (2013a); De Vocht et al. (2014b)). For this kind of data, an exploratory visualization scenario to academic metadata is applicable and useful De Vocht et al. (2014a). Exploratory visualization is the process of creating maps and other interfaces while dealing with relatively unknown data Kraak (2008).

5.3.4 Visualizing Social and Bibliography Data

Aligning event data (COLINDA[36]), social media data (Twitter) and publication data (DBLP[37]) forms the foundation for combining recognized conferences tags and Twitter accounts in a single visualization. This is driven by the result that conferences and people can be accurately recognized and interlinked with corresponding authors (see Softic et al. (2015b); De Vocht et al. (2014b)). In presented approach, exploratory analysis methods are used to visualize the network around researchers. Here introduced experimental visualization achieves aligning traditional research networks and networks as they emerge based on data from social media, providing a unique perspective of researchers multi-modal interactions.The screenshot in figure 5.14 depicts the network of a researcher. The scholar is centered with the blue

[36] http://colinda.org, last access: 2017–05–29
[37] http://dblp.uni-trier.de, last access: 2017–05–29

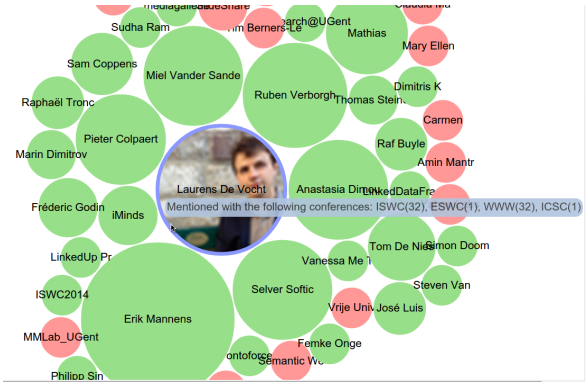


Figure 5.14: The scholar is centered in the middle and the network is visualized in nodes around the central (blue with picture) node. As published in De Vocht et al. (2015).

node and around it are other related people (the more co-mentions, the more nearby they are positioned). The size of the scholar is in the middle between the minimum and maximum size of a node. The more publications someone co-authored with the scholar, the bigger the node. After the researcher has signed in with their Twitter account on ResXplorer they can check recent interactions. A video is available at http://youtu.be/QopnPvWIFzw. A tooltip displays facts about the collaborations (e.g. co-authorships and mentions), i.e. the number of mentions for a specific conference and the the number of co-publications. For the purpose of implementation of the experimental prototype research oriented datasets such as DBLP and COLINDA have been interlinked with data from social media containing information about conferences and social profiles of researchers. This last data was extracted on-the-fly just before generation of the visualization, this way the visualization always shows the latest results. As in approaches so far published in Softic et al. (2010); De Vocht et al. (2011); Softic et al. (2013a); De Vocht et al. (2014b) and described in sections 4.3, 5.1 and 5.2 common vocabularies (such as FOAF[38], SIOC[39], SWRC[40] , and the Dublin Core[41]) have been used to annotate (model) tweets from user profiles. Always the latest 200 tweets have been filtered from the user's timeline and home-timeline, to find those

[38] http://xmlns.com/foaf/spec/, last access: 2017–05–29
[39] http://rdfs.org/sioc/spec/, last access: 2017–05–29
[40] http://ontoware.org/swrc/, last access: 2017–05–29
[41] http://dublincore.org/documents/dcmi-terms/, last access: 2017–05–29

containing matching hashtags corresponding with conference abbreviations (provided by COLINDA). Furthermore, each user that is mentioned in a tweet or is the creator from a tweet is identified and linked as a person. For the interlinking, have been used same techniques as the ones in prior works to model and align researcher profiles with data from web collaboration tools Softic et al. (2010); De Vocht et al. (2011); Softic et al. (2013a); De Vocht et al. (2014b) (see also sections 4.2, 4.3, 5.1 and 5.2). After extracting and converting the tweets, mentions and hashtags have been identified which have been subsequently interlinked with the researcher's bibliographic record (on DBLP). Each researcher has been matched together with one of their co-authors, if mentioned in a tweet, with the researcher's bibliographic record. The result is a graph of people containing links to the conferences and co-occurrences of mentions. This graph is transformed to an ordered list as preparation of the visualization. The way the visualization is configured is that it starts from a node in the center and then adds the other nodes around it, as specified by the input data. The generated ordered list starts with the scholar and continues with the people mentioned most frequently in an online thread or tweet. Included parameters for each visual item (here researcher bubble in the visualization) are: the conferences with number of mentions; number of collaborations (expressed as co-authorships); and whether the scholar has already connected to the person mentioned in the entry or not.

5.3.5 Results

The evaluation of the visualization was focused on two aspects: **relevance**, by observing the *precision* and *recall* of the visualization, essential to validate that the presented information is sufficiently applicable to the user; and **acceptance**, a user survey to verify that the visualization is *usable*, *useful* and might lead to more *effective* scholarly networking. For test purposes we selected a group of 10 researchers from the computer science field who tweet more often and visit conferences in their field of interest. The size of the group is more representative and allows limited observations on applicability and acceptance and some retrospective on already achieved results for the tests conducted for mining/profiling architecture and COLINDA

reported in previous chapters. As results will show previous findings on mining and profiling efforts are throughout re-confirmed and a new aspect as acceptance was sensed through the test group. Test users tended mainly to perceive the visualization as 'useful'.

Relevance

To measure precision and recall a group of *10* researchers in computer science (who visited and contributed to at least one computer science related conference in 2014 and use Twitter) has been asked to complete a set of tasks where they indicated how they judged the visualized nodes. The visualization for all test-users combined resulted in *217* recognized people and *29* recognized conferences. The response of the test users brought high precision for conferences (*0.97*), because almost all detected conferences were correct. There was still a low recall (*0.56*), as many conferences were missing according to the users. Also notable is the moderate recall (*0.83*), a relatively large number of people in the network (co-authors especially) were missing because they were not mentioned together - however users expected them in the visualization. The precision for recognizing people is high (*0.92*): people who where connected indeed belonged to the researchers network or the users considered adding them to their network. Noteworthy is that overall test-users in total, discovered *19* (out of *217*) people they considered adding to their network. Not everybody was presented the way test-users expected: they indicated that *37* people were missing. This implies that the coverage does not extend to people that do no have a Twitter account. However, users could increase the number of people in their visualization by actively tweeting and interacting with the people they consider relevant. This implies that for conferences where Twitter is not common, the results are definitely less interesting for the users.

Acceptance

To test the acceptance a set of statements by applying the Technology Acceptance Model has been created to measure effectiveness, usefulness and usability. The same *10* test-users completed the survey by answering

the questions on a Likert-scale from 1 to 5, indicating the degree to which users agree with the statement or find it likely. Figure 5.15 shows the trend where the min and max are the lowest and highest point of the lined and the majority is in the box for each parameter.

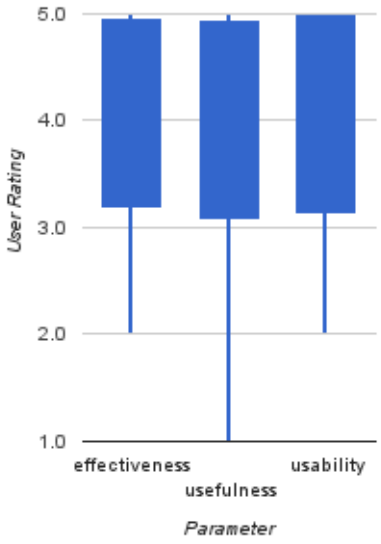


Figure 5.15: The users response on the acceptance was mixed, but the general tendency was they agreed to most of the statements. As in intial version of De Vocht et al. (2015).

The users response is varied, but on average and the majority (scores between 3 and 5) agreed or found most of the statements likely for all of the parameters measured. A small portion of the users were not immediately convinced by the usefulness (scores between 1 and 3), mainly because for them the visualization returned few results. Some indicated this was because they were only passively using Twitter, or Twitter was not used at all during the recent conferences they visited.

5.3.6 Conclusions and Future Work

The information presented in the visualization has high precision, both for people and conferences. However, the recall is moderate for people and low for conferences. This is due to the many missing conferences, typically because Twitter was not used very often or because there was a lot of noise (unrelated tweets) preventing the detection of the relevant context hashtag. The strength of the visualization lies in its fairly effective mapping as perceived by users. The more scholars use Twitter and use it to interact with others in the context of conferences, the more relevant results they will see in the visualization. Consideration to extend the number of tweets taken into consideration, could lead to obtaining larger networks, especially for users who tweet often.

5.3.7 Contribution to Existing Work

Presented visualization finds its application in "scientometrics", the study of measuring and analysing science, technology and innovation (see Van Raan (1997)). The novelty in presented approach in this context lies in combining Twitter data with co-authorship and conference data (see De Vocht et al. (2012)). It specifically relates to the challenge of detecting interesting people in a community of interest where it is useful to have a common research focus and thereby using Twitter as a real-time source. This includes identifying how a researcher's network is structured through collaborations (i.e. co-authorship) and how this is reflected in online interactions and who is joining the conversation that might be relevant, before, during and after conferences. Furthermore the relevance of the content presented to the user has been verified and validated the acceptance of the way it was visualized.

5.4 Finding and Exploring Commonalities Between Researchers Using the ResXplorer

5.4.1 Statement to Own Contribution

Effort described in this section is inspired and resides on previous achievements elaborated in De Vocht et al. (2011, 2012, 2013a,b,c) on semantically and Linked Data driven search and user interfaces for Research 2.0. This section explains how researchers can find and visually explore commonalities about other researchers within their interest domain using the user interface of "ResXplorer"and underlying search infrastructure. Detailed description of "ResXplorer" will be presented in section 5.5. This section introduces only briefly discussion of the most important components of "ResXplorer" relevant for commonalities detection and evaluates the commonality finding features. Section is closing up with discussion of results and conclusions for future work.

5.4.2 Introduction

Research 2.0 introduced by Ullmann et al. (2010) as adaptation of the Web 2.0 for researchers defines researchers as main consumers of information. Led by this idea presented implementation uses Linked Data knowledge base as infrastructure to resolve the connections and commonalities existing between researchers and visualize them in a Web 2.0 interface. The interface uses a search engine which relies on Linked Data knowledge base containing semantically modeled research related and personal information. The scope of this section focuses feature-wise on the use of interactive visualization to enable exploratory knowledge discovery from Linked Data. Data from Linked Data Knowledge Base originates from scholar repositories like DBLP(L3S)[42], COLINDA[43] and other relevant commonly used Linked Open

[42] http://dblp.l3s.de/, last access: 2017–05–29
[43] http://colinda.org, last access: 2017–05–29

Data, in our case from DBPedia[44] and GeoNames[45].

5.4.3 Interface

In initial step a real-time keyword disambiguation guides researchers by expressing their needs. User selects the correct meaning from a type-ahead drop down menu. Query expansion of terms happens in real-time, which has been emphasized as useful feature during the early stages of the search as reported in White and Marchionini (2007). Figure 5.16 shows the type-ahead expansion of "ResXplorer" in action.

In behind the back-end Everything is Connected Engine (EiCE) (work

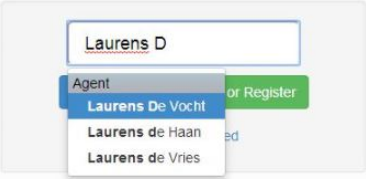


Figure 5.16: Mapping of keywords. As published in Softic et al. (2014b).

related to EiCE originates from De Vocht et al. (2013a)) connects the scholar resources and ranks them according to the entered context. At the same time background modules also fetch neighbor links which match the selected suggestion. As result, selection of various resources is then presented to the researchers. In case they have no idea which object or topic to investigate next, they get an overview of possible objects of interest (like points of interest on a street map) within radial interface.
Features like color, shape (icons) and size of the items are used to enhance the guidance of the user during the search and exploration process (see De Vocht et al. (2013b,c)). Whole process around finding the scholar artifact that serves the commonality detection is depicted in figure 5.17.

[44] http://dbpedia.org, last access: 2017–05–29
[45] http://geonames.org, last access: 2017–05–29

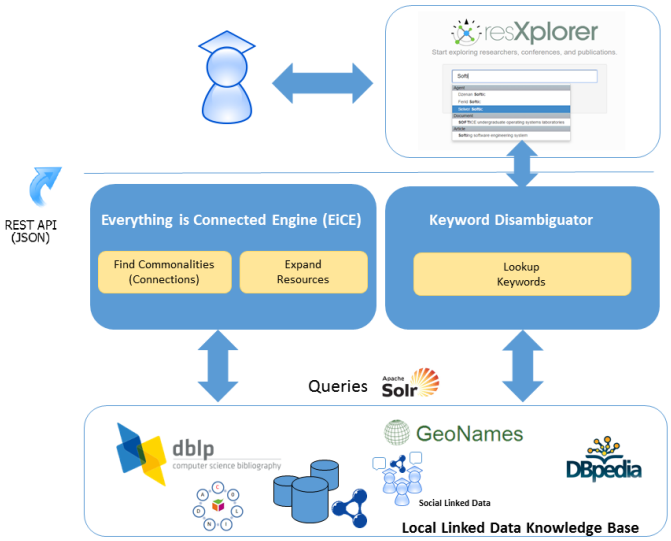


Figure 5.17: ResXplorer concept for finding scholar artifacts necessary to reveal the commonalities. Adapted from Softic et al. (2014b)

5.4.4 Visual Exploration of Commonalities

The visualization emphasizes commonalities by showing, on a specially developed radial map initially introduced by Yee et al. (2001), how the current focused entity is related to the other found entities. It is based on the concept of affinity that can be appropriately expressed in visual terms as a spatial relationship: proximity (see Pintado (1995)). Additionally the amount of unexpectedness is expressed as *novelty* of a resource in each particular search context. A typical example is illustrated in figure 5.18.

Each time a combination of various resources is visualized, the application suggests new queries: they are generally most useful for refining the system's representation of the researcher's need. In case the researchers have no idea which entity to focus on or what topic to investigate next they get an overview of possible entities of interest, like points of interest on a street map. By profiling their activities and contributions on Social Media and other platforms such as their own research publications, the affinity with the proposed resources is enhanced for this perspective. Efforts on this topic has been presented in sections 5.1, 5.2 and 5.3 and to them related

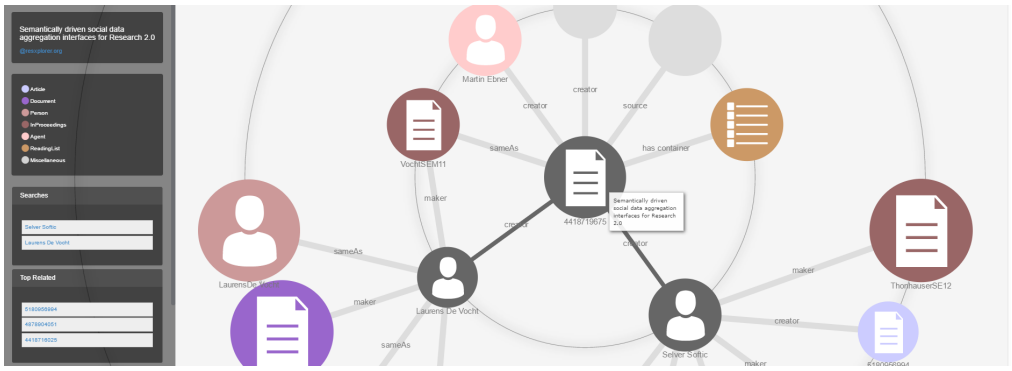


Figure 5.18: Visual depiction of commonality between *Laurens De Vocht* and *Selver Softic* based on common publications such as the highlighted "*Semantically...Research 2.0*". Adapted from Softic et al. (2014b).

publications in Softic et al. (2010); De Vocht et al. (2011, 2012); Softic et al. (2013a); De Vocht et al. (2014b, 2015).
The user expands the query space by clicking the results retrieved by initial keyword based search. Additional query expansion happens either through adding further keywords as well as through keyword combinations already entered where the back-end tries to deliver additional results based upon connection paths between the resources.

5.4.5 Experiment Design, Evaluation and Results

Setup

For evaluation of the module responsible to find commonalities, a set of ten queries shown in table 5.5 has been chosen consisting from the name pairs of authors for which is to be expected that they will deliver results, and that author profiles already exist in the DBLP bibliography archive. This set of queries is selected for reason to easier determinate relevance of results. Measurement of recall is left out intentionally because of the size of search space (hundreds of millions of potentially relevant resources).

Table 5.5: Set of queries, for finding of commonalities between researchers. As published in Softic et al. (2014b).

Query	Keywords
Q1	Selver Softic, Laurens De Vocht
Q2	Selver Softic, Erik Mannens
Q3	Martin Ebner, Selver Softic
Q4	Martin Ebner, Laurens De Vocht
Q5	Erik Mannens, Martin Ebner
Q6	Erik Mannens, Laurens De Vocht
Q7	Laurens De Vocht, Rik Van De Walle
Q8	Rik Van De Walle, Selver Softic
Q9	Rik Van De Walle, Martin Ebner
Q10	Rik Van De Walle, Erik Mannens

Measures

Definition represented in equation bellow expresses precision as combination of *true positives* (TP), *false positives* (FP) results. Links discovered along traversing path of algorithm which lead to scientific resources (publications, persons and events) relevant for one of the both authors represent true positives. All other unresolvable or repeating links are false positives.

$$precision = \frac{TP}{TP + FP} \tag{5.4}$$

Results

Table 5.6 summarizes preliminary results of our tests. The experiment measured precision of retrieved commonalities, path length between the two resources entered as terms of the query, and total count of discovered commonalities per query. The precision values range from **0.7** up to **0.95**. This precision rate is unexpectedly high although it was known that test queries represent authors who know and work with each other. These results are partly influenced by the well-connectedness of graph structures in the Linked Data Knowledge base. Path lengths are very short as expected and range from **2** up to **4** hops. Total count of detected commonalities ranges from **4** up to **11** except in query *Q10*. The explanation for this outlier is that relation in *Q10* is the strongest one because of the length of common period

Table 5.6: Precision, path length, commonalities count along the detection path for test queries. As published in Softic et al. (2014b).

Query	Precision	Path length	Commonalities
Q1	0,75	2	4
Q2	0,86	4	7
Q3	0,78	2	9
Q4	0,75	2	4
Q5	0,82	4	11
Q6	0,83	2	6
Q7	0,83	2	6
Q8	0,7	4	10
Q9	0,7	4	10
Q10	0,95	3	37

of collaboration between those two researchers and the number of together published works. They also have a bigger social network of collaborators which allows finding more alternative connection paths within semantic graphs than in the case of other queries. Evaluation of precision versus the path lengths in figure 5.19 reveals that; there is no linear dependency between the path lengths and precision. At least in our evaluation, results with shorter path lengths reach in average better precision then the ones with long paths.

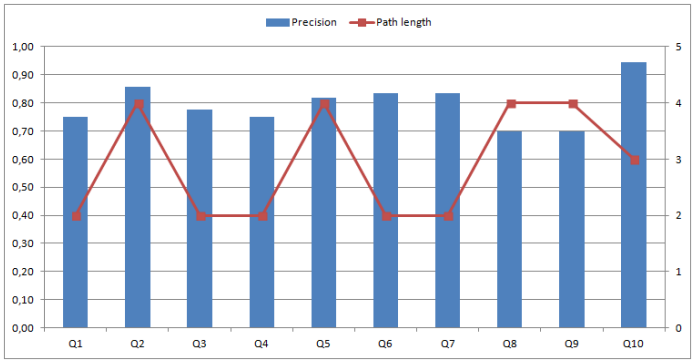


Figure 5.19: Precision vs. Path lengths. As published in Softic et al. (2014b).

Figure 5.21 shows that changes of total number of retrieved commonalities does not have any immediate significant impact on the precision score. This is not surprising since the precision depends directly on the ratio of true

positives and false positives.

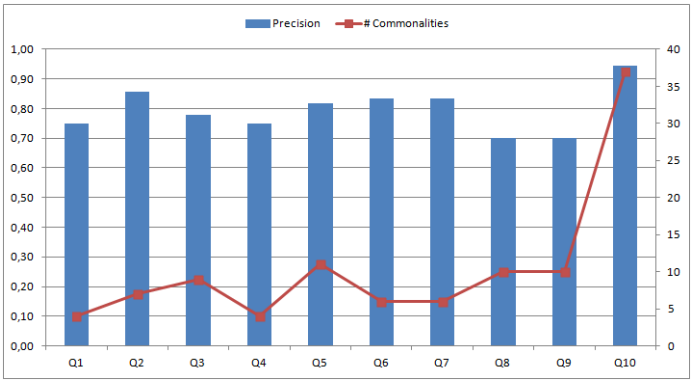


Figure 5.20: Precision vs. total count of Commonalities. As published in Softic et al. (2014b).

The most interesting finding reveals figure 5.21 where path lengths face the total counts of detected commonalities. The results depicted here discount the assumption that the length of a path traversed by algorithm within a graph structure which is well-connected implies inductively the increase of detected commonalities by each new hop. Even the outlier in the *Q10* proves this assumption wrong. This confirms once again the latter findings that solely quality of the detected commonality links determinate the precision and do not correlate strongly with changes of path lengths and total count of discovered commonalities. This finding is potentially influenced by the specific form of data graph structures in the Linked Data Knowledge Base, however this assumption is not confirm able with current results.

Quantitative reasons for the high precision are visible in figure 5.22 where total count of detected commonalities faces the count of true positives and false positives. The count of true positives almost correlates with the total count of commonalities which is a strong indicator for high precision.

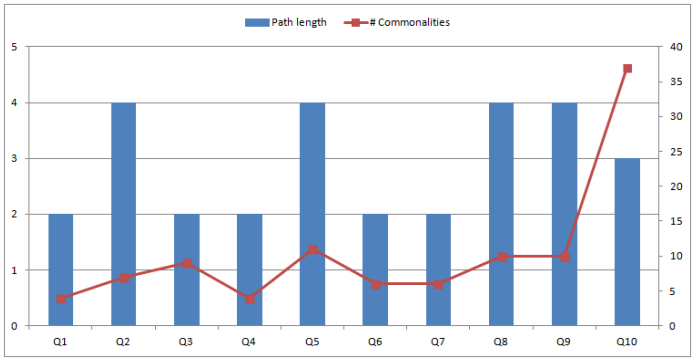


Figure 5.21: Path lengths vs. total count of Commonalities. As published in Softic et al. (2014b).

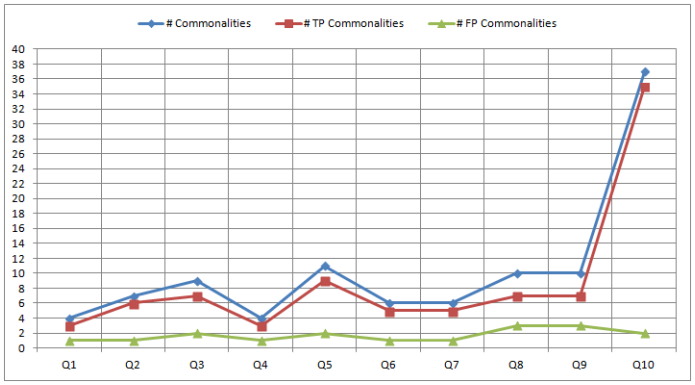


Figure 5.22: Total count of Commonalities vs. TP Commonalities vs. FP Commonalities. As published in Softic et al. (2014b).

5.4.6 Conclusion on Contribution of Presented Work

The main contribution of conducted experiment is, besides retrieving resources from Linked Data knowledge repositories, allowing researches to interactively explore relations between the resources and entities like events, places, publications or persons related to their work and discover commonalities between them. Results on detecting commonalities between two researchers perform very precise, although the experiment is very small sized, initial results are very promising. In the next section the volume of experiments and the exhaustive description of "ResXplorer" will be presented in order to show the full outreach of research work on done around the

"ResXplorer".

5.5 ResXplorer - Enhancing Exploratory Search for Researchers using Semantic Modeling of Data

The main concept of ResXplorer[46] resides on the idea of an interactive search interface which leads the researcher through the process of expansion and exploration of results to the hidden implicit valuable information discoveries which are uncovered in such a process.

5.5.1 Statement to Own Contribution

The user interface and search infrastructure introduced in following subsection is a product of a long term common research together with IDLab (former Data Science/Multimedia Lab) at Ghent University accumulated through through the years from 2011 until today and partly also described in publications Softic et al. (2010); De Vocht et al. (2011); Softic et al. (2013a); De Vocht et al. (2012, 2013b,c, 2014b); Softic et al. (2014b, 2015b). Following subsection present fragments of the texts from a journal publication De Vocht et al. (2016) published in Computer Science and Information Systems Journal[47]. Related content with same focus was also published in De Vocht et al. (2017). Parts of presented work strongly contribute to the areas such as: Semantic Data Modeling an Mining, Profiling Researchers through Social Media, Exploratory Semantic Search and Research 2.0 Interfaces. I was strongly involved in delivered pre-work to ResXplorer on semantic modeling, mining and profiling researchers. Further, I contributed on conceptualization of use case, evaluation and implementation of ResXplorer especially by user interface and in the area of underlying profiling knowledge base. Implementation of the search engine and all the infrastructure was done by my co-author Mr. De Vocht and infrastructure was provided by IDLab at Ghent University.

[46] http://resxplorer.org

[47] http://www.comsis.org/, last access: 2017–05–29

5.5.2 Introduction

Currently, there are several online bibliography archives available on the web including peer-revied publications and related meta-data. This information is useful for researchers to gain information on new interesting publications, scientific contacts and collaboration opportunities and relevant and topic related events. Available bibliography archives and digital archives with publications have an API or allow a certain access to their structured content. Some of those archives are even available as Linked Data e.g. DBLP (Digital Bibliography and Library Project)[48] or Springer [49]. Binding to such interfaces into own search requires high proficiency for technologies. As Ebner and Reinhardt (2009) report in their paper researchers use social media such as Twitter and Facebook to comment and discuss during scientific events and share their research related materials. Many of them have accounts for academic researcher social networks such as Mendeley[50], Research Gate[51], Google Scholar[52], Academia[53] or the like. According to Van Noorden (2014) such platforms already have millions of regular users from scientific community. The big challenge for researchers is to use the existent resources for their own benefit in a efficient way. Most of the resources interesting for researchers are not easy to explore because existing platforms rarely support identifying, linking, and selecting of research related resources for further investigation. Experiment and actions related to implementation and evaluation of ResXplorer are focused on how researchers are finding information they need such as: conferences, people and publications of their interest. The experiment envisions the idea of Research 2.0 (see section 2.3). ResXplorer experimental implementation represents a personalized interactive semantic search environment based on previously implemented search infrastructure and data from diverse open Linked Data repositories including scientific publication archives and Social Media.

