

Knowledge Diffusion on the Web

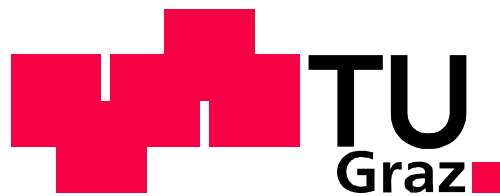
DISSERTATION

submitted to the
Graz University of Technology,
Faculty of Computer Science,
for the attainment of the degree of
Doctor of Engineering Sciences (Dr. techn.)

by

Anwar us Saeed

Institute for Knowledge Management (IWM)
Graz University of Technology



Graz University of Technology

First Assessor and Advisor: **Univ. -Prof. Dr. Klaus TOCHTERMANN**

Second Assessor: **Assoc. Prof. Dr. Andreas HOLZINGER**

Graz, 21 June, 2010

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.

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 entnommene Stellen als solche kenntlich gemacht habe.

Place, Date

Anwar us Saeed

Abstract

Recent developments in the Web termed ‘Web 2.0’ or ‘Social Web’ have brought up new user-generated content and metadata resources in the form of Wikis and blogs as well as social tagging and bookmarking applications, and they transformed the Web into an efficient channel for knowledge diffusion. At the same time W3C, bringing order and structure to the Web, has initiated Linked Open Data (LOD) movement. LOD is a community effort that motivates people to publish their information in a structured way. These new developments of the Web have profound effects regarding how we create, diffuse and consume knowledge and hence need to be researched.

This work intends to probe the applications and opportunities for the diffusion of knowledge on the Web. The thesis makes contributions in four areas:

- 1) Collaborative knowledge creation and diffusion in Wikis
- 2) Measuring knowledge diffusion using tags and bookmarks
- 3) Expertise mining and visualization in scientific communities
- 4) Accelerating knowledge discovery through simplified, user-friendly keyword search interface of LOD

The initial parts of the dissertation describe basic concepts, elaborate the state of the art, and outline the challenges for creation and diffusion of knowledge in open and collaborative Web applications. The subsequent parts propose new approaches and techniques. The thesis proposed and provided a prototypical implementation of new content aggregation and personalization features in Wikis, using a novel sub-document content tagging approach. The effectiveness of tagging and bookmarking was shown by rating and measuring diffusion. The interlinking of digital scientific resources with social digital libraries provided the means of discovering related resources and newly evolving fields and concepts. A multifaceted approach of mining the expertise was able to rank more accurately the experts for scientific knowledge systems. This system is being used by the administration of the digital “Journal of Universal Computer Science”. At the end this work implemented a simplified search interface and keyword-based search mechanism for Linked Open Data which will remove the semantic query requirement and present the information in more logical aggregation. This can enhance global discovery across LOD data sets and hence will advance diffusion of Knowledge.

Kurzfassung

Jüngste Entwicklungen im Web, die als "Web 2.0" oder "Social Web" bezeichnet werden, haben neue benutzergenerierte Inhalte in Form von Wikis und Blogs sowie soziale Tagging- und Bookmarking-Anwendungen hervorgebracht und das Web in einen effizienten Kanal für die Diffusion von Wissen verwandelt. Gleichzeitig hat das W3C, um Ordnung und Struktur ins Web zu bringen, die "Linked Open Data" (LOD) Bewegung initiiert. LOD ist ein Gemeinschaftsprojekt, das die Menschen motiviert, ihre Informationen in strukturierter Weise zu veröffentlichen. Diese neuen Entwicklungen im Web haben tiefgreifende Auswirkungen darauf, wie wir Wissen erstellen, verteilen und konsumieren, und müssen daher erforscht werden.

Diese Arbeit beabsichtigt, die Anwendungen und Möglichkeiten für die Diffusion von Wissen im Web zu sondieren. Die Arbeit leistet Beiträge in vier Bereichen:

- 1) Gemeinschaftliche Wissenserstellung und -diffusion in Wikis
- 2) Messung der Diffusion von Wissen mit Hilfe von Tags und Bookmarks
- 3) Erkennung und Visualisierung von Expertise in wissenschaftlichen Communities
- 4) Beschleunigung der Wissenserschließung durch vereinfachte und benutzerfreundliche stichwortbasierte Suchschnittstellen zu LOD

Die ersten Teile der Dissertation beschreiben die grundlegenden Konzepte, erläutern den aktuellen Stand der Forschung und beschreiben die Herausforderungen für die Schaffung und Diffusion von Wissen in offenen und kollaborativen Web-Anwendungen. Die nachfolgenden Teile beschreiben neue Ansätze und Techniken. Die Arbeit lieferte eine prototypische Implementierung neuer Features zur Aggregation und Personalisierung von Inhalten in Wikis mit einem neuartigen Ansatz, um einzelne Teile eines Dokuments zu taggen. Die Wirksamkeit von Tagging und Bookmarking wurde durch Bewertung und Messung der Diffusion gezeigt. Die Vernetzung von digitalen wissenschaftlichen Ressourcen mit sozialen digitalen Bibliotheken lieferte die Mittel, um verwandte Ressourcen und neue, sich entwickelnde Felder und Konzepte zu entdecken. Ein mehrdimensionaler Ansatz zur Erkennung von Expertise ermöglichte eine genauere Reihung der Experten für wissenschaftliche Wissenssysteme. Dieses System wird von der Verwaltung der digitalen Fachzeitschrift "Journal of Universal Computer Science" eingesetzt. Zum Schluss entwickelte diese Arbeit eine vereinfachte Suchoberfläche und einen stichwortbasierten Suchmechanismus für Linked Open Data, der eine semantische Abfrage überflüssig macht und Informationen in einer logischen Aggregation präsentiert. Dies kann zu einer verbesserten Wissenserschließung über LOD-Datensätze hinweg führen und damit die Diffusion von Wissen vorantreiben.

Acknowledgements

First of all, I would like to praise Almighty Allah Who provided me not only this golden opportunity to enhance my skills but also blessed me with the most visionary and able supervisor Professor Dr. Klaus Tochtermann.

I thank Professor Dr. Klaus Tochtermann for giving me the exceptional opportunity of doing a PhD with him. Prof. Tochtermann extended to me sustained guidance and motivation which enabled me to bring out my best. I learnt from him many things like critical thinking and scientific approaches. He gave me the freedom to explore different ideas, linked me with best researchers like Prof. Dr. Denis Helic and Dr. Markus Strohmaier. Without Prof. Tochtermann's help and guidance it was not possible for me to complete this research.

I am indebted to Professor Dr. Andreas Holzinger for accepting the role of the second reviewer of my dissertation and being the part of evaluation committee. The comments and feedback provided by him were very helpful in refining my dissertation.

I also wish to express my dearest thanks to my parents for their support and confidence. Their encouragement, inspirations and prayers have always been with me. I should not forget to acknowledge my wife for her endurance and understanding. She took care of my children in my absence. I extend my love and acknowledgement to my children Fatima, Aamna and Abdullah for their patience during my research even when they were missing me a lot.

I would like to thank all my colleagues and friends at Know-Center and IWM. Especially, I am grateful to Dr. Alexander Stocker, Patrick Höfler, Syed Khuram Shahzad and Aatif Latif as they extended every possible help and spared time whenever I needed them. They dug me out whenever I felt stuck. I also want to thank Ms. Anke Beckmann, Ms. Anita Grießer and Martin Mayer for their continuous support in office and administrative matters. I extend my thanks to my collaborators in IICM like Dr. M. Tanvir Afzal and Dr. Keith Andrews.

I am highly obliged to the Higher Education Commission of Pakistan for awarding me a fully funded scholarship. Without this financial support I would have not been able to pursue my research. I am also grateful to Dr. Syed Arif Ahmad DG DOS and to my parent organization PAEC for granting me the permission to avail the HEC scholarship and study abroad. I would also like to acknowledge Austrian Exchange Service for their administrative support of the scholarship

Table of Contents

| | |
|--|-----------|
| CHAPTER 1: INTRODUCTION | 1 |
| 1.1 A Note on Terminology | 1 |
| 1.1.1 Knowledge | 2 |
| 1.1.2 Knowledge Sharing, Transfer and Diffusion | 2 |
| 1.1.3 Other Diffusion Concepts | 3 |
| 1.2 Motivation..... | 4 |
| 1.3 Research Trends and Challenges | 6 |
| 1.4 Thesis Objective and Contributions..... | 9 |
| 1.4.1 Foundation of the Dissertation..... | 11 |
| 1.4.2 Research Contributions..... | 12 |
| 1.5 Thesis Organization | 13 |
| CHAPTER 2: BASIC CONCEPTS AND LITERATURE REVIEW | 15 |
| 2.1 Web 2.0: The Brave New Web..... | 15 |
| 2.2 Social Software | 17 |
| 2.2.1 Blogs | 17 |
| 2.2.2 Wikis | 19 |
| 2.2.3 Tagging and Social Bookmarking | 21 |
| 2.2.4 RSS and Syndication..... | 22 |
| 2.3 Linked Open Data (LOD) | 23 |
| 2.3.1 Linked Data Design Principles | 24 |
| 2.4 Citations and Bibliometric | 24 |
| 2.4.1 Citation..... | 24 |
| 2.4.2 Citation Analysis and Bibliometrics | 24 |
| 2.5 Related Work | 25 |
| 2.5.1 Collaborative Knowledge Creation and Diffusion | 26 |
| 2.5.2 Measuring Knowledge Diffusion..... | 27 |
| 2.5.3 Multifaceted Expertise Mining | 29 |
| 2.5.4 Global Discovery on LOD through Simplified Interfaces..... | 30 |

| | |
|---|-----------|
| CHAPTER 3: COLLABORATIVE KNOWLEDGE CREATION /DIFFUSION AND SCIENTIFIC SCHOLARSHIP..... | 33 |
| 3.1 Digital Scientific Publishing..... | 35 |
| 3.1.1 The Open Access Movement:..... | 37 |
| 3.1.2 Shift in Scientific Publishing Paradigm -- Participatory Content..... | 39 |
| 3.2 Wikis Vs. Blogs---- Collaborative Vs. Expert / Personal Knowledge..... | 41 |
| 3.2.1 Collaborative Knowledge Creation in Wikipedia..... | 42 |
| 3.3 Encyclopedia of Life (EOL) | 43 |
| 3.3.1 Comparison: Wikipedia Vs. EOL..... | 44 |
| 3.4 Dynamically Creating Wiki Pages Using Section Tagging..... | 47 |
| 3.4.1 Content Creation and Information Restructuring | 47 |
| 3.4.2 Prototype Application | 49 |
| 3.4.3 Implementation Aspects..... | 52 |
| 3.5 Concluding Remarks..... | 54 |
| CHAPTER 4: MEASUREMENT AND DIFFUSION OF KNOWLEDGE USING BOOKMARKS AND TAGS..... | 57 |
| 4.1 Introduction..... | 58 |
| 4.2 Social Bookmarking and its Potentials in Measuring Knowledge Diffusion . | 60 |
| 4.2.1 Empirical Relationship (Bookmarks/Tags vs Citations)..... | 62 |
| 4.2.2 Author's and Co-authors' Network | 64 |
| 4.2.3 Findings from the Study..... | 65 |
| 4.2.4 Paper Rank Models..... | 66 |
| 4.3 Linking Contextual Resources from CiteULike Using Tags..... | 69 |
| 4.3.1 WWW'06 dataset..... | 71 |
| 4.3.2 CiteULike dataset..... | 71 |
| 4.3.3 Matching Author's Keywords with CiteULike Tags..... | 71 |
| 4.3.4 Recommending Relevant Tags for Research Papers | 72 |
| 4.4 Concluding Remarks..... | 73 |
| CHAPTER 5: MULTIFACETED EXPERTISE MINING AND TOPICAL VISUALIZATION OF EXPERTS | 76 |
| 5.1 Research Overview | 79 |
| 5.2 A Multi-Faceted Expert Profile | 80 |
| 5.2.1 Number of Publications | 81 |

| | | |
|--|---|------------|
| 5.2.2 | Citations Received | 81 |
| 5.2.3 | Reviewer Profile Records | 81 |
| 5.2.4 | Experience..... | 81 |
| 5.3 | Data Extraction | 83 |
| 5.3.1 | Weights Assigned to Experts | 84 |
| 5.4 | Information Visualization | 85 |
| 5.4.1 | Extended Hyperbolic Visualization | 85 |
| 5.5 | Concluding Remarks..... | 88 |
| CHAPTER 6: GLOBAL DISCOVERY THROUGH SIMPLIFIED INTERFACE OF LINKED OPEN DATA..... | | 90 |
| 6.1 | Global Discovery | 91 |
| 6.2 | Linked Open Data | 93 |
| 6.2.1 | URI Dereferencing..... | 94 |
| 6.2.2 | Ontology Classification | 94 |
| 6.3 | Semantic Search Mechanism and SPO Logic for LOD | 94 |
| 6.4 | Proposed Keyword Search Mechanism | 96 |
| 6.5 | Concept Aggregation Framework..... | 96 |
| 6.5.1 | Aggregation Knowledge Bases Layer | 97 |
| 6.5.2 | Property Aggregation Layer | 99 |
| 6.5.3 | Inferred Aspects Layer..... | 99 |
| 6.6 | System Architecture..... | 100 |
| 6.6.1 | Auto-Suggestion Module | 100 |
| 6.6.2 | Information Retrieval Module | 100 |
| 6.6.3 | ‘Search within Properties’ Module | 101 |
| 6.7 | Concluding Remarks..... | 102 |
| CHAPTER 7: SUMMARY AND OUTLOOK | | 103 |
| 7.1 | Results and Discussions..... | 103 |
| 7.2 | Future Prospects..... | 106 |

List of Figures

| | |
|---|-----|
| Figure 1.1: Knowledge Flows..... | 3 |
| Figure 1.2: Thesis Foundation | 12 |
| Figure 2.1: A "meme map" of Web 2.0 (Adopted from [O'Reilly 2005]) | 16 |
| Figure 3.1: Progress Flow of Chapter 3..... | 35 |
| Figure 3.2: Aspects of Openness | 39 |
| Figure 3.3: Novice and expert view [www.eol.org] | 45 |
| Figure 3.4: ST form module..... | 54 |
| Figure 3.5: Architectural diagram of the ST module..... | 55 |
| Figure 4.1: Progress Flow for the chapter 4..... | 58 |
| Figure 4.2: Modules of the study design..... | 63 |
| Figure 4.3: Tag cloud comparison of heavily cited and tagged papers. | 67 |
| Figure 4.4: Comparison of recommended tags for particular author keywords and their relevant CiteULike tags | 74 |
| Figure 5.1: Progress flow of the chapter..... | 78 |
| Figure 5.2: Expert Profile | 82 |
| Figure 5.3: A sample paper XML File..... | 83 |
| Figure 5.4: Hyperbolic Visualization..... | 86 |
| Figure 5.5: Discovery of Potential Reviewers | 87 |
| Figure 5.6: ACM Topic Search Facility | 88 |
| Figure 6.1: Flow of Chapter 6..... | 92 |
| Figure 6.2: Concept Aggregation Framework | 97 |
| Figure 6.3: System Architecture | 101 |

List of Tables

| | |
|--|----|
| Table 2.1: Key distinctions between blogs and wikis (adapted from Fichter 2005a, Wagner & Bolloju 2005, Szybalski 2005)..... | 20 |
| Table 4.1: Heavily bookmarked papers in 2006 got heavy citations in 2007..... | 66 |
| Table 4.2: Top 5 Ranks of Papers with respect to bookmarking and their respective other Ranks | 69 |
| Table 4.3: Top 5 Ranks of Papers with respect to bookmarking and their respective citation Ranks | 69 |
| Table 4.4: Comparison of citation prediction models based on LR | 70 |
| Table 6.1: Person's property list..... | 98 |

Introduction

Knowledge being the primary catalyst for economic and social development the diffusion of knowledge holds an important role in the creation and distribution of knowledge boons. This high potential of knowledge to transform economies and societies attracted the interest of researchers towards understanding the dynamics of its diffusion. Understanding the diffusion of knowledge leads to more efficient strategies for all stake-holders interested in the dissemination of this valued asset.

Knowledge diffuses through channels. These channels may be the paper based publishing, networks, mass media or Internet. Recently, the explosive growth of the Internet and bandwidth has triggered new evolutions in the World Wide Web leading to the provision of ‘the fast lanes of information and knowledge flows’. The focus of this dissertation is to research the collaborative and participatory applications of the WWW for knowledge diffusion. During the course of this research prototype applications are built as the proof of the proposed concept.

This chapter provides a short description of knowledge diffusion terms and related concepts. Then it gives an account of the current challenges in the efficient diffusion of knowledge on the Web. Furthermore, it describes the objective motivations and contributions of the thesis.

This chapter is divided in five sections. The first section describes the terminologies used in this work. The second section gives a note on objective motivations. Third section provides a brief overview of the research trends and challenges. Contributions of the thesis are discussed in the fourth section. In the last section thesis structure is presented.

1.1 A Note on Terminology

This section explain basic terminology related to this work

1.1.1 Knowledge

The vagueness in the use of term knowledge and its different modalities along with the dynamic and fluid nature of knowledge has created a ‘semantic and taxonomic’ fog [Cowan et al, 2000]. Regarding properties of its diffusion, knowledge is mainly classified in two types. Polany calls them tacit knowledge and explicit knowledge. According to Polany, “...tacit knowledge is what is in our heads and explicit knowledge is what we have codified”. [Polanyi, 1976] [Molapo, 2007]

As the tacit knowledge is knowledge that resides in heads the easiest and the only way to disseminate this type of knowledge is through personal interactions and depends upon the holder of the knowledge. The second type of knowledge is explicit knowledge. Explicit knowledge ‘has been or can be articulated, codified and stored in certain media’ [Hoffmann, 2008]. Explicit knowledge is organized and structured. It is available in documents, databases, training videos and other traditional knowledge sharing channels like in the World Wide Web.

There are some other discussions too in the literature, like knowledge vs. information or data, but we do not intend to refer to that ongoing discussion. Within the scope of our work we agree with [Sorenson and Singh, 2006] that "science ... appears to facilitate the codification of knowledge" and this codification of scientific knowledge along with its open availability on the Web are considered to be a major cause of its rapid diffusion. As the knowledge is inherently non-rivalrous, the amount of codified knowledge is not reduced by its consumption. Furthermore, knowledge even grows in value, when consumed, allowing the regeneration of codified knowledge. This property of dissemination and value relationship establishes the motivation for the knowledge holder to diffuse it.

1.1.2 Knowledge Sharing, Transfer and Diffusion

From the knowledge management perspective, we can identify three different types of knowledge flows (1) knowledge transfer, (2) knowledge sharing and (3) knowledge diffusion as shown in Figure1.1

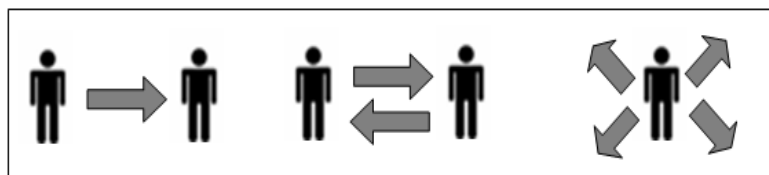


Figure 1.1: Knowledge Flows

With reference to [Puntschart and Tochtermann, 2006], knowledge transfer is the uni-directional targeted transfer of knowledge from a sender to a recipient. Knowledge sharing is an extension to knowledge transfer, where knowledge flows in both directions, from one person to the other. However, apart from transfer and sharing, the concept of knowledge diffusion can be described as the undercurrent (not directly apparent), flow of knowledge irrespective of the direction of flow. Knowledge Diffusion is less specific than directed transfer or sharing of knowledge. Its efficiency is more related to ‘the norm of openness’ [Sorenson and Singh, 2006].

1.1.3 Other Diffusion Concepts

This section provides some other closely related diffusion concepts researched in the social, library and health care sciences.

1.1.3.1 Diffusion of Innovations

A lot of work has been done on the diffusion of innovation, principally by economists, market researchers, and historians. However, innovation has been defined in most cases as technology in use, not scientific knowledge. Some quantitative work has been done, using measurable features of technology, especially statistics for manufacturing, sales, and usage. There is a heavy focus on new product development and marketing, as well as economic impact.

Innovation diffusion, first defined by Rogers (1983) in studies of the agricultural extension agent in the 1950s, has most often been used to refer to the spread of information about innovations (a particular technology, procedure, or organized body of information), resulting in individual adoption of innovative practices and procedures. Diffusion of health care practices among physicians and other professionals has been the subject of many studies under this topic heading. [Backer, 1991]

According to [Rogers, 2003] himself "Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social System." It is also used in the place of Technology diffusion.

1.1.3.2 Web Information Diffusion

The phenomenon of document forwarding or transmission between various web sites is denoted as Web information diffusion. In other words, documents are diffused or transmitted between web sites frequently. Some documents are directly copied or forwarded from one web site to another web site without any changes, and other documents are forwarded between web sites after minor revisions, e.g., addition or deletion of some texts, or rewriting of some sentences. [Wan and Yang, 2007]

There are some other types of diffusion like diffusion of culture as it relates to anthropology.

1.2 Motivation

The Web was originally conceived and developed as a knowledge and information sharing channel between scientists working in different universities and institutes all over the world. 'The basic idea of the WWW was to merge the technologies of personal computers, computer networking and hypertext into a powerful and easy to use global information system'. The growth of the broadband Internet and the scalable architectures transformed the Web into the information highway and an efficient medium of knowledge diffusion.

As the web evolves its purpose and nature of its use are changing. This work focuses on the two important evolutions of the Web, Web 2.0 or specifically Social Software and Linked Open Data. The thesis will probe the applications and opportunities for the diffusion of knowledge on the Web. These evolutions of the Web have profound effects regarding how we create, diffuse and consume knowledge; hence need to be researched.

The term Web 2.0, which has attracted a lot of attention in the Internet world, has been coined to describe the changes that the Web is currently going through [O'Reilly, 2007]. Most of them are caused by the vast growth of the web together with the rise of new collaborative technologies, marked under the umbrella of Social Software, reaching out for a richer user experience. Web 2.0 is, at the same time, a social phenomenon, causing users to interweave their communication and interaction processes with the web.

Users have continually begun to assemble in new types of online communities which are emerging all over the web [Sorenson and Singh, 2006], accompanied by changing their traditional role from mainly using the Internet as a source of information to actively participating in the content creation process. The social phenomenon is enabled by the technical revolution, where new rising technologies including content syndication, semantic annotation and richer user interfaces like wikis and blogs are tempting social interaction, thus resulting in the emergence of new types of collaborative knowledge structures on the web.

Entry barriers of using the web have been reduced mainly due to, amongst others, the radical simplification of interactive user interfaces and easy access to huge pools of knowledge. This has changed the way in which the knowledge is managed and diffused on the Web. Inspired by these open and collaborative trends of Web 2.0 and social software, this thesis proposes that Wikis, Blogs, Bookmarking and Tagging systems can provide an ecosystem of scientific knowledge creation and diffusion. This work introduces a novel sub-document level tagging called selection (or section) tagging. A prototype implementation is presented in a wiki environment with personalization plug-in. We also proposed a tag based recommendation of related scientific resources from social bookmarking service CiteULike¹.

On the other hand, the bulk of the data currently residing on the Web is unstructured or semi-structured at best. Therefore, the W3C launched the Linking Open Data² (LOD) movement, a community effort that motivates people to publish their information in a structured way³. LOD not only “semantifies” different kinds of open data sets, but it also provides a framework for interlinking. This framework is based on the rules described by Tim Berners-Lee [Berners-Lee]. Tim Berners-Lee explains the impact of these semantic technologies will be huge ‘An emerging successor to the web, the Semantic Web , will likely profoundly change the very nature of how scientific knowledge is produced and shared, in ways that we can now barely imagine’. The LOD bears the vision of Tim Berners-Lee and has amassed, as of April 2010, about 13 billion RDF triples, which are interlinked by around 150 million RDF/OWL links [LDOW 2010]. Motivated from this, the thesis makes contributions by simplifying semantic search on LOD using keyword search mechanism. This will enhance global

1 www.citeulike.org

2 <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

3 <http://www.w3.org/RDF/>

discovery across distributed scientific knowledge databases and hence will further the diffusion of knowledge.

The potentials of these web initiatives will be discussed in the next chapter.

1.3 Research Trends and Challenges

The structures and properties of knowledge diffusion in scientific domain have been mainly investigated in past by referring to the diffusion of published (codified) scientific knowledge. In the conducted empirical studies, citations were mainly used as an indicator for the level of diffusion. There are many diffusion studies but three major categories of empirical studies regarding citation analysis of scientific research can be recognized as : (1) Diffusion in networks (e.g. study of co-authorship networks), (2) geographical (e.g. diffusion of knowledge along the supply chain across the borders), and (3) technological (e.g. how university research results are diffusing to industry) contexts.

The diffusion study of scientific work provides researchers with an understanding of its usage and generates evidence for the impact of research on the scientific and economic development from different perspectives.

The patent citation analysis is used in technology diffusion research as indicated in [MacGarvie, 2005][Park and Park, 2006][Maurseth et al. 2002] whereas the academic research citation analysis is used to measure the impact of research [Garfield, 1955], as well as, to study the diffusion of knowledge between science and technology [Branstetter, 2003]. More recent studies have even provided insights of the knowledge flow within blog-networks [Anjewierden et al, 2005]. They frame a research field dealing with the new forms of social structures emerging on the web.

[Scharnhorst and Wouters,, 2006][Day, 2008] have recommended that in addition to studying the diffusion of (codified) scientific knowledge through citations, web based indicators may also be encouraged for assessment of different aspects of science and technology. Looking into this challenge this work tries to find empirical and analytical similarities of citation link structure with web based social tagging and bookmarking links and meta-data. Along with this we see that tags can also be used for recommending scientific articles to improve 'browsing experience' just like references or citations do. We extracted the scientific paper specific tags from CiteULike for the whole set of accepted papers of WWW06 conference. These tags are the hyperlinks to the set of relevant papers

in CiteULike which a user visiting the scientific paper of WWW06 can access by clicking on these tags.

According to [Godwin-Jones, 2003], the purpose of a Wiki site is to become a shared repository of knowledge, with the knowledge base growing over time. [Kristine et al, 2010] notes that transformation considering the ‘dramatic changes in the way that scientific information is collected and disseminated’ due to the Web 2.0 user generated content.

Despite its success in tempting the millions of volunteers, Wikipedia; an open and collaborative bottom up authoring system, still suffers the issues like credibility of content, vandalism and hence failed to inspire the scientific community.

[Roberta et al, 2010] points out that scientific community, regarding scientific publishing, are not catching up with new collaborative trends. It further outlines this challenge that to encourage the scientific community and the business models of the scientific publishing industry towards the adoption of the collaborative revolution of Web 2.0 one should consider two soft drivers

- the certification abilities of publishers and
- the need for reputation of authors

The certification comes from the review process while the citation counts are the currency of reputation for authors. This work provides empirical evidence of similarities among citations and bookmarks. Hence the bookmark reputation of a scientific resource will bring value to its authors like the citations do.

Sanger the cofounder of Wikipedia [Sanger, 2009] pointed out the similar two factors as the explanation for the consistently mediocre quality of most of the Wikipedia articles. He explains that ‘without granting experts any authority (even if it is soft one) to overrule aggressive people’ who have time and hotly guard their articles, ‘there is no reason to think that Wikipedia’s articles are on a vector toward continual improvement’. The second factor he listed is the ‘Wikipedia’s commitment to anonymity’ which deny value to the contributor and hence further drives off good contributors. What role an expert can have is a discussion more towards the social sciences research but we suggest that one useful and soft role of experts may be to color-highlight the content which is against some fact or not credible.

This dissertation contributes to the solution of this challenge by implementing the application for automatic multifaceted discovery and topical visualization of experts in a scientific knowledge system. The visualization can be used to initiate other type of collaborations too. Facets can be grouped from open and conventional knowledge repositories which are now available on Linked Open Data cloud.

The digital information made available by the Web is indexed by different search engines like Google, Yahoo and MSN. These Web search engines further provide search interfaces over the indexed Web pages. One of the most successful search engines, Google, indexed over 26 million Web pages in 1998. The index number reached one billion Web pages in the year 2000. Then by the year 2008, Google achieved a milestone by indexing 1 trillionths (1,000,000,000,000) unique Web page [GoogleBlog, 2008].

This exponential growth in the size of the Web has posed several challenges. One of the biggest challenges is that the indexed information is either semi-structured or not structured at all. Subsequently, this prevents the development of quality services for users and makes it difficult to provide them with the intended information. Some initiatives have been taken to cope with this situation. One of the biggest initiatives is the Semantic Web. The goal of semantic Web is to structure the indexed web pages. The semantic Web focuses on creating an environment where software agents would be able to collect required and accurate information from multiple resources to process them autonomously. However, The Semantic Web is not a separate Web but an extension of the current Web with intentions to provide well-defined meaning to the existing one. This will enable computers and people to work in cooperation [Berners-Lee, 2001]. One of the major success story of Semantic Web is Linked Open Data (LOD). Linked Data (LOD) was launched by W3C in 2006. This movement has motivated people to publish their information in a structured way (RDF). LOD semantifies openly available datasets of various domains and provides a framework for interlinking similar concepts in these datasets. Currently, LOD cloud consists of over 13 billion RDF triples, which are interlinked by about 150 million RDF/OWL links [LDOW 2010]. This initiative paved a way for different kinds of applications to discover more structured (meaningful) and interconnected data to overcome the problem of information supply. Some key challenges related to Linked Data have been pointed out in [Latif et al 2009].

[Wojick et al, 2008] explained that a search mechanism which can query across distributed and diverse databases will 'illuminate often obscure databases

and speed access to scientific information, which will in turn increase the probability of further and more rapid innovation and discovery'. The same principle of 'global discovery' is at the heart of new Socio-Semantic Web movement Linked Open Data (LOD). LOD holds potential not only to enhance the global discovery but also to extend its definition towards the more generic knowledge and data integration principles. [Losoff, 2009] notes that with the rise of digital access to data, the data has become more valuable than the published paper itself. It points out, that the datasets from the Human Genome Project "have more value than any single publication that was derived from an analysis of them" [Carlson, 2008]. But this value of datasets is undermined if they are not searchable by the common users. In the LOD data search applications, it is tedious job for a user to sort out important pieces of information without having the knowledge of underlying ontologies and basic RDF facts. The same gap between semantic search and end user applications has also been identified by [Chakrabarti 2004]. There is a lack of user friendly interfaces and end users usually need to deal with complex semantic mechanisms to explore information.

In solutions to these issues, this work suggested architecture of keyword search mechanism which will hide semantics in order to reduce the cognitive load of the users. It also proposed the Concept Aggregation Framework which conceptualizes the most relevant information of a resource in an easily perceivable construct.

1.4 Thesis Objective and Contributions

As discussed in the previous section, thesis deals with the challenges in diffusion of knowledge from four aspects.

- Collaborative knowledge creation and its diffusion in Wikis
- Popularity (Tag and bookmarks as measures of diffusion): Measurement and Diffusion of Knowledge using bookmarks and tags
- Credibility (through soft reviews by experts): Multifaceted Expertise mining in scientific repositories and their topical visualization
- Global Discovery: by providing web user friendly keyword search interface for semantic data sets in LOD.

This section provides short description of the each of them.

A) Collaborative Knowledge Creation and Diffusion

This study provides a brief description of new evolving scientific knowledge diffusion platforms alongside it describes web 2.0 applications like Wikipedia and EOL by comparing them. Further more this work proposes a novel combination of granular tagging in Wiki systems for rapid restructuring and information import. The underlying consideration is that the lowering of editing barriers can speed up the content generation in wiki systems. This section proposes a prototype of an application for efficient aggregation of resource snippets from diverse sources (wiki pages in this prototype) using section tagging and bookmarking to build dynamic wiki pages in Austria Forum. It is further discussed in chapter 3.

