Graz University of Technology
Faculty of Computer Science
Institute for Information Systems and Computer Media

Doctor of Philosophy Dissertation

Context Aware Information Discovery for Scholarly e-Community

by

Muhammad Tanvir Afzal

First Reader: Univ.-Prof. Dr.Dr.h.c.mult. Hermann Maurer
Second Reader: Univ.-Prof. Dr. Denis Helic

February 2010
Graz, Austria

to my parents, wife and children

Acknowledgments

First of all, I would like to thank the Almighty Allah, for His divine guidance and providence. His support, blessings, goodness, and kindness were always with me. He blessed me with motivation, passion, and hard work. It was his blessings which made me able to plan, visualize, and execute my dreams into the reality. I would like to dedicate this achievement to Him and to my parents, whose immeasurable sacrifices have led to what I am today.

Every Ph.D student dreams for a visionary supervisor. I am extremely thankful to Allah, for connecting me up with the best possible supervisor Professor Dr. Herman Maurer. Prof. Maurer's guidance, inspirations and motivations enabled me to bring out my best. I learnt many things from him. The list is quite long but to be short, Prof. Maurer introduced me to the scientific writing, guided me to the art of critical thinking, taught me to work hard, gave me freedom to explore different ideas, provided me funding for some conferences, linked me up with the best researchers like Prof. Wolf-Tilo Balke, Prof. Narayanan Kulathurmaiyer, and Prof. Denis Helic, and provided me the best working environment. I pay my deep regards to him and his greatness.

I am indebted to Prof. Denis Helic for being part of my dissertation evaluation committee and taking the role of second reader. The discussions/comments of Prof. Helic were very useful for my research and thesis. When I arrived in Graz, apart from my supervisor, there was another personality, Prof. Narayanan Kulathurmaiyer, who helped me in my initial time both academically and socially. Prof. Kulathurmaiyer was always there for me whenever I need him. He always encouraged me, motivated me, commented on my work, and taught me many research lessons. I cannot forget Doner Kebab which we often eat together and discuss on different aspects of life, and off course about research topics. Thanks to Prof. Maurer and to Prof. Kulathurmaiyer for arranging my trip to Malaysia. It was very nice time for me where I presented my research ideas and got enough feedbacks and a possibility of collaborations under different projects.

During my PhD studies, I got another opportunity to work under the guidance of one of the best researchers Prof. Wolf-Tilo Balke at Braunschweig, Germany. It was fantastic time when I groomed academically. His research lessons made me able to even think more deeply. His suggestions, comments and feedbacks enabled me to reshape some of my ideas. Subsequently, we had good publications. "Thank you!" for arranging accommodation there in Braunschweig. Thanks to Thomas Mack, my flatmate in Germany, for providing me nice company and to all other staff of Institute for Information Systems, technical University Braunschweig. I got another incredible opportunity to work with Dr. Muhammad Abulaish, one of the best researchers and graduates from Indian Institute of Technology, India. His feedbacks on ontologies part were very useful for my research.

I am grateful and thankful to all the colleagues and friends at IICM. I would like to say "heartiest Thanks" to: Maria Luise Lample and Gabriele Leitner, for providing me the best administrative assistance; Dana Kaiser, for assisting me in the matters of J.UCS and sometimes proof-reading my research work; Ingeborg Schinnerl, Dana kaisar and Maria Luise Lample, for making efforts to reissue my visa while I forgot my visa card in Graz and went back to Pakistan. Thanks to all of them for translating my problem to visa lady in German language and making it possible for me to come back Graz on time; Christian Gütl, for valuable comments and suggestions, Keith Andrews, for proposing a nice visualization technique for expertise data and giving me a privilege to write a nice paper with him; Helmut Leitner and Robert Hoffmann, for providing me guidance in Hyperwave and J.UCS technical issues; Nermin Kajtazovic, for developing a piece of software, during his bachelor thesis which helped me to expand my research activities; Karlheinz Trummer, for providing services as a system administrator; and to Franjo Bratic, Christian Safran, Nick Scerbakov, Walter Schinnerl, Christoph Trattner, Katharina Ziegler for providing nice company and encouragements. I would like to thank Bilal Zaka, Salman Khan, and Mohammad Al-Smadi for their moral support and motivations. I would like to thank my collaborators in Know-Centre like Prof. Klaus Tochtermann, Anwar us Saeed, Atif Latif, Alexander Stocker, and Patrick Hoefler.

I was lucky enough to find very loving and caring family of Pakistani scholars in Graz. They always made me happy while I was sad. Thanks to all who arranged and accompanied me in: playing cricket, watching movies, doing gym, arranging parties, and encouraging me all the time. I am unable to mention names of all scholars yet all of You are special to me. It was very nice stay at ElisabethHeim, thanks to all of my hostel fellows. It was a memorable stay at Hafnerriegel, and especially D-12 played a vital role in providing a platform for discussions and entertainment. I am speechless and feel proud to be a part of such a community.

I am very obliged to the Higher Education Commission of Pakistan, for awarding me a fully funded scholarship. Without it, I would not be able to pursue my studies. In the same spirit, I would like to thank Austrian Exchange Service for

their administrative support. After getting this scholarship, it was not easy for me to avail it because of some issues. I would like to thank my ex-boss Wg. Cdr. Muhammad Tahir Siddiqui, my friend Khalid, HEC project director Ishfaq Anwar, and all others who supported me to avail this opportunity and help me out to resolve the issues.

I am thankful to Atif Latif, Ilhem Benzaoui, Syed Nadeem Ahsan, and Shahzad Saleem for proof-reading some parts of this thesis and/or helping me to write it in LaTeX.

With my deep heart, I acknowledge my wife, who took care of my children while I was busy in my research activities. Her encouragement, motivations, support and understanding made me able to work on my research activities and to write this thesis. I would like to acknowledge my children: Muhammad Ahmad, Muhammad Umar and Aisha Tanvir for their patience while I was away for this noble cause.

Muhammad Tanvir Afzal
Graz, Austria, February 2010

Abstract

In digital environments such as electronic journals, finding context specific and task related information is a big challenge, simply due to the availability of the huge amounts of data. There are some of techniques to provide the intended information to users of digital communities (e-communities). However, all available techniques have some inherent problems. Thus, users are often frustrated.

This dissertation defines and implements a framework where most related information resources are linked and provided to users in a timely fashion. The framework of linking is composed of three sub-processes as follows: (i) discovering most relevant resources for linking, (ii) linking related resources, and (iii) supplying discovered and linked information to the users.

The relevant information resources are discovered using new techniques, mainly of heuristic nature. The proposed techniques often outperform existing ones and are able to discover highly relevant resources more effectively. Existing resources are annotated with links to the newly discovered resources. The resources discovered are supplied to users based on the users' local context and task at hand.

For a digital journal, this framework contributes in the following four areas: (i) Finding most related papers from multiple sources, (ii) Discovering and visualizing the relationships between experts, (iii) Linking digital journals with digital libraries maintained by some community leading to serendipitous discoveries of other relevant resources, and (iv) harvesting pertinent resources from Linked Data and hence helping the users efficiently.

The dissertation starts by summarizing the state-of-the-art in the field. It proposes and implements new techniques and heuristics which prove to be better than existing approaches. For example, the newly developed citation mining technique is able to overcome limitation of Google Scholar, CiteSeer and the ISI web of knowledge.

The heuristics for acquiring information from the Web can reduce the number of choices from millions to a few related resources for users; based on a multi-faceted approach, the expertise mining technique is able to rank experts for scholarly e-community more accurately.

The interlinking of digital journals with other digital libraries provides a platform for discovering newly evolving fields and concepts. Linking digital journals with semantic resources (Linked Data), using a concept aggregation framework provides a coherent view of informational aspects of authors in one place.

The findings of this research have been implemented for a digital journal which is known as Journal of Universal Computer Science step by step since 2007, and are now either in productive or prototype use.

Zusammenfassung

In digitalen Bibliotheken von z.B. Zeitschriften ist das Auffinden von kontextspezifischen Zusammenhängen und aufgabenorientierten Informationen durch die riesigen Datenmengen eine große Herausforderung. Es gibt eine Reihe von Techniken, um Informationen für die Nutzer von "Digital Communities" (E-Communities) zur Verfügung zu stellen. Aber alle verfügbaren Techniken und Methoden haben spezifische Probleme. Durch diese ist es oft schwierig, ohne sehr viel manuellen Aufwand die gewünschten Informationen zu finden.

Diese Dissertation stellt ein System vor, welches die relevantesten Informationsquellen miteinander verbindet und den Benutzern rechtzeitig zur Verfügung stellt. Dieses System besteht aus drei Modulen: (1) Ermittlung der wichtigsten Ressourcen für die Verknüpfung, (2) Verknüpfung von relevanten Ressourcen, und (3) Bereitstellen der ermittelten und verknüpften Informationen für den Benutzer.

Die relevanten Informationsquellen werden mit neuen Techniken und Heuristiken ermittelt. Diese Verfahren haben sich in bestimmten Szenarien als effizienter als bekannte Ansätze herausgestellt und konnten hochrelevante Ressourcen effektiver ermitteln. Die vorhandenen Ressourcen werden mit den Links zu den neu entdeckten Ressourcen versehen. Die entdeckten Ressourcen werden dem Benutzer präsentiert, und zwar angepasst an den lokalen Kontext und die aktuelle Aufgabe angepasst.

Für digitale Zeitschriften ergibt dies Vorteile in vier Bereichen dar: (1) Die relevantesten Veröffentlichungen werden aus verschiedenen Quellen ermittelt; (2) Experten werden ermittelt und Gruppen von Experten entsprechend visualisiert; (3) Digitale Zeitschriften werden mit digitalen Bibliotheken zur weiteren Recherche verknüpft; (4) wichtigen Ressourcen werden aus dem verknüpften Material zusammengetragen.

Diese Dissertation beschreibt zunächst den gegenwärtigen Stand der Forschung in diesem Aufgabenbereich. Dann werden innovative Techniken und Heuristiken vorgeschlagen und implementiert. So wurde zum Beispiel die neu entwickelte Citation-Mining-Technologie entwickelt, welche in der Lage ist, manche Grenzen

von Google Scholar, CiteSeer und dem ISI Web of Knowledge zu überwinden.

Die Heuristiken für den Erwerb von Informationen aus dem Internet verringern die dem Benutzer zur Verfügung stehenden Quellen von mehreren Millionen auf einige wenige, hochrelevante. Basierend auf diesem vielschichtigen Ansatz wurde die Expert-Mining-Technologie zur Einschätzung der Qualität von Experten in wissenschaftlichen digitalen Communities entwickelt.

Die Vernetzung von digitalen Zeitschriften mit gemeinschaftlich verwalteten digitalen Bibliotheken stellt eine Plattform zur Verfügung, um aktuelle und in Entwicklung befindliche Forschungsgebiete besser verfolgen zu können. Durch die Verknüpfung von digitalen Zeitschriften mit semantischen Ressourcen (Linked Data) kann das Concept Aggregation Framework einen kohärenten Blickwinkel auf die Informationsbedürfnisse eines Autors bieten. Die Forschungsergebnisse dieser Dissertation wurden für die digitale Zeitschrift Journal of Universal Computer Science umgesetzt. Dieses System wurde ab 2007 nach und nach in Betrieb genommen.

Contents

Chapter 1

Introduction

This chapter gives a short overview of the history of digital libraries, their importance and current challenges. Furthermore, it describes the objective, motivations and contributions of the thesis to the field of digital libraries.

This chapter is divided into four sections. The first section describes the history of digital libraries. In the second section, terminologies used in this thesis are discussed. Current research trends and challenges related to digital libraries are listed in the third section. The objectives, motivations and contributions of the thesis are discussed in the forth section. The overview of the structure of the thesis is presented thereafter.

1.1 History

Before starting to discuss aspects of modern digital libraries, it is useful to mention ideas in the past. Many ideas that are now starting to be accepted go back a surprisingly long time. Thus, the field of digital libraries is much older than one might think at a first glance. It has undergone an incremental progress with the contributions of numerous people. Some of them remain prominent because their ideas have inspired future generations. Among the number of outstanding scholars, Vannevar Bush, J. C. R. Licklider and Ted Nelson are worth mentioning here. Now we will cite some key notions related to the scope of this thesis from the aforementioned scholars. Subsequently, we will mention how different prototypes and real systems came into existence. This also highlights the vision of digital libraries and sheds deeper insights about "what we should have" and "what we actually have".

When World War II was near its end, Vannevar Bush, the former director of the wartime office of Scientific Research and Development, wished that scientists shift their energies from war effort to the process of constructing a huge repository of human knowledge. According to Bush, this repository should be made accessible to scientific community and provide extended functionalities that were not common at that time. In his historical paper published in The Atlantic Monthly in 1945 entitled *"As we May Think"* [Bush 1945], he described one of the unconventional library systems called *"memex"*. The Infostructure he described paved the way for Hypertext and helped in the realization of what we now know as the Internet. Bush described the term "record" as follows:

"A record if it is to be useful to science must be continuously extended, it must be stored, and above all it must be consulted".

The dynamic nature of a record is obvious from the aforementioned statement. The record must remain up-to-date and needs to be extended. A record must not be interpreted as the term used in the field of databases where *"record"* refers to a single entry. However, here *"record"* means object or resource stored in digital libraries. We will refer to this term as *"information object"* hereafter.

In the 1960s, several people at the Massachusetts Institute of Technology (MIT) investigated the power of modern digital computing and studied how it could transform libraries of that time. One of the prominent scholars was J. C. R. Licklider. Although the main concern of both Bush and Licklider was the literature of science, Licklider focused on the advent of modern computing. Thus, he predicted upcoming trends that have subsequently occurred [Arms 2000].

In 1965, Licklider wrote a book entitled "The Library of the Future". Unfortunately, this book is not available on the internet and is less famous than the article published by Bush. Licklider explained the required research and development needed to construct a really functional digital library. Although digital computing was not so powerful at that time, he yet predicted the developments which could be made in the next thirty years. In 1994, the predictions proved extremely perfect as an overall vision. Generally, Licklider expected less about what would be achieved by using large amounts of cheap computer power, and expected more about how much developments could be made from Artificial Intelligence (AI) and Natural Language Processing (NLP) [Arms 2000].

Another famous scholar in the field of digital libraries is Theodor Holm Nelson, shortly known as Ted Nelson. He explained a special logic structure for a world of universal digital media, where digital media can be annotated freely, viewed and linked side-by-side. Today's World Wide Web is only a partial implementation of his overall vision. He coined the term Hypertext [Nelson 1965]:

"Let me introduce the word "hypertext" to mean a body of written or pictorial material interconnected in such a complex way that it could not conveniently be

presented or represented on paper. It may contain summaries, or maps of its contents and their interrelations; it may contain annotations, additions and footnotes from scholars who have examined it. Let me suggest that such an object and system, properly designed and administered, could have great potential for education, increasing the student's range of choices, his sense of freedom, his motivation, and his intellectual grasp".

To fully realize the overall vision, Nelson dreamed of a visionary project called Xanadu ® [Xanadu 2009]. This project was started in the 60's and is still under the process of development. In 1998, the source code of Xanadu was released as project Udanax and in 2007, an initial working system called XanaduSpace 1.0 was released [Nelson et al. 2007].

Although ideas for managing world's digital media were floating since 1945 yet there was no real implementation of a working system. Douglas Engelbart, the inventor of the mouse, was another independent researcher who invented the computer-based hypertext system. In the 50s, Engelbart started thinking in different directions how to enhance human intellect with computers. Early in the 60s, he began building what became NLS (oNLine System), one of the first two computer-based hypertext systems. The online demo of the system is available in [Engelbart, 1968].

Andy van Dam and his colleagues at Brown University developed three notable hypertext systems: HES [Carmody et al. 1969], FRESS [DeRose and Dam 1999], and Intermedia [Yankelovich et al. 1987]. Andy van Dam working with his friend Ted Nelson developed the mid 60's system called HES (Hypertext Editing System). HES was originally going to be built on Ted's hypertext vision, but the project turned to an emphasis on printout and formatting, and can be seen as a prototype of the word digital systems of today. After meeting with Douglas Engelbart, Andy van Dam developed in the late 60's, a system called FRESS (File Retrieval and Editing System). FRESS was sponsored by a National Endowment for the Humanities and was used for over two decades at Brown University for personal hypertext libraries and courses. In the 1980s, Norm Meyrowitz, after collaborating with Andy Van Dam developed Intermedia. Intermedia was probably the best-realized and most sophisticated hypertext system of the mid-eighties. This system was used extensively in Brown University classes in a variety of topics.

In his presentation, at Ars Electronica 2009 entitled *"Before the Internet"*, Hermann Maurer explained the precursor activities of the Web [Maurer 2009]. In Europe, it all started with interactive Videotex. Videotex was based on the simple observation that most families have a TV set and a telephone. The idea was to use network of computers (servers) for information/services, to get access via phone and modem. A decoder connects to TV as display device and remote control serves as input device. The idea was proposed by Sam Fedida, British Telecom, in 1976 under name PRESTEL which was then known as interactive videotex. In Austria (1979-1982), Maurer and Posch decided to develop a Z80 based color graphics computer called MUPID [Maurer and Posch 1982] [Maurer 1982]. MUPID stands for Multipurpose Universally Programmable Intelligent Decoder. MUPID was useable as full-fledged personal computer since it allowed to download what was then called *"telesoftware"*, today known as *"JAVA Applets"*. More than 50, 000 MUPIDs were produced in Austria during 1982-1989. The networked learning was already available in 1986 as 500 hours of *"COSTOC"* lessons using MUPID. The MUPID system was pushed out by PCs and the WWW.

In the mid-end 80's, a number of attempts to use the emerging Internet as a basis for networked systems started. Three of the systems are worth mentioning here: Gopher, Hyper-G and WWW. The development of all of three mentioned systems took place simultaneously. In 1990, Tim Berners-Lee and Robert Cailliau made a proposal titled *"Information Management: A Proposal"* at CERN. This proposal aimed at managing the distribution of physics documents rather than creating an interactive medium for managing human knowledge. The proposal was accepted by the director Mike Sendall and work started. Meanwhile, Hermann Maurer and Ivan Tomek laid the basis for Hyper-G [Maurer and Tomek 1990]. The Hyper-G system was inspired from the ideas of Ted Nelson and the aim was to make a real hypertext system. At that time Mark McCahill was making his efforts in the development of a TCP/IP protocol for distributing, searching, and retrieving documents over the Internet which became the basis for Gopher system. The Gopher combined document hierarchies with collections of services, including WAIS, the Archie and Veronica search engines, and gateways to other information systems such as ftp and Usenet. In 1991, The Gopher server was released by Mark McCahill and his colleagues at the University of Minnesota. In the same year (1991), the first WWW server was released at CERN [Cailliau 2006] and the first Hyper-G application was released in Graz [Maurer 1996]. In 1993, a group at Illinois University wrote the first Web-client called *"Mosaic"* which popularized the Web and started a boom in 1994. The complete version of Hyper-G system was released in 1994. However, the developers of Mosaic implemented only a subset of the hypertext-model defined by Nelson [Nelson 1965]. The ideas of bi-directional links, annotations and document version management were not incorporated in the design of the Web because the development was originally for a one way distribution of physics documents with limited functionality. Therefore, it is now almost impossible for users to make annotations to the documents stored on web-servers. This, however, may be achieved by extending the standard implementations. The ideas described in the hypertext-model such as: bi-directional links, integrated metadata management, links management and information clusters were implemented in the Hyper-G system [Maurer 1996]. In view of the fact that the defined features are essential for a rather practical digital library, Hyper-G technology was opted instead of simple web-server technology to host a digital journal *"Journal of Universal Computer Science"* (J. UCS) [Calude et al 1994]. J. UCS is explained in details in section 1.2.4.

Some of the largest electronic libraries are libraries of learning modules such as in the ARIADNE project by Eric Duval, one of the most successful early projects started in 1996 and still growing [ARIADNE 2009]. The core of the ARIADNE infrastructure is a distributed network of learning repositories called Knowledge Pool System (KPS) which has actively been used in academic and corporate already for a considerable time.

1.2 A Note on Terminology

This section explains basic terminologies used in this dissertation.

1.2.1 Digital Library

The terms Electronic or Virtual library are often used synonymously for the term digital library. There is no standard definition of "digital library" because people from different disciplines are working in this field. As a result, different jargons are used for the same meaning. To understand digital library as an overall vision, various definitions of the term belonging to different contexts, disciplines and understandings are summarized here. As illustrated in the following definition, it is not easy to define the term digital library. In the 1990's, there were many initiatives started in the development of digital libraries. Edward A. Fox [Fox 1993] prepared a Source Book on digital libraries for the National Science Foundation and concluded the following:

> *"One group was supposed to define the library. It came back with a statement that a digital library is a distributed technology environment which dramatically reduces barriers to the creation, dissemination, manipulation, storage, integration and reuse of information by individuals and groups. They suggested the national initiative should contain some specific testbed projects, but gave no guidance on what these should be. In other words, they not only failed to define the collection, they didn't really even describe the system that would hold it".*

In 1995, the Association of Research Library recapitulated the definitions of a digital library as given below [Association of Research Library 1995]:

- A digital library is not a single entity

- A digital library requires technology to link the resources of many digital libraries

- Linkages between the many digital libraries and information services are transparent to the end users
- Universal access to digital libraries and information services is a goal
- Digital library collections are not limited to document surrogates: they extend to digital artifacts that cannot be represented or distributed in printed formats

There are two main streams of communities working in the area of digital library 1) research community and 2) practice community. In the research community, most of the members who have a computer science background are focused on the experimentation and developmental research to deal with technology applications in various areas. Although there is a perception that this research will result in working and functional digital libraries yet the objective is not connected to the concrete operations but to research. On the other hand, the practice community is focused on building operational digital libraries, maintenance of digital libraries, and providing services to users. The approach is extremely practical with less research work involved [Saracevic 2000].

There exists no standard definition of digital libraries for research community. The following represents the closest summary of the approaches adopted by the research community as proposed by [Lesk 1997].

"Digital Libraries are organized collection of digital information. They combine the structure and gathering of information which libraries and archives have always done, with the digital representation that computer have made possible".

The Digital Library Foundation (DLF) is an organization in the United States which was formed in 1995 and which represents the practical community. The declared objective of DLF is *"to establish the conditions necessary for the creation, maintenance, expansion, and preservation of a distributed collection of digital materials accessible to the scholars and the wider public"* [DLF 1998]. After substantial efforts, the DLF agreed on a working definition of a digital library, denoting a definition of the practice community.

"Digital Libraries are organization that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities." [DLF 1998].

In 1999, Borgman [Borgman 1999] presented a complex definition of digital libraries which included extensive discussions. This definition can be treated as a bridge between research community and practical community as stated below:

- Digital Libraries are set of electronic resources and associated technical capabilities for creating, searching, and using information ... they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium ...The content of digital libraries includes data, [and] metadata ...
- Digital libraries are constructed, collected, and organized, by (and for) a community of users and their functional capabilities support the information needs and uses for that community.

In 2000, William Y. Arms provides an informal definition of digital libraries as follows:

"A digital library is managed collection of information, with associated services, where the information is stored in digital formats and accessible over the network. The crucial part of this definition is that the information is managed" [Arms 2000].

European Commission, in the frame of the Information Society Technologies Programme (IST), founded a Network of Excellence on Digital Libraries known as DELOS. The DELOS network is focused on the incorporation and organization of ongoing research activities of the major European teams working in Digital Library or in associated areas with the objective of developing the next generation Digital Library technologies [DELOS 2009].

In 2007, DELOS prepared a Digital Library Reference Model and defined the term Digital Library as follows:

"An organization, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies". [Candela et al. 2007].

The Association for Computer Machinery categorized Digital Libraries (H.3.7) as a specialization of Information Storage and Retrieval (H.3) in Information Systems (H) [ACM-CCS 1998].

1.2.2 Information Object

Digital libraries archive, process, maintain, and deliver different kinds of resources including text, audio, video, metadata, annotations, 3D scenes, interactive contents, and geographical data etc. The metadata was part of Hyper-G system when it was defined in 1991 [Maurer 1996] but it was only taken up much later by Dublin Core [Dublin Core Metadata Initiative 2001], and has only started to be considered important recently. The ARIADNE project described above also uses metadata for learning objects to share and reuse information [Duval 2000a] [Duval 2000b]. There exists no standard term for the resources stored in a digital library. [Arms 2000] suggests material or item, but different terms like electronic document, entity or object can be found in the literature. DELOS recently described it as content where content is defined as an umbrella concept used to aggregate all forms of information objects that a Digital Library collects, manages and delivers. [Candela et al 2007]. In this thesis, we will refer to this term as information object.

1.2.3 Electronic Publishing

According to Alliance for Telecommunications Industry Solutions [ATIS 2000], electronic publishing is *"the process of creating messages, distributing them, and reproducing them entirely online, often with a capability for feedback. Note: Unlike desktop publishing, electronic publishing does not usually generate hard copy."*

Schmaranz also describes the term 'Electronic Publishing' [Schmaranz 1998]. Electronic Publishing is seen as a kind of *"Electronic Paperware"*, that is the content of the documents is mostly text-based and non-interactive. There are numerous advantages of electronic publishing over traditional publishing including searching and hyperlinks. Any type of document can be published and further navigational capabilities can be incorporated by hypermedia systems. Schmaranz entitles this approach as dynamic interactive hypermedia publishing [Schmaranz 1998].

Hitchcock [Hitchcock 2002] discussed different perspectives in electronic publishing. According to his analysis, there exist two views on electronic publishing: one view can be described as publishing with the help of new technology [Graham 2001]; the other one is to treat the media incorporation of modern technology or 'Multimedia'. According to Hitchcock, both views misinterpret the impact of modern technology on the published product. In reality, the product is experienced by the end users rather than the publishers. According to users' concerns, electronic publishing primarily in the form of networked publishing on the internet, and World Wide Web-based publishing, will be publishing reinvented. Online publishing eliminates the restriction of physically packaged products such as books, journals and CD-ROMs.

1.2.4 Electronic Journals

Electronic journals publish manuscripts online and offer indexing, searching, interactive visualizations to users, and a number of functionalities. Online publishing has reduced the price of journals significantly as well as it has made searching easy for its users. The published output of the worldwide scholarly community has risen exponentially. Odlyzko [Odlyzko 1994] has shown the effect in the field of mathematics where over a period of more than a century the number of published papers annually doubled every 20 years, but after World War II that doubling happened every 10 years. One of the reasons for having electronic journals is this information overload. This huge information need to be processed at lower cost [Odlyzko 1998]. This is why electronic journals progressed so fast. If we look at the history of scientific journals, the first one appeared in 1665 in London (Philosophical Transactions of the Royal Society of London) and Paris (LeJournal des Scavans) simultaneously [Guedon 2001]. Therefore, the publication of the first journal took about 200 years since the invention of print, while it took only 20 years to get first electronic journal since such a possibility was discussed [Senders 1977]. According to a survey [Hitchcock 2002], at the end of 1995, there were only 100 peer-reviewed electronic journals and in the year 2001, there were more than 10,000 available electronic journals.

The possibility of making this huge repository accessible to all scientific community was discussed in [Marchionini and Maurer 1995]. This was followed by an open access movement [Roberts et al 2001]. Before this movement, readers had to pay subscription fee to access human knowledge. But it was argued by [Marchionini and Maurer 1995] [Roberts et al 2001] to make scientific knowledge accessible to all users of the Web free of charge. With this movement, open access journals have subsequently emerged. The Journal of Universal Computer Science [J. UCS 1994] is one of them and is explained in details in the next paragraph. However, the open access is not an ideal situation as there is a danger that low quality manuscripts are accepted to make money from authors. The belief that all can be done without money at high quality is just not correct. Someone hast to pay, the reader, the author, or some institution. The high quality material needs an effort that has to be financed. In the case of J.UCS, different institutions contribute to its survival, but there is a danger that open access journals may suffer when no such funding is available.

The Journal of Universal Computer Science (J.UCS) is a high-quality electronic publication that deals with all aspects of Computer Science [Calude et al. 1994]. J. UCS has been appearing monthly since 1995 with uninterrupted publications. According to the survey paper on electronic journals [Liew and Foo 2001], J.UCS has incorporated innovative features such as the enabling of semantic and extended search and its annotative and collaborative features. It was one of the first electronic published journals to have implemented features such as personal and public annotations, multi-format publications, multi-categorization, etc. These features have made J.UCS a rather unique electronic journal. Readers of such high-quality electronic journals expect and anticipate highly sophisticated features, such as automatic reference analysis, similarity search between documents and other features using knowledge management technology [Krottmaier 2003]. Some of the features mentioned are included in the scope of this thesis as explained in section 1.4.2.

1.3 Research Trends and Challenges in Digital Libraries

The growth of digital information has increased exponentially. It has become a challenge to manage the huge quantity of information efficiently without having a real implementation of original hypertext-model. As the information objects are not annotated and do not contain bi-directional links, it is very difficult to find context specific information related to user local context. Although the intended information exists, the users battle the problem of finding context specific information. To judge how much new information is produced every year, we need to look at the statistics from the following research study.

The study [Lyman and Varian 2003] was performed by the faculty and students at the School of Information Management and Systems at the University of California at Berkeley in the year 2000 and was revised in the year 2003. The study sets to find how much digital information is produced every year. According to statistics presented in the study, nearly 5 exabytes of unique information are produced every year worldwide, which is around 800 megabytes per person living in the world. An exabyte is equivalent to billion gigabytes. More than 90% of this abundant annual output is stored digitally. No more than 0.01% of this information is stored in printed form. Little of this information is made available through Digital Library collections [DELOS 2001].

Some visionary ideas and challenges in Digital Libraries related to teaching and learning have been pointed out in [Marchionini and Maurer 1995]. The argument of authors is that digital libraries still lack a number of sophisticated functionalities. They should provide better links between related materials, support for informal learning, offer automatic answering services (e.g., Ackerman's Answer Garden system for handling XWindows questions [Ackerman 1993]). Digital libraries should serve teachers on demand, and provide sophisticated information access to novices and professionals. The novice user may seek guidance from professional users provided that both have access to the same information. All type of learners should mutually share and explore information and expertise. The information access rights and intellectual property laws hindered the availability of large information repositories to scientific community. However, in recent years, this issue has been resolved to some extent by the open access movement [Roberts et al 2001].

In a meeting at Beijing, the presentation "Beyond Digital Libraries" by Hermann Maurer shed deeper insights on key problems in the field of digital libraries [Maurer 2001]. The digital libraries should play an important role in information society by extending the standard functionalities and by incorporating sophisticated services. Some important points discussed were: making search mechanisms more powerful and interactive, incorporating metadata, annotations, active documents and creating cross-references. The search functionalities should consider grammatical/linguistic (stemming, fill-words, synonyms, natural language processing) and semantic features to provide all relevant but unnecessarily many results. The incorporation of metadata is helpful for getting all relevant results, yet the metadata is usually generated manually which needs to be automated. The chapter 3 of the thesis will discuss metadata extraction and processing in more details. The documents in digital libraries should be active documents [Heinrich and Maurer 2000]. The idea of active documents is to provide a help/guidance service to users. For example, users may ask any question related to a document and the system can generate an answer that suits the users. The basic functionality of this idea has been implemented in Hyperwave [Maurer 1996]. Digital libraries should also support annotations. Readers, authors and editors should be able to comment on a published work at any point in time. To create cross-references, one of the visionary ideas discussed by Maurer is Links into the Future. This feature will provide users with all relevant information that has been made available after the publication date of the source content. The idea of Links into the Future is described in more details in chapter 3.

Some more thoughts on *"What we expect from Digital Libraries"* are discussed by Dreher et al [Dreher et al 2004]. For a single library, it is not possible to store each and every scientific document. This requires creating cross-references between different libraries regardless of the storage-location. Personalization features can make user interaction more productive with the system. Dreher et al discussed intelligent search, e-learning support, conceptual search, white lists, interactive visualization, adaptive user interfaces as key areas of research and demanding functionalities in digital libraries.

The World Wide Web, is most probably the largest digital archive enabling wide range of different communities to make available large sets of diverse resources and information. This information is further used and linked by different digital libraries in their local settings. 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 contents. 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 pages [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. This becomes an inherited problem when systems like digital libraries want to utilize this huge information efficiently. 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 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 and 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 (will be referred to as Linked Data hereafter). 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 of similar concepts in these datasets. Currently, LOD cloud consists of over 4.7 billion RDF triples, which are interlinked by around 142 million RDF/OWL links [Auer et al 2009]. This initiative paved a way for different kinds of applications to discover more structure (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]. It is no trivial task to search the intended information from the mentioned big repository of Linked Data. There is a lack of friendly user interfaces and end users usually need to deal with complex semantic mechanisms to explore information.

Some key challenges in the field of Digital Library are summarized by DELOS. As explained in section 1.2.1, DELOS is a network of Excellence on Digital Libraries. In 2001, DELOS conducted a workshop with leaders in the field of digital library research. In a brainstorming session future directions were debated [DELOS 2001]. The participants of this meeting agreed on the following visionary statement:

"Digital Libraries should enable any citizen to access all human knowledge anytime and anywhere, in a friendly, multi-modal, efficient, and effective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices".

To make a strong foundation for the future-oriented digital library systems, DELOS identified the need for a Reference Model for Digital Library Management Systems. This reference model provides a formal and conceptual framework describing the characteristics of this class of information systems. One of the main objectives of this reference model was to study user requirements and the current functionality provided by digital library systems.

In 2006, DELOS prepared a reference manual entitled "Current Digital Library Systems: User Requirements vs. Provided Functionality" [Candela et al 2006]. This document is based on surveys from digital library users. The study analyzed five classes of functionalities:

1. Functions for locating information,
2. Functions for presenting resources,
3. Functions for personalization of content and services,
4. Facilities for communicating and collaborating with other DL users,
5. Other common DL functions (such as Social navigation support, Multilingual support, Personal annotation, notification/alerting services, Glossaries, Thesaurus, and Dictionaries, Printing / Print preview facilities, and Downloading / uploading facilities)

The DL functions at the provider site identified by users as of highest priority include:

1. Organizing resources;
2. Archiving resources;
3. Storing metadata about resources (creator, content, technical requirements, etc.);
4. Locating resources;
5. Creating cross reference links between similar resources, and
6. Storing metadata about resources was classified homophonous by stakeholders as of highest importance.

The key results of the study are as follows:

- In terms of locating information objects, higher scores were given to those associated with Search (e.g. keywords search, parametric search) and cross-referencing of information objects.

- In terms of personalization of information objects and services: higher ratings were given to functions supporting the Presentation of objects according to profiles, and to Bookmarks facility (i.e. Favorites);
- In terms of communicating and collaborating with other DL users: high scores were recorded for shared annotation facilities (e.g. peer reviews), and to e-mail services.

It is obvious from the above mentioned results that users want good search facilities, better cross-references between related "information objects", information object presentation according to user profile, bookmark facilities and shared annotation facilities. This thesis proposes and implements novel ideas in the aforementioned highly ranked functionalities in a domain of a digital journal, Journal of Universal Computer Science (J.UCS). Chapters 3, 4, 5, and 6 of this thesis talk about the context aware discovery of resources, creating cross-references between resources, and visualizing the resources to users by looking at the users' local context and task at hand for the following information objects:

- Papers
- Experts
- Shared annotations
- Authors/Experts' profiles.

The detail of each can be found in section 1.4.2.

The process of creating cross-references can be considered as a three tier process.

1. Locating related information objects to link.
2. Creating links.
3. Presentation to users.

To establish useful cross-references between relevant information objects, one needs to find both the potentially related information objects and a process of creating links between them. The process of accurate linking requires good algorithms to discover and link related information objects. The created links are then presented to the users. This thesis automates the process of creating links between related information objects for a digital journal and contributes to: 1) discovering potentially related information objects 2) Linking them and 3) presenting them to the users in context-aware fashion. The first two steps for different information objects (papers, experts, social tags and author's profiles from semantic resources) are explained in details in the chapter 3 onward. The third step, presenting the discovered related information objects, employs information supply paradigm. The conventional search technology employs a pull model which requires explicitly specified search terms by users. The relevant information is subsequently retrieved based on users' search terms. The results, however, depend upon the correctly formulated search queries. The thesis uses information supply model as proposed by [Maurer and Tochtermann 2002] and Broder [Broder 2006] to make available the related information objects whenever users need them.

1.4 Motivations, Thesis Objective and Contributions

1.4.1 Motivation

The scientific community agreed on the clearest statement and requirements for accessing published research papers - a complete collection that can be indexed, searched and linked efficiently. According to a recent paper published in Science, Roberts et al urged journal publishers, their editors and scientist to make available complete versions of scientific papers to public. However, the open access is not an ideal situation as discussed in section 1.2.4. Nevertheless, these open access journals have subsequently occurred. This huge repository of human knowledge further opens new ways for knowledge discovery. Related information objects can further be linked to make a 'dynamic digital archive' [Roberts et al. 2001]. This thesis focuses on searching and linking innovations that can be applied to openly available scientific literature. The thesis reflects the visionary thoughts of Roberts et al that *"will enable researchers to take on the challenge of integrating and interconnecting the fantastically rich, but extremely fragmented and chaotic, scientific literature"*.

Ted Nelson's vision of creating a hypertext system has already been explained in section 1.1. One of the main features of the hypertext model is that information objects should contain bi-directional links. In today's online environment, when an information object is created, it may point to previously available information objects yet, the old information objects do not necessarily contains a link to newly occurred content. Inspired by the hypertext model, Hermann Maurer suggests an idea of Links into the Future [Maurer 2001]. This thesis contributes to the creation of 'Links into the Future' for scientific literature. Links into the Future treat published papers as dynamic documents which are continuously extended with relevant links as soon as new papers are made available online. One important aspect of creating Links into the Future is linking documents based on citations. However, autonomous citation mining in itself poses some challenges [Giles et al 1998]. This thesis further develops a sophisticated autonomous citation mining technique. Furthermore, this technique is used to create Links into the Future.

With the availability of voluminous information online, searching has become a challenge. To reduce users' cognitive load of searching, Maurer and Tochtermann presented the push idea [Maurer and Tochtermann 2002]. The authors explained a new model for knowledge management. According to this model, the system can generate and offer knowledge without being explicitly asked by the users. Users do not need to make explicit queries instead related knowledge is pushed to users by observing user context and activities. This idea was further realized by Yahoo's vice president for search and technology, Andrei Broder, who described the 4th generation of Web search [Broder 2006]. This search generation uses a push model which requires no search queries from user, but the best suited information objects are pushed/supplied to users by observing users' local context and task at hand instead. Motivated from this, the thesis makes use of the push model where the most relevant information objects (paper, experts, social tags and authors' profiles) are pushed to users' local contexts.