[48] http://dblp.l3s.de/d2r/, last access: 2017–05–29
[49] http://lod.springer.com/, last access: 2017–05–29
[50] https://www.mendeley.com/, last access: 2017–05–29
[51] https://www.researchgate.net, last access: 2017–05–29
[52] http://scholar.google.com, last access: 2017–05–29
[53] http://academia.edu/, last access: 2017–05–29

5.5.3 Main Goals of Research Around ResXplorer

The first use cases, data architectures, mesh-up concept studies and prototypes on aligning the social web with semantics in the context of research, were introduced in 2011 by De Vocht et al. (2011). The data modeling concepts were discussed in Softic et al. (2010, 2015b) (see also 4.3) while the back-end used for ResXplorer was investigated, a framework for discovery of chains of links between resources was introduced in Softic et al. (2013a); De Vocht et al. (2013b). The aligning and matching of research related semantic resources was the main scope of the previous work on dynamic alignment of scientific resources such as web collaboration tools and digital archives De Vocht et al. (2014b); Softic et al. (2014b). The first prototypes of the ResXplorer search interface were presented at conferences in late 2013 in De Vocht et al. (2013b) and 2014 Softic et al. (2014b). One of the first live versions participated at the Semantic Web Challenge 2013 (see De Vocht et al. (2013c)). The goal of all these publications was to evolve the concept, demonstrate the interface and visualization, trigger discussion and gain insight on the exploration work-flow. ResXplorer is one of the first prototypical solutions combining the Social Web and the Semantic Web in an interactive search environment that visually emphasizes and represents the search context and results. The aim is to show in an example how interactive visualizations enable Knowledge Discovery in Linked Data, which can be invaluable to researchers. One of the use cases the solution supports, focuses on the end-user usability of semantically enriched researcher profiles. In this use case, the experimental prototype "ResXplorer" shows relations between researchers based upon the semantic analysis of researcher's tweets and aligned with information about conferences and proceedings, users are presented how they are (indirectly) related based on their institutions, visited locations, and conferences they contributed to. As a measure of usability that was investigated was the ability of search interface to support the construction of a good cognitive model of the underlying data and the relations within the data. Finally, the effectiveness and productivity of the interface have been measured by checking to which extent end-users carry out knowledge-intensive and analytical tasks. All insights regarding these investigations are presented in form of summary in 5.5.6. The summary represents a short version of findings presented in De Vocht et al. (2016).

5.5.4 Interacting with Research and Social Media Data

ResXplorer basically uses information from resources that have been already explored by experienced researchers. Such information should primary serve as starting point for researchers who are looking for next practical piece of information that could enhance their domain knowledge. In combination of keyword based Search with link clicks out of results presented on the screen (query expansion for selected item) and background driven search for the "connection paths" between the items (authors, events, locations and publications) researchers are able to select and expand the 'intended' search goal over several iterations. When users are looking for new leads, they get an overview of possible objects of interest (like points of interest on a street map) by having their activities and contributions linked on Social Media and other platforms such as their own research publications profile. Example of an "in search situation" is depicted in figure 5.23.

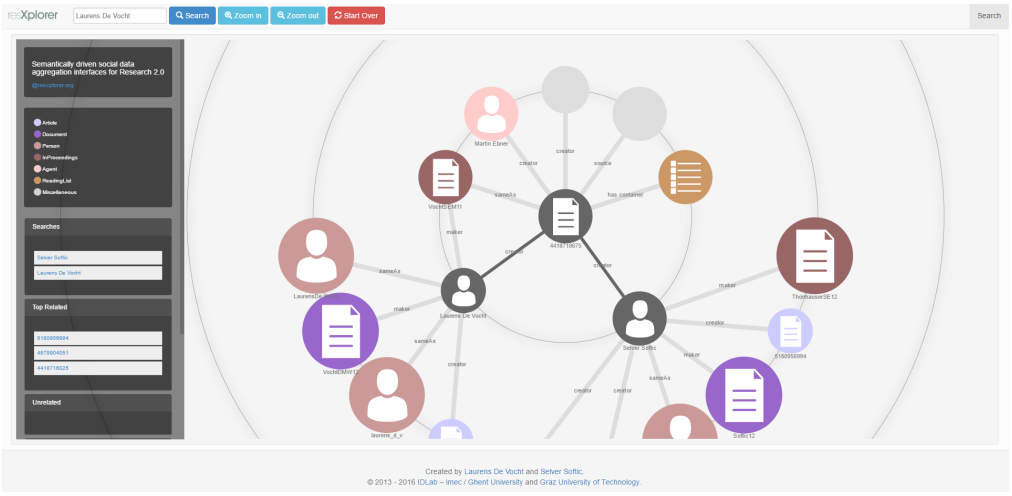


Figure 5.23: Situation when Selver Softic and Laurens de Vocht have been searched by keywords. The found connection is the common paper. The connection is populated when the paper is clicked.

Data Model

The data model for search and matching researcher profiles includes: Research Linked Data and Social Media Linked Data.
Research Linked Data
The used Research Linked Data is described with state-of-the-art vocabularies according to findings in previous works reported by Softic et al. (2010); De Vocht et al. (2011, 2012); Softic et al. (2013a, 2015b) and according to the recommendations of the Semantic Web Community. Research Linked Data includes: the "Digital Bibliography and Library Project" (DBLP)[54] (see Ley (2002)) that provides bibliographic information on major computer science journals and proceedings. It includes approx. 2.3 million articles. DBLP also links to home pages of computer scientists. The Conference Linked Data (COLINDA) data is used to link conferences with user profiles, other conferences, venues and resources (presented in 4.2 and in Softic et al. (2015b)). COLINDA describes conferences with Semantic Web for Research Communities (SWRC)[55] ontology presented in Sure et al. (2005).
Social Media Linked Data
The experimental implementation used own made annotated set of extracted conference hash tags mentioned in tweets of researchers which would be associated with corresponding tweets and which can be used for further mining tasks like label based matching of scientific events in Linked Data sets e.g. COLINDA, DBLP. The reasons for linking data from Social Media are manifold. Most significant out of perspective of this work are: discovering new links between the users, attaching timely and personalized context to the search and finally revealing the relation between researchers and related resources and events.

Both data sets are previously introduced in 5.2.4.

5.5.5 Interactive Search

The overview over technical concept of interactive exploratory search was already briefly introduced in section 5.4.3 (see also Softic et al. (2014b)). Basi-

[54] http://dblp.l3s.de, last access: 2017–05–29
[55] http://ontoware.org/swrc/, last access: 2017–05–29

cally, ResXplorer is an experimental interface for search in a Research Linked Data knowledge base which is using the latest Linked Data technologies with an advanced indexing and path finding system. The back-end resides on earlier work performed by IDLab (University of Ghent) on "Everything is Connected" engine (EiCE) introduced by De Vocht et al. (2013a) and Web 2.0 technologies (such as JQuery and Django). The interface (front-end) is a realization in HTML5 and JavaScript, which makes advanced use of JQuery UI[56] in combination with the 'Javascript Information Visualization Toolkit'[57].
Within search process users basically combine keyword-based disambiguation (through underlying Linked Data Knowledge Base) combined with visual refinements through expansion queries on the semantic entities which the back-end system recognizes as facets displayed in search interface. The facets should offer always and at each step a complete understanding of why certain results are showed. In this way the system hinders the algorithm assume things about researchers preferences. Since it is meant to be an exploratory search, the point is to involve the researchers on the base of input-output principle in a guided approach through facets and three dimensions: shapes, colors and size. They choose what they want to see and get that result delivered. As a parallel process in the back-end, the engine discovers additional relations between the search results and presents them as alternatives to the already acquired information. In this way the facet range available to the researcher is automatically expanded or narrowed down. This leads the researchers through the data by offering them at each point in time exploration and involvement of new and already found items into the search.

Front-end

A real-time keyword disambiguation assists in front-end the researchers in expressing their research needs. Users are allowed to select the correct meaning from a drop down menu that appears below the search box. According to White and Marchionini (2007) presenting candidate query

[56] https://jqueryui.com/, last access: 2017–05–29
[57] http://philogb.github.io/jit/, last access: 2017–05–29

expansion terms in real-time, as users type their queries, can be useful during the early stages of the search. In this is case it is very important that the users understand meaning of the suggested terms. Therefore an as straightforward as possible representation of the keyword mappings is used as shown in Figure 5.24. Researchers can define and select their 'intended'

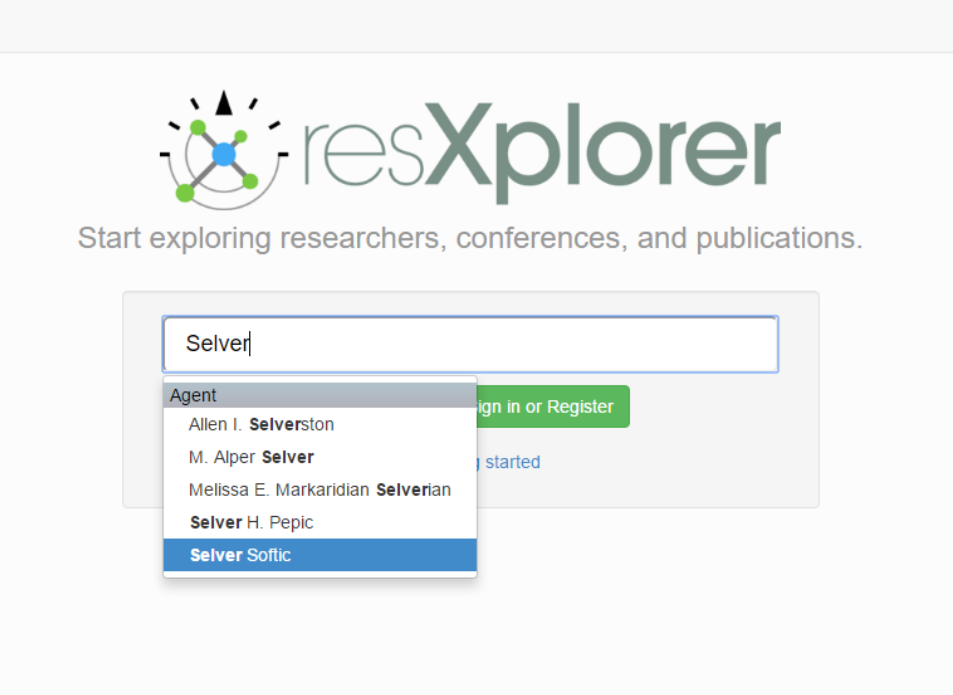


Figure 5.24: Mapping of keywords to Linked Data entities. Adapted from example in De Vocht et al. (2016).

search goal over several iterations. A combination various resources is then presented to the researchers. In case they have no idea which object or topic to investigate next, they get an overview of possible objects of interest (like points of interest on a street map). Researchers define a search query for their research and have it parsed by the back-end system.
Based on the ability of humans to rapidly scan, recognize, recall images and detect changes in size, color, and shape as pre-attentive attributes, the interface aims to enhance the guidance of users during their search by using several visual aids. Figure 5.26 shows how researchers can track the history of their search: the explored relations are marked and clearly

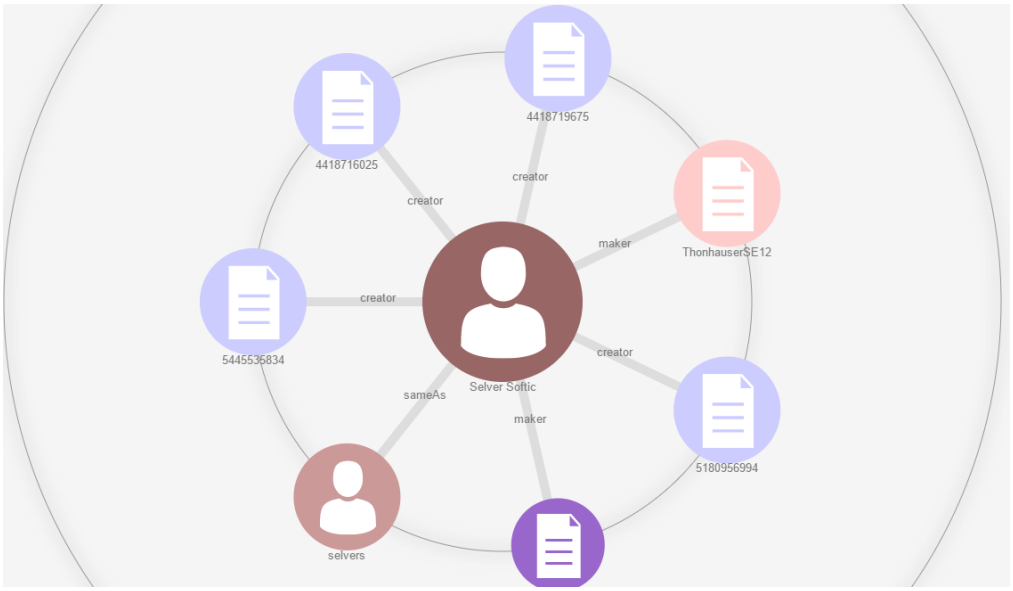


Figure 5.25: Different icons (shapes) and color to distinguish types and different sizes to guide the user's focus. Adapted from De Vocht et al. (2016).

highlight the context of a search. This is a good example of how our system adapts to the users and their environments. It shows one of the ways how to build a model of the goals and knowledge of an individual user (see Brusilovsky (2003)), and the model is used throughout the interaction the user. Researchers can click on a list of resources they have searched to focus the visualization. A screencast of the search interface is available online[58]. The screencast shows how researchers interact with the search interface and the above described visualization.

Back-end

The back-end supports the search process in ResXplorer with two main tasks:

- discovering links between two resources
- ranking of found links and resources.

[58] http://youtu.be/tZU97BQxE-0, last access: 2017–05–29

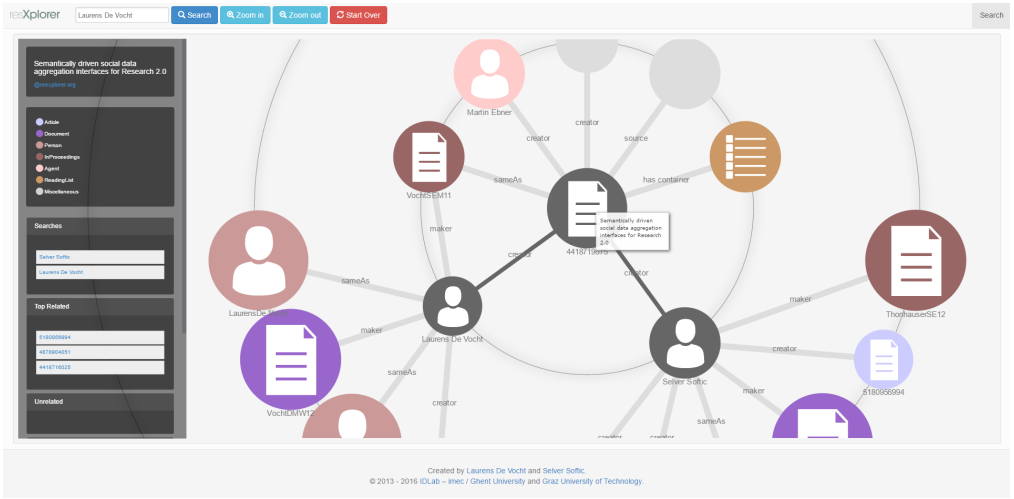


Figure 5.26: The explored relations are marked in same color. Adapted from De Vocht et al. (2016)

The architecture behind the back-end is depicted in figure 5.27. With the delivery of first results, the back-end expands the query and enhances the context by path finding and neighbors resolution within the "Everything is Connected engine" (EiCE) (see De Vocht et al. (2013a)). It uses the 'distance' to the first query as a measure for ranking the result. The EiCE Engine is used here to compute heuristically optimized minimum cost paths between pairs of researchers, publications, conferences or mixed pairs. The heuristics take into account the rarity of resources to avoid common resources (that have many in- or outgoing links) and the semantic relatedness between resources. Each time a user adds another resource to the results, the visualized path between the resources takes these factors into account.

5.5.6 Evaluation Summary

The conducted evaluation addressed two main groups of users. The primary target group of users are non-Linked Data researchers, and in this particular Research 2.0 use case, academic researchers. They interact visually with underlying Linked Data model. Another group of users are domain experts, as they are likely to have a very good understanding of data structure

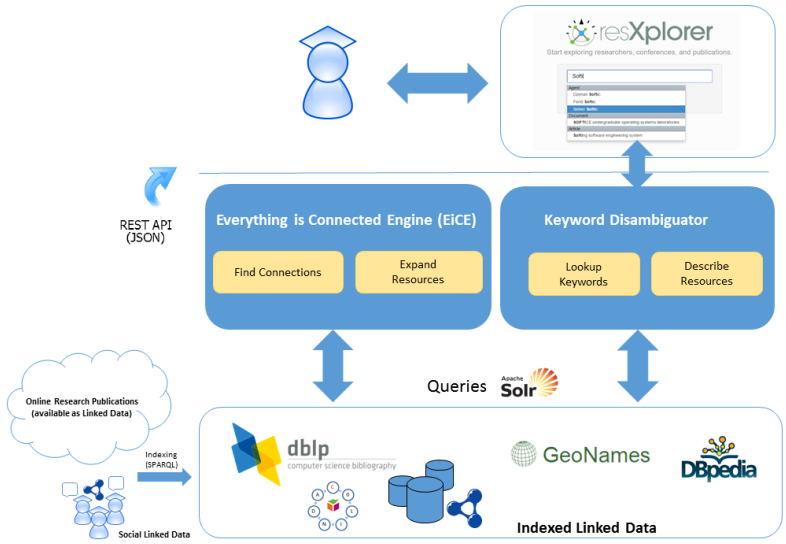


Figure 5.27: ResXplorer uses the Everything is Connected (EiCE) engine for finding relations between resources. Adapted from De Vocht et al. (2016).

and content in their domain, and bring this knowledge to guide both browsing research and targeted searches. All summarized results referenced in following subsection are elaborated into detain in De Vocht et al. (2016).

Methodology

Since exploratory search represents "a shift from the analytic approach of query-document matching toward direct guidance at all stages of the information-seeking process" (see White et al. (2006)), where users can at all stages see immediate impact of their decisions, conduction of such searches should be possible with minimal interruptions. The evaluation setup to test the concept of Exploratory Semantic Search in ResXplorer used two particular methods: a lean user test and expert user reviews. These methods give insight in how the users perceive the tools and show quickly potential bottlenecks Graves (2013). They also deliver insight on how precise the solution performs in comparison to the existent state of the art solutions of industry as well as academia. All of the compared solutions in evaluation

setup target the same audience. They differ in implementation and interface design, but more importantly, they have more or less valuable stock of users. The choice of such applied evaluation methodology came out of consideration of relevant aspects of already existing achievements in this field introduced in Kraaij and Post (2006); White and Marchionini (2007); Faulkner (2003). Since ResXplorer aims to offer a solution for research community users, a user centered methodology plays a decisive role in our evaluation process.

Lean User Study

Within lean user study a controlled experiment was conducted with 16 participants. This is according to Faulkner (2003) sufficient number to reach nearly high level of certainty for discovery of the most of the existing usability problems. The assignment was to find a relevant persons or events. They were asked to execute specific assignments and afterward to fill in a questionnaire with qualitative questions about their experience during the test. The users had to mark all found sources. Hereby they had a choice between three actions: searching, adding top related resources; this is done through disambiguated keyword based search on topics knowingly related to the initial search or expanding neighbors of found resources. While completing tasks (in time period of 30 up to 45 minutes) they had to think aloud. This was recorded and their comments have been tracked by an observer. Selected tests group representatives that participated in evaluation are researcher/experts in computer science and digital media. Obtained results lead to the conclusion that, based on the experience of previous work presented in De Vocht et al. (2015); Dimou et al. (2014a) (see also section 5.3.5), there was a good match between the test users and our target audience. The conducted lean user study measured the *Effectiveness* and *Productivity* of ResXplorer. The results showed that 'adding a top related resource' was not done often by the users and added only a couple of resources to the result set. However, it proved to be the most effective action as the users marked (50%) of the visualized resources relevant in this case. The 'adding top related resources' resulted in a result set that contained +12% more relevant nodes as before adding top related nodes, even though it has higher effectiveness (50%). This means that the impact

of each added resource when searching is much bigger, because the quality of the result set was not relatively high at the moment users decided searching. On average less than 31% of the resources, which would result in an increase in productivity if of the newly added resources at least 31% was marked relevant according to users. The effectiveness of expanding resources (32%) is about the same as searching for a resource (31%). As the user actions resulted in about as many new resources in the case of searching and expanding, this is a very reliable comparison. Expanding the direct neighbors is the most productive (+6%) expansion. Expanding further related neighbors retains the quality of the result set and barely impacts it, but the productivity is still positive (+1%).

Feature Impact Survey

Additional a survey has been conducted to measure the impact of the most two important features of ResXplorer: *personalization* (using social media data) and *pathfinding* (with EiCE (Everything is Connected Engine)). The results have shown there is no clear positive impact in cases when EiCE is enabled and a rather negative impact when personalization is enabled for simple queries. Over 60% of the users agrees that for complex queries the results when using the EiCE are preferred. For personalization the ratio is 45% positive against 36%, the bias is less positive here, but clearly better than the case for personalization with simple queries. When looking at enabling both features vs. disabling both features, nearly 66% prefers the results with both personalization en EiCE enabled and 56% in case of the simple queries.

Expert User Reviews

As already previously applied in Kraaij and Post (2006) for exploratory search expert user reviews used task based approach. To compare ResXplorer against industry reference academic search interfaces (Microsoft Academic Search, Google Scholar) and related academic projects (ARnet Miner, Falcons and Faceted DBLP) two researchers - search interface experts - independently reviewed the performance of each of these search interfaces.

They were familiar with all of the tools beforehand. A set of six representative tasks supported by these systems has been selected for the evaluation process. The search tasks designed for expert user review are optimized for state-of-the-art search engines and for ResXplorer and they are either simple (e.g. single fact or source) or complex (combinations of facts and sources). For each of the tasks have been measured the *average precision* (between 0 and 1) and the *efficiency* (expressed as number of actions needed).
Average Precision measures the average of the search precision over all the required actions in certain task. Thereby, the precision (see Powers (2011)) corresponds in this case to the effectiveness of the kth search action as defined for the lean user evaluation:

$$\textbf{precision} = P_k = \frac{TP}{TP + FP} \tag{5.5}$$

and the average precision over all actions A in certain task:

$$\textbf{average precision} = AP = \sum_{k \in A} \frac{P_k}{\mid A \mid} \tag{5.6}$$

However, the actions are different so a direct comparison for ResXplorer between the user action effectiveness and the precision measured here is not possible. It also would make no sense as the user tests focused on lean users while the experts are specialized in search interfaces.
To verify that the expert reviews are similar enough to be considered, the inter-rater agreement was measured among them. Therefore, for this purpose was selected the *chance corrected agreement* (κ) measure ($-1 < \kappa < 1$) initially introduced in Hripcsak and Rothschild (2005). The inter-rater agreement of the results between the experts is substantial ($\kappa = 0.61$ and F-measure 0.83) according to the scale in Landis and Koch (1977). Based on results of evaluation it has been shown that the ResXplorer is situated in the mid-range in terms of mean average search precision (mean value lies by 0.76) and requires relatively lots of action (in comparison to compared solution) from the user. However, ResXplorer is best when the task consisted of relating resources that are not directly related or when at least the user is not aware of how they are related. That is precisely the goal aimed with ResXplorer and the methods and techniques that drive it. Furthermore, this pinpoints, once again, to the importance and the need of user-centered evaluation concept within conducted measurements.

5.5.7 Discussion On Findings and Conclusion

According to the findings searching by keywords for resources increases the result set with the most new relevant resources, while it is on average as effective as expanding existing resources in the result set. The most effective user action was 'adding top-related nodes' to the visualization. The balanced choice of comparable solutions: from industry (Microsoft Academic Search[59] and Google Scholar[60]) and from research domain (ARNet Miner now known as Aminer[61], Falcons[62] and Faceted DBLP[63]); this allows good positioning and review of ResXplorer in a qualitative manner. Having visually more advanced solutions like Microsoft Academic Search and ARNet Miner and those with less search interface interactivity possibilities like Google Scholar and Falcons, the intention behind the experiment was to cover the essential aspect in evaluation for user driven search applications which considers the visual representation and analysis of search results and interaction possibilities on search interface. With ResXplorer users can combine any searches and interact the results that exposes relationships between them. This is a feature not found in conventional search interfaces. It offers search for publications, as well as supports relation visualization on author level. The ResXplorer visually emphasizes discovered types of entities and relations. Unlikely to the current existing solutions, ResXplorer can use the snapshot of social content published by researchers on social media and collaborative platforms like Twitter and Mendeley to make a pre-set for exploratory search. This feature is unique to ResXplorer. Furthermore, the method by which the context-based results are generated differs from ARnet Miner (Aminer) because the process do not rely on data mining and machine learning techniques to resolve the research related information. ResXplorer uses affinity based ranking derived from the social context and search process itself. Since pre-sets of the search reside on actualized Social Media content of the user (Semantically Modeled Data), presented solution adapts better on changes of information and trends from Social Media

[59] http://academic.research.microsoft.com/, last access: 2017–05–29
[60] https://scholar.google.at/, last access: 2017–05–29
[61] https://aminer.org/
[62] http://ws.nju.edu.cn/falcons/, last access: 2017–05–29
[63] http://dblp.l3s.de/, last access: 2017–05–29

as any other presented approach. This aspect differs strongly from the conventional approaches mentioned here.