B) Measurement and Diffusion of Knowledge using Bookmarks and Tags

Based on multiple empirical studies this research element explored the potentials of the bookmarking applications in the diffusion of knowledge and its estimation. It further probes their similarities to citations which are a conventional measure of diffusion of knowledge. Above that, tagging practices have an added advantage to augment the understanding of knowledge diffusion by providing an additional element – the user context in tagging a resource of knowledge. Moreover, it shows how the relevant concepts and papers from these socially maintained reference management systems can be linked to scientific papers in other digital repositories by mining and using contextually relevant tags from these systems.

C) Multifaceted Expertise Mining and Topical Visualization of Experts

[Sanger 2009] argues that if open and bottom up resources, like Wikipedia, are to become authoritative there must be some role, may be softer, for expert overview of the facts in the content.

The focus of this research component is that how we can assign experts (as reviewers) automatically to the topics of the content. We propose an innovative automated technique which incorporates multiple experience atoms to judge the overall expertise of an individual in providing a more representative assessment of expertise. For the prototype application, proposed in this research, we used the online Journal of Universal Computer Science (J. UCS) database for mining expertise. The chapter 5 further elucidates this approach.

D) Global Discovery through Simplified Search Interface of LOD

In this part of research we will explain the concept of Global discovery and its importance for knowledge diffusion. Then we will detail that how Linked Data framework enables Web of Data and Global Discovery. Furthermore we propose keyword based search architecture and a concept aggregation framework to bridge the gap between end user and semantic search on LOD.

The Linked Data best practices, otherwise termed as the design principles, hold great potential to enhance global discovery by integrating digital scientific data with scholarly literature. We proposed and implemented keyword search mechanism to reduce the cognitive load of the users. We also proposed the Concept Aggregation Framework conceptualizes the most relevant information of a resource in an easily perceivable construct. Chapter 6 provides further detail on this topic.

1.4.1 Foundation of the Dissertation

The foundation of this dissertation is a selected set of publications authored or co-authored by the author of this thesis over a period of about three and half years. Their relation and their arrangement in the dissertation are depicted in Figure 1.2

The focus of this dissertation is to research the collaborative and participatory applications of the WWW for knowledge diffusion. The thesis makes contributions broadly in four areas 1) Collaborative knowledge creation and diffusion in Wikis 2) Measuring knowledge diffusion using tags and bookmarks The thesis analyses the potential of the tagging and bookmarking metadata resources for the study of knowledge diffusion by finding its empirical relationship and similarities with citation (an established indicator of knowledge diffusion).3) Expertise mining in scientific communities and its visualization 4) Accelerating Global Discovery through simplified user friendly search Interface of LOD

These research areas are discussed in chapter 3, 4, 5, and 6 respectively. Every chapter is based on a set of 1-3 publications as depicted in Figure 1.2.

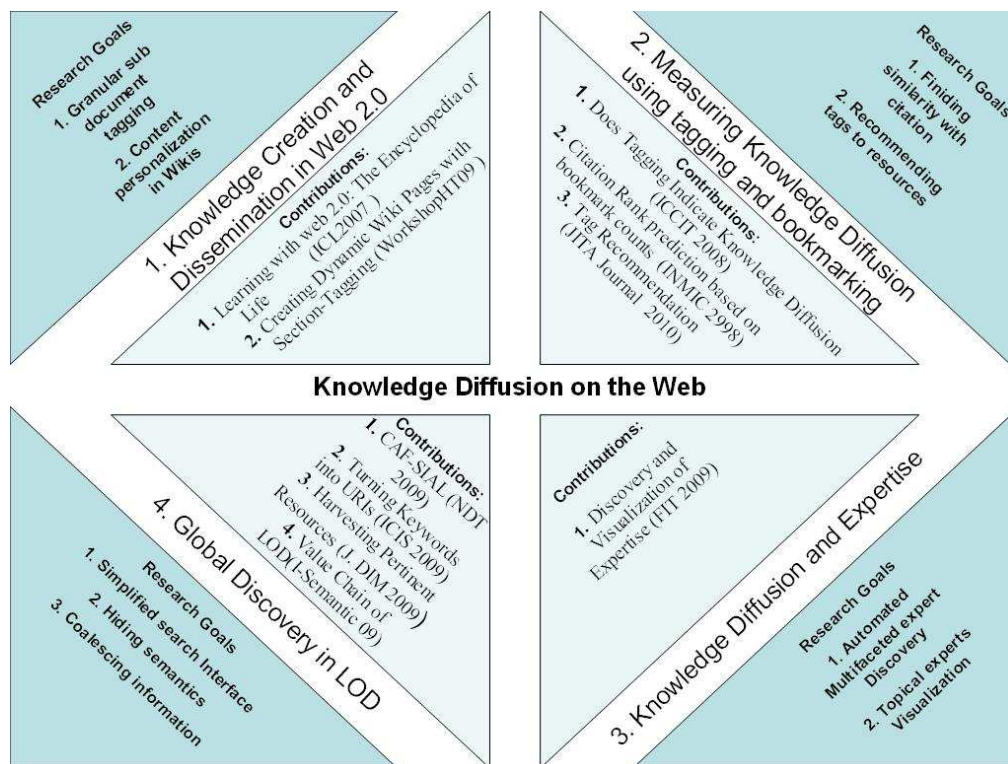


Figure 1.2: Thesis Foundation

1.4.2 Research Contributions

The author of the thesis has published following research papers during the period of three and half years.

- Us Saeed, A., Afzal, M.T., Latif, A., Tochtermann, K., Recommending tags for scientific resources, accepted for publication in the Journal of IT in Asia (JITA), 2010
- Us Saeed. A, Afzal, M.T.,Latif, A., Stocker, A., Tochtermann, K., Does Tagging indicate Knowledge diffusion? An exploratory case study, In Proc. of 3rd ICCIT pp.605-610 , 2008
- Us Saeed, A., Afzal, M.T., Latif, A., Tochtermann, K., Citation rank prediction based on bookmark counts: Exploratory case study of WWW'06 papers, INMIC 2008. IEEE International pp. 392 - 397, Dec. 2008
- Us Saeed, A., Stocker, A., Hoefler, P., Tochtermann, K., "Learning with the Web 2.0: The Encyclopedia of Life", in Conference ICL2007, Villach, Austria, 2007.

- Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K., Harvesting Pertinent Resources from Linked Open Data. To appear in the Journal of Digital Information Management, 2009.
- Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K., CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data. Proceedings of NDT 2009, Ostrava, Czech Republic, July 2009.
- Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K., Turning Keywords into URIs: Simplified User Interfaces for Exploring Linked Data. Accepted for: ACM proceeding of ICIS 2009, Seoul, Korea, November 2009. ISBN: 978-1-60558-710-3
- Latif, A., Hoefler, P., Stocker, A., Us-saeed, A., Wagner, C (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: Proceedings of I-Semantic. Graz, Austria.
- Afzal, M. T., Latif, A., Us Saeed, A., Sturm, P., Aslam, S., Andrews, K., Tochtermann, K., Maurer, H. (2009). Discovery and Visualization of Expertise in a Scientific Community. In: Proceeding of International Conference of Frontiers of Information Technology, Islamabad, Pakistan, 16-18, Dec. 2009.
- Helic, D., Us Saeed, A., Trattner, C., Creating Dynamic Wiki Pages with Section-Tagging , in HT09 workshop New Forms of Xanalogical Storage and Function, 2009

1.5 Thesis Organization

The current chapter serves as an introduction explaining some important definitions which are fundamental to the thesis work. It also identifies the research challenges and contributions in this field. The rest of the content is organized as follows:

Chapter 2 describes briefly the two evolving structures of the Web as well as it provides the literature review and state-of-the-art in the respective research. Chapter 3 provides comparison of two successful but different kinds of Wikis, Wikipedia and EOL. Afterwards it explains the prototype implementation of a new granular tagging approach and personalization in Wiki environment. Chapter 4 probes potential of tagging by empirical comparison with citations. Further on it implements a novel approach to link the socially maintained libraries (CiteULike) with papers in other scientific repositories. Chapter 5 elaborates the discovery and visualization of expertise in scientific communities. Chapter 6 explains the concept of Global Discovery and its importance in knowledge diffusion. It also

give details for the architecture and implementation of an innovative keyword based search interface for exploring semantic data in LOD. Therefore the system evaluation in terms of usefulness of the system is illustrated in Chapter 7. The thesis ends with conclusion and future work as highlighted in chapter 7

Basic Concepts and Literature Review

This chapter briefly describes the basic concepts related to the work of this thesis and the existing state-of-the-art systems. In the beginning of the chapter concepts like Web 2.0, social software, Linked Open Data and citation analysis are explained. While in the later sections an account is provided for the past research in regard to the each of four aspects of knowledge diffusion as mentioned in previous chapter.

2.1 Web 2.0: The Brave New Web

The concept of "Web 2.0" was coined in 2004 in a brainstorming session by Dale Dougherty, web pioneer and vice president of O'Reilly Media. They pointed out that even if the dot com bubble has "crashed", the web has become more important than ever as the exciting new applications and sites were popping up with surprising regularity. Furthermore, they noted that 'the companies that had survived the collapse seemed to have some properties in common' [Oreilly 2004]. Although the term Web 2.0 itself is confusing as it indicates a kind of technological or software up gradation but on the contrary it is not characterized by a new step of technology like in the case of Semantic Web [Berners-Lee et al. 2001]. Instead of defining Web 2.0 on the technological basis Tim O'Reilly defined the web 2.0 principles or otherwise known as design patterns.

Today, the term Web 2.0 is used to describe web applications that distinguish themselves from previous generations of software by a number of principles (see figure 2.1). These design patterns as described by O'Reilly are listed below.

- Use of the Web as a platform
- Harnessing collective intelligence
- Data is the next Intel Inside
- Perpetual beta
- Lightweight programming models
- Software above the level of a single device
- Rich user experience

The term Web 2.0 itself was debated in research as confusing because it indicates a kind of technological or software up gradation but on the contrary it is not characterized by a new step of technology as is the case of the Semantic Web [Berners-Lee et al. 2001] [Ullrich et al. 2008].

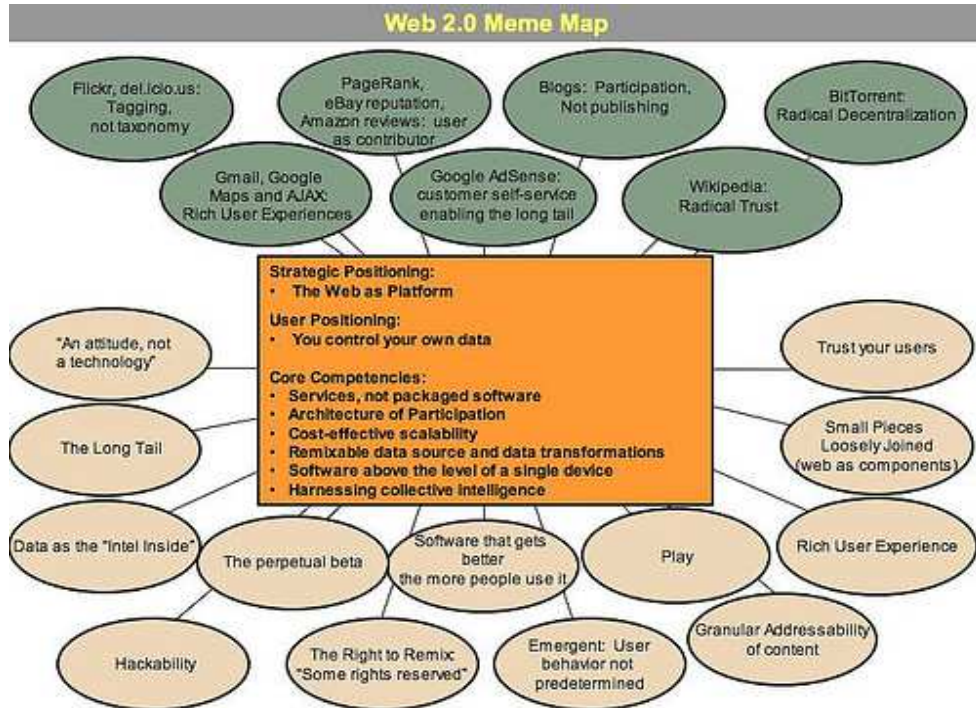


Figure 2.1: A "meme map" of Web 2.0 (Adopted from [O'Reilly 2005])

- The major transformations of the Web were the change from a medium to a platform, from a read-web to a read-write-web and it also entered a new, more social and participatory phase. These trends have led to a feeling that the Web is entering a 'second phase'—a new, 'improved' Web version 2.0 [Anderson 2007]. [Vossen and Hagemann 2007] describe this as the evolution of web and divided it into three streams;
 - The application stream
 - The technology stream and
 - The user participation and contribution stream

They also advanced the perception that the future evolution of the Web will be driven by these streams too. In line with these streams, [Ankolekar et. al. 2008] described that even that the Web 2.0 and the Semantic Web are considered to be separate and competing visions but 'the core technologies and concerns of these two approaches are complementary and that each field can and must draw from the other's strengths'. They predicted that the new Web will breed the socio-

semantic applications using Web 2.0 for front end to attract rich user interaction and Semantic web as the backbone for providing computational intensive data services.

The scope of this work encircles only some of social software applications, a subset of Web 2.0 applications, and Linked Open Data (LOD) framework. Description of LOD will be provided in later sections of this chapter while in the next section we will provide a brief introduction of Social software and related Web applications.

2.2 Social Software

Social software has emerged as a major component of the Web 2.0 movement. The idea using networked computing to connect people in order to boost their knowledge and their ability to learn dates as far back as the 1960s. But this is only recently, during the past few years, that this vision is supposed to be emerged practically through a group of Web projects and services which are perceived as especially connective. These applications are marked under the title of “social software” [Alexander 2006].

[Anderson 2005] notes that social software is a very difficult concept to define. It further points out that the term not only includes scalable interlinking technologies but also the social effects created due to the combined or interlinked usage of these technologies [Dalsgaard 2006]. The examples of social software technologies which will be discussed in this work include weblogs, wikis, social bookmarking, and syndication RSS/Atom feeds. It is, however, important to note that social software is in no way limited to these specific technologies.

2.2.1 Blogs

The term web-log, or blog, was coined by Jorn Barger in 1997 [Anderson 2007]. A blog is a simple webpage consisting of posts arranged chronologically with the most recent first, in the style of an online journal. Posts in a blog are brief paragraphs generally representing the opinions of the blog holder. These posts are like personal diary entries and may contain any information or links [Doctorow et al, 2002]. Most blogs allow visitors to add a comment below a blog entry. Blogs are also penetrating to the practice of educational institutions. [Holzinger et al, 2009a] shows that blogs can support to improve learning performance by supplementing traditional lecturing.

The posting and commenting process in blogs is also called a ‘weighted conversation’ as this important feature of blogs provides a generic feedback from the community on the Web about the opinion of the primary author. Blogging provides a channel for exchange of views.

Bloggings’ informal chronologically organized diary like postings provide a sense of immediacy, since ‘blogs enable individuals to write to their Web pages in journalism time –that is hourly, daily, weekly – whereas the Web page culture that preceded it tended to be slower moving: less an equivalent of reportage than of the essay’ [Benkler, 2006].

Blogging applications allow authors to tag each post with a keyword or two. These tag terms are then used for the categorization of the subject of the post within the system. Such a categorization is helpful in organizing the older posts into a standard theme-based menu system. Clicking on a post’s description, or tag (which is displayed below the post), will take the user to a list of other posts the same author which use the same tag [Anderson 2007].

Another important aspect of blogging is linking as it deepens the conversational nature of the blogosphere and its sense of immediacy. Linking also helps to manage the information retrieval and referencing on different blogs but some of these are not without inherent problems. Below we provide a short description of these linking mechanisms.

2.2.1.1 The Permalink

It is a permanent URI (universal resource identifier) which is generated by the blogging system and is applied to a particular post. The permalink don’t change during achieving or any other change in the blog. There is no version control, and using a permalink does not guarantee the content of a post.

2.2.1.2 Trackback (or Pingback)

This linking method allows the system to notify that another blogger have referenced or commented on posts of first one. System also creates automatically a record of the permalink of the referring post. Trackback only works when it is enabled on both the referring and the referred blogs. Some bloggers deliberately disable trackback to avoid spamming.

2.2.1.3 The Blogroll

It is a list of links to other blogs that a particular blogger likes or finds useful. It is similar to a blog 'bookmark' or 'favorites' list. Blog software also facilitates syndication, in which information about the blog entries, for example, the headline, is made available to other software via RSS and, increasingly, Atom. This content is then aggregated into feeds, and a variety of blog aggregators and specialist blog reading tools can make use of these feeds.

The large number of people engaged in blogging has given rise to its own term – **blogosphere** – to express the sense of a whole 'world' of bloggers operating in their own environment. As technology has become more sophisticated, bloggers have begun to incorporate multimedia into their blogs and there are now photo-blogs, video blogs (vlogs), and, increasingly, bloggers can upload material directly from their mobile phones. [Nardi et al, 2004] provides an account of the reasons why people blog, the style and manner of their blogging and the subject areas that are covered [Anderson 2007].

2.2.2 Wikis

'Wikis are radically different than blogs and require a fundamentally different orientation towards truth and knowledge to be successful. By simply removing the traditional author-reader relationship, knowledge-building via wikis becomes a community effort, which requires a substantial paradigmatic shift from traditional views of truth and knowledge' [Gijsbers, 2004].

As compared to blogs which serve as interactive personalized publishing platforms, wikis provide the foundation of the collaboration platform [Holzinger et al. 2009]. The term Wiki is adopted from Hawaiian term 'wiki wiki' meaning "quick". The Wiki systems are websites that allow users to easily add, delete and edit website content. [Ebersbach et al, 2006] defines a wiki as a webpage or set of web pages that can be easily edited by anyone who is allowed access. The popular success of Wikipedia exhibit that the concept of the wiki, as a collaborative tool for facilitating group work, is widely understood [Anderson 2007]. Simple, hypertext-style linking between wiki pages is used to create a navigable set of pages. Each wiki page typically contains a concept (a title/name) and a description of that concept (an article). The edit button displayed on the Wiki pages allows users to change or even delete the contents of the page in question. This ease of access, intuitive interface and a single repository make wikis an efficient and effective tool of mass collaboration and hence a "living document".

As Wikis are web of interlinked pages created by users, they are ‘freeform, informal and emphasize content over form’ [Cooney 2006]. The barriers of entry in Wikis are kept very low giving users as much power as possible to change the content. Wikis have become a tool for online collaboration and community building. They differ from blogs in several ways. The fundamental difference is that wikis do not contain chronological posts, and are otherwise not a tool for recording chronological data. Wikis generally have a ‘history function’, which allows storage of previous versions for later content examination. They also possess a ‘rollback function’, which restores previous versions if required [Anderson 2007]. Old versions of pages and recent changes of pages are all well documented and manageable by users and/or administrators [Cooney 2006].

Table 2.1: Key distinctions between blogs and wikis (adapted from Fichter 2005a, Wagner & Bolloju 2005, Szybalski 2005)

| Feature | Blog | Wiki |
|------------------------------|--|--|
| Focus: | Currency (Most recent information takes precedence and pushes other content down) | Importance: (Most important information takes precedence and remains in focus) |
| Organization: | Chronological | Topical |
| Mode of distribution: | One-to-many (Or few-to-many) | Many-to-many |
| Attribution: | Single author (or small group) | Community (and largely anonymous) |
| Content control: | Centralized (Only author(s) can create content) | Decentralized (Anyone in the community can create and manipulate content) |
| Version management: | Not offered (content is not typically modified once posted) | Full version and complete change history |
| Personality / Point of view: | Author or group, personality plays a key role | Generally neutral, or multiple points of view, personality is minimized |
| Development cycle: | Content is published quickly in final form | Content generally continues to evolve long after initial publication |
| Best suited for: | Short, time-sensitive material like diaries, journals, news, opinions and reviews | Documents with longer life, expected to be edited and refined over time, e.g. knowledge management, FAQs, best practices, etc. |

2.2.3 Tagging and Social Bookmarking

Tagging systems are increasingly becoming popular in the web. The reason for increasing success of these systems is the fact that no specific skills are needed for participating, and that these tools yield immediate benefit for each individual without too much overhead. These tagging systems enable the users to add keywords (tags) to web resources (web-pages, images, documents, papers) without having to rely on a controlled vocabulary [Marlow et al. 2006]. One of the first large-scale application of tagging was seen with the introduction of Joshua Schacter's del.icio.us website, which launched the 'social bookmarking' phenomenon [Anderson 2007].

Social bookmarking systems possess a number of common features [Millen et al, 2005]: The users can create lists of 'bookmarks' or 'favourites', which are stored centrally on a server rather than within the client browser. These applications also allow users to share their bookmarks with other users of the system. These bookmarks can also be tagged with keywords.

The concept of tagging has proved to be contagious on web and have spread to diverse resource sharing services like Flickr(photos), YouTube (video) and Odeo (podcasts) which allow a variety of digital artefacts to be socially tagged. A particularly important example within the context of academics is Richard Cameron's CiteULike (www.Citeulike.org), a free service to help users to store, organize and share the academic papers they are reading. When you see a paper on the Web that interests you, you click a button and add it to your personal library. CiteULike automatically extracts the citation details, so you don't have to type them in [Anderson 2007]. CiteULike, Del.icio.us (www.delicious.com) and Bibsonomy (www.Bibsonomy.org) were used for the research in this thesis.

Below we provide short description of the terms which are frequently used in the discussions of tagging and bookmarking.

- **Folksonomy**

These new socially maintained resource and link management systems use free form tags for dynamic categorization of resources. This unstructured (or better, free structured) approach to classification with users assigning their own labels is variously referred to as a 'folksonomy' [Hammond et al, 2005]. The word 'folksonomy' is a blend of the words 'taxonomy' and 'folk', and stands for conceptual structures created by the people. Folksonomies are thus 'a bottom-up complement to more formalized Semantic Web technologies, as they rely on

emergent semantics which result from the converging use of the same vocabulary' [Hotho et al, 2006]. The main difference to 'classical' ontology engineering approaches is the simplicity avoiding any formal modeling overhead on the part of the common user. Intelligent techniques may reside under the interface layer of the system and should be hidden from the user. Overall, these systems provide a very 'intuitive navigation through the data' [Hotho et al, 2006].

- **Personomy**

The collection of all tag assignments of a user is called his personomy. The collection of all personomies results in an overall folksonomy [Hotho et al, 2006].

- **Tag cloud**

Tag clouds are groups of tags (tag sets) from a number of different users of a tagging service, which collates information about the frequency with which particular tags are used. This frequency information is often displayed graphically as a 'cloud' of terms in which tags with higher frequency of use are displayed in larger text [Anderson 2007]. Tag clouds are the visualization pattern for personomies and folksonomies.

Tagging holds potential to improve the search on the web. Tagging systems introduce new forms of social communication and generate new opportunities for data mining.

2.2.4 RSS and Syndication

RSS is term used to represent a family of formats which allow users to get updates to the content of RSS-enabled websites, blogs or podcasts (a series of digital media files , either audio or video, that are released episodically and often downloaded through web syndication) without actually visiting the parent site. Typically, a new story's title and synopsis, along with the originating website's name is collected within a feed which uses the RSS format. These content feeds are 'piped' to the user in a process known as syndication.

A software tool known as an aggregator or feed reader is required to be installed by the users on their desktops in order to collect these feeds. With aggregator users can subscribe to multiple feeds from different websites. The feed reader will then periodically check for updates to the RSS feed and keep the user informed of any changes.

In its earliest incarnation the term RSS was understood to stand for Rich Site Summary as it was used by Netscape to extend users the feature to create custom Netscape home pages with regularly updated data flows. Later on Netscape lost interest, and the technology was carried forward by Dave Winer's company under the name 'Really Simple Syndication' [O'Reilly 2005].

Technically, RSS is an XML-based data format for websites to exchange files that contain publishing information and summaries of the site's contents.

In 2003 a new syndication system was proposed and developed under the name Atom in order to clear up some of the inconsistencies between RSS versions and the problems with the way they interoperate. This consists of two standards: the Atom Syndication Format, an XML language used for Web feeds, and the Atom Publishing Protocol (APP), a HTTP-based protocol for creating and updating Web resources. The two most important differences between the two are, firstly, that the development of Atom is taking place through a formal and open standards process within the IETF(Internet Engineering Task Force), and, secondly, that with Atom the actual content of the feed item's encoding (known as the payload container) is more clearly defined . Atom can also support the enclosure of more than one podcast file at a time and so multiple file formats of the same podcast can be syndicated at the same time [Anderson 2007].

2.3 Linked Open Data (LOD)

The World Wide Web can be seen as a huge repository of networked resources. Due to its exponential growth, it is a challenging task for search engines to locate meaningful pieces of information from heavily redundant and unstructured resources. The semantic paradigm of information processing suggests a solution to the above problem. Semantic resources are structured, and related semantic metadata can be used to query and search the required piece of information in a very precise manner. On the other hand, the bulk of the data currently residing on the Web is unstructured or semi-structured at best. Therefore, the W3C launched the Linking Open Data⁴ (LOD) movement, a community effort that motivates people to publish their information in a structured way⁵. LOD not only "semantifies" different kinds of open data sets, but it also provides a framework

4 <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

5 <http://www.w3.org/RDF/>

for interlinking. This framework is based on the rules described by Tim Berners-Lee.

2.3.1 Linked Data Design Principles

Tim Berners-Lee in his article [Berners-Lee et al, 2006] described Linked Data publishing guidelines or principles which he himself called rules. These rules are as follow:

1. Use URIs as names for things:
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs so that they can discover more things.

These principles have provided a scalable architecture of linking and accessing structured data on the Web.

2.4 Citations and Bibliometric

In technological terms, scholarly communication is being transformed through the use of personal and portable computers, electronic mail, word processing software, electronic publishing, digital libraries, the Internet, the World-Wide Web, mobile phones, wireless networks, and other information technologies. Bibliometrics offers a powerful set of methods and measures for studying the structure and process of scholarly communication [Borgman & Furner, 2002].

2.4.1 Citation

Citation is a relationship between two published papers or articles where normally the author/s of 'citing' paper infer/s from and refer/s to the part of 'cited' paper used to extend or create knowledge published in the 'citing' paper. Such citations can be counted as measures of the usage and impact of the cited work [Garfield, 1998] [Moed, 2005]. This is called citation analysis which is one of the bibliometric methods.

2.4.2 Citation Analysis and Bibliometrics

Bibliometrics offers a powerful set of methods and measures for studying the structure and process of scholarly communication [Borgman & Furner, 2002]. Citation analysis, the best known of bibliometric approaches, has become more

sophisticated over the time. On the other hand the advent of networked information technologies has led to quantitative and qualitative advances in other bibliometric methods. More content is available online in digital libraries, and more of it is in full text (and in other media including still and moving images, sound, and numeric data). More connections exist between documents, both in the form of citations and in the form of active hyperlinks that allow an information seeker to move between related documents. Bibliometrics is being applied in new ways, to ask new questions. New analytical approaches like “cybermetrics” (the title of an electronic journal) and “webometrics” [Almind & Ingwersen, 1997] have emerged with the rising popularity of the Web.

Citation can be considered as the connection or a type of link between documents. In general terms the methods of link analysis, then, are those employed in studies in which data are collected primarily in the form of counts of links — pointers to, references to, or citations of “target,” “cited,” or “later” documents made in the text of “origin,” “source,” “citing” or “earlier” documents. There are two general purposes for which link analyses may be conducted: contextualization and evaluation [Borgman & Furner, 2002].

Evaluative citation analysis for determining the quality of research is not without controversy. It is sometimes impugned on the basis that “quality” — the characteristic that citation counts are used to measure — is not an attribute that may be evaluated objectively at all, but one whose values depend on the subjective opinions of individuals [Borgman & Furner, 2002]. These proponents talk about negative citations, self-citations, and methodological papers. They claim that a high citation count can be achieved by publishing low-quality work that attracted a lot of criticism. This raises a more fundamental question that what facet of scientific performance do citation counts measure. [Garfield, 1979] argues that citation counts are a measure of scientific activity. Usually, in science and technology citations are considered as an indicator of diffusion of a published work.

2.5 Related Work

In this thesis work we propose that in the ecology of Web 2.0 and LOD applications a combination of wikis, personal WebPages or web-logs and tagging/bookmarking systems with semantic LOD data integration can provide a base platform for the future socio-semantic knowledge applications with high potentials of knowledge diffusion. In a coherent web 2.0 publishing environment Wikis can provide a diverse content aggregation or publishing platform with

embedded certification processes while blogs and bookmarking systems can provide author reputation mechanisms. In this section we will review the literature about these applications in the light of four aspects of knowledge diffusion, as stated in previous chapter.

2.5.1 Collaborative Knowledge Creation and Diffusion

The popularity of social software has brought up new user generated content and metadata resources in the form of wikis, blogs, social tagging and bookmarking applications. These new systems have emerged as a major force reshaping the information spaces on the World Wide Web in order to better serve both collaborative and personalized information needs of users. In social software applications Web has drifted towards users' content creation as a major contributing factor to the Web resources instead of the commercial content. For instance, wikis are used for sharing, management, and organization of knowledge. Wikipedia is a user-created encyclopedia and a well known example of a wiki system. Wiki systems are asynchronous, collaborative authoring and content versioning systems where any user can add and edit content. A new version of the page is stored in the system after each editing operation [Désilets et al, 2005].

Wikipedia has been highlighted as a success story of low-cost collaborative knowledge systems. The openness of Wikipedia to new users has been cited as both a source of strength and weakness [Hafner, 2006]. One of its key strengths based on its open editing model lies in attracting contributions from new users who may make few edits. This suggests a kind of "wisdom of crowds" effect [Surowiecki, 2004] in which quality of its content is derived by a large number of people making small contributions.[Kittur et. al, 2007]

In wiki systems, user's content-creation/authoring processes involve laborious tasks like information selection from diverse resources, restructuring, modification, and adaptation of information object according to the perceived context [Nelson et al, 2008]. The reuse of existing content in the form of copy-paste mechanisms in order to restructure and create new documents is applied by authors frequently. For example, a typical editing workflow in wiki systems involves investigating volumes of information wherein fact only small part of that information is relevant to the current user need. Thus, the user has to browse all the resources again and again to review the related pieces of information from their relevant or selected resources. This typically requires a lot of effort and time.

On the other hand resource organization with tagging and bookmarking services like Delicious, CiteULike or Bibsonomy have received community focus

due to ease of use and information discovery mechanisms [Hotho et al, 2006]. In social tagging and bookmarking applications users assign free form keywords and annotations to the addresses (URLs) of an information resource(e.g., a web page) [Hammond et al, 2005]. These keywords relate the current user's context to the content of a tagged resource.

As [Ames and Naaman , 2007] suggests the user motivation to tag a resource might be organizational or communicational but in general the users tag resources for their personal use and/or to share them with others. For example, users who tag resources for their personal use in an organizational sense use social tagging applications to organize interesting, important, and related resources according to their current needs. The tags are applied as a support for later search and retrieval of tagged resources via search or navigating the tag cloud. Typically, the tag cloud provides an overview of defined tags showing only the tags themselves but not the actual content of the tagged resources. The resources are represented via navigable links. Another motivation of using tags is to share them with other users and in such a scenario tags are typically used in a communicational sense to send signals to other users about resources that might be of interest in a more general case.

In this work we proposed a new subdocument tagging named section tagging to create dynamic wiki pages by coalescing and restructuring the tagged content from different wiki pages. The prototype proof of concept application is implemented in Austria Forum experimental wiki environment.

2.5.2 Measuring Knowledge Diffusion

In this aspect of knowledge diffusion we probed the potential of social bookmarking and tagging in relation to the citation uses for measuring diffusion and linking high value resources. Below we provide an account of the related work regarding citation analysis and social bookmarking and tagging.