1.4.2 Scope of the Thesis

As discussed in section 1.3, the thesis discovers, creates and supply links for four information objects: papers, experts, shared annotations, authors/experts' profiles. This section provides details for each of them. For the experimentation, the Journal of Universal Computer Science was used as a source dataset. The links were created within J.UCS, to the Web documents, CiteULike repository, and Linked Data resource.

Links into the Future

The thesis extends and implements the idea of Links into the Future [Maurer 2001]. The idea was implemented for all papers published in J.UCS. The system exploits multiple sources to retrieve relevant papers that have become available after the publication date of the focused paper. Links are further created and visualized to users. The system employs two techniques for creating these links: metadata extraction technique and citation mining technique. The metadata extraction technique presents ontological representation for Links into the Future. The ontological framework is further used to extract Links to future related resources from J.UCS and the Web. Moreover, a new autonomous citation mining technique named TIERL (Template based Information Extraction using Rule based Learning) is presented and implemented. The details of this system can be found in chapter 3.

Discovering Relevant Information from Socially Maintained Digital Libraries

Based on multiple studies, the thesis explores social bookmarking and studies relationships between tags and citations. Furthermore, relevant concepts from socially maintained digital libraries are discovered for J.UCS papers and related resources are visualized to users by observing users' local context. It has been discussed in chapter 4 in details.

Discovery and Visualization of Expertise

The thesis employs an automatic technique for discovering experts and expertise. This technique uses multiple experience-atoms to judge the overall expertise of an individual. The discovered experts are visualized using extended hyperbolic tree visualization and experts are also pushed to users in their local contexts. The chapter 5 elaborates this in details.

Harvesting Pertinent Resources from Linked Data

The thesis explains how the Semantic Web can add value to digital journals. By discovering relevant Linked Data resources in an intelligent and efficient way, the thesis establishes links between authors of the journal and authors profiles available in Linked Data. The thesis employs an innovative URI discovery technique for Linked Data resources. Subsequently, the extracted information is integrated with the help of presented Concept Aggregation Framework. This helps in hiding underlying complex semantic mechanics and helps users of journals to discover information instantly. Further details can be found in chapter 6.

1.4.3 Research Questions

This section describes research questions addressed in this thesis. Wherever applicable, we will break down these questions into more specific ones. First two research questions are addressed in chapter 3, research question 3 and 4 are fulfilled in chapter 4, and research question 5 and 6 are investigated in chapter 5, while research question 7 and 8 are explored in chapter 6. The research question 9 is about context aware delivery of resources as discovered from previous objective so it remains active throughout the thesis.

RQ1. How can a system for creating Links into the Future from multiple sources be developed?
RQ2. Can we improve existing citation mining techniques?
RQ3. How are tags and citations related? Do tags hold potential for measuring research popularity, if yes then how?
RQ4. How can important tag terms from social bookmarking be exploited by digital journals?
RQ5. How can experts be discovered and ranked in scientific community. Which metrics are the important ones?
RQ6. How can experts be visualized?

RQ7. How can we retrieve relevant information from Linked Data resources? Can we hide underlying semantic mechanics from end users?
RQ8. How can digital journals consume information from Linked Data resources?
RQ9. How can user be given only required and relevant information whenever they need it?

1.4.4 Foundation of the Dissertation

The foundation of the dissertation is a set of publications selected form the ones that have been authored or co-authored by the author of thesis over a period of some three years. Their relation and their arrangement in the dissertation are depicted in Figure 1.1 The thesis primarily discovers related information objects, creates cross-references for scientific literature and visualizes them to users by observing users' local context. The thesis makes contribution broadly in four areas 1) creating Links into the Future 2) discovering relevant social tags and attached resources from socially maintained digital libraries 3) discovering and visualizing expertise and 4) harvesting relevant informational aspects from Linked Data and their visualization. These research areas are discussed in chapter 3, 4, 5, and 6 respectively. Every chapter is based on set of 3-7 publications as depicted in Figure 1.1

1.4.5 Thesis Contributions

The thesis deals with the context aware information discovery of academic resources for scientific community. The thesis makes five contributions. First, the thesis implements and extends the idea of 'Links into the Future' where future related documents from multiple sources are pushed to users by looking at the user's local context. The system is able to reduce millions of generic search results to a few relevant ones. The user feedbacks show that the knowledge discovery by this system is very useful and has reduced the user cognitive effort required to find relevant resources. Second, it proposes and implements a citation mining technique named as 'Template Based Information Extraction using Rule based Learning' (TIERL). The comparisons with leading citation indexes show that the system can proactively find citations with high accuracy. The citation mining technique retrieves citations that would not otherwise be viewed. Third, it discovers potentially high profiled authors (experts) based on multi-faceted approach. The experts are visualized through extended hyperbolic visualization. The discovered experts are further pushed to users in their local context whenever users need them. The system is able to discover potential reviewers by aggregating multiple experience-atoms of experts. Fourth, based on multiple studies, it explores the shared metadata infrastructures of web 2.0 like tagging and bookmarking for making serendipitous discoveries of relevant popular concepts. The discovered re-

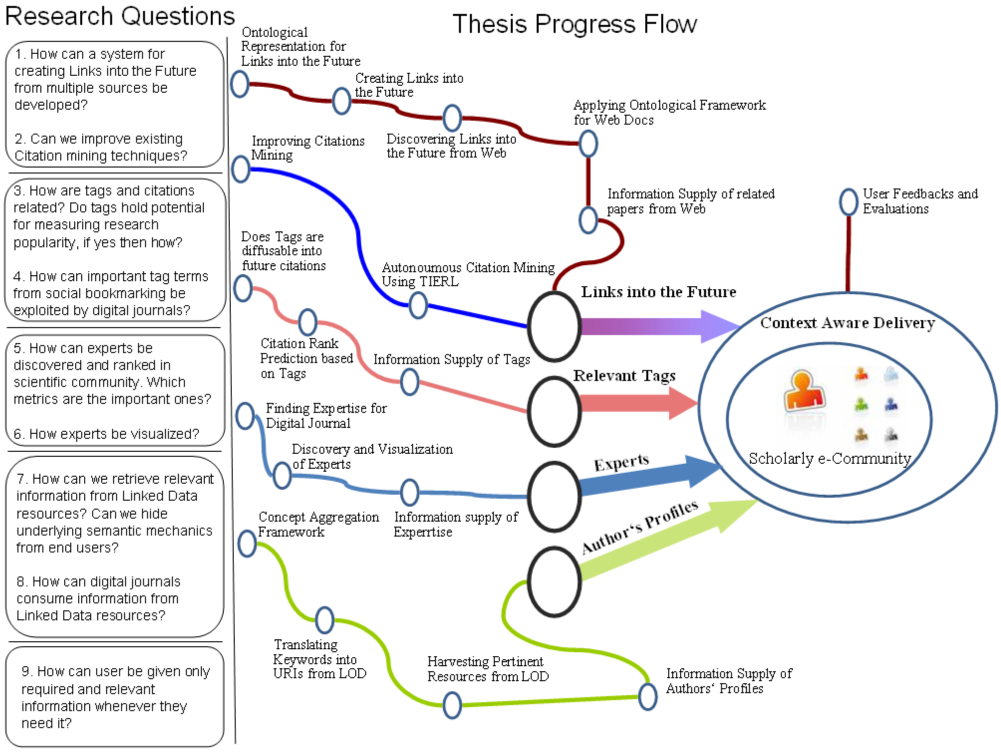

Research Questions
1. How can a system for creating Links into the Future from multiple sources be developed?
2. Can we improve existing Citation mining techniques?
3. How are tags and citations related? Do tags hold potential for measuring research popularity, if yes then how?
4. How can important tag terms from social bookmarking be exploited by digital journals?
5. How can experts be discovered and ranked in scientific community. Which metrics are the important ones?
6. How experts be visualized?
7. How can we retrieve relevant information from Linked Data resources? Can we hide underlying semantic mechanics from end users?
8. How can digital journals consume information from Linked Data resources?
9. How can user be given only required and relevant information whenever they need it?
Thesis Progress Flow
Ontological Representation for Links into the Future
Creating Links into the Future
Discovering Links into the Future from Web
Applying Ontological Framework for Web Docs
Information Supply of related papers from Web
User Feedbacks and Evaluations
Improving Citations Mining
Autonoumous Citation Mining Using TIERL
Links into the Future
Does Tags are diffusable into future citations
Citation Rank Prediction based on Tags
Information Supply of Tags
Relevant Tags
Context Aware Delivery
Scholarly e-Community
Finding Expertise for Digital Journal
Discovery and Visualization of Experts
Experts
Concept Aggregation Framework
Information supply of Experrtise
Author's Profiles
Translating Keywords into URIs from LOD
Harvesting Pertinent Resources from LOD
Information Supply of Authors' Profiles

Figure 1.1: Thesis Foundation

sources are further linked with the resources available in a digital journal. Fifth, the system characterizes a framework for utilizing semantic resources from Linked Data (LOD). It provides a user friendly interface for LOD named as CAF-SIAL (Concept Aggregation Framework for Structuring Information Aspects of Linked Open Data). Furthermore, using the framework, the system establishes links between J.UCS authors and their relevant resources available in LOD. The system is able to aggregate, structure and present authors' informational aspects from LOD in a comprehensive way.

All of the above defined research tasks were practically implemented and now either in productive or prototype use. The idea of Links into the Future (based on metadata extraction technique and citation mining technique) was implemented in the Journal of Universal Computer Science (J.UCS). The system is up and running since 2007. The ideas of expertise finding and discovery of relevant resources from socially maintained digital libraries have been made available in J.UCS since 2009. The system for linking J.UCS authors with Linked Data resources is in prototype use.

1.4.6 Thesis Organization

The current chapter serves as an introduction to the thesis explaining the history of digital libraries, research challenges and contribution in this field. The remaining parts of the thesis are categorized as follows. Chapter 2 describes state-of-the-art work in the respective research areas by highlighting existing methods, techniques and developed systems/prototypes. Chapter 3 elaborates the system of Links into the Future where metadata extraction technique is employed using ontological framework and new citation mining technique is introduced. How digital journals can be linked with socially maintained digital libraries is described in chapter 4. Using consolidated expert profile, the discovery and visualization of expertise in scientific community is elaborated in chapter 5. Chapter 6 sheds deeper insights on the required processes and techniques for discovering relevant information from Linked Data. The thesis implements all described ideas for a digital journal. 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.

Chapter 2

Related Work

The research presented in this thesis is novel in 1) discovering relevant information for contents stored at digital journal from various sources (journal itself, the Web, CiteULike, Linked Data) using multiple algorithms and heuristics, and 2) the incorporation, realization and supplying of the discovered information in users' local context. This chapter briefly describes existing state-of-the-art systems and highlights existing problems.

Section 2.1 introduces the idea of Links into the Future. The idea was architected and developed using ontologies and citation mining technique: first, a brief overview of ontologies followed by existing systems related to our work will be given. Then, an overview of citation mining techniques is elaborated along with problems faced by existing systems to find all relevant citations. Section 2.2 explains tagging and bookmarking system. It also points out existing metrics for measuring knowledge diffusion and prevailing systems for measuring citation rank. Section 3.3 briefly elaborates popular systems in discovering experts and expertise profiles. Section 3.4 discusses Linked Data, one of the most successful projects of the Semantic Web. It provides state-of-the-art for locating and consuming information from Linked Data.

2.1 Creating Links into the Future

Most related work explores the servicing of users within a present-context, making use of limited information, captured in-vivo. Our work, on the other hand, describes the augmentation and annotation of documents created in the past with information that became available later. In this way, a research paper is not seen as a static document, but rather one that is constantly updated and kept up-to-date with relevant links.

A number of past studies make use of users' context and activity to provide them with the most relevant information. For example, typical search engines return relevant results based on the small amount of information from user queries and a measure of website popularity, rather than considering individual user interest and context [Speretta and Gauch 2005]. Spretta and Gauch employed user profiles based on user queries, search activities and the snippets of each examined result to refine search result rankings. With this context specific ranking of search results, an improvement of 34% in the rank order has been obtained

Rhodes and Maes [Rhodes and Maes 2000] described a new class of software agents called Just-in-Time Information Retrieval Agents (JITIRs), which has an ability to proactively present potentially valuable information based on a person's local context in a non-intrusive manner.

Another related work pushes the most relevant Web URLs based on the user activity and context. User context is determined by examining active personal desktop documents [Chirita et al 2006]. Similarly by observing user activity and context while reading a particular article, our notion of 'Links into the Future' presents the most related papers of the same team of authors within a local context. This paper discus how this concept can be extended to the WWW as a mechanism for contextual information supply for academic publications along a temporal dimension.

Existing approaches for finding related papers uses citation analysis, text similarity, bibliographic analysis and context based relatedness. For example, CiteSeer has employed three methods for finding related papers a) word vectors b) string distance c) and citations [Giles et al 1998]. PubMed [PubMed 2009] on the other hand computes the relatedness between two papers using text-based similarities between paper titles, abstracts, and assigned *"Medical Subject Headings"* (MeSH) terms [MeSH 2009]. For the focused paper, PubMed provides a list of related papers according to their relatedness scores.

Ratprasartporn [Ratprasartporn and Ozsoyoglu 2007] have made use of context (topics) of research publications to determine the related papers. An ontology consisting of terms were utilized as a context of publications. A publication is assigned to one or more contexts with the context represents the publication topics.

Digital libraries are traditionally built largely by a massive human effort. Example of these include INSPEC for engineering, PubMed for medicine and DBLP for Computer Science. Alternatively automated approaches are being employed to construct large citation indices. Examples of these efforts include CiteSeer and Google Scholar. The limitations of these automatic approaches are that human effort is often required in verifying entries in the index. Fully automated techniques have problems in disambiguating entries while traditional constructed digital libraries are limited in their number of scientific publication.

2.1.1 Ontologies

The word ontology has been borrowed from Philosophy, where it means *"a systematic explanation of being"* and has been used in the field of Natural Language Processing (NLP) for quite sometime now to represent and manipulate meanings of texts. Ontology plays a key role in building the Semantic Web, by providing a source of shared and precisely defined terms that are being used to describe web resources and their contents and improve their accessibility to automated processes. The knowledge engineering community has adopted ontology as a key enabling technology in order to realize the notion of Semantic Web [Berners-Lee 1998]. Gruber [Gruber 1993] defined an ontology as ”an explicit specification of a conceptualization”, which has become one of the most acceptable definitions to the ontology community.

Ontologies are used to define explicit formal conceptualized specification in a particular domain [Natalya and Deborah 2001]. Ontology provides an abstract view of a set of world objects. Ontology specifies the key concepts in a domain and their inter-relationships to provide an abstract view of an application domain [Fensel et al 2001]. Usually a concept in ontology is defined in terms of its mandatory and optional properties along with the value restrictions on those properties. Along with concept descriptions, it provides a taxonomic classification of concepts in the world to be used as semantic primitives.

With the support of ontology, both user and system can communicate with each other by the shared and common understanding of a domain. Ontologies are built by knowledge engineers with inputs from domain experts. Since ontologies provide a framework for unambiguous representation of a set of domain concepts and their inter-relations, therefore they can help in intelligent web-based information retrieval wherein it is possible to overcome the heterogeneity of web resources through the use of a shared conceptualization.

A typical ontology for the Web has a taxonomy defining the classes and their relations and a set of inference rules implementing reasoning functions. Classes and relations in ontology are usually domain specific. Use of a domain-specific ontology improves the accuracy of Web searches. Ontology enables knowledge engineers to provide structural and semantic annotations of web-document contents. These annotations help in conducting:

- Intelligent context-based search instead of keyword search
- Query answering instead of simple information retrieval
- Information exchange among distributed applications through ontology mapping
- Defining views on documents

2.1.2 Application Areas of ontologies

The use of ontological models to access and integrate large knowledge repositories in a principled way has an enormous potential to enrich and make accessible unprecedented amounts of knowledge for reasoning [Crow and Shadbolt 2001]. Ontologies are being used by Artificial Intelligence labs to semantic technologies now a day. People make ontologies to conceptualize things in a particular domain of knowledge [Natalya and Deborah 2001]. Andreasen et al. [Andreasen et al 2004] have proposed a system for content-based querying of texts based on the availability of an ontology that describes domain concepts. In their system the retrieval of text passages is based on matching descriptors from the text against descriptors from the noun phrases in the query using taxonomic reasoning with sub- and super-concepts. Snoussi et al. [Snoussi et al 2002] have proposed an ontology-based approach that facilitates the formalization and the extraction of data from different sources. The extracted data is converted into a coherent structure so that users and agents can query them regardless of their origin. Liddle et al. [Liddle et al 2003]] have proposed an ontology-based data extraction system, which uses an application ontology that describes a data-rich, ontologically narrow domain in a conceptual fashion. With inputs from a domain knowledge facilitator who can provide the knowledge for creating application ontology in an appropriate format, the system automatically generates a single wrapper that can be applied to any page relevant to the application domain.

The Artequakt project [Alani et al 2003] aims to implement a system that searches the Web and extracts knowledge about artists, based on an ontology describing that domain, and stores this knowledge in a KB to be used for automatically producing tailored biographies of artists. The Artequakt project aims to dynamically link a knowledge extraction tool with an ontology to achieve continuous knowledge support and guidance to the extraction mechanism. The ontology can provide a domain knowledge classification in the form of concepts and relations. The extraction tool searches online documents and extracts the knowledge that matches the given classification structure, and provides it in a machine-readable format to be automatically maintained in a Knowledge Base.

Handschuh et al. [Handschuh and Ciravegna 2002] have proposed S-CREAM (Semi-automatic CREAtion of Metadata) framework that uses ontology to support knowledge extraction. Vargas-Vera et al. [Vargas-Vera et al 2001] have proposed a Semantic Annotation Tool for extraction of knowledge structures from web pages through the use of simple user-defined knowledge extraction patterns. The semantic annotation tool contains: an ontology-based mark-up component which allows the user to browse and to mark-up relevant pieces of information; a learning component (Crystal from the University of Massachusetts at Amherst) which learns rules from examples and an information extraction component which extracts the objects and relation between these objects.

We, being the pioneer in implementing the idea of links into the future for WWW, have assembled some general information composition for this domain. One of the benefits of Ontologies is to reuse the ontologies as discussed in [Bontas et al 2005] where Bontas et al. have demonstrated the challenges related to the reuse process on the basis of two scenarios in the domains of eRecruitment and medicine, which aim at building domain ontologies by reusing existing knowledge sources. In the same way our Ontologies could be reused by digital journal or any digital resource manger to implement the concept of links into the future for their own digital resources.

2.1.3 Citation Mining and discovery

Another possibility to link forward in time is mining the papers' references. Subsequently, the cited papers can be linked further to the cited-by papers. Citation management is an important task in managing digital libraries. Citations provide valuable information e.g., used in evaluating an author's influences or scholarly quality (the impact factor of research journals). But although a reliable and effective autonomous citation management is essential, manual citation management can be extremely costly. Automatic citation mining on the other hand is a non-trivial task mainly due to non-conforming citation styles, spelling errors and the difficulty of reliably extracting text from PDF documents. Existing citation indexing and mining systems are explained below.

ISI citation index is the premier service provided by the ISI Web of Knowledge[1]. It indexes about 9,000 international and regional selected journals and book series. The selection of a journal by ISI dependents on the impact factor of the journal and on a number of factors[2]. This citation index is further used for the ranking of journals [Garfield 1972]. It is a manually created index making it extremely expensive. Some thoughts and issues on this manual approach are discussed in [Garfield 1964]. In searching for a particular paper's citations, ISI offers different databases such as *"Web of Science"*, *"Current Contents Connect"*, and *"ISI Proceedings"*. One can also select all the databases to be searched for all citations for a given paper.

CiteSeer[3] on the other hand provides an autonomous citation indexing service automating the entire process from crawling to extraction of citations from the Web [Giles et al 1998]. Although the primary focus area of CiteSeer is limited to computer and information science, it has nevertheless indexed about 1,077,967 documents and 20,328,278 citations. CiteSeer extracts titles and authors information from a citation entry programmatically. References are used to find the identical match within the collection to ascertain a citation. This service claims

[1] http://www.isiknowledge.com/

[2] http://scientific.thomsonreuters.com/free/essays/selectionofmaterial/journalselection/

[3] http://citeseer.ist.psu.edu/

that 80% of the titles can be extracted correctly from a number of citations. CiteSeer removes standard words and delimiters such as *"-&( )[ ], pp, pages, in press, accepted for publication, vol., volume, no, number et al, isbn, conf, conference, proc., proceeding, international society, transactions, technical reports"*. Word and phrase matching is subsequently performed on the extracted references (with an error margin of 7.7%) [Giles et al 1998].

Google Scholar[4], an open source multi disciplinary citation indexing service, was established in fall 2004 as a beta release. Its citations are indexed and extracted autonomously and cover a wide range of scientific literature. Google Scholar claims that it covers *"peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations"*[5]. As its search is not restricted to pre-defined journals and conferences, Google Scholar can be applied for the tracking of citations across most open access scholarly documents. One major limitation of Google Scholar is that it considers false positives including citations to press releases, resumes, and even links to bibliographic records for cookbooks [Price 2008]. It has gradually improved its algorithm and has been able to overcome previously encountered problems of finding citations backward in time [Jacsó 2008]. Its algorithm, however, has not been made known publicly.

Apart from the aforementioned citation indexes, there have been some other systems developed for a local dataset to extract references. For example Day [Day et al 2007] briefly described various systems and introduced a new hierarchical representation framework based on the template mining technique. This survey categorized existing systems into two broad categories *"Machine learning"* approach and *"Rule based"* approach. The template mining approach involves a Natural Language Processing (NLP) technique to extract data from text when data exists in recognizable patterns [Ding et al 1999]. If a text form matches a template pattern then the data is extracted by using instructions associated with that template. In the current work, we extract references from research papers by employing a template mining approach. As research papers fit into a well defined template, we have used a template-based reference extraction of research papers.

Machine learning approaches discover patterns from a dataset as discussed in [Agichtein and Ganti 2004] [Borkar et al 2001]. Such approaches as used for CiteSeer [Giles et al 1998] take advantage of probabilistic estimation, based on training sets of tagged bibliographic data. Although this technique has a good adaptability, it needs a huge set of labelled sample data for training. This requires a great effort in manually tagging substantial amounts of data.

The rule based approach on the other hand is based on rules defined by an expert in the field. Ding [Ding et al 1999] has discussed a template-based mining

[4] www.scholar.google.com

[5] http://scholar.google.at/intl/en/scholar/about.html

technique applying pattern matching and pattern recognition in natural language to extract information components. We have augmented our template-based technique by employing heuristic rules to extract the information components from extracted references. Rule-based approaches are straight forward to implement but they are not adaptable and it is often difficult to work with a system with many features. A generalised set of common heuristics has been proposed to overcome this limitation.

2.2 Tagging and Social bookmarking

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

Tagging is already a driving component in the fields of emergent semantic techniques [Mika 2005], Information Retrieval [Wu et al 2006] [Hotho et al 2006] and user profiling [Huang et al 2008].

Wu et al has shown that *"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 [Mika 2005] has studied the tagging behaviours and their usage in delicious, 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.

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] 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.

Citation Prediction system was selected as winner of KDD Cup 2003 Task-1 [Manjunatha et al 2003]. 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 locking 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 analyzed the relationship between the citation rate of an article and the extent of collaboration [Figg et al 2006]. 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.

We intend to use the bookmarking behaviours to model the citation rank prediction and we will compare this with the similar model developed from co-author network rank of publications with respect to the diffusion mechanisms of knowledge and their contexts.

2.3 Expertise Mining systems

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 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 2009] 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 [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 et al [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 tree.

2.4 Semantic Web - The Linked Data

2.4.1 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 2006] 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 2008] 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 [6] 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.

2.4.2 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], Disco[7] , Zitgist data

[6] http://www.sameas.org

[7] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

viewer[8], Marbles[9], Object Viewer[10] and Open link RDF Browser[11] 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 [Berrueta and Phipps 2009] [Chimezie 2009] an 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 to explore 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 2007] 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.

Chapter 6 deals with the highlighted problems in this section. An innovative URI mapping technique has been built which finds the exact URI of the intended resource. A concept aggregation framework has been proposed and implemented which extract informational aspects of a resource from Linked Data. Furthermore, the Journal of Universal Computer Science has been linked with Linked Data. The system is able to show different informational aspects of authors in a consolidated view.

[8] http://dataviewer.zitgist.com/

[9] http://beckr.org/marbles

[10] http://objectviewer.semwebcentral.org/

[11] http://demo.openlinksw.com/rdfbrowser/index.html

Chapter 3

Links into the Future

In chapter 1, the idea of *"Links into the Future"* was introduced as proposed by Maurer [Maurer 2001]. The idea is explained with the following fact: if a paper A was written e.g. in 1990 and a new contribution B in 2009 refers or is related to A, then a digital library may have a link from B to A e.g. If A refers to B, i.e. a *"Link into the Past"*. However, a link from A to B can also be obtained, thus providing a *"Link into the Future"*, specifically from 1990 to 2009 [Maurer and Tochtermann 2002]. This chapter explains the idea in more details, presents techniques/heuristics to develop a working system for creating Links into the Future. This chapter addresses the following research questions.

RQ1. How can a system for creating Links into the Future from multiple sources be developed?

RQ2. Can we improve existing citation mining techniques?

The first research question is further sub-divided into the following questions

RQ1.1 How can the concept of Links into the Future be represented using ontological framework.

RQ1.2 How can the idea of Links into the Future be applied for papers published within Journal of Universal Computer Science (J.UCS).

RQ1.3 How can the relevant Links be created from papers published in J. UCS to the papers available on the World Wide Web.

The figure 3.1, based on published contributions, explains the progress flow for the system. This thesis implements the idea with two techniques. The first is

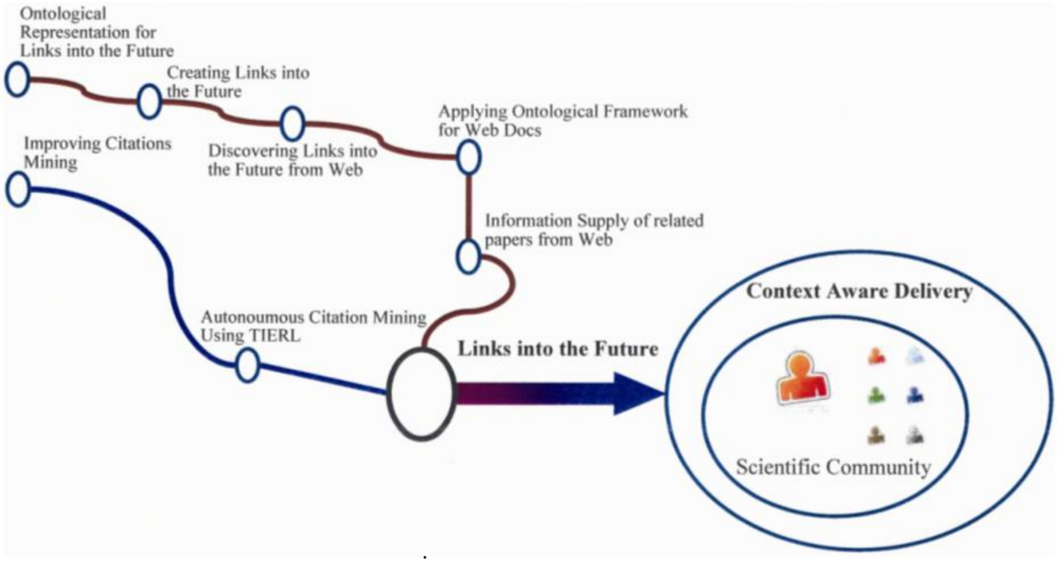

Ontological Representation for Links into the Future
Creating Links into the Future
Discovering Links into the Future from Web
Applying Ontological Framework for Web Docs
Improving Citations Mining
Information Supply of related papers from Web
Autonoumous Citation Mining Using TIERL
Links into the Future
Context Aware Delivery
Scientific Community

Figure 3.1: Progress flow for Links into the Future system based on published contributions

by exploiting metadata of publications (we refer to this as metadata extraction technique) while the second technique is by going through the references of all papers (we refer to it as citation mining technique). Based on the references, mutual links to the papers are created that are referred to. The figure 3.1 explains thesis progress for both mentioned techniques. The figure highlights progress flow of this chapter based on published contributions. For metadata extraction technique, the work was initiated by proposing an ontological representation of the idea [Afzal and Abulaish 2007] followed by techniques/heuristics for the implementation of the idea for papers published within J. UCS [Afzal et al 2007], then those published on the Web [Afzal 2009a] [Afzal 2009b][Afzal 2009c]. For citation mining technique, the job started by employing an autonomous technique for papers published within J. UCS [Afzal et al 2009a]. The technique was further tested on generic dataset, proving its efficiency [Afzal et al 2009b]

3.1 The Concept of Links into the Future

Initially, the idea of Links into the Future was proposed by Maurer [Maurer 2001]. In this section we explain this idea in details by defining different scenarios and contexts. When a user is looking on a particular content on the Web, the Links into the Future feature provides users with the most relevant information that has been made available on the Web after the publication date of the focused content. This can be realized in number of situations. For example a user is reading particular news; Links into the Future provides the user with the relevant news articles that were made available afterwards. The discovered news may contain other news articles expressing positive or negative sentiment about the focused news. If a user is reading a book, Links into the Future provides user with the books that are the extended version of the focused book. It may also provide positive and negative reviews about the focused book from different sources. However, in the scope of this thesis, we are talking about Links into the Future in the domain of digital journals. If a user is reading a research paper, he/she is provided with the most relevant research papers that were published in the future dates in the same area.

The number of digital publications is growing exponentially. Users expect to instantly get access to the information that is relevant to them. The management of such digital publications has to take into consideration the anticipated needs of users in providing highly customized services. It is currently possible to locate a paper in a previously published journal for a relevant research area, if a citation has been made explicitly. However, the user will not be able to view future relevant works from within the same paper based on a citation list alone. In order to achieve this, a user has to go through a citation index [CiteUlike 2006] [CiteSeer 2006] [DBLP 2006] and then find a future paper that has cited the former. We explore the possibility of producing a shortcut for the user, to enable links into the future to be accessed from within the paper itself. This has been achieved, to some extent, by employing typed-linking technology in the context of digital journals [Maurer 1996].

In the past, an initial attempt has demonstrated the idea of Links into the Future to some extent [Krottmaier 2003]. This concept has been, however, only partially realized so far. In this chapter, we shall extend the idea further by providing details on its full realization and its implementation. We explore the discovery of Links into the Future to be incorporated within a previously published paper in the Journal of Universal Computer Science [J.UCS 2009]. J.UCS is a unique in being the first electronic journal to implement this idea for enhancing a user's ability to gather future information on published papers in the same area. A published contribution is typically a static document; an author is not allowed to add or edit the published work. By implementing this idea, published papers will not have to remain as static documents, since it becomes possible to record new related developments as they get published. The previously published paper in the same area will get a link to the new paper as well. This is however, only a part of described visions of dynamic publications in a modern digital library [Dreher et al 2004]. Note, however, that the body of a published paper is never changed; only notes in the sense of links to other publications are made available, additionally.

While writing a research publication, researchers can cite any previously published paper. But of course they can not cite future contributions in the same area. While reading a paper, a reader may overlook papers published later by the same authors dealing with the similar topic. The readers would then need to make a deliberate effort to access the citation index [Citeseer 2006] in order to access later publications. However, there are inherited problems in existing citation mining approaches which have been described in section 3.3. In this chapter we propose a novel rule-based autonomous citation mining technique, to address this important task. We define a set of common heuristics that together allow to improve the state of the art in automatic citation mining. Moreover, by first disambiguating citations based on venues, our technique significantly enhances the correct discovery of citations. Our experiments show that the proposed approach is indeed able to overcome limitations of current leading citation indexes such as ISI Web of Knowledge, Citeseer and Google Scholar. As illustrated by [Dreher et al 2004], another limitation of typical citation systems is that they are constrained by the implementation of unidirectional links. In such systems when a document *"A"* cites document *"B"*, there is a link between A to B. There is however no link specified between *"B"* to *"A"*. Thereby the reader of article *"B"* may not know of *"A"*. One can claim that it is possible to create automatically two links: from article *"A"* to article *"B"* and vice-versa. However, this is very often not possible as a 'write access' would be needed to the system where article *"B"* was published. This problem, however, can be tackled in hypertext system such as Hyperwave Information Server [Hyperwave 2009]. Hyperwave provide a link-database, i.e. links are stored separately from contents. No write-access restriction is imposed to create links to a document [Dreher et al 2004].

Considering citations as the only measure for creating Links into the Future has another limitations. We illustrate the potential problem with one example: An author has published a paper *"A"* in the past, and subsequently published two other papers *"B"* and *"C"* in the same area. In paper *"B"*, the author cited the paper *"A"* but in *"C"* the author did not cite the paper *"A"*. While reading the paper *"A"*, if one wants to see future relevant publications, one goes to the citation index and may be able to find paper *"B"*, but would not find the paper *"C"* which may be more relevant to paper *"A"*. Thus, by using the citation index, a reader is able to find relevant papers only to a certain extent depending on how many references have been provided by authors. This has led us to explore the idea of incorporating Links into the Future using metadata extraction technique. By utilizing publications metadata, the technique creates links for papers published in J. UCS to other future related papers published in J.UCS and on the Web. Current search engines require the explicit specification of queries in retrieving related materials. Based on personalized information acquired over time, such retrieval systems aggregate or approximate the intent of users. In this case, an aggregated user profile is often constructed, with minimal application of context-specific information. However, in our case, the information captured based on an individual's current activity is applied for discovering relevant information along a temporal domain. This information is further pushed directly to the users' local contexts. This chapter, as such, presents a framework for the characterization and discovery of highly relevant documents.

The idea of Links into the Future gives two benefits: It enables the user to have a shortcut to access future work from within the paper and it makes sure that the user gains access to all most relevant research papers published in the future in the same area. This will make the papers dynamic in the sense that readers may be able to see all potentially similar publications in the same area for a particular publication.

The formal definition of Links into the Future is as follows: A future link from paper *"a"* to paper *"b"* (FutureLink (a, b)) exists, if a semantic relationship can be established between them. For example: if paper *"b"* is written by the same team of authors of paper *"a"* and the topics of both papers are similar, then paper *"b"* is considered to be related to paper *"a"*. Alternatively if there exists a citation from paper *"b"* to *"a"*, there is a highly likely relationship. Current systems tend to perform similarity matches without considering semantic similarity, based on the task characteristics. Equation 3.1 describes the definition of *"Links into the Future"* as used in our research.

$$[Authors(b) \in Authors(a) \wedge Topics(b) \in Topics(a)] \vee Citation(b,a) \rightarrow Future\_Link(a,b) \tag{3.1}$$

3.2 The Process of Creating Links into the Future

Each J.UCS paper is published with a set of topics which describe the area of the paper. The topics are basically the categories of the ACM classification system [ACM-CCS, 1998] (J.UCS has explicit permissions to use them) with minor extensions reflecting the growth of the field. A paper may belong to more than one topic like: D.2.1, H.5, and K.2. etc.

The process of creating Links into the Future can be summarised as follows:

1. For a particular paper which one is interested in, identify target items to be considered as potential Links into the Future. This will require the retrieval of all other papers of authors and co-authors. In order to validate the relevance of a future link, we check the mutual relatedness of topics and date of the publications.

2. Each of the above target items (potentially the related papers) will then be verified and validated to establish Links into the future.

3. Incorporate the links into the focused paper and make them accessible to readers. This step involves the determination of an effective way to incorporate the Links into the Future system.

We will describe two different techniques to establish Links into the Future. One possibility (we will refer it as citation mining technique) would be to look at the references of a paper, as proposed by [Krottmaier 2003] and check whether the list of papers contains publications in J.UCS. For a paper found in J.UCS, there is a great likelihood that the cited work (by the author) is of the same topic. As such, links are created from the former to these future papers. Citation indexes [CiteUlike 2006], [CiteSeer 2006] may also perform a similar operation. The citation mining technique [Krottmaier 2003] will also place future relevant links within the paper for future usage. This, however, cannot be achieved via a citation index alone. The details of citation mining technique are provided in section 3.3.

Another possible approach (we will refer it as metadata extraction technique) is to examine a particular paper and its authors and check for other papers in J.UCS and on the Web with the same author or one of the authors. In this case, we need to explicitly ensure that both papers are in the same field or same topic published in J.UCS or acquired from the Web. After validation, links to these future papers are created. This is not done by citation indexes [Citeseer 2006] in a direct way. The details of metadata extraction technique have been provided in section 3.4.

3.3 Citation Mining Technique

In recent years, the number of citations a paper is receiving is seen more and more (maybe too much so) as an important indicator for the quality of a paper, the quality of researchers, the quality of journals, etc. Based on the number of citations a scholar has received over his lifetime or over the last few years various measures have been introduced. The number of citations (often without counting self-citations or citations from "minor" sources, in whatever way this may be defined), or some measurement based on the number of citations (like the h- or the g-factor) are being used to evaluate scholars; the citation index of a journal (again with a variety of parameters) is seen as measuring the impact of the journal, and hence the importance one assigns to publications there, etc. The number of measurements based on citation numbers is steadily increasing, and their definition has become a science in itself. However, they all rest on finding all relevant citations. Thus, "citation mining tools" used for the ISI Web of Knowledge, the Citeseer citation index, Google scholar or softwares such as the "publishorperish.com" software based on Google scholar, etc., are the critical starting points for all measurement efforts. In this section, we show that the current citation mining techniques do not discover all relevant citations. We propose a technique that increases accuracy substantially and show numeric evaluations for one typical journal and for a set of generic dataset of citations. It is clear that in the absence of very reliable citation mining tools, all current measurements based on citation counting should be considered with a grain of salt. Furthermore, the developed citation mining technique is applied for creating Links into the Future.

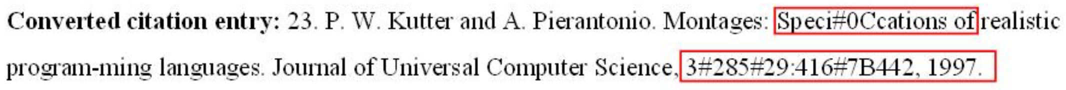
3.3.1 Citations and their Importance in Digital Libraries

Digital libraries (DL) collect, organize, and provide access to large collections of diverse knowledge resources. A well-managed digital collection of electronic published works and artefacts is of great importance in providing a strong impact for forthcoming new research that may otherwise not be possible without "standing on the shoulders of giants". Citations allow authors to refer to past research in a formal and highly structured way [Garfield 1955], to systematically construct a citation network that then serves as a means of valuation for published works.