5.6 Concluding Remarks on this Chapter

All presented sections describing the joint efforts on revealing advantages and disadvantages of semantic modeling and mining of sparse semi-structured information from tweets in combination with Linked Data ended up in a common journal publication **?** showing the application and efficiency of proposed methods in the use case of researcher profiling, user interfaces and exploratory semantic search for research purposes. As it can be seen reading through the chapters the idea of finding useful information in form of semantically modeled entities such as: events (conferences), publications and persons; evolved into creation of exploratory browsing interfaces, visualizations and a knowledge base of scientific resources enhanced with Social Linked Data which at the end offered an alternative approach for exploratory semantic search for research related artifacts on a personalized level that according to the achieved results performs regarding efficiency and usefulness at least as good as existing solutions. In some fields such as discovery of top related resources experimental implementation even excels in comparison to competitive solutions considered in evaluation process.

6 Potentials for Knowledge Discovery in Online Educational Communities

This chapter introduces experiments, prototypes, findings and concepts published as separate scientific works listed in 1.6. The text from the listed publications has been used to describe the methodology and concept, implementation, experiment conduction, evaluation and conclusion in form of a prototype. Description of these experiments reflect answers to research questions specified in 1.3.

6.1 Detecting Educational Communities on Twitter Using Basic Similarity Measures

6.1.1 Statement to Own Contribution

All ideas, prototypes and experiments conducted and described in following section are authors own work and they originate from Softic (2012). The main intention of this contribution is to test the eligibility of state of the art natural language processing NLP similarity measures for sparse semi-structured forms of text. As preparation of the usage of such measures for semantic modeling and detection in the area of education and technology enhanced learning I used tweets.

6.1.2 Motivation

This subsection reports about profiling collaborative learning groups of persons within the social micro-blogging platforms like Twitter that share potentially common interests on special topic. Hereby, the focus is held on spontaneously initiated collaborative learning in Social Media and detection of collaborative learning groups based upon their communication dynamics. Research questions targeted to be answered are: are there any useful data mining algorithms to fulfill the task of pre-selection and clustering of users in social networks, how good do they perform, and what are the metrics that could be used for detection and evaluation in the realm of this task. Recent research has shown that social interactions with people who share the same affinities can contribute progress in research and learning (see Mejas (2005)). Basic approach presented here uses as preamble hypothesis that users and their interests in Social Networks can be identified through content generated by them and content they consume. Special focus is held on topic oriented approach as least common bounding point. Those should be also the basic criteria used to detect and outline the learning groups. The aim of this action is to deliver a study on demands on implementation of recommendation systems using Social Network metrics and content features of Twitter users for the purposes of better learning group communication and research collaboration using Twitter. Processes that happen spontaneously are mostly initiated by adequate stimuli. As necessary precondition for stimuli of this kind as fundamentally important indicator a familiar ambiance will be assumed. All methodologies represented in following subsections will use this hypothesis as preamble.

6.1.3 Methodology

Thinking in manner of solving such complex task as collaborative learning content consummation inside of heterogeneous information ambient as Social Networks are, the first task that has to be solved is to identify the information stakeholder relevant for the process of collaborative learning with respect to information consumer. In order to achieve this first task area

of semantically-lexical analysis combined with NLP and data mining can deliver the proper tools and techniques.

However before the clustering process can be done, data has to be pre-processed and formed in a manner acceptable for common clustering algorithms. Then significant features of content should be used to determinate least common relation. In the case of Twitter this would be mentions denoted in micro text fragments with ”@someusername” and hash tags denoted using "#sometopic” .Hash tags are expected to contribute content related clustering while mentions will be used to discover relatedness in social context. This methodology follows the logic of item based filtering of recommendation systems design. To the best of authors knowledge no similar comparison or evaluations has been done so far in the area on similarity measures as preamble of item based recommendation of learning groups. Using these two common features as base for clustering and identification of potential collaborative groups makes sense since the persons who communicate about same topic and persons belong potentially to the same interest area. On the other hand persons mentioning the same communication actors also share implicitly an interest on the content generated from particular source. Tweets as short as they are, brought into a proper context can delivery astonishing results. Their usage as "social sensors” is applicable for several purposes. Lately some work on tracking the sentiment inside the "electronic word of mouth” as tweets were described has been published by Jansen et al. (2009) with respect to e-commerce area of appliance. This is a pre-assumption that has to be necessarily done before the context of learning groups in the manner of E-Learning can be considered. Therefore for now the focus of this paper remains on this pre-condition. Aim in this realm was targeted primary at evaluation of similarity measures needed for clustering of collaborative groups. As data source for intended experiments serves Grabeeter, previously introduced in section 2.5.2.

6.1.4 Definitions and Detection Procedure

Considered as simple concept a collaborative learning group can be primary treated as a "Interest Group". Let us define a potential "Interest Group” in

a more formal way. Let G be the a set of "Group Candidates" defined as follows:

$$G = \{G_i\} \texttt{ where i = 1 ... n and } n \in N$$

And a single member of this set $G_i = \{C_j, L_k\}$ is a pair of items where C_j is a vector of top content items and S_k a vector of top social references: ($\texttt{where j,k = 1 ... n and } n \in N, \texttt{ and where } j \neq k$). Items of both sets can be either single values or tuple of values. In current observation single values and value pairs depending on similarity function will be used (e.g. #hashtag or {#hashtag , 2} where 2 represent the occurrence). Also j and k indexes are of the same length, which means that we assume j = k. Let H be a single reference "Reference Candidate" of type "Group Candidate" as previously defined:

$$H = (C_r, L_r) \texttt{ where r = 1 ... n and } n \in N$$

Note that indexes j,k and r are the same length! Further T a pair of real value thresholds between 0 and 1 will be defined as follows:

$$T = \{t_c, t_l \in \texttt{ R } | \texttt{ } 0 \leq t_c \leq 1 \texttt{ and } 0 \leq t_l \leq 1\}$$

Intersection between the corresponding item sets C_j,C_r and L_r,L_k delivers a subset μ:

$$\mu = H \cap G_i = (C\mu, L\mu)$$

This subset delivers input for a similarity ration function α. This function delivers either correspondence ratio in percent between significant content or social reference items from intersection set μ respectively the "Group Candidate" vectors as a value between 0 und 1.

$$\alpha(\mu) = \{x \subseteq \texttt{R } | \texttt{ } 0 \leq x \leq 1\}$$

As final step a threshold based clustering function δ is applied on a to determinate whether a "Group Candidate" G_i belongs to an "Interest Group" or not.

$$\delta(\alpha, T) = 1 \texttt{ if } 0 \leq t_c \texttt{ or } t_l \leq \alpha, 0 \texttt{ otherwise.}$$

Hence "Interest Group" `I` is defined through following factors:
$I = (G, H)$ `where` $\delta(a(\mu), T) = 1$ or the matter of evaluation one additional measure will be defined called λ or "acceptance ratio". This is a ratio between the count of accepted and considered "Group Candidates".

$$\lambda = \texttt{\#accepted } G_i / \texttt{\#considered } G_i \ , \ 0 \leq \lambda \leq 1$$

As similarity function in the context of group detection Cosine Similarity was used for single valued vectors while Euclidian Distance was used as pair value vectors similarity measure.

Cosine Similarity: This ratio can be used as a similarity measure between any two vectors representing documents, text fragments, snippets or the like. Cosine Similarity represents the angle between two vectors that reflects their diversity. As the angle between the vectors becomes shorter, the cosine angle approaches the value of 1, which means that the two vectors are getting closer regarding their similarity. Total diversity is represented through 0. Cosine Similarity is defined as:

$$Sim(A, B) = \cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \bullet \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

Euclidean Distance: Euclidean distance is most often used to compare profiles of respondents across variables. In other words, Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. Note that the formula treats the values of X and Y seriously: no adjustment is made for differences in scale. In order to hold the scaling convention some correlations and scaling for the purposes of evaluation respectively expressing the similarity in percent as value between 0 and 1 has been made.

$$d(A, B) = \sqrt{\sum_{i}^{n} (A_i - B_i)^2}$$

6.1.5 Data Set Preparation and Measurement Process

As reference data for evaluation set top 100 results for persons from Grabeeter accounts register who used "elearning" or "e-learning" keyword

in their tweets were taken. This is done in order to compare the ratio of similarity respectively the size of candidate group. For evaluation purposes always the last 250 tweets of a specific user hast been taken into account. This was the biggest number of tweets that could be obtained for all test user. Out of them top 5, 10 and 20 hash tags and mentions per each user were generated and compared using similarity measures: Cosine Similarity and Euclidean Distance. Vectors are all of same length. Dynamical vector size adjusting was intentionally left out since the main point of matter rather whether the approach delivers promising results than the scalability of algorithm presented here. All measurement made respectively the detection of potential "Interest Groups" were made using a specially designed Similarity API based upon Grabeeter tool. Similarity API was implemented in PHP[1] using the Grabeeter database as primary data source. Results are delivered in JSON (see figure 6.1) format and finally processed into results using the statistic functions inside the API.

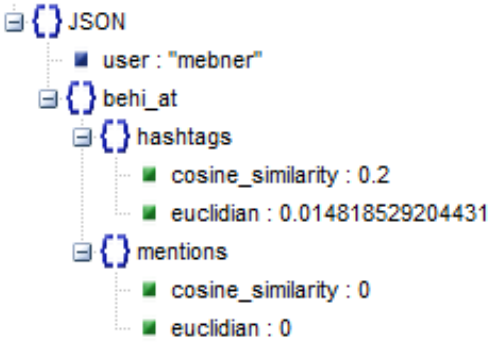


Figure 6.1: Similarity API as in Softic (2012).

Cosine Similarity and Euclidean Distance were used as similarity measure since they has been shown in various research works before (see Huang (2008)) as reliable indicators for detection of text based similarity. Distances used here belong in two different groups. Cosine Similarity uses only simple items to calculate the similarity angle among two text terms while Euclidian distance is calculated using the text item and their occurrence. Upon these results clustering using simple thresholds in percent in the range from 10% and 20% has been applied on similarity results. As a reference candidate for

[1] http://www.php.net/, last access: 2017–05–29

target learning group **@mebner** account was used since this account can be considered as one of the key competence bearer for E-Learning area. Each simulation consisted as described above out of similarity calculation and calculation of δ ratio function which checks if the result which is calculated for similarity reaches the threshold. "Interest Group" potential was reflected by the number of acceptable group candidates respectively the number of observed group candidates or as defined in λ ("acceptance ratio"). Values presented in the results section represent a median value of retrieval ratio. To get a deeper insight also the number of top hash tags and mentions was varied from 5 to 10 to 20 in order to evaluate how the length of parameter vector influences the result. Expectancy of presented measurement relies on the idea that comparison of different similarity measures should deliver first hints on building the collaborative groups techniques and an evaluation which of the measure fits best for proposed effort. It has to be considered that the test group was quite small but as it will be shown in the result section it delivers very encouraging results. Hereby, it has to be mentioned that the choice of keywords for filtering the users for candidate group as well as choice of reference candidate had a decisive influence on similarity level ratio as most important clustering criteria.

6.1.6 Preliminary Results and Discussion

At the beginning of this section it has to be mentioned that all of the observation made respectively simple clustering of the potential "Interest Group" are aiming at the evaluation of proposed methodology and system dynamics more than at qualitative analysis of retrieved results. "Interest group" detection is meant to be as pre-step for building the qualitative "Collaborative Groups". Described methodologies in this paper are meant to act as "sieves" and can be used as tools to simplify the task of building "Collaborative Groups" by reducing the number of potential candidates.

Single Valued Measurement Results with Cosine Similarity

Evaluation of "hashtag" vectors
Evaluation results for Cosine Similarity measure applied on "hash tags"

vectors of different length (5,10,20) with thresholds of 0,1 (10%) and 0,2 (20%) can be seen in figures 6.2: The 10% threshold can be easily reached

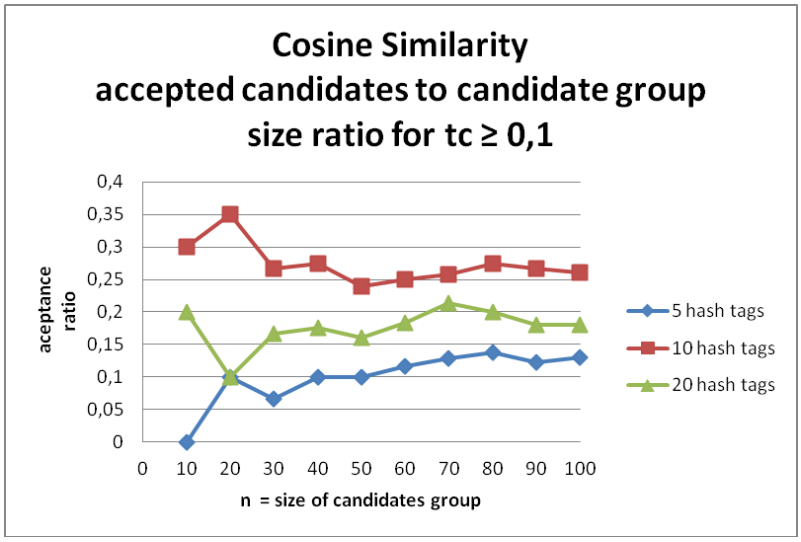


Figure 6.2: Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $t_C \geq 0{,}1$. As in Softic (2012).

result retrieved in figure 6.2 and their diffusion between 0 and 0,35 (or 0% and 35%) does not come surprisingly. The same can be observed for 20% threshold (figure 6.3) according the dynamics, although test with 10% boundary drifts more stable hand in hand with candidate group size, both of them tend to converge against a median value. Linear behavior of both systems relies on distribution of correspondences across the test set and on the nature of similarity function. Threshold with 10% is reached easily and causes less oscillation. Real nature can be recognized for candidate sets n > 80. Systems tend to stability as the candidate group increases. In figure. 6.2 there are some deviations for vectors of size 20. The reason is the structure of data set and its potential regarding the variation of vector size. Same can be observed for figure 6.3 and 5 "hashtags" sized vectors. It is obvious that significantly corresponding hashtags in test data set are placed at top 5 positions.

Evaluation of "mentions" vectors

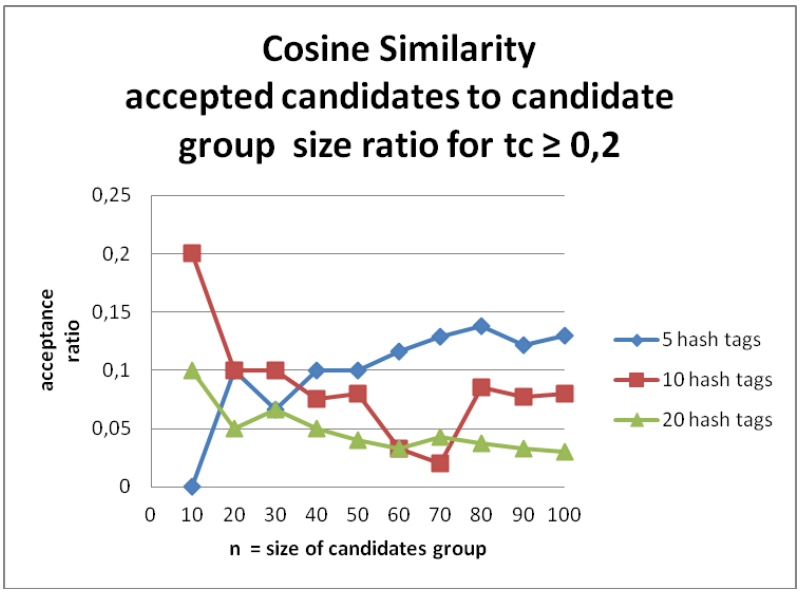


Figure 6.3: Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates $t_C \geq 0{,}2$. As in Softic (2012).

Figures 6.4 and 6.5 reflect the results of appliance of Cosine Similarity on "mentions". Same as in the case with "hashtags" the size of vectors was varied starting by 5 over 10 up to 20. For 10% matching threshold however the values of λ ("acceptance ratio") seem to perform better than for "hash tags" ($0{,}05 < \lambda < 0{,}45$). This fact points to the consistence better distribution and quality of "mentions" retrieved from test data. Same case can be observed also for the 20% threshold ($0 \leq \lambda \leq 0{,}3$). This is also reflected in the trend of λ which changes consistent together with the growth of number of group candidates. The same observation as for the "hashtags" can be concluded for the appliance of Cosine Similarity on the mentions in the case of linear dependency of "acceptance ratio" from the candidate group size. Dynamics of the system as already mentioned relies of distribution of interesting "mentions" and on the nature of similarity function. Deviation regarding the vector size are caused as in the case of "hash tags" by the placement of relevant "mentions" inside the vector. Interpreting the course and form of "acceptance ration" it can be easily concluded that in observed data set the mentions are distributed more equally all-over the data set.

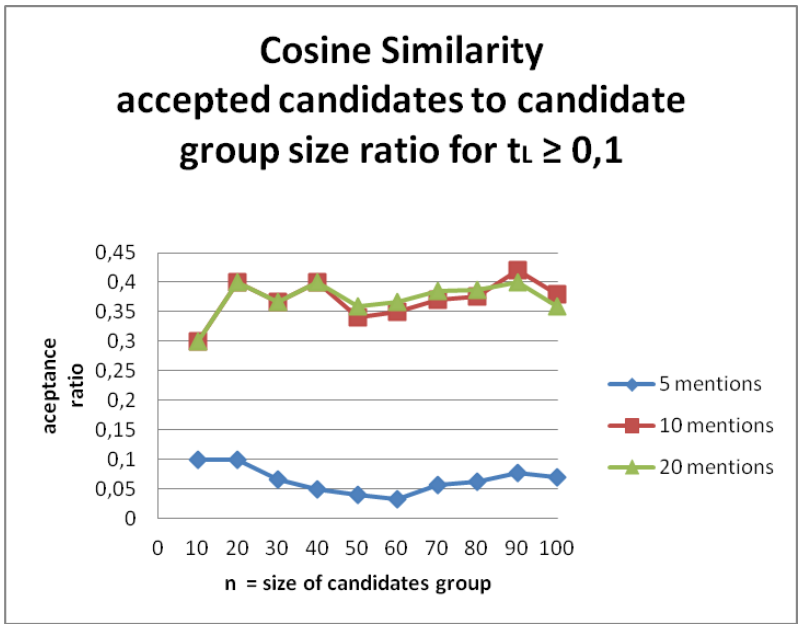


Figure 6.4: Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates $t_L \geq 0{,}2$. As in Softic (2012).

Pair Valued Measurement Results with Euclid Distance

Evaluation of "hashtag" vectors with occurrences

In following figures results based upon Euclidean Distance will be presented. Additionally to sole "hashtags" also their occurrences are taken into account by calculation of Euclidean distance. Occurrence frequency aspect as it will be shown lead to more stable behavior of "acceptance ratio" course.

Figures 6.6 and 6.7 are representing the results for thresholds of 10% and 20%. It is significant that larger number of "hashtags" in vector for the the case of 10% threshold also increases the "acceptance ratio" ($0 \leq \lambda \leq 0{,}3$). For the 20% thresholds this happens after the size of candidate group exceeds the count of 70 with approximately half lesser "acceptance ratio" ($0 \leq \lambda \leq 0{,}12$). Except two deviating values for n = 50 and n = 60 observations made

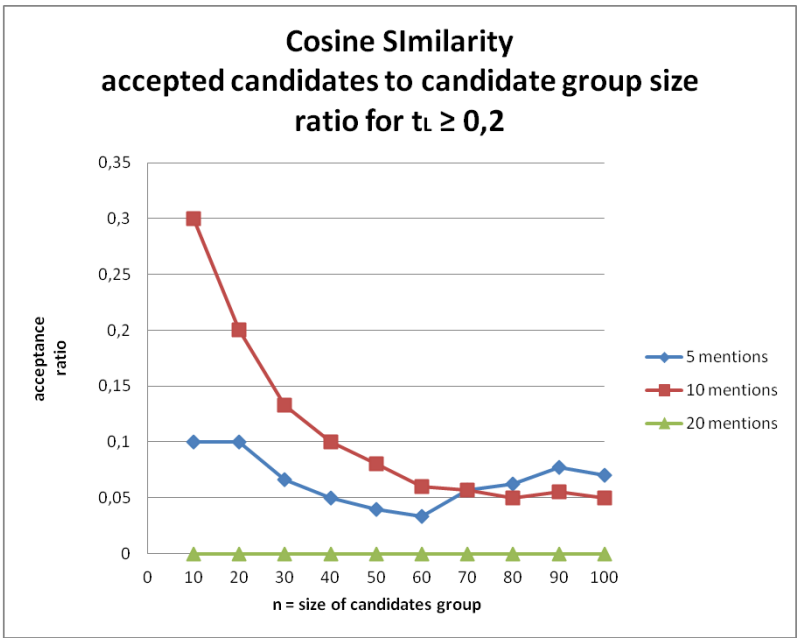


Figure 6.5: Cosine Similarity – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates $t_L \geq 0{,}2$. As in Softic (2012).

by 10% thresholds mainly correspond with the 20% case. It is also evident especially for the 10% that when a "acceptance ratio" reaches its nearly median value it hardly deviates heavily. Depending obviously on threshold this convergent behavior is reached at different count of candidates.

Evaluation of "mentions" vectors with occurrence frequency
Hardly different behave threshold based clustering based upon Euclidan Distance for input vectors consisting out of "mentions" and their occurrences which is clearly depicted in Figures 6.8 and 6.9. Once again size of input vector here filled with "mentions" and occurrence frequency influences the rate of "acceptance ratio". For threshold of 10% "acceptance ratio" varies in dependence on size of vectors between 0 in single case of 10 candidates and 10 "mentions" up to high rate of 0,4. Same characteristics are also measured by the 20% threshold. However here is the highest "acceptance ratio" value by 0,3. In comparison to the "hashtags" Euclidean Distance measurements

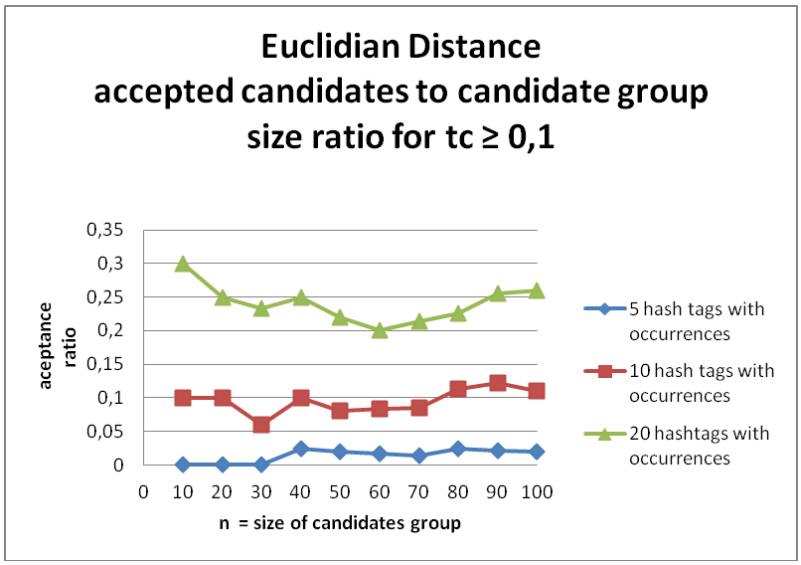


Figure 6.6: Euclidean Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $t_C \geq 0{,}1$. As in Softic (2012).

with same clustering threshold "acceptance ratio" does not decreases by the same coefficient. The reason for this behavior relies most probably on more equally dissemination of relevant vector items ("mentions") in test data set than the one of "hashtags" as in the case of Cosine Similarity for the same observation. Same as in the case of "hashtags" here even more evident the course of "acceptance ratio" values deviates lesser as the number of candidates increases.