Bookmarking is provided as a popular personalization feature which allows researchers to organise their resources on the Web but now these applications also provide bibliography export in multiple formats (bibtext, EndNote, RDF etc.) which is an added advantage.

Tagging is already a driving component in the fields of emergent semantic techniques [Mika et al, 2005], Information Retrieval [Wu et al, 2006] [Hotho et al, 2006] and user profiling [Huang et al, 2008]. Information retrieval and textual mining is already being used in many decision making systems, such as in the

case of medical sciences [Holzinger et al, 2008]. Tagging can also be helpful in these systems as ‘in a collaborative tagging system, tags codify the knowledge of relationships among documents and concepts represented by the tags. Harvesting individuals through folksonomies, therefore, can benefit the whole society’ [Wu et al, 2006]

[Mika et al, 2005] has studied the tagging behaviors and their usage in delicious, an emerging bookmarking service. He used actor, concept, and instance nodes as a tripartite graph to explain the emergence of ontologies from social context where he considers tags as a socially represented concept.

Citation prediction has also been of interest to the link analysis research. A citation is a directed link from citing paper to cited paper. [Popescul and Ungar, 2003] presented an ‘upgrade’ model of Standard Logistic Regression with the name of Structural Logistic Regression. They combined the standard logistic regression with feature generation from relational data. They demonstrated the effectiveness of their techniques by applying the method to link prediction in the citation network of CiteSeer. They extracted features from the CiteSeer relational database and applied learning models to decouple the feature space and predict the link. They also rediscovered evidences for some common old features and concepts like bibliographic coupling, co-citations and hub documents.

[Manjunatha et al, 2003] Citation Prediction system was selected as winner of KDD Cup 2003 Task-1. The goal of KDD cup2003 was to understand and realize applications to solve contemporary learning problems using past experience data. The arXive dataset was provided for developing the citation prediction models. The winning candidates modeled on the basis of quarterly (in 3 months) changes in citations and calculated the parameters of regression function from the training set of changes in citations on quarterly basis.

Co-authorship and co-author collaborative networks are considered as proxy for high citation counts and are also studied in citation prediction models. Citation prediction models are also interesting for the Link analysis and statistical modeling techniques. The correlation of citing behavior with bookmarking has not yet been explored. The bookmarking of a publication can safely be assumed as the interest of a researcher in a particular (related to his context) publication. Many researchers have explored that the increase in number of authors per publication may increase the number of citations per paper. But very few have experimented with the co-author network in this regard, although the co-author network volume is a direct representation of that authors collaborating behavior.

[Figg et al, 2006] analyzed the relationship between the citation rate of an article and the extent of collaboration. They analyzed the data from 6 leading journals for the years 1975, 1985, and 1995. They found that a correlation exists between the number of authors and the number of times an article is cited in other articles. They suggested that the researchers who are open produce high impact research acquiring higher number of citations.

In [Goldfinch et al, 2003] Goldfinch used negative binomial regression model by taking citations as dependent variable and predicting the citation behaviors and its dependence on co-authorship, number of authors, number of institutions involved, number of international authors. It uses the publication data of Crown Royal Institutes using ISI web of data to retrieve citations. The results vet that co-authorship and involvement of institutions especially international ones inflates citations heavily.

Having the potential to improve the search on the web, tagging and bookmarking systems introduce new forms of social communication and generate new opportunities for data mining and resource sharing. However, we found that tagging systems were not very popular until 2006.

2.5.3 Multifaceted Expertise Mining

Expertise finder systems in the past have been innovatively applied in helping PhD applicants for finding relevant supervisors [Liu and Dew 2004] and also in identifying peer-reviewers for a conference [Rodriguez and Bollen]. The former made use of a manually constructed expertise profile database while the latter employed reference mining for all papers submitted to a conference. In the latter, a co-authorship network was constructed for each submitted paper making use of a measure of conflict-of-interest to ensure that papers were not reviewed by associates.

Cameron [Cameron et al 2007a] employed a manually crafted taxonomy of 100 topics in DBLP [DBLP] covering the research areas of a small sample of User researchers appearing in DBLP. They proposed the need for automatic taxonomy creation as a key issue in finding experts. Mockus et al [Mockus and Herbsleb 2002] employed data from a software project's change management records to locate people with desired expertise in a large organization. Their work indicated a need to explicitly represent experiential characterization of individuals as a means of providing insights into the knowledge and skills of individuals. Yimam [Yimam 1999] have further shown that a decentralized approach can be applied for information gathering in the construction of expertise profiles. [Tho et

al 2007] employed a citation mining retrieval technique where a cross mapping between author clusters and topic clusters was applied to assign areas of expertise to serve as an additional layer of search results organization.

There are also expertise detection systems that were based entirely on an analysis of user activity and behavior while being engaged in an electronic environment. [Krulwich and Burkey 1995] have analyzed the number of interactions of an individual within a discussion forum as a means of constructing an expert's profile. Although such an approach is useful in monitoring user participation, measures such as number of interactions on a particular topic is in itself not reflective of knowledge levels of individuals.

Information visualization techniques have been used to visualize large datasets to support exploration and in finding hidden patterns [Card et al 1999]. To visualize large hierarchal structures, the hyperbolic tree was developed by Xerox [Lamping and Rao 1996]. The principle of Focus plus Context is supported by a detailed view for the focused part of the data in the center of the display, while the overall hierarchal structure of data remains visible around the edges. In computer science, ACM categories are widely used to organize scientific work. ACM categories can be seen as a hierarchal taxonomy and can be visualized using a hyperbolic tree. To visualize experts in a proper ranking for a specific ACM category, spiral visualization is appropriate. The RankSpiral was used by [Spoerri 2004] to maximize information density and minimize occlusions for large documents. We have applied a similar approach for the visualization of experts around a particular node in the ACM category hyperbolic tree.

2.5.4 Global Discovery on LOD through Simplified Interfaces

In this section we will describe the state-of-the-art related to the search applications on LOD

2.5.4.1 URI Retrieval State of the Art

A) DBpedia

DBpedia is currently one of the most promising knowledge bases, having a complete ontology along with Yago (Suchanek et al 2007) classification. It currently describes more than 2.6 million things, including at least 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films, and 20,000 companies (Auer et al 2009). The knowledge base consists of 274 million pieces

of information (RDF triples). The openly available RDF dumps make DBpedia an interesting subject of study. There has been valuable work done on studying the reliability of Wikipedia URI's (Hepp et al 2008) that are being used by DBpedia. This study suggests that the meaning of a URI stays stable approximately 93% of the time. Its heavy interlinking within the LOD cloud makes it a perfect resource to search URIs. For our current prototype, we concentrated on the part of DBpedia that encompasses data about people.

B) Sindice

Sindice (Tummarello et al 2007) provides indexing and search services for RDF documents. Its public API allows forming a query with triple patterns that the requested RDF documents should contain. Sindice results very often need to be analyzed and refined before they can be directly used for a particular use case. Similar kinds of services are provided by semantic search engines like Falcon (Cheng et al 2007) or Swoogle (Ding et al 2004). We used Sindice in our work due to its larger indexing pool and the ease provided in use of public API.

C) SameAs

SameAs from RKB explorer provides a service to find equivalent URIs. It thereby makes it easier to find related data about a given resource from different sources.

2.5.4.2 Linked Data Consumption

A) Linked Data Browsers

The current state of the art with respect to the consumption of Linked Open Data for end users is RDF browsers (Berners-Lee et al 2006)(Kobilarov and Dickinson 2008). Some tools such as Tabulator (Berners-Lee et al 2006), Disco⁶, Zitgist data viewer⁷, Marbles⁸, Object Viewer⁹ and Open link RDF Browser¹⁰ can explore the Semantic Web directly. All these tools have implemented a similar exploration strategy, allowing the user to visualize an RDF sub-graph in a tabular fashion. The sub-graph is obtained by dereferencing

⁶ <http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/>

⁷ <http://dataviewer.zitgist.com/>

⁸ <http://beckr.org/marbles>

⁹ <http://objectviewer.semwebcentral.org/>

¹⁰ <http://demo.openlinksw.com/rdfbrowser/index.html>

(Berrueta and Phipps 2009) (Chimezie 2009) a URI, and each tool uses a distinct approach for this purpose. These tools provide useful navigational interfaces for the end users, but due to the abundance of data about a concept and the lack of filtering mechanisms, navigation becomes laborious and bothersome. In these applications, it is a tough task for a user to sort out important pieces of information without having the knowledge of underlying ontologies and basic RDF facts. Keeping in mind these issues, we suggest a keyword search mechanism to reduce the cognitive load of the users.

B) SPARQL Query Tool

Regarding the problem of searching and filtering in the Web of Data, a number of approaches and tools exist. One approach is to query a SPARQL endpoint that returns a set of RDF resources. There are a few tools that allow exploring a SPARQL Endpoint. NITELIGHT (Russell et al 2008), iSparql [Kiefer et al 2007], Explorator (Samur and Daniel 2009) are Visual Query Systems (VQS) [Catarci et al 1997] allow visual construction of SPARQL queries and differ mainly in the visual notation employed. However, in order to use these tools, the user must have comprehensive knowledge of the underlying RDF schemata and the semantic query languages (e.g. SPARQL). In summary, current tools allow users to manipulate the raw RDF data and do not provide user-friendly interfaces.

C) Faceted Search Tools

Contrary to VQS applications, Freebase Parallax [Hildebrand et al. 2006], the winner of Semantic Web challenge 2006, is based on the idea of faceted search. Freebase Parallax is a browser for exploring and presenting the structured data in a centralized infrastructure. Similar faceted search application YARS2 [Harth et al 2006] explores distributed datasets using SPO constructs. To the best of the authors' knowledge, the approach presented here is the first one that uses arbitrary data accessible via SPARQL and aggregates important facts on the basis of informational aspects.

Collaborative Knowledge Creation /Diffusion and Scientific Scholarship

Technology, beginning with Gutenberg's printing press and more recently leading to digital publishing on the Web; has always played a major role in knowledge creation and diffusion. In the field of science and technology 'publishing' in the formal journal or conference is considered as a 'hallmark of good research'. The aim of scientific publishing is to disseminate new research knowledge and findings as widely as possible in a timely and efficient manner [Hersh and Rindfleisch, 2000]. The conventional publishing paradigm of the scientific journals and paper publishing has been unsatisfactory to fulfill its promises of efficient diffusion of research. This is due to restricted journal access, rising journal costs and long delays in publication time. Above that the traditional research paper has obvious limitations in regards to the type of information that can be conveyed through such formats. Not only video and audio data can not be integrated into traditional research papers but also the huge amounts of data that may be collected in the research process can not be communicated through them.

In mid of 1990s World Wide Web (WWW) revolutionized the way in which knowledge was disseminated. The digital publishing on the Web offers the opportunity to publish new forms of data and can blur the barriers of the research group with global network effect of the Web. The Web also provides a 'global review base' for receiving feedback on research.

The latest developments in the Web termed 'Web 2.0' or 'Social Web' enhanced the open collaborative knowledge creation and its diffusion. New social web tools and applications enabled users to be the masters of their information. The unbounded number of content creators have spurred a new age of information and knowledge flows. [Kleinberg, 2004] argued that the web will bring evolution in future in the ways of scientists' work and their communication. The recent web-based open and collaborative publishing especially in wikis holds the potential to blur the boundaries of formal and informal scientific communication. Such an application called The Encyclopedia of Life is a global repository for all kinds of information related to life on earth. It builds upon the vision of Wikipedia and enhances it with Web 2.0 and semantic technologies along with a concept for assuring high quality content. The applications like the 'Encyclopedia

of Life' (EOL) have the potential to become very popular future data publishing platforms for scientists. The EOL plans to provide data access through portable devices. Any researcher working in the field of biosciences can immediately verify if he has discovered new specie by comparing the DNA scan with the dataset of EOL. The submission will be simple too as the wiki system will enable him to submit his findings at the same spot and start a new specie page in the system owned and managed by him. But this is not the only way of contributing to EOL, any citizen scientist will be able to upload and contribute any information like photos and videos of animals to their respective specie pages. The submitted information will be available online after the review process.

This chapter provides a brief description of new evolving scientific knowledge diffusion platforms and web 2.0 applications like Wikipedia and EOL along with their comparison. This chapter mainly focuses on the collaborative scientific knowledge creation and its diffusion especially in the Wikipedia and Encyclopedia of Life. Further more this work proposes a novel combination of granular tagging in Wiki systems for rapid restructuring and importing information. Considering that the lowering of editing barriers can speed up the content generation in wiki systems this chapter proposes a prototype of an application for efficient aggregation of resource snippets from diverse sources using section tagging and bookmarking to build dynamic wiki pages in Austria Forum.

This chapter addresses the following research questions.

RQ.1. How 'social software' applications of Web 2.0 like wikis, social bookmarking and tagging are leading the paradigm shift in digital scientific publishing and knowledge diffusion?

RQ.2. How can we lower the editing barriers by removing tedious copy paste requirements for content import and restructuring operations?

The research questions 1 and 2 are further subdivided into following sub-questions

RQ.1.1. What is the role of digital publishing and collaborative authoring applications in scientific knowledge creation and its diffusion?

RQ.1.2. How EOL is changing the scientific publishing and encouraging citizen scientists' contributions?

RQ.1.3. Why Wikipedia lacks confidence of the scientific community while another wiki Encyclopedia of Life (EOL) doesn't?

RQ.2.1. How social tagging and bookmarking to sub-document (i.e. section and selection) levels can provide power to users for content aggregation and restructuring in wiki environments

RQ.2.2. How dynamic wiki pages, created by snippets selected and tagged by a user, can add to rapid content creation and personalization feature.

Based on two published [Us Saeed et al, 2007] [Helic et al, 2009] works, figure 3.1 explains the progress flow for this chapter.

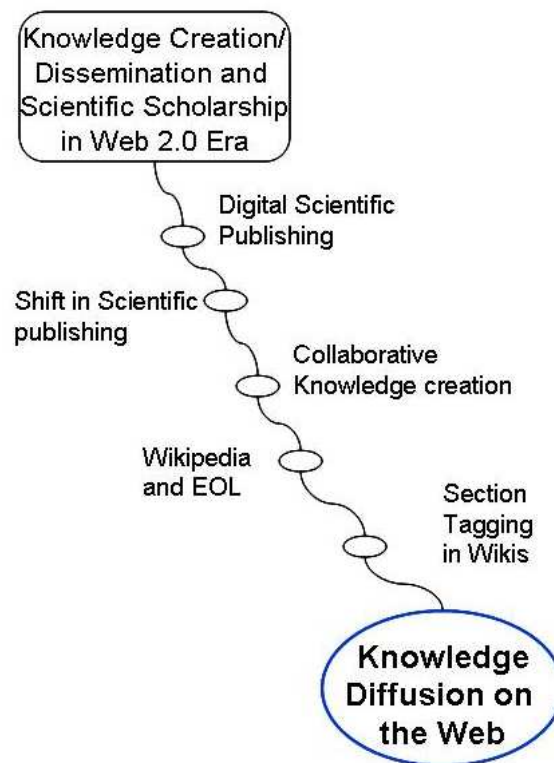


Figure 3.1: Progress Flow of Chapter 3

3.1 Digital Scientific Publishing

The term, Digital or Electronic publishing, is primarily used for online and web-based production of text and other media types. However, it is also used to describe the user interaction with regard to computer-based text and media production. Digital publishing also includes the publication of e-books and

electronic articles, as well as the development of digital libraries and catalogues [Lancaster, 1995] [Lambert, 2003]. Digital publishing has also become common in scholarly publications as the medium offers easy management of content and its fast diffusion. It is also argued that digital publishing is in the process of replacing peer reviewed paper based scientific journals. [Ng, 2009]

The traditional journals and publishing systems are not an ideal disseminating platform as they require about a year to publish an article after it is written. In this scientifically advanced era, the scientific discoveries and clinical findings are emerging at faster pace. [Odlyzko, 1994] mentioned that the growth of scholarly literature with the rapidly increasing power and availability of electronic technology, will lead the change towards the digital publishing and that the present scholarly publication system is not satisfactory. He mentioned the popularity of preprints at that time as an alternate mechanism of time stamping an innovation. He also pointed out following factors which would make a change to electronic publishing feasible in the next years.

- The predicted costs for digital publishing will be negligible compared to those of traditional print journals.
- The publications delays will disappear, and reliability of the literature will increase.
- Processor and transmission speeds are increasing at rates far higher than the growth rates of scholarly literature.

Regardless of all these predictions and the rise of digital versions of the scientific journals, the review process and delays along with rising costs of access still hinder the growth of science. The publishers are able to hold the scientific community in their monopoly because the scientific journals still play an important and unique role in quality control, archiving papers and establishing scientific credit and credibility. Traditional scholarly publishing systems, until now, have failed the academic and research communities because of their high costs and restrictive policies. These factors have resulted in limited access to information, research, innovation, academic discussion and exchange of ideas. [Ng., 2009]

The other most important issue in this regard is the copyright transfer where publishers do not pay academic authors; instead they often require authors to transfer copyright when they submit their work. All the related services like refereeing or reviewing along with paper or content authoring are provided by the scientific community free of any cost while the publishers use these services to sell back the same research to scientific community.

But now with alternative electronic publishing systems researchers have greater expectations that some of these problems will be solved. In mid nineties with the rising popularity of the World Wide Web (WWW), there was a big rush into electronic publishing with its promises of speed, efficiency and limitless accessibility. [Lawrence, 2001] provided statistical evidence that electronic publishing enabled wider diffusion of information. A number of journals have established electronic versions or even migrated entirely for electronic publication while retaining their peer review process. But still remains the access problem as digital versions apply restriction policies and rising costs for access. In recent times, several scholars and institutions have started to blame the current configuration of the publishing industry that permits commercial publishers to make money from government-funded research by restricting access to the research. This group of researchers, institutes, libraries and other such research organizations brought up the Open Access Initiative.

3.1.1 The Open Access Movement:

The main objective of Budapest Open Access Initiative (BOAI) was to accelerate the international efforts in order to make research articles in all academic fields freely available on web. The BOAI declaration proposes an alternative system of free access journals and self-archiving set-up in parallel to the commercially published journals (www.soros.org/openaccess). Open access journals are the journals that use a funding model which provide access to the readers free of any cost. From the BOAI definition of “open access”, a journal must provide users the right to “read, download, copy, distribute, print, search, or link to the full texts of these articles” (www.earlham.edu/~peters/fos/boifaq.htm).

This philosophy has been further streamered in two main routes: the gold and the green routes to Open Access. The gold road to Open Access leads to the establishment of “a new generation of journals” that do not charge subscription or access fees from readers. In these journals the author or author’s institution pay a fee to the publisher to publish a peer-reviewed research. The impact of Open Access on the economic sustainability of publishers is still an open question. The second or “green” route to Open Access states that authors should be free to self-archive or deposit a digital copy of their publication to a publicly-accessible domain.

To date, there are more than 4814 open access journals listed in the Directory of Open Access Journals (DOAJ) (<http://www.doaj.org/> accessed on March 12 2010). Below we will discuss briefly the Open Access initiatives like

Open Access Journals, Open Archive Initiative and Open Educational Resource (OER) Movement.

3.1.1.1 Open Access Journals

Open access journals provided research content freely in electronic form. Examples of open access resources/organizations are Journal of Universal Computer Science (J.UCS), Scholarly Publishing and Academic Resources Coalition (SPARC), Public Library of Science (PLoS), and Author Self-archiving.

Since its conception, open access has generated a lot of controversies among the stakeholders, especially the publishers, librarians, scientists, funding agencies and consumers. Its implications have been hotly debated [Oppenheim, 2008] [Ng., 2009].

3.1.1.2 The Open Archive Initiative:

OAI (<http://www.openarchives.org/>) or the Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. OAI has its roots in the open access and institutional repository movements. The Open Archive Initiative has substantially enlarged and improved availability and access to digital resources in various areas. The OAI provides two very important standards for data sharing named OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and OAI-ORE (Open Archives Initiative Object Reuse and Exchange).

3.1.1.3 Open Educational Resources (OER) Movement

Although learning resources are often considered as key intellectual property in a competitive higher education world, more and more institutions and individuals are sharing digital learning resources over the Internet openly and without cost, as open educational resources (OER).

The [OECD, 2010] defined OER as “digitized materials offered freely and openly for educators, students and self-learners to use and reuse for teaching, learning and research”. Such resources are accumulated assets that can be enjoyed without restricting the possibilities of others to enjoy them. This means, as shown in the figure 3.2, that these resources should be non-rival (public goods), or that the value of the resource should be enlarged when used (open fountain of goods). Furthermore, to be “open” means that the resources either provide non-

discriminatory access to the resource or can also be contributed to and shared by anyone.

3.1.2 Shift in Scientific Publishing Paradigm -- Participatory Content

Nature research article ‘Internet encyclopedias go head to head’ was the first to note the shift in scientific publishing paradigm. The article states that Wikipedia, the encyclopedia that relies on volunteers to pen its millions of entries, is about as accurate in covering scientific topics as Encyclopedia Britannica. The finding was based on a side-by-side comparison by an expert review process for articles covering a broad swath of the scientific spectrum. This brought into light the importance and marvelous rapid growth of knowledge which can be achieved by the open collaborative authoring systems like wikis at very low costs.

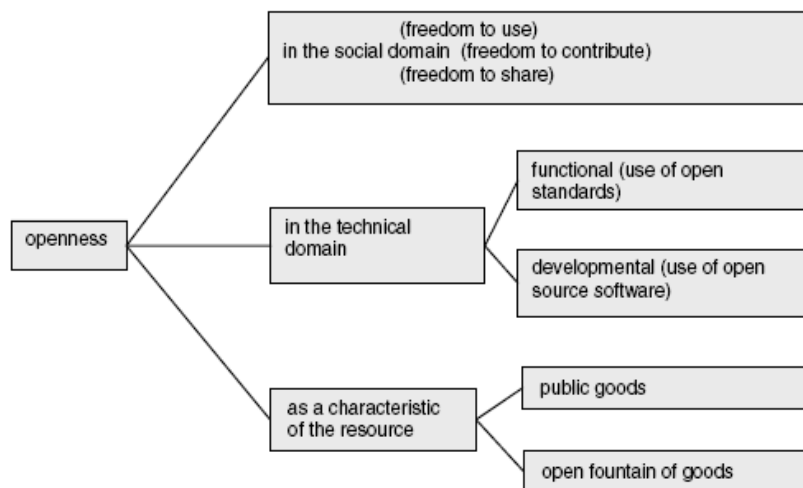


Figure 3.2: Aspects of Openness

With the advent of new participatory Web 2.0 enthusiasm, the collaborative and open way of generating, organizing, and managing knowledge has been on the rise in several fields of knowledge and life. The computer science field was the first to be affected by this collaborative revolution: free/open source software initiatives are a well known example of this. In the light of Web 2.0 principles, the WWW represent the most common platform through which people interact and collaborate in order to create, share and disseminate knowledge. [McAfee, 2006] states the Web as a social platform, adding a new layer of information interactivity based on tagging, social networks, user-created taxonomies and content. This interactivity was backed by the distributed tools and

applications aimed at supporting the collective production, sharing and maintenance of various streams of knowledge such as text, photos, and videos . While some examples of the collaborative practices of scientific communities can be observed in applications like CiteULike etc, overall the scientific content publishing seems to be still far from catching up with the new collaborative, participatory and user centered solutions. Looking into the reasons for slow response of scientific community [Cuel,R. et al, 2010] pointed out that Web 2.0 lags two soft drivers of the scientific scholarship which need to be considered

(a) the knowledge certification abilities of publishers and

(b) the need for reputation of authors

Although the Web 2.0 applications provide a socio-technical mechanism for reputation and quality assurance but it lags strong review processes present in the current practice of scientific scholarship. Because of these reasons such systems lack confidence of the scientific communities. The current review processes, on the other hand, have also been criticized for their inherently delaying and nontransparent nature [Casati et al, 2009]. It also pointed out that a significant obstacle to change is that ‘people respected in the community are successful in the current system, and hence are not very interested in changing it’.

On the other side within the Web 2.0, [Giles, 2005] mentions, the certification process of a certain piece of work is left to the auto-adjustment of the system. While this may be true for open first generation wiki systems but this is not true in a regulated wiki environment like EOL. The detailed discussion of EOL system is provided in the next sections.

The second factor ‘author reputation’ is derived from the citation index of the publisher or the publishing journal. This indexing in return gives power to the publisher to restrict the access to knowledge and charge heavier costs for access. Many researchers have shown their discontent on publishers earning money from government funded research. The result was the open access movement as mentioned above. The journal indexing systems depend on citations and research suggests that they may not be flawless [Glänzel et. al,2004] [Figg et. al,2006]. The webometrics are considered as an alternate to these citation systems which will provide popularity to a piece of work and hence to its authors. The Web 2.0 applications like CiteULike provide open metadata and databases based on tagging and bookmarking of scientific resources. These resources can be exploited to introduce new rich webometrics based knowledge diffusion and popularity indicators.

We propose that in the ecology of Web 2.0 applications a combination of wikis, personal WebPages or web-logs and tagging/bookmarking systems can provide a base platform for the future scientific publishing environments. Wikis can provide a diverse content aggregation or publishing platform with embedded certification processes while blogs and bookmarking systems can provide author reputation and publication mechanisms. The next section will compare the potentials of wikis and blogs while the role of tagging and bookmarking systems in knowledge diffusion will be discussed in the chapter 4.

3.2 Wikis Vs. Blogs---- Collaborative Vs. Expert / Personal Knowledge

Wikis and Blogs (Web logs) are two successful Web 2.0 collaborative authoring systems often mentioned together in literature. Although both are collaborative systems but they have stark differences regarding their working and applications. The scope of this chapter is limited to the collaborative publishing in wikis but we will discuss here the comparison of wikis with blogs so that their working and applications can be understood well.

The collaborative environments which have sparked the most intense interest in recent years are blogs and wikis. Blogs can be characterized as traditional, centralized, one-to-many communications, where one or few selected authors have the authority to create and edit content, and publish, “push,” or broadcast the content to a community of readers (or audience) at predictable intervals [Szybalski, 2005] [John and Walker, 2006]. This in return gives popularity and hence the value to the authors in a community. Some blogs allow readers to post comments about the content but these contributions always remain secondary to the main content, which remains central. These comments provide an opportunity for feedback and improvement in the blogging entries and process. Blogs are similar to diaries or journals, and are often used to convey a singular point of view about a given topic (e.g. science, technology, music, movies, food, etc.). Blog entries represent author’s unique personality or distinctive point of view serving a central role in defining the content and attracting an audience. Higher the readership of the blogs more popular it is. Wikis, on the other hand, are decentralized; many-to-many communications where the entire community may create, delete or manipulate content incrementally overtime [Szybalski, 2005]. Wikis are intensely collaborative, with the focus being on the development of the content, not the authors, who often remain anonymous, and they espouse a neutral point of view, or a blend of voices, through perpetual collaboration and negotiation [Fichter, 2005a].

Blogs are chronologically organized and their focus is on currency. The most current entries in blogs are always on top, displacing previous content or pushing it down, regardless of importance. Contrary to blogs, wikis are content or topic-centric, and organized according to importance, with the most relevant or important part of any article usually remaining on top and in focus [Szybalski, 2005]. This is a key distinction. Bloggers must continuously create fresh new topical content to keep their readers interested and engaged, while a wiki community may continuously make improvements to the same page, in addition to creating new content.

‘Wikis are radically different than blogs and require a fundamentally different orientation towards truth and knowledge to be successful. By simply removing the traditional author-reader relationship, knowledge-building via wikis becomes a community effort, which requires a substantial paradigmatic shift from traditional views of truth and knowledge’ [Gijsbers, 2004]. On the other hand, blogs operate much the same way as traditional learning methodologies, with an author, teacher or expert imparting his/her knowledge to the public through mainly one-way mediums or communication channels. Blogs have readers, or audiences, while wikis strive to attract participants and collaborators.

It is evident from the above discussions that blogs and wikis have different roles in the ecology of Web 2.0 knowledge sharing applications.

3.2.1 Collaborative Knowledge Creation in Wikipedia

In the Web 2.0 era of collaborative technologies, the rapid mass production of content and websites brought in mountains of partly new/redundant and distributed information which demands a great effort to develop an understanding on evolving new topics. Therefore, a need for the rapidly updating and current encyclopedic aggregation of knowledge about new concepts has never been felt stronger than in the Web 2.0 age, even in the presence of ‘googling’ technologies. Wikipedia was the first to take up this challenge and with the enthusiasm of ‘social text’ gathered more than 15 million entries in 270 languages up till now (March 2010), achieving a milestone in this regard.

The conventional encyclopedias lag behind the fast lanes of knowledge and information growth as they require periodic updating cycles or new paper editions. Wikipedia is a multilingual, web-based online collaborative encyclopedia. The entries in Wikipedia are written collaboratively by largely anonymous internet volunteers who write without pay. Anyone with internet access can contribute to Wikipedia articles. Users do not even have to register to

edit content in Wikipedia. This low cost of participation is also considered its most distinctive feature.

Wikipedia has been highlighted as a success story of low-cost collaborative knowledge systems. The openness of Wikipedia to new users has been cited as both a source of strength and weakness [Hafner, 2006]. One of its key strengths based on its open editing model lies in attracting contributions from new users who may make few edits. This suggests a kind of “wisdom of crowds” effect” [Surowiecki, 2004] in which quality of its content is derived by a large number of people making small contributions.[Kittur et. al, 2007]

Due to its open editing policy people of all ages and cultural and social backgrounds can write Wikipedia articles as most of the articles can be edited by anyone with access to the Internet. The expertise or qualifications of the user is usually not considered. This openness has also attracted lot of controversy [Helic et al, 2008]. Critics have raised concerns whether multiple unpaid editors can match paid professionals for accuracy. The question was taken up by the Nature’s research team who performed .an expert-led investigation using peer review to compare Wikipedia and Britannica’s coverage of science. The astonishing results of investigation showed that the difference in accuracy was not particularly great: the average science entry in Wikipedia contained about 4 inaccuracies while Britannica contained about three. The only major criticism brought by investigating reviewers was about readability. They commented that the Wikipedia articles they reviewed were poorly structured and confusing. Another major problem complained about is the frequent occurrence of vandalism and misinformation in new entries of Wikipedia. Due to these problems scientific communities still feel reluctant to adopt Wikipedia as a resource for scientific scholarship. Recently in 2008 the EOL another wiki encyclopedia of life sciences was built upon the vision of Wikipedia. EOL enhances the vision of Wikipedia with semantic technologies along with a concept for assuring high quality content with the expert review process. EOL has been broadly welcomed by the scientific communities.

3.3 Encyclopedia of Life (EOL)

The desire to understand life forms on our planet is not new. The success of the Genome project, ‘one of the most significant achievements of modern science’ [NPACI, 2010] and the technological advancement in biology and informatics provide the foundation for ‘a leap for all life’ the Encyclopedia of Life (EOL). It is envisioned as the first major encyclopedia of the Web 2.0 that will cover the

breadth and depth of authentic and comprehensive information as ‘a macro scope for biodiversity and an entry point into virtually all of biological knowledge’ [Patterson, 2007]. It also aims to ‘combine the authority of a traditional print behemoth with the collaborative spirit of the Web’s user-created Wikipedia’ [ScienceNews, 2007] to create a separate web page for each species on earth.

The vision of the EOL is not a new one: Already in the 1990s, Daniel Janzen from the University of Pennsylvania was among the first to address species pages. More than 10 years later, E.O Wilson articulated Janzen’s idea in his essay “The Encyclopedia of Life” [Wilson, 2003] and became one of the leading proponents of the EOL. As stated, the goal of the EOL is to serve as an online reference source and database for each and every of the 1.8 million species that are known and named today, and for those who are still to be discovered. A comparable knowledge pool has never been available to the scientific community or society before. The vision of EOL was possible only because in the recent years, crucial tools like semantic technologies and wiki-style editing have proven mature enough to be used on a grand scale.