The citation count, which refers to the number of citations a particular paper receives, is used in evaluating bibliometrics such as the quality of a paper, the quality of researchers, the quality of journals, etc. It has been used for knowledge diffusion studies [Hu and Jaffe 2003], network studies [Dorogovtsev and Mendes 2002] and in finding relationships between documents [Small 1973]. Impact factor measurements, as derived from citation counts have been applied in making important decisions such as hiring, tenure decisions, promotions and the award of grants [PLoS Editorial 2006]. As such the determination of precise citation counts is of utmost importance.

Citation mining refers to the process of discovering citation counts. This task in itself is not trivial as it involves extensive text analysis to determine the exact intended citation of authors to published works. Owing to the large number of publications, this task involves a great amount of human effort if done manually. Alternatively, an approach for autonomous citation discovery can be applied. This approach, however, tends to be prone to omissions and mistakes [Giles et al 1998]. Fully autonomous citation mining as such has to rely on community effort for the verification and regular updating of citation records (e.g. Citeseer [Giles et al 1998]).

In order to address this important task, this chapter proposes a novel rule-based autonomous citation mining technique, called Template based Information Extraction using Rule based Learning (TIERL). A two-phase approach is used whereby the system first disambiguates citations based on venues. Subsequently detailed rule-based mining is performed on a much smaller collection of data within the particular venue. The heuristic approach employed is described in the following sections. We illustrate the benefits of this approach by studying the



Converted citation entry: 23. P. W. Kutter and A. Pierantonio. Montages: Speci#0Ccations of realistic program-ming languages. Journal of Universal Computer Science, 3#285#29:416#7B442, 1997.

Figure 3.2: Problems in PDF to text conversion

enhancements to current state of the art by applying our methods to the dataset of the Journal of Universal Computer Science (J.UCS)[1].

3.3.2 Problem Statement

The state-of-the-art citation mining systems have been explained in chapter 2. Here we summarize the problem statement:

Citation mining can be viewed as a three tier process:

1. Reference (citation entries) extraction from documents.

2. Metadata extraction from citation entry.

3. Linking citation entry to the cited paper.

Most scholarly works reside in digital libraries as PDF documents. For extracting references, these PDF documents are further converted into plain text. This conversion process may result in errors as shown in figure 3.2:

While the original citation entry looks like:

23. P. W. Kutter and A. Pierantonio. Montages: Specifications of realistic program-ming languages. Journal of Universal Computer Science, 3(5):416-442, 1997.

The automated extraction of metadata sub field, such as title and authors, from a citation entry is not at all a trivial issue as:

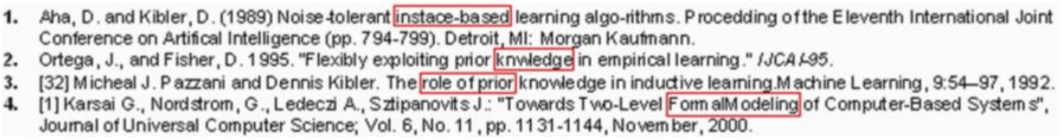
1. All publishers have their own style guide which needs to be considered while extracting sub fields from a particular reference entry.

2. There are times when authors inadvertently do not follow the style guides properly.

While citing a paper, authors tend to also make mistakes as illustrated in figure. 3.3. These mistakes may then lead to improper citation linking.

Apart from spelling mistakes made by authors, re-wording of titles also occurs e.g in the 3th entry, the word "utility of" was replaced by "role of prior". These types of errors are made mainly because authors simply copy citations from existing references. Mistakes may also arise due to carelessness or negligence.

[1] http://www.jucs.org/



1. Aha, D. and Kibler, D. (1989) Noise-tolerant instace-based learning algo-rithms. Procedding of the Eleventh International Joint Conference on Artifical Intelligence (pp. 794-799). Detroit, MI: Morgan Kaufmann.
2. Ortega, J., and Fisher, D. 1995. "Flexibly exploiting prior knwledge in empirical learning." *IJCAI-95*.
3. [32] Micheal J. Pazzani and Dennis Kibler. The role of prior knowledge in inductive learning.Machine Learning, 9:54–97, 1992.
4. [1] Karsai G., Nordstrom, G., Ledeczi A., Sztipanovits J.: "Towards Two-Level FormalModeling of Computer-Based Systems", Journal of Universal Computer Science; Vol. 6, No. 11, pp. 1131-1144, November, 2000.

Figure 3.3: Badly formatted references by authors

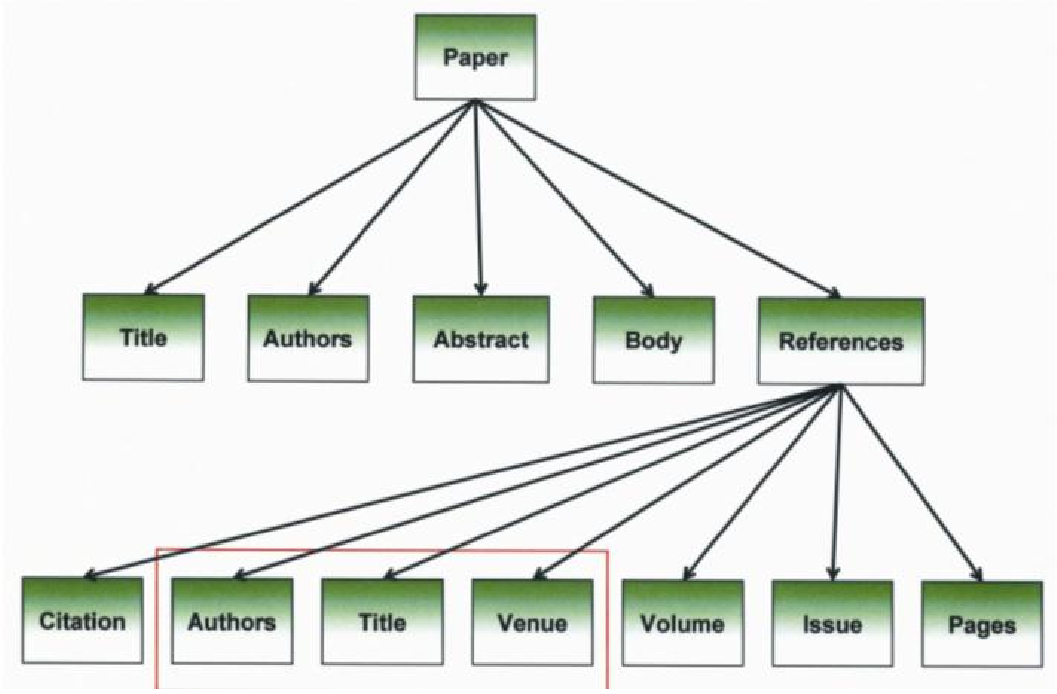


Figure 3.4: Template based Information Extraction

3.3.3 Template based Information Extraction using Rule based Learning (TIERL)

We propose the Template Based Information Extraction using Rule Based Learning (TIERL) technique to increase accuracy of citations obtained. We could make a full text search to link the citations but due to the problems defined in section 3.3.2, we have introduced a systematic way of citation linking. The system architecture for TIERL is shown in figure 3.5. TIERL is a layered approach where Template based Information Extraction (TIE) refers to the treatment of a paper as a template from which reference entries are extracted. Rule Based Learning refers to the usage of heuristic rules applied to extract the data and in dealing with uncertainty and the approximate matching of citations. Research papers are represented as a template structure as shown in figure 3.4. From a given citation string, authors, title and venue information will be used to link citations.

TIERL Algorithm

The generic rules to identify a citation entry are depicted below:

Step 1. Extract references from each document using template based information extraction technique.

Step 2. Tokenize each citation string and extract citation components (title, authors, and venue) using FLUX-CIM [Cortez et al 2007].

For each citation string repeat step 3 to step 8.

Step 3. Disambiguate extracted venue in step 2 from DBLP for the focused citation string using rule based approach as given in the next section.

Step 4. Select all papers (their titles and authors) from DBLP which are published in the disambiguated venue in step 3.

Step 5. Apply direct match between the extracted title in step 2 and the titles of the papers selected in step 4.

If (exact match is found) **then** link the citation, focus the next citation entry and go to step 3.

Else if (direct match fails) then continue to step 6.

Step 6. Remove stopwords from extracted title in step 2 and focused titles in step 4.

Step 7 Apply approximate matching (App. Match) on the paper's titles returned by step 6.

$$\text{App. Match}=\frac{\textit{number of words found in the compared title in a sequence}}{\textit{max number of words of a papers title in extracted or compared title}}\times 100.$$

If (match) > threshold then link the citation, focus the next citation entry and go to step 3.
Else if (match of more than one records) > threshold then select all matched papers as candidates and go to step 8.
Else if (match) < threshold then select max. matched paper as candidate and go to step 8.
Step 8. Match author's list of both extracted and candidate papers.

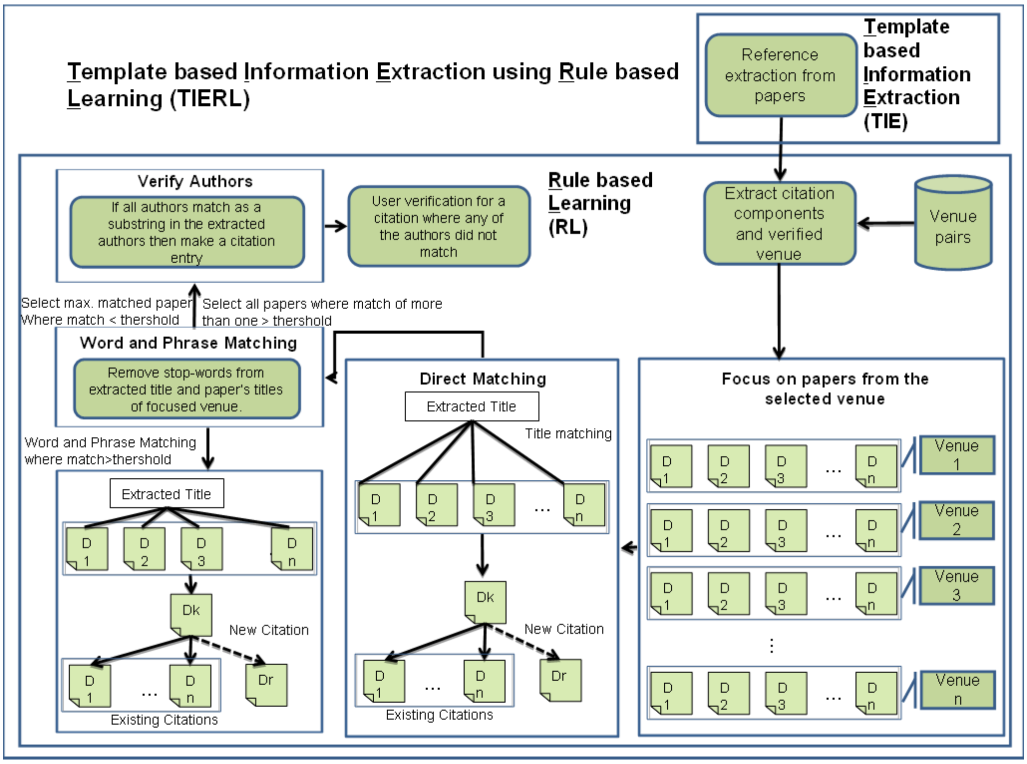

Template based Information Extraction using Rule based Learning (TIERL)
Reference extraction from papers
Template based Information Extraction (TIE)
Verify Authors
If all authors match as a substring in the extracted authors then make a citation entry
User verification for a citation where any of the authors did not match
Rule based Learning (RL)
Extract citation components and verified venue
Venue pairs
Select max. matched paper Where match < thershold
Select all papers where match of more than one > thershold
Word and Phrase Matching
Remove stop-words from extracted title and paper's titles of focused venue.
Word and Phrase Matching where match>thershold
Extracted Title
D 1
D 2
D 3
D n
Dk
New Citation
Dr
Existing Citations
Direct Matching
Extracted Title
Title matching
Focus on papers from the selected venue
Venue 1
Venue 2
Venue 3
Venue n

Figure 3.5: System architecture for TIERL

If (all authors matched) then link the citation, focus the next citation entry and go to step 3.

Else show to user/community for verification, focus the next citation entry and go to step 3.

Different techniques for the extraction of citation components have been proposed and used in the past. For our experiments, we used a technique proposed quite recently [Cortez et al 2007]. This technique gives precision and recall of more than 94% on a generic dataset. This technique uses a knowledge base (KB) which contains pairs of (m_i, o_i) where m_i is metadata field like author, title, and venue, and o_i is different occurrences of this field. This KB is used to calculate the field frequency. A citation string is split into blocks on the occurrence of any character other than the characters A,..., Z, a,...,z, 0,...,9. For each block field, frequency is calculated as shown in equation 3.2.

$$\text{FF(b,m}_i) = \frac{\sum\limits_{t \in T(m_i) \cap T(b))} fitness(t, m_i)}{|T(b)|} \quad (3.2)$$

Where fitness (t, m$_i$) is defined as follows:

$$\text{fitness (t, m}_i) = \frac{f(t,m_i)}{N(t)} \times \frac{f(t,m_i)}{f_{max}(m_i)} \quad (3.3)$$

The block b is associated with the field which gives the maximum value of FF. More details about the technique can be found in [Cortez et al 2007].

Searching Articles by Venue

Venue disambiguation is an important task for citation indexes like Thomson ISI, Google Scholar, and CiteSeer. Accurately disambiguated venues are further used for user interfaces and for performing data mining of research literature. We try to cleverly use this venue information to accurately link the *"cited"* and *"cited by"* paper. Hall et al [Hall et al 2008] have recently suggested an unsupervised method for venue disambiguation. They assume that venues tend to focus on particular research areas and these areas are reflected in the titles of the published papers in a venue. Consequently, they made a venue over title model and disambiguate venues based on Dirichlet process mixture. This model works fine when the venue is focused. They also applied this model to two venues which share the same *"acronym"* like ISWC (International Semantic Web Conference and International Symposium on Wearable Computing). The venues were accurately disambiguated because the focus of both venues was quite different. But if the venues share the same acronym and the focus of the venue is also the same, then it becomes difficult to disambiguate. These types of venues are listed in Table 3.1. Moreover, venues which are not focused are also difficult to disambiguate like the venues *"Communications of the ACM"*, *"IEEE Computer"*, *"Journal of Universal Computer Science"* etc.

In DBLP[2] , the venues are indexed by acronym along with the full venue title. There are more than 5000 unique venues listed in DBLP.

A knowledge base was built which comprises of a set of pairs of the form KB= (a$_i$, f$_i$), ... (a$_n$, f$_n$) in which each a$_i$ is an acronym and f$_i$ is a full name of the

[2] http://www.informatik.uni-trier.de/~ley/db/

Table 3.1: Venues sharing same acronym with almost same research focus

ID	Venue Acronym	Venue Full Name
1	ICIS	International Conference on Information Systems
		IEEE/ACIS International Conference on Computer and Information Science
2	ICDM	Industrial Conference on Data Mining
		IEEE International Conference on data Mining
3	AIPR	Applied Imagery Pattern Recognition Workshop
		Artificial Intelligence and Pattern Recognition
4	PDCS	Parallel and Distributed Computing Systems (IASTED)
		Parallel and Distributed Computing Systems (ISCA)

venue where a_i and f_i both are pointing to the same venue. A typical example of this pair is a venue pair where a_i is "AAAI" and f_i is "National Conference on Artificial Intelligence".

The rules to disambiguate venue is illustrated here:

Step 1. Make venue pairs from DBLP as (a_i, f_i) where a_i (acronym) and f_i (full name) are pointing to the same venue.

Step 2. Remove stopwords from the extracted venue (in step 2 of above mentioned algorithm).

Step 3. Apply direct match between the cleaned venue string from step 2 with the pairs (a_i,f_i).

If (one match is found) then note the corresponding DBLP venue and exit.

Else if (more than one venues in (a_i, f_i) share the same a_i) then go to step 4.

Else if LD (substring (venue in step 2),any value in pairs (a_i, f_i)) =1 OR LD(substring (any value in pairs (a_i,f_i)),venue in step 2) = 1 (where LD is Levenshtein distance) then note the corresponding DBLP venue and exit.

In step 3, treat the words (Journal, International, National, European, Asian, publishers like (IEEE, ACM, WSEAS, Springer, and Elsevier etc) as general words, if they match in a sequence then okay, otherwise they will be ignored while matching.

Else if all patterns of a venue in step 2 match in a sequence as a substring with any pair of (a_i,f_i) then note the corresponding DBLP venue and exit.

Step 4. select all papers from the venues which share the same acronym. Disambiguate venue and citation based on matched titles of the paper as described in above mentioned algorithm.

The matching of patterns in the extracted venue string means that it should match as a substring in a sequence with any of the venue pair (a_i, f_i). For example in the case of venue *"Journal of Universal Computer Science"*, all of the following extracted venues will find their match: *"Jour. Univers. Comp. sci."*, *"J. Uni. Comp. Science"* and *"J. Uni. Computer Sci."* etc.

Dataset

For our initial experiments, we collected texts of citations already hand-clustered into groups referring to the same paper from Cora[3]. For this dataset, we collected the extracted citation components. Our main task was to disambiguate venue and link the citation accurately. Within this dataset, we further focused on the venues listed in DBLP. In this dataset, there were 7 unique venues with different strings mentioning the same venue. These venues belonged to a focused area where venue over title model may work fine [Hall et al 2008]. This dataset was enhanced with three further venues. One of the venues is *"Journal of Universal Computer Science"* which belongs to a list of venues that publish papers in broad categories. Two remaining venues belong to the similar focus area and share the same acronym, i.e. ICIS (International Conference on Information Systems), IEEE/ACIS (International Conference on Computer and Information Science). In this way, we have approximately 400 citation strings which were used to disambiguate venues and then accurately linked with cited papers.

From the citation strings, we first need to extract the part of the venue string which is actually referring to some venue. Stop-words like (’proc’, ’proceedings’, years, months, ’in’, ’.’, ’:’, ’published’, numeric values, corresponding alphabets for numeric values like eleventh, twelfth etc., ’of’, ’the’, ’(’, ’)’, ’’, ’’, ’[’, ’]’, ’-’, ’to appear’, ’accepted’, ’vol’, ’issue’, ’no’, leading and trailing spaces) are removed. By means of this process, we clean the venue string. However, it may still contain some discrepancies along with typographical errors.

In the first run of matching a cleaned venue string with the venue pair (a_i,f_i), 89% of the venues were matched. The remaining 9% venues were found during step 3 and 4 of venue disambiguation algorithm. 8.5% of the venues found their match in step 3 resulting in LD (s, t) = 1 while comparing individual strings.

For a citation entry, we focused only on the paper's titles published in the extracted and verified venue. The results are shown in Table 3.2. This algorithm achieved an overall accuracy of 99.23%. A small fraction (0.77%) of the citations were unidentifiable as authors wrongly recorded venue information in their citations e.g. The paper "Learning subgoal sequences for planning" was actually published in venue ’IJCAI’ but was wrongly cited as being published in ’AAAI’.

[3] http://www.cs.umass.edu/~mccallum/code-data.html

Table 3.2: TIERL algorithm results on Cora Dataset

Matching Steps	Accuracy
Direct Matching	89.05%
Approximate matching > threshold	7.38%
Author's verification where approximate matching < threshold	2.80%
Overall accuracy	99.23%

Added Value

The extraction of venues and focusing on the papers published in particular venues was significant in linking the citations properly. For example, the same team of authors has written the following two papers in two different venues with a slight change in title. *"Instance-Based Learning Algorithms"*, published in *Machine Learning.*

"Noise-tolerant instance-based learning algorithms", Published in *"IJCAI"*.

Although the authors are the same and title of the paper is also similar, it was successfully disambiguated because of the focused dataset (searching for articles within the articles published in the verified venue). For another citation string, the cited title was *"Instance-Based Learning."* instead of *"Instance-Based Learning Algorithms"*, published in *Machine Learning.* Without focusing by venue resulted in 62 unique records from DBLP dataset where this title was matched 100% as a substring. Focusing by venue then significantly helped by reducing the choices to only three candidate papers to select. As a result CiteSeer which selects citation strings of similar lengths from its huge index [Giles et al 1998] gets too many similar records. This makes it very difficult to disambiguate.

Sometimes, while making a citation, authors write some additional words or omit or change some words from the title e.g. paper *"Instance-Based Learning Algorithms"* was cited as *"Instance-Based Learning Methods"*. During an approximate matching process, 67% was matched and then citation was derived based on the matched authors´ list. It was noted that there was not a single false positive citation. This is predictable as the same team of authors normally do not submit a paper with almost same title to the same venue.

3.3.4 Experimental Case Study

The Journal of Universal Computer Science (J.UCS) was considered to be a suitable journal to be used for this case study, based on its broad coverage of Computer Science and Information Technology areas. Because of its broad coverage, there is no particular community which is only publishing in J.UCS. Thus, authors from different backgrounds publish their articles which makes it an interesting dataset for this case study. J.UCS has published more than 1200 peer reviewed papers. J.UCS also provides a large enough document collection to illustrate the workings of the proposed approach.

We applied Template based Information Extraction (TIE) to extract references from PDF versions of J.UCS papers. To perform TIE, we need the full text of all papers in a digital form. The papers are currently available in PDF format and were downloaded automatically from the J.UCS server. Many PDF to text converter tools were tested in terms of accuracy and speed. These include PDFBox [4] , Ghostview[5] and PDFTextStream[6]. Based on its performance, PDFBox (open source java PDF library) was selected for conversion. We then explored the use of layout information of a paper to discover detailed information regarding its structure. For example, a reference starts with the term "references", followed by a delimited list of citation entries. We used three styles of writing a reference entry, which would start from any of the following styles: ’[author’s years]’, ’[1]’, ’1’ ". Each citation entry is also expected to have a fixed format.

We used intrinsic pattern mining of documents.13.5% of the papers were editorial columns. Almost 78% out of 86.5% of the papers’ references were extracted resulting in over 15 thousands citation entries. 3.5% of the papers have bad references (not complying with any of the templates). 5% of the papers were not compliant with the conversion tool, and were thus not converted correctly into plain text. These 5% papers were not recognized as PDF documents even by the professional converters like INTRAPDF[7]. We propose the use of the postscript and HTML versions of these documents for future experiments.

For the current case study, we focused the citations from J.UCS to J.UCS papers. There were two reasons for the focused dataset (1) J.UCS is indexed by ISI. ISI indexes only a selected number of journals and if we compare the citation out degree for J.UCS then the comparison would not be interesting enough because not all journals and conferences may be indexed by ISI. But if we focus on citations from J.UCS to J.UCS then it is sure that ISI should have all the citations. CiteSeer also claims that it indexes open access journals and tracks when new issues are published. Then the comparison is meaningful to know either CiteSeer index all papers of J.UCS if yes then either it is able to find all citations with an error margin of 7.7% as of their claims [Giles et al 1998]. (2) Second reason for selecting the dataset was the manual effort required for comparison with the citation indexes because these citation indexes provide free services for community to explore the citations for a focused article most of the time manually. But they (ISI and Google Scholar) do not give their whole data free of charge which could lead to developing an automatic program to compare the results. Consequently,

[4] http://www.pdfbox.org/

[5] http://pages.cs.wisc.edu/ ghost/index.html

[6] http://snowtide.com/

[7] http://www.intrapdf.com/

Table 3.3: TIERL algorithm results on J.UCS dataset

Matching Steps	Accuracy
Direct Matching	69.17%
Approximate matching > threshold	24.06%
Author's verification where approximate matching < threshold	3.76%
Overall accuracy	97%

it is a herculean effort to compare each and every paper with ISI, Google Scholar and CiteSeer for checking the citations.

We used the "FLUX-CIM" technique described in [Cortez et al 2007]. The knowledge base (KB for short) for this was built from all published papers in J.UCS. We extracted the citation components from citation strings where the venue block was represented as J.UCS. The details of venue disambiguation are already explained. In this way we extracted citation components from 133 J.UCS to J.UCS citations. This technique when applied on a generic dataset [Cortez et al 2007] gives a precision of 95.85% and recall of 96.22% for CS domain. This, however, depends on the complete knowledge base where each and every token represented in the citation string could find its match. In our case, we have focused on the KB built from J.UCS. This is why all tokens found their match in the KB and we were able to extract all the titles and authors of J.UCS citations. But of course the accuracy of results for a venue for which one does not have complete bibliographies to compare with the extracted token would not be 100%. The results of our TIERL algorithm (as described in section 3.3.3) on J.UCS dataset gives the results as shown in Table 3. 3% citations were unidentified. On manual inspection, it was found that 2.25% were referring to papers which were not indexed by DBLP but in fact were published by J.UCS. This is however not the fault of our algorithm. While the match for 0.75% (only one record) was less than threshold. Subsequently, list of extracted authors for the maximum matched paper was compared to DBLP. However, all authors did not find their match and the system was not able to automatically link the citation. This citation was further shown to the user for feedback and on user's response, the citation was linked. Nevertheless, we revised the same pattern that we did not find any 'False Positive'.

After the citation mining for J.UCS articles was completed, we performed comparisons with existing citation indexes. For a comparison with ISI, we selected all of the available databases (*"Web of Science"*, *"Current Contents Connect"*, and *"ISI Proceedings"*). To compare with CiteSeer and Google Scholar, we used their standard websites. We have a total of 133 citations from J.UCS to J.UCS but while comparing we found 13 more citations which were missed by TIERL. The reasons for these missed citations by TIERL are explained in next sections. So now we have total 93 unique J.UCS papers with 146 citations within J.UCS.

3.3.5 Experimental Results

The measurements selected to compare the citations with other citation indexes were subject to answer three questions. (1) Out of the 146 citations, how many are indexed by each citation index? (2) What was the total missed percentage by each citation index regardless of indexing (the paper or cited by paper). (3) Out of these 146 citations, how many papers and their 'cited by' papers were both indexed by each citation index but the citation index has failed to find the citation. The effect of this would be studied by calculating the total number of citations for those papers received within J.UCS. The initial experiment was done during April, 2008 and revised in March 2009.

Indexed Papers

The numbers of papers indexed on different citation indexes are listed here. ISI indexes 38% of the papers, CiteSeer indexes about 53% of the papers while Google Scholar indexes 100%. TIERL indexes 98% because overall 2% J.UCS papers were not indexed by DBLP. If these citation indexes do not recognize these J.UCS papers then how they can include them for finding citations. The comparison is shown in figure 3.6.

Overall Missed Citations

Different citation indexes were compared with the focused citations dataset. The figures represent the percentage of the data missed by citation indexes. These are the overall missed percentages regardless whether the paper is indexed or not. The percentage of missed citations was surprisingly high for the major citation indexes like ISI, Google Scholar and CiteSeer as can be seen in figure 3.7.

Missed Citations within the Index

Here we focused on missed citations if both the *'cited'* and *'cited by'* paper are indexed by the citation index. For example, in the case of ISI, J.UCS was not indexed until 2001. But if we evaluate the missed citations by ISI from 2001, there were a total of 42 articles in J.UCS since 2001 which have been cited by other J.UCS articles. According to our experiments, these 42 articles received 58 citations within J.UCS. All of these 'cited' and 'cited-by' papers are indexed by ISI. Out of 58 citations, 17 were missed by ISI. This gives an error rate of 29.3%. This is surprisingly high for an established citation index. The comparison with all citation indexes is shown in Table 3.5 and the missed percentages are shown in figure 3.8.

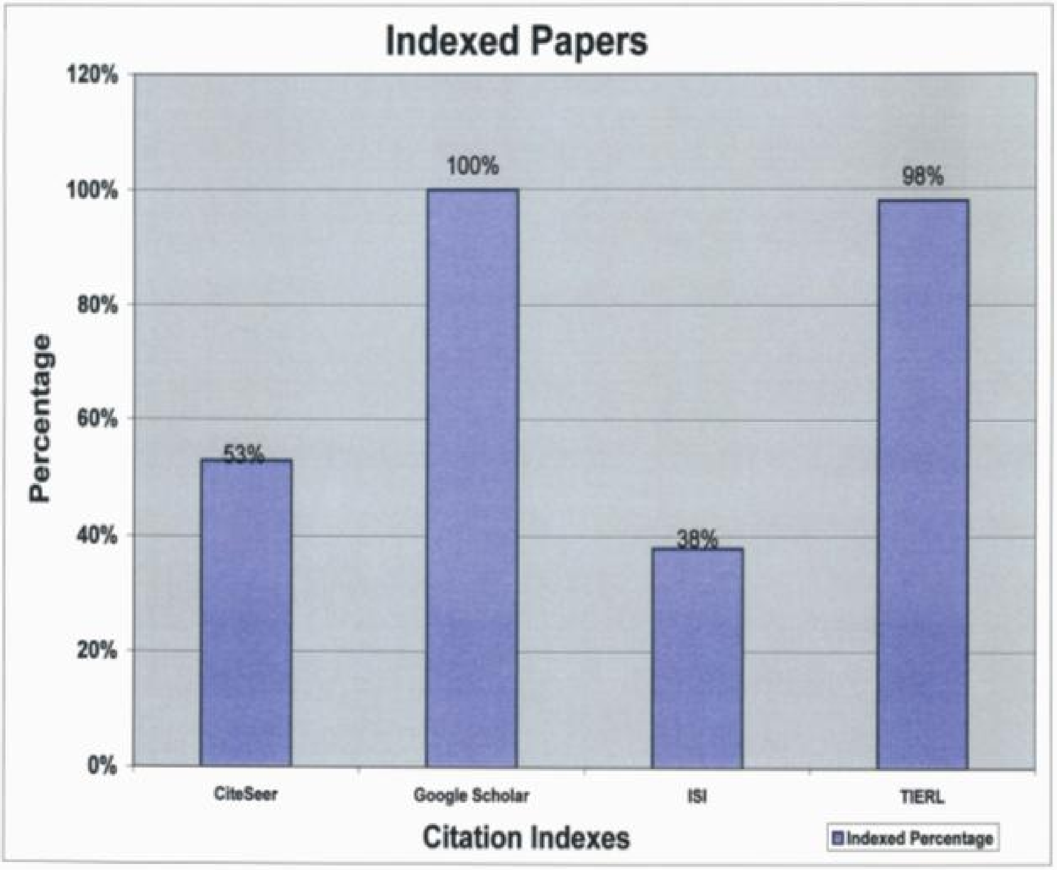


Figure 3.6: Indexed papers

Table 3.4: Impact factor of J.UCS in 2005

Cites in 2005 to articles published in: 2004 = 26 Cites in 2005 to articles published in: 2004 = 33 Sum = 59	Number of articles published in: 2004 = 89 Number of articles published in: 2004 = 86 Sum = 175
Impact factor=59/175 = 0.337	

Misleading Impact Factor

Being an authority in measuring impact factors of journals, Thomson ISI publishes a Journal Citation Report every year. Thomson ISI calculates an impact factor for a particular venue in a given year based on the citations for the papers published in the last two years. For example the impact factor of J.UCS in 2005 would be the number of citations made by the papers in 2005 (which are published in ISI indexed venues) to papers published in J.UCS during the years 2003 and 2004 divided by the total number of papers published in J.UCS during 2003 and 2004. The impact factor of J.UCS in 2005 by ISI is shown in table 3.4.

But within our small focused dataset of citations from J.UCS to J.UCS articles, it has been observed that there were extra 4 citations in J.UCS papers

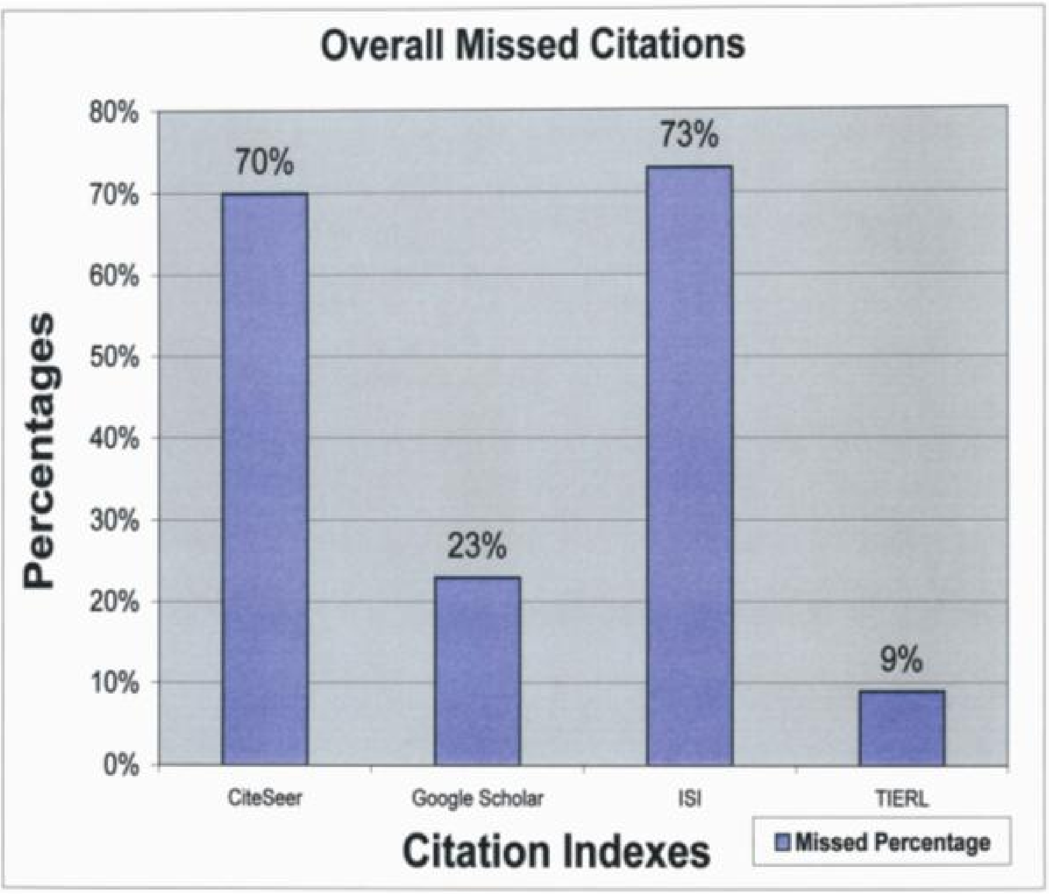


Figure 3.7: Missed citations

Table 3.5: Found citations within the index by Citation Indexes

Citation Index	Indexed papers	All Citations within J.UCS	Found by Citation Index
ISI	42	58	41%
GS	93	146	113
CiteSeer	53	78	44
TIERL	91	143	133

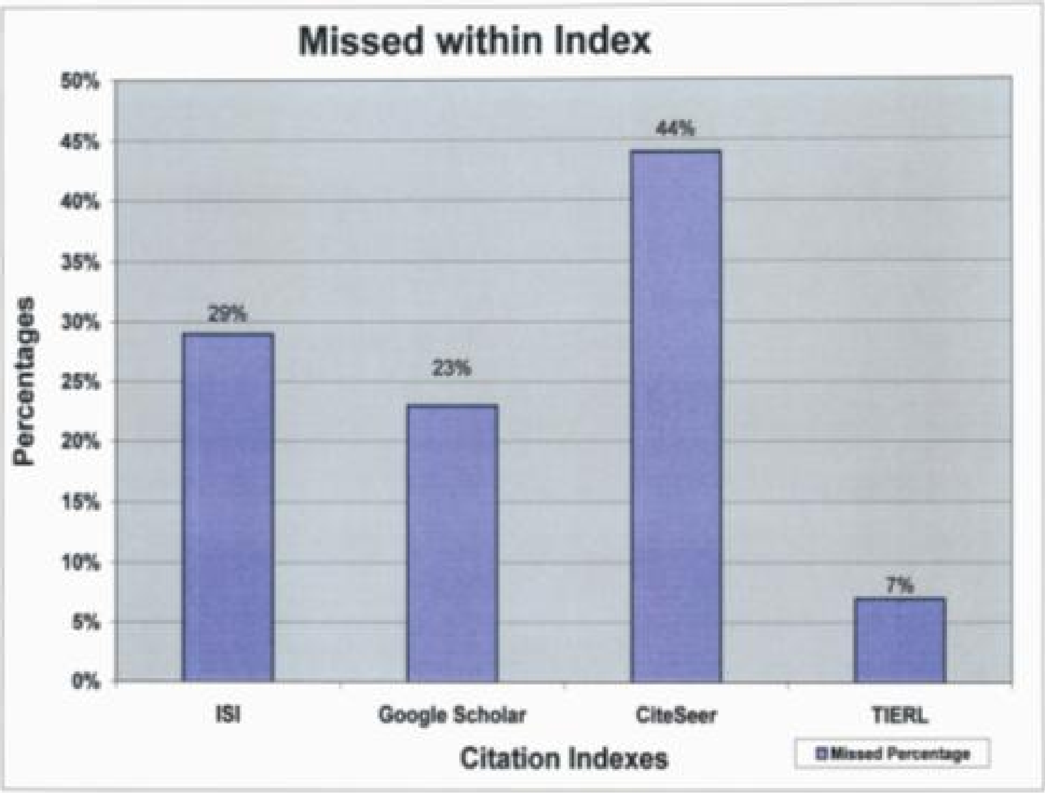


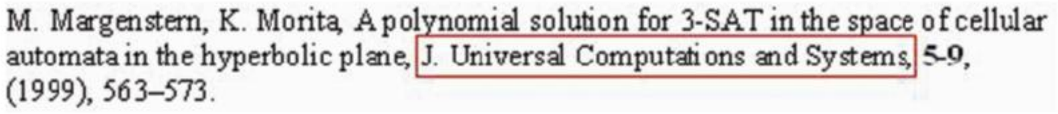
Figure 3.8: Missed citations within citation index and their overall impact

published in 2005 to J.UCS articles published in 2004. With this small information the actual impact factor of J.UCS for the year 2005 becomes 0.36 instead of 0.337. But it has been shown that the overall impact of missed J.UCS citations by ISI within their index was 29.3%. And if ISI is missing citations to J.UCS papers by the same ratio for other sources then the impact factor of J.UCS should be 0.48 instead of 0.336 i.e. is almost equivalent to the J.UCS impact factor in 2003.