6.1.7 Conclusion and Future Work

Despite the test set included only 100 candidates and one reference candidate conclusion dynamics of similarity measures based threshold driven clustering could be evaluated and observed with some valuable answers. Although no qualitative evaluation has been made, and "acceptance ratio" as such is clearly inaccurate indicator of the precise distinction of discovered “interest groups”, it was sufficient to approve the significance of the intention behind the usage of similarity based approach for organizing and

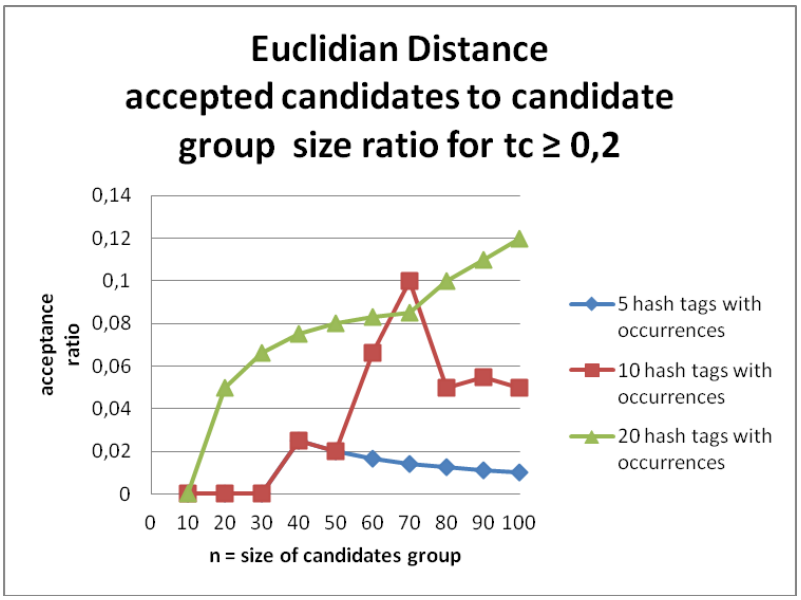


Figure 6.7: Euclidean Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $t_C \geq 0{,}2$. As in Softic (2012).

steering of targeted information exchange between the persons that have same interests participating in Social Networks as Twitter. Results presented in previous section are showing us that this approach looks promising even on very small data sets, which is encouraging for future works. The choice of parameters approved the initial expectancy of setting the first steps in right direction. Further it made possible the comparison of two approaches. Details from measurement also clearly outlined the facts about the stability of single measures. Euclidean Distance performed more stable and consistent in comparison to Cosine Similarity at least according the presented measurement. Some instability characteristics of Cosine Similarity can be explained by not equally dissemination of relevant matching items across the data set. Therefore, this measurement demonstrates even better realistic circumstances. It would be too optimistic to claim that the presented approach could be the end concept towards building collaborative learning groups however it seems to be a small step in right direction. It would be more interesting for future work to extend the measurement on more application cases and reference users from different areas. Additionally in order to en-

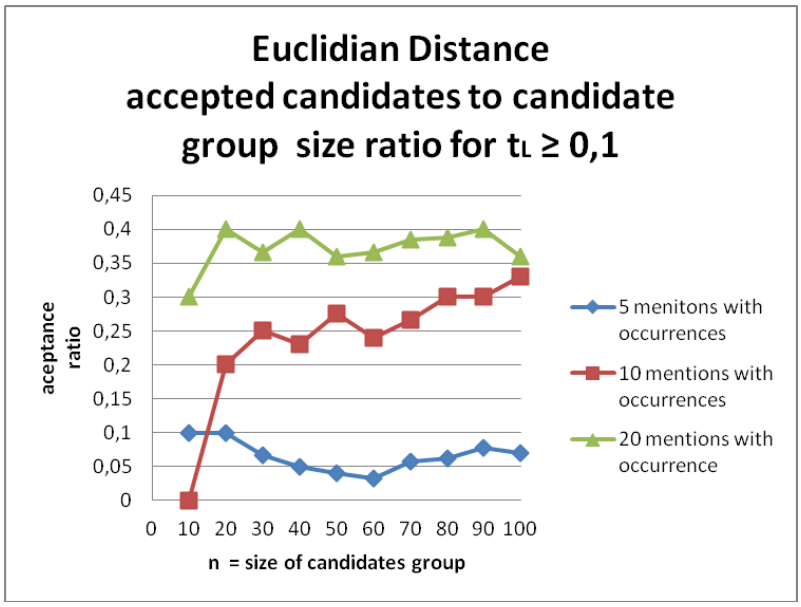


Figure 6.8: Euclidian Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $t_L \geq 0{,}1$. As in Softic (2012).

able more accurate and qualitative evaluation of clustering single matching similarities should be considered, clustered and re-evaluated more precisely during the measurement process. Also some other approved similarity measures like Pearson or Jaccard could be considered as extension to the current experiment setup. In this way it could be possible to determinate the level of quality of each single similarity method. Such extension of presented approach would contribute the reliability of the initial idea. Improvements towards preparation of more extended test data set are aimed to be done with expectation to re-approve the results. Nevertheless presented results confirm the basic intention of the current work made by author and other researchers towards improving organized collaboration and information placement and exchange in Social Networks and underlines the claims that such effort is based upon realistic expectations. Most encouraging about this approach is awareness that current scientific technologies, methods and techniques can be used to deliver complete solutions and answers to addressed challenges in a very near future.

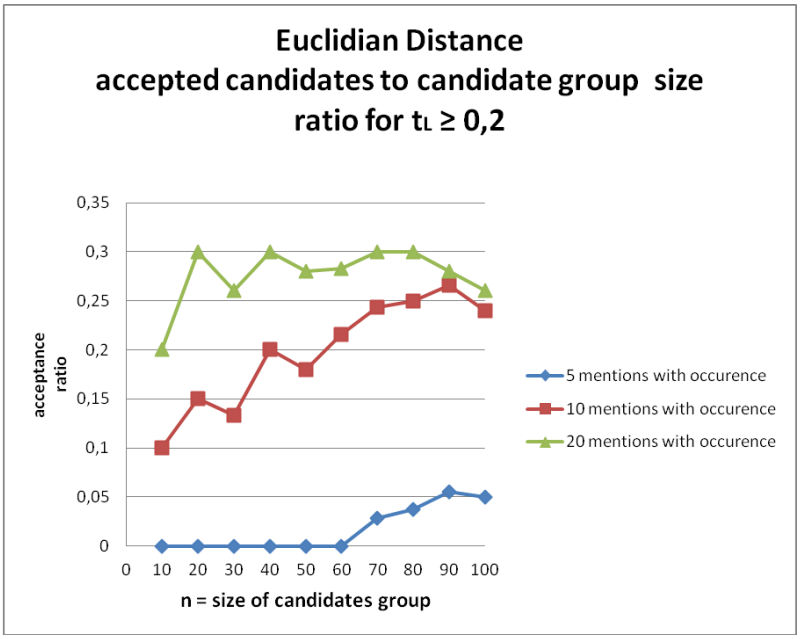


Figure 6.9: Euclidian Distance – ratio that reflects the percentage of accepted number of candidates to the total number of evaluated candidates for $t_L \geq 0{,}2$. As in Softic (2012).

7 Semantic Modeling and Mining Approach for Tracking Learners

The content of this chapter considers the user logs from PLE (personal learning environment) at Graz University of Technology as semi-structured text fragments. The approach described here uses the same methodology which includes semantic modeling of data and its preservation as RDF as well as SPARQL based retrieval as introduced in use case for research field. However, the usage of the results is different. The user system logs oriented part addresses Visual and Learning Analytics as main consumers of the mining process targeted on improvement of learning environments.

7.1 Modeling and Querying Learner Behavior in a Personal Learning Environment with Semantics

The overall idea is to analyze the learner's behavior in detail, the widgets they use or stop using for the given reasons and to map their monitored actions to the context of confidence regarding the system design.

7.1.1 Statement to Own Contribution

The concept of semantic modeling as well as preparation of data for querying, mining and visualization was made by myself. Mr. De Vocht and other co-authors contributed partly to the generation of visual results and choice of visualizations and partly to conception of analytic approach in particular

in the part of mocking up the initial version of proposed dashboard. The text in following section was originally published in Softic et al. (2013b) and resides on pre-work done in publications Softic et al. (2013c); Taraghi et al. (2013).

7.1.2 Motivation and Challenges

The Web 2.0 introduced intensive and wide-spread participation in online activities: the Social Web, where web users act as main content generators, became a reality and results of such circumstances are visible nowadays in form of social networks (e.g. Facebook, Twitter), resource sharing platforms or interactive collaborative environments for problem solving (see Ebner et al. (2007); Pohl et al. (2008)). The transformation of internet from consuming into interacting medium along with the corresponding web technologies influences strongly how we think, inform ourselves, organize our every day activities but also how we learn. This evolution is bringing new approaches to education. Massive Open Online Courses (MOOCs) for example aim for large scale worldwide participation. This became possible on the one hand thanks to advances in the technology and on the other hand by challenges resulted by organizing the education in general in order to adapt to the needs of modern learners with adequate time contemporary environments. The idea about open knowledge and open access also contributed to the developments in this direction. E-Learning platforms turned to be more efficient for tackling the problem of organizational and cost-effective matter[1]. Since the Web became not only consuming but also a producing medium evolving problem of Big Data is one of the next challenges for E-Learning to tackle in the near future. Limited availability of resources along with a time efficiency focus forces the designers and decision makers of learning platforms to revise their methodologies and techniques in order to respond the challenges of time and the needs of their targeted groups. On the other side learners are expecting a focused and simple way to organize their learning process, without losing time on information and actions which could disturb or prolong their learning, which also has a strong impact on acceptance of such platform (see Holzinger et al. (2011)). Therefore, nowadays

[1] https://tinyurl.com/lodcl7d, last access: 2017–05–29

learning process became more individual, versatile and activity driven with the tendency to ad-hoc initiated collaboration and information exchange. Such circumstances imply the need for a scalable, adaptive learning environment enriched with multimedia supportive materials, communication channels, personalized search and interfaces to external platforms from Social Web like e.g. Slideshare, Youtube channels etc. All these parameters increase the complexity of online learning platform design and organization. Dynamics involved in this process require nowadays shorter optimization cycles in adaptation process of Learning Management Systems and Personal Learning Environments. In order to provide the learners an attractive surrounding and to tackle the named problems use of Learning Analytics for optimization of learning process and design of learning surrounding emerges as the time passes by. Personalized learning dashboards with focus to the learning objectives are necessary. Additionally learning platforms need a more focused view on overall Learning Management System performance and activities. Growth of data produced as monitoring material to the common state of the art learning platforms reveals a new dimension of optimization possibility to monitor the usage of learning artifacts and learning activities of users individually and overall aiming at the analysis of emotion and affective data in learning environments. Such data contributes to the personalization and adaptation of the learning process and delivers new interfaces for Learning Analytics. The matter of observation in following research will be widget based Personal Learning Environment (PLE)[2] developed for the needs of Graz University of Technology. The PLE serves currently more then 4000 users, mostly students. The usage, activities and the use of the learning widgets has been tracked. Widget-based interfaces have been considered by Reinhardt et al. (2009b) to cope with learner awareness requirements as they allow dynamic addition of functionality (see Reinhardt et al. (2011)). For the purpose of analysis data was collected over 2 years in order to generate Learning Analytics services with visualization support, which reflects the overall usage and process within the PLE. Inspired by the research trends of previous years by Santos Odriozola et al. (2011); Pardo and Kloos (2011) current work wants to gain insights as in Mazza and Milani (2005) to optimize PLE and adapt the PLE to the learners by using more personalized methods of learning possibilities e.g.

[2] http://ple.tugraz.at, last access: 2017–05–29

through recommendations respectively consideration also introduced by Drachsler et al. (2010) with focus on learning widgets. Following subsections introduce the findings and concepts based upon semantic modeling of user behavior applied for visual data exploration to improve Learning Management Systems through the prism of parameters on a user, widget and activity centered level (see also Rosen et al. (2011)). A PLE does not intend to substitute a Learning Management Systems (LMS), but it is an additional learning environment to support self regulated learning. Presented model and analysis does not actually improve LMS, but it may have a role to improve the quality of learning by supporting students in their personal learning process through better services and functionality. The model of the learning context was done using combination of existing domain specific ontologies. Open and accessible interfacing and extendability on machine and human level offers advantages such as the possibility to enrich the analysis results and functionality with external sources and systems e.g. through Linked Data[3].

7.1.3 Concept Considerations

Use Case

PLE has no defined roles on teachers and learners, producers and consumers like in Learning Management Systems (LMS). PLE lies in the category of self regulated learning where students have the whole control over the services and resources they may need and would like to use. Teachers may recommend their students to use some widgets or resources in PLE as they may recommend them to read some books, but they provide nothing to PLE.

Semantic Modeling of Learner Logs

Concept Modeling considerations treated following dimensions for the PLE: reflection (by tracking users), prediction (tracking activities) and un-

[3] http://linkeddata.org/, last access: 2017–05–29

veiling hidden information (tracking widgets - learning objects in our case). All three dimension are directly in relation to each other which implies that reflection influences prediction and vice versa. The hidden information regarding the learning objects (widgets) is derived from these bidirectional bounds. This implication relies on modeling and the native concept of widget as learning object as it will be shown in further text. Revealing hidden information enables to find out how the learning process is going on in general and individually for each student in respect of what learners are learning: how often are they learning and whether it is continuously or not. This shows which learning objects are mostly used and hence is a possible indicator for usefulness. Considerations regarding prediction imply following the activities of learners. Assumed that we can extract some patterns within activities (what they do and also what goals do they have and to which extent they achieve a goal) teachers can predict the overall performance of their learners according to their activities. Plotting the overall activities of learners reflects their learning process within PLE: and in given context this is reflection.

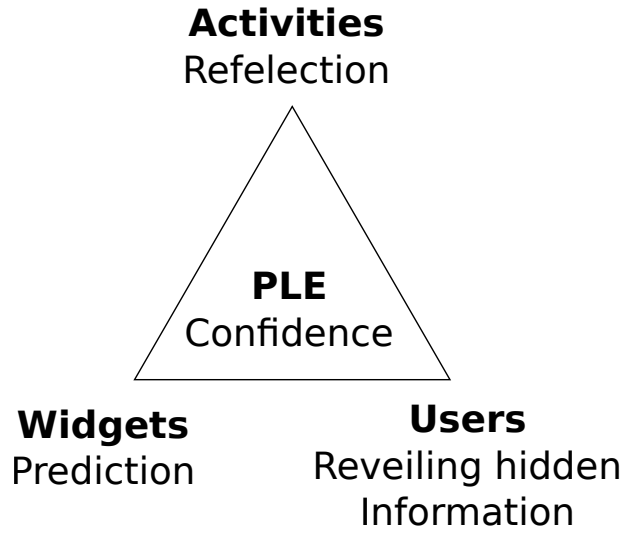


Figure 7.1: `Dimensions of PLE` Measuring confidence by monitoring widgets, activities and users as published in Softic et al. (2013b).

Purpose All statistics combined establish confidence in modules/widgets as interface between teacher (knowledge provision) and learner (knowledge consumer). The context of widgets is important to achieve reliable outcomes of the analysis of learner's activities. Figure 7.1 depicts the analysis of learner's activities ensures the optimization of the PLE focus to cover three modeling dimensions maximally by constructing a coherent view to support a call to action with high confidence.

Application Specific use cases based on statistics learned on the modeled learners logs should: contribute to better understanding of PLE usage, and reveal favored designs of the widgets. Further, intention that should be covered with this investigation is to orchestrate the insights into a recycling feedback loop to increase the overall acceptance of PLE as useful learning environment. Last targeted but not less important appliance of lesson learned should deliver initial information for improvements in recommendation system for widgets already integrated in PLE.

Dashboard for Analytics

As overall entry point to reflection within PLE should serve a prototype of a dashboard.

Concept To get an overview, PLE administrators/teachers have access to the "Dashboard" facilitating browsing the Learning Analytics from the PLE as shown in Figure 7.2. The dashboard contains views containing a graph visualization on the modeled information. The view is split in a summary which displays several graphs of measures derived from the raw statistics data to monitor the confidence and the balance of the learning environment.

Purpose The dashboard is a collection of indicators for administrators/teachers to get to-the-point feedback. They can deduct new views, broader or narrower; based on actions in the existing views because it should be allowed intelligently adding new views on the statistics data to the dashboard.

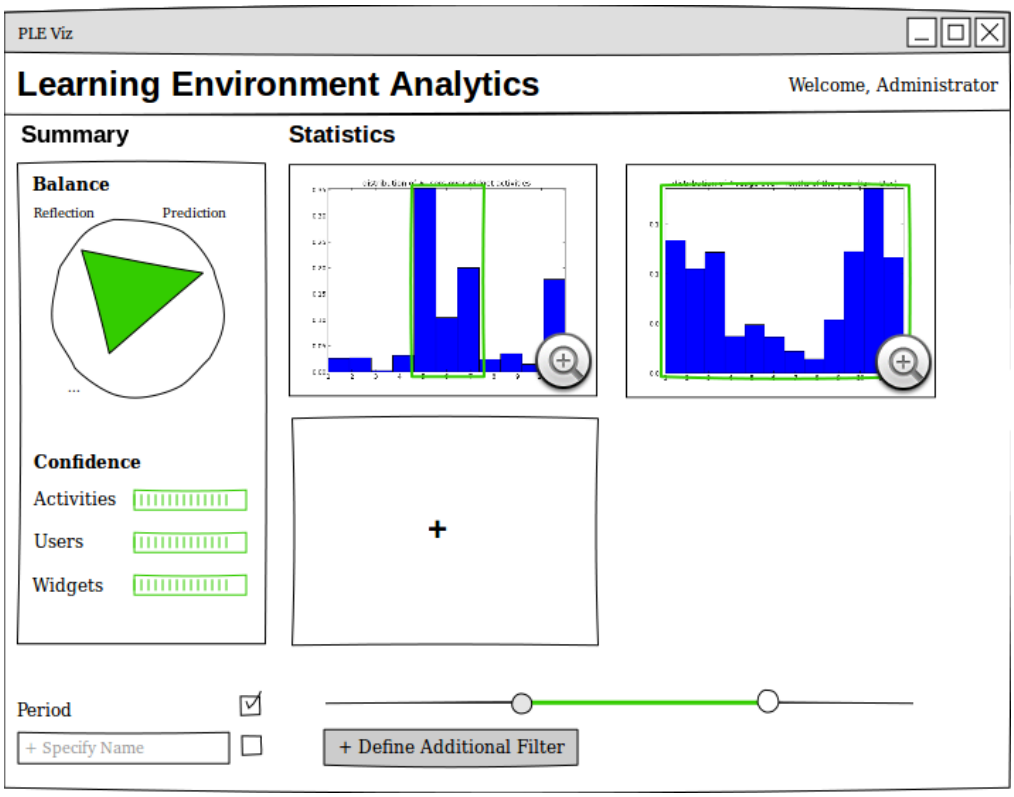


Figure 7.2: PLE Analytics Dashboard Overview of the planed available statistics and measures of the PLE as presented in Softic et al. (2013b).

The combination of different views and visualizations of analytics based learner's log data encourages administrators/teachers to take action and further optimize the learning environment/process.

Application The widget based interface for the dashboard guides users in constructing complex queries and revealing hidden correlations among the datasets through parameters and control interface elements as buttons, sliders and the like. It is an excellent way for putting analytics into context using categories, assumptions, and visually supported reasoning towards relating perspectives. In a broader context it should be also possible trough the addition and linking of multiple data resources.

7.1.4 Semantics for Learner Logs in PLE

Modelling

Final application of semantic modeling of PLE user's behavior is aiming at visualization and overview of three different kind of monitoring aspects interesting for optimization of PLE:

- **User centric** view where relations between the learner and the learning surrounding along with aligned activities should be outlined.
- **Activity centric** view where activities bound to the widgets that a learner is using are tracked.
- Finally **widget centric** where the whole perspective is reflected out of the sight of learning widgets.

With this purpose the data that was collected was tracked out of the PLE using simple log files which included information about a user (in anonymous way), about widget and activities related to the learning widget with additional time stamp when this logging event happened. Simple logging of data is unstructured and not easy query-able, the same problem is also with maintenance of such data. Generating specific visualization would in unstructured form imply formatting data into the form of visualization interfaces and requires additional efforts for each new visualization framework that would be used for implementation of such monitoring dashboard. In order to provide flexible data model that also delivers all wide accepted formats as e.g. XML or JSON as final output since those formats are very wide spread as input in visualization libraries our consideration lead us towards more operable and flexible data modeling framework and standards, for maintenance of tracking data. The intention was to make the data model extensible and scalable, and to additionally enrich the data with the context reflection in which such data was collected. Since the Semantic Web offers a flexible and scalable approach to modeling, formatting data in this way was the next logical step. SPARQL as retrieval technology driven by the efforts of W3C community reached mature level comparable to common occurrences. Output of SPARQL frameworks support XML, JSON or comma separated values. The challenge is to choose an adequate modeling vocabulary (in our case Ontology) since RDF offers only the framework how the data is

aligned and organized in such constructions. Fortunately current research in *IntellLEO* EU project resolved modeling dilemmas. One of the main goals of this project is building an *"innovative ontological framework for learning representation which includes learners, context and collaboration models, serving to achieve the targeted synergy"*[4]. In the realm of the *IntellLEO* project inside the provided ontology framework two special ontologies are eminent. The first is the Activity Ontology which offers a vocabulary to represent different activities and events related to them inside of a learning environment with possibility to describe and reference the environment (in this case PLE) where these activities occur. The second contribution from current Ontology research work in *IntellLEO* project is the Learning Context Ontology which describes the context of a learning situation. The PLE logs include the

```
@prefix ao: <http://intelleo.eu/ontologies/activities/ns/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix lc: <http://www.intelleo.eu/ontologies/learning-context/ns/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix um: <http://intelleo.eu/ontologies/user-model/ns/> .

<https://ple.tugraz.at/ns/activity/#Viewing> a ao:Viewing .

<https://ple.tugraz.at/ns/users/#FSKSN> a um:User;
    foaf:name "FSKSN" .

<http://ple.tugraz.at/ns/events/log/#7912> a ao:Logging;
    ao:performedBy <https://ple.tugraz.at/ns/users/#FSKSN>;
    ao:timestamp "2012-10-04T07:52:52" .

<https://ple.tugraz.at/ns/widgets/#LatexFormulaToPngWidget>
    a ao:Enivironment;
    rdfs:label "LaTeXFormulaPNG Converter" .

<http://ple.tugraz.at/ns/learningcontext/#7912> a lc:LearningContext;
    lc:activityRef <https://ple.tugraz.at/ns/activity/#Viewing>;
    lc:environmentRef
    <https://ple.tugraz.at/ns/widgets/#LatexFormulaToPngWidget>;
    lc:eventRef <http://ple.tugraz.at/ns/events/log/#7912>;
    lc:userRef <https://ple.tugraz.at/ns/users/#FSKSN> .
```

Listing 7.1: `LearningContext` in N3 RDF notation. As in Softic et al. (2013b).

events about learners who use a PLE while performing different learning activities in a certain period of time. Their activities comply to our use cases very well, which implicitly solved modeling vocabulary dilemma stated before. Representation of log entries from PLE as instance of a learning context concept can be seen in N3 RDF Notation in Listing 7.1. As stated in listing 7.1 depicted instance of `lc:LearningContext` class describes in compact N3 RDF Notation that a `ao:Logging` event occurred which tracked

[4] http://intelleo.eu/index.php?id=5, last access: 2017–05–29

the learning activity of `ao:Viewing` by certain `um:User` inside the learning widget named *LatexFormulaToPngWidget* as `ao:Enviroment` at certain time.

Querying

Beside the scalability and flexibility of data models Semantic Web also includes the advantage of traceability of such models using SPARQL query language[5]. Common storage and retrieval systems for semantic data instances support the exposure of so-called SPARQL endpoints, where the data from the storage (RDF triple store) can be easily retrieved by simple SQL like queries defined by SPARQL W3C standard. Additional advantage of such endpoints is that most of them deliver result data in common formats like XML,JSON or comma separated values. This functionality is essential for processing the retrieved results for visualization dashboard. Also very important function is that the endpoints offer implicitly standardized interfaces based upon RDF for data exchange to other platforms. Interoperability over the data is much easier then in the case if the log data would be stored in specific structure without standardization. In this way humans and machines readable, reusable activity knowledge artifacts has been produced with broader appliance field then a simple tracking log entry.

Listing 7.2 depicts in the best way how easily a question like: "Which users performed viewing in *LaTeXFormulaPNG Converter* widget after the first of January 2011?" can be answered by simple SPARQL query. This approach obviously enables easy pre-processing and thanks to SPARQL endpoints output configuration, the desired inputs for visualizations can be delivered in the same step. Semantic Web uses a "closed world" representation which means if there are no results when there is no answer possible in the system. The advantages of Semantic Web technologies combined with adequate vocabularies and ontologies do not only support easy and flexible analysis, it extends the repositories to the outside world while implementing implicitly many interoperability options for external analytic systems.

[5] https://www.w3.org/TR/rdf-sparql-query/, last access: 2017–05–29

```
PREFIX ao: <http://intelleo.eu/ontologies/activities/ns/> .
PREFIX foaf: <http://xmlns.com/foaf/0.1/> .
PREFIX lc: <http://www.intelleo.eu/ontologies/learning-context/ns/> .
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
PREFIX um: <http://intelleo.eu/ontologies/user-model/ns/> .

SELECT ?user WHERE
{
 ?x a lc:LearningContext;
      lc:activityRef <https://ple.tugraz.at/ns/activity/#Viewing>;
      lc:environmentRef
      <https://ple.tugraz.at/ns/widgets/#LatexFormulaToPngWidget>;
      lc:eventRef ?e;
      lc:userRef ?u.

 ?e a ao:Logging;
      ao:timestamp ?date.

 ?u a um:User;
      foaf:name ?user.

 FILTER ( ?date > "2011-01-01T00:00:00Z"^^xsd:dateTime )
}
```

Listing 7.2: SPARQL query filtering Viewing action on LatexFormulaPNG widget. As in Softic et al. (2013b).

7.1.5 First Results - Visualization of Statistics for PLE Dashboard

In this subsection some possible statistics visualizations have been presented that can be generated for the first prototype of PLE Analytics Dashboard. According to the PLE measuring confidence triangle described before, the statistics has been modeled into three dimensions. These dimensions are illustrated through some examples. The dataset used to generate the following statistics contains the user log data of about last two years in PLE.

User Centered Examples

Each widget in PLE is associated to one or more activities depending on the functionality that is provided by the widget. For instance, *Twitter* widget is associated to the activities *Reading*, *ContentSharing*, *DiscussAsynchronouly*, *Viewing* and *Search*. The other defined activities in PLE are *Authoring*, *Learning*, *Game*, *Quizzing*, *Computing* and *Listening*. Figure 7.3 depicts the distribution of users over all activities in PLE. The diagram illustrates that most of all users are engaged in the activities *Reading* (4290 users) followed by *Authoring* (2461 users) and *Search* (2156 users). In contrast *Listening* (33

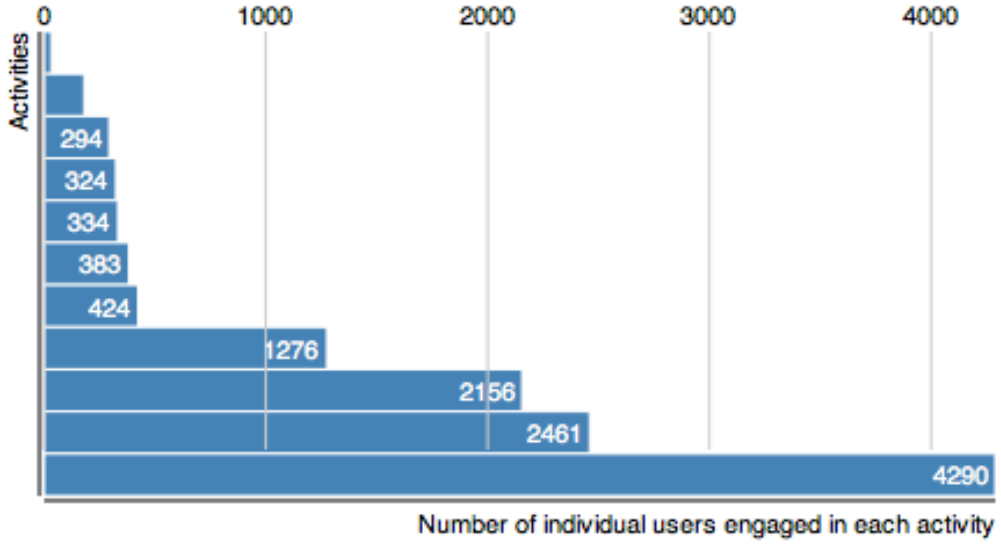


Figure 7.3: PLE `statistics` Distribution of users over activities in PLE. As published in Softic et al. (2013b).

users), *Computing* (181 users) and *Quizzing* (294 users) are rarely popular for users.