3.3.1 Comparison: Wikipedia Vs. EOL

In this section, the concept of the EOL will be compared with the concept of the Wikipedia. Wikipedia is chosen mainly because of two reasons: It is freely available and uses a similar authoring environment. Although both EOL and Wikipedia claim to be encyclopedias, they strongly differ in their goals. In general Wikipedia aims to build a widespread base of knowledge; in contrast EOL focuses to gather all the knowledge in the field of biology, creating a repository of the expert knowledge. As a result, articles in Wikipedia are numerous covering the breadth of knowledge, but most of the time missing a detailed level, while EOL focuses on a particular topic, hence articles are expected to be on a consistently detailed level covering the depth in that topic.

The comparison focuses on the three aspects content, stake holder and technologies. Wikipedia is a grown up encyclopedia, addressing the phenomenon of mass authoring to the area of content creation in a wiki environment. Everybody may contribute to any subject in the Wikipedia regardless of his knowledge in the particular field. Wikipedia is suitable for providing an overview of a topic of interest towards a knowledge-seeker who can be anybody, including scientists.

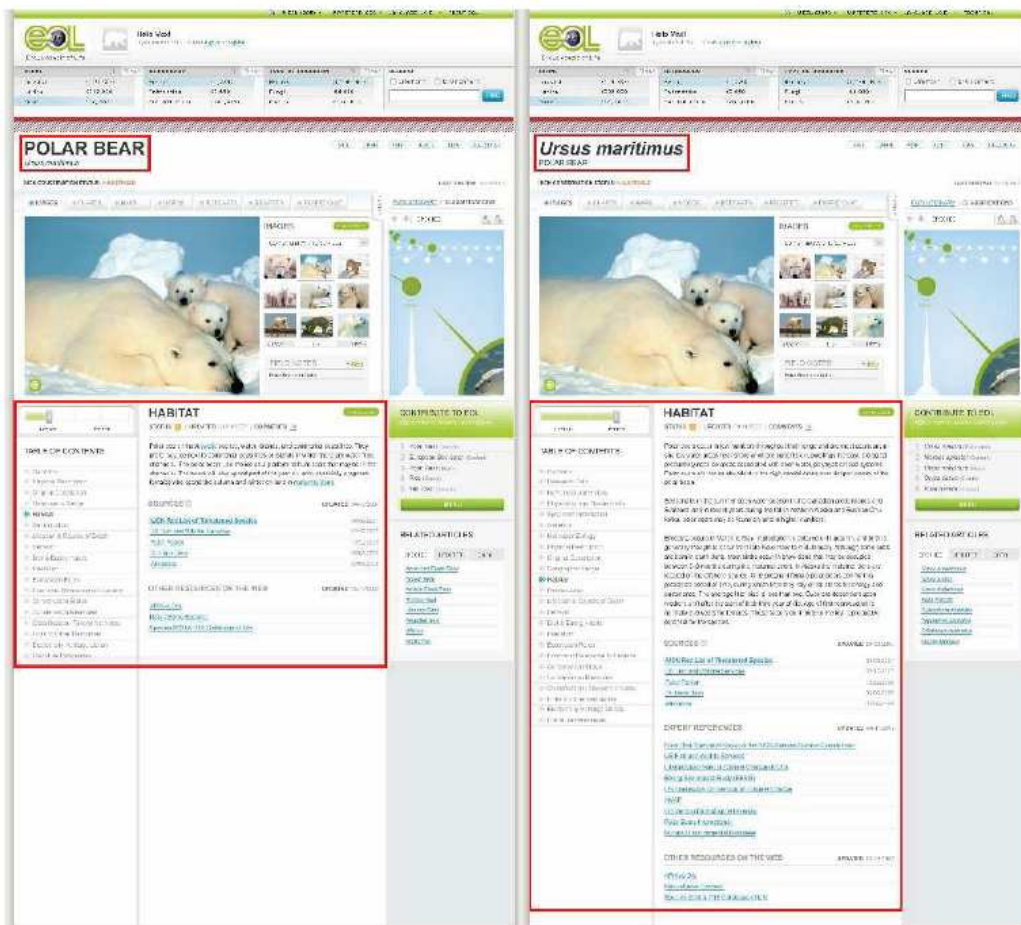


Figure 3.3: Novice and expert view [www.eol.org]

However, the usage of the content of the Wikipedia for scientific purposes is very limited, because of the lack of validity from the scientific community. In Wikipedia, plain text is dominating, multimedia content is scarce. Due to the collaborative nature of content creation, quality of content in Wikipedia is inconsistent and can easily be vandalized or falsified. Wikipedia lacks a workflow for quality assurance. A wiki-based discussion forum is aimed to support the collaboration of the authors and changes may be (but do not necessarily have to be) discussed there before they are conducted. Wikipedia also lacks personalization features and a bulk of content has to be browsed to find the relevant information on a topic.

Contrary to Wikipedia, the EOL uses a well defined workflow for information structuring and validation of content. Moreover, EOL is enriched with personalization features to facilitate end-users to organize the content in the form they like it. For scientists, motivation of publishing within the EOL is high

because, unlike Wikipedia, EOL holds the potential for reputation for the content creator. Due to the peer reviewed nature, EOL may even become a major platform for scientific publishing in biology in future.

The EOL incorporates pre-authenticated content as well as fresh content, which has to be peer reviewed by scientists, before being accessible to the public. The peer review is a formal authentication process, conducted by scientists, who are experts in the respective species. The pre-authenticated content is drawn from data-providers consisting of well-established research organizations from all over the world. Contributors for the fresh content may range from scientists to anybody with an interest in the domain of biology and biodiversity. Content in EOL will differ from content in Wikipedia regarding multimedia enrichments including images, audio and videos.

Both Wikipedia and EOL use a wiki-like environment for the creation and usage of the content. Wikipedia supports the collaborative content creation with technologies like discussion wikis for each article and a revision control to counteract vandalism. Wikipedia offers no tools for reusing its content in different environments, except a simple option to download the whole Wikipedia in a huge file. Wikipedia lacks in current technologies including Tagging, Ajax or semantic ones. The Semantic Media Wiki (http://ontoworld.org/wiki/Semantic_Media_Wiki) tries to enhance the Media Wiki, which is the underlying wiki for the Wikipedia, with Semantics. The EOL presents itself in an aesthetically pleasing way, offering vast multimedia support for the learner, out-rivaling the Wikipedia in the way the information is presented. The EOL is based on an interactive wiki-like environment. In the front-end the content elements are dynamically structured depending on the knowledge level of the learner by using a skill slider to select the expertise level. The figure 3.3 shows that the content in the novice level is more compact and easier to understand than in the expert level. When the slider is moved, both the available subtopics and the content of the article itself change according to the new skill level. News-feeds (RSS), podcasts and expert chats are provided to build a better understanding and up to date information on the topics of interest e.g. the latest scientific publications on a particular species. The EOL allows personalization of the content regarding the special needs of the learner by using bookmarking, tagging and widgets.

The EOL search is different to common search tools in the web, providing fine tuned semantic search mechanisms to cater for large and diverse set of end users. Due to the semantic algorithms, based on the underlying biological taxonomy, search is smarter and more relevant search results are retrieved. As an

example, if one searches for the term ‘habitat polar bear’ the search result presented will be the corresponding content on the habitat of the polar bear. A taxonomic map visualized as a graph will show links between the polar bear and its related species.

Contrary to Wikipedia where the content can only be dumped to a file, EOL provides sophisticated tools for reuse and mash-up of content. Based on the EOL content, modules can be developed allowing interested parties like research facilities or learning institutions to customize the interfaces or to conduct data mining according to their respective needs.

3.4 Dynamically Creating Wiki Pages Using Section Tagging

Authoring and editing processes in wiki systems are often tedious. Sheer amount of information makes it difficult for authors to organize the related information in a way that is easily accessible and retrievable for future reference. Social bookmarking systems provide possibilities to tag and organize related resources that can be later retrieved by navigating in so-called tag clouds. Usually, tagging systems do not offer a possibility to tag sections of resources but only a resource as a whole. However, authors of new wiki pages are typically interested only in certain parts of other pages that are related to their current editing process. This work describes a new approach applied in a wiki-based online encyclopedia that allows authors to tag interesting wiki page sections. The tags are then used to dynamically create new wiki pages out of tagged sections for further editing.

3.4.1 Content Creation and Information Restructuring

The popularity of social software has brought up new user generated content and metadata resources in the form of wikis, blogs, social tagging and bookmarking applications. These new systems have emerged as a major force reshaping the information spaces on the World Wide Web to better serve both collaborative and personalized information needs of users. In social software applications Web has drifted towards users’ content creation instead of the commercial content as a major contributing factor to Web resources. For instance, wikis are used for sharing, management, and organization of knowledge. Wikipedia is a user-created encyclopedia and a well known example of a wiki system. Wiki systems are asynchronous, collaborative authoring and content versioning systems where any user can add and edit content. A new version of the page is stored in the system after each editing operation [Désilets et al, 2005].

In wiki systems, user's content-creation/authoring processes involve laborious tasks like information selection from diverse resources, restructuring, modification, and adaptation of information object according to the perceived context [Nelson et. al, 2008]. The reuse of existing content in the form of copy-paste mechanisms in order to restructure and create new documents is applied by authors frequently. For example, a typical editing workflow in wiki systems involves investigating volumes of information wherein fact only small part of that information is relevant to the current user need. Thus, the user has to browse all the resources again and again to review the related pieces of information from their relevant or selected resources. This typically requires a lot of effort and time.

On the other hand resource organization with tagging and bookmarking services like Delicious, CiteULike or Bibsonomy have received community focus due to ease of use and information discovery mechanisms. In social tagging and bookmarking applications users assign free form keywords and annotations to the addresses (URLs) of an information resource(e.g., a web page) [Hammond et. al, 2005]. These keywords relate the current user context to the content of a tagged resource. The weighted set of keywords (tags) assigned to a resource by all users within a system is called the tag cloud. Tag cloud is a visual representation of tag terms in which their font is scaled according to their frequency weights.

As [Ames and Naaman, 2007] suggests the user motivation to tag a resource might be organizational or communicational on one hand, and on the other hand the users tag resources for their personal use and/or to share them with others. For example, users who tag resources for their personal use in an organizational sense use social tagging applications to organize interesting, important, and related resources according to their current needs. The tags are applied as a support for later search and retrieval of tagged resources via search or navigating the tag cloud. Typically, the tag cloud provides an overview of defined tags showing only the tags themselves but not the actual content of the tagged resources. The resources are represented via navigable links. Another motivation of using tags is to share them with other users and in such a scenario tags are typically used in a communicational sense to send signals to other users about resources that might be of interest in a more general case.

Regardless of users tagging scope- personal resource organization or sharing it with others- they have to tag the whole resource. This, however, does not always fulfill the users need. For example, users are often viewing content and are interested only in one part of the whole content. For future use users tag and bookmark it with a keyword that would be helpful later to retrieve the content. In

this case users tag the whole content with a navigational keyword useless to represent the context of resource but a useful one for them to reach the content section of their interest. This unrelated navigational tag in tag cloud will create noise. But users have no option to tag a particular interesting section within the whole resource. Such an option of tagging apart of resource may increase the user efficiency for later content retrieving, as well as help reducing noise from document tag cloud and providing a separate content-focused section tag cloud. To overcome above mentioned problems we present a novel modified social tagging approach. The benefit of such an approach has been illustrated in a wiki system on the example of simplifying the editing process. We call this new approach section-tagging as it supports users to assign keywords and annotate sections of a wiki page.

3.4.2 Prototype Application

To practically implement and test the idea, we extended the functionality of an online encyclopedia called Austria-Forum with section tagging along with the conventional social tagging. The Austria Forum was selected for its similarities with the example case of EOL. Similar to EOL, Austria forum presents an environment of a regulated wiki where the content quality is ensured by the editorial board.

The next sections describe in more details the Austria-Forum system, the idea of section tagging in Austria-Forum, how it may be used to support content retrieval, simplification of atypical editing workflow and the implementation of section-tagging idea within Austria-Forum.

3.4.2.1 Austria Forum

Austria-Forum (<http://www.austria-forum.org>) is a networked information system that manages a very large repository of information items, where new information items are easily published, edited, checked, assessed, and certified, and where the correctness and a high quality of each of these items is backed by a person that is accepted as an expert in a particular field. Consequently, each of the information items is citable as any other editorially checked content and might be used in education, scientific research, or journalism. The content of Austria-Forum is always related to Austria – as such Austria-Forum might be seen as an Austrian online encyclopedia.

In the first experimental phase of Austria-Forum the system had an editorial board of more than 20 editors and a growing community of users. The

number of users who contributed with the content was more than 100. The number of unique users who have visited the site is around 4000 each month. The current number of contributions is around 80000 (including pictures and videos as well as the content converted from the well-known Austrian cultural information system AEIOU, <http://aeiou.iicm.tugraz.at.>, visited on March 30, 2009), out of which around 6000 are user-generated contributions –approximately 8% of all contributions. Most of these user contributions are pictures and photos, with a small number of blogs, discussion forum posts, and comments. Although these numbers are quite substantial for a site that has been online experimentally a more active community involvement is desired. Community tools and facilities are already present in the system. However, as a number of users suggested, usability and a better integration of different community tools with the main system needs to be improved.

Therefore, the original system that was technically based on an in house developed content-management system has been replaced by open-source wiki software called JSP Wiki (<http://www.jspwiki.org>). The idea here is that more users will be attracted to a well-known collaborative authoring tool such as wiki. Moreover, the intention is to offer a number of community tools that will support users in retrieving information quickly and reduce the complexity of editing workflow. Among such tools is also the above presented section-tagging tool.

Even if the Austria-Forum wiki is still under development, it nearly offers ideal environment to test the concept because a huge amount of test data is available.

3.4.2.2 Section Tagging and Personalization

Section tagging is a novel social tagging approach which allows users to annotate the content of interest within a resource using free form keywords. The implemented approach differs from existing tagging and bookmarking services in the following way. First, it allows the tagging of subdocument level content. Second, tag retrieves not merely the set of links annotated by tag keyword but also the actual content of the tagged sections. Thus, when the user clicks on a tag all sections from wiki pages that have been tagged with the particular term by the specific user are dynamically loaded and presented to the user in the form of a standard wiki page.

The section of a wiki page is a self explaining piece of information about some topic of interest. Tagged content snippets in the case of section tagging have conceptual relationship to perceived structure of an information object that the

user relates to the tag terms. Hence, the context of information snippet of user's interest is more relevant to the user perception of an information object in relation to the tag terms. The underlying idea of such an approach is based on personalized content aggregation from different wiki pages because the wiki system may not hold the required information in one page but typically in various pages. Personalization in Austria-Forum refers to the content annotation and aggregation from different wiki pages according to users' intent. A typical personalization scenario involves users collecting, customizing, and modifying diverse text snippets from different wiki pages within an informational focus being described by the given tag keyword.

System offers two levels of personalization:

- Users can tag and annotate sections of wiki pages as well as full pages and hence personalize the content of interest.

- A dynamic personalized wiki page content view is created for a user by aggregating all sections tagged by him with a particular keyword. The aggregated sections are retrieved from the same versions of wiki pages which were used while tagging. The rank of a particular section within this aggregated set is determined by the frequency of same tag assigned by other users to this section.

The resulting dynamic personalized wiki page can further be collaboratively edited to create a logically complete information object reflecting the particular user context. After user has completed the editing they can publish it on the wiki where everyone can improve it further if needed. The system facilitates further the personal/collaborative knowledge creation and management. Dynamic wiki pages created by collecting snippets of information from diverse wiki pages allow users to restructure and organize information on multiple axes of personalization. Currently, the section tagging is primarily used for supporting editing workflow in the system. For example, suppose that an author is writing a new contribution on the Mozart's birth house. Before writing about the birth house the author wants to have an introductory section about Mozart that includes the basic biographical information, the list of Mozart symphonies and a picture of the Mozart monument in Vienna. The basic biographical information is included in the first section of the page on Mozart biography, the list of symphonies is described in the page on Mozart's work and the Mozart monument is depicted in the page that talks about monuments in Vienna. Thus, the author tags all the appropriate section in pages in question with a tag "Mozart". In the personal section-tag cloud the tag "Mozart" is now visible. When the author clicks on that tag a new dynamic wiki page including three tagged sections from three different

wiki pages is created on the fly. The author chooses to save the dynamically created page in the system. Now, the author can access the new page as any other wiki page and edit it by restructuring sections and adding new sections about Mozart's birth house.

3.4.3 Implementation Aspects

As described before, the core of the section-tagging mechanism is to allow users to tag not only a whole wiki page, but also to tag a particular section (identified with a heading). In this way users add semantic information to arbitrary sections of different wiki pages. In the next step, it is possible to extract sections referred by a particular tag and to create a new personalized wiki page out of tagged content snippets. The implementation of the section concept is comprised of two functional modules, called Section-Tagging (ST) and Personalized-Content-Creation (PCC) module. The JSP Wiki system is based on a clean and extensible plug-in and filter architecture that allows easy addition and configuration of new modules. The filter mechanism allows on the fly parsing and modifying of wiki pages before they are rendered.

On the other hand, the plug-in mechanism allows server-side code to be referenced from within a wiki page. This code dynamically produces wiki content that can be included in the wiki page that refers to the plug-in. Thus, technically the ST module is a filter module as it inserts section-tagging functionality into already existing wiki pages by pre-processing them; the PCC module is a plug-in module that dynamically creates a new wiki page according to the selected tag and the tagged sections from various wiki pages.

3.4.3.1 Section Tagging Module

ST module is a filter for pre-processing of rendered wiki pages. This unit is responsible for extending document object model (DOM) of a rendered wiki page via a JavaScript module called ST form module. As shown in Figure 3.4, this module supplies a simple to use pop-up form (red colored box in front of section) that visualizes particular semantic section information by an onmouseover effect and letting the user tag a section using the onclick event. Moreover the ST form module also supplies the database connector module with information about the currently tagged section number and page version.

The actual centerpiece of the ST module is a unit called ST plug-in. It loads and manipulates the data from the ST data storage backend module, extracts

user data from the ST security module and handles data sent by the ST form module via XMLHttpRequest (see Figure 3.5).

As a data storage module the open-source content-management system Scuttle (<http://sourceforge.net/projects/scuttle>) is deployed. The database itself is not accessed by the API which the system offers but by the database connector module which extracts user data such as username and IP address directly from the JSP Wiki user session module. This user data record is stored together with a special section URI to the Scuttle database by the plug-in module every time a section is tagged by the user, in order to guarantee an unambiguous relationship between user and tagged sections.

In order to have a clear relationship between page sections, page versions and corresponding tags and still offer a readable URI without changing the database structure itself, the well known(X)HTML method of creating links within a hypertext document was adopted in the following form:

`http://<URI>#<section ID>_<version>`

Thus a section of a wiki page can be easily addressed to a tag and vice versa by adding a fraction identifier holding information about the section ID (<section ID>) and page version (<version>).

3.4.3.2 Personalized-Content-Creation Module

The PCC module is implemented as a plug-in that can be included in any wiki page. Currently, this module is included in a personalized wiki page that is shown on the right-side of the user screen. It shows a standard tag cloud with tags assigned by a particular user to wiki page sections of interest. When a user clicks on a tag the PCC module retrieves all tagged sections using the appropriate wiki page versions. The sections are then dynamically combined into a wiki page that is shown to the user. The user has then the possibility to edit and modify this new wiki page using the standard wiki editor and to save the editing operations in a completely new wiki page for later retrieval.

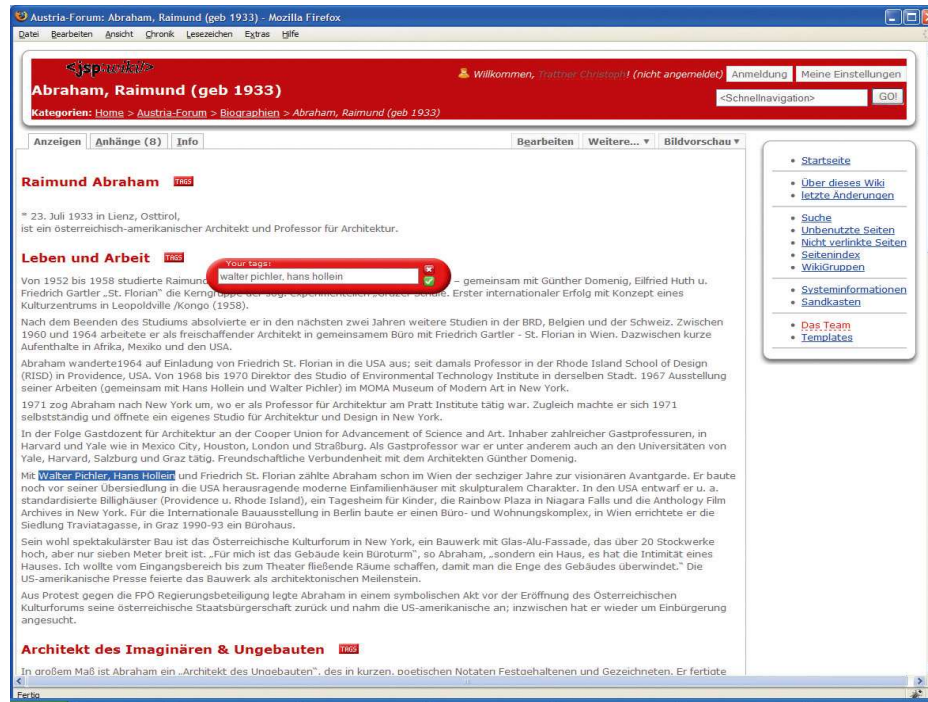


Figure 3.4: ST form module

Moreover, the dynamic page can be still retrieved at all times by simply clicking on the appropriate tag. Note that the dynamic page is always created on the fly, thus whenever the user adds tags to sections of some other wiki pages this will be reflected in the dynamic page as the page will include the new sections.

3.5 Concluding Remarks

By lowering, and often removing the technical barriers of entry and participation, wikis can not only drastically reduce the cost of knowledge creation and management, but they can also vastly improve the process of creating and disseminating information from the bottom-up, where the community itself creates, organizes and disseminates the information that it wants and needs [Dickerson,2004].

The chapter presented an overview of open Wikipedia and regulated EOL Wiki environments regarding scientific scholarship. A novel approach for tagging sections of wiki pages has been presented which lowers the barrier of content editing, restructuring and importing new content from other pages. This approach is able to personalize the users' content in an efficient way. It has reduced the manual effort required to author a wiki-page about a topic. Often, the wiki system may not have the required information in one page but typically in various pages.

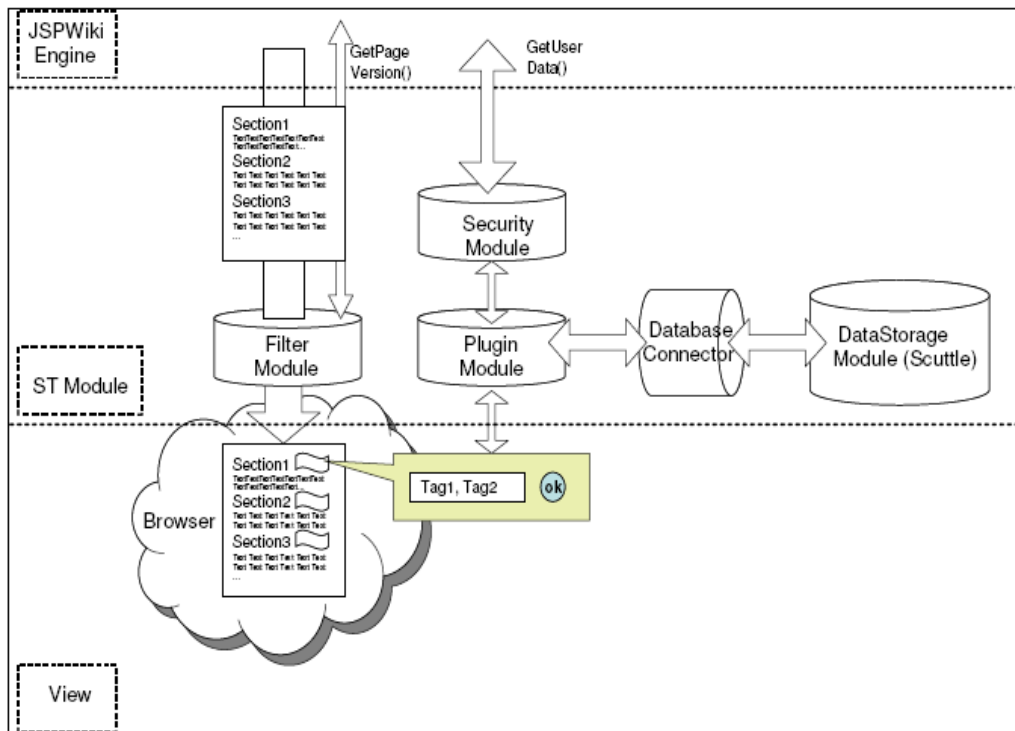


Figure 3.5: Architectural diagram of the ST module

Therefore, a combination of the social tagging approach with the wiki concept in an innovative manner facilitates an easy retrieval of the relevant content in the form of a new dynamically created wiki page. Such dynamic wiki pages created by collecting snippets of information from diverse wiki pages allow users to restructure and organize information on multiple axes that best fit their current needs.

It is assumed that the tool (granular tagging, content import and personalization plug-in) will provide rapid content creation and will add to the fecundity of Wiki environment. Higher the growth of wiki knowledge base greater will be the diffusion. As this is a prototype application and the scope of tagging and aggregating content is within the wiki environment, it proves the implementation concept only. The true power of importing and coalescing content for rapid knowledge creation will come with the extension of this approach to selection tagging browser plug-in.

Future work, in this regard, will be extended to implement and study the following:

- Interesting aspects of global section-tag clouds will be the tag and section selection strategy in the case that there are numerous sections tagged by a particular tag. A collaborative filtering approach taking into account the user profiles might be needed to limit the sections only to those that are most relevant.
- Extending the section-tagging approach to arbitrary web resources with selection tagging plug-in. This can be implemented as browser plug-in in future which will gather the tagged content in a dynamic wiki system as a web-based service. This will increase the diffusion of content from the Web to the Wiki.

Diffusion of Knowledge using Bookmarks and Tags

Social bookmarking and tagging services are popular web-based systems that allow users to share, classify, and discover interesting resources on the web. Recently, such applications are gaining high popularity in scientific communities. The applications like Bibsonomy (www.bibsonomy.org), CiteULike (www.citeulike.org) and Connotea (www.connotea.org) are some examples of such systems. This chapter explores the potential of tagging and bookmarking systems to indicate diffusion of knowledge. It further probes their similarities to citations which are a conventional measure of diffusion of knowledge. The following research questions are addressed in this chapter.

RQ.1. Does tagging/bookmarking indicate knowledge diffusion?

RQ.2. What similarities do exist between tags and citations and how can we use them?

The research questions 1 and 2 are further subdivided into following sub-questions

RQ.1.1. Is there any positive correlation among bookmark counts and citations?

RQ.1.2. Can citation rank be predicted from bookmark count rank?

RQ.2.1. What information do tag terms hold about citations?

RQ.2.2. How the important and related resources in bookmarking systems can be linked to other scientific resources in digital scientific repositories.

The figure 4.1 explains the progress flow for this chapter based on multiple published works. The parts of this chapter are published in two conference [Us Saeed et al 2008a] [Us Saeed et al 2008b] and one journal paper [Us Saeed et al 2010]. [Us Saeed et al 2008a] study the statistical correlation among bookmarks and citations. It also discovers that tag keywords appear

frequently in citing titles. [Us Saeed et al 2008b] compare a citation rank prediction from bookmarks and coauthor network. [Us Saeed et al 2010] propose a tag recommendation system for scientific resources.

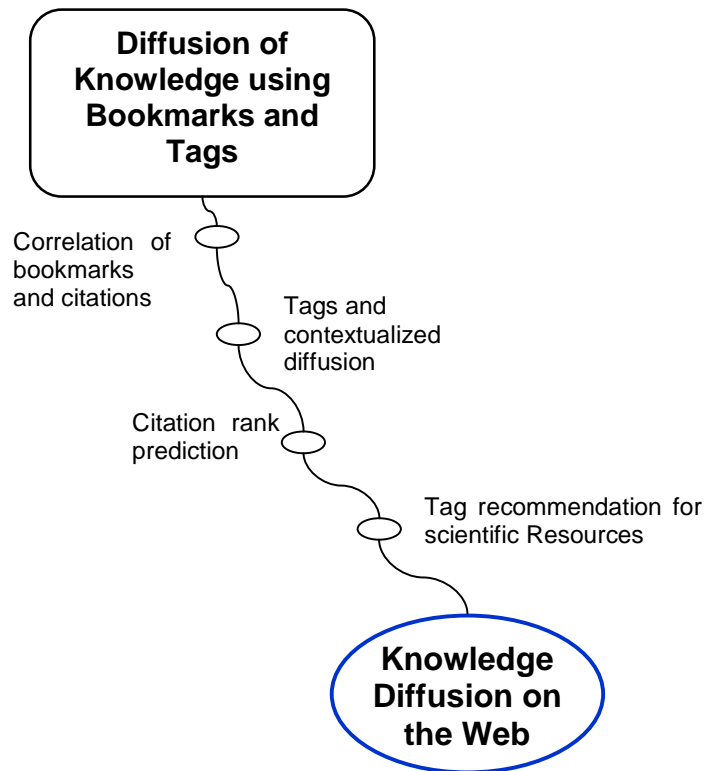


Figure 4.1: Progress Flow for the chapter 4

4.1 Introduction

Knowledge is of prime importance for economic and social development. The diffusion of knowledge holds an important role in the creation and distribution of knowledge boons. The diffusion of published (codified) scientific knowledge has been mainly investigated in the past to study the structures and properties of knowledge diffusion in scientific domain. In science and technology citations are considered as an indicator for volume of diffusion of a published work. Citation is a relationship between two published papers or articles where normally the author(s) of ‘citing’ paper infer(s) from and refer(s) to the part of ‘cited’ paper used to extend or create new knowledge published in the ‘citing’ paper. Citations are also used to measure the impact of research. It is considered that, to some extent, collaborative behavior may affect the citations of a paper or an article. Usually researchers collaborate and jointly report in their research publications. The new ideas and findings of research are established after conversations among

them. When more than one authors share a published work, they are called coauthors. Co-authorship analysis and citation analysis are the popular techniques used to assess diverse aspects of knowledge, in science and technology. Knowledge diffusion in general is analyzed using diffusion of innovations, epidemiology, collaboration network analysis (co-authorship analysis) and citation analysis techniques.

In addition to the study of the diffusion of (codified) scientific knowledge through citations, the need of web based indicators for assessment of different aspects of science and technology has also been pointed out in [Scharnhorst and Wouters 2006] [Day 2008]. The latest developments in the Web termed ‘Web 2.0’ or ‘Social Web’ has provided access to open source data and metadata resources. Kleinberg argues that the web will ‘bring evolution in future in the ways of scientists’ work and their communication’ [Kleinberg 2004]. Furthermore, the recent trends of contributory web and inflated web-based publishing have the potential to blur the boundaries of formal and informal scientific communications. The applications like the ‘Encyclopedia of Life’ (EOL) may become very popular future publishing platforms for scientists [Us Saeed et al 2007]. Every day, the research work is getting more and more convoluted with the emerging structures of web. It is feared that the dynamics of diffusion of scientific literature on the web in future may not be assessable by conventional techniques alone. This emphasizes the need for a particular type of web indicators, one of which may be bookmarking /tagging , which are within the streams of this new form of web evolution. The research in this chapter intends to explore the potentials of these bookmarking applications in the diffusion of knowledge and its estimation. Tagging practices have an added advantage to augment the understanding of knowledge diffusion by providing an additional element – the user context in tagging a resource of knowledge (to understand the better reason about the usage of knowledge).