Missed Citation Snippets

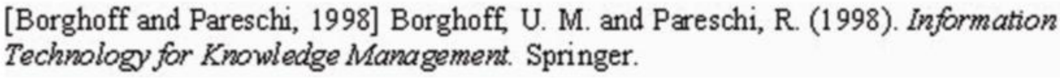
This section first describes the reasons for missed citations from TIERL and then by other systems.

As discussed earlier that TIERL had missed 13 citations which is 9% of the total. These are the following reasons for: 1.5% was due to unspecified venue information or citing a venue wrongly. 7.5% were due to bad conversion from PDF to text. The reason for this failed conversion was due to PDF files encoding that prevented editing. But Google Scholar was able to find these citations as it had indexed HTML versions of these documents. For future experiments we will consider PS and HTML versions to overcome this limitation. Typical missed citations by TIERL are shown below:



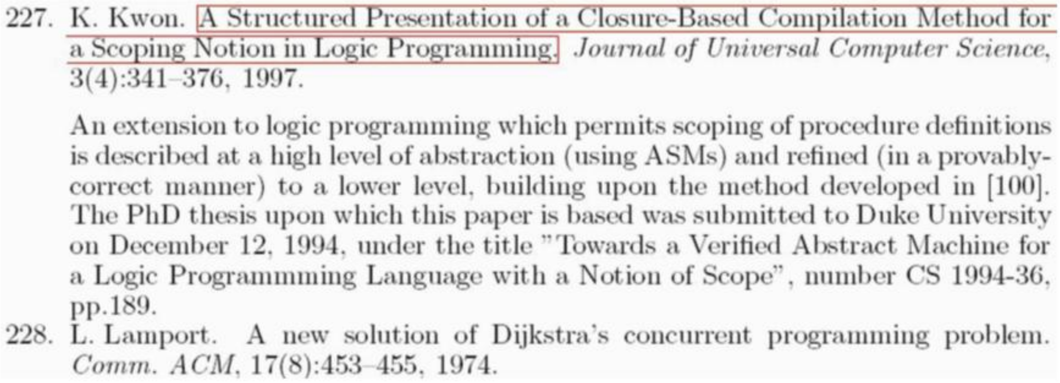
M. Margenstern, K. Morita, A polynomial solution for 3-SAT in the space of cellular automata in the hyperbolic plane, J. Universal Computations and Systems, **5-9**, (1999), 563–573.

Figure 3.9: Wrong venue information



[Borghoff and Pareschi, 1998] Borghoff, U. M. and Pareschi, R. (1998). *Information Technology for Knowledge Management*. Springer.

Figure 3.10: Missing venue information



227. K. Kwon. A Structured Presentation of a Closure-Based Compilation Method for a Scoping Notion in Logic Programming. *Journal of Universal Computer Science*, 3(4):341–376, 1997.

An extension to logic programming which permits scoping of procedure definitions is described at a high level of abstraction (using ASMs) and refined (in a provably-correct manner) to a lower level, building upon the method developed in [100]. The PhD thesis upon which this paper is based was submitted to Duke University on December 12, 1994, under the title "Towards a Verified Abstract Machine for a Logic Programmming Language with a Notion of Scope", number CS 1994-36, pp.189.

228. L. Lamport. A new solution of Dijkstra's concurrent programming problem. *Comm. ACM*, 17(8):453–455, 1974.

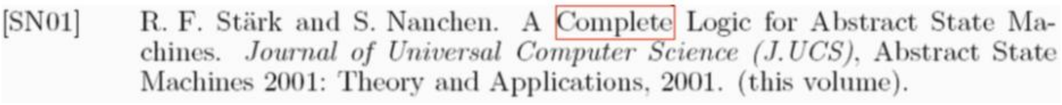
Figure 3.11: Discussions within references

In the figure 3.9, the authors have specified each component correctly but the venue is cited wrongly. The article was published in Journal of Universal Computer Science but while citing authors have written J. Universal Computations and Systems.

In figure 3.10, authors have not provided the venue information and that is why the citation was not found by TIERL.

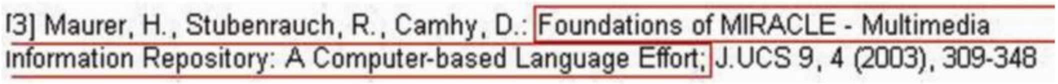
If we carefully look at the missed citations by major citation indexes then we will find some interesting patterns. For example, in figure 3.11, authors have included explanations within the reference section. Here the authors have written an explanation after the reference entry 227. Usually, it is not expected that authors would write some explanation within the references. But in this case the reference entry 227 would be considered until the next entry 228 starts although the actual reference entry is only the first three lines. But in this case the 227 reference entry is assumed to comprise 10 lines. When this reference entry would be compared in the citation index, it will not find a match with any reference entry. For example in the case of CiteSeer which arranges citations according to the length.

In figure 3.12, The authors have made a mistake while writing the title. The



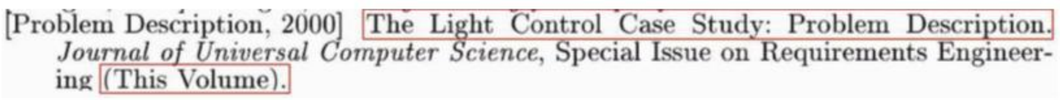
[SN01] R. F. Stärk and S. Nanchen. A Complete Logic for Abstract State Machines. *Journal of Universal Computer Science (J.UCS)*, Abstract State Machines 2001: Theory and Applications, 2001. (this volume).

Figure 3.12: Extra word in the title



[3] Maurer, H., Stubenrauch, R., Camhy, D.: Foundations of MIRACLE - Multimedia Information Repository: A Computer-based Language Effort; J.UCS 9, 4 (2003), 309-348

Figure 3.13: Wrong title information



[Problem Description, 2000] The Light Control Case Study: Problem Description. *Journal of Universal Computer Science*, Special Issue on Requirements Engineering (This Volume).

Figure 3.14: Volume and issue number is missing

word "complete" was added additionally which means that the citation may not be found.

In figure 3.13, the authors have made two errors while writing a title. ":" is replaced with "-". However, this is not a big problem. But the other mistake is crucial: "Computer-supported" is replaced by "Computer-based". Thus it becomes difficult to identify the corrected cited article when the comparison is made within the huge index. Our word and phrase matching algorithm working on a focused subset of the huge index has discovered the correctly cited article.

In figure 3.14, the title of the paper seems correct but still it did not find a match within the existing citation index. The reasons for this are that after the venue name, there is no volume and issue number. It is written as *"This Volume"* which did not find its match. But our technique first identified the venue and then checked for the title as a substring in this entry and found it correctly.

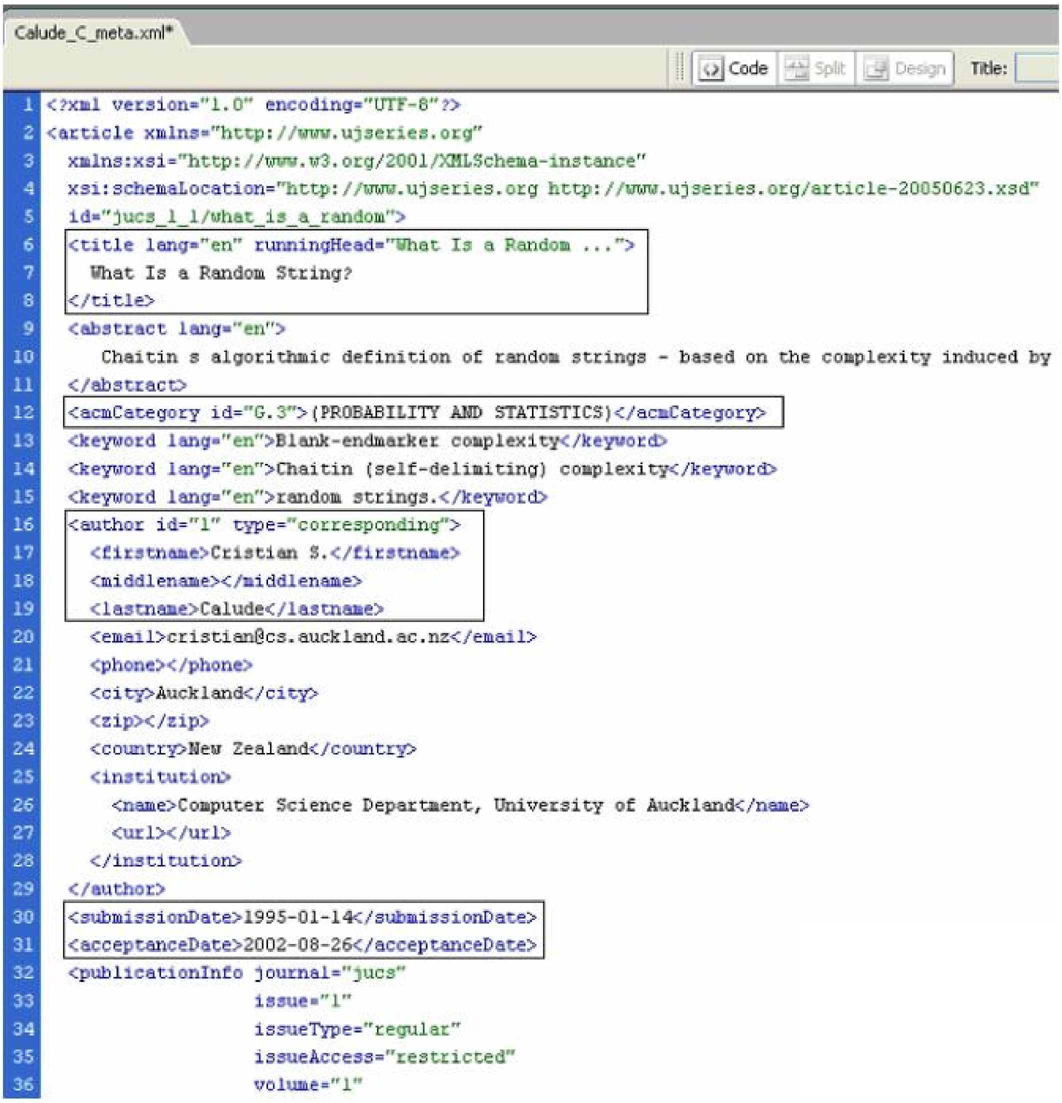
The results of citation mining are also questionable as the citation indexes have difficulties in distinguishing individuals precisely. For example, Ann Arbor, Walton Hall and Milton Keynes (the name of cities) were wrongly classified as actively cited authors [Postellon 2008].

All discovered citations are further used in the Links into the Future system.

The next section explains the other technique "Metadata Extraction Technique" for creating Links into the Future.

3.4 Metadata Extraction Technique

When authors submit papers to J.UCS for publication, metadata (in XML) files are created as they upload a paper. This file is maintained in a hierarchical



```
<?xml version="1.0" encoding="UTF-8"?>
<article xmlns="http://www.ujseries.org"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.ujseries.org http://www.ujseries.org/article-20050623.xsd"
  id="jucs_1_1/what_is_a_random">
  <title lang="en" runningHead="What Is a Random ...">
    What Is a Random String?
  </title>
  <abstract lang="en">
     Chaitin s algorithmic definition of random strings - based on the complexity induced by
  </abstract>
  <acmCategory id="G.3">(PROBABILITY AND STATISTICS)</acmCategory>
  <keyword lang="en">Blank-endmarker complexity</keyword>
  <keyword lang="en">Chaitin (self-delimiting) complexity</keyword>
  <keyword lang="en">random strings.</keyword>
  <author id="1" type="corresponding">
    <firstname>Cristian S.</firstname>
    <middlename></middlename>
    <lastname>Calude</lastname>
    <email>cristian@cs.auckland.ac.nz</email>
    <phone></phone>
    <city>Auckland</city>
    <zip></zip>
    <country>New Zealand</country>
    <institution>
      <name>Computer Science Department, University of Auckland</name>
      <url></url>
    </institution>
  </author>
  <submissionDate>1995-01-14</submissionDate>
  <acceptanceDate>2002-08-26</acceptanceDate>
  <publicationInfo journal="jucs"
                   issue="1"
                   issueType="regular"
                   issueAccess="restricted"
                   volume="1"
```

Figure 3.15: Metadata XML File

representation of Volumes and Issues. This file stores the information on names of authors, submission-date, acceptance-date, title of the paper etc. For exploring Links into the Future, the attributes title, authors, date and topic need to be examined. These metadata or attributes are shown in figure 3.15. We have written an XML parser that parses these XML files to populate our database in proper order. We then search all the papers of the authors in the same topic (including later published papers) and create links from a paper to the selected papers.

The metadata extracting technique can be described as follows:

(I) Select Candidates as Potential Links (into the Future)

1. Select a paper to be considered for creating Links into the Future.
2. b) Find references to authors and co-author's names from the entire list of publication in the metadata file. Extract the entries that contain their names.

(II) Links Verification and Validation

Validate an author's publication (as relevant and from the future) by examining metadata such as date and topic of each entry extracted in (b) A publication is considered a link into the future if:

- The age of publication is less than original document of source and
- The document has the same topic.

We suggest that the use of document similarity checking as a means of finding relevant documents should also be investigated. A user profile will be maintained for all users, to be able to allow the visualisation of types of links (to the future) the user wants to see.

(III) Realisation and Incorporation of Links

1. Construct an internal representation to highlight all discovered information about the author. We have developed publication ontology (see figure 3.16) which will represent currently known information about authors and their publications, together with information about discovered links. The ontological representation of all ontologies is explained in section 3.4.1. As new issues are published, these ontologies are examined and updated accordingly, instead of repeating the metadata extraction all over.
2. Perform visualization of the discovered links to be incorporated into the system.

3.4.1 Ontological Framework to Represent Future Links

In this section we present an ontological framework as a basis for creating Links into the Future. Here we have introduced some basic conceptualization specification for the domain. In the upcoming sections, we have shown an ontological framework by linking these basic ontologies. First of all we want to make it clear that authors, papers and their relationship may contain many specific conceptualizations for different tasks. But here we are interested in formal specification of author and paper through which Links into the Future concept can be realized

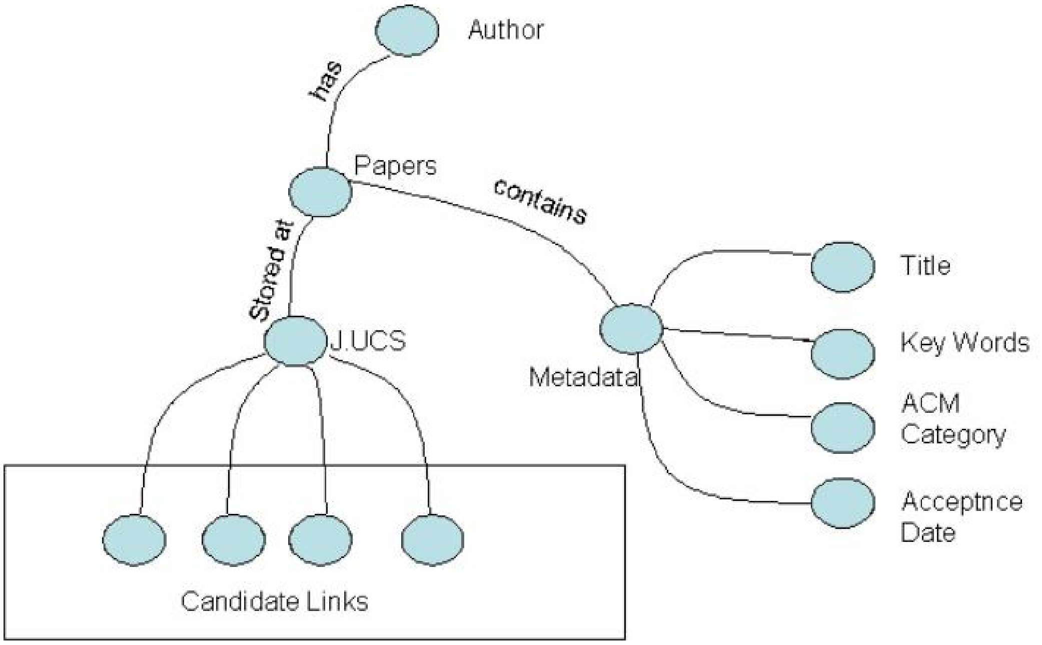


Figure 3.16: Publication ontology

and we are dropping all other concepts that may exist but are out of the scope of this novel idea.

Author's Publication

Authors may have written multiple different papers that can reside on WWW, DBLP, and CiteSeer etc. (see figure 3.17 (a)). This ontology conceptualizes that authors' papers are stored at different mentioned sources.

Paper's Metadata

Papers stored at J.UCS contain detailed metadata file but we are interested in this specific metadata (see figure 3.17 (b)). In J.UCS system, we are maintaining this metadata in an XML file and we have written a service to extract this metadata for the focused document from the J.UCS server. We need the same metadata for other papers residing in WWW to find either they are related or not. For that we are using some SOAP methods to extract information from Web, we are using Google search APIS [Google API 2009], Yahoo search [Yahoo API 2009], and Microsoft Live search [MSN API 2009]. We have built full service oriented architecture and model for using these services to implement the concept of Links into the Future. This has been discussed in details in section 3.4.2.

Author's Onamasticon (Lexicon of Author's Names)

An author is cited as different names in different papers (see figure 3.17 (c)). For example Hermann Maurer may be refereed as H. Maurer, M, Hermann or Hermann Maurer etc. This ontology presents different names sets for a particular author. In the above example for Hermann Manure, there are three name sets. One can claim that M, Hermann could mean Mark Hermann or Maurer Hermann. To resolve this issue we have develop another author's specialization ontology which will match the area of specialization of a particular person before generating link.

Author's Specialization

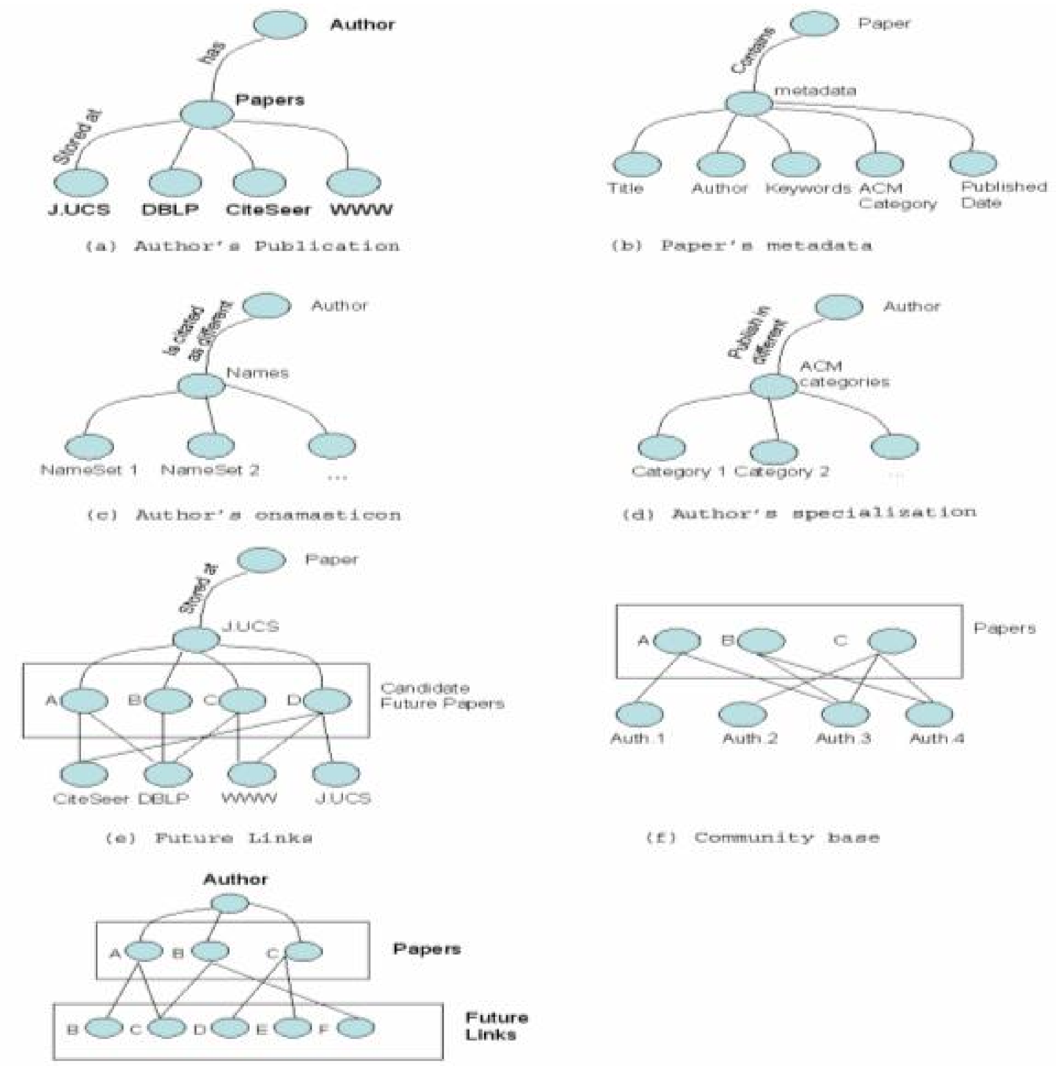
An author has some specialized field of interest (see figure 3.17 (d)). In the field of Computer Science, ACM categories [ACM-CCS, 1998] are good reference for some specialization of an author. While an author submits a paper to J.UCS, authors has to select ACM categories to which it belongs to. Those ACM categories are stored as a metadata of the paper in J.UCS server. As paper belongs to some author, so by combining author's papers and paper's ACM categories, it becomes author's specialization. We have extracted that metadata as depicted in figure 3.15. However, we are getting this information from WWW by using different SOAP APIs as discussed in section 3.4.2.

Future Links

This ontology is elaborating the concept that how a paper stored at J.UCS may have some candidate future links (see figure 3.17 (e)). Candidate papers may be stored at different sources like within J.UCS itself, at CiteSeer, DBLP and WWW. During searching related papers from different sources, a paper is considered more relevant and accurate future paper if the same paper is founded by more than one sources. For example Paper "D" founded by three sources in figure 3.17 (e) is considered more relevant than paper "B" that is founded only by one source.

Community (Co-authors)

This ontology is describing the concept that one paper may be written by several authors (see figure 3.17 (f)). Because in finding future links for a paper, we will find all the related papers by the same team of authors. For example finding future papers for a paper "C" in figure 3.17 (f). we have to find all the papers that are written by Author 2, 3 and 4 in the same ACM category (author's specialization ontology) in future dates as compared to the published date of paper C.



(a) Author's Publication

(b) Paper's metadata

(c) Author's onamasticon

(d) Author's specialization

(e) Future Links

(f) Community base

Figure 3.17: Ontologies for the concept of Links into the Future

Author's Future Papers

After finding future related papers for a particular author. This ontology (see figure 3.17 (f)) helps in implementing whether this discovered link is already linked to the author's papers or not. If it is already linked to the author's paper then the discovered link is dropped otherwise new link is created.

Ontology Merging to Realize the Concept of Links into the Future

Ontologies describe formally concepts in the domain of discourse [Natalya and Deborah 2001]. As our system is composed of different ontologies as discussed in previous sections. We need to merge those ontologies to realize the full strength

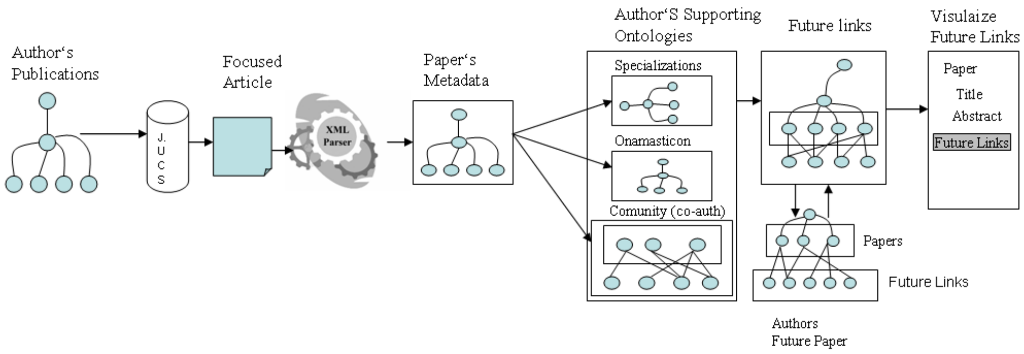


Figure 3.18: Connected ontologies to form the system of Links into the Future

of the system. By connecting ontologies we mean that we are going to show how the system of Links into the Future can be implemented by using these ontologies and at which point, which ontology will be doing what? The complete system can be visualized in figure 3.18. The process can be summarized as follows:

1. Authors ontology will be used to select some particular article stored at J.UCS.

2. Then after extracting article's metadata by using some XML parser, paper's Metadata ontology is built for the selected paper.

3. Then by using Paper's Metadata ontology, related information is fetched from WWW by using community (co-author ontology) which conceptualizes that how many authors have written this paper and subsequently authors supporting ontologies (authors onamasticon and specialization) are checked to verify the validity and similarity of the author.

4. Discovered future links are conceptualized by using Future links ontology.

5. Discovered links are verified from Author's future links ontology, Only new discovered links are incorporated to the author's paper.

6. Future links are visualized then to user.

Experimental Setup

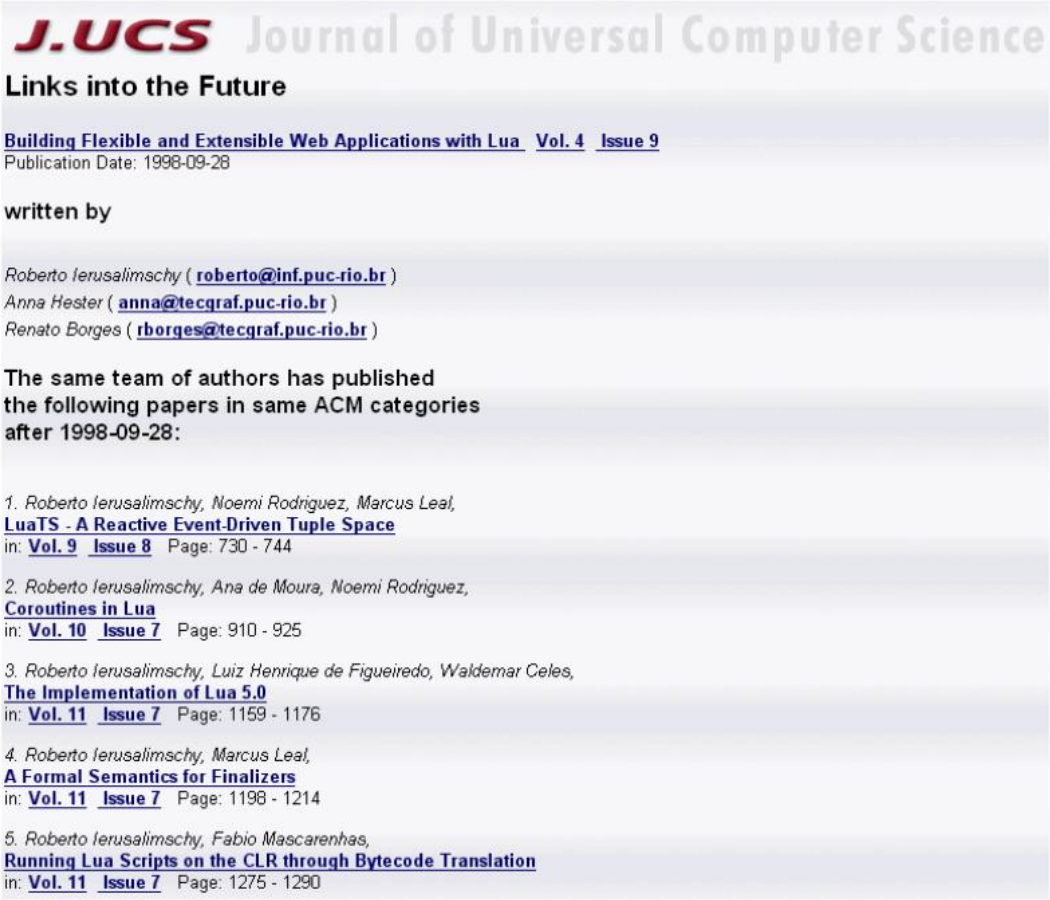
J.UCS has been running over Hyperwave information system since 1994. Hyperwave is one of the leading tools for knowledge management. As we did not want to interfere with the running server, first we build a test server and migrated all data and templates along with document classes from the running server to this test server. When a paper is published in J.UCS, an XML file to represent metadata of the paper is created. That was the starting point for us to write a parser that converted more than 1500 XML files to our knowledge Base (KB) in proper format and order. By traversing paper by paper in the knowledge base, we have created future links for those papers that were also published in J.UCS. Then we visualize the links to user by instantiating a servlet running on tomcat server which takes the selected paper reference as parameter and display a page for the future links for that particular paper after querying the KB (see figure 3.19). We have pre-computed future links in our KB. It would not be a good idea for getting all future links dynamically for a paper by querying external sources because it would slow down the whole process and there is a case that external sources may prevent the access to their server in future. So it is a better idea to have future links from those sources once and place them in the knowledge base and update it on some periodic basis.

We have discussed in the previous section the identification and validation of the candidate documents to be linked. In this section, we are going to talk about realisation and incorporation of the concept of Links into the Future. Here we will discuss some of the results produced by our system. On the first page of a paper in J.UCS, we have introduced a button titled "Links into the Future". When a user is viewing some particular paper and wants to see related future papers, the user simply clicks the button and all related future papers for the same team of authors in the same topic are shown to the user.

Some of the results produced by our system have been shown in figure 3.19. For example, user was viewing a paper titled "Building Flexible and Extensible Web Applications with Lua". It was written by three authors and was published in J.UCS on 28, Sep, 1998. It belongs to "D.2" and "H.5" ACM category. Any paper written by any of the three authors in the same ACM category (we are dealing with only first level of the ACM category like "D" and "H" in this case to get maximum related papers) after or on the same date are shown to user (see figure 3.19). On clicking a discovered future paper title in figure 3.19, user is redirected to the selected paper residing at J.UCS server.

3.4.2 System architecture for Web Documents

In this section we present the system architecture for creating Links into the Future. Ontology based knowledge extraction from Web documents has been focused in project Artequakt [Alani et al 2003]. Alani et al. proposed an ontology-based knowledge extraction from text documents in artist domain. Much information on web exists in natural language documents. To extract some domain specific information from web documents has been challenge for a decade. We have proposed system architecture that how information from web can be acquired in the domain of finding Links into the Future. The proposed system can be seen in figure 3.20. As an overall vision, we are interested in finding related future papers



Links into the Future

Building Flexible and Extensible Web Applications with Lua Vol. 4 Issue 9
Publication Date: 1998-09-28

written by

Roberto Ierusalimschy (roberto@inf.puc-rio.br)
Anna Hester (anna@tecgraf.puc-rio.br)
Renato Borges (rborges@tecgraf.puc-rio.br)

The same team of authors has published the following papers in same ACM categories after 1998-09-28:

1. *Roberto Ierusalimschy, Noemi Rodriguez, Marcus Leal,*
LuaTS - A Reactive Event-Driven Tuple Space
in: Vol. 9 Issue 8 Page: 730 - 744

2. *Roberto Ierusalimschy, Ana de Moura, Noemi Rodriguez,*
Coroutines in Lua
in: Vol. 10 Issue 7 Page: 910 - 925

3. *Roberto Ierusalimschy, Luiz Henrique de Figueiredo, Waldemar Celes,*
The Implementation of Lua 5.0
in: Vol. 11 Issue 7 Page: 1159 - 1176

4. *Roberto Ierusalimschy, Marcus Leal,*
A Formal Semantics for Finalizers
in: Vol. 11 Issue 7 Page: 1198 - 1214

5. *Roberto Ierusalimschy, Fabio Mascarenhas,*
Running Lua Scripts on the CLR through Bytecode Translation
in: Vol. 11 Issue 7 Page: 1275 - 1290

Figure 3.19: Discovered Future Links

of the same team of authors from different sources like Web, CiteSeer, DBLP. However, the focus of this thesis is limited to find Links into the Future from Web documents. There are mainly three modules of the system: knowledge extraction, ontology framework and visualization to the user.

The identification of future links from the web includes the following steps: query formulation, removing duplicates, filtering papers only, similarity algorithm and determining future links. The description of each is shown in figure 3.21.

Knowledge Extraction

Finding some specific information and relationship between them from text/XML document stored on the Web is already a great challenge [Alani et al 2003]. But finding documents that are Links into the Future for a paper is a different task because we are interested mainly in PDF, PS, DOC documents on the Web that also belong to authors who have already published papers in J.UCS. Knowledge extraction uses the ontology framework to extract related information from the

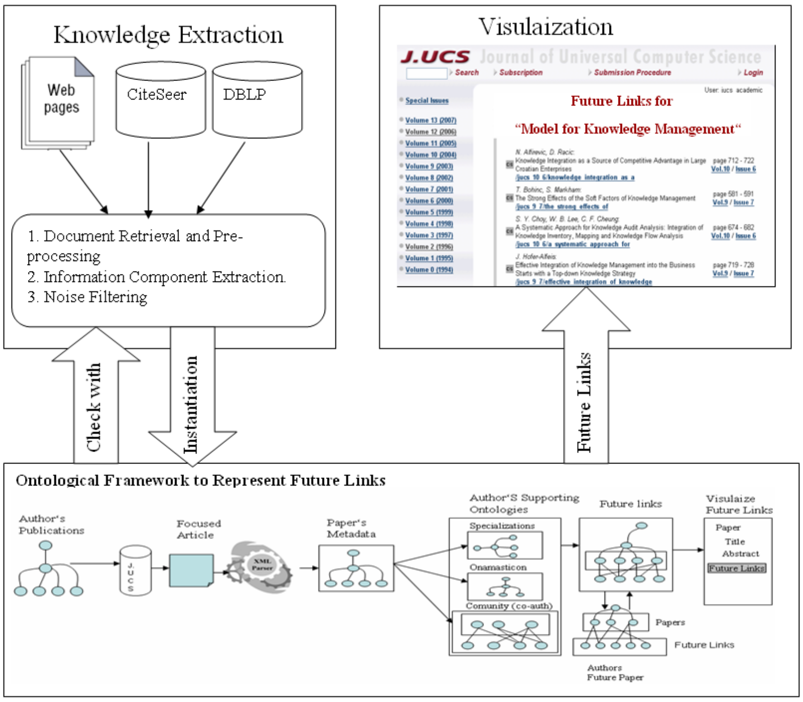
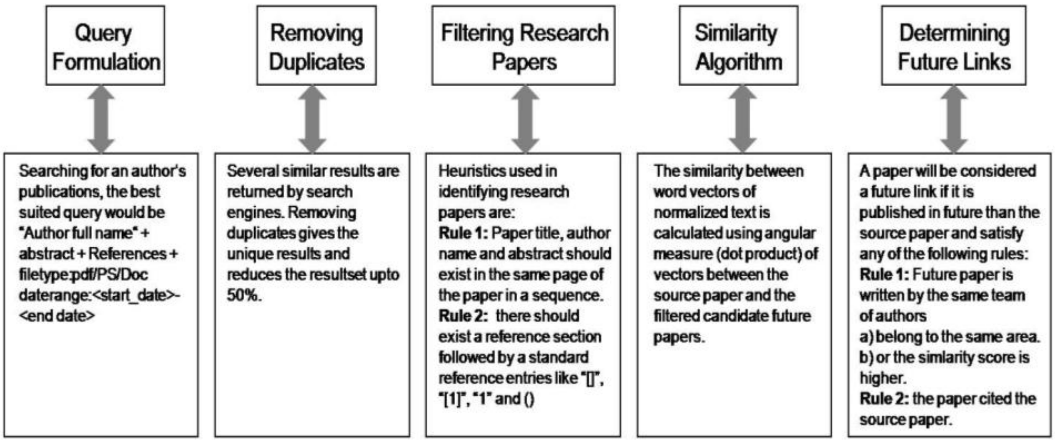


Figure 3.20: System Architecture for Links into the Future

web. For this purpose, we have subdivided knowledge extraction module into three sub modules.

(A) Document Retrieval and Pre-processing

Document retrieval and pre-processing module is responsible for cleaning the raw documents coming from the Web. We are querying Web to find related papers. We use Google search APIS [Google API 2009], Yahoo search [Yahoo API 2009], Microsoft Live search [MSN API 2009] to extract information from the Web. To filter out the raw documents from these sources, we have defined some particular formats of documents to be searched. As we are interested in finding related future research papers of some author, these research papers may be in PDF, PS, and DOC format but may not be in video files, image files, XML documents and all other formats. So this module defines all formats for research papers. We are using a formulated query for searching papers from Web. The J.UCS authors from "author's publication ontology" are used to query. Furthermore, author's Onamasticon ontology is used to describe different name variations. The "community (co-authors) ontology" is also used to query for all co-authors. The retrieved documents may contain duplicate records which are further removed.



Searching for an author's publications, the best suited query would be "Author full name" + abstract + References + filetype:pdf/PS/Doc daterange:<start_date>-<end date>

Several similar results are returned by search engines. Removing duplicates gives the unique results and reduces the resultset upto 50%.

Heuristics used in identifying research papers are:
Rule 1: Paper title, author name and abstract should exist in the same page of the paper in a sequence.
Rule 2: there should exist a reference section followed by a standard reference entries like "[]", "[1]", "1" and ()

The similarity between word vectors of normalized text is calculated using angular measure (dot product) of vectors between the source paper and the filtered candidate future papers.

A paper will be considered a future link if it is published in future than the source paper and satisfy any of the following rules:
Rule 1: Future paper is written by the same team of authors
a) belong to the same area.
b) or the simlarity score is higher.
Rule 2: the paper cited the source paper.

Figure 3.21: Heuristics to identify Links into the Future from the Web

This module can be divided further into two sub modules:

(i) Link Extraction

When querying a search engine, the formulation of query terms strongly affects the results. SOAP APIs have been used by our Web search service to seek Web documents. In performing a search we found that the use of all available semantic information was able to narrow down search space significantly. The effects of query formulation and choice of query terms is shown in Table 3.6. For example for author Hermann Maurer the typical query looks like this abstract references "Hermann Maurer" "H Maurer" "Maurer H" filetype:PDF. This query formulation helped us to retrieve documents which include abstract, references and author name and hence reduced undesired hits from general search engine.