Widget Centered Examples

Figure 7.4 demonstrates an example for widget centered statistics. It shows how often each widget is used in each period of time in PLE. The widgets *ZID News* (representing the actual news related to the Central Informatic Service), *TUGraz online* (Administration System), *TUGraz Newsgroups* (News groups), *TUGMail* (E-Mail service) and *TeachCenter Courses* (LMS platform) are listed on the top as the most frequent used widgets in the last two years. All these widgets represent university services that students daily use. According to the visualized statistics the highest range of user activity can be monitored from October (begin of the winter semester) until July (end of the summer semester). On the first week of January as well as in

summer holidays no active usage cam be seen in PLE. That is actually expected. The visualization helps to detect widgets that are not popular at all or have been rarely used over the whole monitored period. Interestingly it can be observed no significant change on this behavior considering different period of times and different users. Widgets *Google Search*, *Address Book*, *Plane-Sweep Algorithmus* and *laengste gemeinsame Teilfolge* (a learning object to support learning the algorithm) are such examples that must be revised in a further development process. The other observations can be taken from this visualization is the development of PLE usage in general during the time. It is obvious that the frequency and quantity of used widgets have been increased in year 2012 in comparison with 2011.

Figure 7.5 demonstrates an example that can be of high interest. It demonstrates the activities of a specific user during a time period (in this example over the whole monitored time). The sorted list of widgets that the user have been actively using can be seen on the diagram. It shows that the user has been constantly using some widgets (*KulturKalender Graz*, *ZID News* and *TUGMail*) since February 2012. *TUGMail* widget is an exception. The user has stopped using it from April to August 2012. Figure 7.6 demonstrates the activity of another user who uses only two widgets: *ZID News* and *TUGraz Newsgroup*. It is obvious that she has been using *ZID News* continuously.

Activity Centered Examples

Figure 7.7 depicts the distribution of user activity over all activities in PLE. This diagram resembles figure 7.3 which depicts the distribution of users over all activities in PLE. The diagram shows that the activities *Reading* (28406 times) followed by *Search* (10588 times) and *Authoring* (9437 times) are most top popular ones. In contrast *Listening* (194 times), *Computing* (295 times) and again *Quizzing* (530 times) are rarely popular for users.

Figure 7.7 depicts the same situation over the whole monitored time period: an overall picture of the activity usage intensity. Again our observations from previous statistics can be confirmed. The list of activities on figure 7.7 are sorted according to the popularity and dominance during the whole monitoring period. The same results can be achieved here. *Reading*, *Authoring* and *Search* are dominant activities, clearly seen in the year 2012

compared with 2011.

7.1.6 Discussion on First Results

The overview over distribution of activities can reflect the overall interest of the learners within PLE. It can be concluded that in case of our PLE users are more consumers that contributors. Visualization of statistics can help to improve the PLE in general. Activities such as *Quizzing* and *Learning* (supported by some learning object widgets) are not quite popular. Conducted investigation showed that the corresponding widgets that support those activities must be revised in regard to some usability issues. The analytic results can help by obtaining a kind of rating/quality measure for the widgets that can be used as an indicator of likely future activity in the PLE. Distribution of usage of widgets over time in PLE showed exactly which widgets have been popular in certain period of time. Widget centered statistics for a specific user reflect user oriented statistics on which widgets are favored by a single user: one can observe if this trend is traceable over time or not. It delivers fast overview of affinities of single user considering the usage of special widgets. It can also be used e.g. as a basis for recommendation of new widgets in the widget store within PLE. Through activity centered statistics target users gain a better insight in the activities done in the PLE and their use. Further insights that could be gained would be e.g. about dominant activities, activity dissemination over time and activity peak usage periods.

7.1.7 Conclusions and Further Steps

Conducted prototyping experiments demonstrated that using semantic technologies enables the extensibility of Learning Analytics dashboards. Semantic approach generates uniform interfaces for information exchange, enables flexibility for Visual Analytics, and also includes the flexibility regarding the enrichment of Learning Analytics data with Linked Data. The spread of applicability covers wide range of analytic methodologies like

prediction, reflection and as result of these the intervention field. Future efforts regarding improvement semantic structure data layer, besides the mentioned Linked Data could also include precisely defined categorization of learning widgets, since PLE includes also this information. Especially the learning widget store as part of PLE could profit from this improvement. The statistics visualization help us to gain deep insight into the behavior of a single user in a certain period of time. Design examples achieved towards PLE Analytics Dashboard have been presented. The statistics examples covered the user, widget and activity centered dimensions of the PLE confidence model introduced in this section. The main question that arises from these results is how the PLE Analytics Dashboard must be further improved to meet these goals.

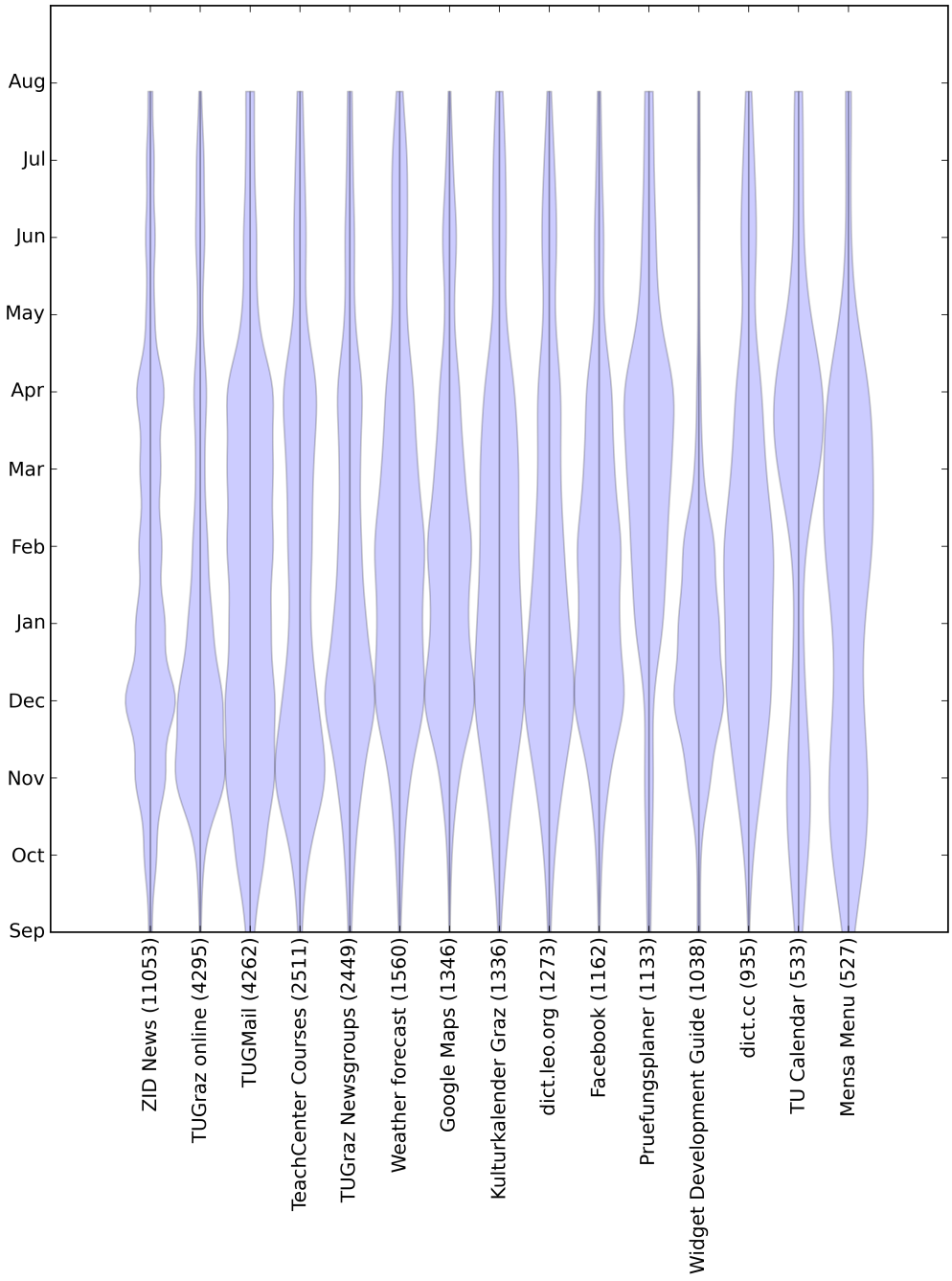


Figure 7.4: PLE statistics Distribution of usage of the 15 most popular widgets each month in PLE as published in Softic et al. (2013b).

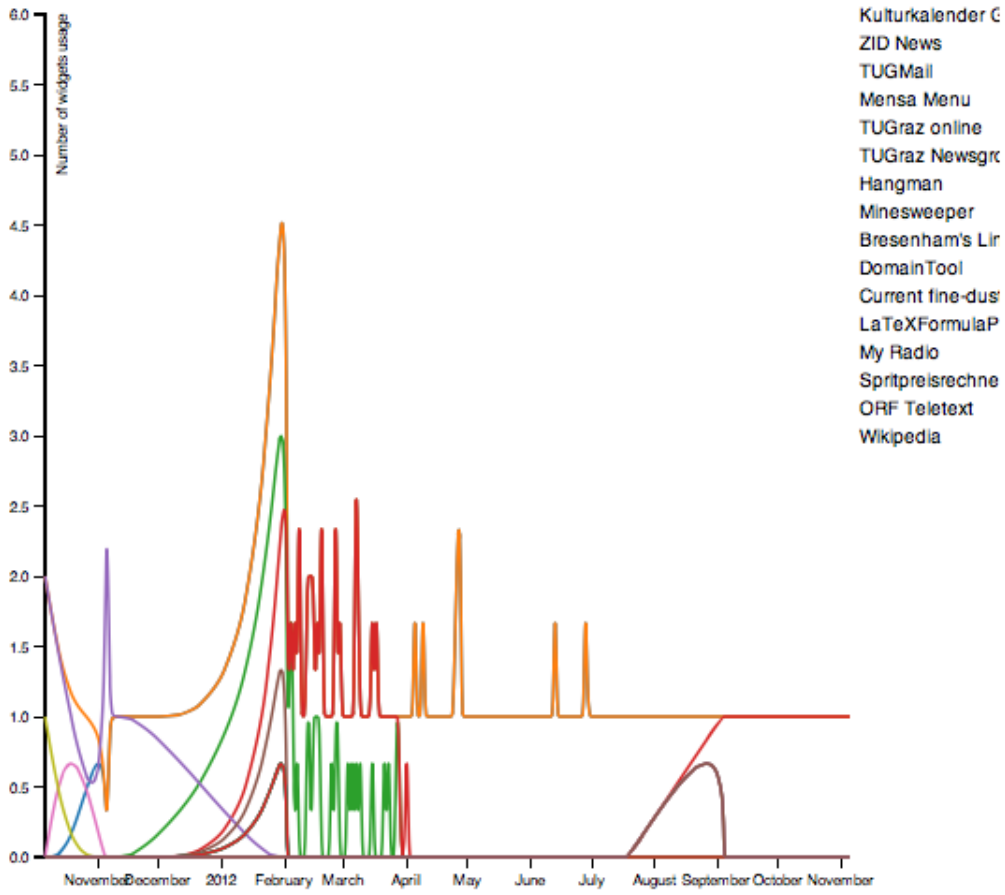


Figure 7.5: PLE `statistics` Distribution of usage of widgets by a sample active user over time in PLE as published in Softic et al. (2013b).

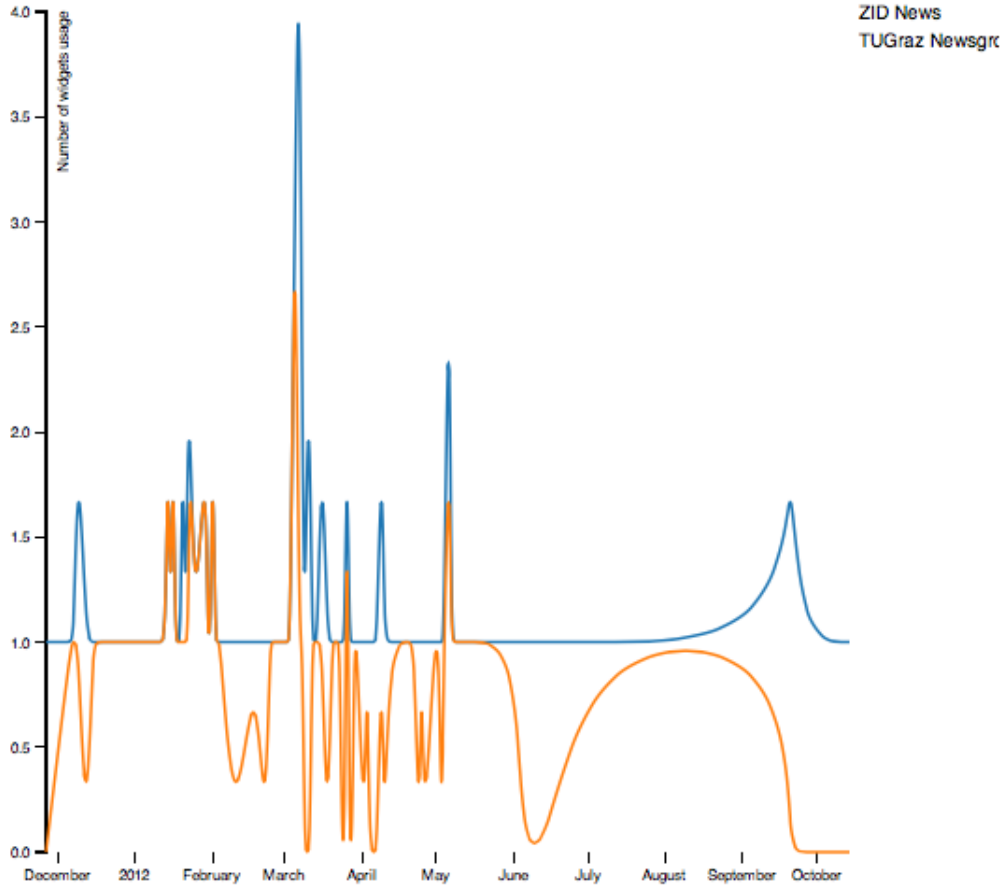


Figure 7.6: PLE `statistics` Distribution of usage of widgets by a sample active user over time in PLE. Widgets: ZID News and TUGraz News as published in Softic et al. (2013b).

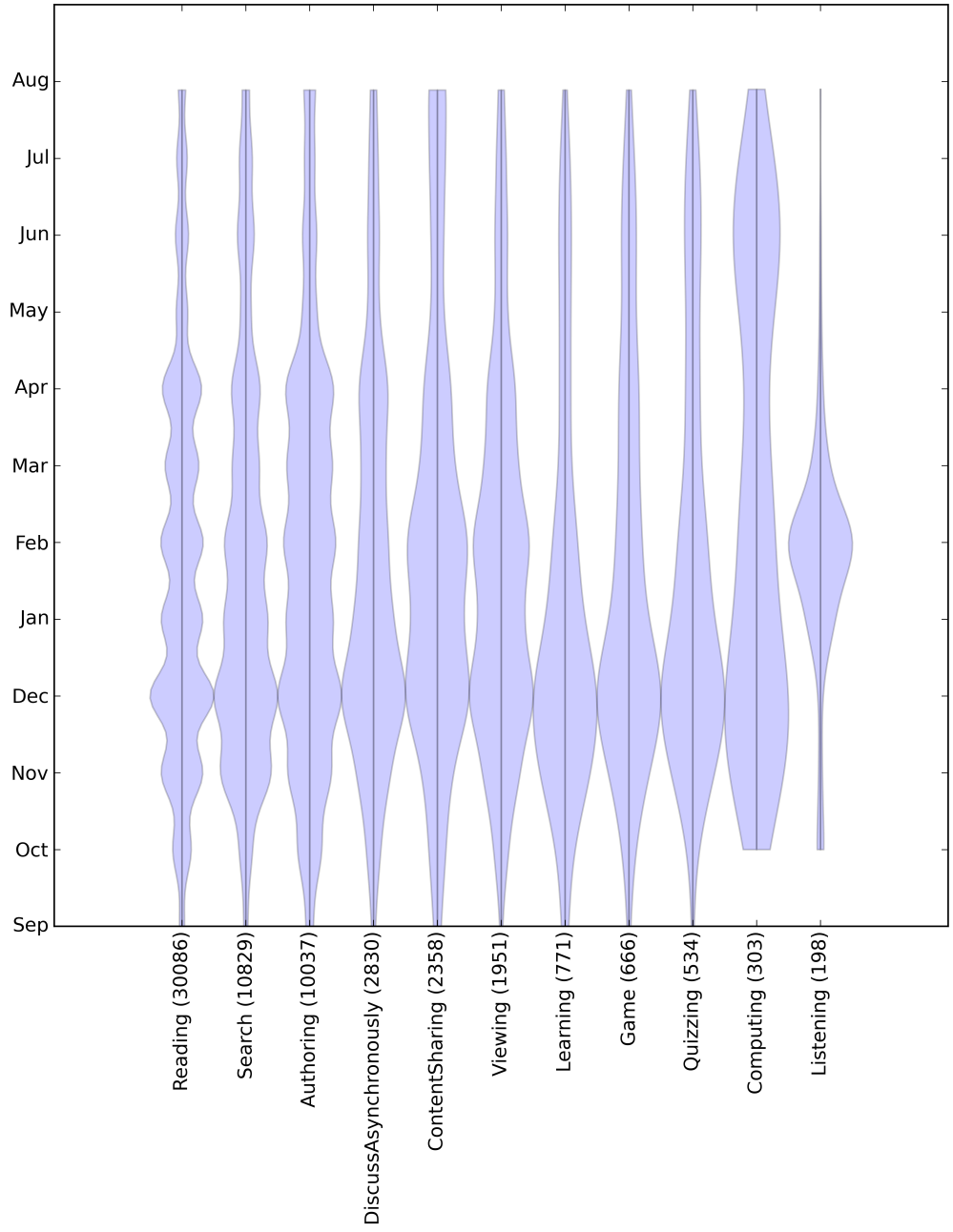


Figure 7.7: PLE statistics Distribution of activities occurrence each month in PLE as published in Softic et al. (2013b).

8 Exploitation of Proposed Approaches for Education

8.1 Leveraging Learning Analytics in PLE using Linked Data

8.1.1 Statement to Own Contribution

Findings and prototyping experiment presented in this section aims at showing the analytic potential of semantic modeling and mining of learner logs represented as Linked Data in Personal Learning Environment (PLE) of Graz University of technologies. The content behind this section was published in Softic et al. (2014a, 2015a) and relies on previous works published in Taraghi et al. (2013); Softic et al. (2013c,b). I am the main author of the used methodology, prototype, and the use cases applied within the published work. The other co-authors are mentioned because parts of the text has been used from the previous publications related to this text as well as because they co-authored the previous publications to the same context.

8.1.2 Introduction to the Efforts Presented in this Section

This section reports on the reflection of learning activities and revealing hidden information based on tracking and visualization of PLE user behavior through mining of learner user logs. Presented approach was previously tested and described in Softic et al. (2013c,b). Approach as such introduces usage of semantic context modeling and creation of Linked Data from

learner user logs in Personal Learning Environment (PLE) at Graz University of Technology with focus on reflection and prediction of trends in form of experimental prototype of Learning Analytics dashboard. The implemented prototype demonstrates the application of semantic modeling of the learner context, from data collected for over two years from widget based Personal Learning Environment (PLE) at Graz University of Technology. The approach models learning activities using adequate domain ontologies, and allows querying them using semantic query language SPARQL as input for visualization in the prototypical dashboard which serves as reflection and prediction tool for potential technical and functional improvements like widget recommendations. As presented, this approach offers easy interfacing and extensibility on technological level and fast insight on trends in e-learning systems like PLE.

8.1.3 Motivation Behind the Proposed Approach

Limited availability of resources along with a time efficiency focus forces the designers and decision makers of learning platforms to revise their methodologies and techniques in order to respond the challenges of time and the needs of their targeted groups. On the other hand, learners are expecting a focused and simple way to organize their learning process, without losing time on information and actions which could disturb or prolong their learning. Nowadays learning process became more individual, multi-faceted and activity driven with the tendency to immediate initiated collaboration and information exchange. These circumstances imply the need for a scalable, adaptive learning environment enriched with multimedia supportive materials, communication channels, personalized search and interfaces to external platforms from Social Web like e.g. Slideshare, Youtube channels etc. All these parameters increase the complexity of online learning platform design and organization. Dynamics involved in this process require shorter optimization cycles in adaptation of learning process. Also maintaining such platforms is intensively changing process demanding from maintainers to actively adapt their systems to the learner needs. Adaptation to learner needs has a strong impact on acceptance of such platforms and should be matter of continuous improvement. Accumu-

lated system monitoring data (e.g. logs) of such environments offer new opportunities for optimization (see Prinsloo et al. (2012)). Such data can contribute the better personalization and adaptation of the learning process but also improve the design of learning interfaces. The idea behind this effort is aiming at gaining insights useful for optimization of PLE relevant for the system designers and teachers and adapting the system to the learners by using more personalization e.g. through recommendation of interesting learning widgets (see also Mazza and Milani (2005)).

8.1.4 Mining Learner Logs for Learning Analytics Dashboard

Dataset

Data used in the case study for implementation of dashboard prototype originates from Personal Learning Environment (PLE) developed for the needs of Graz University of Technology which serves currently more than 4000 users. The data was collected during two years period in order to generate analytics reports with visualization support for overall usage and process view on PLE environment following the research trends of previous years presented by Santos et al. (2011); Pardo and Kloos (2011). More details on PLE are also described in section 2.5.4.

Data Modeling for Mining Learner Logs

Meaningful mining of learner trends requires meaningful modeling of data. Such action assumes the choice of appropriate vocabulary or ontology. The RDF standard as such offers only the generic framework how to: organize, structure and link data. The Activities Ontology introduced by IntelLEO research project already analyzed by my prior research work Softic et al. (2013c,b) introduced in section 7.1.3 and elaborated in related work in sections 2.9 and 2.9.1 offers a vocabulary to represent different types of activities and events related to them. Further this ontology also supports the description of the environment (in our case PLE) where these activities occur as well the description of actors within (learners). The Learning Context ontology serves as shown in figure 8.1 as container model to link

PLE usages as event to the widget as execution environment where this event happens as well to the user who performed the action related to this context.

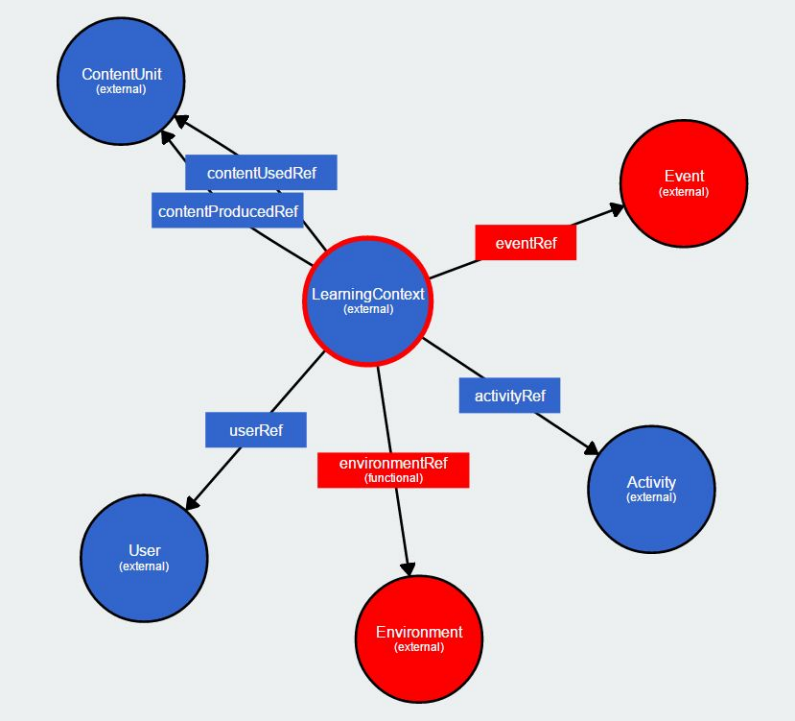


Figure 8.1: Visualisation by WebVOWL beta 3.0 of a LearningContext ontology concepts and properties used to model the PLE log data. As published in Softic et al. (2014a).

Every learner action of a PLE widgets creates a logging entry which produces a RDF construct similar to the presented example in listing 7.1. Such constructs represent Linked Data instances which are then stored in a RDF memory store: graph database for Linked Data with SPARQL Endpoint (an interface where Linked Data can be queried). The general concept of this process is shown in figure 8.2.