Initial probing shows that bookmark counts in CiteULike mines the interest of researchers in a particular scientific resource. The bookmark counts are correlated positively with the citations of that resource. This result can be used to establish the popularity and hence citation count or quality of that resource. The research also concludes that the tag terms assigned by users to a particular scientific paper of WWW‘06, in social bookmarking applications, frequently re-occur in the titles of its citing papers. This shows that tag terms hold the context of diffusion of a scientific research.

4.2 Social Bookmarking and its Potentials in Measuring Knowledge Diffusion

Social bookmarking and tagging has become a very successful phenomenon in the web and getting more popular day by day. Systems adhering to these principles transform the way in which users manage and disseminate content in the conventional web environment. These systems enable the users to add keywords (tags) to web resources (web-pages, images, documents, papers) without having to rely on a controlled vocabulary [Marlow et al 2006]. It's potential to improve the search on the web, resulted in new forms of social communication and generated new opportunities for data mining. This research probes bookmarking and tagging as a medium to measure the knowledge diffusion. The past research identifies the inability of tagging systems to have control on the users for specifying relevant tags to the resource and handling manipulation of these tags to various contexts. One approach adopted here in the proposed recommendation system for filtering the tags with the author key words as seeds can also be effective to resolve this vocabulary problem.

In the fields of emergent semantic [Mika, 2005], Information Retrieval [Wu et al. 2006], [Hotho et al, 2006] and user profiling [Huang et al, 2008] tagging is considered as a driving component [Michlmayr et al. 2007]. Information retrieval and textual mining is already being used in many decision making systems, such as in the case of medical sciences [Holzinger et al, 2008]. Tagging can also be helpful in these systems as 'in a collaborative tagging system, tags codify the knowledge of relationships among documents and concepts represented by the tags. Harvesting individual through folksonomies therefore can benefit the whole society' [Wu et al 2006]. Mika [Mika 2005] has studied the tagging behaviors and their usage in del.icio.us, an emerging bookmaking service. He used actor, concept, and instance nodes as a tripartite graph to explain the emergence of ontologies from social context where he considers tags as a socially represented concept.

This study intends to compare the tagging behaviors with the knowledge diffusion mechanisms and their corresponding contexts. It is also used for effective tags extraction and resource recommendation for scientific papers. Literature has shown that 'context' became an important consideration in any discussion of codified knowledge [Cowan et al 2000]. However, in previous works there were very limited explicating instances about the usage of context in diffusion studies. For example, Tsai described the contextual flow of knowledge within scope of an organization [Tsai 2001] and Chen used context in the

geospatial distribution of diffusion [Chen et al 2007]. Heterogeneity of context in reuse of knowledge implies the need for an indicator in which the constituent parts can be rendered commensurably. Tags may augment the context of the knowledge being used by different users [Wu et al 2006]. The Figure 4.3 in the next section shows that how tagging can be used to identify the diffusion context.

Previously many constructs has been employed to measure the Knowledge diffusion, one of the popular and important one is Citations. Citations are studied in different ways like scientific fronts, a service provided by ISI since Feb 2008 which performs a co-citation analysis within different subfields of a broad subject. They built subfields by extracting keywords from titles of highly co-cited papers. But there is a lack of a standard taxonomy for a particular field. For example if someone want to study subfields for computer science, one may suggest that ACM standard taxonomy can be used, but research has shown that a large amount of documents in digital libraries are not categorized according to this taxonomy and then mapping of papers to this classification becomes problematic when the paper is not explicitly stated into a particular category which is the case in most of the papers [Cameron et al 2007]. Previous research showed that there are certain limitations of citations like 1). citations of existing papers do not necessarily mean that the cited-by paper is regenerating knowledge by using knowledge from the cited papers 2) Citations inability to highlight the real context of the citing paper for example citations are made to just give a broad level background study and the context of cited paper is not always clear by reading the citing paper. 3) Citation analysis may not always predict the contextual use of the knowledge 4) Limitation of citations to just understand the codified knowledge. For example in the case of applied research, knowledge is not often used to create new knowledge, thus receives a fewer citations but is used practically in various fields. This knowledge for practice, however, cannot be measured by citations.

By taking these limitations in account, this research has proposed that bookmarking/tagging got a potential to be used as a supplementary measure in predicting and estimating the contextualized knowledge diffusion. Tagging may explicate the contexts of diffusion in a more convincing way as compared to citations because tags are explicitly specified by the users in their own context when viewing a particular paper. For example a user tags a particular paper most of the time as “Web 2.0”, but at the same time other contexts of users for that particular paper will also be a part of its tag cloud. As investigated by Mika, these tags and their proportional percentages can be used to make an automatic taxonomy [Mika 2005].

This work explores the potential of bookmarking and tagging with safe assumption, that people tag something: 1) if they conceptually understand the content and 2) if they perceive it to be useful in their own context (of work).

4.2.1 Empirical Relationship (Bookmarks/Tags vs Citations)

The exploratory case studies [Us Saeed et al 2008a] [Us Saeed et al 2008b] were performed to find potential of tagging and bookmarking systems. The published 84 papers of the conference World Wide Web 2006 (WWW'06) were analyzed. The WWW'06 was selected as a dataset because of its special focus and popularity. Papers presented at the WWW conference series generally discuss the future evolution of the web. That's why, the expectation was to find WWW papers both frequently cited and tagged in social bookmarking applications. The higher numbers of citations show the large scale of volumetric knowledge diffusion and high impact of scientific resources. The citation ranks for research papers are usually predicted using various factors. These factors include multi-author publications, geographical positions of co-authors, co-authors' network, and multi-institutional involvement in a publication. However, with the evolution of the Web 2.0, bookmarking and tagging applications may provide a popularity measure for scientific resources. As the focus of study was to compare different citation prediction models, a dataset of research papers from a conference which is popular and within a particular focus related to the web was selected (so that the potential research community is already integrated within the bookmarking systems). Considering all these factors, World Wide Web conference was selected.

The event from the year 2006 is taken, because tagging applications were not popular before the year 2006. The assumption was that a certain degree of popularity would be required for representing real tagging behaviors. The event from 2007 or 2008 was not select because normally it takes 1-2 years to enable the regeneration of the new knowledge.

The selected papers are explored in three common social bookmarking and tagging systems CiteULike¹¹, BibSonomy¹² and Del.icio.us¹³. Although BibSonomy and Del.icio.us give access to their search APIs, yet our initial experiments showed that searching a particular paper which have some special

¹¹ <http://www.citeulike.org/>

¹² <http://www.bibsonomy.org>

¹³ <http://del.icio.us/>

characters (like : , - _ ‘ “ & vs. / etc.) in its title does not find its match in the tagging application. It was found that sometime the same user (who tags a resource) is listed repetitively for one paper in these applications. It was also found that sometimes same user tags the same paper with different tags in different times. This leads to miscount of the total number of users for a paper. By considering all of these limitations, the bookmark counts, tags and the users in these applications were safely explored. Citations were acquired from Google scholar¹⁴ manually because Google Scholar does not provide open access API to explore the citations. The dataset was tabulated year wise from bookmarks/tags and citations with the paper numbers as ‘ids’ and their titles extracted from WWW’06 website¹⁵. The ids are maintained in the order of paper titles listed on the website. Figure 4.2 depicts various modules of the study design for the research.

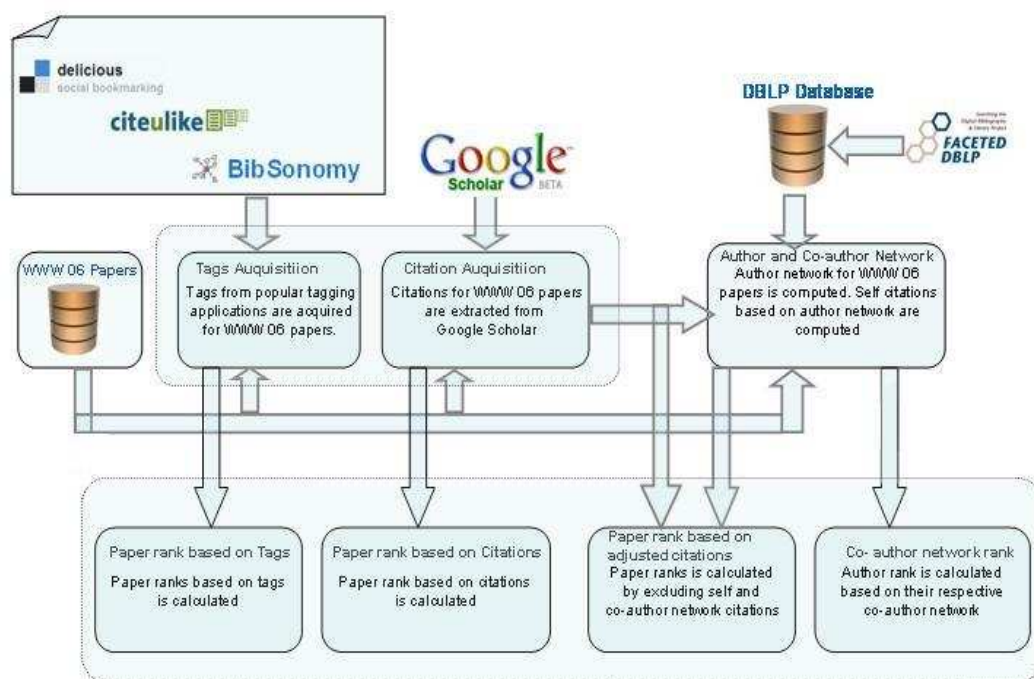


Figure 4.2: Modules of the study design

¹⁴ <http://scholar.google.com/>

¹⁵ <http://www2006.org/>

The next section explains how the data sets for bookmarks, citations, co-authors' network were acquired prior to computing different citation prediction models.

Tags and bookmarks for WWW'06 papers were collected from the CiteULike, BibSonomy and Del.icio.us based on their popularity in the Web research community. The total bookmarks for the 84 papers were 1051. Citations for WWW'06 papers were acquired using Google Scholar. Although Google Scholar does not provide a search API for citation extraction, but Google Scholar was chosen because of its large index. Google Scholar index covers "peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations" [About Google Scholar 2009]. Google Scholar also finds some false positive citations like citations to press releases, resumes, and links to bibliographic records for cookbooks [Price 2004]. But all citations for WWW'06 papers were extracted manually. The total citations for the 84 papers were 1165.

4.2.2 Author's and Co-authors' Network

The citation rank studies are usually based on co-authors' network. In this study the citation rank for WWW'06 papers was computed based on number of bookmarks and co-authors' network. To build a co-authors' network, a dataset of DBLP++ [Diederich et al 2007] was selected. This is an enhanced dataset of DBLP (a digital library for computer science publications). DBLP indexes WWW'06 conference in particular and contains 1,048,576 publication records in general. DBLP is managed manually. Due to this, it does not include the inherited problems of autonomous systems. This module performs four tasks:

- 1) Finds authors of papers of WWW'06 conference.
- 2) Finds citing authors for all papers of WWW'06.
- 3) Computes a co-authors' network based on the original authors of the paper. The co-authors' network is computed up to 2 degrees of separation. The average co-authors' network for WWW'06 authors was 119.
- 4) Computes self citations and citations by a co-author's network.

There were 1165 overall citation for WWW'06 conference papers. Self citations were 208. Citations, in the first level co-authors' network, were 60 and citations, in the second level co-authors' network, were 26. These figures also indicate that self citations and citations in co-authors' network (up to 2 levels) accumulatively were only 25% of all citations.

4.2.3 Findings from the Study

- **Bookmark counts positively correlate to citations**

It was found that a positive correlation ($r=0,65$, $p=2.133 \text{ e-}11$) exists between the total number of bookmarks and the total number of citations from May 2006 to May 2008 for all the papers. This finding indicates that the bookmarking and tagging behavior somehow matches with the citation behavior.

- **Bookmarking may have the potential to foretell the future volume of knowledge diffusion**

The average number of users, in table 4.1, was calculated by adding all the users from three tagging applications for a particular paper and dividing it by three (i.e. number of tagging applications). It was observed that if the average is higher than 6, then the tagged paper also gets reasonable number of citations (>6). See table 4.1. For such papers the major number of citations came from the year 2007. However, for the same papers, the major number of user's bookmark counts came from the year 2006.

This is logical, because the bookmarks/tags will come earlier in time than the citations. The regeneration of knowledge needs more time than the selection of a piece of knowledge. This makes the case interesting for tagging analysis, because it shows a possible potential of the bookmark counts to forecast the future volume of knowledge diffusion.

- **Tagging may have the potential to foretell the context of future knowledge diffusion**

A lightweight tool was developed to create tag-clouds. Using this tool, two tag-clouds for each paper were created: 1) Tag-cloud of the tag terms from all tagging applications. 2) a second tag-cloud was generated by selecting the matched tag terms of first tag-cloud in the titles of the respective citing papers. The font size of second tag-cloud is assigned on the matching frequency of the terms in the titles of citing papers. The trend for heavily tagged and cited papers is visualized in Figure 4.3.

Table 4.1: Heavily bookmarked papers in 2006 got heavy citations in 2007

| Paper ids | Avg. No. of users per tagging application (>6) | Total user bookmark counts (06) | Citations in 2006 | Citations in 2007 | Total Citations |
|-----------|--|---------------------------------|-------------------|-------------------|-----------------|
| 9. | 7 | 7 | 11 | 44 | 61 |
| 10. | 8 | 20 | 3 | 6 | 12 |
| 17. | 9 | 13 | 4 | 11 | 18 |
| 23. | 49 | 80 | 9 | 37 | 49 |
| 24. | 11 | 18 | 5 | 15 | 23 |
| 25. | 7 | 14 | 1 | 19 | 23 |
| 31. | 7 | 7 | 1 | 7 | 8 |
| 50. | 40 | 100 | 10 | 24 | 43 |
| 51. | 32 | 37 | 4 | 32 | 39 |
| 69. | 30 | 41 | 34 | 68 | 112 |
| 73. | 21 | 21 | 5 | 24 | 33 |

The results showed that about 16 to more than 22 percent tagged terms matched with the title terms of the citing papers. This result is in line with our assumption that tagging may forecast the context of knowledge diffusion.

4.2.4 Paper Rank Models

Bookmarks, citations and co-authors' network are further used to establish different models for paper rank.

a) Paper rank based on bookmarks

This model ranks papers based on their popularity on Web (tagging and bookmarking applications), the number of users who bookmarked a paper are aggregated from different applications to form a total user count for a particular paper. The large number of users ranks a paper on top in this model.

b) Paper rank based on citations

This model ranks papers based on their citation counts. The extracted citations are used to rank paper in this model. The high number of citations ranks a paper on the top in this model.



Figure 4.3: Tag cloud comparison of heavily cited and tagged papers.

4.2.4.1 Citation Ranks Prediction Models

Based on the collected bookmarks, citations and co-authors' network for WWW'06 conference papers, citation rank model was explored by applying different variables and then comparing the results. Linear regression analysis was applied to find out relationship. Linear regression is a form of regression analysis in which the relationship between one or more independent variables and another variable, called dependent variable, is modeled by a least squares function, and represented by a Linear Regression (LR) equation. The details of citation rank model based on different variables are depicted below.

a) Citation rank prediction model based on bookmarks

In this model bookmarks are used as an independent variable while citations are taken as a dependent variable. The linear regression equation model is as follows:

$$0.69 * \text{variable (bookmark - rank)} + 6.21 \quad (1)$$

In the model equation (1) 0.69 is called the regression coefficient. It explains the behavior of change in the value of dependent variable for small change in bookmark rank. The term 6.21 is called the disturbance or noise term.

b) Citation rank prediction model based on co-author network

In this model co-author's network (calculated in section 4.2.2) is used as an independent variable while citations are taken as a dependent variable. The linear regression equation model is as follows:

$$0.46 * \text{variable (coauthor rank)} + 30.27 \quad (2)$$

In the model equation (2) factor 0.46 is called the regression coefficient. It explains the behavior of change in the value of dependent variable for small change in co-author counts. The term 30.27 is called the disturbance or noise term.

c) Citation rank prediction model based on adjusted citations

In this model bookmarks are used as an independent variable while citations are taken as a dependent variable. The citation counts are adjusted by excluding self citations. The linear regression equation model is as follows:

$$0.69 * \text{variable (bookmark rank)} + 6.85 \quad (3)$$

In the model equation (3) factor 0.69 is called the regression coefficient. It explains the behavior of change in the value of adjusted citation rank for small change in bookmark rank. The term 6.85 is called the disturbance or noise term.

The correlation coefficient established on WWW'06 papers by bookmarking count model is 0.6003 which is considered as a fair correlation, while it is 0.1559 by co-authors' network model. This is not so good. This correlation coefficient is enhanced up to 0.6657 by excluding the self citations

The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. It was 5.3727 by bookmark model while this mean error was much higher (18.1428) in co-authors' network. This error is reduced up to 4.3821 with the self citation adjustment.

Our results have proved that citation rank prediction based on bookmark ranks of papers have got fairly good results than co-author network model (see

Table 4.3). The citation loops like self citations are considered in this research (see Table 4.2). This furthermore improves the correlation coefficient and reduces the mean absolute error (see Table 4.4). However, these results are obtained for WWW'06 conference papers and further studies are necessary to their generalization.

Table 4.2: Top 5 Ranks of Papers with respect to bookmarking and their respective other Ranks

| Paper ID | Bookmark Rank | Citation Rank | Adjusted Citation Rank |
|-----------------|----------------------|----------------------|-------------------------------|
| 23 | 1 | 3 | 3 |
| 50 | 2 | 5 | 7 |
| 51 | 3 | 6 | 5 |
| 69 | 4 | 1 | 1 |
| 73 | 5 | 7 | 6 |

4.3 Linking Contextual Resources from CiteULike Using Tags

In previous sections, it has been shown that there exist a positive correlation between bookmark counts and citations. Citation count also inflates diffusion by increasing popularity of research and is considered as an indicator for establishing the quality of research. On the other hand, researchers use citations or references to search the connected and related resources hence increasing diffusion of interlinked knowledge.

Table 4.3: Top 5 Ranks of Papers with respect to bookmarking and their respective citation Ranks

| Paper ID | Paper Rank based on coauthor count | Citation Rank |
|-----------------|---|----------------------|
| 49 | 1 | 6 |
| 23 | 2 | 3 |
| 50 | 3 | 5 |
| 69 | 4 | 1 |
| 65 | 5 | 26 |

Table 4.4: Comparison of citation prediction models based on LR

| LR | Prediction model based on bookmark rank | Prediction model based on Co-author network | Prediction model based on adjusted citations |
|-----------------------------|--|--|---|
| Correlation coefficient | 0.6003 | 0.1559 | 0.6657 |
| Mean absolute error | 5.3727 | 18.1428 | 4.3821 |
| Root mean squared error | 6.6213 | 20.8102 | 5.5976 |
| Relative absolute error | 75.6676 % | 99.4605 % | 71.1488 % |
| Root relative squared error | 79.9746 % | 98.7775 % | 74.6248 % |
| Total Number of Instances | 84 | 84 | 84 |

Based on these potential uses of citations in the research community and results of past research, which shows that citation count and bookmark counts are positively correlated, it is argued here that bookmark counts of research papers can be used in a similar way as an alternative popularity indicator. The next section also proposes a tag based resource recommender system for scientific papers. These contextual tags provide a link to the most related resources which gives two benefits: 1) these resources will be directly related to the content and context of diffusion of that paper which is implicitly derived from the tags extraction mechanism 2) the researcher can explore the interlinked and related resources using tags (hyperlink) as they use references or citations.

The tag based scientific paper recommender system exploits author keywords of scientific publications to link these resources with tags in CiteULike which is a social bookmarking and tagging application. For a focused resource, the tags extracted from CiteULike based on author keywords were compared with the corresponding tag cloud of CiteULike. The result shows that system extends the authors keyword set with social tags providing links to rich and focused resources in CiteULike. This also enhances the serendipitous discovery of emerging concepts related to that resource. Such a system may enhance the discovery of related and popular resources for researchers hence furthering the diffusion of knowledge.

This section explains how scientific papers can be linked to relevant resources (tags and papers) for papers published within digital journals or libraries. For this exercise, WWW'06 was taken as a source data set. The social bookmarking system used in these experiments was CiteULike. CiteULike is a social bookmarking system where a huge number of users share scientific papers and tag them accordingly. Major task is to find the most relevant resources from CiteULike for all papers published within WWW'06. On the WWW'06 side, every paper is assigned with suitable keywords by the authors of the paper, while on CiteULike side, papers are tagged with some keywords by the users of the CiteULike. To find relevant resources for WWW'06 papers, authors' assigned keywords were used to mine the tags from CiteULike. The papers at WWW'06 are further annotated with the matched tags.

4.3.1 WWW'06 dataset

This dataset is comprised of all published papers in the conference World Wide Web 2006.

Total papers published in WWW'06 = 84

Total Keywords for all papers = 5129

Unique Keywords = 107

4.3.2 CiteULike dataset

The dataset of CiteULike was acquired in August, 2009. The statistics for tags and papers is shown below.

Total tag assignments in CiteULike = 6.5 million

Total Papers in CiteULike = about 2 million

Unique tags = 348420

4.3.3 Matching Author's Keywords with CiteULike Tags

To match papers' keywords of WWW'06 with CiteULike tags, a two-tier approach was adopted. First, the tag extraction tries to find an exact match between papers' keywords and CiteULike tags. Subsequently, a partial match between both datasets was checked. The partial match enhanced discovery of

relevant tags but also introduced some noise. Afterwards, some heuristics were used to clean the noise and the discovered tags were used to annotate the corresponding papers.

1) Direct Match

WWW papers for which at least one keyword is matched = $52/84 = 62\%$

Unique Keywords of WWW'06 matched = $102/107 = 95\%$

2) Partial Match

WWW'06 Papers for which at least one tag is matched = $52/84 = 62\%$.

Total results of WWW'06 Keywords matched with CiteULike = 5129

Total CiteULike unique tags matched = 4228/348420

In the direct match, the system found one exact tag from CiteULike for each of 102 unique keywords of WWW'06. The knowledge discoveries are significantly enhanced by employing partial match. The partial match found a total of 5129 matching tags from CiteULike. This becomes a basis for recommending relevant tags for the focused paper. The partial match enhances the system discoveries significantly for example; the author keyword 'visualization' was found and matched in the related popular concepts (GeoVisualization, DataVisualization, NetworkVisualization, SoftwareVisualizatuion, GraphVisualization, TreeVisualization, etc).

4.3.4 Recommending Relevant Tags for Research Papers

The contribution of this research can be structured into two aspects: 1) discovery of focused set of tagged resources in social bookmarking applications 2) leads to serendipitous discoveries of relevant and evolving concepts.

The intention of this research is to discover and recommend a set of most relevant and focused tags from social bookmarking applications for scientific resources. It is a common practice of researchers to explore the resources through interlinked chains as through references or citations. The socially annotated libraries like CiteULike also provide an interlinking of resources by using hyperlinked tags. For example CiteULike provides a list of tag terms for a user search keyword 'visualization' as shown in figure 4.4 (only top 20 are shown).

These terms are computed from the tag co-occurrence, for example, terms related to ‘visualization’ search keyword are the terms which same users assigned to resources along with tag term ‘visualization’. This tag list is organized on the basis of frequency of term occurrence in CiteULike. From the figure 4.4, if a user want to explore further resources from CiteULike related tag terms of visualization search keyword, say by clicking on Clustering tag in the list, then the user will get a list of all resources annotated with tag term ‘Clustering’. There might be some resources related to main focus (visualization) somewhere in the list but the returned recourses will be sorted based on clustering keyword rather than visualization keyword which put an extra burden on user to find focused resources. However in our case, system extracts the tag terms from CiteULike tags based on direct and partial match of authors’ keywords of a particular research paper. In this way the highly relevant discovered tags are linked with the paper. These tags were then compared in CiteULike tags by using direct and partial match. The extracted tag terms for ‘visualization’ and ‘tags’ are shown in figure 4.4. The extracted tags for visualization remains in the same focus and will link the resources in CiteULike which will often be related to the scope of visualization. Now if a user visits this paper he will see these related tags organized according to author keywords as hyperlinks. For further navigation if a user selects any tag from the extracted list, he/she is directed to the associated resources in CiteULike.

For example for the WWW’06 paper ‘ Visualizing tags over time’ authors provided keywords are ‘visualization’, ‘tags’, ‘flickr’, temporal evolution’ and ‘interval covering’. The second contribution of this research is an overall extension of the author keyword concepts with the social meta-data of tagging along with some serendipitous discoveries of relevant or evolving concepts. It is obvious from the figure 4.4 that the tags extracted for keyword ‘visualization’ are its subfields like data-visualization, its application areas like network-visualization and evolving concepts like social visualization. These lists of keywords signify an overall picture of popular research in related fields within the focus of a research paper.

4.4 Concluding Remarks

This research intended to discover a relationship between bookmarks/tags and citations. The case study shows that there exist a positive correlation between bookmark counts and citations. Tag terms also reoccur in the titles of the citing papers. Furthermore, the ranking of papers based on bookmark counts can predict citation counts better than the co-author network. This also means that Bookmark

popularity has the potential to become web-based indicator for knowledge diffusion (at least within CiteULike application scenario). Further more it provides very useful contextual information about diffusion by the tags.

| Paper number 23: Visualizing Tags over Time | | | |
|--|---|--|---|
| Extracted keywords for visualization | Citeulike keywords for visualization | Extracted keywords for "TAGS" | Citeulike keywords for "TAGS" |
| information-visualization, geovisualization, visualizations, data-visualization, network_visualization, social-visualization, informationvisualization, graph-visualization, information_visualization, software-visualization, volume-visualization, tree_visualization, software_visualization, tag-visualization, network-visualization, flow-visualization, graph_visualization, search-result-visualization | Information, software, data, network, project-email, infovis, hci, analysis, graph, bioinformatics, data-mining, Clustering, communication, email, collaboration, evaluation, design, networks | Tags, no-tag, geotags, expressed-sequence-tags, metatags, update-tags, tail-tags, skin_tags, unsure-tags, encoding_tags, qtags, rating-tags, meta-tags, affinity_tags, affectivetags, searchandtags, smart_tags, penntags, etags | Tagging, folksonomy, social, pixi, end-user-programming, plurality, flickr, tag, Folksonomies, delicious, collaboration, networks, citeulike, eni, social-software, toread, web, classification, location, recommendation |
| Paper Number 69: Semantic Wikipedia | | | |
| Extracted tag keywords for "Wikipedia" | Citeulike keywords for "Wikipedia" | Extracted keywords for "RDF" | Citeulike keywords for "RDF" |
| Wikipedia, used_for_wikipedia, web-characterization-wikipedia, historywikipedia | Wiki, semantic, quality, collaboration, ontology, visualization, semantic_web, cooperation, paper, social, social-network, trust, web, Tagging, reputation, wikis, collaborative, 2009, community, conflict | Squirrelrdf, computingrdf, translationrdf, ontologyrdf, bio2rdf, rdfa, krdf, analytic_brdf, sw-rdf, rdfs, brdf, -rdf, squirrelrdf-hpl2007-rdf | Semantic, semantic_web, Owl, Web, xml, ontology, p2p, semanticweb, semantic-web, sparql, database, knowledge, ontologies, query, graph, semantics, metadata, kr, rss, iswc |
| Extracted keywords for "WIKI" | | Citeulike keywords for "WIKI" | |
| lkewiki, aclwiki, kawawiki, acewiki, bowiki, sweetwiki, biowiki, xowiki, pmwiki, annotation-on-wiki, engineswiki, engineeringwiki, sitesxwiki, toolwiki, mapwiki, ots-wiki, traduwiki, wiki, wikis, semanticwiki, mediawiki, semwiki, twiki, semantic_wiki, bizwiki, wikid, geowiki, semperwiki, ow2wiki, ontowiki, sbwiki, creationcustomer-centricityknowledgemanagementope nsourcewiki | | Collaboration, is366c, Wikipedia, semantic, blog, web20, learning, awareness, community, web, education, social, collaborative, knowledge, online, blogging, internet, socialsoftware, elearning, koelpu, | |

Figure 4.4: Comparison of recommended tags for particular author keywords and their relevant CiteULike tags

As most of databases and meta-data resources related to scientific focus are joining LOD, in future an overall tags and bookmarks of a resource can be aggregated from LOD which will provide a more comprehensive assessment of this aspect.

Afterwards, it was found that the tags not only hint the content of the paper but also the context of future diffusion. WWW'06 papers were linked with CiteULike papers through tags. For this purpose, authors' assigned keywords to

WWW'06 papers were used as seed to find relevant tags from CiteULike by direct and partial match. The system was able to recommend popular tags for WWW'06 papers and a user had an option to find other relevant resources (papers) that are annotated with the same or similar tag. The result shows that system extends the authors keyword set with social tags providing links to rich and focused resources in CiteULike. This also enhances the serendipitous discovery of emerging concepts related to the focused resources. The dataset for tag based scientific resource recommendation has been made available for posterity at <http://www.student.tugraz.at/anwar.ussaeed/datasets.html>.

Multifaceted Expertise Mining and Topical Visualization of Experts

In numerous contexts and environments, it is necessary to identify and assign (potential) experts to subject fields. Although, apparently, the phenomenal success of Wikipedia and other open and bottom-up knowledge systems seems to undermine and challenge the role of experts in the knowledge industry. But even in open dynamic authoring systems like Wikipedia articles, which are good, are there because those are written by the experts of that field. Although Wikipedia claims to have an anti-expert bias but still [Sanger 2009] explains that ‘it is not wholly free of deference to expertise’. The quality of content in Wikipedia articles comes from experts but ‘the prerogatives of expertise are respected voluntarily’ the only difference is that experts are not being imposed by the system. But on the contrary, the same virtue of least interference of experts is also reported to be the reason for low credibility of the Wikipedia content.

As explained in chapter 3 the collaborative authoring systems like Wikipedia have the huge potential to lure the heart of those who intend to build highly prolific knowledge systems with low costs and hence these systems may play the central role in future knowledge creation and diffusion on the Web. These systems have become popular because of their fecundity and openness attracting huge volunteered participation. The only issue that encircles such systems is the mediocrity and credibility of the content which they contain as a resource of knowledge. [Sanger 2009] argues that if such resources are to become authoritative there must be some role, may be softer, for expert overview of the facts in the content.

In this chapter we shall not enter into the discussion of ‘what role an expert may have in Wikipedia like systems’ as the role may vary to the requirements of such systems and the discussion is related to the social science research. However, we propose that one of the minimal roles for an expert may be that of color highlighting the parts of content which are against a fact. In this way the viewer of the page will know about the credible parts of wiki-page. In this chapter research will be concerned with the fact that how we can assign experts (as reviewers) automatically to the topics of the content. The anonymity in open

systems impedes, due to their inherent nature, the discoveries of experts from within these systems but still some features can be used to define facets to find pseudo-named experts from within the system. Consideration must be taken while deciding features like the prolific authors. We suggest that the experts can be recommended from the established digital scientific resources like ACM, IEEE and BioMed as they have already opened their meta data resources on the web through Linked Open Data framework and this will enhance the confidence of scientific community in Wiki systems, hence, encouraging more participation from them. A hybrid approach will be possible by using bookmarking and tagging meta-data resources. As the already existing review process in scientific Journals is free of cost (the researchers extend their services as reviewers without any financial benefit), the popularity of authors as experts will remain the only currency in the new system too. This reduces, in the case of our prototype implementation, the problem of finding experts assigned to a content topic automatically in the online digital libraries. We propose an innovative automated technique which incorporates multiple facets in providing a more representative assessment of expertise. For the prototype application, proposed in this chapter, we used the online Journal of Universal Computer Science (JUCS) database for mining expertise.