(ii) Duplicates Removal

As a further pre-processing step, duplicates are filtered reducing the results by more than 50%. Documents are then downloaded in parallel into java threads. The importance of removing duplicates is shown in Table 3.7.

For example, for author Hermann Maurer, the formulated query returned 112 results from Google and 495 from Yahoo. However, after removing duplicates, we left with 75 and 86 records from Google and Yahoo respectively.

(B) Noise Filtering

Extracted dataset contains author's future papers and some documents that are not research papers for example CV, business card, publication list of authors etc. The retrieved documents normally contain:

Table 3.6: Query Formulation for accessing the Web Docs.

Query	Google Hits	Yahoo Hits	MSN Live Hits
Hermann Maurer	1,680,00	1,260,000	4,480,00
"Hermann Maurer"	25,600	92,800	27,000
abstract references "Hermann Maurer"	918	1,720	446
abstract references "Hermann Maurer" filetype:PDF	193	775	114

1. Theses supervised by the author.
2. Curriculum Vitae, Home page and Business cards of the author.
3. Conference programmes where the author's name was mentioned.
4. Documents edited by the author.
5. Presentation files
6. The author's publication list.
7. The author may be listed in the reference entries or in the acknowledgement section of a research paper.

As we are only interested in actual research papers at this point, a further filtering step was performed. This process is important in potentially automating the discovery of Web-pages and publication lists.

Docments in PS and DOC file formats are first converted to PDF using MiKTeX [8] and Openoffice tool[9] respectively. Then pdfbox[10], a java library, is used to convert PDF to plain text for further analysis.

A heuristic approach is applied in the actual identification of research papers. The heuristics used are as follows:

[8] http://www.miktex.org/

[9] http://www.openoffice.org/

[10] http://www.pdfbox.org/

Table 3.7: Links into the Future results for selected authors

Author	Focus Paper in JUCS	Search Engine	Formulated	After Duplicate	Classified	Unique Paper	Actual Features
Maurer H	Digital Libraries as Learning and Teaching Support vol. 1 Issue 11	Google	112	75	12	23	17
		Yahoole	495	86	19		
Abraham A.	A Novel Scheme for Secured Data Transfer Over Computer Networks, Vol. 11 Issue 1	Google	148	62	13	33	22
		Yahoole	263	87	41		
Bulitko V	On Completeness of Pseudosimple Sets, Vol.1 issue 2	Google	21	21	7	17	3
		Yahoole	45	28	13		
Shum S. B	Negotiating the Construction and Reconstruction of Organisational Memories, Vol. 3 issue 8	Google	103	81	11	28	21
		Yahoole	546	104	26		
Abecker A.	Corporate Memories for Knowledge Management in Industrial Practice: Prospects and Challenges, Vol. 3 Issue 8	Google	69	59	9	17	15
		Yahoole	335	65	14		

1. Title of the paper followed by author/s name and abstract should exist in the same page. (need not be in the first page). Authors' full name is then searched to disambiguate author/s names.

2. The word "reference" (or "references") is found followed by a proper sequence starting with one of the them "[author]", "[1]", "1" and ().

Documents that were classified as research papers are shown in Table 3.7 for selected authors. These authors were selected randomly from J. UCS author index for this experiment. The used heuristics were found enough to classify retrieved contents as a research paper. Paper classification module gives no false positive. Furthermore, a union set of all retrieved papers is formed by discarding duplicate papers.

(C) Information Component Extraction

This module uses the ontological framework to extract the relevant information component. Extracting information component from J. UCS papers is straightforward as the required metadata is stored in xml file as explained earlier. But when we locate papers from the Web, documents are not categorized according to the ACM topics, and metadata cannot be expected to be found. We then performed similarity detection to automatically discover topics of documents. We measured similarity by taking dot product of vectors from the source and the candidate paper.

The results were, however, not satisfactory due to the following reasons: 1) Author's writing style was usually the same in his/her set of documents. A similar use of common terms produced an impression of being a larger similarity between documents 2) Paper's headers share similar text such as author name, affiliation etc 3) The Reference List at the end of both documents make use of similar text.

To overcome these problems, we pre-processed the text removing the paper's header (section before abstract) and the reference section of the paper to focus only on the original text. We performed Yahoo Term Extraction[11] to extract key terms. This extraction scheme has been used in the number of past studies for extracting facet terms [Dakka et al 2006] [Dakka and Ipeirotis 2008] and building expertise profile [Aleman-Meza et al 2008]. In our case, the results from Yahoo Term Extraction was seen to be not convincing until we removed the header and the references sections. The similarity measured on these terms was able to filter the most relevant papers as can be seen in Table 3.7 and Figure 3.22. For example, in Table 3.7, for the author "Vadim Bulitko", the relevant papers are only 3 out of 17 unique candidate papers found by the paper classification module. The manual inspection revealed that these three were the only papers in the same area.

Working of Ontological Framework

In this section we present how ontological framework is applied to discover Links into the Future from J.UCS and Web documents [Afzal et al 2007a]. We already introduced basic conceptualization specification of different ontologies in previous sections. However, we here concentrate on how these ontologies play their role in the overall system.

(a) Author's Publication Ontology

This was initially populated from authors and their papers published in J.UCS. There were more than 1400 papers that were published in J.UCS. The total number of authors who contributed in J.UCS was 2100. This ontology is further used to cross-check whether the paper found from the Web is different from the papers already stored at J.UCS. If the same paper is found again from Web then that paper is ignored. Otherwise the paper is linked in this ontology according to the source of finding.

[11] http://developer.yahoo.com/search/content/V1/termExtraction.html

(b) Paper's Metadata

As explained earlier, when a paper is published in J.UCS, a detailed metadata file is generated. In the context of Links into the Future, we are interested in the metadata like paper's title, list of authors, keywords, topic of the paper and its publication date. All the metadata files were downloaded automatically from J.UCS server. The paper's metadata ontology was populated from these metadata files [Afzal et al 2007a]. There were more than 1400 papers along with their metadata fields in J.UCS. This served as an initial dataset to populate Paper's metadata ontology. This metadata is further used to find related papers from Web. We use this metadata for finding other papers residing at WWW to find whether the papers are related or not.

(c) Author's Onamasticon (Lexicon of Author's Names)

Author disambiguation is an important task when we are talking about finding papers written by the same team of authors. The same author can be cited or referred to with different name variations. For example Hermann Maurer can be referred to as "Hermann Maurer", "Maurer, H.", "H. Maurer". If we generalize it, an author can be represented in three different ways: "Author full name", "First intial., last name", "Last name, first initial." Author's Onamasticon ontology represents these different variations of an author. Author's Onamasticon ontology was populated with the described set of variations for all authors represented in "author's publication ontology". There were more than 2500 unique authors. All of these authors and their Onamasticon were used to populate this ontology. But this ontology alone is not sufficient for authors' disambiguation. For example if we search for the papers written by "H. Maurer" then we may retrieve some false positives like papers by Henry Maurer or so. This aspect has been solved by author's specialization ontology discussed below along with some general rules as mentioned before.

(d) Author's Specialization

When a paper is published in J.UCS, it is annotated with ACM categories by the authors of the paper. These ACM categories describe an author's specialization. This ontology was populated from the metadata files of J.UCS papers. All of J.UCS authors along with their area of publications ware stored in this ontology.

(e) Future Links

This ontology conceptualizes all candidate Links into the Future for all J.UCS papers. The same paper may be acquired from different sources which enhance its importance and is used to rank accordingly.

(f) Community (Co-Authors)

This ontology describes the concept that one paper may be written by several authors. In finding Links into the Future for a J.UCS paper, we find all the related papers by the same team of authors. For example if a paper is written by three authors then all papers written by any of these three authors in the same area in future dates would be considered Links into the Future for this focused paper. The J.UCS papers and authors are represented in this ontology which helped to find Links into the Future within J.UCS and from the Web.

(g) Author's Future Papers

This ontology extends the concept of "author's publication ontology" where every paper is further linked with the future papers found from the Web. The "future links ontology" is used to cross- check whether the discovered link is new or old. The newely discovered links are updated in "author's Future Papers ontology". While finding Links into the Future from J.UCS to J.UCS papers, more than 500 Links into the Future were found for 250 unique papers. This ontology is dynamically updated from "future links ontology" when a new Links into the Future is found from Web.

Case Study

Figure 3.22 represents an example of a source paper and its candidate future papers. All of these candidates are acquired from Web by using SOAP APIs as discussed earlier. Candidates "C1, C2, C7, C11, C18" were published within J.UCS. The remaining 18 papers were published outside J.UCS. figure 3.22 has been created by using "Graphviz" java toolkit. The link distance between source "S1" and candidate "Cn" node is inversely proportional to the term similarity. The figure is further annotated using key terms from the associated papers. Based on the visual representation it is possible to manually ascertain a threshold for candidate papers that belong to the same area. The threshold for this example has been represented by a dotted circle from source paper to the candidate future papers. In this way, it filters 17 papers out of 23. Here the source paper belongs to the topical areas of E-Learning, digital libraries and teaching support. It is obvious that the papers within the closed circle also belong to the topics of the source paper. The threshold can be altered to refine the closeness of fit of target documents based on usage or application.

Figure 3.23 represents the user interface for this feature. The user viewing the source paper entitled "Digital Libraries as Learning and Teaching Support" at[12] in J.UCS envirnoment, clicked on *"Links into the Future"* button and was shown the screen as in figure 3.23 In the figure, the future links from J.UCS

[12] http://www.jucs.org/jucs$_{11}$1/$digital_libraries_as_learning$

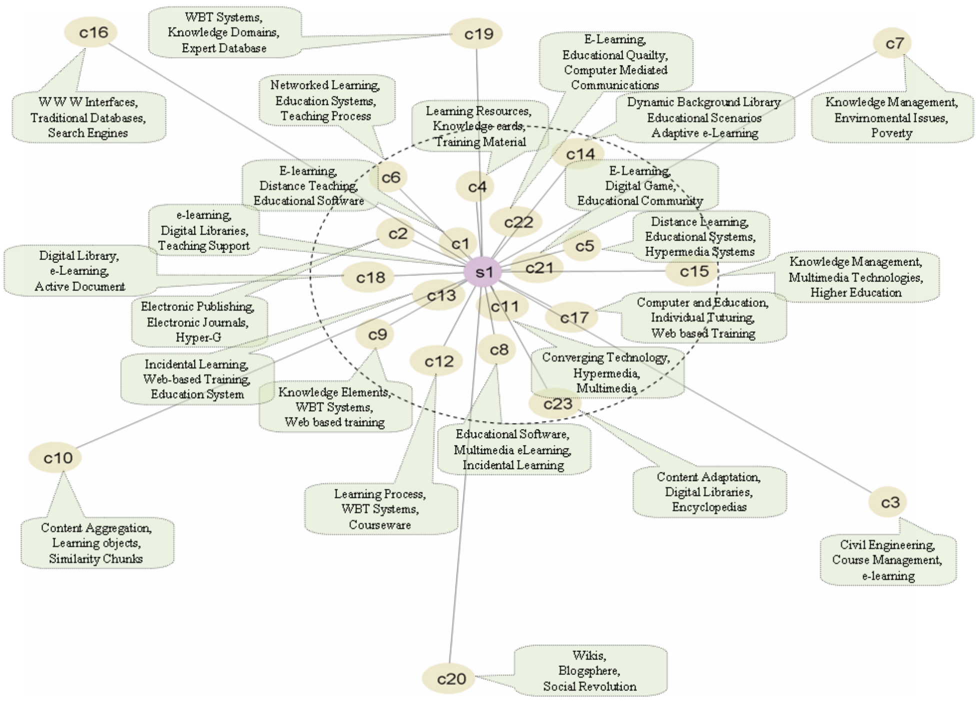


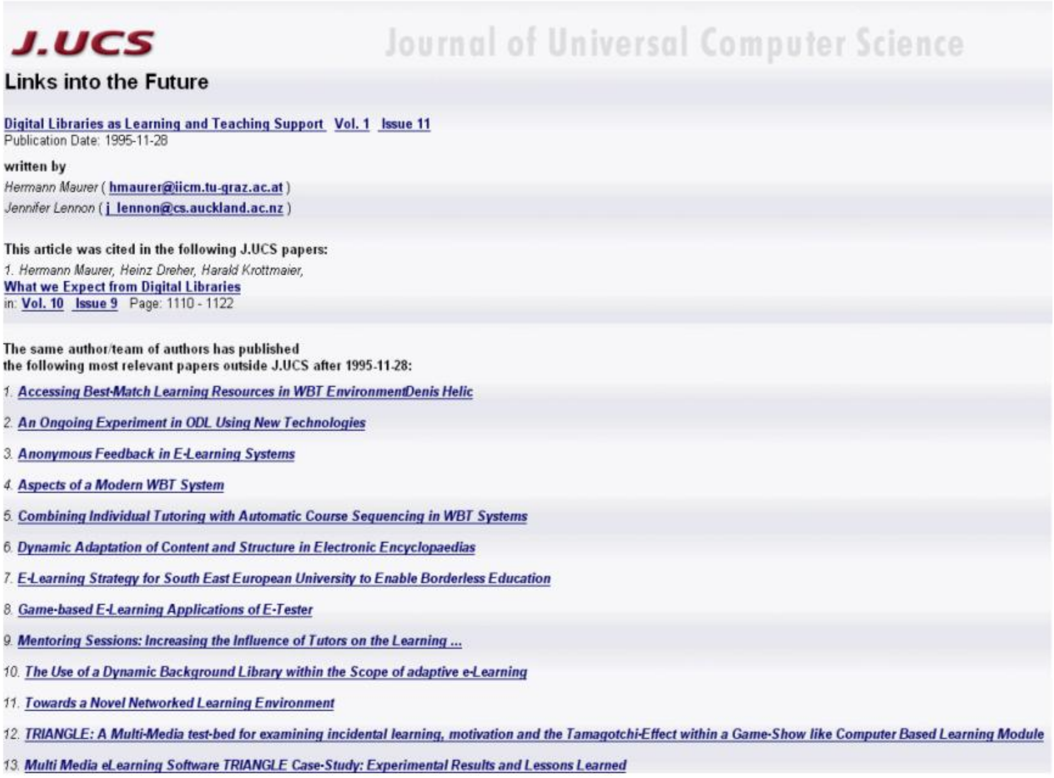
Figure 3.22: Similarity measure score for a source paper and its candidate future links

database (based on metadata similarity and citations) are consolidated with the future links extracted and filtered from web (as shown in figure 3.23). Readers are encouraged to explore this feature in Journal of Universal Computer Science (http://www.jucs.org).

This feature is currently fully implemented for the J.UCS papers and it suggests future related papers that are also published in J.UCS or cited in J.UCS papers. As we are also extending Links into the Future for documents published outside J.UCS, this prototype is being updated.

Update Problems

The execution of the metadata extracting technique and citation mining technique has to be performed incrementally to ensure that all future links are discovered. Since this is not a static repository, either a periodic bulk update or a regular update when new papers come in, has to be performed. The current implementation of the technique has created future links for all papers published until volume 15 issue 4 and monitors every new paper that comes into the system and creates future links into the existing relevant papers pointing to the new paper as soon as a new paper is published.



Links into the Future

Digital Libraries as Learning and Teaching Support Vol. 1 Issue 11
Publication Date: 1995-11-28

written by

Hermann Maurer (hmaurer@iicm.tu-graz.ac.at)
Jennifer Lennon (j_lennon@cs.auckland.ac.nz)

This article was cited in the following J.UCS papers:

1. *Hermann Maurer, Heinz Dreher, Harald Krottmaier,*
What we Expect from Digital Libraries
in: Vol. 10 Issue 9 Page: 1110 - 1122

The same author/team of authors has published
the following most relevant papers outside J.UCS after 1995-11-28:

1. *Accessing Best-Match Learning Resources in WBT EnvironmentDenis Helic*
2. *An Ongoing Experiment in ODL Using New Technologies*
3. *Anonymous Feedback in E-Learning Systems*
4. *Aspects of a Modern WBT System*
5. *Combining Individual Tutoring with Automatic Course Sequencing in WBT Systems*
6. *Dynamic Adaptation of Content and Structure in Electronic Encyclopaedias*
7. *E-Learning Strategy for South East European University to Enable Borderless Education*
8. *Game-based E-Learning Applications of E-Tester*
9. *Mentoring Sessions: Increasing the Influence of Tutors on the Learning ...*
10. *The Use of a Dynamic Background Library within the Scope of adaptive e-Learning*
11. *Towards a Novel Networked Learning Environment*
12. *TRIANGLE: A Multi-Media test-bed for examining incidental learning, motivation and the Tamagotchi-Effect within a Game-Show like Computer Based Learning Module*
13. *Multi Media eLearning Software TRIANGLE Case-Study: Experimental Results and Lessons Learned*

Figure 3.23: Links into the Future interface

3.5 Discussions

For creating Links into the Future, the system employs two techniques: 1) metadata extraction 2) autonomous citation mining. The citation mining technique was able to present valuable information by showing links considered to be relevant by the author while publishing. It may be used to provide directions for further exploration because citations may contain links to other journals. The metadata extraction technique, however, does not take into consideration the information stored in the cited papers. The metadata extracting technique may in future be enhanced by incorporating user specified references.

The advantage of the metadata extraction technique is that it does not depend on the correct formatting of reference section as compared to the citation mining technique. Efforts in enforcing compliance need to be strengthened to further enhance the citation mining technique in the future.

In extracting the metadata from the reference section, one cannot achieve hundred percent results. For example, a reference entry use abbreviations of first and middle name of authors and to extract it accurately from reference entry is a non trivial job. However, this information is useful in validating authors in finding their potential future papers. Using the metadata extraction technique, we are able to do that in a better way because authors are represented with their full names both in XML files of J.UCS, and in the papers acquired from the Web

Using citation mining technique alone in finding all Links into the Future documents may have a shortcoming. The author's decision not to cite a relevant paper in the past will lead to a paper being not represented in the future links section of some other paper. The Metadata extracting technique overcomes this problem.

The metadata technique was able to disambiguate authors by looking for author's full name in the text of paper and focusing on authors' specialization. This approach also avoids the mistaken identity of names of places as author of scientific publications as discussed earlier.

When a user performs a query on search engines, he/she is normally returned with millions of generic hits. The discussed heuristic technique was able to reduce noise at various levels and filters only a small number of most relevant documents.

Alternatively a user has an option to explore citation indexes to search for related papers. But there are two issues 1) times when papers do not exist on these citation indexes like the source paper in our case study was not indexed by CiteSeer. While Google Scholar indexes it but suggests hundreds of related papers. As shown earlier in J. UCS case study that CiteSeer index only 53% of papers 2) a deliberate effort is thus needed to find related papers outside the user's local context.

3.6 Concluding Remarks

We have introduced and implemented a useful new feature within the context of a particular journal. Links to the past already exist in the form of citations. But the concept of *"Links into the Future"* is a new idea which opens more horizons for digital resources. We have illustrated this concept to animate static published contributions to automatically be linked to the previously or later published papers of the same team of authors in a related area. We will explore the expansion of this feature to also find papers for the same area that are written by other authors. The metadata extraction technique for J.UCS is able to support the realisation of Links into the Future. Users are encouraged to browse J.UCS e.g. "Software patents and the Internet" to see some of the Links into the Future that have been created by our system.

This work further describes the extension of the idea of Links into the Future to cover research papers from the Web. The results are promising in providing candidates for future links. The formulated query enables us to retrieve relevant contents from the Web. Furthermore, a set of heuristics helped to filter unique research papers from the retrieved contents. The key term similarity detection has additionally discovered the most relevant papers for a focused paper. The discovered Links into the Future are supplied to users of a digital journal. This information supply is based on the user local context and the task at hand. As further works, we are also exploring the discovery of future related papers from digital libraries like DBLP and CiteSeer.

The citation mining technique 'TIERL' has focused on venue-specific articles prior to determining citations, it was able to disambiguate papers much more efficiently. However, this technique will not work if authors do not specify venues or provide wrong venue information. Our experiments revealed that the error rate in specifying venues was small (1.5% for J.UCS case study and 0.8% for generic experiments). These figures have indicated that although authors make many mistakes when citing references, mistakes in writing venue strings are not as significant. Our experiments have shown that the proposed approach was able to overcome limitations of current citation mining approaches by providing a layered citation discovery. As the implications of not finding correct citation counts can be serious, this approach should be useful for both autonomous systems such as Citeseer and manual approaches such as ISI. All the experimental and statistical data shown in this chapter has been made available at (http://www.jucs.org/jucsinfo/downloads/onlinematerial.rar).

Chapter 4

Linking Digital Journals with Social Bookmarking

This chapter explores shared metadata infrastructure like tagging and bookmarking and finds relationships between tags and citations. This work then investigates how these socially maintained digital libraries can add value to digital journals in finding relevant resources (tags and papers).

The following research questions are addressed in this chapter.

RQ3. How are tags and citations related? Do tags hold potential for measuring research popularity, if yes then how?

RQ4. How can important tag terms from social bookmarking be exploited by digital journals?

The research question 3 is further sub-divided into the following questions:

RQ3.1 What relationship exists between the total number of bookmarks counts and the total number of citations counts for scientific papers?

RQ3.2 Does Tag cloud capture the context of diffusion?

RQ3.3 Can Bookmark counts be used as a proxy for Citation counts?

RQ3.4 What effect/relationship self /coauthor citations have on bookmark count based citation prediction model?

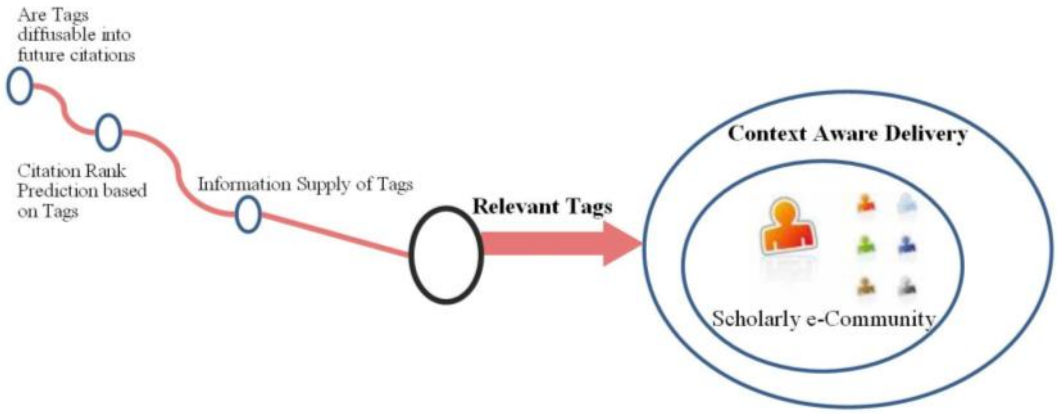

Are Tags diffusable into future citations
Citation Rank Prediction based on Tags
Information Supply of Tags
Relevant Tags
Context Aware Delivery
Scholarly e-Community

Figure 4.1: Progress flow of the chapter based on published contributions

The research approach adopted for the research question 3, was statistical analysis of citation and tagging data. For each of the mentioned question, the following research methods were used:

RQ3.1: Correlation between Bookmarks and citations counts.
RQ3.2: Frequency of tagging keywords reflected in citing titles.
RQ3.3: Linear regression citation rank prediction model based on bookmark counts.
RQ3.4: Correlation analysis with adjusted citations where self citation and coauthor citations are subtracted.

The progress flow of the research is shown in the figure 4.1 which is based on multiple published/accepted contributions [UsSaeed et al 2008a] [UsSaeed et al 2008b] [Afzal et al 2010a].

Sections 4.1 to 4.5 provide answers to the research question 3, after which we exploit important bookmark terms (tags) for the digital journal J. UCS. This task addresses research question 4 and is explained from section 4.6 onward. To address research question 3, we have studied tags and citations behaviours to measure the research popularity and knowledge diffusion. The next section explains knowledge and its diffusion.

4.1 Knowledge and its Diffusion

The vagueness in the use of the term knowledge and its different modalities along with the dynamic and fluid nature of knowledge flow has created a 'semantic and taxonomic' fog [Cowan et al 2000]. We do not intend to refer to that ongoing discussion on knowledge vs. information. Within the scope of our work, we agree

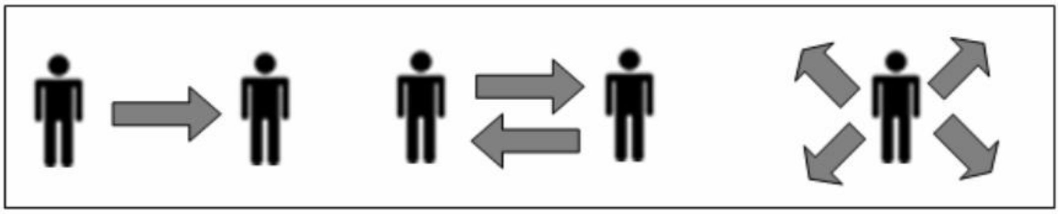


Figure 4.2: Knowledge transfer, sharing and diffusion

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 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.

From a knowledge perspective, we can identify three different types of knowledge flows: (1) knowledge transfer, (2) knowledge sharing and (3) knowledge diffusion as shown in figure 4.2. 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].

In current research, we propose that the knowledge diffuses in two streams: (1) we speak of a 'regenerative knowledge' when its diffusion evolves new (codified) knowledge. (2) We speak of 'knowledge for practice', when people apply knowledge within their practices but do not evolve new codified knowledge.

4.2 Knowledge Diffusion Studies

Knowledge being the primary catalyst for economic and social development of the diffusion of knowledge, it therefore holds an important role in the creation and distribution of knowledge boons. Understanding the diffusion of knowledge leads to more efficient strategies for all stake-holders interested in the dissemination of this valued asset as well as in its measurement.

The structures and properties of knowledge diffusion in scientific domain have been mainly investigated in the past by referring to the diffusion of published (codified) scientific knowledge. 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 knowledge published in the 'citing' paper. Citations are also used to measure the impact of research. It is considered that, to some extent, that citations of a paper or an article are affected by collaborative behaviour. Usually researchers collaborate with each other to establish new ideas and findings of research which they jointly report in their research publications. In most of the publications more than one author share a published work and are called coauthors. Citation analysis and co-authorship 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.

The Office of Scientific and Technical Information (OSTI) of the US Department of Energy, under its strategic initiative 'Innovations in Scientific Knowledge and Advancement', is searching for the 'fast lanes for knowledge diffusion to propel researchers toward scientific discovery'. They are using epidemiological models for modeling knowledge diffusion. It is termed 'epidemiological' after the epidemic diseases. These models were first developed to cope with epidemics. In [Garfield 1980], Garfield E. explains his friends' Bill Goffman and Vaun Newill's model of *"intellectual epidemics"*. He gave the base line SIR (Susceptible, Infections, Recovered) model and its analogies of intellectual 'susceptible' such as researchers or students; intellectually 'infectious material' such as research ideas which are either communicated informally in workshops conferences, discussions etc. or through publications or journals; intellectual 'removals' consisting in those researchers who have died or are not doing research anymore. The OSTI team adapted it and used the SEIR (Susceptible Exposed, Infected, Recovered) epidemic model. Using citations they modeled the collaboration relationship and infection rates. They observed the growth of science in some particular fields by taking the measure of overall growth of the publications related to a particular field or area of research [Bettencourt et al 2006]. OSTI also provided federated deep web search to boost global discovery of scientific knowledge.

There are many Knowledge diffusion studies but three major categories of empirical studies regarding citation analysis of scientific research can be recognized as follows: (1) Diffusion in networks (e.g. study of co-authorship networks), (2) Geographical context (e.g. diffusion of knowledge along the supply chain across the borders), and (3) Technological context (e.g. how are university research results diffused to industry).

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 and Verspagen 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.

In addition to studying 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', opened new horizons for open source data and metadata resources. Kleinberg argues that the web will then bring future evolution in the ways scientists work and in the ways they communicate [Kleinberg 2004]. In addition, this web-based publishing holds the potential to blur the boundaries of formal and informal scientific communication, when for example applications like the 'Encyclopedia of Life' (EOL) may become a very popular future publishing platform for scientists [Us Saeed 2007]. With this transformation of the web as a major communication medium, the research work is getting convoluted with the emerging structures of the web. It is feared that the dynamics of diffusion of scientific literature on the web in future may not be assessable only by the conventional techniques. This emphasizes the need for a particular type of web indicators, one of which may be *"tagging"*, which is within the streams of this new form of web evolution.

However, knowledge diffusion through the informal platforms like EOL may be indicated by measuring the contextualized tagging behaviour of the knowledge seeking of users. We assume that the emerging posting and tagging practices will provide insights in the information seeking behaviour of potential researchers which may publish their related work.

Considering the fact that the web is becoming more and more social, our intention is to probe the potential of tagging according to the knowledge diffusion. We performed an experiment to examine whether tagging holds a potential to indicate the level of regenerative diffusion of knowledge like citations do. 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 which gives a better understanding of the reasons for the usage of knowledge. This chapter contributes to the knowledge diffusion discussion by studying the potential impact of tagging and is based on the results of an exploratory case-study.

4.3 Social Bookmarking Systems and their Potential in Measuring Knowledge Diffusion

Tagging systems are increasingly becoming popular in the web. They 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]. Having the potential to improve the search on the web, tagging systems introduce new forms of social communication and generate new opportunities for data mining. However, we found that tagging systems were not very popular until 2006. One reservation of using tagging systems as a supplementary measure for knowledge diffusion is that these systems have no control on the users for specifying a relevant tag to the resource and are easy to manipulate. This can be true for tagging non scientific content but in our experiments it has been noticed that users do tag a document only after having some understanding of the content and its future use in their particular personal context of its application. Meanwhile, some further efforts may be needed to enhance the tagging applications to make them more strict systems.

Tagging is already a driving component in the fields of emergent semantic techniques [Mika 2005], Information Retrieval [Wu et al 2006] [Hotho et al 2006] and user profiling [Huang et al 2008] [Michlmayr et al 2007]

Wu et al. have shown that *"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 behaviours 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.

We intend to compare the tagging behaviours with respect to the diffusion mechanisms of knowledge and their contexts.

Literature shows that *"context"* becomes an important consideration in any discussion of codified knowledge [Cowan et al 2000]. But previous work shows limited instances of explicating the usage context indication in the diffusion studies. Tsai describes the contextual flow of knowledge within limited scope of an organization [Tsai 2001]. Then there are other studies which take into account the context in 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]. We have shown in Fig. 4.4 that how tagging can be used to contextualize the knowledge diffusion.

One of the existing measures for knowledge diffusion is citations but in this thesis we have explored that tagging may also be used as a supplementary measure in this regard. Citations of existing papers do not necessarily mean that the cited-by paper is regenerating knowledge by using knowledge from the cited papers. About 15 different purposes of citing a particular paper have been identified by Garfield [Garfield 1964]. Some of them can actually be used for studying the contextual knowledge diffusion but not all of them. Sometimes, citations are made to just give a broad level background study for the focused problem and the context of cited paper is not always clear by reading the citing paper. Citation analysis may predict the contextual use of the knowledge if all the documents have a uniform classification which is not the case. The use of citations is also limited to just understanding 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, nevertheless it is used practically in various fields. This knowledge for practice, however, can not be measured by citations.

Citations are studied in different ways like scientific fronts[1], 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 we 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]. On the other hand tagging may tackle the situation in a more convincing way 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. These tags and their proportional percentages can be used to make an automatic taxonomy [Mika 2005].

We explore the potential of tagging with our 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).

[1] http://esi-topics.com/erf/index.html

4.4 Study Framework

We decided to perform an exploratory case study. For this, we have chosen to investigate the accepted 84 scientific papers of the WWW '06 conference, because of the special focus of this conference and its degree of popularity. The intention of the WWW conference series is to discuss and debate the future evolution of the web. We expected to find WWW papers both frequently cited, appearing in citation indexes, and numerously tagged in tagging systems. The higher number of citations indicates the large scale of volumetric knowledge diffusion and high impact of scientific resources. The citation ranks for research papers are normally predicted and considered to be based on different factors. These factors include multi-author publications, geographical positions of co-authors, co-authors' network, and multi-institutional involvement in a publication. On the other hand, bookmarking and tagging applications are considered as the popularity measure for scientific resources. As we are studying and comparing different citation prediction models, we need a dataset of research papers which is within a particular focus related to the web (so that the potential research community is already integrated within the bookmarking systems) and is rich with respect to citations, co-authors' network and its popularity on the Web (bookmarking applications). Taking all these factors into consideration, we have chosen the most highly ranked conference i.e. World Wide Web conference 2006 1. The focus of this conference is the future evolution of Web and it covers all kind of research in the domain of Web. The papers published in this conference are highly cited and popular in tagging and bookmarking applications. The author's network of this conference is also large. We selected all accepted 84 papers from WWW 06 conference.

We took the event from the year 2006, because tagging seemed to be not so popular until 2006 and we assumed that a certain degree of popularity is needed for representing real tagging behaviours. We did not select the event from 2007 or 2008, as a minimum of 1-2 years may be needed to enable the regeneration of the new knowledge.

For the above mentioned study, we explored those papers in three common tagging systems citeulike[2], BibSonomy[3] and del.icio.us[4]. Although BibSonomy and del.icio.us provide search API, our preliminary experiments show that searching a particular paper having some particular characters $(like :, -\backslash_{'"/vs.etc.})$ in its title does not find its match in the tagging application when the whole title of the paper is compared. Another problem also arises when dealing with these tagging applications, which is the repetition of users, moreover even if the same user tags the same paper, he may provide different tags in different times, which leads to miscounting the total number of users for a paper. By keeping these limitations

[2] http://www.citeulike.org/

[3] http://www.bibsonomy.org/

[4] http://del.icio.us/

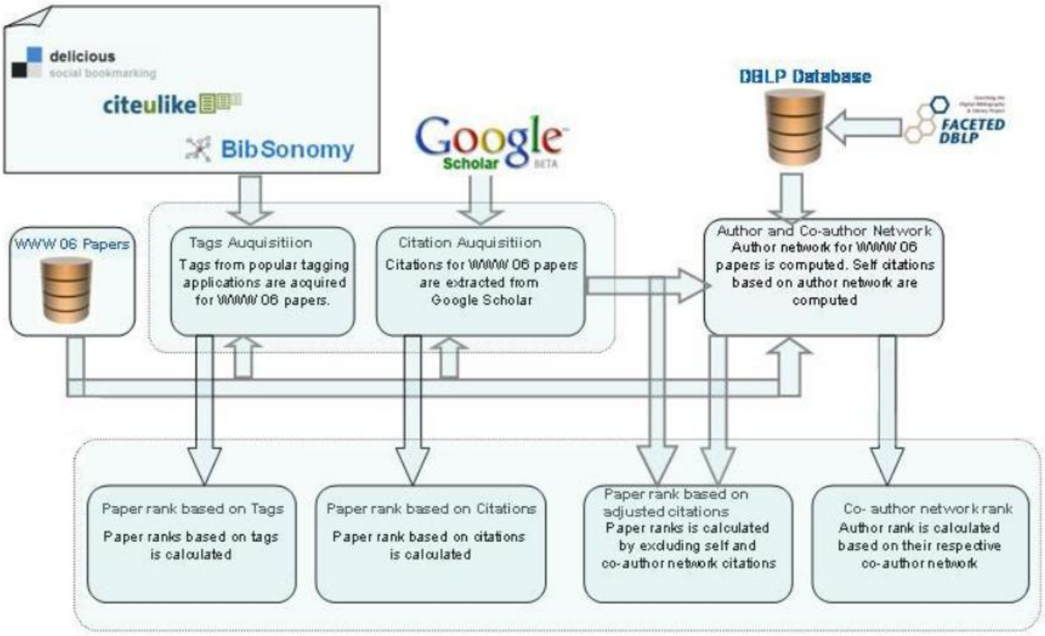

delicious
social bookmarking
citeulike
BibSonomy
Google Scholar BETA
DBLP Database
FACETED DBLP
WWW 06 Papers
Tags Auquisitiion
Tags from popular tagging applications are acquired for WWW 06 papers.
Citation Auquisitiion
Citations for WWW 06 papers are extracted from Google Scholar
Author and Co-author Network
Author network for WWW 06 papers is computed. Self citations based on author network are computed
Paper rank based on Tags
Paper ranks based on tags is calculated
Paper rank based on Citations
Paper rank based on citations is calculated
Paper rank based on adjusted citations
Paper ranks is calculated by excluding self and co-author network citations
Co- author network rank
Author rank is calculated based on their respective co-author network

Figure 4.3: System design for tags and citation analysis

in mind we safely explored the tags and the users in these applications. Meanwhile, we are in a process to employ some heuristic approach to overcome these issues. Citations for these papers were collected by using Google scholar[5] manually as Google Scholar does not provide open access API to explore the citations. We tabulated the dataset year wise from tags and citations with the paper numbers as 'ids' along with their titles taken from WWW 06 website[6]. The ids are maintained in the order of paper titles listed on the website. Figure 4.3 explains different modules of the study design for the current research.

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

4.4.1 Tags Acquisition

Tags and bookmarks for WWW 06 papers were acquired from different tagging applications. We selected CiteULike2, BibSonomy3 and De1.icio.us4 based on their popularity in the Web research community. CiteULike provides dump for publications which can be used by the research community. BibSonomy and Del.icio.us provide search APIs to explore the tagged resources. One can extract tags for a specific paper and number of users who tagged it.

[5] http://scholar.google.com/

[6] http://www2006.org/

Our preliminary experiments show that finding a specific paper with a specific character set (Like - ' vs. I) in its title does not find its match in these applications when the whole title of the paper is compared. By considering these issues, we manually explored a number of users who bookmarked a specific paper. To overcome these issues in the future, we are in a process of developing some heuristic approach. The total bookmarks for the 84 papers were 1051.

4.4.2 Citation Acquisition

Citations for WWW 06 papers were acquired using Google Scholar5. Google Scholar does not provide a search API for citation extraction. Nevertheless Google Scholar was selected because of its large index. Although Thomson ISI is a premier citation index and is considered as an authority in citation indexes, it indexes only a selected number of journals. On the other hand, 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 considers some false positive citations like citations to press releases, resumes, and links to bibliographic records for cookbooks [Price 2004]. But we have safely extracted all citations manually for WWW 06 papers. The total citations for the 84 papers were 1165.