In listing 7.1 a sample instance of **lc:LearningContext** links the usage log event denoted as instance of class**ao:Logging** , a subclass of an **ao:Event**, which occurred at certain time point inside the learning widget named **LatexFormulaToPNGWidget** represented through class **ao:Enviroment**. As

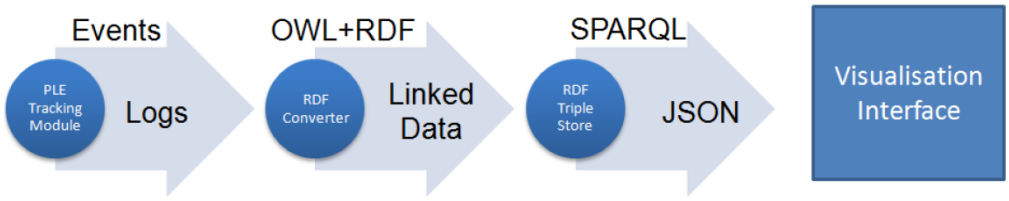


Figure 8.2: Sample simplified mining pipeline for PLE learner logs. As published in Softic et al. (2014a, 2015a).

demonstrated in example explained above, vocabularies and ontologies which suits well to specific use case, enrich the analytic process with a high level of expressiveness in a very compact manner.

Querying the Semantic Model Instances

Usage logs data presented as Linked Data graph can be queried using SPARQL. In this way we are able to answer the questions like: *"Show me the top 10 used widgets?"*. Figure 8.3 represents exactly this question stated in the manner of SPARQL syntax. The benefit of this query is visible for instance in figure 8.4 where the results of this query (see figure 8.3) influence the widget arrangement in the widget store. Such direct impact on system with functional interoperability on machine level would not be possible without standards like SPARQL and RDF.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
PREFIX lc: <http://www.intelleo.eu/ontologies/learning-context/ns/> .

SELECT DISTINCT ?widget ?ref (COUNT(?ref) as ?count)
{
 ?link rdf:type lc:LearningContext;
       lc:environmentRef ?ref.

 ?ref rdf:label ?widget.
}
GROUP BY ?ref
ORDERBY DESC (?count)
LIMIT 10
```

Listing 8.1: Querying the usage of top 10 widgets in PLE. As in Softic et al. (2014a).

8.1.5 Extended Results, Conclusion and Outlook

Presented approach allows us mining the trends of PLE widgets usage overall time periods like presented in figure 8.3. This pie chart graph depicts the visual answer of the query from listing 8.1. The overview over distribution of widget usage can reflect the overall interest of the users within PLE for different periods of time.

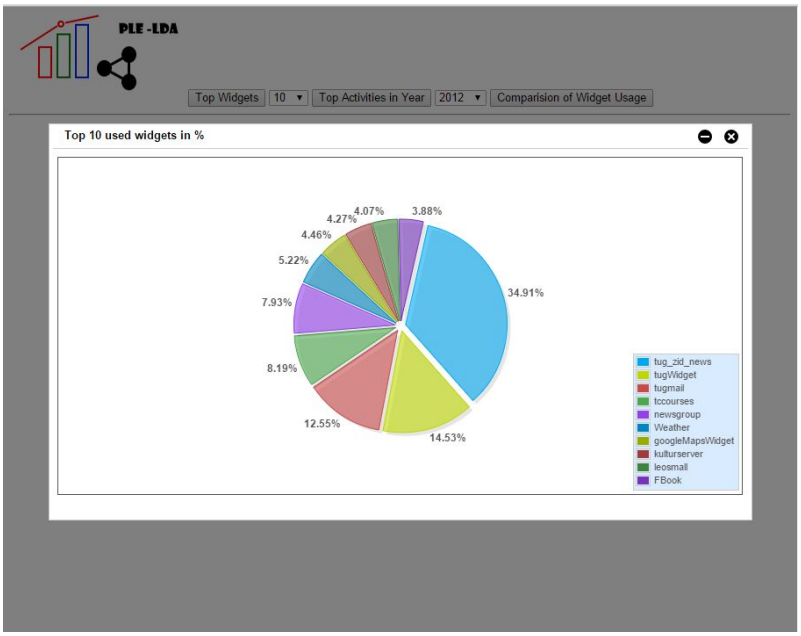


Figure 8.3: Visualising top 10 used widgets in PLE. As published in Softic et al. (2014a, 2015a).

Such outputs implicitly support the improvement of the quality of services for students and teachers. The same results from query in figure 8.3 are also used as input for ranking of widgets in widget store depicted in figure 8.4.

Beside widget centered reflection and trend monitoring the experimental implementation of Learning Analytics dashboard supports activity centric statistics as shown in figure 8.5. These examples show the manifold application of presented approach. The PLE becomes, in technical manner,

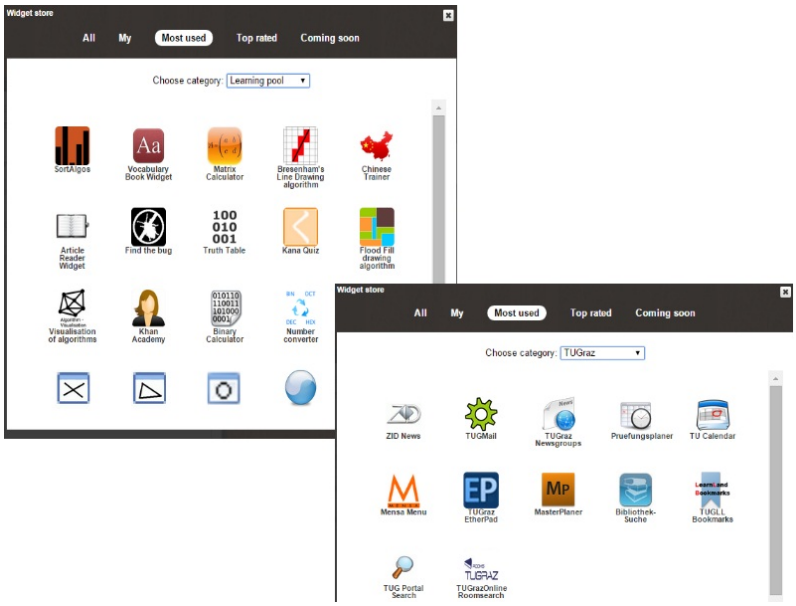


Figure 8.4: Optimized widget store based on Linked Data statistics. As published in Softic et al. (2014a, 2015a).

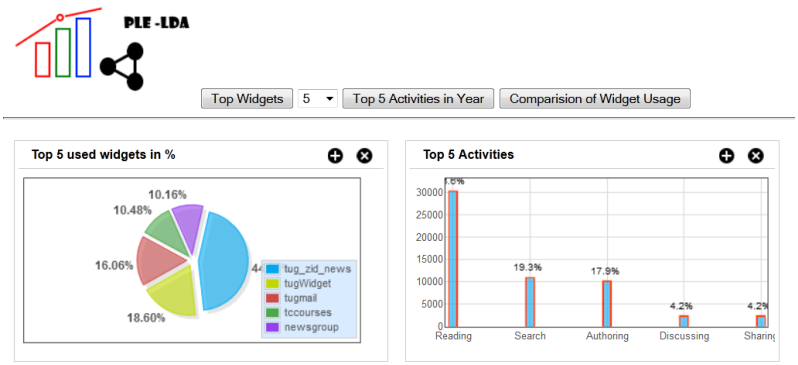


Figure 8.5: Visual Analytics Dashboard. As published in Softic et al. (2014a, 2015a).

extensible and well connected by standardized and intelligent interfaces and available for other web based tools and services. Future efforts will focus on user wise statistics of learning widgets, since PLE can also provide this information. Especially the learning widget store as part of PLE could profit from this improvement. Mostly used and favored widgets by users will

be ranked higher and recommended by the store itself as shown in figure 8.4. This process will be personalized as soon as the user information is included. The presented Learning Context ontology as such have foreseen such option already. By tracking the usages on user level the teachers will be able to draw conclusions about the popularity and quality of their learning widgets, on more granular and personal level.

8.2 Linked Data Driven Visual Analytics for Tracking Learners in a PLE

8.2.1 Statement to Own Contribution

The content from this section represents the final prototype of the idea of analytic dashboard of semantically modeled system logs from PLE at Graz University of Technology developed as idea initially in Taraghi et al. (2013); Softic et al. (2013c,b) and conceptualized as experimental prototype in Softic et al. (2014a, 2015a) (described in section 8.1) with the final application on the use case of the Visual Analytics. The original text was published by my master student Mr. Salkic who implemented the final prototype implementation of the dashboard and presented it in Salkic et al. (2015). My contribution was focused on the concept, use case and semantic data modeling as well as on co-authorship of the publication. The latest screenshots from the dashboard originate from Softic et al. (2016).

8.2.2 What is This Section About

This section introduces necessary steps and actions for implementation of analytic application with purpose on analyzing and visualizing information gathered by tracking user (learner) behavior and actions in educational system called Personal Learning Environment (PLE)[1] (for details refer section 2.5.4). PLE at Graz University of Technology has been running for several

[1] http://ple.tugraz.at, last access: 2017–05–29

years (2011-until today) and the need for getting deeper insight into the usage and behavior of system as well as for understanding the users and their needs has emerged over the time. All steps as well as requirements of our analytic application from pre-processing and cleaning the gathered log data to the final stage where the results are revealed, are described in detail. Furthermore, this text presents a novel Semantic Web driven approach, for modeling of learning and activity based context using eligible domain specific ontologies (elaborated in Softic et al. (2013b,c, 2014a, 2015a) and described in sections 7.1 and 8.1), as well as for retrieving modeled data depending on the value of interests demonstrated by learner himself. The intention of this work lies on closing the learning analytic cycle introduced by Clow (2012) for PLE, and for that purpose the requirements and implementation steps of Visual Analytic Dashboard have been defined which shall give us necessary knowledge for improvement.

8.2.3 Motivation for Semantically Driven Analytics Dashboard

Each learning process has only one goal, to fulfill the needs of learners. To reach that goal, information about the learners, their behavior and their actions, is needed. Additionally appropriate analysis and interpretation of that information makes it necessity because that is the only way of understanding the learners. With that information the learning process can be improved to fulfill learner's needs. After launching the Personal Learning Environment (PLE) at TU Graz University of Technology in 2010 such need was emerged. The tracking module was first step towards improving the learning process which was introduced already in 2011 through Taraghi et al. (2011). The major goal was to track or record user behavior, widget usage and user activities. Since we already have the data, next logical and most important step is processing and interpreting captured data into distinctive form suitable for analytics. In this step the semantic modeling of the data from tracking module shall be used in order to gain meaningful information. Semantically modeled information shall be visualized by Visual Learning Analytic Dashboard. From results of analytic step, new actions or interventions for improvement of learning process in PLE shall

be derived. The goal of visual analytic approach is to gain deeper insight in usage of PLE based on tracking of learners, to reveal hidden implicit information and to use it for improvement of PLE. The dashboard as analysis toll should enable predictive learner support and widget recommendation, where widget can be considered as single learning object in a virtual learning environment.

8.2.4 Extension of the Idea of Linked Data Driven Learning Analytics

Sections 7.1 and 8.1 reported so far on modeling learner behavior from the semi-structured text fragments in form of user (learner) system logs. Further the idea of mining and visualizing them was elaborated within previous sections as well. The contribution in this part of the thesis extends the proposed idea with complete visualization and data mining pipeline and final version of a Linked Data driven Learning Analytics dashboard as tool for visually supported decision making also known as Visual Analytics. For detailed related work on this discipline and previous related fields please refer sections 2.5.4, 2.9, 2.9.1 and 2.4.

8.2.5 Methodology

Methodology section in this part of the work describes all steps of the realization of visualization analytic dashboard. The description starts with description of mining pipeline followed by description of semantic data modeling and finally the close up deals with the description of analytic tasks based on SPARQL executable on such constructs.

8.2.6 Visualization Data Mining Pipeline

The visualization pipeline for gathering and preparing data for knowledge extraction and visualization can be seen on figure 8.6 .

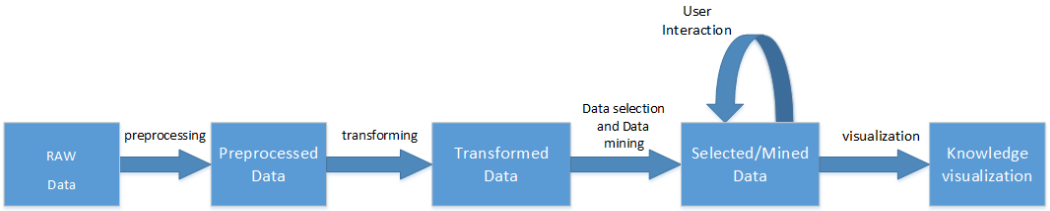


Figure 8.6: Visualization pipeline as published in Salkic et al. (2015).

In order to be able to perform any kind of analytics or data mining there must be some kind of data or information provided. Provided raw data will be used as input for application. Next step is pre-processing or cleaning the data from all unnecessary parts. This step is important because the raw data may contain outliers, redundant data or have missing values. Having clean preprocessed data improves the quality of results and reduces the possibility of getting a bias in results. To get the most of gathered data we need a flexible and scalable model to represent it. For that purpose in next step the data shall be transformed to semantic data model. The data model shall use eligible Ontology to describe the context in which it occurred. To reveal only the relevant data from the semantic model, as in experiments described in 7.1 and 8.1 the authors used SPARQL as data retrieval technology. And the final step is visualization of processed and selected data. The data visualization shall give deeper and easy understandable insight into existent user data from target Learning Management System. The visualization depends on value of interest of the user and only selected part will be visualized. The application shall answer to basic need of Visual Analytics which is interactive visualization. For that purpose each time when user changes the value of interest, application shall perform the data selection and data mining step again and visualize changed data accordingly. Example of value of interest would be "Top 5 widgets" or "Most active users" etc.

In elaborated use case the initial data are ordinary system logs including information which user when used a certain widget. This data is pre-processed using the tracking module into JSON[2] format by splitting the information about user, environment and event into separate blocks. As next step in mining pipeline is the transformation which requires data modeling.

[2] http://json.org/, last access: 2017–05–29

Figure 8.7: Impemented visualization pipeline as published in Salkic et al. (2015).

This step produces then instanced data which is dynamically processable in order to generate responsive visualizations needed for tracking of learners within PLE. Implemented pipeline is in the figure 8.7.

Data Modeling

The details on data modeling were already introduced in section 8.1.4 and do not differ in the final design of the dashboard implementation experiment.

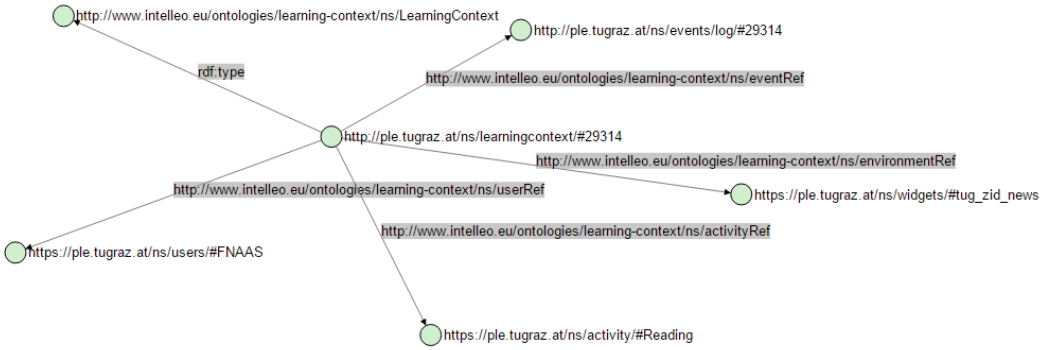


Figure 8.8: Sample instance of Learning Context displayed using Visual RDF - http://graves.cl/visualRDF. As published in Salkic et al. (2015).

Figure 8.8 is sample RDF instance of *lc:LearningContext*, produced by PLE learner tracking pipeline, which describes a log event denoted as instance of class *ao:Logging*, a subclass of an *ao:Event*, which occurred at certain time point inside the learning widget named **tug_zid_news** represented through class *ao:Enviroment* performed by user **#FNAAS** (*um:User*). The action which was registered is **Reading** represented through class *ao:Activity*.

Running Analytic Tasks

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
PREFIX lc: <http://www.intelleo.eu/ontologies/learning-context/ns/> .

SELECT DISTINCT ?activity ?ref (COUNT(?ref) as ?count)
{
 ?link rdf:type lc:LearningContext;
       lc:activityRef ?ref.

 ?ref rdf:label ?activity.
FILTER (?date > "2012-10-01T00:00:00"^^xsd:dateTime && ?date < "2015-12-31T00:00:00"^^xsd:dateTime)
}
GROUP BY ?ref
ORDERBY DESC (?count)
LIMIT 5
```

Listing 8.2: Querying the LearningContext for top 5 widgets. As in Salkic et al. (2015).

Usage logs data presented as *lc:LearningContext* Linked Data graph instances can be retrieved using SPARQL queries. In this way we are able to answer the questions like: "Show me the top 5 activities in PLE for a certain time period?". Listing 8.2 represents exactly this question stated in the manner of SPARQL syntax. The benefit of this query is visible for instance in figure 8.11 where the results of this query influence the widget arrangement in the widget store. Such direct impact on system with functional interoperability on machine level would not be possible without standards like SPARQL.

8.2.7 Final Results and Analysis

Presented approach allows us tracking of PLE learner activity trends over time periods, user wise, widget wise or like presented in figure 8.9 per performed activity. The improved final version of dashboard prototype in figure 8.9 shows the top activities in PLE for a certain time period which can be specified in the sidebar. The different charts on the right of the screen represent different views on the analytical data. Top 5 activities as pies or bar charts represent the visual answer of the query from listing 8.2. Alternative as general overview shown as line chart with zooming function in located in the lower part of dashboard. Same dashboards are also available for user centric and widget centric view. The overview over distribution of widget usage can reflect the overall interest of the users within PLE for different

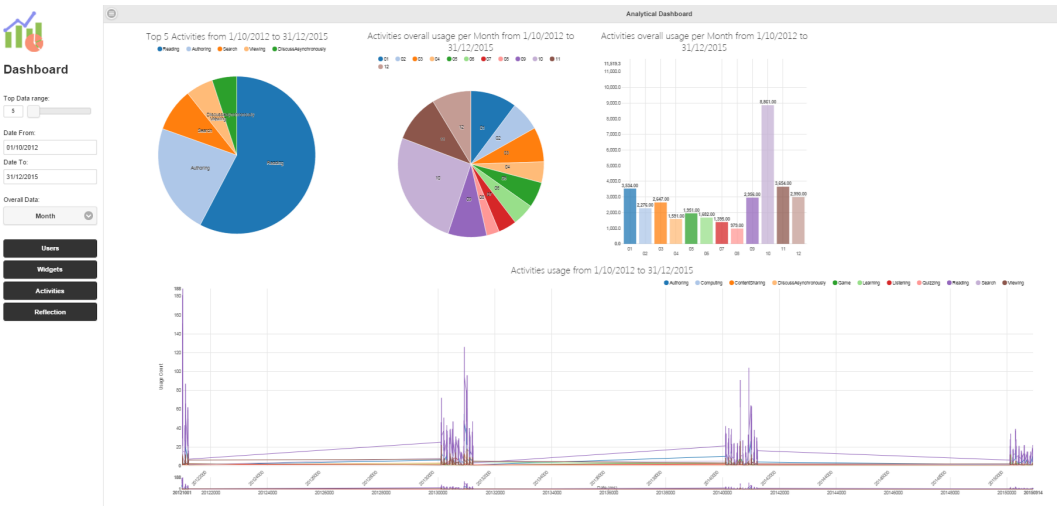


Figure 8.9: Visual Analytics Dashboard, top activities. As published in Salkic et al. (2015).

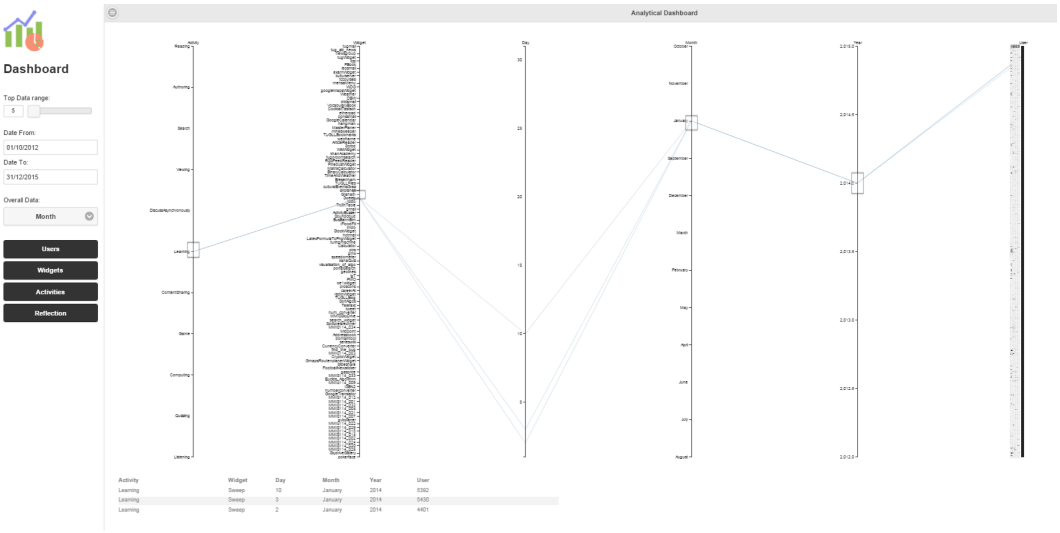


Figure 8.10: Visual Analytics Dashboard, reflection through parallel coordinates. As extension to published text in Salkic et al. (2015).

periods of time or based on type of actions they perform (see figure 8.9). Such outputs implicitly support the improvement of the quality of services for students and teachers. Figure 8.10 also shows the implemented reflection part of the semantic mining dashboard in form of parallel coordinates for the

case of activity of *Learning*. Each horizontal line represents one dimension and it is zoom-able. The dimensions are: time (as day, moth and year), activity, widget and user.

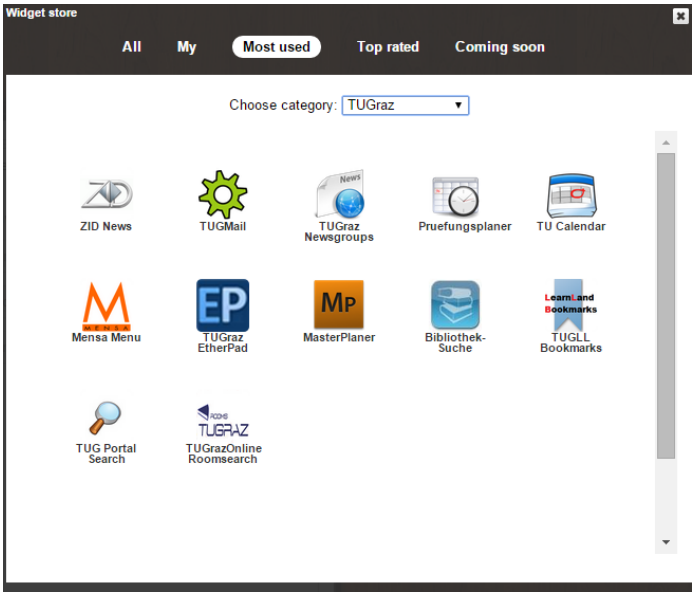


Figure 8.11: Widget store. As published in Salkic et al. (2015).

8.3 Concluding Remarks on Achievements

While remarkable prior initiative such as Learning Object Metadata (LOM) introduced by Hodgins et al. (2002) in 2002 aims to "facilitate search, evaluation, acquisition, and use of learning objects, for instance by learners or instructors or automated software processes", approaches presented here, although intersecting in some goals with intentions behind the LOM, focus merely on modeling user behavior in the learning environments for further practical analytic use and less on the standardized exchange, retrieval and description of learning artifacts as such. Visually supported, semi-automated and automated analytic tasks, relying on semantic models of user system logs, that reflect user (learner) activity, should primary serve

the monitoring and design improvements of the Online Learning System, also to recommendation of relevant learner artifacts and to overall reflection of the Online Learning System state. Mature generic interchange frame such as RDF and retrieval standard such as SPARQL are only means for the purpose and not the overall goal.

9 Scientific Implication of Achieved Results

With my thesis I made a contribution to Knowledge Discovery, with special focus on mining data from semi-structured text fragments for the purposes of research and education. As mining sources depending on the application area and specific use case author used tweets and user (learner) logs from a specific learning system. The contributions focused in case of research on Research 2.0 use case and Researcher Profiling and in the area of Education the scope of the contributions was on Visual (Learning) Analytics in online Online Learning Systems such as Personal Learning Environment (PLE) at Graz University of Technology. The targeted use cases as well as related contribution fields: Semantic Modeling, Data Profiling, Data Mining, Research 2.0, Technology Enhanced Learning, Visual (Learning) Analytics and Exploratory (Semantic) Search, have been discussed in chapters 3, 4, 5 for Research related application area and in chapters 6, 7 and 8. These chapters represent the summary of publications listed in 1.6 where to each paper are shortly stated my contribution, referred topics (contribution fields), referred research questions and chapters where the text from the papers were used. Additionally to this at the beginning of each of chapters: 3, 4, 5, 6, 7 and 8, there is a section named **"Statement to Own Contribution"** where in short summary form my own contribution is elaborated more detailed regarding the corresponding chapter and publication involved within.

9.1 Contribution to Research

Contribution of each single paper from the list presented in 1.6 to the application area of Research and Knowledge Discovery related fields related

to the corresponding use case of Research 2.0 can be seen in figure 9.1.