In organizational practices and digital libraries, both manual and automated approaches are employed for expertise mining. These approaches have their own pros and cons. The quality of data is good in manual approaches but they need extensive human efforts. On the other hand, the quality of service may not as good in automated approaches as in manual ones but they are faster and don't need human efforts. The current automated approaches normally use only one metric to measure the expertise of an individual, e.g., the number of publications etc. This chapter proposes and implements an automated approach for measuring expertise profile in academia based on multiple metrics for measuring an overall expertise level.

In the context of an academic journal for computer science (J.UCS), papers and reviewers are classified using the ACM classification scheme. We used this topical classification due to its ready availability in the Journal system. The tagging classification or emergent semantics can also be used for topic clustering of resources.

This work describes a system to identify and present potential experts/reviewers for each category from the entire body of paper's authors. The topical classification hierarchy is visualized as a hyperbolic tree and currently

assigned reviewers are listed for a selected node (computer science category). In addition, spiral visualization is used to overlay a ranked list of further potential reviewers (high-profile authors) around the currently selected category. This new interface eases the task of journal editors in finding and assigning reviewers. The system is also useful for users who want to find expert research collaborators in specific research areas.

This chapter addresses the following research questions:

RQ.1. Which facets are important for ranking experts in scientific communities and how they can be discovered in scientific resource datasets?

RQ.2. Which visualization approach will be suitable for viewing rated experts in diverse topics of knowledge?

Based on published contribution [Afzal et al 2009], the progress flow of the research is shown in the figure 5.1.

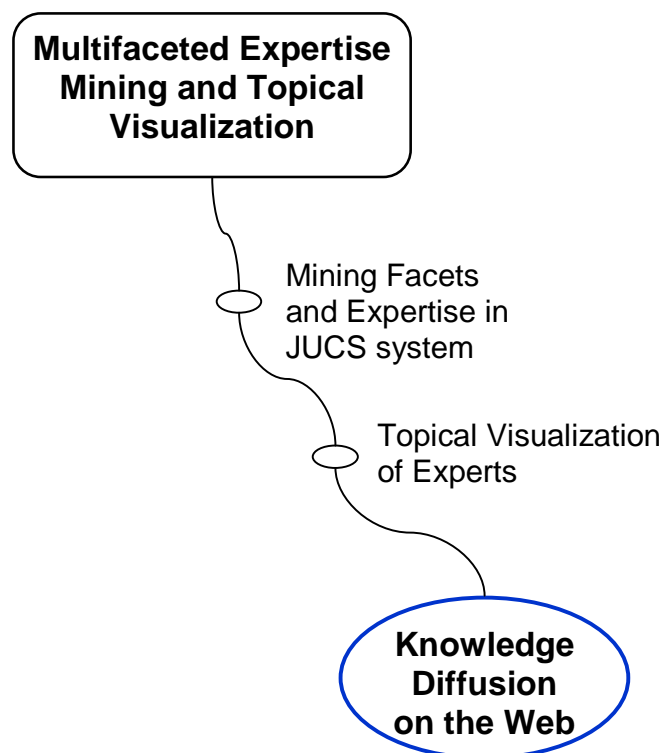


Figure 5.1: Progress flow of the chapter

5.1 Research Overview

The discovery of expertise is crucial in supporting a number of tasks. Expertise finder systems in the past have been innovatively applied in helping PhD applicants in finding relevant supervisors [Liu and Dew 2004] and also in identifying peer-reviewers for a conference [Rodriguez and Bollen 2008]. The former made use of a manually constructed expertise profile database while the later employed reference mining for all papers submitted to a conference. In the later, a co-authorship network was constructed for each submitted paper making use of a measure of conflict-of-interest to ensure that papers were not reviewed by associates.

[Cameron et al 2007] employed a manually crafted taxonomy of 100 topics in [DBLP] covering the research areas of a small sample of researchers appearing in [DBLP]. They proposed the need for automatic taxonomy creation as a key issue in finding experts. [Mockus and Herbsleb 2002] employed data from a software project's change management records to locate people with desired expertise in a large organization. Their work indicated a need to explicitly represent experiential characterization of individuals as a means of providing insights into the knowledge and skills of individuals. [Yimam 1999] have further shown that a decentralized approach can be applied for information gathering in the construction of expertise profiles. [Tho et al 2007] employed a citation mining retrieval technique where a cross mapping between author clusters and topic clusters was applied to assign areas of expertise to serve as an additional layer of search results organization. There are also expertise detection systems that were based entirely on an analysis of user activity and behavior while being engaged in an electronic environment. [Krulwich and Burkey 1995] have analyzed the number of interactions of an individual within a discussion forum as a means of constructing an expert's profile. Although such an approach is useful in monitoring user participation, measures such as number of interactions on a particular topic is in itself not reflective of knowledge levels of individuals.

A variety of tools have been implemented within organizations to find experts and expertise for different scenarios. Most related works make use of explicitly specified expert profiles constructed manually. The problem with such manually constructed profiles is that they tend to be developed for particular projects and constantly need to be updated e.g. [Pipek et al 2002].

Using an entirely automated mechanism for determining user expertise may also not be adequate in itself. As an illustration, Google Scholar employed an

automated approach and wrongly identified names of places such as Ann Arbour, or Milton Keynes as cited authors [Postellon 2008]. This also highlights the non-trivial nature of expertise mining and the difficulty faced in the disambiguation of individuals.

We propose an automated technique which incorporates multiple facets in providing a more representative assessment of expertise as explained in Section 5.2. To overcome automation errors mentioned above, we used an innovative citation mining technique [Afzal et al 2009a]. We see these facets as providing multiple sources of evidence for a more reflective perspective of experts. We present the combination of both tangible and intangible metrics to shed deeper insights into the intensity of expertise. The system mines multiple facets for an electronic journal and then calculates expertise' weights. The overall weight is further used to rank experts in the respective topic. The measures provided are, however, not absolute indicators of expertise as the discoveries are limited by the coverage of the database of publications and expert profiles used.

The system discoveries can be enhanced by visualizing the mined data [Shneiderman 2002]. In order to enhance the knowledge discoveries, we have visualized experts by using hyperbolic tree visualization technique. The proposed technique is based on focus plus context with extended focus to represent the statistical data as explained in section 5.3. The aforementioned technique is useful especially for journal administration to find high profile authors (experts) who can be assigned as editors/reviewers for the respective topics. Such applications may also help users to establish expert collaborations in their respective area.

5.2 A Multi-Faceted Expert Profile

In exploring a comprehensive characterization of expertise, we proposed a multifaceted approach for mining the expertise for a digital journal. The multiple facets are represented by the following measurements: number of publications, number of citations received, extent and proportion of citations within a particular area, expert profile records, and experience. We have thus incorporated the use of user-defined profiles, “experience atom” (as proposed by [Mockus and Herbsleb 2002] to indicate fundamental experiential units), reference mining results and a characterization of expert participation as facets of an expert profile. In a comprehensive characterization of expertise, the following measurements have been proposed:

5.2.1 Number of Publications

Number of publications is used to account for the overall activity of an author. Further more they help in deciding the proficiency of an author in a particular field. This means that one may have more publications in one topic than he has in others.

5.2.2 Citations Received

The simplified assumption is that only the quality papers are cited. Citation counts are already used by authors as currency. Therefore, it can safely be assumed that citations are indicative of the impact of publications and as a result can be applied to reflect the impact of expert.

5.2.3 Reviewer Profile Records

J.UCS has an editorial board consisting of its 300 members and their expert profiles represent the specified area of their expertise based on ACM categories. These experts are selected manually. These selections are not updated very frequently and remain intact for longer periods. During this time, however, the recent research focus and activities of these selected members may be changed. There can be new evolving research areas in their respective fields for which they may not be considered as experts. To overcome these issues they are assigned a proportionate weight scheme so that the current and working researchers may not be overwhelmed by these manually selected reviewers.

5.2.4 Experience

The measures of experience are always very complex. The measures that can be acquired with regards to the assessment of experiences may include but not limited to: period of publishing in a particular area, list of projects participated in, assessments of mentoring activities, etc. The factors from Open collaborative systems, like how prolific one is, how many articles are bookmarked and tagged etc can also be taken into account in this facet. In the current work, we have taken into consideration the publication age factor only.

Combining all these factors provides a better indication of expertise with regards to a particular topic. Figure 5.2 shows the consolidated view of expert profile construction as applied in our research.

In our research, there are two main sources of information used to construct an expert profile: 1) user inputs and 2) system discoveries. User inputs are taken from reviewers of the journal J.UCS. The J.UCS has over 300 reviewers on its editorial board. The expertise of these reviewers are specified and maintained according to the ACM classification scheme [ACM-CCS, 1998]. This information was extracted from J.UCS and used to populate the expert profile database.

The second source for constructing expert profiles is computed by the system. The computation considers the number of publications of an individual, the number of citations that a person receives, and the person's duration of publication in the respective area.

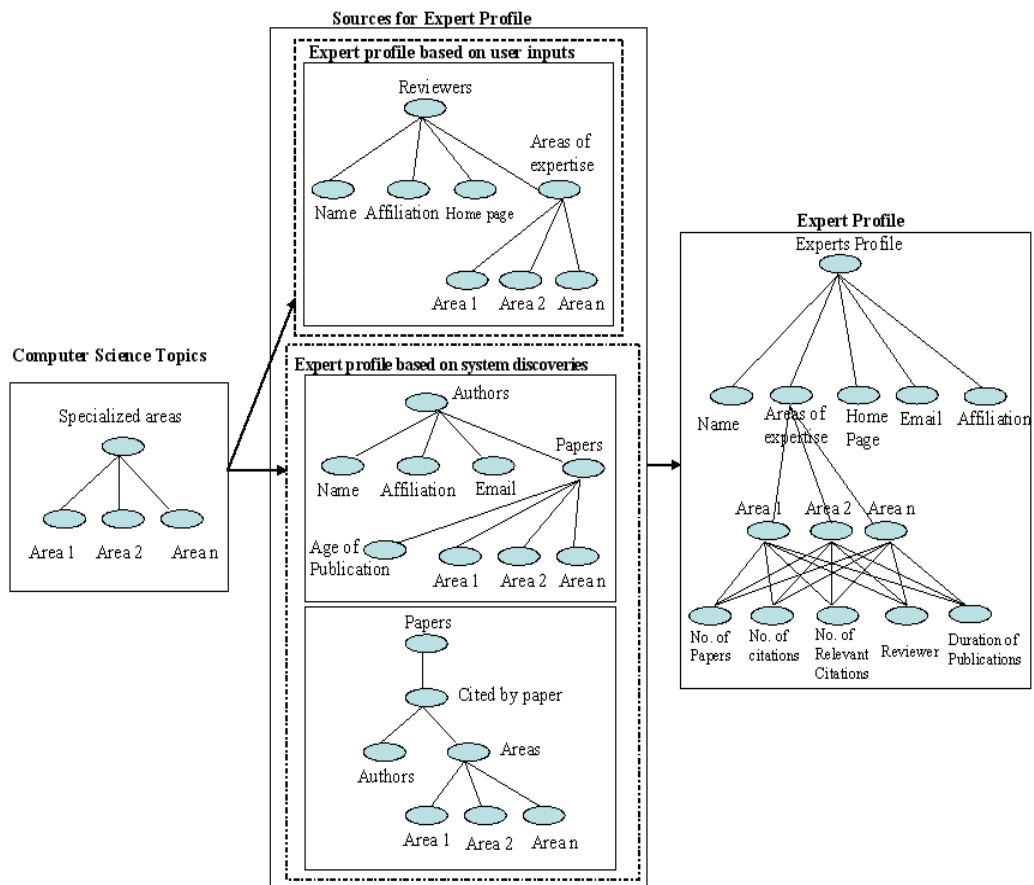



Figure 5.2: Expert Profile

5.3 Data Extraction

Within J.UCS, ACM topics, editors, and every individual paper are represented in an XML notation, which needs to be parsed to extract metadata. A typical XML file for J.UCS papers can be seen in Figure 5.3. The metadata (paper title, authors, ACM topic, etc.) related to a paper is stored inside the XML file.

The extracted data was used to populate a relational database. The database presents a coherent view of all data with relationships (category, paper, authors, and citations). The data from this database was then used to calculate and visualize experts within the J.UCS environment.



```
Calude_C_meta.xml*
1 <?xml version="1.0" encoding="UTF-8"?>
2 <article xmlns="http://www.ujseries.org"
3   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4   xsi:schemaLocation="http://www.ujseries.org http://www.ujseries.org/article-20050623.xsd"
5   id="jucs_1_1/what_is_a_random">
6   <title lang="en" runningHead="What Is a Random ...">
7     What Is a Random String?
8   </title>
9   <abstract lang="en">
10    Chaitin s algorithmic definition of random strings - based on the complexity induced by
11  </abstract>
12  <acmCategory id="G.3">(PROBABILITY AND STATISTICS)</acmCategory>
13  <keyword lang="en">Blank-endmarker complexity</keyword>
14  <keyword lang="en">Chaitin (self-delimiting) complexity</keyword>
15  <keyword lang="en">random strings.</keyword>
16  <author id="1" type="corresponding">
17    <firstname>Cristian S.</firstname>
18    <middlename></middlename>
19    <lastname>Calude</lastname>
20    <email>cristian@cs.auckland.ac.nz</email>
21    <phone></phone>
22    <city>Auckland</city>
23    <zip></zip>
24    <country>New Zealand</country>
25    <institution>
26      <name>Computer Science Department, University of Auckland</name>
27      <url></url>
28    </institution>
29  </author>
30  <submissionDate>1995-01-14</submissionDate>
31  <acceptanceDate>2002-08-26</acceptanceDate>
32  <publicationInfo journal="jucs"
33    issue="1"
34    issueType="regular"
35    issueAccess="restricted"
36    volume="1"
37    date="1995-01-28"
38    managingEditorColumn="no"/>
39  <pageInfo from="48" to="66" number="19"/>
40 </article>
```

Figure 5.3: A sample paper XML File

5.3.1 Weights Assigned to Experts

In our system, experts are grouped into one of two categories: 1) editors (persons currently manually assigned as reviewers for a particular ACM topic) and 2) high-profile authors (persons flagged automatically as experts in a particular topic). Reviewers are selected by the editor-in-chief based on their expertise in the respective ACM topical area. Reviewers for a particular ACM category are visualized without any further calculation. High-profile authors are calculated based on weights assigned to them. The facets defined in Figure 5.2 are used to assign the weights. The weights used in our system are `publication_weight`, `citation_weight`, and `editor_weight`.

5.3.1.1 Publication Weight

The publication weight of an author in a particular topic is the ratio between the number of the author's publications and the number of publications' years (duration of publications). The current activity factor is exhibited in the publication age which in our system is the last five years of publication. The number of years is calculated from the year of a first publication (within last five years) until the current year. A larger publication weight would mean that author is very prolific for a certain research area in last five years.

Publication Weight = No. of publications / duration (No. of years).

5.3.1.2 Citation Weight

Citations, in general, are considered as the research measure or currency for the authors and are also used as requirement to hire in lucrative research positions. Although there may exist some reservations to the extent of research measure they hold but they are frequently used in this respect for evaluations and allocation of funds in the field of science and technology.

Citation Weight = No. of citations received by an author in a topic / total No. of citations in an ACM topic

5.3.1.3 Editor Weight

The editors' weight represents the proportion that how many of the authors are the reviewers. It's a ratio of number of J. UCS editors with the total number of J. UCS authors. This weight is assigned only to the reviewers in J. UCS system. In this way, the expertise factor of reviewers is accounted for and they get an edge over the other authors.

Editor Weight = No. of J. UCS editors / Total no. of J. UCS Authors.

The total weight is defined as the sum of the above defined weights:

Total weight = publication weight + citation weight + editor weight.

High-profile authors are then ranked according to their total weight.

5.4 Information Visualization

While considering different visualization schemes, extended hyperbolic visualization was selected for topical representation due to its comprehensiveness of nodes visibility or context visibility. This feature helped the view of all topics in one place as context of visualization where one can easily navigate to a particular topic by dragging or by clicking on the node and can see editors and potential experts belonging to that topic. Due to these amiable and user-friendly features, we have chosen the hyperbolic browser which is based on “focus+context” technique [Lamping and Rao 1994] [Lamping et al 1995] [Lamping and Rao 1996]. The hyperbolic browser was overlaid with ranked spiral visualization for a topical node in focus where each node on the spiral represents a respectively rated expert. This, compound visualization scheme, helped the J. UCS administrators to focus on any particular topic and select an automatically suggested expert as reviewer while the overall context of its rating remains visible. The details of hyperbolic visualization can be found in the next section. This visualization can also be very helpful in initiating collaboration with experts whenever one needs. The remaining parts of this section explain both of the aforementioned visualizations.

5.4.1 Extended Hyperbolic Visualization

Reviewers are essentially attached to a node within the ACM classification hierarchy. For each node within the ACM classification hierarchy, a ranked list of high-profile authors (potential reviewers) was calculated as shown in earlier section. The hyperbolic browser [Lamping and Rao 1994] [Lamping et al 1995] [Lamping and Rao 1996] is an efficient visualization technique for large hierarchies. A hyperbolic browser is used to provide intuitive navigation within the ACM classification hierarchy. For any selected node in the ACM hierarchy, a spiral is used to visualize the ranked list of high-profile authors for that node. The spiral is simply superimposed upon and around the selected node. This builds on past work with GopherVR [McCahill and Erickson 1995], PRISE [Cugini et al

1996], and RankSpiral [Spoerri 2004] which both use spiral representations to display ranked search result lists.

The user interface is shown in Figure 5.4. This is implemented in Java. A hyperbolic browser is used to visualize the ACM classification hierarchy, using the freely available Hypertree package [Hyperbolic Package 2009]. Both categories of experts are visualized by superimposing upon the hyperbolic view. Reviewers are shown in a simple list and high profile authors are shown in spiral visualization. To draw the spiral, a package called Turtle Graphics is used [Turtle Graphics 2009]. With Turtle Graphics, simple commands are used to move and draw on the graphical surface. With these commands, the spiral is drawn and the names of the experts are written at constant angular steps. To visualize the reviewers of a specific ACM topic, a simple JList is used. A maximum of 10 reviewers are shown in the JList.

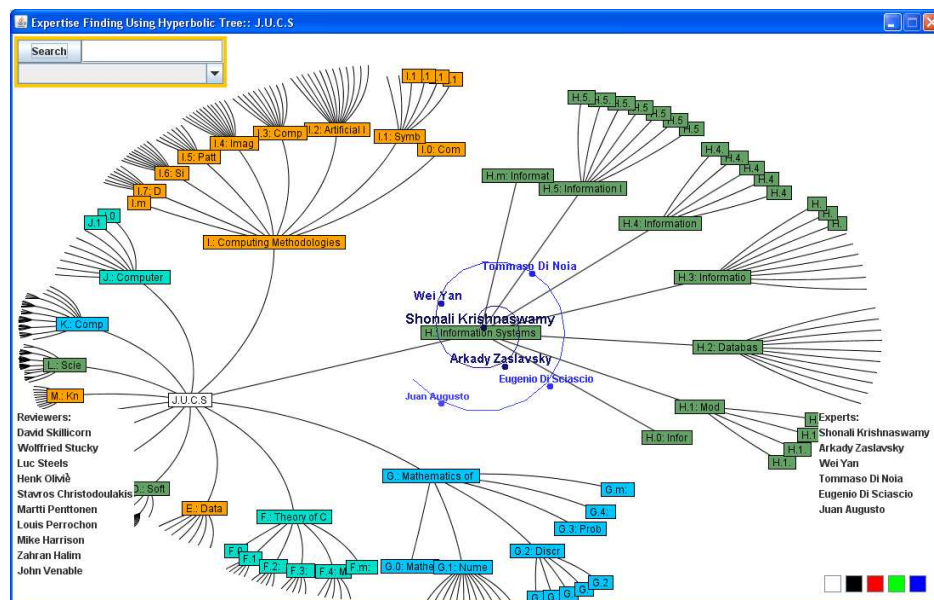


Figure 5.4: Hyperbolic Visualization

The JList, spiral, and Hypertree are placed in JPanels inside a frame, and are ordered with a JLayeredPane. One can arrange the JPanels horizontally and vertically and even manipulate the z-order. The Hypertree is drawn in the back. When an ACM topic is clicked, the list of reviewers is shown in the bottom left and the spiral of high-profile authors is overlaid over the ACM topic in the top layer, as shown in Figure 5.4. When there are neither reviewers nor high-profile authors, no list or spiral is drawn. In the bottom right of the window, there are five colored buttons. When clicked, the spiral is redrawn with the new color. It is

possible to choose black, red, green, or blue. Users can hide both the spiral and the reviewers list if required by clicking white button. The spiral moves along with the focused node whenever a user drags a particular node.

Figure 5.5 shows the visualization for ACM category “H. Information Systems”. The reviewers are shown in the bottom left corner. When a user clicks on the node “H. Information Systems”, a spiral is drawn around the selected node. The high-profile authors are placed in the spiral in descending order of their total weight (the highest weighted in the centre of the spiral).

This visualization is useful for journal administering. For example, in J.UCS there are some topics with very few assigned reviewers. J.UCS administration can instantly find potential reviewers based on the high-profile authors shown by the system. For example, the topic ‘M.8 Knowledge Reuse’ has no reviewers at the moment (this is a new topic added by J.UCS). Potential reviewers are easily found in the visualization, as shown in Figure 5.5. This type of discovery is very helpful for J. UCS administrators to locate potential reviewers for any selected area.

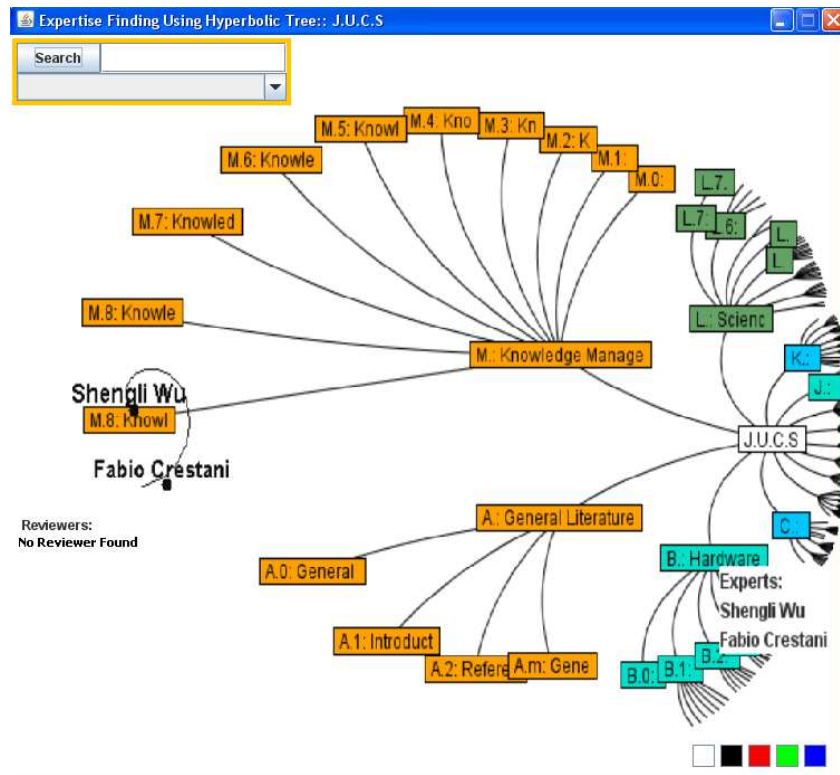


Figure 5.5: Discovery of Potential Reviewers

Although it is convenient to explore the topical hierarchy with the hyperbolic tree, users sometimes know the name of a topic and want to navigate directly to it. The search facility in the top left corner of the main interface (see Figure 5.6) supports this task. For example, if a user searches for the term “Information”, then a combo box is filled with all topics containing the term “Information” as a substring. The 13 topics containing the term “Information” are shown in Figure 5.6. The user can select any ACM topic from the search result list and the hyperbolic tree is redrawn to show the selected topic centered in the window.

5.5 Concluding Remarks

This chapter presented a new system to identify and visualize current and potential experts in topical areas of a scientific discipline. It is used in the context of a computer science journal to identify and assign reviewers to areas of computer science, but can easily be generalized to other scientific communities.

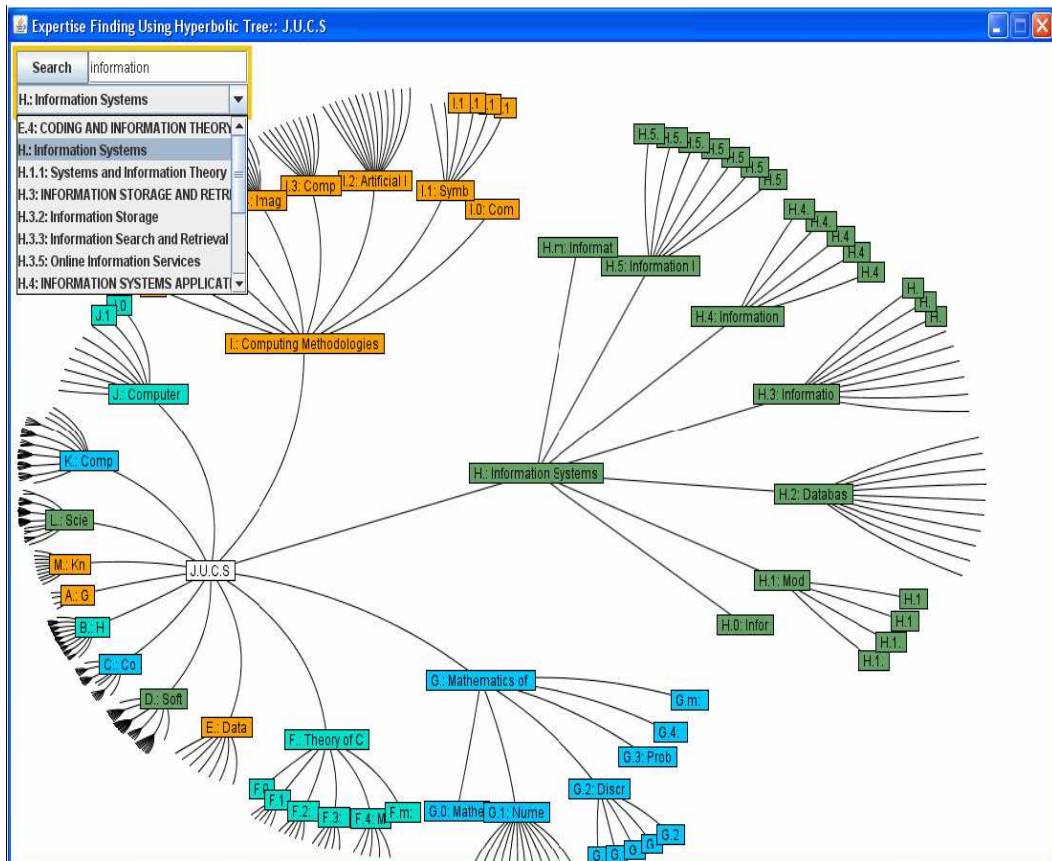


Figure 5.6: ACM Topic Search Facility

The main contributions of this chapter are:

- A methodology for automatically identifying potential experts from assembled profiles.
- A combined visualization of a topical classification hierarchy and a ranked list of potential experts at each level in the hierarchy.

Global Discovery through Simplified Search Interface of Linked Open Data

“The established system of journals for communicating the results of scientific research is already being challenged by the existence of the web. But we are only in the early days of a new Internet revolution, one which will have a deeper and more disruptive impact on scientific, and other, web publishing, and have profound implications for the web itself. An emerging successor to the web, the Semantic Web, will likely profoundly change the very nature of how scientific knowledge is produced and shared, in ways that we can now barely imagine”. [Berners-Lee and Hendler, 2001]

The participatory enthusiasm of the users in response to the rise of open and collaborative trends has spurred a new age of knowledge, data and information flows on the Web. [Bizer et al, 2010] marks that the low barriers to publishing, open access, hypertext linking, capabilities of search engines to infer potential relevance to users' search queries by analyzing the structure of hypertext links between documents [Brin & Page, 1998] and extensible nature of the Web [Jacobs & Walsh, 2004] are the hallmark principles which enabled unconstrained growth of the Web of interlinked documents. While the Web has been a success in managing, linking and exploiting document resources it has recently entered to the Web of linking data and exploiting their semantics, realizing the vision of the Web of Data. The huge amount of knowledge, especially scientific knowledge like in the Genome project, resides in the databases on the Web but cannot be globally searched as the tools and search technologies don't provide search across these databases with one query. Such is the case with other useful public or government databases. Search engines, such as Google, rely upon automated crawlers and are great for finding Web pages but they typically cannot reach information within a database. On the other hand, the hypertext HTML data format is not sufficiently expressive to enable individual entities described in a particular document to be connected by typed links to related entities' [Bizer et al, 2010].

However, in recent years, the adoption of the Linked Data best practices has augmented the Web with a global data space connecting data from diverse domains such as people, companies, books, scientific publications, films, music, television and radio programmes, genes, proteins, drugs and clinical trials, online communities, statistical and scientific data, and reviews. The Linked Data applications provide capabilities to query across this unbound number of databases through the Web. This brings added value to user and to these applications as they can ‘deliver more complete answers’ with the addition of every new data source on the Web. [Bizer et al, 2010]

The focus of this work is limited to the discussion of semantic search mechanism and user friendly search interface for Linking Open Data (LOD)¹⁶ so that the end user can query these databases without the prior knowledge semantic structures.

In this chapter following research questions are addressed.

RQ.1. How we can simplify the (Subject - Predicate - Object) SPO logic of semantic search with a keyword search mechanism?

RQ.2. How we can coalesce and present the resource information in a user friendly structure by hiding the ontology hierarchy?

Based on multiple coauthored publications [Latif et al, 2009] [Latif et al, 2009a] [Latif et al, 2009b] [Latif et al, 2009c], the flow of this chapter is presented in figure 6.1

In this chapter we will explain the concept of Global discovery and its importance for knowledge diffusion. Then we will detail that how Linked Data framework enables Web of Data and Global Discovery. Further on we propose a keyword base search architecture and a concept aggregation framework to bridge the gap between end user and semantic search on LOD.

6.1 Global Discovery

Term ‘Global Discovery’ was used in the diffusion of scientific knowledge ‘to address the complexity of search issue’ [Wojick et al, 2006] specifically related to the distributed research databases on the Web. With this term they intend to

¹⁶ <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

explain that a federated search function which can query multiple scientific databases of diverse community focuses can enhance diffusion of scientific knowledge and hence the growth of science.

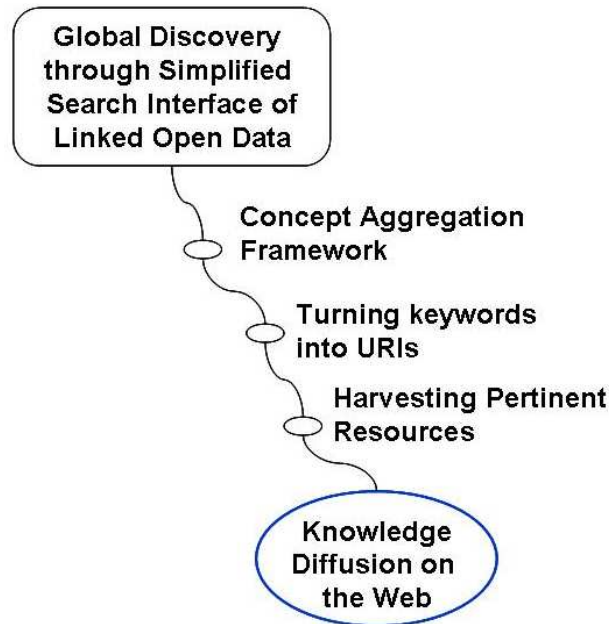


Figure 6.1: Flow of Chapter 6

As the huge amounts of authoritative science information reside in databases within the deep Web and conventional webpage search engines have limitations in regard to querying such databases, researchers are left with the tedious and time-consuming task of searching “door to-door” in only the scientific communities and databases with which they are already familiar. Their access remains within the particular community knowledge due to interaction dynamics and technological search function constraints. [Wojick et al, 2006] argues that ‘if scientists can easily discover the initial breakthroughs being made in communities other than their own, then scientific knowledge diffusion would be greatly accelerated. A search mechanism which can query across these databases will ‘illuminate often obscure databases and speed access to scientific information, which will in turn increase the probability of further and more rapid innovation and discovery’. Thus, global discovery itself has recently become a necessary focus area for research. The same principle of Global discovery is at the heart of new Socio-Semantic Web movement named Linked Open Data (LOD). LOD holds potential not only to enhance the Global discovery but also to extend its definition towards the more generic knowledge and data integration principles.