4.4.3 Author's and Co-authors' Network

As citation rank studies are mainly based on co-authors' network. We will compute citation rank for WWW 06 papers based on a number of bookmarks and co-authors' network. To build a co-authors' network, we selected a dataset of DBLP++ [Diederich et al 2007]. This is an enhanced dataset created from 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. DBLP also solves the author's disambiguation problem. We have developed a module which performs four tasks:

1) It finds authors of papers of WWW 06 conference. 2) It finds citing authors for all papers of WWW 06. 3) It computes a coauthors' network based on the original authors of the paper. The Coauthors' network is computed up to 2 degrees of separation. The average co-authors' network for WWW 06 authors was 119. 4) It computes self citations and citations by a co-author's network. As already mentioned there were 1165 overall citation found 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.

4.5 Findings from the Study

4.5.1 Tagging Positively Correlates to Citations

In the initial state of our study, we found a positive correlation (r=0,65, p=2.133 e-11) between the total number of tags and the total number of citations from May 2006 to May 2008 for all the papers. This finding indicates that the tagging behaviour somehow matches with the citation behaviour.

4.5.2 Tagging may have the Potential to Foretell the Future Volume of Knowledge Diffusion

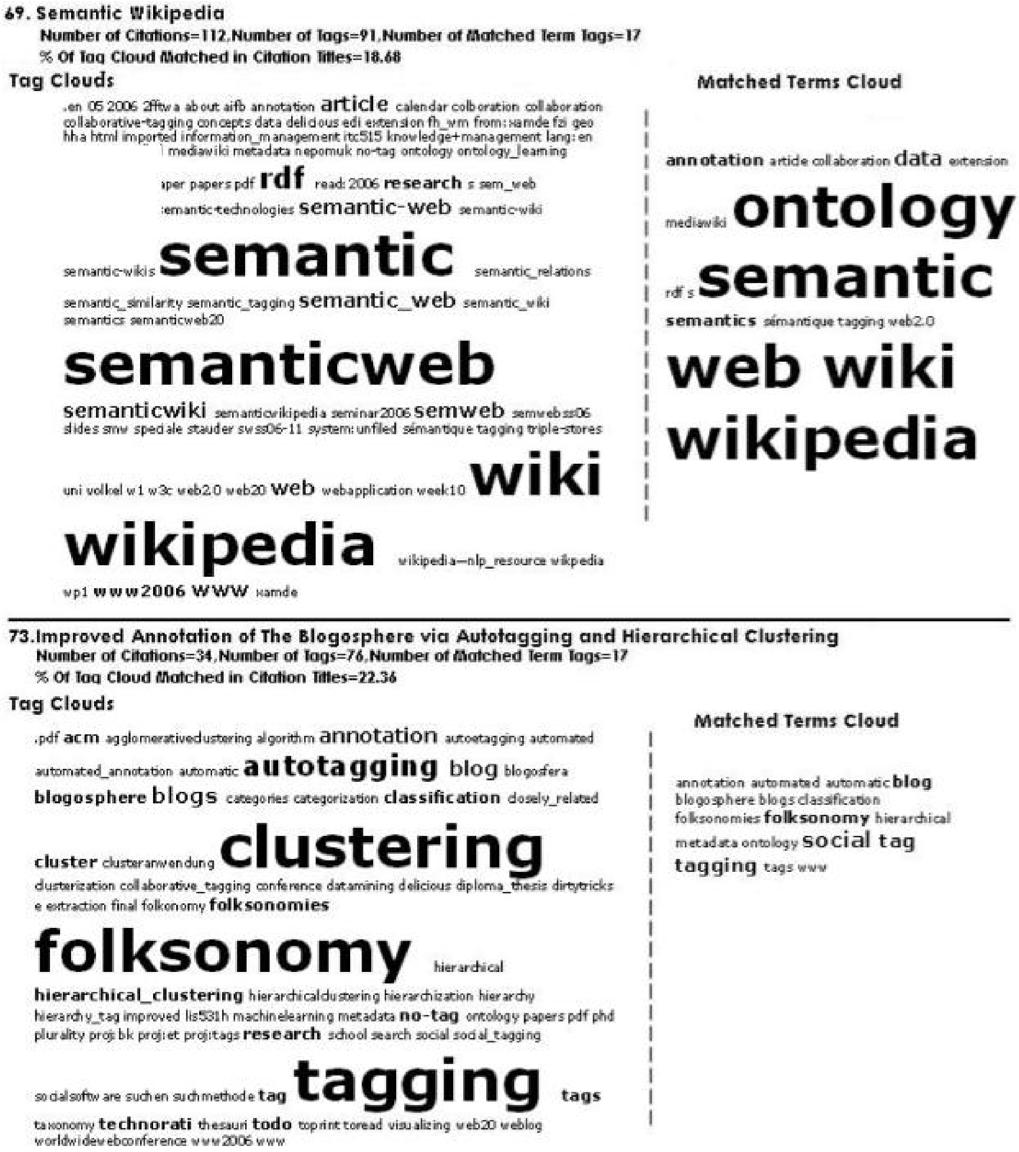
We calculated the average number of users in table 4.1 by adding all the users from three tagging applications for a particular paper and dividing it by three (i.e. number of tagging applications). We observed that if the average is higher than 6, then the tagged paper also gets reasonable number of citations (=7). 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 tags came from the year 2006.

This is logical, because the 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 tags to forecast the future volume of knowledge diffusion.

4.5.3 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, we created two tag-clouds for each paper: 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 paper. 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 4.4.

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. We found that the bigger portion of the tags represent the content of the paper being tagged, while the rest represents the context of future use.



69. Semantic Wikipedia

Number of Citations=112,Number of Tags=91,Number of Matched Term Tags=17

% Of Tag Cloud Matched in Citation Titles=18.68

Tag Clouds

.en 05 2006 2fftwa about aifb annotation article calendar colboration collaboration collaborative-tagging concepts data delicious edi extension fh_wm from:xamde fzi geo hha html imported information_management itc515 knowledge+management lang:en mediawiki metadata nepomuk no-tag ontology ontology_learning aper papers pdf rdf read:2006 research s sem_web emantic-technologies semantic-web semantic-wiki semantic-wikis semantic semantic_relations semantic_similarity semantic_tagging semantic_web semantic_wiki semantics semanticweb20 semanticweb semanticwiki semanticwikipedia seminar2006 semweb semwebss06 slides smw speciale stauder swss06-11 system:unfiled sémantique tagging triple-stores uni volkel w1 w3c web2.0 web20 web webapplication week10 wiki wikipedia wikipedia—nlp_resource wikpedia wp1 www2006 WWW xamde

Matched Terms Cloud

annotation article collaboration data extension mediawiki ontology rdf s semantic semantics sémantique tagging web2.0 web wiki wikipedia

73.Improved Annotation of The Blogosphere via Autotagging and Hierarchical Clustering

Number of Citations=34,Number of Tags=76,Number of Matched Term Tags=17

% Of Tag Cloud Matched in Citation Titles=22.36

Tag Clouds

.pdf acm agglomerativeclustering algorithm annotation autoetagging automated automated_annotation automatic autotagging blog blogosfera blogosphere blogs categories categorization classification closely_related cluster clusteranwendung clustering clusterization collaborative_tagging conference datamining delicious diploma_thesis dirtytricks e extraction final folkonomy folksonomies folksonomy hierarchical hierarchical_clustering hierarchicalclustering hierarchization hierarchy hierarchy_tag improved lis531h machinelearning metadata no-tag ontology papers pdf phd plurality proj:bk projet proj:tags research school search social social_tagging socialsoftware suchen suchmethode tag tagging tags taxonomy technorati thesauri todo toprint toread visualizing web20 weblog worldwidewebconference www2006 www

Matched Terms Cloud

annotation automated automatic blog blogosphere blogs classification folksonomies folksonomy hierarchical metadata ontology social tag tagging tags www

Figure 4.4: Tag cloud comparison of heavily cited and tagged papers

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

Paper ids	Average No. of users per tagging application (>6)	Total user tagged (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

4.5.4 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 in section 4.4.2 are used to rank paper in this model. The high number of citations ranks a paper on the top in this model.

(c) Paper Rank based on Adjusted Citations

As explained in chapter 2, there are some previous studies which talk about the adjustment of scientific impact based on co-authorship and its network. There is a need to adjust the citations by excluding self citations and citation loops [Ioannidis et al 2008]. There is evidence that, to some extent, sharing of self citations may be inflated by co-authorship [Glänzel and Thijs 2004].

(d) Co-authors' Network Rank

In this model, we compute the network of an individual author for all authors of WWW 06 conference. Author's network is computed up to 2 levels. An author is selected for each publication in WWW 06, his co-authors' count is added to form the author's network count. Furthermore 2nd level of coauthors' count is also added to the original author's network count. In this way, the author's network count is calculated for each author of WWW 06 conference. Authors are ranked based on their respective co-authors' count. All authors' network counts for a particular publication are added to form the absolute count for a paper. This model assumes that the papers with high number of authors' and coauthors' count will receive high citations and hence the higher rank.

4.5.5 Results and Discussions

Based on the collected bookmarks, citations and co-authors' network for WWW 06 conference papers, we have explored citation rank model by applying different variables and then compared the results. We have applied linear regression analysis. 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.

(i) 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:

$$0.69 \times \text{variable (bookmark rank)} + 6.21$$

(ii) Citation Rank Prediction Model based on Co-author

In this model co-author's network (calculated in section 3.3) is used as an independent variable while citations are taken as a dependent variable. The linear regression equation model is as follows:

$$0.46 \times \text{variable (coauthor rank)} + 30.27$$

(iii) 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

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

citations. The linear regression equation model is as follows:

$$0.69 \times \text{variable (bookmark rank)} + 6.85$$

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.

The existing studies of citation rank predication are mainly based on formal structure like citations. These studies have considered the factor like multi-author publication, geographical positions of co-authors, co-authors' network, and multi-institutional involvement to predict a citation rank. But with the evolution of Web and bookmarking/tagging applications, it is now possible to study informal structures like bookmarks which are considered as the popularity measures for a publication. 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.

4.6 J.UCS Case Study

In previous sections, it has been shown that there exist a positive correlation between tags and citations. A paper starts getting tags from the users of the social

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

bookmarking system immediately after its publication. This section explains how digital journals can get relevant resources (tags and papers) for papers published within digital journals. For this exercise, we have focused on J.UCS as a source data set. The social bookmarking system used in our experiments was CiteULike. The CiteULike is a social bookmarking system where a huge number of users share scientific papers and tag them accordingly. Our task is to find the most relevant resources from CiteULike for all papers published within J.UCS. On the J.UCS 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 J. UCS papers from CiteULike, we used authors' assigned keywords and compared them with CiteULike tags. The papers at J. UCS are further annotated with the matched tags. Furthermore, the tags are pushed to users by looking to their local context and tasks at hand.

4.6.1 J.UCS Dataset

The dataset for J. UCS was acquired until volume 15, issue 7. The statistics are shown below:

J.UCS total papers = 1460 (until volume 15, issue 7)
J.UCS papers other than managing editor column = 1271
J.UCS papers having one or more author's keywords = 1187
Total keywords for 1271 papers were 5397.
Unique Keywords were 3935.

4.6.2 CiteULike Dataset

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

Total tags in CiteULike = 6.5 million
Total Papers in CiteULike = about 2 million
Unique tags = 348420

4.6.3 Matching Author's Keywords with CiteULike Tags

To match papers' keywords of J. UCS with CiteULike tags, a two-tier approach was adopted. First we tried to find an exact match between papers' keywords and CiteULike tags. Subsequently, a partial match between both datasets was checked. The partial match added lots of value but also introduces some noise. Afterwards, some heuristics were used to clean the noise and the discovered tags were used to annotate the corresponding J. UCS papers.

Direct Match

J.UCS Papers for which at least one keyword is matched= 665/1187 = 56%

All J.UCS Keywords matched = 760/3935 = 19

Partial Match

J.UCS Papers for which at least one tag is matched = 683/1187 = 58% (Collectively)

Total J.UCS Keywords matched =797/3935= 20% (Collectively)

Total CiteULike unique tags matched = 91766/348420

The Partial Match enhances the system discoveries significantly for example the author keyword 'wiki' has found its match in the related popular concepts (Wikis, Semantic Wikis, Wikipedia, Wikification, GeoWiki, WikiNews, wikiproteins, wikipedia-mining, wiki-engine etc).

It was good to find partial match of authors' keywords in CiteUlike tags and the way around was not good as tagging systems use free vocabularies. But authors' assigned keywords are sensible. Although there is a need to clean some abbreviated keywords like CAD, CAD will find its match with all tags having 'CAD' as a substring which is not desired. We need to remove this type of noise as explained in the next section.

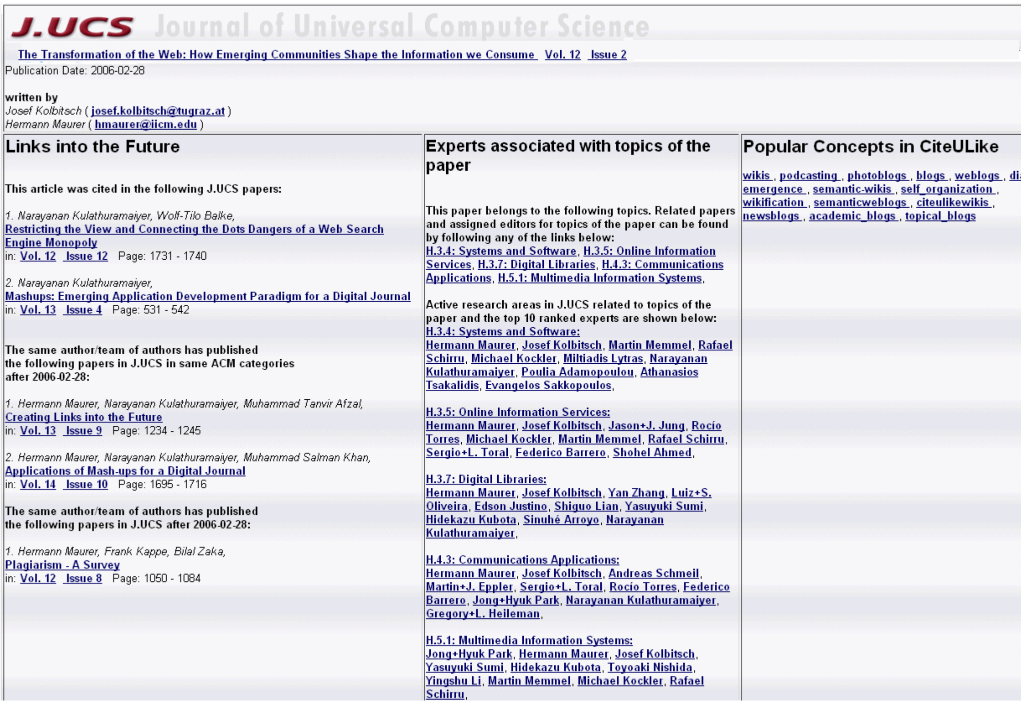
4.6.4 Removing Noise

Based on manual inspection of discovered resources, we noticed that there were some noisy tags. We made some heuristics to clean the data. Tags having length more than 30 or equal to 1 were marked and deleted. Some heuristics worked fine for cleaning abbreviated tag terms.

4.6.5 Pushing Relevant Tags to User's Context

The relevant tags are shown to users by observing their local context. For example a user was viewing a paper entitled *"The Transformation of the Web: How Emerging Communities Shape the Information We Consume"*. The user clicks on *"Links into the Future"* and the user is redirected to a screen like figure 4.5. The popular tags (concepts), ranked with respect to frequency, relevant to the focused paper are displayed to users.

By following any concept, a user has an option to view top ranked relevant papers from CiteUlike. For example, when the user clicks on *"Semantic-wikis"*, he/she is redirected to CiteULike for further knowledge discovery as shown in the figure 4.6. The user not only can see the top ranked related papers but also can view the attached tag cloud. The visualization in both Figure 4.5 and Figure 4.6 increases the knowledge discovery of related papers/tags/concepts for the user.



J.UCS Journal of Universal Computer Science

The Transformation of the Web: How Emerging Communities Shape the Information we Consume Vol. 12 Issue 2
Publication Date: 2006-02-28

written by
Josef Kolbitsch (josef.kolbitsch@tugraz.at)
Hermann Maurer (hmaurer@iicm.edu)

Links into the Future

This article was cited in the following J.UCS papers:

1. Narayanan Kulathuramaiyer, Wolf-Tilo Balke,
Restricting the View and Connecting the Dots Dangers of a Web Search Engine Monopoly
in: Vol. 12 Issue 12 Page: 1731 - 1740

2. Narayanan Kulathuramaiyer,
Mashups: Emerging Application Development Paradigm for a Digital Journal
in: Vol. 13 Issue 4 Page: 531 - 542

The same author/team of authors has published
the following papers in J.UCS in same ACM categories
after 2006-02-28:

1. Hermann Maurer, Narayanan Kulathuramaiyer, Muhammad Tanvir Afzal,
Creating Links into the Future
in: Vol. 13 Issue 9 Page: 1234 - 1245

2. Hermann Maurer, Narayanan Kulathuramaiyer, Muhammad Salman Khan,
Applications of Mash-ups for a Digital Journal
in: Vol. 14 Issue 10 Page: 1695 - 1716

The same author/team of authors has published
the following papers in J.UCS after 2006-02-28:

1. Hermann Maurer, Frank Kappe, Bilal Zaka,
Plagiarism - A Survey
in: Vol. 12 Issue 8 Page: 1050 - 1084

Experts associated with topics of the paper

This paper belongs to the following topics. Related papers and assigned editors for topics of the paper can be found by following any of the links below:
H.3.4: Systems and Software, H.3.5: Online Information Services, H.3.7: Digital Libraries, H.4.3: Communications Applications, H.5.1: Multimedia Information Systems,

Active research areas in J.UCS related to topics of the paper and the top 10 ranked experts are shown below:
H.3.4: Systems and Software:
Hermann Maurer, Josef Kolbitsch, Martin Memmel, Rafael Schirru, Michael Kockler, Miltiadis Lytras, Narayanan Kulathuramaiyer, Poulia Adamopoulou, Athanasios Tsakalidis, Evangelos Sakkopoulos,

H.3.5: Online Information Services:
Hermann Maurer, Josef Kolbitsch, Jason+J. Jung, Rocío Torres, Michael Kockler, Martin Memmel, Rafael Schirru, Sergio+L. Toral, Federico Barrero, Shohel Ahmed,

H.3.7: Digital Libraries:
Hermann Maurer, Josef Kolbitsch, Yan Zhang, Luiz+S. Oliveira, Edson Justino, Shiguo Lian, Yasuyuki Sumi, Hidekazu Kubota, Sinuhé Arroyo, Narayanan Kulathuramaiyer,

H.4.3: Communications Applications:
Hermann Maurer, Josef Kolbitsch, Andreas Schmeil, Martin+J. Eppler, Sergio+L. Toral, Rocío Torres, Federico Barrero, Jong+Hyuk Park, Narayanan Kulathuramaiyer, Gregory+L. Heileman,

H.5.1: Multimedia Information Systems:
Jong+Hyuk Park, Hermann Maurer, Josef Kolbitsch, Yasuyuki Sumi, Hidekazu Kubota, Toyoaki Nishida, Yingshu Li, Martin Memmel, Michael Kockler, Rafael Schirru,

Popular Concepts in CiteULike

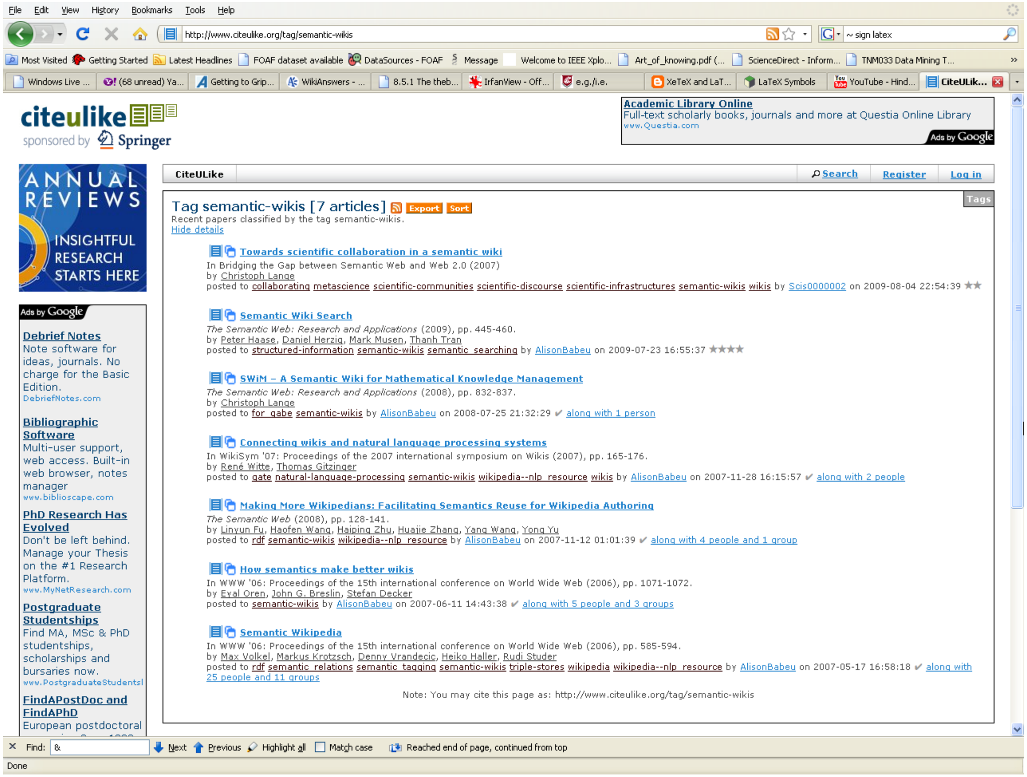
wikis , podcasting , photoblogs , blogs , weblogs , di
emergence , semantic-wikis , self_organization ,
wikification , semanticweblogs , citeulikewikis ,
newsblogs , academic_blogs , topical_blogs

Figure 4.5: Tags/Concept visualization

4.7 Concluding Remarks

In this chapter, we found a relationship between tags and citations. The case study shows that there exist a positive correlation between tags and citations and tags terms reoccur in the titles of the citing papers. Furthermore, the ranking of papers based on tags counts are comparable and sometime better than the co-authors based ranking.

Afterwards, we found that there are some tags which only show the context of future diffusion but a high percentage of tags shows the content of the paper. We linked J. UCS papers with CiteULike papers. For this purpose, we used authors' assigned keywords to J.UCS papers and found relevant tags from CiteULike by direct and partial match. The system was able to find popular tags for J. UCS papers and a user had an option to find other relevant resources (papers) that are annotated with the same or similar tag.



Academic Library Online
Full-text scholarly books, journals and more at Questia Online Library
www.Questia.com

Tag semantic-wikis [7 articles]

Recent papers classified by the tag semantic-wikis.

Hide details

Towards scientific collaboration in a semantic wiki
In Bridging the Gap between Semantic Web and Web 2.0 (2007)
by Christoph Lange
posted to collaborating metascience scientific-communities scientific-discourse scientific-infrastructures semantic-wikis wikis by Scis0000002 on 2009-08-04 22:54:39 ★★

Semantic Wiki Search
The Semantic Web: Research and Applications (2009), pp. 445-460.
by Peter Haase, Daniel Herzig, Mark Musen, Thanh Tran
posted to structured-information semantic-wikis semantic_searching by AlisonBabeu on 2009-07-23 16:55:37 ★★★★

SWiM – A Semantic Wiki for Mathematical Knowledge Management
The Semantic Web: Research and Applications (2008), pp. 832-837.
by Christoph Lange
posted to for_gabe semantic-wikis by AlisonBabeu on 2008-07-25 21:32:29 ✓ along with 1 person

Connecting wikis and natural language processing systems
In WikiSym '07: Proceedings of the 2007 international symposium on Wikis (2007), pp. 165-176.
by René Witte, Thomas Gitzinger
posted to gate natural-language-processing semantic-wikis wikipedia--nlp_resource wikis by AlisonBabeu on 2007-11-28 16:15:57 ✓ along with 2 people

Making More Wikipedians: Facilitating Semantics Reuse for Wikipedia Authoring
The Semantic Web (2008), pp. 128-141.
by Linyun Fu, Haofen Wang, Haiping Zhu, Huajie Zhang, Yang Wang, Yong Yu
posted to rdf semantic-wikis wikipedia--nlp_resource by AlisonBabeu on 2007-11-12 01:01:39 ✓ along with 4 people and 1 group

How semantics make better wikis
In WWW '06: Proceedings of the 15th international conference on World Wide Web (2006), pp. 1071-1072.
by Eyal Oren, John G. Breslin, Stefan Decker
posted to semantic-wikis by AlisonBabeu on 2007-06-11 14:43:38 ✓ along with 5 people and 3 groups

Semantic Wikipedia
In WWW '06: Proceedings of the 15th international conference on World Wide Web (2006), pp. 585-594.
by Max Volkel, Markus Krotzsch, Denny Vrandecic, Heiko Haller, Rudi Studer
posted to rdf semantic_relations semantic_tagging semantic-wikis triple-stores wikipedia wikipedia--nlp_resource by AlisonBabeu on 2007-05-17 16:58:18 ✓ along with 25 people and 11 groups

Note: You may cite this page as: http://www.citeulike.org/tag/semantic-wikis

Debrief Notes
Note software for ideas, journals. No charge for the Basic Edition.
DebriefNotes.com

Bibliographic Software
Multi-user support, web access. Built-in web browser, notes manager
www.biblioscape.com

PhD Research Has Evolved
Don't be left behind. Manage your Thesis on the #1 Research Platform.
www.MyNetResearch.com

Postgraduate Studentships
Find MA, MSc & PhD studentships, scholarships and bursaries now.
www.PostgraduateStudentsl

FindAPostDoc and FindAPhD
European postdoctoral

Reached end of page, continued from top

Figure 4.6: Adapted from CiteULike for tag 'Semantic-Wiki'

Chapter 5

Discovery and Visualization of Expertise

Finding experts in academics as well as in enterprises is an important practical problem. Both manual and automated approaches are employed and have their own pros and cons. On one hand, the manual approaches need extensive human efforts but the quality of data is good, on the other hand, the automated approaches normally do not need human efforts but the quality of service is not as good as in the manual approaches. Furthermore, the automated approaches normally use only one metric to measure the expertise of an individual. For example, for finding experts in academia, the number of publications of an individual is used to discover and rank experts. This chapter illustrates both manual and automated approaches for finding experts and subsequently proposes and implements an automated approach for measuring expertise profile in academia. The proposed approach incorporates multiple metrics for measuring an overall expertise level. To visualize a rank list of experts, an extended hyperbolic visualization technique is proposed and implemented. Furthermore, the discovered experts are pushed to users based on their local context. This chapter addresses the following research questions:

RQ5. How can experts be discovered and ranked in scientific community. Which metrics are the important ones?

RQ6. How can experts be visualized?

Based on multiple published contributions, the progress flow of the research is shown in the figure 5.1. Initially, we proposed an automated approach which was able to measure the overall expertise level by using multiple experience-atoms

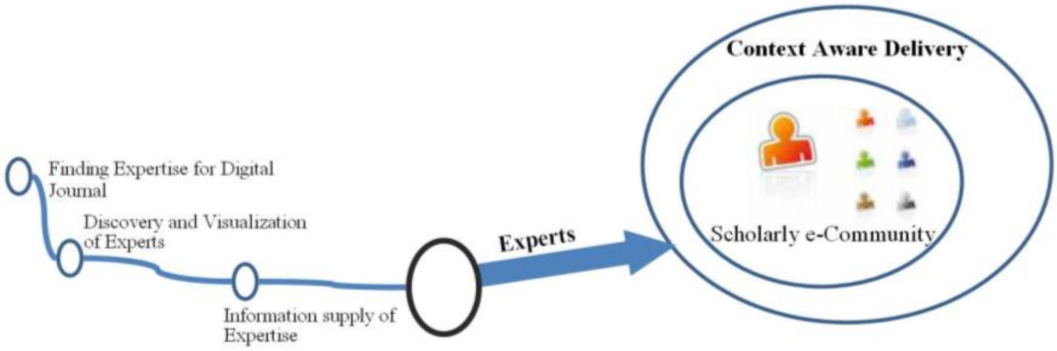

Finding Expertise for Digital Journal
Discovery and Visualization of Experts
Information supply of Expertise
Experts
Context Aware Delivery
Scholarly e-Community

Figure 5.1: Progress flow of the chapter based on published contributions

[Afzal et al 2008]. Subsequently, we developed an extended hyperbolic tree visualization [Afzal et al 2009]. This visualization was helpful in finding new experts (high profiled authors) who could be assigned reviewing duties. The discovered experts are further pushed to users in their local contexts [Afzal 2010].

5.1 Research Overview

The discovery of expertise is crucial in supporting a number of tasks. Finding appropriate experts is a key to unprecedented success in enterprises as well as in academia. Finding an appropriate expert is very helpful when one needs guidance on a subject matter, or needs to fill a vacancy based on relevant expertise, or needs to boost the overall productivity especially in enterprises, or needs to find research collaborators working in similar areas, or needs to find editors/reviewers in peer-review setting etc. Therefore, the expertise finding systems can increase overall productivity and can decrease critical delays due to ineffective work. There are different application areas like Software Engineering [Mockus and Herbsleb 2002], Enterprise [Balog et al 2006], Medicine [Sun and Giles 2007] and Research [Liu and Dew 2004] which employ various techniques to find appropriate experts using both manual and automated approaches.

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. Automated approaches normally use one facet to judge the overall expertise level of an expert. For example, in a discussion forum analyzing only the number of interactions of an individual is used to judge expertise level [Krulwich 1995].

In the peer-review setting, appropriate and capable reviewers/committee-members/editors are discovered by computing their profiles, usually based on the overall collection of their publications [Cameron 2007]. However, the publication quantity alone is insufficient to get an overall assessment of expertise. To incorporate the publication quality in the expertise profile, Cameron used the impact factor of publications' venues (journals, conferences etc.) [Cameron 2007]. However, the impact factor in itself is arguable [Seglen 1997] [Hecht et al 1998]. All publications in a high impact venue do not necessarily get high number of citations. The impact factor of a journal is calculated by considering the number of citations received by all publications published in the journal for a typical period of time [Garfield 1972]. However, Hirsch, a physicist, proposed another metric, the *"H-Index"*, to rank individuals [Hirsch 2005]. The H-Index of an author is calculated by considering the number of citations received by his/her most cited publications. To be precise, a scholar with an index h means that the author has published at least h papers each of which has been cited by others at least h times. However, this index works fine only for comparing scientists working in the same field because citation conventions differ widely among different fields [Hirsch 2005]. Therefore, to measure the quality of one's work in the same field, it is better to calculate the number of citations a person receives rather than just considering the impact factors of journals/conferences where the publications were published. In our approach, additionally we incorporate the number of citations received by an author in a particular topic to make an overall assessment of expertise.

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 during citation mining process as mentioned above, and as described in chapter 3, we introduced an innovative citation mining technique [Afzal et al 2009b]. 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 5.5. 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. To facilitate users of J. UCS, the mined experts are further pushed to users by observing users' local context and task at hand. For example, when a user is viewing a paper, he/she will instantly know about assigned editors and highly active experts associated with topics of the paper. This helps users to establish collaborations in their respective area.

Visualizing a rank list of experts is not enough, users would have an option to explore more aspects of experts, for example short biographies, recent publications, contact information, and affiliations of experts. To support this task, editors are linked with their profiles represented in J. UCS. However, the actively emerging experts (potential reviewers) are further linked with FacetedDBLP. The FacetedDBLP provides a search interface of the huge repository of DBLP. The search interface provides different facets like publication years, co-authors, venues (journal/conference/book series etc) for the selected author. By this means, the users not only know about experts in the respective area, but they can also explore other recent publications of experts and their co-authors, indexed in DBLP.

5.2 A Multi-faceted Expert Profile

In exploring a comprehensive characterization of expertise, we proposed a multi-faceted approach for mining the expertise for a digital journal [Afzal et al 2008]. 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:

Number of publications: This describes the overall expertise areas of a person. The intensity of expertise, however, can be represented by the extent of publications. The number of publications can also be used to indicate the topic specific expertise intensity of researchers.

Number of citations received: Citations are indicative of the impact of publications and as a result can be applied to reflect the impact of expert.

Extent and proportion of citations within a particular area: This further indicates the actual interest of citing authors and the overall contribution in a specific area.

Expert Profile Records: J.UCS has expert profiles for its 300 members in its editorial board representing the specified area of expertise based on ACM categories. This input can be useful as a source in identifying a person as an expert in the area. There are however a number of issues to be considered: areas of interest may change and the research area in itself may evolve.

Experience: Other experiential measures of a person can also be applied in representing one's expertise. Measures that can be acquired with regards to the assessment of experiences include: period of publishing in a particular area, list of projects participated in, assessments of mentoring activities, etc. 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. The extraction of all publications (over 1,400) along with authors and co-authors of the publication is described in [Afzal et al 2007] with a set of over 15,000 references [Afzal et al 2009b].

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). For citation extraction, a technique called Template-based Information Extraction using Rule-based Learning (TIERL) was developed [Afzal et

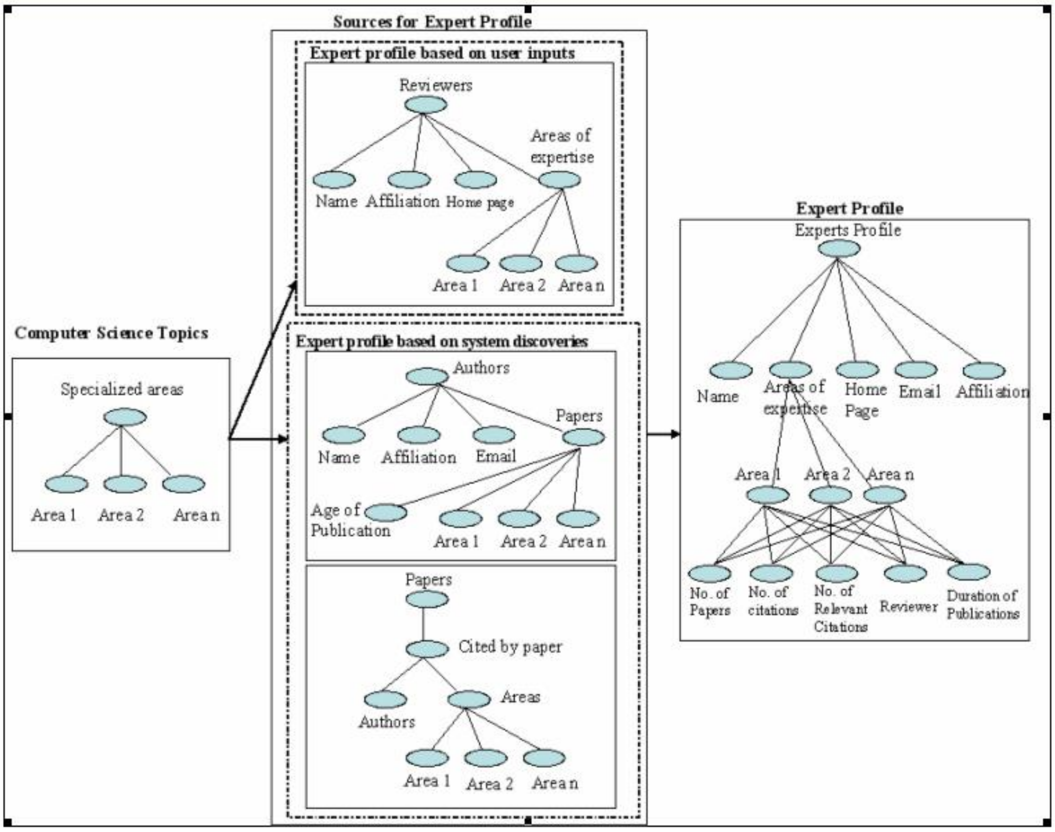
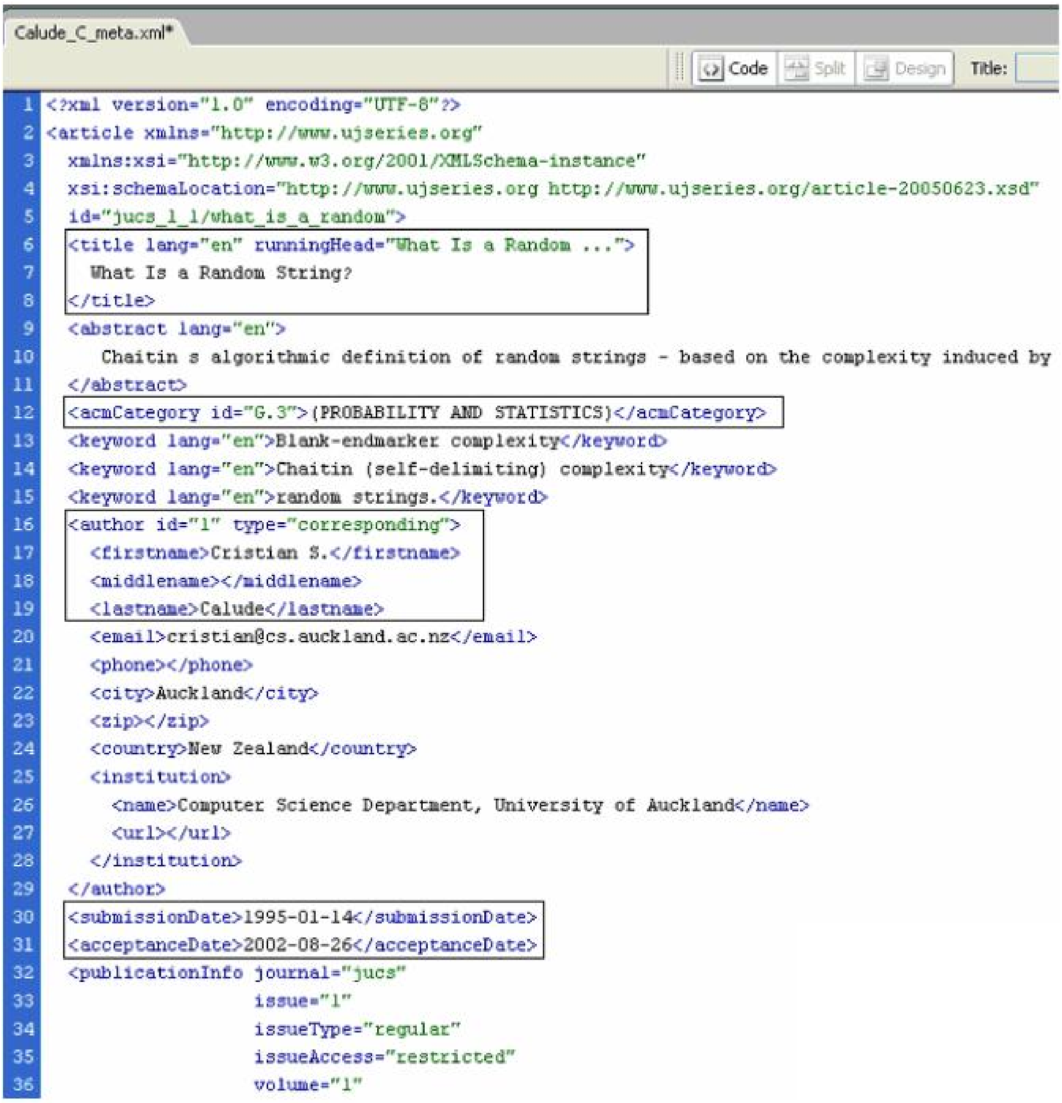


Figure 5.2: Sources for expert profiles

al 2009b] as explained in chapter 3. The TIERL outperformed existing citation extraction approaches (like ISI, Google Scholar, and CiteSeer). The data from this database was then used to calculate and visualize experts within the J.UCS environment.