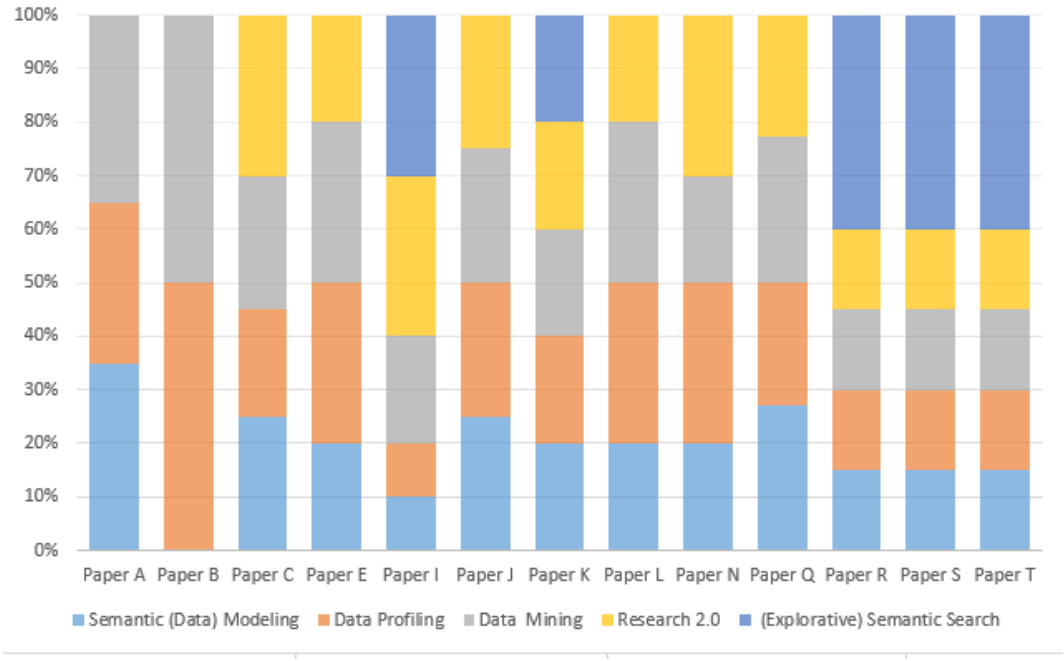


Figure 9.1: Contribution of each paper to each research area in research use case.

9.2 Contribution to Education

How and which single paper from the list presented in 1.6 contribute the application area of Education and Knowledge Discovery related fields related to the corresponding use case of Online Learning Systems (here Personal Learning Environment) is shown in figure 9.2.

9.3 Contribution to the Scientific Community

Concluding this chapter it is very significant to mention that work on this thesis resulted in co-supervision of three master thesis. Two of them by Mr.

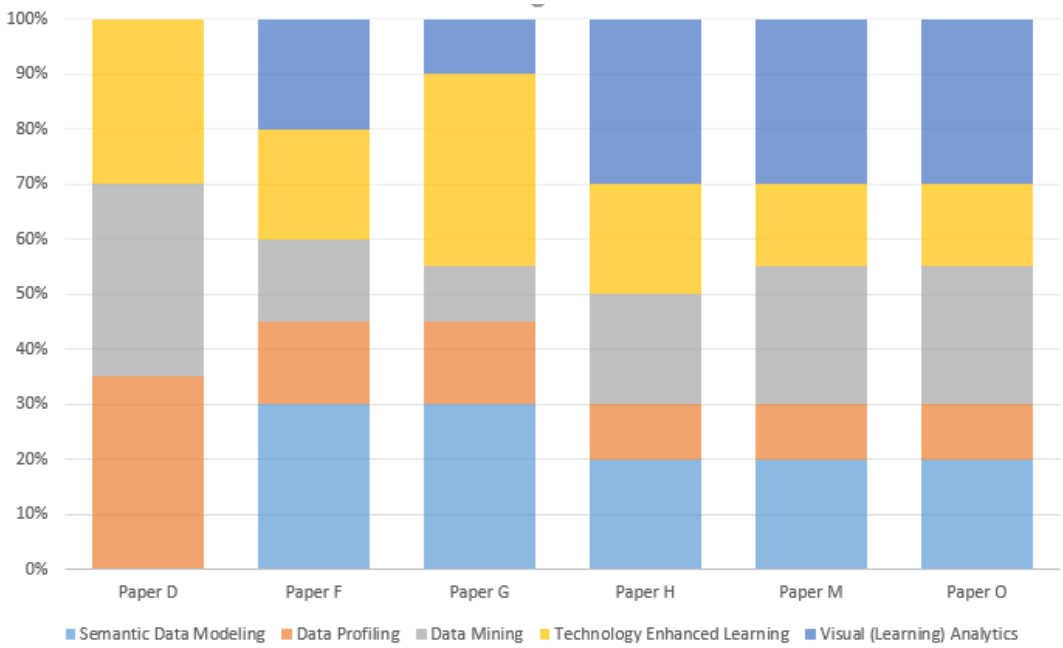


Figure 9.2: Contribution of each paper to each research area in education use case.

Patrick Thonhauser with master thesis entitled "Semantic Recommendation Systems in Research 2.0" in 2012 and Mr. Senaid Salkic with his thesis entitled "Linked Data Driven Visual Analytics for Tracking Learners in a PLE" in 2016 have been conducted at Graz University of Technology and the third one at the University of Leuven by Mr. Laurens De Vocht entitled "Researcher Profiling based on Semantic Analysis in Social Networks" in 2011. Also three higher educational institutions have been implicitly or explicitly involved. Beside the Graz University of Technology there were University of Leuven (Faculty of Engineering) and Ghent University in Belgium (IDLab- Internet Technology and Data Science Lab[1]). I published 20 papers (listed in 1.6) related to this work at different conferences and in journals and participated in program committees as member and reviewer of highly appreciated and ranked scientific events such as: BigScholar

[1] http://idlab.ugent.be, last access: 2017–05–29

2015[2], BigScholar 2016 [3], SemPub 2015[4], SemPub 2016[5] (my Linked Data set COLINDA[6] which is part of this thesis was used as reference data set for the challenge), and 4th European Immersive Education Summit[7]. Thanks to achieved results I also reviewed papers for the International Journal of Emerging Technologies in Learning (iJET)[8]. The Achievements implemented during the work on my thesis in collaboration with Ghent University were also shortlisted in scientific competitions such as Semantic Web Challenge 2013[9] at International Semantic Web Conference 2013 and LinkedUp Challenge at International Semantic Web Conference 2014[10]. Furthermore, all included publications are of high quality (peer or blind reviewed) and some of them are have been published in Springer Lecture Notes and in Web of Science indexed journals with impact factor or highly rated workshops at the World Wide Web Conference, European Semantic Web conference and International Semantic Web Conference among others. For sure, one of the tools that was strongly contributed through research work at this thesis was ResXplorer[11].

[2] http://thealphalab.org/bigscholar/2015/, last access: 2017–05–29
[3] http://thealphalab.org/bigscholar/2015/, last access: 2017–05–29
[4] https://github.com/ceurws/lod/wiki/SemPub2015, last access: 2017–05–29
[5] https://github.com/ceurws/lod/wiki/SemPub2016, last access: 2017–05–29
[6] http://colinda.org, last access: 2017–05–29
[7] http://immersiveeducation.org/EUROPE, last access: 2017–05–29
[8] http://online-journals.org/index.php/i-jet, last access: 2017–05–29
[9] http://challenge.semanticweb.org/2013/submissions/, last access: 2017–05–29
[10] http://linkedup-project.eu/2014/09/30/vici-the-shortlist/, last access: 2017–05–29
[11] http://resxplorer.org, last access: 2017–05–29

10 Conclusion, Limitations and Future Work

Concluding this thesis requires reflection on research questions in 1.3 and hypotheses in 1.4 as well on obtained results and insights. As shown in figure 10.1 the thesis started and was motivated by a real-world problem of Knowledge Discovery within sparse semi-structured text fragments for the purposes of Research and Education.

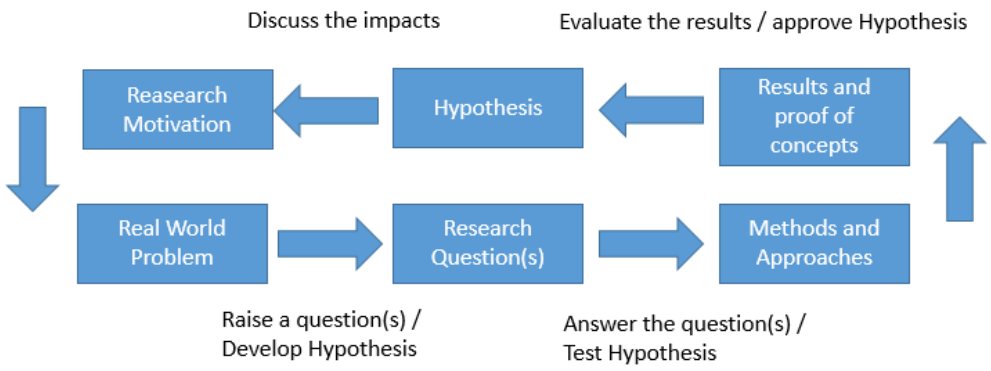


Figure 10.1: Research methodology cycle of this thesis.

Based upon standardized semantic technologies as RDF, OWL, SPARQL (see 2.6.1 for details), Linked Data as well on common Natural Language Processing and Data Mining techniques I elaborated in presented cumulative work the whole spectrum of Knowledge Discovery for two specific use cases, one each per application area. In case research application area that would be Research 2.0 and (Visual) Learning Analytics aiming at improving

Online Learning Systems in particular Personal Learning Environment at Graz University of Technology for education application area. All conducted experiments, implemented prototypes and methods and concepts developed have been tested, discussed and evaluated in mentioned scenarios regarding benefits for addressed fields. Very special benefits have been achieved in the area of research in Research 2.0 interfaces, Semantic Modeling and Profiling, also visually, of the researchers using the data from Twitter and Mendeley in combination with Linked Data Knowledge Bases and publication archives, and leaned on this insights, in the field of Exploratory Semantic Search through ResXplorer. Those achievements have been presented in chapters 3, 4 and 5. In the application area of education, significant achievements have been reached in the field of Semantic Modeling and Mining of Learner Behavior in the Personal Learning Environment of Graz University of Technology and in visual analysis with targeted toward Learning Analytics. Those achievements have been presented in chapters 6, 7 and 8. In following concluding sections each achievement has been discussed regarding the stated research questions and application area, use cases and addressed scientific fields. The sections also include concluding remarks on related research questions and an short outlook how the achieved insights might be used in future research.

10.1 Concluding Remarks, Limitations and Outlook to RQ1

The RQ1 was stated as:
Do sparse semi-structured text fragments (as tweets and user logs) contain information useful for better exploration of research related and learning resources?

According to results of experiments conducted within the research related to this thesis described in sections: 3.2 and 6.1 the potentials of sparse semi-structured text fragments (as tweets and user logs) bear a huge potential for further Data Mining and Data Profiling that can serve also higher level applications as mesh-ups or search systems. Of course the raw data from the addressed sparse semi-structured text fragments requires an initial

pre-processing that can be managed either through text mining techniques or tools or through data modeling strategies which include retrieval and clustering potentials. Specifically chosen data sources for the two application areas of research and applications showed already in feasibility experiments a high potential for discovery and mining new insights based upon their content.

10.2 Concluding Remarks, Limitations and Outlook to RQ2

The RQ2 was stated as:
How such content can be semantically modeled and explored with machines using semantic (web) technologies and Linked Data?
As "such content" hereby are meant sparse semi-structured text fragments as tweets or system logs from PLE at Graz University of Technology. Proposed method of Semantic Modeling of semi-structured text fragments as tweets or system logs enables very specific wide range of applications regarding the user generated context. Presented methods are as effective as usual retrieval methods in focused applications as demonstrated through e.g. PLE Visual Analytics Dashboard, Researcher Affinity Browser or in ResXplorer and its components. The strongest appliance of Semantic Modeling was represented through profiling the researchers on Twitter for the purposes of revealing the connection within the research networks (see sections 10.1, 4.1, 4.3, 5.1, 5.2 and 5.3) and improvement of specific exploratory semantic search tasks for researcher. Further appliance was in tracking and profiling learner behavior (see sections 7.1, 8.1 and 8.2). Experiments and evaluation on this topic have been in particular described in 4.3.6, 5.1.6, 5.2.8 and 5.3.5). For research application area (focus on researcher community and Research 2.0) and in 7.1.5, 8.1.5 and 8.2.7 education application area (focus on profiling and tracking learners and learn objects related activities). As essential contribution of my work I offered a novel concept of mining architecture based upon semantically modeled data for both application area and related use cases and appliance fields, and approved them through experimental implementations and related evaluations. Based upon ex-

perimental implementations and their evaluation we can draw conclusion that proposed approach of Semantic Modeling contributes to areas as User Profiling, Learning Analytics, Exploratory Semantic Search and Knowledge Bases suitable for Data Mining at least as good as existing solutions based upon other technologies. However degree of details used for specific experimental implementations are very high and specific what implies a very focused range of re-use without previous reconfigurations.

10.3 Concluding Remarks, Limitations and Outlook to RQ3

Finally the RQ3 was stated as:
How such modeled data can be used to profile researchers on Twitter and explore related resourced on the Web or in case of system logs for reflections and improvements of learning systems?
By "such modeled data" are primary meant semantically modeleds parse semi-structured text fragments as tweets or system logs from PLE at Graz University of Technology but also partly especially in research application area data from other Social Media platforms such as Mendeley. Following subsections discuss the three main beneficial fields impacted through the results achieved in presented thesis.

10.3.1 Researcher Profiling and Exploration

Semantic Modeling of researcher Twitter and other Social Media profiles in combination with reliable Linked Data source as e.g. COLINDA introduced in 4.2 became approved source for Data Mining and Profiling of similar researchers on the Web and related resources and events bound to them. This has been confirmed through recognition from research community where COLINDA was used as reference data set for mining tasks at diverse workshops and challenges and also through citations in scientific works published after introducing COLINDA. The efficiency within experiments reaches a high level comparable to advanced conventional state of the

art search systems with a very high precision of results. This findings are specifically supported by results from exploitation examples described in 4.3.6, 5.1.6, 5.2.8 and 5.3.5. Especially results as "Researcher Affinnity Browser" interface and evaluation conducted related to it presented in section 5.1 demonstrates very practically these findings. The same was also confirmed further on the sample of visualization component of ResXplorer presented in 5.3. Very specific limitation to achievements reached through the conducted experiments is that increasing the performance of such profiling infrastructures requires high level on details what makes such approaches very tightly bound to the specific use cases. However their practical usefulness has been approved through users on examples on of "Researcher Affinity Browser" and "ResXplorer".

10.3.2 Exploratory Semantic Search for Researchers

Although implemented experimental mining and profiling infrastructures targeted specific use case their contribution through involvement of social data from researchers that was semantically modeled and interlinked with reliable Linked Data sources into knowledge bases that enhanced search as additional mining reference has been enormous. This is in particular confirmed through ResXplorer[1] with its components as exploratory search interface, research knowledge base and researcher network visualization. The practical benefit for researcher was verified by a series of experiments regarding usability, usefulness described in 5.1.6, 5.3.5 and 5.2.7 and retrieval quality described in 5.4.5, and 5.5.6. Since meanwhile there are several solutions for search of research related resources, potential future efforts that may arise from the presented work could consider for instance creation of an initiative for definition of common exchange data model of research related search entities containing information about persons, locations, events and publications in order to enable better exchange of search results between the existing solutions and to foster creation of uniform views of search results delivered by them. Also a commonly built Linked Data knowledge base as foundation for more sophisticated search approaches in this field would be a thinkable option in the same direction. ResXplorer related future

[1] http://resxplorer.org, last access: 2017–05–29

work could be pointed towards use of history based, critique-based, and conversational recommendations. They could be created through tracking of search session of the users. The idea would be to use their experience to enhance and improve the adding of top related resources in the exploratory search.

10.3.3 Visually Supported (Learning) Analytics

Beside presented practical benefits from using Semantic Modeling and Linked Data in Learning Analytics demonstrated on the example of PLE at Graz University of Technology (see sections 7.1.5, 8.1.5 and 8.2.7) it is important to pinpoint how such approach differs from conventional methods and to which extent extends them. While conventional Learning Analytics focus more on monitoring and drawing conclusions from analyzed and derived results Linked Data driven Learning Analytics approach delivers and derives the results on demand and in in-time. This is made possible through automated conversion of events happening in PLE represented as system logs into semantically modeled instances which can be then queried with single query through SPARQL and retrieved in commonly spread exchange formats as XML, JSON or comma or tab separated values. This option offers technical extensibility and claims interoperability by default, opening the interfaces toward other web platforms, sources and internet technologies. Flexibility through SPARQL standard for interactive querying allows the dynamic generation of inputs for hosting platforms, included visualizations and analytic dashboards. Such actions require data models with certain degree of expressiveness, and well-thought-out constrains. Main challenge lies in choice or construction of proper model, as well as in the decision about the granularity degree of chosen model. Sometimes, this process is limited by the quality and variety of provided data. Very important advantage of such models is their adaptability to extensions, reductions and changes of model schema. The nature of RDF, RDFS and OWL allows also to inference based on logical rules. This is especially useful for asking sophisticated questions about the context of modeled data. Leveraging Learning Analytics with Linked Data supports standardized interfaces for information exchange, offer flexibility for visual other kinds

of analytics, and also can enrich the learning system's data with Linked Data sources from the Web. The spread of applicability covers wide range of analytic methodologies like prediction, reflection and as outcome of these the intervention field. Presented work based on case study of data from a PLE outlines the contribution of Semantic Technology Stack and Linked Data to Learning Analytics. The idea including Semantic Modeling of usage data in form of Linked Data is promising and delivers great results with very low effort because of standardized approach for data description, representation and retrieval, what makes it especially valuable for analytical tasks targeted on improvement of learning environments like PLE.

Bibliography

Abney, S. P. (1992). *Principle-Based Parsing: Computation and Psycholinguistics*, chapter Parsing By Chunks, pages 257–278. Springer Netherlands, Dordrecht. 80

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749. 60

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2008). DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer. 37, 110

Bacchelli, A., Cleve, A., Lanza, M., and Mocci, A. (2011). Extracting structured data from natural language documents with island parsing. In *26th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 476–479. 2

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7:112–118. 16

Bakharia, A. and Dawson, S. (2011). Snapp: a bird's-eye view of temporal participant interaction. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 168–173, New York, NY, USA. ACM. 49

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM 11, pages 65–74, New York, NY, USA. ACM. 21

Beck, J. and Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In Woolf, B., Aïmeur, E., Nkambou, R., and Lajoie, S., editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 353–362. Springer Berlin Heidelberg. 17

Bell, R. M. and Koren, Y. (2007). Lessons from the netflix prize challenge. *SIGKDD Explorations Newsletter*, 9(2):75–79. 53, 59

Berners-Lee (1996). WWW: Past, present, and future. *COMPUTER: IEEE Computer*, 29. 24

Berners-Lee, T. (2006). Linked data. 37, 38, 93, 110

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43. 24

Berrueta, D., Brickley, D., Decker, S., Fernández, S., Görn, C., Harth, A., Heath, T., Idehen, K., Kjernsmo, K., Miles, A., Passant, A., Polleres, A., Polo, L., and Sintek, M. (2007). Sioc core ontology specification. W3c member submission, W3C. 34

Bizer, C. and Cyganiak, R. (2006). D2r server - publishing relational databases on the semantic web. Poster at the 5th International Semantic Web Conference (ISWC2006). 38, 39

Bizer, C., Cyganiak, R., and Heath, T. (2007). How to publish linked data on the web. Web page. Revised 2016. Accessed 10/08/2016. 110, 113

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22. 38

Bizer, C., Heath, T., Idehen, K., and Berners-Lee, T. (2008). Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1265–1266, New York, NY, USA. ACM. 38

Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. In *Proceedings of the 1st International*

Conference on Learning Analytics and Knowledge, LAK '11, pages 110–116, New York, NY, USA. ACM. 49

Bojars, U., Passant, A., Breslin, J. G., and Decker, S. (2008a). Data portability with sioc and foaf. In *XTech 2008*, Dublin, Ireland. 35, 112, 113

Bojars, U., Passant, A., Cyganiak, R., and Breslin, J. (2008b). Weaving SIOC into the Web of Linked Data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*. 35

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences*, volume 0, pages 1–10, Los Alamitos, CA, USA. IEEE. 20, 97, 113

Breslin, J. G., Decker, S., Harth, A., and Bojars, U. (2006a). Sioc: an approach to connect web-based communities. *International Journal of Web Based Communities (IJWBC)*, 2(2):133–142. 47, 48, 97

Breslin, J. G., Decker, S., Harth, A., and Bojars, U. (2006b). Sioc: an approach to connect web-based communities. *International Journal of Web Based Communities (IJWBC)*, 2(2):133–142. 132

Breslin, J. G., Harth, A., Bojars, U., and Decker, S. (2005). Towards semantically-interlinked online communities. In Gomez-Perez, A. and Euzenat, J., editors, *European Semantic Web Conference (ESWC)*, volume 3532 of *Lecture Notes on Computer Science*, pages 500–514. Springer. 47, 48, 97, 112

Brickley, D. and Miller, L. (2004). FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project. http://xmlns.com/foaf/0.1/. 34, 112

Brusilovsky, P. (2003). Adaptive navigation support in educational hypermedia: the role of student knowledge level and the case for meta-adaptation. *British Journal of Educational Technology*, 34(4):487–497. 165

Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):261–272. 43

Chen, J., Huang, H., Tian, S., and Qu, Y. (2009). Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, 36(3):5432–5435. 58

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., and Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670. ACM. 52

Chen, Z., Jiang, Y., and Zhao, Y. (2010). A collaborative filtering recommendation algorithm based on user interest change and trust evaluation. 52

Choudhury, S. and Breslin, J. G. (2011). Extracting semantic entities and events from sports tweets. In *In Proceedings of the 1st Workshop on Making Sense of Microposts (MSM2011) at ESWC*, pages 22–32. 46, 48, 77

Chu-Carroll, J., Prager, J. M., Czuba, K., Ferrucci, D. A., and Duboué, P. A. (2006). Semantic search via xml fragments: a high-precision approach to ir. In Efthimiadis, E. N., Dumais, S. T., Hawking, D., and Järvelin, K., editors, *SIGIR*, pages 445–452. ACM. 43

Clow, D. (2012). The learning analytics cycle: closing the loop effectively. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 134–138, New York, NY, USA. ACM. 217

De Pessemier, T., Deryckere, T., and Martens, L. (2009). Context aware recommendations for user-generated content on a social network site. In *7th European Conference on Interactive Television*, pages 133–136. Association for Computing Machinery (ACM). 60

De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., and Van de Walle, R. (2013a). Discovering meaningful connections between resources in the web of data. In *Proceedings of the 6th Workshop on Linked Data on the Web (LDOW2013)*. 150, 151, 163, 166

De Vocht, L., Dimou, A., Breuer, J., Van Compernolle, M., Verborgh, R., Mannens, E., Mechant, P., and Van de Walle, R. (2014a). A visual exploration workflow as enabler for the exploitation of linked open data. *Proceedings of the 3rd Workshop Intelligent Exploration of Semantic Data*. 144

De Vocht, L., Softic, S., Dimou, A., Verborgh, R., Mannens, E., Ebner, M., and Van de Walle, R. (2015). Visualizing collaborations and online social interactions at scientific conferences for scholarly networking. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 1053–1054. xix, 18, 142, 145, 148, 153, 168

De Vocht, L., Softic, S., Ebner, M., and Mühlburger, H. (2011). Semantically driven social data aggregation interfaces for research 2.0. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, pages 43:1–43:9, New York, NY, USA. ACM. xviii, 18, 38, 47, 68, 83, 95, 100, 109, 112, 115, 117, 118, 119, 120, 122, 123, 128, 129, 132, 133, 140, 141, 142, 144, 145, 146, 150, 153, 158, 160, 162

De Vocht, L., Softic, S., Mannens, E., , Ebner, M., and Van de Walle, R. (2013b). A search interface for researchers to explore affinities in a linked data knowledge base. In *Proceedings of the 12th International Semantic Web Conference (ISWC 2013), Posters & Demonstrations Track, Sydney, Australia, October 23, 2013*, pages 21–24. 18, 142, 150, 151, 158, 160

De Vocht, L., Softic, S., Mannens, E., Ebner, M., and Van de Walle, R. (2014b). Aligning web collaboration tools with research data for scholars. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 1203–1208, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. xviii, xix, 18, 47, 128, 129, 132, 134, 135, 136, 137, 138, 139, 142, 144, 145, 146, 153, 158, 160

De Vocht, L., Softic, S., Mannens, E., Van de Walle, R., and Ebner, M. (2013c). Resxplorer: interactive search for relationships in research repositories. In *Semantic Web Challenge 2013*. 18, 142, 150, 151, 158, 160

De Vocht, L., Softic, S., Verborgh, R., Mannens, E., and Ebner, M. (2016). Resxplorer: Revealing relations between resources for researchers in the web of data. *Computer Science and Information Systems*, 14(1):25–50. xix, 18, 158, 160, 164, 165, 166, 167

De Vocht, L., Softic, S., Verborgh, R., Mannens, E., Ebner, M., and Van de Walle, R. (2017). Social semantic search: A case study on Web 2.0 for

science. *International Journal On Semantic Web and Information Systems*, 13(3). 18, 158

De Vocht, L., Van Deursen, D., Mannens, E., and Van de Walle, R. (2012). A semantic approach to cross-disciplinary research collaboration. *iJET*, 7(S2):22–30. 128, 129, 142, 144, 149, 150, 153, 158, 162

Dietze, S., Sanchez-Alonso, S., Ebner, H., Qing Yu, H., Giordano, D., Marenzi, I., and Pereira Nunes, B. (2013). Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program*, 47(1):60–91. 43

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., and Taibi, D. (2012). Linked education: Interlinking educational resources and the web of data. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 366–371, New York, NY, USA. ACM. 43

Dimou, A., De Vocht, L., Van Compernolle, M., Mannens, E., Mechant, P., and Van de Walle, R. (2014a). A visual workflow to explore the web of data for scholars. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 1171–1176, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. 168

Dimou, A., Sande, M. V., Colpaert, P., Verborgh, R., Mannens, E., and de Walle, R. V. (2014b). RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014*. 33

Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., Lindstaedt, S., Stern, H., Friedrich, M., and Wolpers, M. (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1(2):2849 – 2858. Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010). 49, 53, 192

Ebner, M., Altmann, T., and Softic, S. (2011). @twitter analysis of #edmedia10 - is the #informationstream usable for the #mass. *Form@re - Open Journal per la formazione in rete*, 11(74). 21, 48, 67, 68, 104, 140

Ebner, M., Holzinger, A., and Maurer, H. A. (2007). Web 2.0 technology: Future interfaces for technology enhanced learning? In *Lecture Notes in Computer Science*, volume 4556, pages 559–568. Springer. 190