[Losoff, 2009] notes that with the rise of digital access to data, the data has become more valuable than the published paper itself. He points out that the data sets, from the Human Genome Project, 'have more value than any single publication that was derived from an analysis of them' [Carlson, 2008]. The Semantic Web design principles envisioned by Tim Berners-Lee, director of the World Wide Web Consortium, creating a universal medium for data, information, and knowledge exchange has led to the extension of the Web with a global data space. These design principles, otherwise termed as the Linked Data best practices, hold great potential to enhance global discovery by integrating digital scientific data with scholarly literature.

Semantic Web provides a standard Resource Description Framework (RDF), a web-based "Semantic Tool" for encoding knowledge, "permitting web sites to publish information as machine-readable, process-able, and in integrated forms" [Tauberer, 2006]. Along with RDF the Web Ontology Language (OWL) provides an agreed-upon published conceptualization of content [De Roure et al, 2003]. These have made possible for applications to explore relationships between data and electronic literature across multiple platforms and databases [Losoff, 2009]. In the next section we will describe Linked Open Data initiative and our prototype search application.

6.2 Linked Open Data

As described in chapter 2, the W3C launched the Linking Open Data (LOD) movement, a community effort that motivates people to publish their information in a structured way. The Linked Open Data movement has been integral to RDF publishing on the Web. As of April 2010, the LOD cloud consists of about 13 billion RDF triples, which are interlinked by around 150 million RDF/OWL links [LDOW 2010]. Although LOD has created huge volumes of data and has attracted the attention of many researchers, it still lacks broad recognition, especially in commercial domains. This is, amongst other reasons, because of complex semantic search and end user applications [Latif et al, 2009c]. The underlying publishing framework of LOD, as explained by Tim Berners-Lee(see section 2.2), demands to publish the data with some ontological structure, with unique URI and HTTP lookup or dereferencing. Below we explain what is meant by URI dereferencing and what the core ontological structures in LOD are.

6.2.1 URI Dereferencing

With regards to providing information about a resource upon a HTTP lookup of its URI is called dereferencing. Emphasis is placed on providing information in RDF and disambiguating identification of information resources (document URIs) from non-information resources (entities described in those documents) [Hogany et al, 2010].

W3 proposed draft standard defines “Information resources are resources, identified by URIs and whose essential characteristics can be conveyed in a message [AWWW]. The pages and documents familiar to users of the Web are information resources. Information resources typically have one or more representations that can be accessed using HTTP. It is these representations of the resource that flow in messages. The act of retrieving a representation of a resource identified by a URI is known as ‘dereferencing’ that URI” [W3, 2010].

6.2.2 Ontology Classification

In the absence of official standards, DBpedia¹⁷ and Yago¹⁸, amongst others, are considered de facto standards for classification. DBpedia is also a central interlinking hub for Linked Data. Facts about specific resources, extracted from the info-boxes of Wikipedia, are structured in the form of properties as defined by DBpedia's ontology [Auer et al 2007]. This ontology is associated with Yago's classification to identify the type (person, place, organization, etc.) of the resource. For instance, a query about Arnold Schwarzenegger returns about 260 distinct properties, encapsulating nearly 900 triples in the raw RDF form. Such semantic data is not (easily) graspable by end users. Representing this bulk of structured information in a simple and concise way is still a challenge.

6.3 Semantic Search Mechanism and SPO Logic for LOD

Recently, a few applications have emerged, which provide user interfaces to explore LOD datasets [Berners-Lee et al 2006a] [Kobilarov and Dickinson 2008]. These applications use SPARQL endpoints to query LOD with Subject-Predicate-Object (SPO) logic. SPO logic represents a triple, which is a building block of RDF. A triple establishes a relationship between two resource types. One resource

17 <http://dbpedia.org>

18 <http://www.mpi-inf.mpg.de/yago-naga/yago/>

is called subject and the other one object. The relationship between subject and object is called predicate. For example, Arnold Schwarzenegger (subject) is governor of (predicate) California (object). Now, in order to exploit LOD resources using SPARQL endpoint with interfaces of recent applications, users have to understand the underlying semantic structures (triples, ontologies, properties).

Each resource that is described by Linked Data can be uniquely identified by its URI [Sauermann et al 2008]. Relations and attributes of this URI can then be queried by use of SPARQL. However, the URI dereferencing provides the power of direct access to the common users and they can explore the resource by HTTP browsers. For example, when a user wants to know something about “Arnold Schwarzenegger”, it is necessary for him to find a URI that represents this person in the Semantic Web e.g. http://dbpedia.org/resource/Arnold_Schwarzenegger. This adds another complex search dimension for common end users. The gap between semantic search and end user applications has also been identified by [Chakrabarti 2004].

The current state of the art with respect to the consumption of Linked Open Data for end users is RDF browsers [Berners-Lee et al 2006a] [Kobilarov and Dickinson 2008]. Some tools such as Tabulator [Berners-Lee et al 2006a], Disco¹⁹, Zitgist data viewer²⁰, Marbles²¹, Object Viewer²² and Open link RDF Browser²³ can explore the Semantic Web directly. All these tools have implemented a similar exploration strategy, allowing the user to visualize an RDF sub-graph in a tabular fashion. The sub-graph is obtained by dereferencing a URI and each tool uses a distinct approach for this purpose. These tools provide useful navigational interfaces for the end users, but due to the abundance of data about a concept and lack of filtering mechanisms, navigation becomes laborious and bothersome. In these applications, it is a tough task for a user to sort out important pieces of information without having the knowledge of underlying ontologies and basic RDF facts. Keeping in mind these issues, we suggest a keyword search mechanism to reduce the cognitive load of the users. We also proposed the Concept Aggregation Framework conceptualizes the most relevant information of a resource in an easily perceivable construct.

¹⁹ <http://www4.wiwiw.fu-berlin.de/bizer/ng4j/disco/>

²⁰ <http://dataviewer.zitgist.com/>

²¹ <http://beckr.org/marbles>

²² <http://objectviewer.semwebcentral.org/>

²³ <http://demo.openlinksw.com/rdfbrowser/index.html>

6.4 Proposed Keyword Search Mechanism

We propose a two-step keyword search process in order to hide the underlying SPO logic. In the first step, users search for a keyword, and the system auto-suggests related entries to exactly specify the subject. In our system users don't need to remember a URI anymore to find resources from LOD. Users enter a keyword, and the system discovers the most relevant resources from LOD. The system employs a two-layered approach. In the first layer, users are automatically suggested with resources matching the entered keywords from a locally maintained LOD resource triple store. In the second layer, the user keyword is matched with metadata of resources indexed by a Semantic Web search engine (Sindice). When the system has identified a correct resource URI, then the system proactively picks up a set of properties related to the selected resource. The most relevant set of properties is grouped together by using the Concept Aggregation Framework. Furthermore, to avoid searching a specific property (predicate) of the selected subject by its name, a keyword based 'search within' facility is provided where the specified keyword is mapped to a certain property or set of properties of the retrieved resource. With this proposed methodology of simplifying semantic search to keyword search on LOD we have contributed in two ways:

1. We introduced a Concept Aggregation Framework which selects a set of properties related to a particular informational aspect of a resource type. This approach conceptualizes the most relevant information of a resource in an easily perceivable construct.
2. We proposed a two step keyword search process in order to hide the underlying SPO logic. In the first step, users search for a keyword, and the system auto-suggests related entries to exactly specify the subject. Then, information related to that subject is structured using the aggregation framework. Furthermore, to avoid searching a specific property (Predicate) of the selected Subject by its name, a keyword based 'search within' facility is provided where the specified keyword is mapped to a certain property or set of properties.

6.5 Concept Aggregation Framework

The Concept Aggregation Framework aggregates relevant concepts from DBpedia and organizes the most important informational aspects related to a resource.

The scope of this application is limited to DBpedia and Yago. DBpedia covers 23 types of resources (places, people, organizations, etc). Initially, we selected the resource type person for the experimentations.

The Concept Aggregation Framework is shown in figure 6.2. The aggregation classification layer is responsible for aggregating the most relevant information related to the person in question. This information is collected based on the list of related properties compiled at the property aggregation layer. The properties are extracted from knowledge bases shown in the aggregation knowledge bases layer.

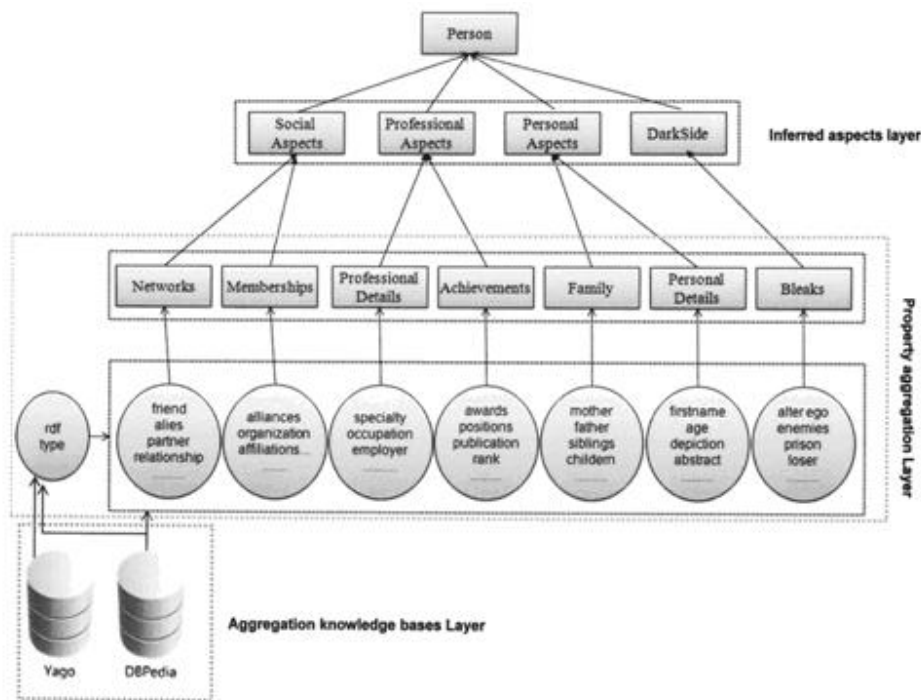


Figure 6.2: Concept Aggregation Framework

6.5.1 Aggregation Knowledge Bases Layer

DBpedia, Yago and Umbel ontologies mainly contribute in the identification and classification of the resources. Two of them (DBpedia and Yago) are considered complete knowledge bases [Suchanek et al 2007]. The underlying mechanism in our system is as follows:

We have generated two knowledge bases, a DBpedia Property Dump and a Yago Classification Dump. The DBpedia Property Dump is built by querying

each type of a person (Artist, Journalist, etc.) from SNORQL query explorer²⁴ (SPARQL endpoint of DBpedia). Then we aggregate all the distinct property sets for each person. Out of 21 queried person types in total, we were able to collect distinct properties of 18, which are presented in Table 6.1. It shows the number of distinct properties in total that we collected for a specific person as well as the number of properties picked by a set of experts, which will be mapped to defined aspects.

To decide which of these properties should be presented to the user, a query is formulated to get the count of every distinct property used for person type. After getting the count, the rank is assigned to each property.

Table 6.1: Person’s property list

| Person Type | Total Properties | Picked Propertie |
|--------------------|------------------|------------------|
| Artist | 2111 | 409 |
| Journalist | 186 | 55 |
| Cleric | 419 | 76 |
| BritishRoyalty | 252 | 47 |
| Athlete | 2064 | 496 |
| Monarch | 337 | 50 |
| Scientist | 421 | 126 |
| Architect | 132 | 41 |
| PlayboyPlaymate | 125 | 37 |
| Politician | 36 | 18 |
| MilitaryPerson | 725 | 158 |
| FictionalCharacter | 599 | 273 |
| Criminal | 287 | 74 |
| CollegeCoach | 282 | 124 |
| OfficeHolder | 1460 | 634 |
| Philosopher | 226 | 71 |
| Astronaut | 168 | 62 |
| Model | 211 | 99 |

The higher the rank, the more prominently the property will be displayed. For example, some of the properties of person type Athlete like “Position” (70939 times), “clubs” (46101 times) and “debutyear” (9247 times) provide interesting stats to organize properties in a more conceivable fashion. The formulated query for this operation is given below:

²⁴ <http://dbpedia.org/snorql/>

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?p
WHERE {
?s ?p ?o .
?s rdf:type <http://dbpedia.org/ontology/Artist> .
}

```

The Yago Classification Dump is built by querying subclasses of Person class from SNORQL query explorer. The query looks like this:

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?s
WHERE {
?s rdfs:subClassOf <http://dbpedia.org/class/yago/Person100007846>.
}

```

6.5.2 Property Aggregation Layer

This layer first identifies the profession type. This works in two steps. In the first step, the resource type (RDF type) is identified by using DBpedia. In the case where in the retrieved set of properties, there is no property mapped within DBpedia knowledge base, the system tries to map the retrieved property to a Yago class. For example if the retrieved property is “AustrianComputerScientist” which is not listed in DBpedia knowledge base, then the system maps it to the Yago hierarchy and can infer that the person belongs to the profession of “Scientist” because “AustrianComputerScientist” is a subclass of “Scientist”.

Based on a resource type, we have extracted all the possible properties from the DBpedia Property Dump. We then have manually identified sets of properties indicating an informational concept (networks, memberships, family, achievements etc.) related to a person. These concepts are aggregated and mapped to the related informational aspect identified in the inferred aspects layer. More than one concept may be mapped to a single informational aspect defined at the inferred aspects layer.

6.5.3 Inferred Aspects Layer

The information for a resource such as person may be organized and viewed in different informational aspects like personal, professional or social. The most popular search engine like Google also tries to present such informational aspects related to a topic in its top results. It has been shown in [Brin and Page 2008] that how Google rank its results to provide the most relevant contents. For

example, in a response to a user query of “Bill Clinton”, Google top ten results are based, amongst other things, on personal information (biography) and his professional career (president, writer). These results, however, depend on the complex link analysis of Web pages (citations to Web pages from different sources) along with weight mechanisms assigned to different factors [Feldstein 2009] [Boykin 2005]. Google is considered as the most popular search engine having 64.2 % share in U.S search market [Lipsman 2009]. Inspired from Google’s success in calculating and presenting the results in diverse and important informational aspects related to a query, we developed a concept aggregation framework where diverse yet important aspects of a person are represented in inferred aspect layer.

6.6 System Architecture

The system architecture is depicted in figure 6.3. The implemented system is divided into four modules called query manager, auto-suggestion module, information retrieval module and search within property module. The query manager is a controlling module of the application. It is responsible in translating the keyword search query into SPARQL queries. The auto-suggestion module helps users to disambiguate entered search term. The information retrieval module is responsible for locating the URIs and extracting related information. The search within property module provides the facility of searching within all retrieved properties of a resource.

6.6.1 Auto-Suggestion Module

The query manager triggers the auto suggestion module by converting the searched keyword of a user into a SPARQL query. This module interacts with the DBpedia person and the DBpedia disambiguation triple store to autosuggest persons with names that match the entered keyword. This module has been discussed in detail in section 6.4. If the user does not select any of the suggested terms, or in case of a distinct query (no auto-suggestions yielded), the searched term is passed on to the information retrieval module for further processing.

6.6.2 Information Retrieval Module

This module is further divided into four processes:

- 1) URI locator
- 2) LOD retrieval
- 3) Parser
- 4) Concept aggregation

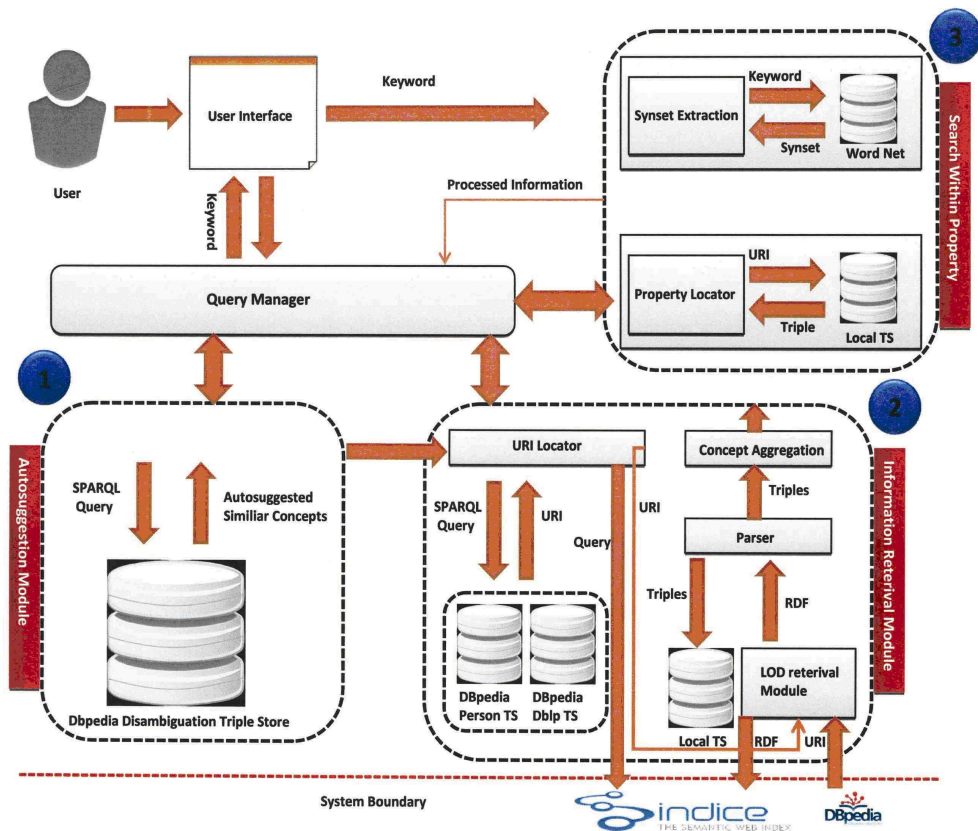


Figure 6.3: System Architecture

The searched term is passed to the URI locator process which will query the locally maintained data sets (i.e. DBpedia Title TS, DBpedia Person Data TS, and DBLP TS) to get a URI. If this fails, a new query is formulated for the SINDICE²⁵ Web service to locate the URI. After locating the URI of a resource, the LOD retrieval process dereferences that URI at the DBpedia server to get the respective resource RDF description. This RDF description is further passed to the Parser process. This process parses RDF description into triples and stored them locally. Then, the concept aggregation process is called to sort out the most important information aspect of the resource and in the end; the output is presented to the user.

6.6.3 'Search within Properties' Module

This module lets the user search within all properties of a resource retrieved from the information retrieval module. If a user enters a search keyword the

²⁵ <http://sindice.com/>

synset extraction process queries wordnet²⁶ to retrieve the synset of search term. This synset is passed to the query manager and for each word in the synset, it query the local triple store through the property locator process. The property locator process matches the keyword as substring in the retrieved property set. All matched properties are then extracted and presented to the user.

6.7 Concluding Remarks

The proposed keyword-based search mechanism has simplified the process of finding information from LOD by hiding underlying semantic logic. With the help of Concept Aggregation Framework, the information related to a resource (consisting of hundreds of properties) was structured in major and most relevant categories of informational aspects. This reduced the users' cognitive load to find the required information. Keyword-URI locating technique was very helpful in identifying a particular resource from huge LOD repository. This work tries to bridge the gap between semantic search and the end user. By simplification to keyword based search we provide ease of use so that common web users can consume data from LOD. This type of search mechanism will increase the global discovery of knowledge by the common users and hence will increase diffusion. The preliminary evaluation of the systems has shown promising results and we plan to extend this application with the help of further users' online evaluations.

²⁶ <http://wordnet.princeton.edu/wordnet/>

Summary and Outlook

This chapter provides a brief overview of the research. It elaborates results and concludes the thesis. The future possible extensions to the work are then explained.

7.1 Results and Discussions

The scientific contributions in this thesis elucidate the potentials and challenges of open collaborative applications regarding the creation and diffusion of scientific codified knowledge. The thesis proposes solutions for many of these problems and also provides prototype implementations as proof of concept. A novel sub-document tagging and content aggregation mechanism is proposed and implemented. A statistical analysis of citation and tagging data is provided. Novel multifaceted expertise mining and visualization application is developed and explained. An innovative keyword search mechanism for LOD is introduced and implemented for providing access to the common web users for the huge semantic resource. Moreover inspired by the open and collaborative trends of Web 2.0 and social software, this thesis suggests that Wikis, Blogs, Bookmarking and Tagging systems can provide an ecosystem of codified (scientific) knowledge creation and diffusion. In this ecosystem each application may act as an independent service. Such a resembling system was identified as EOL and compared with Wikipedia for its scientific value and other features. We proposed that the blogs or personal websites may replace scientific publications whereas the role of scientific journals may be assumed by the wiki environments where the reviewed content will be aggregated from diverse blogs or websites. Tagging and bookmarking will provide the indicators of popularity and hence the impact of scientific work. In order to become attractive to the scientific communities these open and collaborative applications need to be modified in a way which ensures credibility of content and popularity of the authors. To this vision we contributed in following ways:

This work proposed and provided a prototype implementation of new content aggregation and personalization features in Wikis. These two are explained in detail in chapter 3. In brief this work introduced a novel sub document level tagging called selection (or section) tagging. Selection tagging proposes that any content selected from web can be tagged and then latter aggregated using these tags into a dynamic wiki environment for further collaborative processing. The reuse of existing content in the form of copy-paste mechanisms in order to restructure and create new documents is applied by authors frequently. This typically requires a lot of effort and time. A prototype implementation of selection tagging is presented in a wiki environment ‘Austria Forum’ with personalization plug-in. The prototype is limited to select and tag a section within the wiki environment. The benefit of such an approach has been illustrated in the wiki system on the example of simplifying the editing process. We call this new approach section-tagging as it supports users to assign keywords and annotate sections of a wiki page which can later be aggregated in another dynamically created wiki page. It is implemented as separate plug-in. The plug-in mechanism allows server-side code to be referenced from within a wiki page. This code dynamically produces wiki content that can be included in the wiki page that refers to the plug-in. This rapid content integration into wiki and its restructuring will lower the barriers of content creation and restructuring in wiki along with the added personalization features.

The second contribution of this work is to find empirical and analytical similarities of citation link structures with web based social tagging and bookmarking links and meta-data. Citations are mainly used as indicator of codified knowledge diffusion in scientific scholarship. Past research [Scharnhorst and Wouters, 2006][Day, 2008] recommends the encouragement of web based knowledge diffusion supplementary indicators along with citations. This is discussed in detail in chapter 4. This work explored that the citations have positive correlation with bookmark counts and the tag terms of a paper appear frequently in its citing titles. This work provides empirical evidence of similarities among citations and bookmarks. Hence the bookmark reputation of a scientific resource will bring value to its authors like the citations do. Further more we see that tag based recommendations of popular scientific resources can also be used for scientific articles to improve 'browsing experience' just like references or citations do. We extracted the scientific paper specific tags based on author keywords from CiteULike for the whole set of accepted papers of WWW06 conference. These tags are the hyperlinks (each tag has a unique URI in CiteULike) to the set of relevant papers in CiteULike. This approach of extending set of author keywords of a paper with the socially assigned tags also facilitated

serendipitous discovery of new evolving concepts and fields related to the resource in focus.

The third contribution of this research is the multifaceted mechanism of automatically discovering potential reviewers or experts in a scientific knowledge system and their topical visualization. The prototype application implements an automated approach for measuring expertise profile in J. UCS database based on multiple metrics for measuring an overall expertise level. The ACM classification scheme is used in the prototype for classifying papers and reviewers. We used this topical classification due to its ready availability in the Journal system. The tagging classification or emergent semantics can also be used for topic clustering of resources. Facets can also be grouped from open bottom-up and conventional knowledge repositories which are now available on Linked Open Data cloud. In prototype application the multiple facets are represented by the following measurements: number of publications, number of citations received, extent and proportion of citations within a particular area, expert profile records, and experience. The topical classification hierarchy is visualized as a hyperbolic tree and currently assigned reviewers are listed for a selected node (computer science category). In addition, spiral visualization is used to overlay a ranked list of further potential reviewers (high-profile authors) around the currently selected category. This visualization can be used to initiate other type of collaborations too.

The fourth and final contribution of this thesis is the proposed simplified search interface and keyword search mechanism for Linked Open Data which will remove the semantic query requirement. Furthermore, it presents the information in more logical aggregation hiding the semantic architecture. Such an interface will allow common web user to explore the world of LOD across multiple data sets hence enhancing global discovery which in turn will accelerate diffusion. The user evaluations have shown that the system was able to find required information and logical grouping was of great use for the users. The system scaled in accuracy with semi-automated systems like FreeBase [Latif et al, 2009].

The detailed architecture of the prototype application is presented in chapter 6. The simplification of semantic search to keyword search on LOD has two layers of simplification:

1. A Concept Aggregation Framework which selects a set of properties related to a particular informational aspect of a resource type. This approach conceptualizes the most relevant information of a resource in an easily perceivable construct.

2. A two step keyword search process in order to hide the underlying Subject-Predicate-Object (SPO) logic. In the first step, users search for a keyword, and the system auto-suggests related entries to exactly specify the subject. Then, information related to that subject is structured using the aggregation framework. Furthermore, to avoid searching a specific property (Predicate) of the selected Subject by its name, a keyword based 'search within' facility is provided where the specified keyword is mapped to a certain property or set of properties.

Concluding the above discussion, the thesis makes contributions broadly in four areas which are:

- Collaborative knowledge creation and its diffusion in Wikis
- Tagging and bookmarks as measures of diffusion: The thesis analyses the potential of the tagging and bookmarking metadata resources for the study of knowledge diffusion by finding its empirical relationship and similarities with citation (an established indicator of knowledge diffusion)
- Automatically finding experts or reviewers: Multifaceted Expertise mining in scientific repositories and their topical visualization
- Global Discovery: by providing user friendly keyword search web interface for semantic data sets in LOD

7.2 Future Prospects

“Imagine a database of research in which new findings are not published in papers that are put into volumes, but appended in various places to a single, collaboratively-managed outline of knowledge. It seems that such collaborative resources might indeed change how journalists and researchers find their sources, and textbooks or their future equivalents might well be parts of such systems. This is pure speculation, but it does seem possible” [Sanger, 2009].

The dissertation explored the potentials of collaborative and social evolutions of the new Web applications in line with the above mentioned vision. The thesis proposed and implemented solutions to the challenges that impede the course of efficient diffusion and creation of scientific knowledge in such applications. Due to the time constraint of the research thesis the exhaustive solutions to the challenges were not possible only the prototype implementations were provided. The future extensions of this research may bring more clarity and objectivity to the diffusion of scientific knowledge on the Web.

The innovative new idea of ‘using simplified keyword search interface across distributed databases of LOD and hiding the semantic details’ can serve as a basis for further research. The automatic ontology integration strategies for the back end data processing and URI mining techniques will gain the focus in the future research of this work. Moreover, the content coalescing and its more user-friendly presentations will also be a rich aspect of research.

Also further investigations can be carried out on the novel idea of ‘selection tagging’. As described earlier in chapter 4 the section-tagging approach can be extended to arbitrary Web resources and is called ‘selection tagging’. This can be implemented as browser plug-in in future which will gather the tagged snippets of content selected from the Web in a dynamic wiki page as a Web based service. The interesting aspects of this research will be:

- Sharing of sub-document tags between users, i.e. not only a personalized selection-tag cloud should be generated but also a global one with tags from all users. This then may be compared with the normal document tagging.
- The relationship of such tags with the extracted text snippets will also be a topic of future research.
- The tempting directions of research on global sub-document tag clouds will be the tag and snippets selection strategy in the case that there are numerous text snippets tagged by a particular tag. A collaborative filtering approach taking into the account the user profiles might be needed to limit the snippets only to those that are most relevant.
- Future extension of this research may explore the use of these tags as they are expected to be more helpful in search and emergent semantic mechanisms.

At the end, while concluding, it can be said that the open, collaborative, participatory and socio-semantic trends of the new Web applications hold great potential for rapid diffusion of scientific knowledge and hence can increase the growth of science. However, it will remain to be seen if such systems will be able to tempt the scientists and researchers in large or conventional systems of publishing and knowledge diffusion will remain their major interest.

Bibliography:

[Afzal et al 2009] Afzal, M. T., Latif, A., Us Saeed, A., Sturm, P., Aslam, S., Andrews, K., Tochtermann, K., Maurer, H. (2009). Discovery and Visualization of Expertise in a Scientific Community. In: Proceeding of International Conference of Frontiers of Information Technology, Islamabad, Pakistan, 16-18, Dec. 2009.

[Afzal et al 2009a] Afzal, M. T., Balke, W. T., Kulathuramaiyer, N., Maurer, H. (2009). Rule based Autonomous Citation Mining with TIERL, Accepted in Journal of Digital Information Management, 2009.

[Alexander 2006] Alexander, B., Web 2.0: A new wave of innovation for teaching and learning?, Educause Review, 41(2) (March/April). Retrieved November 2006 from <http://www.educause.edu/ir/library/pdf/ERM0621.pdf>

[Almind & Ingwersen, 1997] Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to “Webometrics”. Journal of Documentation, 53, 404-426

[Ames and Naaman, 2007] Ames, M. and Naaman, M. 2007. Why we tag: motivations for annotation in mobile and online media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 971-980. DOI= <http://doi.acm.org/10.1145/1240624.1240772>.

[Anderson 2005] Anderson, T. , Distance learning – social software's killer ap? ODLAA 2005 Conference.
[<http://www.unisa.edu.au/odlaaconference/PPDF2s/13%20odlaa%20-%20Anderson.pdf>]

[Anderson 2007] Anderson, P., What is Web 2.0? Ideas, technologies and implications for education. Technical report, JISC, 2007

[Anjewierden et al, 2005] Anjewierden, A., de Hoog, R., Brussee, R. & Efimova, L. “Detecting knowledge flows in weblogs”, in 13th International Conference on Conceptual Structures, University of Kassel, Kassel, 1-12, July 2005.

[Ankolekar et al. 2008] Ankolekar ,A., Krötzsch , M., Tran , T., Vrandecic, D., The two cultures: Mashing up Web 2.0 and the Semantic Web,

Web Semantics: Science, Services and Agents on the World Wide Web, v.6 n.1, p.70-75, February, 2008

[Auer et al, 2009] Auer, S., Bizer, C., Idehen, K (2009). DBpedia Knowledge Base, retrieved on 22, Sep.2009 from <http://dbpedia.org>

[Backer, 1991] Backer, TE., Knowledge Utilization: The Third Wave Science Communication.1991; 12: 225-240

[Benkler 2006] Benkler, Y., The Wealth of Networks: how social production transforms markets and freedom, 2006, Yale University Press: USA.