5.4 Weight Assigned to Experts

There are different ways to calculate expertise for different tasks as explained earlier. Our focus is to measure expertise profile in a scientific community, more specifically for finding a program committee or for finding research collaborators. There is no standard and no absolute definition for calculating expertise. The debate for defining suitable scale for overall assessment of expertise is ongoing. Some argue that publication data alone is insufficient to accurately capture expertise [Seglen 1997] [Hecht et al 1998]. Others counter that bibliographic data is reasonable as experimental facts support their value [Cameron et al 2007]. However, the belief that the quantity of publications is proportional to expertise is not universally true. In a very recent and related work, Cameron explained this problem



```
<?xml version="1.0" encoding="UTF-8"?>
<article xmlns="http://www.ujseries.org"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.ujseries.org http://www.ujseries.org/article-20050623.xsd"
  id="jucs_1_1/what_is_a_random">
  <title lang="en" runningHead="What Is a Random ...">
    What Is a Random String?
  </title>
  <abstract lang="en">
     Chaitin s algorithmic definition of random strings - based on the complexity induced by
  </abstract>
  <acmCategory id="G.3">(PROBABILITY AND STATISTICS)</acmCategory>
  <keyword lang="en">Blank-endmarker complexity</keyword>
  <keyword lang="en">Chaitin (self-delimiting) complexity</keyword>
  <keyword lang="en">random strings.</keyword>
  <author id="1" type="corresponding">
    <firstname>Cristian S.</firstname>
    <middlename></middlename>
    <lastname>Calude</lastname>
    <email>cristian@cs.auckland.ac.nz</email>
    <phone></phone>
    <city>Auckland</city>
    <zip></zip>
    <country>New Zealand</country>
    <institution>
      <name>Computer Science Department, University of Auckland</name>
      <url></url>
    </institution>
  </author>
  <submissionDate>1995-01-14</submissionDate>
  <acceptanceDate>2002-08-26</acceptanceDate>
  <publicationInfo journal="jucs"
                   issue="1"
                   issueType="regular"
                   issueAccess="restricted"
                   volume="1"
```

Figure 5.3: A sample XML File for a paper

with an example [Cameron 2007]. He picked two experts in the field of databases in a scenario where one has a long list of publications in the field while the other has only fewer. In this scenario, on one hand, 'E.F. Codd', inventor of the relational database model, and recipient of the ACM Turing Award in 1981 and 1994, has only 49 articles in DBLP, on the other hand, 'Hector Garcia-Molina', an ACM Fellow too, recipient of the ACM SIGMOD Innovations Award in 1999, had 248 publications in DBLP until 2003 (the year of Codd's death). This example highlights a situation in which a researcher having a large list of publications, may by default, be ranked more prolific than his associates having fewer publications, in spite of publication quality. If one considers publication quantity alone as a measure of expertise, the statistics would conclude 'Garcia-Molina', as far more prolific in the field of databases. However, considering magnificent contributions of 'E.F. Codd' to the field, many may regard it astonishing. To measure a better rank of experts, Cameron employed 'publication impact' as an additional measure to incorporate the quality of the published manuscripts.

However, the impact factor in itself is arguable [Seglen 1997] [Hecht et al 1998] [PLoS Medicine Editors 2006]. The impact factor does not work well since a small number of publications are cited much more than the majority of publication in a particular venue. For example, the well known journal Nature has analyzed the citations of individual papers in Nature and found that 89% of the impact factor was generated by just 25% of the papers [Nature Editorial 2005]. Alternatively, if a publication is of great quality then it will receive a reasonable number of citations. Therefore, to rank experts in a field, it is better to calculate the number of citations of all publications of an individual [Hirsch 2005]. This also applies the above defined scenario. As per Google Scholar database, the most cited paper by 'E.F. Codd' has received 5140 citations as of November 2009 while the most cited paper by 'Hector Garcia-Molina' has received 1408. Therefore, using citations of researchers' publications directly rather than using the impact factor as Hirsch did in calculating H-Index [Hirsch 2005] would be better. In our approach, we have applied the number of publications and citations in an innovative way to calculate overall expertise as explained in the next sections. Apart from publications and citations lists, there might be different measures that can be integrated into the overall weight of the experts. For example, one can use the fact that if a person is serving as a reviewer or on editorial board of some journals and conferences.

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 publicationweight, citationweight, and editorweight.

5.4.1 Publication Weight

In a particular research area, the publication weight of an author is obtained by dividing the number of the author's publications by the number of publications' years (duration of publications). To find active experts, we consider the publications of an author that have been published in the last five years. The number of years is calculated from the year of a first publication (within last five years) until the current year. For example, if an author has published four contributions in the last four years then the publication weight of the author would be one. Consequently, for a specific research area, authors having a larger publication weight would get an edge over their counterparts having fewer publications.

$$\text{Publication Weight}=\frac{\textit{No. of publications}}{\textit{duration (No. of years)}}$$

5.4.2 Citation Weight

The citation weight reflects the author's impact in the growth of a particular research area. For example if all papers in a research area have received 1000 citations collectively and an author's papers in that specific area have received 100 citations, then the citation weight of this author would be 0.1.

$$\text{Citation Weight}=\frac{\textit{No. of citations received by an author}}{\textit{total No. of citations in an ACM topic}}$$

5.4.3 Editor Weight

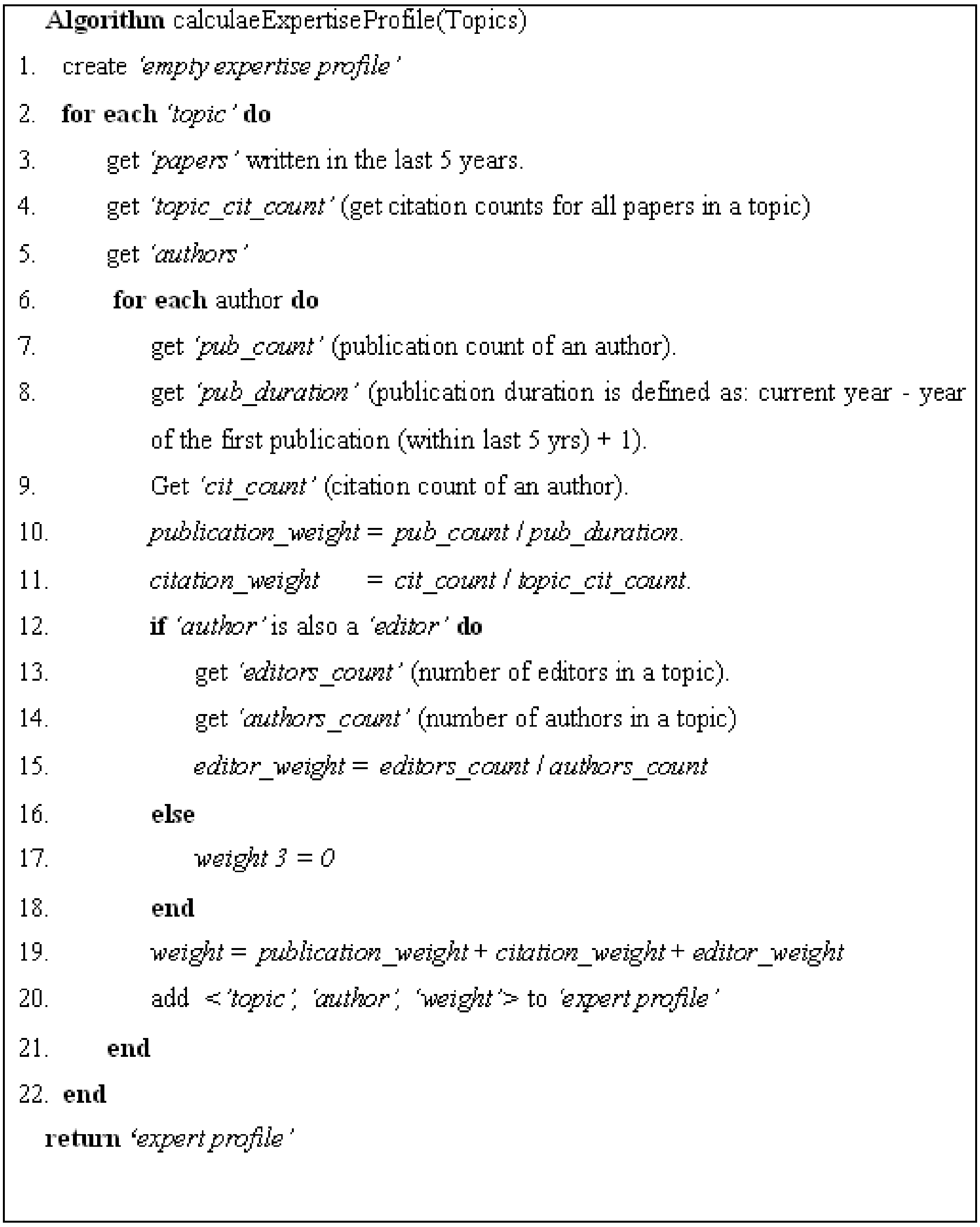
The editors' weight is calculated by dividing the number of J.UCS reviewers by the total number of J.UCS authors. This weight is assigned to only those authors who are also working as editor/reviewers. In this way, reviewers (already acclaimed experts) get an edge over the other authors.

$$\text{Editor Weight}=\frac{\textit{No. J.UCS editors}}{\textit{Total no. of J.UCS Authors}}$$

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

$$\text{Total weight} = \text{publication weight} + \text{citation weight} + \text{editor weight}.$$

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



```
Algorithm calculaeExpertiseProfile(Topics)
create 'empty expertise profile'
for each 'topic' do
    get 'papers' written in the last 5 years.
    get 'topic_cit_count' (get citation counts for all papers in a topic)
    get 'authors'
     for each author do
        get 'pub_count' (publication count of an author).
        get 'pub_duration' (publication duration is defined as: current year - year
        of the first publication (within last 5 yrs) + 1).
        Get 'cit_count' (citation count of an author).
        publication_weight = pub_count / pub_duration.
        citation_weight     = cit_count / topic_cit_count.
        if 'author' is also a 'editor' do
            get 'editors_count' (number of editors in a topic).
            get 'authors_count' (number of authors in a topic)
            editor_weight = editors_count / authors_count
        else
            weight 3 = 0
        end
        weight = publication_weight + citation_weight + editor_weight
        add <'topic', 'author', 'weight'> to 'expert profile'
    end
end
return 'expert profile'
```

Figure 5.4: Algorithm for Computing Expertise Profile

5.4.4 Algorithm to construct an expert profile

The algorithm for calculating an expert profile is shown in figure 5.4. The algorithm takes: topic, papers, citations, and reviewers as input and returns an expert profile for all topics.

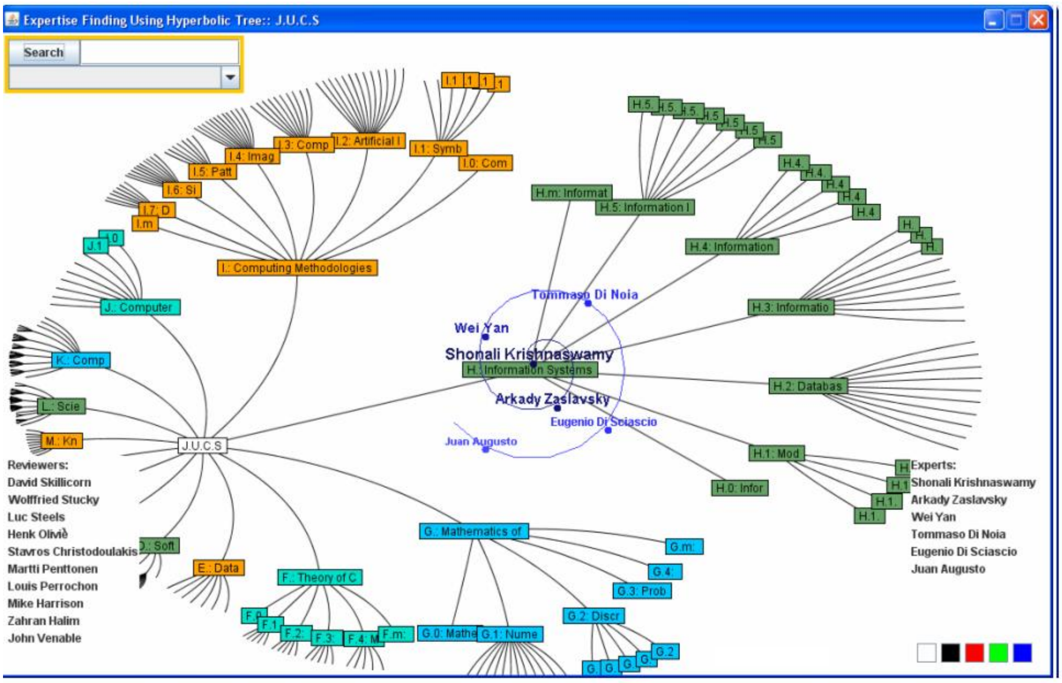
5.5 Information Visualization

Two different visualizations were developed based on measured expertise, one for the journal administration and the other for users of this journal. The visualization for the journal administration is based on the assumption that all topics should be visible in one place where one can easily navigate to a particular topic and can see editors and potential experts belonging to that topic. To make it user-friendly, we have chosen a 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 further extended with a spiral representation of potentially ranked experts. This makes the job of administrator to focus on any particular topic while the overall context remains there. The details of hyperbolic visualization can be found in the next section. The second visualization was developed for users of the journal. This visualization is based on the assumption that the user should have an access to experts whenever he needs them. For the current implementation, when a user is looking on a particular paper and clicks 'Links into the Future', then he/she is shown active experts associated with the topics of the focused paper along with the similar papers written in the same area. The details of 'Links into the Future' can be found in chapter 3. The remaining parts of this section explain both of the aforementioned visualizations.

5.5.1 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 earlier in section 5. 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 5.5. 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



Reviewers:
David Skillicorn
Wolffried Stucky
Luc Steels
Henk Olivié
Stavros Christodoulakis
Martti Penttonen
Louis Perrochon
Mike Harrison
Zahran Halim
John Venable

Experts:
Shonali Krishnaswamy
Arkady Zaslavsky
Wei Yan
Tommaso Di Noia
Eugenio Di Sciascio
Juan Augusto

Figure 5.5: Hyperbolic visualization

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.

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 5.5. 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 coloured buttons. When clicked, the spiral is redrawn with the new colour. 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. When a user drags a particular node, the spiral moves with the focused 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).

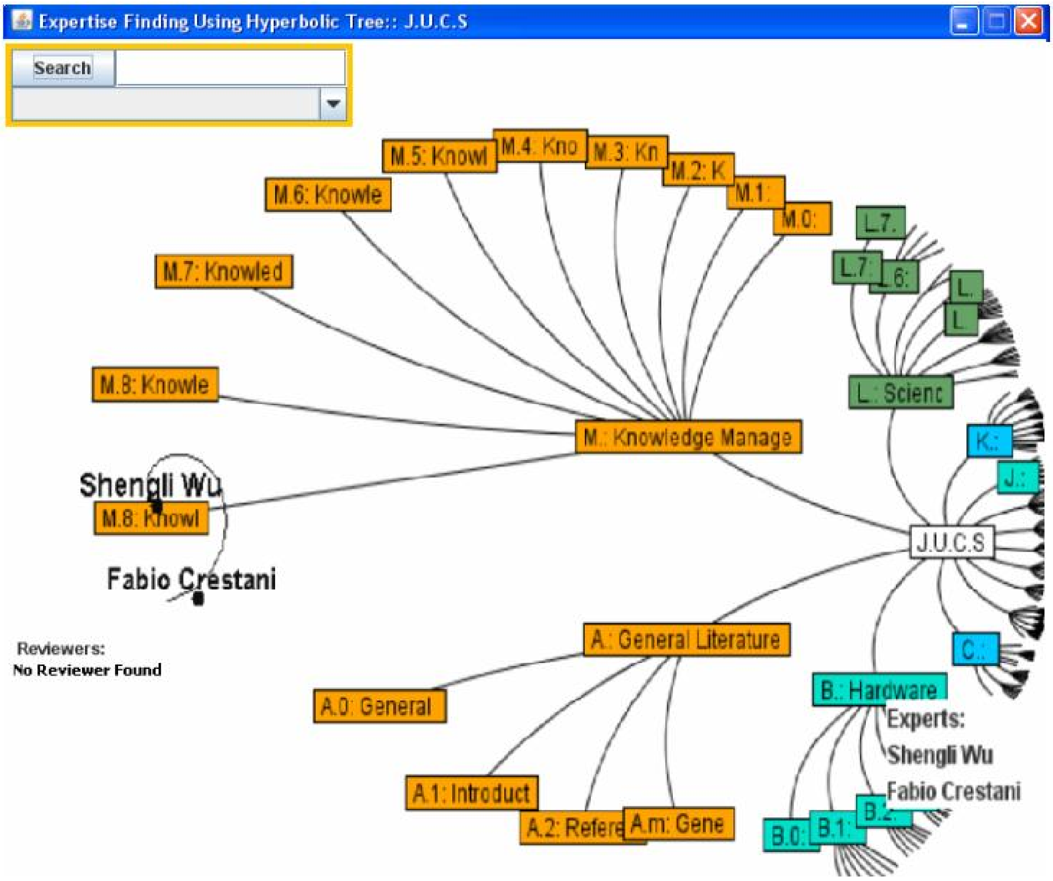


Figure 5.6: Discovery of potential reviewers

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 5.6. This type of discovery is very useful for administrators to locate potential reviewers for any selected area.

Table 5.1 shows a case of one author "Hermann Maurer". The author is already a reviewer for ACM topics: A., H.5.1, K.3, and K.4, in addition, he can be considered as a reviewer for ACM topics: H.1, H.3, H.4, H.5 based on author's contributions to these ACM topics.

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 5.7)

Table 5.1: A comparison between user defined and system discovered expertise

Name: Hermann Maurer:		Total publications in J.UCS = 29		
Already reviewer for:	A. (7%)	H.5.1 (10%)	K.3 (21%)	K.4 (21%)
Can be considered for:	H.1 (14%)	H.3 (17%)	H.4 (24%)	H.5 (14%)

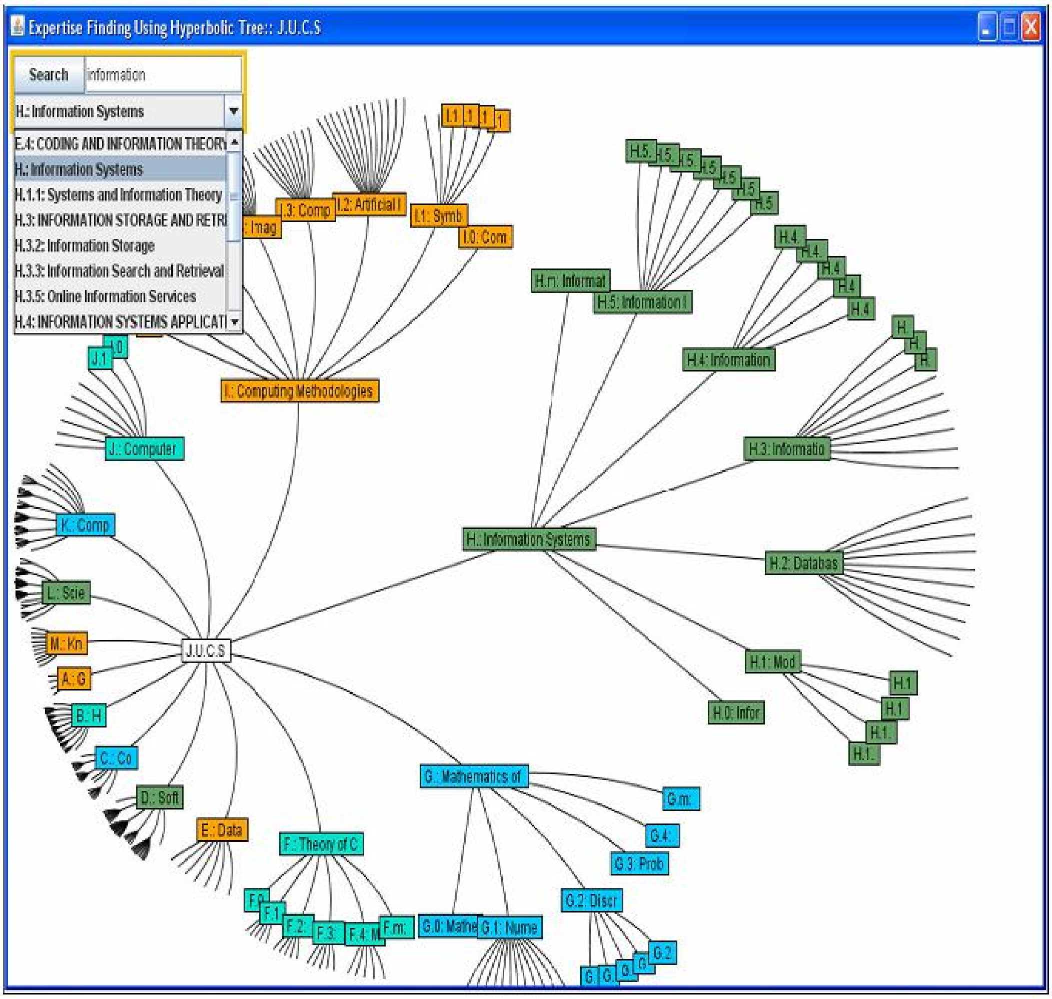
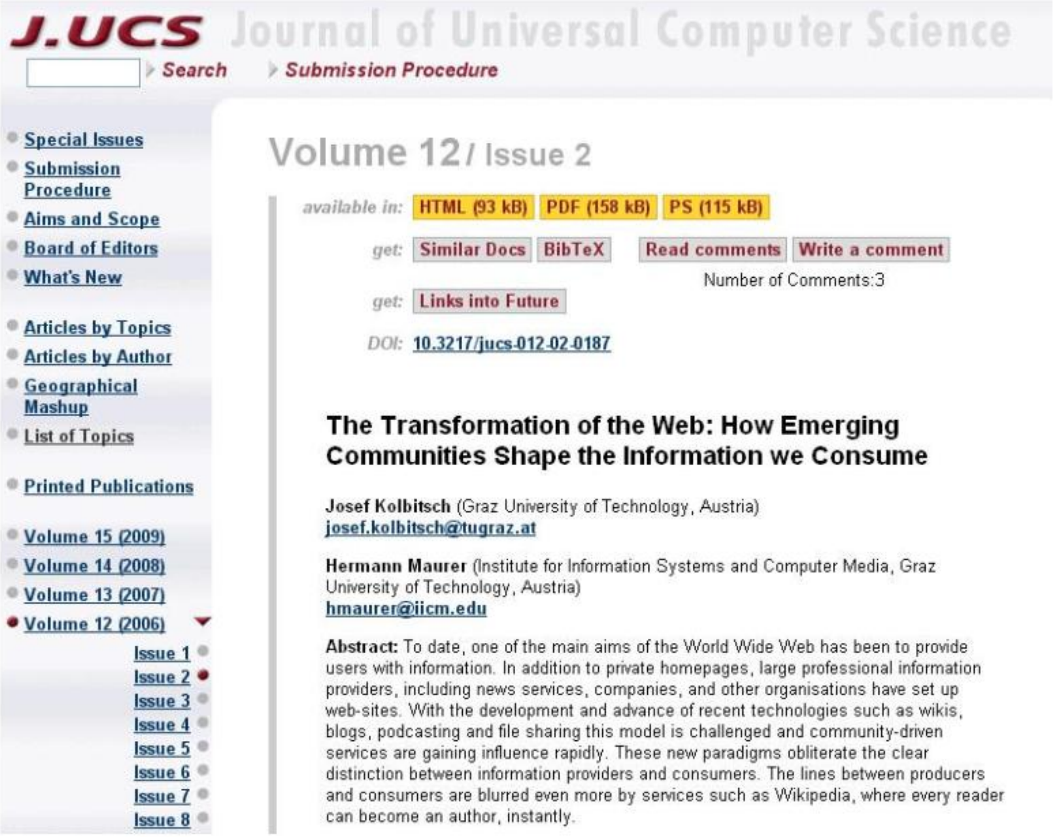


Figure 5.7: Topic search facility

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 5.7. The user can select any ACM topic from the search result list and the hyperbolic tree is redrawn to show the selected topic centred in the window.



The Transformation of the Web: How Emerging Communities Shape the Information we Consume

Josef Kolbitsch (Graz University of Technology, Austria)
josef.kolbitsch@tugraz.at

Hermann Maurer (Institute for Information Systems and Computer Media, Graz University of Technology, Austria)
hmaurer@iicm.edu

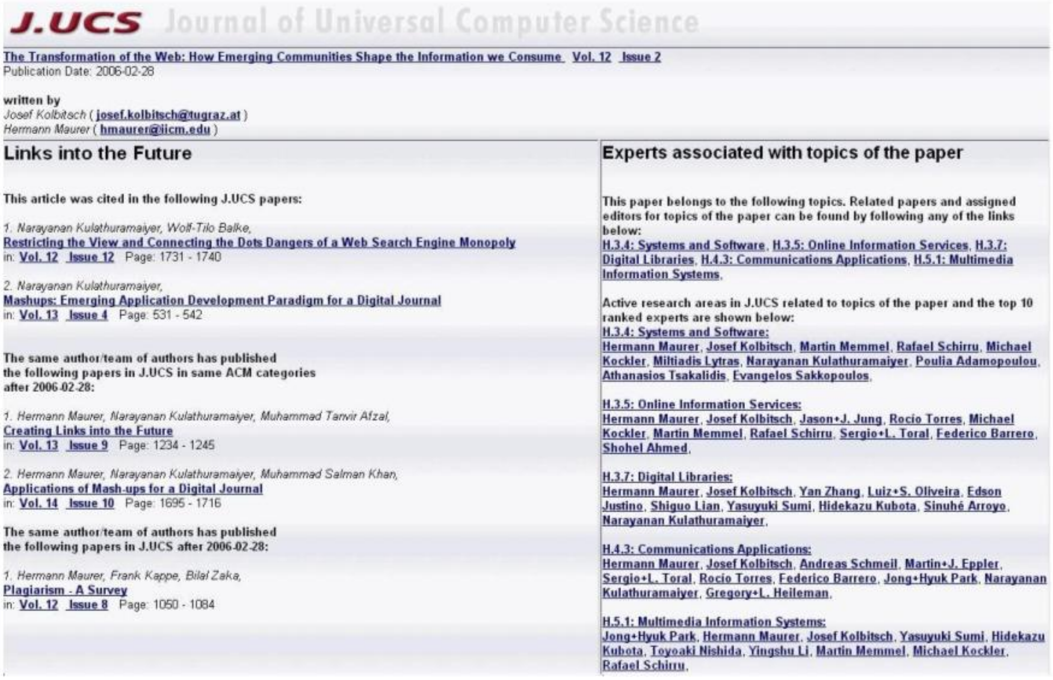
Abstract: To date, one of the main aims of the World Wide Web has been to provide users with information. In addition to private homepages, large professional information providers, including news services, companies, and other organisations have set up web-sites. With the development and advance of recent technologies such as wikis, blogs, podcasting and file sharing this model is challenged and community-driven services are gaining influence rapidly. These new paradigms obliterate the clear distinction between information providers and consumers. The lines between producers and consumers are blurred even more by services such as Wikipedia, where every reader can become an author, instantly.

Figure 5.8: J.UCS interface for viewing a paper

5.5.2 Visualization for Users of J.UCS

The measured experts for topics of the paper are pushed to users by looking at user's local context. For example a user is viewing a paper titled 'The Transformation of the Web: How Emerging Communities Shape the Information we Consume' as shown in figure 5.8.

When the user clicks on the 'Links into Future' button, he is redirected to a screen as shown in figure 5.9. This was implemented using a java servelet. The servelet receives a reference of the viewing paper as a parameter. The servelet subsequently fetch data from different database tables. On the top of this visualization, the focused paper and its metadata are shown. The lower part of the screen is divided into two columns, the left part is dedicated to visualize 'Links into the Future' i.e. related papers written in the same area in future dates as compared to the publication date of the focused paper as explained in chapter 3, while the right part of the screen visualizes the experts associated with the topics of the focused paper. As already mentioned, experts are categorized into two categories: 1) the editors (reviewers) assigned by the editor-in-chief, and 2) the



J.UCS Journal of Universal Computer Science

The Transformation of the Web: How Emerging Communities Shape the Information we Consume Vol. 12 Issue 2
Publication Date: 2006-02-28

written by
Josef Kolbitsch (josef.kolbitsch@tugraz.at)
Hermann Maurer (hmaurer@iicm.edu)

Links into the Future

This article was cited in the following J.UCS papers:

1. Narayanan Kulathuramaiyer, Wolf-Tilo Balke,
Restricting the View and Connecting the Dots Dangers of a Web Search Engine Monopoly
in: Vol. 12 Issue 12 Page: 1731 - 1740

2. Narayanan Kulathuramaiyer,
Mashups: Emerging Application Development Paradigm for a Digital Journal
in: Vol. 13 Issue 4 Page: 531 - 542

**The same author/team of authors has published
the following papers in J.UCS in same ACM categories
after 2006-02-28:**

1. Hermann Maurer, Narayanan Kulathuramaiyer, Muhammad Tanvir Afzal,
Creating Links into the Future
in: Vol. 13 Issue 9 Page: 1234 - 1245

2. Hermann Maurer, Narayanan Kulathuramaiyer, Muhammad Salman Khan,
Applications of Mash-ups for a Digital Journal
in: Vol. 14 Issue 10 Page: 1695 - 1716

**The same author/team of authors has published
the following papers in J.UCS after 2006-02-28:**

1. Hermann Maurer, Frank Kappe, Bilal Zaka,
Plagiarism - A Survey
in: Vol. 12 Issue 8 Page: 1050 - 1084

Experts associated with topics of the paper

This paper belongs to the following topics. Related papers and assigned editors for topics of the paper can be found by following any of the links below:
H.3.4: Systems and Software, H.3.5: Online Information Services, H.3.7: Digital Libraries, H.4.3: Communications Applications, H.5.1: Multimedia Information Systems,

Active research areas in J.UCS related to topics of the paper and the top 10 ranked experts are shown below:
H.3.4: Systems and Software:
Hermann Maurer, Josef Kolbitsch, Martin Memmel, Rafael Schirru, Michael Kockler, Miltiadis Lytras, Narayanan Kulathuramaiyer, Poulia Adamopoulou, Athanasios Tsakalidis, Evangelos Sakkopoulos,

H.3.5: Online Information Services:
Hermann Maurer, Josef Kolbitsch, Jason+J. Jung, Rocio Torres, Michael Kockler, Martin Memmel, Rafael Schirru, Sergio+L. Toral, Federico Barrero, Shohel Ahmed,

H.3.7: Digital Libraries:
Hermann Maurer, Josef Kolbitsch, Yan Zhang, Luiz+S. Oliveira, Edson Justino, Shiguo Lian, Yasuyuki Sumi, Hidekazu Kubota, Sinuhé Arroyo, Narayanan Kulathuramaiyer,

H.4.3: Communications Applications:
Hermann Maurer, Josef Kolbitsch, Andreas Schmeil, Martin+J. Eppler, Sergio+L. Toral, Rocio Torres, Federico Barrero, Jong+Hyuk Park, Narayanan Kulathuramaiyer, Gregory+L. Heileman,

H.5.1: Multimedia Information Systems:
Jong+Hyuk Park, Hermann Maurer, Josef Kolbitsch, Yasuyuki Sumi, Hidekazu Kubota, Toyoaki Nishida, Yingshu Li, Martin Memmel, Michael Kockler, Rafael Schirru,

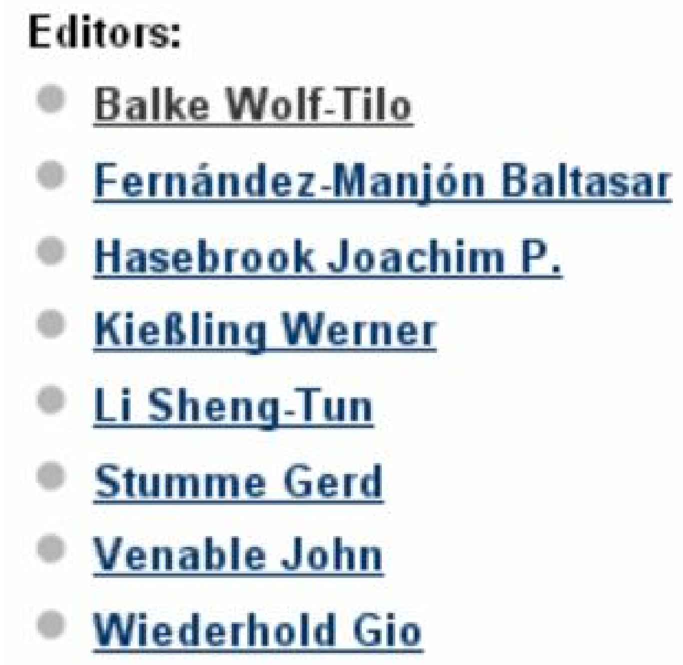
Figure 5.9: Visualization of experts

potential experts flagged by the system. Both categories of experts are shown in this visualization. To find more information about experts, the experts are further linked within J. UCS and with FacetedDBLP [Diederich et al 2007]. The current section and section 5.5.3 gives details about reviewers' linkage within J.UCS while section 5.5.4 explains how discovered experts are linked with FacetedDBLP.

There are more than 300 editors serving as reviewers for J. UCS. There is a many-to-many association between editors and topics. Every editor is usually assigned to multiple topics and each topic is assigned to multiple editors. According to current statistics, there are 45 topics in J. UCS which have more than 10 editors associated. If we visualize all editors for all topics of the paper at one place, then it would become a problem to locate required information for users. To avoid such a situation, an indirect way was used. Initially, the topics of the paper are visualized as shown in figure 5.9. The user can follow any topic to look for all associated editors and published papers in the focused topic. For example from figure 5.9, a user is interested in finding editors of 'H.3.5: Online Information Services'. On click, the user is redirected to the screen as shown in Figure 5.10.

5.5.3 Linking Editors' Profiles in J.UCS

The information about editors is maintained by the J. UCS administration in a highly structured way. This information normally includes: short biography,



Editors:

- Balke Wolf-Tilo
- Fernández-Manjón Baltasar
- Hasebrook Joachim P.
- Kießling Werner
- Li Sheng-Tun
- Stumme Gerd
- Venable John
- Wiederhold Gio

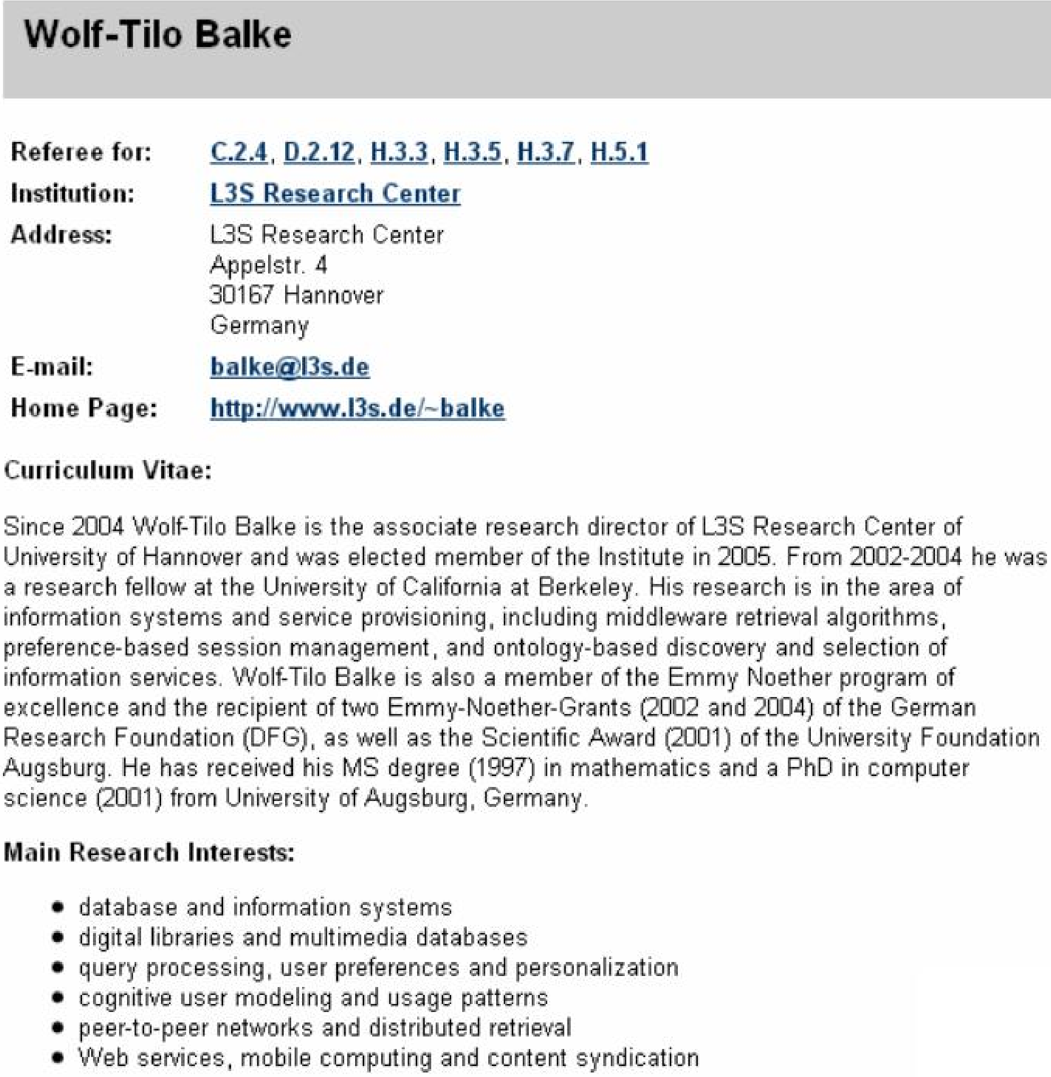
Figure 5.10: Editors associated with topic H.3.5

assigned topics for review, institution, address, email and homepage of editors. For example when a user clicks on *"Balke Wolf-Tilo"* in figure 5.10 to view the details, he is redirected to the screen as shown in figure 5.11. The user is then able to read a brief biography of the expert and can follow to expert's homepage for recent contributions in the area.