Ebner, M., Muehlburger, H., Schaffert, S., Schiefner, M., Reinhardt, W., and Wheeler, S. (2010a). Getting granular on twitter: Tweets from a conference and their limited usefulness for non-participants. In Reynolds, N. and Turcsányi-Szabó, M., editors, *Key Competencies in the Knowledge Society*, volume 324 of *IFIP Advances in Information and Communication Technology*, pages 102–113. Springer Berlin Heidelberg. 21, 68, 97, 111

Ebner, M. and Reinhardt, W. (2009). Social networking in scientific conferences - Twitter as tool for strengthen a scientific community. In Cress, U., Dimitrova, V., and Specht, M., editors, *Learning in the Synergy of Multiple Disciplines, Proceedings of the EC-TEL 2009*, volume 5794 of *Lecture Notes in Computer Science*, Berlin/Heidelberg. Springer. 159

Ebner, M., Scerbakov, N., Taraghi, B., Nagler, W., and Kamrat, I. (2010b). Teaching and learning in higher education an integral approach. In Gibson, D. and Dodge, B., editors, *Proceedings of Society for Information Technology & Teacher Education International Conference 2010*, pages 428–436, San Diego, CA, USA. AACE. 23

Ebner, M. and Taraghi, B. (2010). Personal learning environment for higher education–a first prototype. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 1158–1166. 23

Fancsali, S. E. (2011). Variable construction for predictive and causal modeling of online education data. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 54–63, New York, NY, USA. ACM. 49

Faulkner, L. (2003). Beyond the Five-User assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3). 168

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54. xvii, 15, 16

Frank, E. and Bouckaert, R. R. (2006). Naive bayes for text classification with unbalanced classes. In *Knowledge Discovery in Databases: PKDD 2006*, pages 503–510. Springer. 58

Fung, G. (2001). A comprehensive overview of basic clustering algorithms. 65

Gemmell, J., Ramezani, M., Schimoler, T., Christiansen, L., and Mobasher, B. (2009). The impact of ambiguity and redundancy on tag recommendation in folksonomies. In *Proceedings of the third ACM conference on Recommender systems*, pages 45–52. ACM. 57

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011a). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics. 61

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011b). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics. 77

Glaser, H., Millard, I. C., and Jaffri, A. (2008). Rkbexplorer.com: a knowledge driven infrastructure for linked data providers. In *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, ESWC'08, pages 797–801, Berlin, Heidelberg. Springer-Verlag. 41

Goldberg, D., Nichols, D., Oki, B., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70. 50

Goldwater, S. and Griffiths, T. L. (2005). A fully bayesian approach to unsupervised part-of-speech tagging. 62, 77

Gong, S. (2010). A collaborative filtering recommendation algorithm based on user clustering and item clustering. *Journal of Software*, 5(7):745–752. 52

Graves, A. (2013). Creation of visualizations based on Linked Data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 41. ACM. 167

Guha, R. V., McCool, R., and Miller, E. (2003). Semantic search. In *WWW*, pages 700–709. 43

Hodgins, W., Duval, E., et al. (2002). Draft standard for learning object metadata. *IEEE*, 1484:1–2002. 223

Holzinger, A., Searle, G., and Wernbacher, M. (2011). The effect of previous exposure to technology on acceptance and its importance in usability and accessibility engineering. *Universal Access in the Information Society*, 10:245–260. 190

Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. *Hawaii International Conference on System Sciences*, 0:1–10. 113

Horn, C. (2010). Analysis and classification of twitter messages. 58, 76

Horn, C., Lex, E., and Granitzer, M. (2010). Who tweets: Detecting user types and tweet quality using supervised classification. 46, 58

Horrocks, I., Parsia, B., Patel-Schneider, P., and Hendler, J. (2005). Semantic web architecture: Stack or two towers? pages 37–41. xvii, 24, 25

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298. 170

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University. 63, 76

Huang, A. (2008). Similarity measures for text document clustering. pages 49–56. 178

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2011). *Recommender Systems An Introduction*. Cambridge University Press. 53, 54, 56, 59

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188. 46, 68, 110, 175

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM. 20, 68, 97, 110, 113

Jeremić, Z., Jovanović, J., and Gašević, D. (2012). Personal learning environments on the social semantic web. *Semantic Web*. 49

Keim, D. A., Oelke, D., Bak, P., Spretke, D., Bertini, E., and Ziegler, H. (2010). Advanced visual analytics interfaces. *Advanced Visual Interfaces 2010*. xvii, 18, 19

Khan, M. S., Ebner, M., and Maurer, H. (2009). Trends discovery in the field of e-learning with visualization. In Siemens, G. and Fulford, C., editors, *Proceedings of EdMedia: World Conference on Educational Media and Technology 2009*, pages 4408–4413, Honolulu, HI, USA. Association for the Advancement of Computing in Education (AACE). 71

Kirchberg, M., Ko, R. K. L., and Lee, B.-S. (2011). From linked data to relevant data – time is the essence. *CoRR*, abs/1103.5046. 50

Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., and Goranov, M. (2003). Semantic annotation, indexing, and retrieval. In Fensel, D., Sycara, K. P., and Mylopoulos, J., editors, *International Semantic Web Conference*, volume 2870 of *Lecture Notes in Computer Science*, pages 484–499. Springer. 43

Knoblock, C. A., Szekely, P. A., Ambite, J. L., Gupta, S., Goel, A., Muslea, M., Lerman, K., and Mallick, P. (2011). Interactively mapping data sources into the semantic web. In Kauppinen, T., Pouchard, L. C., and Keßler, C., editors, *LISC*, volume 783 of *CEUR Workshop Proceedings*. CEUR-WS.org. 38

Koren, Y. (2009). The bellkor solution to the netflix grand prize. 53, 59

Kraaij, W. and Post, W. (2006). Task based evaluation of exploratory search systems. In *SIGIR 2006 workshop, Evaluating Exploratory Search Systems*. 168, 169

Kraak, M.-J. (2008). Exploratory visualization. In *Encyclopedia of GIS*, pages 301–307. Springer. 144

Kraker, P., Schlögl, C., Jack, K., and Lindstaedt, S. N. (2014). Visualization of co-readership patterns from an online reference management system. *CoRR*, abs/1409.0348. 18

Kraker, P., Wagner, C., Jeanquartier, F., and Lindstaedt, S. N. (2011). On the way to a science intelligence: Visualizing tel tweets for trend detection. In Kloos, C. D., Gillet, D., García, R. M. C., Wild, F., and Wolpers, M., editors, *EC-TEL*, volume 6964 of *Lecture Notes in Computer Science*, pages 220–232. Springer. 45, 73

Kudo, T. and Matsumoto, Y. (2001). Chunking with support vector machines. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics. 63

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174. 170

Laniado, D. and Mika, P. (2010). Making sense of twitter. In Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., and Glimm, B., editors, *International Semantic Web Conference (1)*, volume 6496 of *Lecture Notes in Computer Science*, pages 470–485. Springer. 21, 45, 48, 56, 140, 144

Letierce, J., Passant, A., Breslin, J., and Decker, S. (2010a). Understanding how twitter is used to widely spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. 21, 68, 69, 113, 140

Letierce, J., Passant, A., Breslin, J., and Decker, S. (2010b). Understanding how twitter is used to widely spread scientific messages. 97

Ley, M. (2002). The DBLP computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval*, pages 1–10. Springer. 162

Li, H., Bhowmick, S. S., and Sun, A. (2011). Affrank: Affinity-driven ranking of products in online social rating networks. *Journal of the Association for Information Science and Technology*, 62(7):1345–1359. 125

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations. item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80. 51, 53

Liskov, B. (1992). Report on workshop on research in experimental computer science held in palo alto, california on 16-18 october 1991. 5

Lockyer, L. and Dawson, S. (2011). Learning designs and learning analytics. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 153–156, New York, NY, USA. ACM. 49

Lohmann, S., Díaz, P., and Aedo, I. (2011). Muto: the modular unified tagging ontology. In Ghidini, C., Ngomo, A.-C. N., Lindstaedt, S. N., and Pellegrini, T., editors, *I-SEMANTICS*, ACM International Conference Proceeding Series, pages 95–104. ACM. 47, 48, 98, 132

Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 1155–1158, New York, NY, USA. ACM. 21, 69

Mazza, R. and Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Workshop on Usage Analysis in Learning Systems, Proceedings of Artificial Intelligence in Education*, Amsterdam. 49, 191, 211

McFedries, P. (2007). Technically speaking: All a-twitter. In *Spectrum, IEEE*, volume 44, pages 84–84. 20, 97

Mejas, U. (2005). A nomad's guide to learning and social software. *The Knowledge Tree*. Edition 7. 111, 174

Milikic, N., Jovanovic, J., and Stankovic, M. (2011). Discovering the dynamics of terms' semantic relatedness through twitter. In Rowe, M., Stankovic, M., Dadzie, A.-S., and Hardey, M., editors, *MSM*, volume 718 of *CEUR Workshop Proceedings*, pages 57–68. CEUR-WS.org. 45

Miller, L. and Brickley, D. (2010). FOAF Vocabulary Specification. 47, 48, 97

Mödritscher, F. (2010). Towards a recommender strategy for personal learning environments. *Procedia Computer Science*, 1(2):2775–2782. 51

Mühlburger, H., Ebner, M., and Taraghi, B. (2010). twitter try out #grabeeter to export, archive and search your tweets. *Research 2.0 approaches to TEL (2010)*. 21, 112

Mulwad, V., Finin, T., Syed, Z., and Joshi, A. (2010). Using linked data to interpret tables. In *Proceedings of the the First International Workshop on Consuming Linked Data*. 49

Nakagawa, T., Kudoh, T., and Matsumoto, Y. (2001). Unknown word guessing and part-of-speech tagging using support vector machines. In *In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 325–331. 63, 76, 77

Niemann, K., Schmitz, H.-C., Scheffel, M., and Wolpers, M. (2011). Usage contexts for object similarity: exploratory investigations. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 81–85, New York, NY, USA. ACM. 49

Ochoa, X., Méndez, G., and Duval, E. (2009). Who we are: Analysis of 10 years of the ed-media conference. In Siemens, G. and Fulford, C., editors, *Proceedings of EdMedia: World Conference on Educational Media and Technology 2009*, pages 189–200, Honolulu, HI, USA. Association for the Advancement of Computing in Education (AACE). 71

Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., and Tummarello, G. (2008). Sindice.com: a document-oriented lookup index for open linked data. *Int. J. of Metadata and Semantics and Ontologies*, 3:37–52. 113

Pardo, A. and Kloos, C. D. (2011). Stepping out of the box: towards analytics outside the learning management system. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 163–167, New York, NY, USA. ACM. 49, 191, 211

Parra Chico, G. and Duval, E. (2010). Filling the gaps to know More! about a researcher. In *Proceedings of the 2nd International Workshop on Research 2.0. At the 5th European Conference on Technology Enhanced Learning: Sustaining TEL,*, pages 18–22. CEUR-WS. 17

Passant, A., Bojars, U., Breslin, J. G., and Decker, S. (2009). The sioc project: Semantically-interlinked online communities, from humans to machines. In Padget, J. A., Artikis, A., Vasconcelos, W., Stathis, K., da Silva, V. T., Matson, E. T., and Polleres, A., editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems V*, volume 6069 of *Lecture Notes in Computer Science*, pages 179–194. Springer. 47, 48

Passant, A., Bojars, U., Breslin, J. G., Hastrup, T., Stankovic, M., and Laublet, P. (2010a). An overview of smob 2: Open, semantic and distributed microblogging. In Cohen, W. W. and Gosling, S., editors, *ICWSM*, pages 303–306. The AAAI Press. 46, 47, 48

Passant, A., Breslin, J. G., and Decker, S. (2010b). Open, distributed and semantic microblogging with smob. In Benatallah, B., Casati, F., Kappel, G., and Rossi, G., editors, *ICWE*, volume 6189 of *Lecture Notes in Computer Science*, pages 494–497. Springer. 46, 47, 48, 132, 141

Passant, A., Hastrup, T., Bojars, U., and Breslin, J. (2008). Microblogging: A semantic and distributed approach. In Bizer, C., Auer, S., Grimnes, G. A., and Heath, T., editors, *Proceedings of the 4th Workshop on Scripting for the Semantic Web*, volume 368 of *CEUR Workshop Proceedings*. 47, 100

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer. 56

Pintado, X. (1995). *Object-oriented software composition*, chapter The affinity browser, pages 245–272. Prentice Hall. 125, 152

Pohl, M., Holzinger, A., Rester, M., Motschnig, R., Ebner, M., and Leitner, G. (2008). Gestaltung von innovativen technologiegestützten lernsystemen

am beispiel von web 2.0-anwendungen. eine herausforderung für hci. *OCG Journal*, (33):20–23. 190

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63. 133, 170

Priem, J. and Costello, K. L. (2010). How and why scholars cite on twitter. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4. 97

Prinsloo, P., Slade, S., and Galpin, F. (2012). Learning analytics: challenges, paradoxes and opportunities for mega open distance learning institutions. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 130–133, New York, NY, USA. ACM. 49, 211

Rajaraman, A., Leskovec, J., and Ullman, J. D. (2014). *Mining Massive Datasets*. 51

Reinhardt, W., Ebner, M., Beham, G., and Costa, C. (2009a). How people are using twitter during conferences. In *Hornung-Prähauser, V., Luckmann, M.(Hg.): 5th EduMedia conference, Salzburg*, pages 145–156. 21, 68, 87, 111, 113, 140

Reinhardt, W., Ebner, M., Beham, G., and Costa, C. (2009b). How people are using twitter during conferences. In *Hornung-Prähauser, V., Luckmann, M.(Hg.): 5th EduMedia conference, Salzburg*, pages 145–156. 68, 97, 110, 191

Reinhardt, W., Mletzko, C., Drachsler, H., and Sloep, P. (2011). Awesome: A widget-based dashboard for awareness-support in research networks. PLE Conference;. 191

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pages 175–186, New York, NY, USA. ACM. 50

Richards, G. and DeVries, I. (2011). Revisiting formative evaluation: dynamic monitoring for the improvement of learning activity design and delivery.

In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 157–162, New York, NY, USA. ACM. 49

Rios, G. and Zha, H. (2004). Exploring support vector machines and random forests for spam detection. In *CEAS*. 76

Ritter, A., Clark, S., Etzioni, O., et al. (2011a). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics. 61

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011b). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics. 76

Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011c). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics. 80

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146. 16, 53

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6):601–618. 16, 53

Rosen, D., Miagkikh, V., and Suthers, D. (2011). Social and semantic network analysis of chat logs. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 134–139, New York, NY, USA. ACM. 49, 192

Rowe, M. (2009). Interlinking distributed social graphs. In Bizer, C., Heath, T., Berners-Lee, T., and Idehen, K., editors, *Proceedings of the 6th International Workshop on Linked Data on the Web*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org. 46, 141

Rowe, M. and Stankovic, M. (2010). Mapping tweets to conference talks: A goldmine for semantics. In *Social Data on the Web Workshop, International Semantic Web Conference*. 47, 68, 110

Russell, M. A. (2011). *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. O'Reilly Media, 1 edition. 61

Salkic, S., Softic, S., Taraghi, B., and Ebner, M. (2015). Linked data driven visual analytics for tracking learners in a PLE. In *DeLFI 2015 - Die 13. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. (GI), München, 1.-4. September 2015*, pages 329–331. xxi, 216, 219, 220, 221, 222, 223

Santos, J. L., Govaerts, S., Verbert, K., and Duval, E. (2012). Goal-oriented visualizations of activity tracking: a case study with engineering students. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 143–152, New York, NY, USA. ACM. 49

Santos, J. L., Verbert, K., Govaerts, S., and Duval, E. (2011). Visualizing ple usage. In *Proceedings of EFEPLE11 1st Workshop on Exploring the Fitness and Evolvability of Personal Learning Environments*. CEUR workshop proceedings. 211

Santos Odriozola, J. L., Verbert, K., Govaerts, S., and Duval, E. (2011). Visualizing PLE usage. In *Proceedings of EFEPLE11 1st Workshop on Exploring the Fitness and Evolvability of Personal Learning Environments,*, pages 34–38. CEUR WS. 191

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer. 51

Seth, A., Zhang, J., and Cohen, R. (2010). Bayesian credibility modeling for personalized recommendation in participatory media. In *User modeling, adaptation, and personalization*, pages 279–290. Springer. 58

Sharkey, M. (2011). Academic analytics landscape at the university of phoenix. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 122–126, New York, NY, USA. ACM. 49

Shinavier, J. (2010). Real-time #semanticweb in <= 140 chars. In *Proceedings of the Linked Data on the Web Workshop (LDOW2010)*, Raleigh, North Carolina, USA. 119

Softic, S. (2012). Towards identifying collaborative learning groups using social media. *International Journal of Emerging Technologies in Learning (iJET)*, 7(S2):15–21. xix, xx, 173, 178, 180, 181, 182, 183, 184, 185, 186, 187

Softic, S., De Vocht, L., Taraghi, B., Ebner, M., Mannens, E., and De Walle, R. V. (2014a). Leveraging learning analytics in a personal learning environment using linked data. *Bulletin of the IEEE Technical Committee on Learning Technology*, 16(4). xxi, 209, 212, 213, 214, 215, 216, 217

Softic, S., Ebner, M., Mühlburger, H., Altmann, T., and Taraghi, B. (2010). twitter mining# microblogs using# semantic technologies. *6th Workshop on Semantic Web Applications and Perspectives*, pages 1–9. xviii, 21, 47, 67, 69, 95, 99, 100, 110, 113, 114, 128, 132, 140, 141, 142, 144, 145, 146, 153, 158, 160, 162

Softic, S., Ebner, M., and Taraghi, B. (2016). Dashboard zur verfolgung von lernaktivitäten in einer personalisierten lernumgebung mittels semantischer modellierung der benutzerdaten. 216

Softic, S., Ebner, M., Vocht, L. D., Mannens, E., and de Walle, R. V. (2013a). A framework concept for profiling researchers on twitter using the web of data. In *WEBIST 2013 - Proceedings of the 9th International Conference on Web Information Systems and Technologies, Aachen, Germany, 8-10 May, 2013*, pages 447–452. 18, 95, 128, 129, 142, 144, 145, 146, 153, 158, 160, 162

Softic, S., Taraghi, B., Ebner, M., De Vocht, L., Mannens, E., and Van de Walle, R. (2015a). *Mining and Visualizing Usage of Educational Systems Using Linked Data*, pages 17–26. Springer International Publishing, Cham. xxi, 209, 213, 214, 215, 216, 217

Softic, S., Taraghi, B., Ebner, M., Vocht, L. D., Mannens, E., and de Walle, R. V. (2013b). Monitoring learning activities in PLE using semantic modelling of learner behaviour. In *Human Factors in Computing and Informatics - First International Conference, SouthCHI 2013, Maribor, Slovenia, July 1-3, 2013*.

Proceedings, pages 74–90. xx, xxi, 190, 193, 195, 197, 199, 200, 204, 205, 206, 207, 209, 211, 216, 217

Softic, S., Taraghi, B., and Halb, W. (2009). Weaving social e-learning platforms into the web of linked data. In *I-SEMANTICS*, pages 559–567. 50

Softic, S., Taraghi, B., and Vocht, L. D. (2013c). Activities and trends analytics in a widget based PLE using semantic technologies. In *CSEDU 2013 - Proceedings of the 5th International Conference on Computer Supported Education, Aachen, Germany, 6-8 May, 2013*, pages 199–203. 190, 209, 211, 216, 217

Softic, S., Vocht, L. D., Mannens, E., de Walle, R. V., and Ebner, M. (2014b). Finding and exploring commonalities between researchers using the resxplorer. In *Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Collaboration - First International Conference, LCT 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II*, pages 486–494. xix, 151, 152, 153, 154, 155, 156, 157, 158, 160, 162

Softic, S., Vocht, L. D., Mannens, E., Ebner, M., and de Walle, R. V. (2015b). COLINDA: modeling, representing and using scientific events in the web of data. In *Proceedings of the 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2015) Co-located with the 12th Extended Semantic Web Conference (ESWC 2015), Protoroz, Slovenia, May 31, 2015.*, pages 12–23. 86, 94, 106, 110, 113, 128, 129, 142, 144, 158, 160, 162

Solskinnsbakk, G. and Gulla, J. A. (2011). Semantic annotation from social data. In *SDoW 2011, Social Data on the Web: Workshop at the 10th International Semantic Web Conference (ISWC 2011)*. 51, 57

Stankovic, M., Wagner, C., Jovanovic, J., and Laublet, P. (2010). Looking for experts? what can linked data do for you? In Bizer, C., Heath, T., Berners-Lee, T., and Hausenblas, M., editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org. 47, 83

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., and Oberle, D. (2005). The swrc ontology - semantic web for research communities. *Progress in Artificial Intelligence*, pages 218–231. 36, 90, 113, 132, 162

Tao, K., Abel, F., Gao, Q., and Houben, G.-J. (2011). Tums: Twitter-based user modeling service. In Garcia-Castro, R., Fensel, D., and Antoniou, G., editors, *ESWC Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 269–283. Springer. 21, 38, 47, 48, 69, 97, 132, 140, 141

Taraghi, B., Ebner, M., and Clemens, K. (2012). Personal learning environment – generation 2.0. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 1828–1835. AACE. 23

Taraghi, B., Ebner, M., and Schaffert, S. (2009). Personal learning environments for higher education: A mashup based widget concept. *Proceedings of the Second International Workshop on Mashup Personal Learning Environments (MUPPLE09), Nice, France*, pages 1613–0073. 23

Taraghi, B., Ebner, M., Till, G., and Mühlburger, H. (2010). Personal learning environment-a conceptual study. *iJET, International journal of emerging technologies in learning*, 5(S1):25–30. 23

Taraghi, B., Softic, S., Ebner, M., and De Vocht, L. (2013). Learning activities in personal learning environment. In *25th World Confernce on Educational Media and Technology (EDMEDIA 2013)*, pages 2466–2475. 190, 209, 216

Taraghi, B., Stickel, C., and Ebner, M. (2011). Survival of the fittest – utilization of natural selection mechanisms for improving ple. *Proceedings of the first Workshop on Exploring the Fitness and Evolvability of Personal Learning Environments*, pages 4–9. 23, 217

Tempelton, M. (2008). Microblogging defined. Technical report. 44

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., and Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science*, 1(2):2811 – 2819. Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010)Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010). 50, 54

Thomas, J. J. and Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr. 18

Thonhauser, P., Softic, S., and Ebner, M. (2012). Thought bubbles: a conceptual prototype for a twitter based recommender system for research 2.0. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '12, pages 32:1–32:4, New York, NY, USA. ACM. xvii, xviii, 18, 48, 73, 75, 78, 79, 81, 82, 140, 144

Tjong Kim Sang, E. F. and Buchholz, S. (2000). Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 127–132, Stroudsburg, PA, USA. Association for Computational Linguistics. 80

Tran, T., Herzig, D. M., and Ladwig, G. (2011). Semsearchpro - using semantics throughout the search process. *Web Semantics: Science, Services and Agents on the World Wide Web.*, 9(4):349–364. 43, 46

Ullmann, T. D., Wild, F., Scott, P., Duval, E., Vandeputte, B., Parra Chico, G. A., Reinhardt, W., Heinze, N., Kraker, P., Fessl, A., Lindstaedt, S., Nagel, T., and Gillet, D. (2010). Components of a research 2.0 infrastructure. In *Lecture Notes in Computer Science,*, pages 590–595. Springer. 17, 150

Uren, V., Lei, Y., Lopez, V., Liu, H., Motta, E., and Giordanino, M. (2007). The usability of semantic search tools: A review. *The Knowledge Engineering Review.*, 22(4):361–377. 43

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature News*, 512(7513):126–130. 159

Van Raan, A. F. (1997). Scientometrics: State-of-the-art. *Scientometrics*, 38(1):205–218. 149

Vander Sande, M., Colpaert, P., Van Deursen, D., Erik, M., and de Walle Rik, V. (2012). The datatank: an open data adapter with semantic output. World Wide Web Conference (WWW2012), Developers Track. 39

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., and Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, LAK '11, pages 44–53, New York, NY, USA. ACM. 49

Verbert, K., Manouselis, N., Drachsler, H., and Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society*, 15(3):133–148. 49

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk - a link discovery framework for the web of data. In Bizer, C., Heath, T., Berners-Lee, T., and Idehen, K., editors, *LDOW*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org. 39

Wagner, C. and Strohmaier, M. (2010). The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*. 44

Weller, K., Dröge, E., and Puschmann, C. (2011). Citation analysis in twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In Rowe, M., Stankovic, M., Dadzie, A.-S., and Hardey, M., editors, *Making Sense of Microposts (#MSM2011)*, pages 1–12. 17, 48, 131

White, R. W. and Marchionini, G. (2007). Examining the effectiveness of real-time query expansion. *Information Processing & Management*, 43(3):685–704. 151, 163, 168

White, R. W., Muresan, G., and Marchionini, G. (2006). Evaluating exploratory search systems. *EESS 2006*, page 1. 167

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition. 63

Wölger, S., Siorpaes, K., Bürger, T., Simperl, E., Thaler, S., and Hofer, C. (2011). A survey on data interlinking methods. STI Innsbruck. 39

Wong, P. C. and Thomas, J. (2004). Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21. 18

Xin, R. S., Hassanzadeh, O., Fritz, C., Sohrabi, S., and Miller, R. J. (2013). Publishing bibliographic data on the semantic web using bibbase. *Semantic Web*, 4(1):15–22. 42

Yee, K.-P., Fisher, D., Dhamija, R., and Hearst, M. (2001). Animated exploration of dynamic graphs with radial layout. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS)*. 152

Zhang, L., Yu, Y., Zhou, J., Lin, C., and Yang, Y. (2005). An enhanced model for searching in semantic portals. In Ellis, A. and Hagino, T., editors, *WWW*, pages 453–462. ACM. 43

Zhao, D. and Rosson, M. B. (2009). How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, GROUP '09, pages 243–252, New York, NY, USA. ACM. 20, 97, 111