[Berners-Lee and Hendler, 2001] Berners-Lee, T. and Hendler, J (26 Apr 2001)., Scientific publishing on the 'semantic web'. Nature 410, 1023 - 1024 (26 Apr 2001)

[Berners-Lee et al 2006] T. Berners-Lee, "Linked Data Design Issues"; (July2006).

[Berners-Lee et al. 2001] Berners-Lee, T., Hendler, J., and Lassila. O.,The semantic web. Scientific American, 284(5):34--43, 2001

[Berners-Lee et al, 2006a] Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., Sheets, D. (2006). Tabulator: Exploring and Analyzing Linked Data on the Semantic Web. In: Proceedings of 3rd International Semantic Web User Interaction Workshop. Athens, Georgia, USA.

[Berrueta and Phipps, 2009] Berrueta, D., Phipps, J (2009). Best Practice Recipes for Publishing RDF Vocabularies. <http://www.w3.org/TR/swbp-vocab-pub/> (accessed 22, Sep.2009).

[Bettencourt et al 2006] Bettencourt, L M. A., Castillo-Chavez,, C., Kaiser, D., Wojick, D.E. Report for the Office of Scientific and Technical Information:Population Modeling of the Emergence and Development of Scientific Fields; http://www.osti.gov/innovation/research/diffusion/epicasediscussion_lb2.pdf ; October 4, 2006

[Bizer et al, 2010] Bizer, C., Heath, T., Berners-Lee, T., Linked Data - The Story So Far. (2010.), WWW2010 workshop: Linked Data on the Web (LDOW2010)

[Bolloju and Wagner, 2005] Bolloju, N., & Wagner, C. . Supporting knowledge management in organizations with conversational technologies: Discussion forums, weblogs, and wikis. *Journal of Database Management*, 2005, 16(2), i-viii.

[Borgman & Furner, 2002] Borgman, C.L., & Furner, J. (2002). Scholarly Communication and Bibliometrics. In B. Cronin (Ed.), *Annual Review of Information Science and Technology*, Vol 36. Medford, NJ: Information Today, pp 3-72.

[Branstetter, 2003] Branstetter, L., “Measuring the impact of academic science on industrial innovation: the case of California's Research Universities”. Columbia Business School Working Paper, 2003.

[Brin & Page, 1998] Brin, S., Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117.

[Cameron et al 2007] Cameron, D., Aleman-Meza, B., Arpinar, I.B. 2007. Collecting Expertise of Researchers for Finding for Relevant Experts in Peer-Review Setting. In the proceeding of 1st International Expert Finder Workshop (Berlin, Germany, Jan 16 2007).

[Card et al 1999]Card, S. K., Mackinlay, J., Shneiderman, B.1999. *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Publishers, San Francisco, CA. ISBN-13: 978-1558605336.

[Carlson, 2008] Carlson, S., How to channel the data deluge in academic research. *The Chronicle of Higher Education*, 2008 [Online]. Available: <http://chronicle.com/weekly/v54/i30/30b02401.htm> [Accessed June 8, 2008].

[Casati et al, 2009]Fabio Casati, Fausto Giunchiglia, Maurizio Marchese Publish and perish: why the current publication and review model is killing research and wasting your money version 1.0, ACM Ubiquity project, 29 November 09

[Catarci et al 1997] Catarci, T., F. Levialdi, M., S. Batini, C (1997). Visual Query Systems for Databases: A Survey. *Journal of Visual Languages and Computing*. 8(2), 215-260.

[Chakrabarti 2004]. Chakrabarti, S., "Breaking Through the Syntax Barrier: Searching with Entities and Relations", In: *Proc. PKDD'2004*, Springer, Berlin Heidelberg, (2004), 9–16.

[Chen et al 2007] Chen, C., Maceachren, A., Tomaszewski, B., MacEachren, A., "Tracing conceptual and geospatial diffusion of knowledge", in *LNCS 4564*, pp.265-274, 2007.

[Cheng et al, 2008] Cheng, G., Ge, W., Qu, Y (2008). Falcons: Searching and Browsing Entities on the Semantic Web. In: *Proceedings of 17th International World Wide Web Conference (WWW)*. pages 1101-1102, Beijing, China.

[Chimezie, 2009] Chimezie, Dereferencing a URI to RDF. <http://esw.w3.org/topic/DereferenceURI> (accessed 22, Sep.2009).

[Cooney 2006] Cooney, L. (2006). Wiki as a Knowledge Management Tool. CERAM Sophia-Antipolis

[Cowan et al, 2000]Cowan, R., Paul A. David, and Dominique Foray "The Explicit Economics of Knowledge Codification and Tacitness", in *Industrial and Corporate Change*, 9(2), pp.211-253, 2000.

[Cugini et al 1996] Cugini, J. V., Piatko, C. D., and Laskowski, S. J. 1996. Interactive 3D Visualization for Document Retrieval. In *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation at CIKM* (Rockville, Maryland, USA, Nov. 12-16, 1996). NPIVM'96. ACM Press.

[Dalsgaard 2006] Dalsgaard, C., Social software: E-learning beyond learning management systems. *European Journal of Open, Distance and E-Learning*, 2006/II, http://www.eurodl.org/materials/contrib/2006/Christian_Dalsgaard.htm.

[Day 2008] Day, M.: "Institutional Repositories and Research Assessment", Supporting Study No. 4, UKOLN, ePrints UK Project, Bath, available at: www.rdn.ac.uk/projects/eprints-uk/docs/studies/rae/rae-study.pdf (accessed 24 April 2008).

[DBLP]Digital Bibliography and Library Project,
<http://www.informatik.uni-trier.de/~ley/db/index.html>

[De Roure et al, 2003] De Roure, D., Jennings, N.R., & Shadbolt, N.R. 2003. The semantic grid: a future e-science infrastructure. In: Berman, F., Fox, G., and Hey, A.J.G., eds. Grid Computing-Making the Global Infrastructure a Reality; New York: Wiley, p.437-470.

[Désilets et al, 2005] Alain Désilets , Sébastien Paquet , Norman G. Vinson, Are wikis usable?, Proceedings of the 2005 international symposium on wikis, p.3-15, October 16-18, 2005, San Diego, California. DOI=<http://doi.acm.org/10.1145/1104973.1104974>

[Dickerson, 2004] Dickerson, C. (2004). Is Wiki Under Your Radar? Infoworld Publishing Group. From <http://search.epnet.com/login.aspx?direct=true&db=afh&an=14969736>

[Ding et al 1999] Ding, Y., Chowdhury, G., Foo, S. (1999). Template mining for the extraction of citation from digital documents. In: Proceedings of the Second Asian Digital Library Conference, Taiwan, pages 47–62.

[Doctorow et al, 2002] Doctorow, C., Dornfest, F., Johnson, Scott, J. , Powers, S., 2002. Essential Blogging,.O'Reilly

[Epidemix Blog,] Epidemix Blog, Why does Wikipedia Suck on Science, <http://epidemix.org/blog/?p=72>, accessed

[Feldstein 2009] Feldstein, A. (2009). Search ranking factors. <http://www.seomoz.org/article/search-ranking-factors> [Boykin 2005] Boykin, J. (2005). Google Top 10 choices for search results. <http://www.jimboykin.com/googles-top-10-choices-for-search-results/>

[Fichter, 2005] Fichter, D. . The many forms of E-collaboration: Blogs, wikis, portals, groupware, discussion boards, and instant messaging. Online (Weston, Conn.), 2005, 29(4), 48-50.

[Figg et al, 2006] Figg , W. D., Dunn, L. Liewehr, D.J., Steinberg, .M., Thurman, P. w., Barrett, J. C., Birkinshaw, J. Scientific Collaboration Results in Higher Citation Rates of Published Articles, doi:10.1592/phco.26.6.759

[Garfield 1980] Garfield, E. "The epidemiology of knowledge and the spread of scientific information.", *Current Contents* 35, pp. 5-10 , 1980

[Garfield, 1955] Garfield, E., "Citation indexes for science: A new dimension in documentation through association of ideas". in *Science*, 122(108-111), 1955.

[Garfield, 1979] Garfield, E. (1979). Perspective on citation analysis of scientists., in *Citation indexing* (pp. 240-252). Philadelphia: ISI Press

[Garfield, 1998] Garfield, E. (1998) The use of journal impact factors and citation analysis in the evaluation of science. 41st Annual Meeting of the Council of Biology Editors, Salt Lake City, UT, May 4, 1998

[Gijsbers, 2004] Gijsbers, V. (2004). Ideals of knowledge: Media from Plato to Wikipedia. Retrieved 10/15, 2005 from <http://www.phys.uu.nl/~gijsbers/writings/0051ideals.pdf>

[Giles, 2005]Giles, Jim, "Internet Encyclopaedias Go Head to Head, *Nature*, December 15, 2005

[Glänzel and Thijs 2004] Glänzel, W., Thijs, B., "Does co-authorship inflate the share of self-citations?", *Scientometrics*, Volume 61, Number 3 / November, 2004.

[Godwin-Jones, 2003]Godwin-Jones R. (2003) Blogs and Wikis: Environments for On-line Collaboration. *Language Learning & Technology*, 7, (2), pp. 12-16.

[Goldfinch et al, 2003]Goldfinch, S., Dale , T., Derouen, K. Jr. Science from the periphery: Collaboration, networks and 'Periphery Effects' in the citation of New Zealand Crown Research Institutes articles, 1995-2000, *Scientometrics*, Vol. 57, No. 3 pp.321.337, 2003

[Hafner, 2006] Hafner, Katie. Growing Wikipedia Refines Its 'Anyone Can Edit' Policy. *New York Times*, June 17, 2006. <http://www.nytimes.com/2006/06/17/technology/17wiki.html>

[Hammond et al, 2005] Hammond, T., Hannay, T., Lund, B., and Scott. J., Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4), April 2005.

[Harth et al 2006] Harth, A., Umbrich, J., Hogan, A., Decker, S (2007). YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In: Proceedings of International Semantic Web Conference (ISWC). Springer, Busan, Korea

[Helic et al, 2008] Helic, D., Maurer, H., White, B., Austria-forum: a citable web Encyclopedia, IADIS International Conference, 2008

[Helic et al, 2009] Helic, D., Us Saeed, A., Trattner, C., Creating Dynamic Wiki Pages with Section-Tagging , in HT09 workshop New Forms of Xanalogical Storage and Function, 2009

[Hepp et al. 2007] Hepp, M., Siorpaes, K., Bachlechner, D (2007). Harvesting Wiki Consensus Using Wikipedia Entries as Vocabulary for Knowledge Management. IEEE Internet Computing. 11(5), pages 54-65.

[Hersh and Rindfleisch, 2000] Hersh WR, Rindfleisch TC. Electronic publishing of scholarly communication in the biomedical sciences. J Am Med Inform Assoc 2000; 7: 324-5.

[Hildebrand et al. 2006] Hildebrand, M., Ossenbruggen, V., Hardman, J (2006). Facet: A Browser for Heterogeneous Semantic Web Repositories. In: Proceedings of International Semantic Web Conference (ISWC). Athens, Georgia, USA.

[Hoffmann, 2008] Hoffmann, A., Proposing a Framework for Frequently used Terms in Knowledge Management, in Proceedings of I-KNOW '08 and I-MEDIA '08, 2008

[Hogany et al, 2010] Hogany , A. Harthy , A. Passanty , A., Deckery , S., Polleres, A., Weaving the Pedantic Web, 2010, WWW2010 workshop: Linked Data on the Web (LDOW2010)

[Holzinger et al, 2008] Holzinger, A., Geierhofer, R., Modritscher, F. & Tatzl, R. (2008) Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses. Journal of Universal Computer Science, 14, 22, 3781-3795.

[Holzinger et al, 2009] Holzinger K., Safran C., Ebner M., Kappe F., Koiner, G. & Holzinger, A. (2009). Geo-Tagging in Archaeology: Practical Experiences with the TUGeoWiki. In: 14th International Congress „Cultural

Heritage and New Technologies“ Vienna, Austria, Workshop 14, Computers and Archaeology, (p. 13). Stadtarchäologie Wien. [m-Learning, Geotagging, GeoWiki, Mobile Technologies]

[Holzinger et al, 2009a] Holzinger, A., Kickmeier-Rust, M.D. & Ebner, M. (2009). Interactive Technology for Enhancing Distributed Learning: A Study on Weblogs. In: Proceedings of HCI 2009 The 23rd British HCI Group Annual Conference, (pp. 309–312). Cambridge University, UK, British Computer Society. [e-Learning, e-Teaching, Collaborative Learning, Distributed Learning, Social Software]

[Hotho et al, 2006] Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G., Information Retrieval in Folksonomies: Search and Ranking. In: Proc. of ESWC 2006, pp. 411--426, 2006.

[Huang et al 2008] Huang, Y.C., Hung, C.C., Hsu, J.Y.: “You Are What You Tag”, in AAAI, 2008.

[Ioannidis et al 2008] Ioannidis JPA (2008) Measuring Co-Authorship and Networking-Adjusted Scientific Impact. PLoS ONE 3(7): e2778. doi:10.1371/journal.pone.0002778.

[Jacobs & Walsh,2004] Jacobs, I., Walsh, N. (2004): Architecture of the World Wide Web, Volume One - W3C Recommendation. Retrieved June 14, 2009, <http://www.w3.org/TR/webarch/>

[John and Walker, 2006] John P. Walker, Jr., Identifying and Overcoming Barriers to the Successful Adoption and Use of Wikis in Collaborative Knowledge Management, University of North Carolina, Chapel Hill, April 2006

[Kiefer et al 2007]Kiefer, C., Bernstein, A., Stocker, M (2007). The fundamentals of iSparql a virtual triple approach for similarity-based Semantic Web tasks. In: Proceedings of International Semantic Web Conference (ISWC). Busan, Korea.

[Kiefer et al 2007]Kiefer, C., Bernstein, A., Stocker, M (2007). The fundamentals of iSparql a virtual triple approach for similarity-based Semantic Web tasks. In: Proceedings of International Semantic Web Conference (ISWC). Busan, Korea.

[Kittur et. al, 2007] Kittur, A.; Chi, E. H. ; Pendleton, B. A. ; Suh, B. ; Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Alt.CHI at CHI 2007; 2007 April 28 - May 3; San Jose, CA.

[Kleinberg, 2004] Kleinberg, J., "Analyzing the Scientific Literature in its online Context". Nature, in Web Focus on Access to the Literature, April, 2004

[Kobilarov et al, 2008]Kobilarov, G., Dickinson, I (2008). Humboldt: Exploring Linked Data. In: Proceedings of Linked Data on the Web Workshop (LDOW). Beijing, China.

[Kristine] Kristine L. Callis, Lindsey R. Christ, Julian Resasco, David W. Armitage,

[Krulwich and Burkey 1995] Krulwich, B., Burkey, C. 1995. Contact Finder: Extracting Indications of Expertise and Answering Questions with Referrals. Technical Report.In the Working Notes of the Symposium on Intelligent Knowledge Navigation and Retrieval. The AAAI Press, 85-91.

[Lambert, 2003] Lambert J. Developments in electronic publishing in the biomedical sciences. Program: Electron Libr Inf Syst 2003; 37:6-15.

[Lamping and Rao 1994] Lamping, J., Rao, R. 1994. Laying out and Visualizing Large Trees Using a Hyperbolic Space. In ACM Symposium on User Interface Software and Technology (Marina del Rey, California, USA, Nov. 2-4, 1994) UISTP'94. 13-14. doi:10.1145/192426.192430.

[Lamping and Rao 1996] Lamping, J., Rao, R. 1996. The Hyperbolic Browser: A Focus+Context Technique for Visualizing Large Hierarchies Journal of Visual Languages and Computing. 7, 1 (Mar. 1996), 33-55. doi:10.1006/jvlc.1996.0003.

[Lamping et al 1995] Lamping, J., Rao, R., Pirolli, P. 1995. A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Denver, Colorado, May 7 - 11, 1995). CHI '95. ACM Press, New York, NY, 401-408. doi:10.1145/223904.223956

[Lancaster, 1995] Lancaster FW. The evolution of electronic publishing. Libr Trends 1995; 43: 518-27.

[Latif et al, 2009] Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K., CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data. Proceedings of NDT 2009, Ostrava, Czech Republic, July 2009.

[Latif et al, 2009a] Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K., Turning Keywords into URIs: Simplified User Interfaces for Exploring Linked Data. Accepted for: ACM proceeding of ICIS 2009, Seoul, Korea, November 2009. ISBN: 978-1-60558-710-3

[Latif et al, 2009b] Latif, A., Afzal, M. T., Us Saeed, A., Hoefler, P., Tochtermann, K., Harvesting Pertinent Resources from Linked Open Data. To appear in: Journal of Digital Information Management, 2009.

[Latif et al, 2009c] Latif, A., Hoefler, P., Stocker, A., Us-saeed, A., Wagner, C (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: Proceedings of I-Semantic. Graz, Austria.

[Lawrence, 2009] Lawrence S. Online or invisible? In: Nature [online] 2001; 411:521. Available at: citeseer.ist.psu.edu/online-nature01/. Accessed March 30, 2009.

[LDOW, 2010] Introductory speech, WWW2010 Workshop Linked data on the Web, 2010, <http://events.linkeddata.org/ldow2010/slides/ldow2010-slides-intro.pdf> retrieved on 31 May 2010

[Lipsman 2009] Lipsman, A. (2009). ComScore Releases April 2009 U.S Search Engine Rankings. http://www.comscore.com/Press_Events/Press_Releases/2009/5/comScore_Releases_April_2009_U.S._Search_Engine_Rankings

[Liu and Dew 2004] Liu, P., Dew, P. 2004. Using Semantic Web Technologies to Improve Expertise Matching Within Academia. In Proceedings of I-Know (Graz, Austria, Jun. 30- July 02 2004). I-Know'04. 370-378.

[Losoff, 2009] Losoff, B., Electronic Scientific Data & Literature Aggregation: A Review for Librarians Issues in Science and Technology Librarianship, 2009, <http://www.istl.org/09-fall/refereed2.html>

[MacGarvie, 2005] MacGarvie, M., “The determinants of international knowledge diffusion as measured by patent citations”, in *Econ. Lett.* 87, pp. 121–126, 2005.

[Manjunatha et al, 2003] Manjunatha, J.N., Sivaramakrishnan, K. R., Pandey, R. K., Murthy, M. N. Citation prediction using time series approach KDD Cup 2003 (task 1), SIGKDD explorations vol 5, issue 2 pp152

[Marlow et al. 2006] Marlow, C., Naaman, M., Boyd, M., Davis, M., “HT06, Tagging paper, Taxonomy, Flickr, Academic article, to read”, in proceeding of the 17th conference on hypertext and hypermedia, in HT, Odense, Denmark, 2006.

[Maurseth and Verspagen 2002] Maurseth, P. B., and Verspagen, B.: "Knowledge Spillovers in Europe: A Patent Citations Analysis" in. *Scandinavian Journal of Economics*, Vol. 104, No. 4, pp. 531-545, 2002 Available at SSRN:<http://ssrn.com/abstract=371854>.

[McCahill and Erickson 1995] McCahill, M. P., Erickson, T. 1995. Design for a 3D Spatial User Interface for Internet Gopher. In *Proceedings of World Conference on Educational Multimedia and Hypermedia (Graz, Austria, Jun. 17-21, 1995)*. ED-MEDIA 95. AACE, 39-44.

[Michlmayr et al 2007] Michlmayr, E., Cayzer, S.: “Learning User Profiles from Tagging Data and Leveraging them for Personalized Information Access”, in *WWW Banff*, Canada, 2007.

[Mika 2005] Mika, P.:”Ontologies Are Us: A Unified Model of Social Networks and Semantics”. In *Proc. of 4th Intl. Semantic Web Conference (ISWC2005)*, 2005.

[Millen et al, 2005] Millen, D., Feinberg, J., Kerr, B. 2005. Social Bookmarking in the enterprise. *ACM Queue*, Nov 2005. Available online at: <http://www.acmqueue.com/modules.php?name=Content&pa=showpage&pid=344>

[Mockus and Herbsleb 2002] Mockus, A., Herbsleb, J. D. 2002. Expertise Browser: A Quantitative Approach to Identifying Expertise. In *Proceedings on the International Conference on Software Engineering (Florida, USA, May 19-25 2002)*. ICSE’02. 503-312. doi:10.1145/581339.581401.

[Moed, 2005] Moed, H. F. (2005) Citation Analysis in Research Evaluation. NY Springer

[Molapo, 2007] Molapo, D. J., Knowledge Dissemination: Determining Impact, at workshop IFLA Durban 2007 Knowledge Management (KM) Best Practices/Lessons Learned

[Nardi et al, 2004] Nardi, B., Schiano, D., Gumbrecht, M., Swartz, L. 2004. Why We Blog, Communications of the ACM. Vol 47, No 12 (Dec 2004) pp. 41–46

[Nardi et al, 2004] Nardi, B., Schiano, D., Gumbrecht, M., Swartz, L. 2004. Why We Blog, Communications of the ACM. Vol 47, No 12 (Dec 2004) pp. 41–46

[Nelson et. al, 2008] Nelson, L., Smetters, D., and Churchill, E. F. 2008. Keyholes: selective sharing in close collaboration. In CHI '08 Extended Abstracts on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 2443-2452. DOI=<http://doi.acm.org/10.1145/1358628.1358701>.

[Ng, 2009] Ng., K. H., Exploring new frontiers of electronic publishing in biomedical science, Singapore Med J 2009; 50 (3) : 230

[NPACI, 2010] National Partnership for Advanced Computational Infrastructure, <http://www.npaci.edu/>, viewed on March 2010

[O'REILLY 2005] O'REILLY, T. 2005. What is Web 2.0: Design Patterns and Business Models for the next generation of software. O'Reilly website, 30th September 2005. O'Reilly Media Inc. Available online at: <http://www.oreilynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> [last accessed 25-05-2010]

[O'Reilly, 2007] O'Reilly, T., Design Patterns and Business Models for the Next Generation of Software, <http://www.oreilynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>, accessed 2007- 08-22.

[Odlyzko 1994] Andrew M. Odlyzko Tragic Loss or Good Riddance? The Impending Demise of Traditional Scholarly Journals, JUCS 1994 pp.3-53

[OECD, 2010] Giving Knowledge for Free THE EMERGENCE OF OPEN EDUCATIONAL RESOURCES

[Oppenheim, 2008] Oppenheim C. Electronic scholarly publishing and open access. *J Inf Sci* 2008; 34: 577-90.

[Oreilly 2004] Oreilly, Tim, What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software, Communications & Strategies, No. 1, p. 17, First Quarter 2007. Available at SSRN: <http://ssrn.com/abstract=1008839>

[Park and Park, 2006] Park, G., Park, Y., “On the measurement of patent stock as knowledge indicators”, in *Technol Forecast Soc Change* 73 (7), pp. 793–812, 2006.

[Patterson, 2007] Patterson, D.J., EOL Biodiversity Informatics: Designing for a Living Encyclopedia of Life, eolinformatics.mbl.edu/Documents/WorkPlan.pdf, accessed 2007-08-22.

[Pipek et al 2002].Pipek, V., Hinrichs, J., Wulf, V. 2002. Sharing Expertise Challenges for Technical Support. In Ackerman, M./Pipek, V./Wulf, V. (eds): *Beyond Knowledge Management: Sharing Expertise*, MIT-Press, Cambridge MA. 111-136.

[Polanyi, 1976] Polanyi, M. (1967), *The Tacit Dimension*, Doubleday, Garden City, NY, ISBN 0-385-

[Popescul and Ungar, 2003] Popescul, A., & Ungar, L. H. (2003). Structural logistic regression for link analysis. In *Proceedings of the Second International Workshop on Multi-Relational Data Mining* (pp. 92–106). Washington, DC: ACM Press

[Postellon 2008] Postellon D.C. 2008. Hall and Keynes join Arbor in the Citation Indices. *Nature*, 452, 282. doi:10.1038/452282b

[Puntschart and Tochtermann, 2006] Puntschart, I. and Tochtermann, K., “Online-Communities and the "un"-importance of e-Moderators”, in *Proceedings of Networked Learning 2006*, Lancaster (UK), April 2006

[Roberta et al, 2010] Roberta Cuel, Diego Ponte, Alessandro Rossi
Towards an Open/Web 2.0 Scientific Publishing Industry? Preliminary Findings
and Open Issues.

[Rodriguez and Bollen 2008] Rodriguez, M.A., Bollen, J. 2008. An
Algorithm to Determine Peer- Reviewers. In the Proceeding of the 17th ACM
conference on Information and Knowledge Management (Napa Valley,
California, USA, Oct. 26-30 2008), CIKM'08. ACM Press. 319-328.
doi:10.1145/1458082.1458127.

[Rogers, 2003]Rogers, E. M. (2003). Diffusion of innovations (5th ed.).
New York: Free Press.

[Rollett et. al, 2005] Rollett, H.; Lux, M.; Strohmaier; M., Dösinger, G.;
Tochtermann, K.; The Web 2.0 way of learning with technologies, in: Int. Journal
of Learning Technology, 2005.

[Sanger 2009] Sanger, L. M. (2009). The Fate of Expertise after
Wikipedia. Episteme, 6(1), 52-73

[Sauermann et al 2008] Sauermann, L., Cyganiak, R., Ayers, D. and Vlkol,
M., "Cool URIs for the Semantic Web. W3C Interest Group Note (2008)",
<http://www.w3.org/TR/2008/NOTE-cooluris-20081203/>

[Scharnhorst and Wouters 2006] Scharnhorst, A., Wouters, P. "Web
Indicators – a new Generation of S & T Indicators", in international journal of
scientometrics, Informetrics and Bibliometrics, Vol. 10, issue 1, 2006

[ScienceNews, 2007] Science News, Extreme Encyclopedia: Every living
thing will get its own page.
<http://www.sciencenews.org/articles/20070512/fob7.asp>, accessed on 2007-08-22

[Shneiderman 2002] Shneiderman, B. 2002. Inventing Discovery Tools:
Combining Information Visualization with Data Mining. Information
Visualization, 1,1 (march 2002), 5-12. doi:10.1007/3-540-45650-3_4

[Sorenson and Singh, 2006] Sorenson, O. and Singh, J., "Science, Social
Networks and Spillovers" (December 26, 2006). Available at SSRN:
<http://ssrn.com/abstract=953731>

[Spoerri 2004] Spoerri, A. 2004. RankSpiral: Toward Enhancing Search Results Visualizations. In Posters Compendium, IEEE Symposium on Information Visualization (Austin, Texas, USA, Oct. 10-12 2004) InfoVis'04. ACM Press. 39-40. doi:10.1109/INFVIS.2004.56.

[Suchanek et al 2007]. Suchanek, F. M., Kasneci, G., and Weikum, G., "Yago: A Core of Semantic Knowledge – Unifying WordNet and Wikipedia." In: Proc. 16th International World Wide Web Conference (WWW 2007), 2007.

[Surowiecki, 2004] Surowiecki, James (2004). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. New York: Doubleday, 2004.

[Szybalski, 2005] Szybalski, A. . Why it's not a wiki world (yet). Retrieved 11/4, 2005 from http://www.stanford.edu/~andyszy/papers/wiki_world.pdf

[Tauberer, 2006] Tauberer, J., What is RDF?, 2006 [Online] Available: <http://www.xml.com/lpt/a/1665> [Accessed May 21, 2008].

[Tho et al 2007] Tho, Q.T., Hui, S.C., and Fong, A.C.M. 2007. A Citation Based Document Retrieval System for Finding Research Expertise, Information Processing and Management, 43, 1 (January 2007), 248-264. doi:10.1016/j.ipm.2006.05.015.

[Tho et al 2007] Tho, Q.T., Hui, S.C., and Fong, A.C.M. 2007. A Citation Based Document Retrieval System for Finding Research Expertise, Information Processing and Management, 43, 1 (January 2007), 248-264. doi:10.1016/j.ipm.2006.05.015.

[Thompson, 2005] Thompson, B., "What is it with Wikipedia?", BBC, December 16, 2005

[Tsai 2001] Tsai, W. "Knowledge Transfer in Intra-Organizational Networks: Effects of Network Position and Absorptive Capacity on Business Unit Innovation and Performance", in Academy of Management Journal, 44(5), 996-1004, 2001.

[Tummarello et al, 2007] Tummarello, G., Delbru, R., Oren, E (2007). Sindice.com: Weaving the open linked data. In: Proceedings of International Semantic Web Conference (ISWC). Busan, South Korea.

[Ullrich et al 2008] Ullrich, C., Borau, K., Luo, H., Tan, X., Shen, L., & Shen, R. (2008). Why Web 2.0 is Good for Learning and for Research: Principles and Prototypes. World Wide Web Conference 2008 (pp. 705-714).Beijing, China: International World Wide Web Committee (IW3C2)

[Us Saeed et al 2007] Us Saeed, A., Stocker, A., Hoefler, P., Tochtermann, K., "Learning with the Web 2.0: The Encyclopedia of Life", in Conference ICL2007, Villach, Austria, 2007.

[Us Saeed et al 2008a] Us Saeed. A, Afzal, M.T.,Latif, A., Stocker, A., Tochtermann, K., Does Tagging indicate Knowledge diffusion? An exploratory case study, In Proc. of 3rd ICCIT pp.605-610 , 2008

[Us Saeed et al 2008b] Us Saeed, A., Afzal, M.T., Latif, A., Tochtermann, K., Citation rank prediction based on bookmark counts: Exploratory case study of WWW'06 papers, INMIC 2008. IEEE International pp. 392 - 397, Dec. 2008

[Us Saeed et al 2010] Us Saeed, A., Afzal, M.T., Latif, A., Tochtermann, K., Recommending tags for scientific resources, accepted for publication in Journal of IT in Asia (JITA), 2010

[Us Saeed et al, 2007] Us Saeed, A., Stocker, A., Hoefler, P., Tochtermann, K., Learning with the Web 2.0: The Encyclopedia of Life, in proceedings of ICL Conference, 2007,

[Vossen and Hagemann 2007] Vossen, G. and Hagemann, S. 2007. Unleashing Web 2.0: From concepts to creativity. Ubiquity 2007, December (Dec. 2007), 1-1. DOI= <http://doi.acm.org/10.1145/1331941.1331942>

[W3, 2010] <http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14.html> extracted on 29 May 2010

[Wagner & Bolloju, 2005] Wagner, C. & Bolloju, N. Supporting Knowledge Management in Organisations with Conversational Technologies: Discussion Forums, Weblogs, and Wikis, Editorial Preface, Journal of Database Management, 16/2 (2005), ABI/INFORM Global, 1

[Wagner & Bolloju, 2005] Wagner, C. & Bolloju, N. Supporting Knowledge Management in Organisations with Conversational Technologies: Discussion Forums, Weblogs, and Wikis, Editorial Preface, Journal of Database Management, 16/2 (2005), ABI/INFORM Global, 1

[Wan and Yang, 2007] Wan, X. , and Yang, J., Learning Information Diffusion Process on the Web, in WWW07, 2007

[Wilson, 2003] Wilson, E.O., The Encyclopedia of Life, in: TRENDS in Ecology and Evolution, Vol. 18, No.2 February 2003

[Wilson, 2007] Wilson, E.O., TED Prize wish: Help build the Encyclopedia of Life, 2007-08-22. <http://www.ted.com/index.php/talks/view/id/83>, accessed 2007-09-03

[Wu et al 2006] Wu, H., Zubair, M., Maly, K.: “Harvesting Social Knowledge from Folksonomies“, in HT, Odense Denmark, 2006,

[Wu et al, 2006] Wu, H., Zubair, M., Maly, K.: “Harvesting Social Knowledge from Folksonomies“, in HT, Odense Denmark, 2006,

[Yimam 1999] Yimam, D. 1999. Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach. In ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop (Copenhagen, Denmark, Sep. 12-16 1999). 276-283.

[Yimam 1999] Yimam, D. 1999. Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach. In ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop (Copenhagen, Denmark, Sep. 12-16 1999). 276-283.