5.5.4 Linking Discovered Experts' Profiles to FacetedDBLP

The potential experts (discovered by the system) are shown after the aforementioned visualization as can be seen in figure 5.9. Only the active research areas (having contributions in the last five years) and their experts are visualized. For example the focused article in figure 5.9 belongs to five topics and all of them remain active research areas in the last five years in J. UCS. The top 10 ranked experts are visualized for each topic of the paper. To gain deeper insights into the experts' contributions, these experts are further linked with FacetedDBLP [Diederich et al 2007]. This FacetedDBLP is build upon the large collection of DBLP data set.

For example a user clicks on the author's name *"Jong Hyuk Park"* (topic H.5.1) in Figure 5.9. The user is redirected to the screen as shown in figure 5.12. This was achieved by querying FacetedDBLP http://dblp.l3s.de by adjusting author's first, middle, and last names using some heuristics. The FacetedDBLP is based on the large repository of DBLP (DBLP currently index more than 1.3 million



Wolf-Tilo Balke

Referee for: C.2.4, D.2.12, H.3.3, H.3.5, H.3.7, H.5.1

Institution: L3S Research Center

Address:
L3S Research Center
Appelstr. 4
30167 Hannover
Germany

E-mail: balke@l3s.de

Home Page: http://www.l3s.de/~balke

Curriculum Vitae:

Since 2004 Wolf-Tilo Balke is the associate research director of L3S Research Center of University of Hannover and was elected member of the Institute in 2005. From 2002-2004 he was a research fellow at the University of California at Berkeley. His research is in the area of information systems and service provisioning, including middleware retrieval algorithms, preference-based session management, and ontology-based discovery and selection of information services. Wolf-Tilo Balke is also a member of the Emmy Noether program of excellence and the recipient of two Emmy-Noether-Grants (2002 and 2004) of the German Research Foundation (DFG), as well as the Scientific Award (2001) of the University Foundation Augsburg. He has received his MS degree (1997) in mathematics and a PhD in computer science (2001) from University of Augsburg, Germany.

Main Research Interests:

- database and information systems
- digital libraries and multimedia databases
- query processing, user preferences and personalization
- cognitive user modeling and usage patterns
- peer-to-peer networks and distributed retrieval
- Web services, mobile computing and content syndication

Figure 5.11: Editor profile maintained by J.UCS

computer science publications). In the figure, there are 65 publications of the author *"Jong Hyuk Park"* found in FacetedDBLP. A user can search using different facets as shown on the left side of the figure like: publication years, publication type (article, proceedings, etc), venues (journals, conferences etc), authors, and the GrowBag graph. Based on the user selection, search results are shown on the right side of the figure. For example a user can search all papers of the focused author which were co-authored with any of the authors shown on the left side. The user can restrict the search results to find papers which appeared only in any of the venues (like: computer communication, The Journal of Supercomputing, etc). The user can characterize the result set in terms of the main research topics and filter it according to certain subtopics. The GrowBag terms may be very useful for the user. For example a user can restrict the result set to see only papers of the focused author which deal with any of the shown GrowBag terms (like security, pervasive computing, privacy protection etc). Therefore, a user may find required information more efficiently and accurately using this interface and instantly becomes aware of the research areas of the authors, his collaborators list, the venues where the author has published, etc.

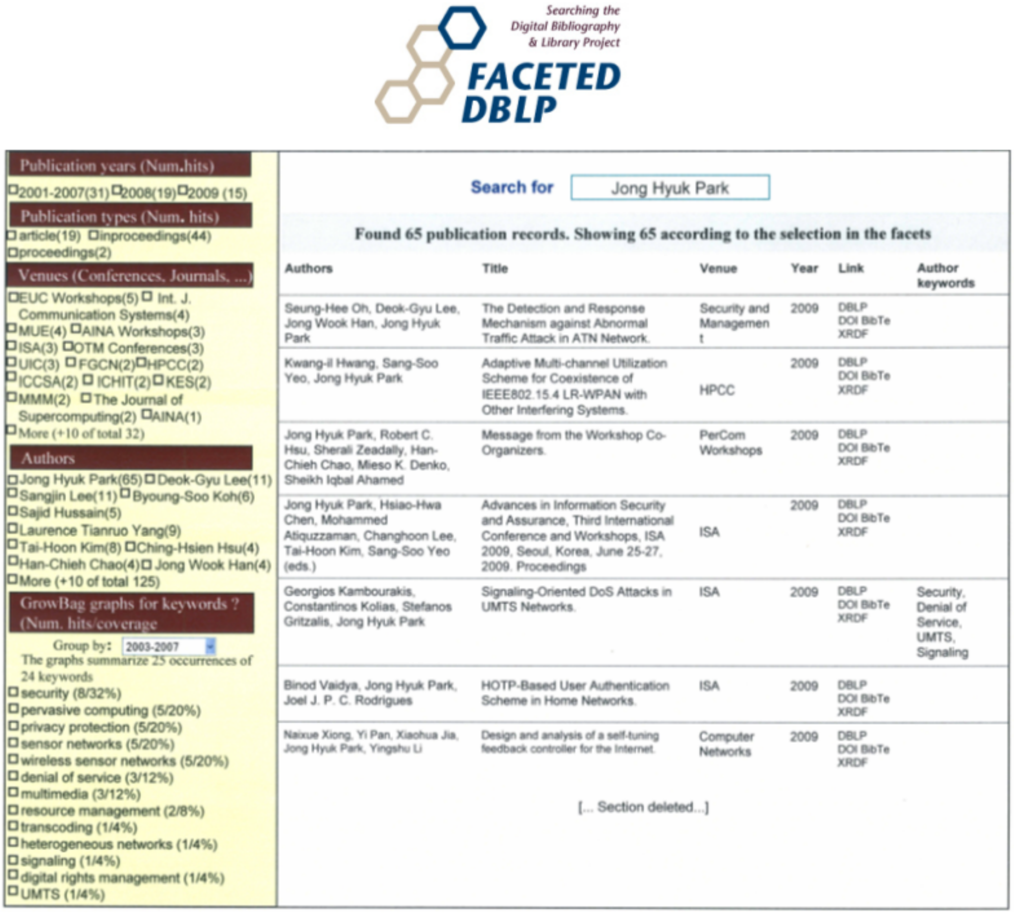
5.6 Concluding Remarks

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.
The main contributions of this chapter are:

1. A methodology for automatically identifying potential experts from assembled profiles.

2. A combined visualization of a topical classification hierarchy and a ranked list of potential experts at each level in the hierarchy.

3. A visualization of experts for users of J. UCS which is further linked with expert's profiles within J. UCS and in FacetedDBLP.



Publication years (Num.hits)

□2001-2007(31) □2008(19) □2009 (15)

Publication types (Num. hits)

□ article(19) □inproceedings(44)
□proceedings(2)

Venues (Conferences, Journals, ...)

□EUC Workshops(5) □ Int. J. Communication Systems(4)
□MUE(4) □AINA Workshops(3)
□ISA(3) □OTM Conferences(3)
□UIC(3) □ FGCN(2) □HPCC(2)
□ICCSA(2) □ ICHIT(2) □KES(2)
□MMM(2) □ The Journal of Supercomputing(2) □AINA(1)
□More (+10 of total 32)

Authors

□Jong Hyuk Park(65) □ Deok-Gyu Lee(11)
□Sangjin Lee(11) □ Byoung-Soo Koh(6)
□Sajid Hussain(5)
□Laurence Tianruo Yang(9)
□Tai-Hoon Kim(8) □Ching-Hsien Hsu(4)
□Han-Chieh Chao(4) □ Jong Wook Han(4)
□More (+10 of total 125)

GrowBag graphs for keywords ? (Num. hits/coverage

Group by: 2003-2007

The graphs summarize 25 occurrences of 24 keywords

□ security (8/32%)
□ pervasive computing (5/20%)
□ privacy protection (5/20%)
□ sensor networks (5/20%)
□ wireless sensor networks (5/20%)
□ denial of service (3/12%)
□ multimedia (3/12%)
□ resource management (2/8%)
□ transcoding (1/4%)
□ heterogeneous networks (1/4%)
□ signaling (1/4%)
□ digital rights management (1/4%)
□ UMTS (1/4%)

Search for: Jong Hyuk Park

Found 65 publication records. Showing 65 according to the selection in the facets

Authors	Title	Venue	Year	Link	Author keywords
Seung-Hee Oh, Deok-Gyu Lee, Jong Wook Han, Jong Hyuk Park	The Detection and Response Mechanism against Abnormal Traffic Attack in ATN Network.	Security and Managemen t	2009	DBLP DOI BibTe XRDF	
Kwang-il Hwang, Sang-Soo Yeo, Jong Hyuk Park	Adaptive Multi-channel Utilization Scheme for Coexistence of IEEE802.15.4 LR-WPAN with Other Interfering Systems.	HPCC	2009	DBLP DOI BibTe XRDF	
Jong Hyuk Park, Robert C. Hsu, Sherali Zeadally, Han-Chieh Chao, Mieso K. Denko, Sheikh Iqbal Ahamed	Message from the Workshop Co-Organizers.	PerCom Workshops	2009	DBLP DOI BibTe XRDF	
Jong Hyuk Park, Hsiao-Hwa Chen, Mohammed Atiquzzaman, Changhoon Lee, Tai-Hoon Kim, Sang-Soo Yeo (eds.)	Advances in Information Security and Assurance, Third International Conference and Workshops, ISA 2009, Seoul, Korea, June 25-27, 2009. Proceedings	ISA	2009	DBLP DOI BibTe XRDF	
Georgios Kambourakis, Constantinos Kolias, Stefanos Gritzalis, Jong Hyuk Park	Signaling-Oriented DoS Attacks in UMTS Networks.	ISA	2009	DBLP DOI BibTe XRDF	Security, Denial of Service, UMTS, Signaling
Binod Vaidya, Jong Hyuk Park, Joel J. P. C. Rodrigues	HOTP-Based User Authentication Scheme in Home Networks.	ISA	2009	DBLP DOI BibTe XRDF	
Naixue Xiong, Yi Pan, Xiaohua Jia, Jong Hyuk Park, Yingshu Li	Design and analysis of a self-tuning feedback controller for the Internet.	Computer Networks	2009	DBLP DOI BibTe XRDF	

[... Section deleted...]

Figure 5.12: Adapted from FacetedDBLP

Chapter 6

Linking Digital Journals with Linked Data

This chapter[1] focuses on how can the Semantic Web add value to digital journals. As explained in chapter 2, Linked Data (LOD) is a big success of the Semantic Web. LOD enabled billions of RDF triples which show relationships between structured contents available on the Web. However, two problems were identified in chapter 2: 1) identifying intended resource URI from LOD, and 2) structuring and presenting information from LOD by hiding complex underlying semantic mechanics. This chapter discusses techniques to address these issues and address the following research questions:

RQ8. How can digital journals consume information from Linked Data resources?

RQ9. How can user be given only required and relevant information whenever they need it?

The research question 8 is further subdivided into the following more specific questions:

RQ8.1. How can intended resource URI be located?

RQ8.2 How can we structure and present information from LOD?

[1] The contents related to CAFSIAL research, mentioned in this chapter, came from [Latif et al 2009b] where author of the thesis contributed 20%.

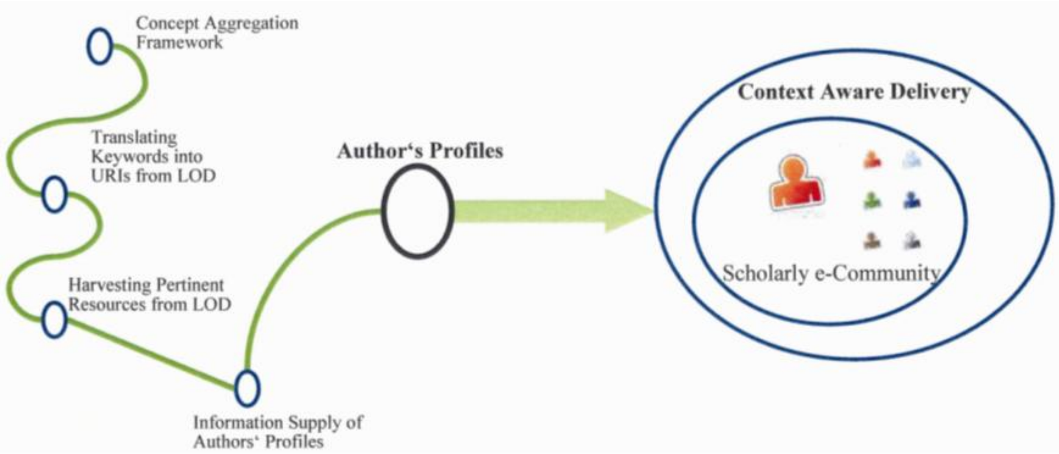

Concept Aggregation Framework
Translating Keywords into URIs from LOD
Harvesting Pertinent Resources from LOD
Information Supply of Authors' Profiles
Author's Profiles
Context Aware Delivery
Scholarly e-Community

Figure 6.1: Progress flow of the chapter based on published contributions

The figure 6.1, based on research contributions [Latif et al 2009][Latif et al 2009b][Latif et al 2009c][Afzal et al 2010b], shows the progress flow of the chapter. An intelligent URI mapping technique was developed. By locating resource using this technique, a concept aggregation framework was developed which is able to structure the information. Furthermore, this concept aggregation framework was used for J.UCS dataset for linking J.UCS authors to their profiles available in LOD.

6.1 Introduction

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 Data[2] (LOD) movement, a community effort that motivates people to publish their information in a structured way (RDF)[3]. 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 [Berner-Lee 2006]. As of May 2009, the LOD cloud consists of over 4.7 billion RDF triples, which are interlinked by around 142 million RDF/OWL links [Auer et al 2009]. Although LOD has created huge

[2] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[3] http://www.w3.org/RDF/

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 2009a].

In the absence of official standards, DBpedia[4] and Yago[5], 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 infoboxes 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.

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 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). The same gap between semantic search and end user applications has also been identified by [Chakrabarti 2004].

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, regular Web users have never even heard of URIs or SPARQL. Therefore, when non-expert users interact with the Semantic Web, the first step is to translate their queries into URIs. For example, when a user wants to know something about "Arnold Schwar-zenegger", it is necessary to find a URI that represents this person in the Semantic Web e.g. http://dbpedia.org/resource/Arnold$_{s}chwarzenegger$.

To overcome the URI discovery, an intelligent Keyword-URI mapping technique has been introduced. 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

[4] http://dbpedia.org/

[5] http://www.mpi-inf.mpg.de/yago-naga/yago/

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). The exploratory evaluations have shown that the system can reduce user's cognitive load in finding required URIs.

When the system has identified a correct resource URI, then it 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. This property set is pre-computed for each resource type. This approach conceptualizes the most relevant information of a resource in an easily perceivable construct.

We also 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. 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.

The state-of-the-art of LOD in different perspectives has already been described in chapter 2. The remainder of this chapter is structured as follows: Section 6.2 presents an intelligent Keyword-URI technique. Section 6.3 elaborates on the Concept Aggregation Framework. Section 6.4 describes the system architecture. Section 6.5 explains the overall use of the system with the help of a use case. Section 6.6 elaborates the process of linking profiles of J.UCS authors with LOD. The summary of the chapter is presented thereafter.

6.2 Keyword-URI Mapping Technique

The design of the Keyword-URI mapping is depicted in figure 6.2. The proposed technique is divided into three parts called Triple Construction, Auto-Suggestion, and Semantic Search Service. The triple construction technique discusses the data acquisition and the process of converting it to triples. The auto-suggestion technique discusses how the suggestions are derived from the local data store and highlights the added value of providing seamless URI mapping. In the semantic search service, the querying and filtering of retrieved results is discussed.

6.2.1 Triple Store Construction

The DBpedia data is maintained locally for guaranteed response and to avoid the negative consequences of a sudden downtime of DBpedia. The "Persondata" dump[6] was downloaded from DBpedia. This dump contains information about

[6] http://downloads.dbpedia.org/3.3/en/persondata$_en.nt.bz2$

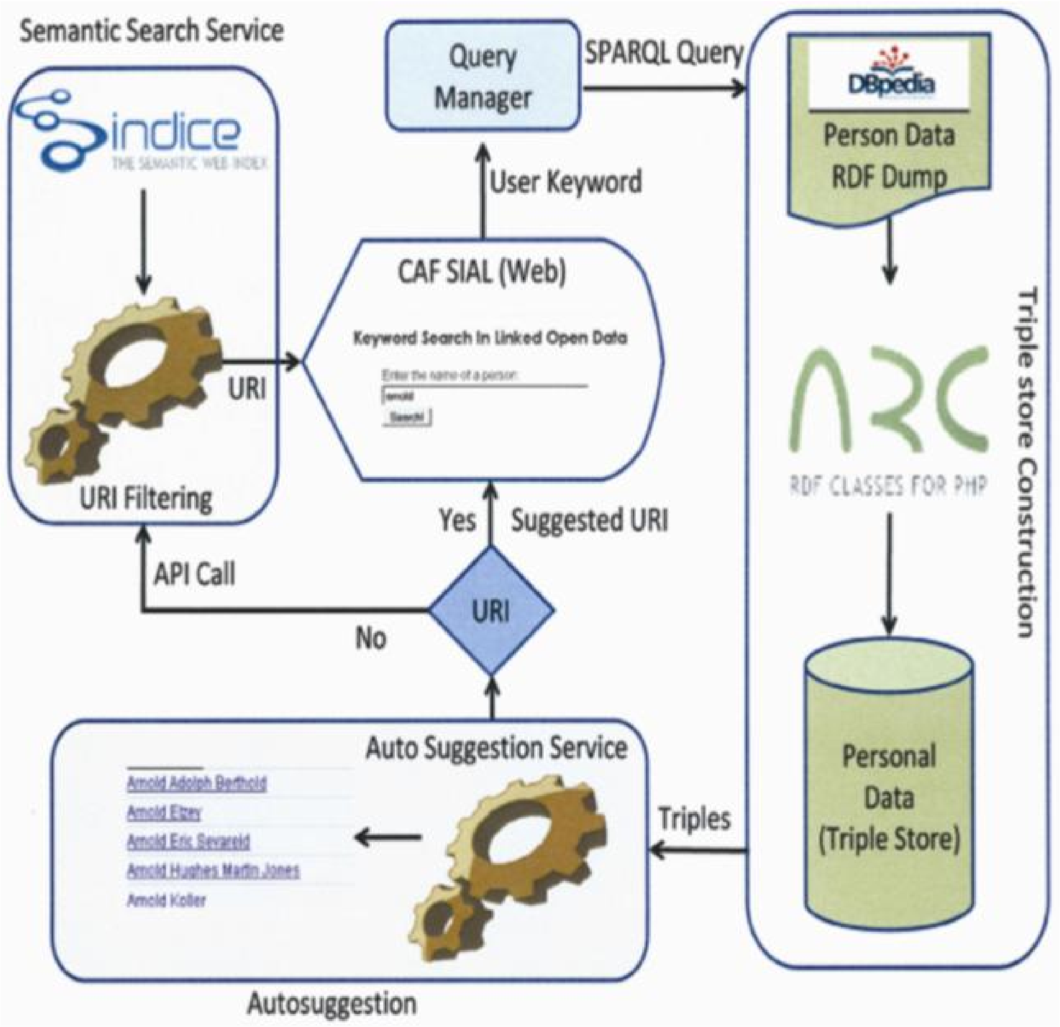


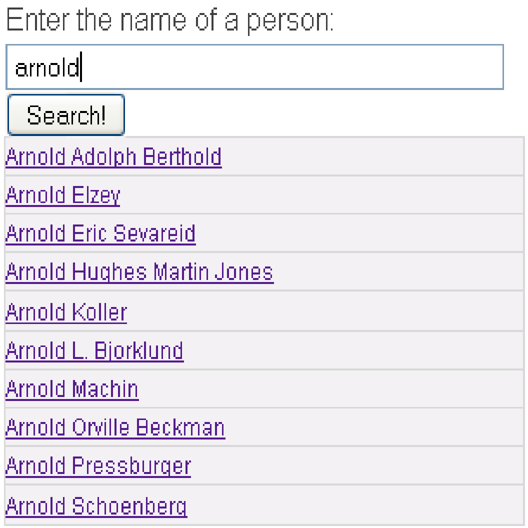
Figure 6.2: Keyword-URI mapping design

persons extracted from the English and German Wikipedia, represented using the FOAF vocabulary ARC[7], a flexible system for RDF data, is then used to import this dump into a local triple store. This triple store provides an interface for the SPARQL queries. At the moment, there are 62,313 URIs of persons stored in the CAF-SIAL triple store.

6.2.2 Auto-Suggestion

When a user starts entering a keyword for a search, a SPARQL query is constructed on the fly on every key press and an AJAX-enabled autosuggestion module is invoked. The autosuggestion module is responsible for finding all possible

[7] http://arc.semsol.org/



Enter the name of a person: arnold

Arnold Adolph Berthold
Arnold Elzey
Arnold Eric Sevareid
Arnold Hughes Martin Jones
Arnold Koller
Arnold L. Bjorklund
Arnold Machin
Arnold Orville Beckman
Arnold Pressburger
Arnold Schoenberg

Figure 6.3: Auto-suggestions

occurrences of an entered person name in the local triple store and returning a list of suggestions. These suggestions help users in the following aspects:

- With auto-complete, users need to type less
- Give user leverage about searching possibilities within dataset
- On-the-fly disambiguation of concepts having similar or the same names
- Selecting the correct concept

On user selection from any of the suggested option, the underlying URI of the selected keyword is passed on for further processing. The presentation of the list of suggestions is shown in figure 6.3.

6.2.3 Semantic Search Service

In case the keyword is not mapped to any concept in the local triple store, the semantic search service is invoked. The public API of Sindice is used for this operation. It returns an RDF file containing number of URIs belonging to different data sources. This file is then parsed into the local triple store by using ARC. Further on, a URI filtering service is called, and the URI matching our set description, i.e. a DBpedia person type resource, is filtered out. If more than one URI belonging to DBpedia person type exist, the first one in the list is picked.

The SINDICE service provides a faster crawling procedure as compared to other semantic search engines. DBpedia releases new data dumps approximately every six months. Hence the newly entered or updated resources will not be part

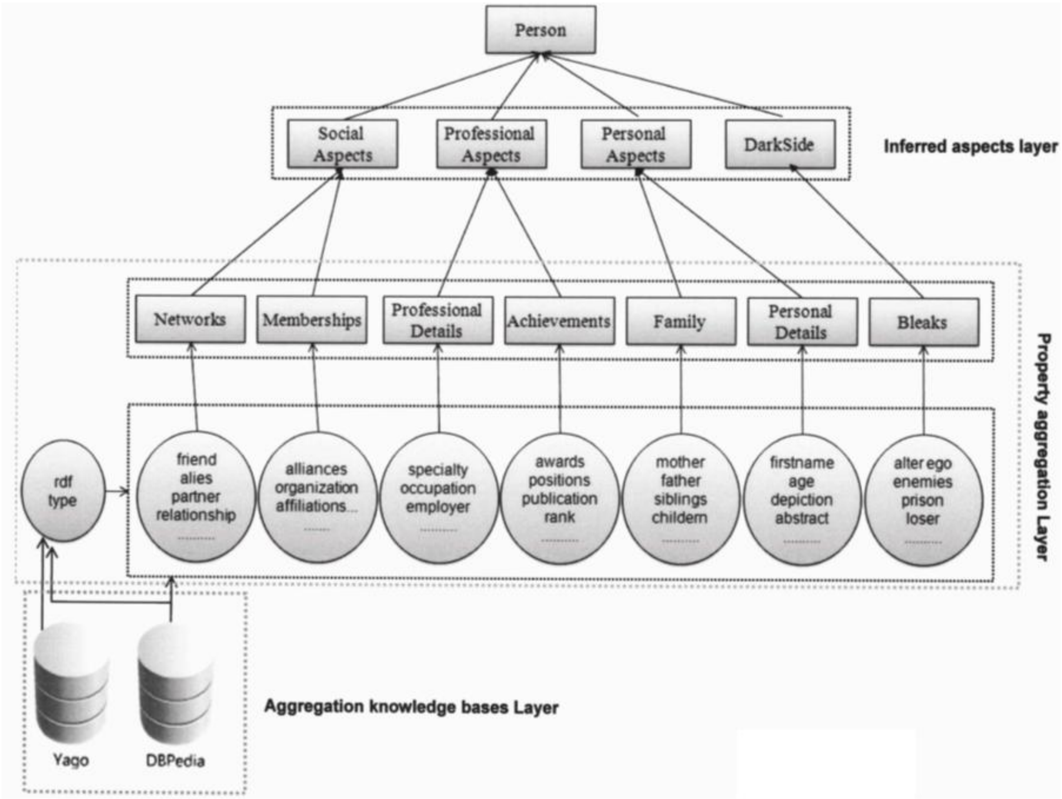

Person
Social Aspects
Professional Aspects
Personal Aspects
DarkSide
Inferred aspects layer
Networks
Memberships
Professional Details
Achievements
Family
Personal Details
Bleaks
Property aggregation Layer
rdf type
friend alies partner relationship
alliances organization affiliations...
specialty occupation employer
awards positions publication rank
mother father siblings childern
firstname age depiction abstract
alter ego enemies prison loser
Aggregation knowledge bases Layer
Yago
DBPedia

Figure 6.4: Concept Aggregation Framework

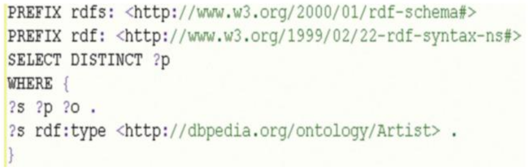
of the older DBpedia dump as well as our local triple store. Meanwhile Sindice, due to its fast crawling procedure, will be having an index of these newly added resources, which may be very useful to locate new resources. This will increase our system's performance and ensure up-to-date URI supply to users.

6.3 Concept Aggregation Framework

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

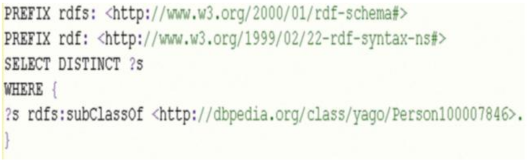
The scope of this application is limited to DBpedia and Yago. DBpedia covers 23 types of resources (places, people, organizations, etc), initially, we selected the resource type person for the experimentations.

The Concept Aggregation Framework is shown in figure 6.4. 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.



```
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> .
}
```

Figure 6.5: Building DBPedia property dump



```
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>.
}
```

Figure 6.6: Building Yago classification dump

6.3.1 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 explorer[8] (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. The formulated query for this operation is given in figure 6.5.

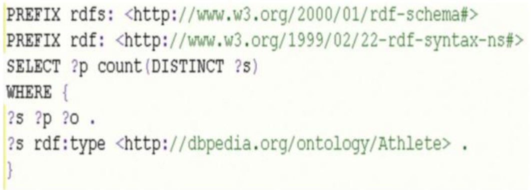
The Yago Classification Dump is built by querying subclasses of Person class from SNORQL query explorer. The query is shown in figure 6.6.

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. 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 to get the count of each distinct property is shown in figure 6.7.

[8] http://dbpedia.org/snorql/

Table 6.1: Selection of persons' properties from DBPedia

Person Type	Total Properties	Picked Properties
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



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

Figure 6.7: Computing property rank

6.3.2 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 concept defined at the inferred aspects layer.

6.3.3 Inferred Aspects Layer

The information for a resource such as person may be organized and viewed in different informational aspects like personal, professional, social etc. 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.

6.4 System Architecture

The system architecture is depicted in figure 6.8. 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.

6.4.1 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 6.3. If the user does not select any of the suggested terms, or in case

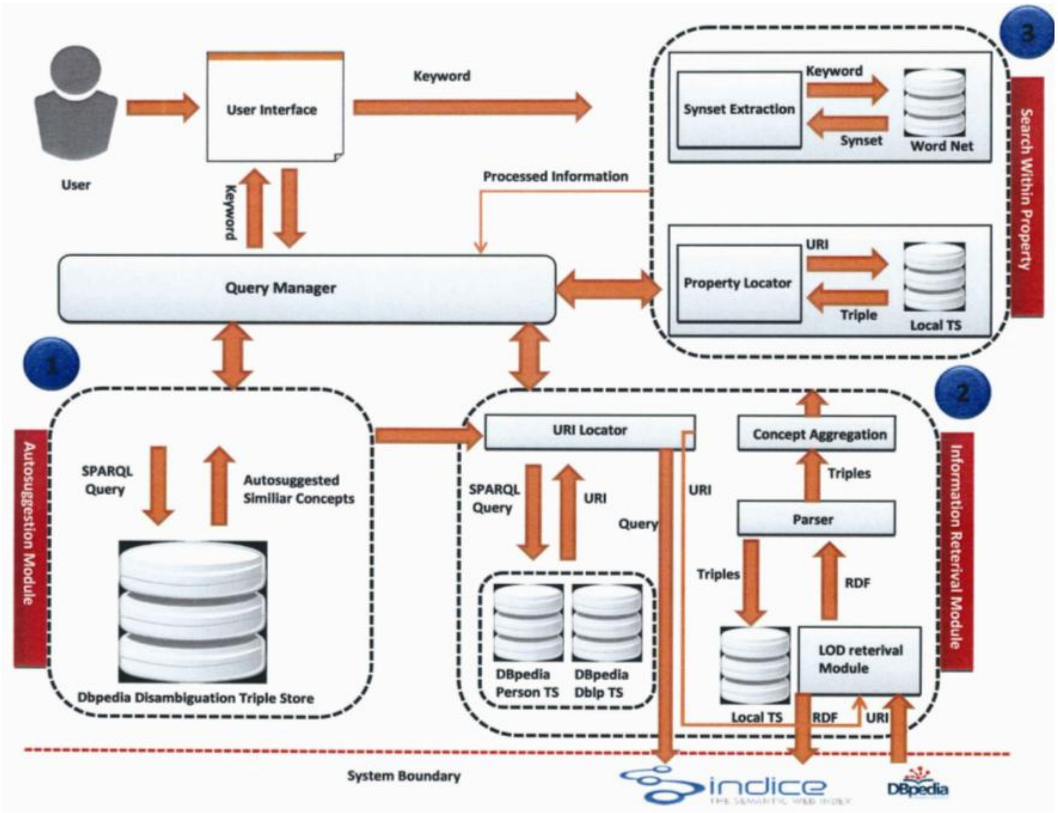


Figure 6.8: System architecture for CAFSIAL

of a distinct query (no auto-suggestions yielded), the searched term is passed on to the information retrieval module for further processing.

6.4.2 Information Retrieval Module

This module is further divided into four processes:

1. URI locator
2. LOD retrieval
3. Parser
4. Concept aggregation

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 SINDICE[9] Web service to locate the URI. After locating the URI of a resource,

[9] http://sindice.com/

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.

6.4.3 Search Within Property Module

This module lets the user search within all properties of a resource retrieved from the information retrieval module. When a user enters a keyword to search some information about a resource, the synset extraction process queries wordnet[10] to retrieve the synset of searched keyword. 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.

6.5 Use Case Scenario

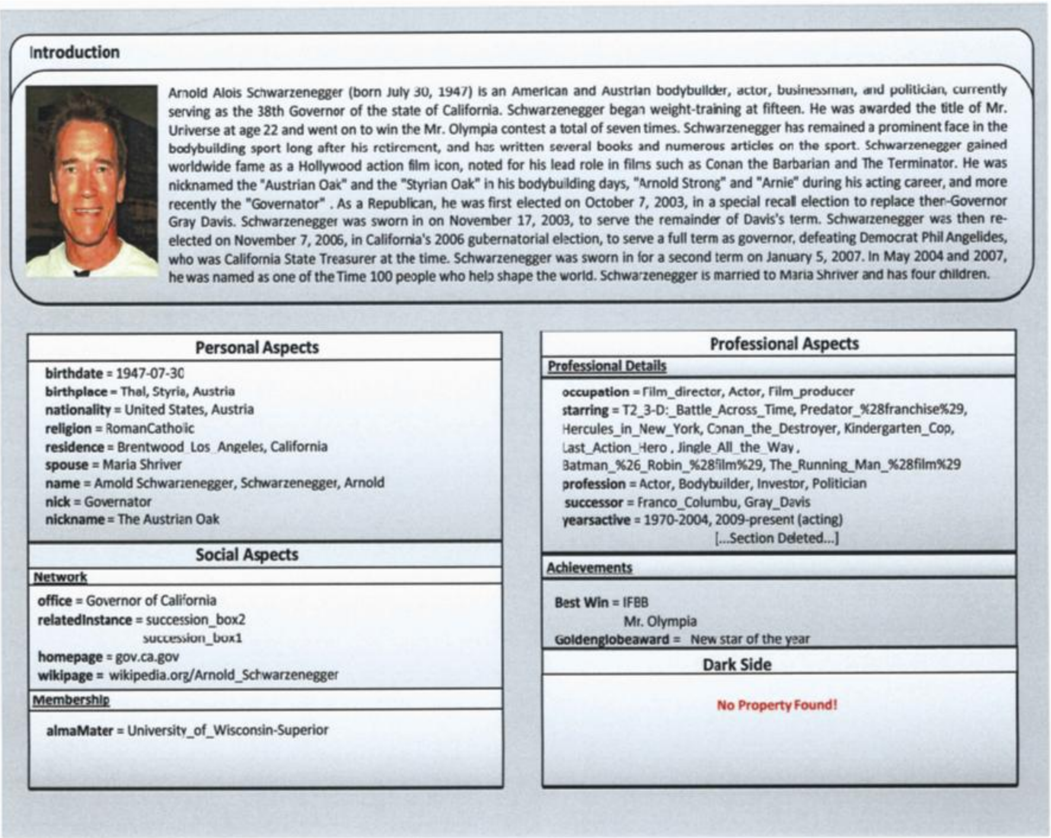
The working of the system is described with the help of a use case scenario. We have selected *"Arnold Schwarzenegger"* for this example. The reason of this selection is that the selected person is affiliated with four interesting and diverse professions along with multiple awards and achievements. This will help in understanding the overall working of the system.

These capabilities make him a distinct person and a suitable choice for the use case. The application flow is explained as follows: User starts typing the search term *"Arnold"*. The persons' names starting with the keyword *"Arnold"* are auto-suggested. For example *"Arnold Bax"*, *"Arnold Bennett"*, *"Arnold Schwarzenegger"* etc. as depicted in figure 6.3.

The user selects *"Arnold Schwarzenegger"* to see his details as shown in figure 6.9. The output is comprised of different informational aspects such as social, personal, and professional. Important properties are shown on the top for each informational aspect. The important property list was prepared manually and the weight to each property is assigned on the basis of its count automatically.

The screenshot shows his important professional details concisely and in easily graspable manner.

[10] http://wordnet.princeton.edu/wordnet/



Introduction

Arnold Alois Schwarzenegger (born July 30, 1947) is an American and Austrian bodybuilder, actor, businessman, and politician, currently serving as the 38th Governor of the state of California. Schwarzenegger began weight-training at fifteen. He was awarded the title of Mr. Universe at age 22 and went on to win the Mr. Olympia contest a total of seven times. Schwarzenegger has remained a prominent face in the bodybuilding sport long after his retirement, and has written several books and numerous articles on the sport. Schwarzenegger gained worldwide fame as a Hollywood action film icon, noted for his lead role in films such as Conan the Barbarian and The Terminator. He was nicknamed the "Austrian Oak" and the "Styrian Oak" in his bodybuilding days, "Arnold Strong" and "Arnie" during his acting career, and more recently the "Governator" . As a Republican, he was first elected on October 7, 2003, in a special recall election to replace ther-Governor Gray Davis. Schwarzenegger was sworn in on November 17, 2003, to serve the remainder of Davis's term. Schwarzenegger was then re-elected on November 7, 2006, in California's 2006 gubernatorial election, to serve a full term as governor, defeating Democrat Phil Angelides, who was California State Treasurer at the time. Schwarzenegger was sworn in for a second term on January 5, 2007. In May 2004 and 2007, he was named as one of the Time 100 people who help shape the world. Schwarzenegger is married to Maria Shriver and has four children.

Personal Aspects

birthdate = 1947-07-30
birthplace = Thal, Styria, Austria
nationality = United States, Austria
religion = RomanCatholic
residence = Brentwood Los Angeles, California
spouse = Maria Shriver
name = Arnold Schwarzenegger, Schwarzenegger, Arnold
nick = Governator
nickname = The Austrian Oak

Social Aspects

Network

office = Governor of California
relatedInstance = succession_box2
succession_box1
homepage = gov.ca.gov
wikipage = wikipedia.org/Arnold_Schwarzenegger

Membership

almaMater = University_of_Wisconsin-Superior

Professional Aspects

Professional Details

occupation = Film_director, Actor, Film_producer
starring = T2_3-D:_Battle_Across_Time, Predator_%28franchise%29, Hercules_in_New_York, Conan_the_Destroyer, Kindergarten_Cop, Last_Action_Hero , Jingle_All_the_Way , Batman_%26_Robin_%28film%29, The_Running_Man_%28film%29
profession = Actor, Bodybuilder, Investor, Politician
successor = Franco_Columbu, Gray_Davis
yearsactive = 1970-2004, 2009-present (acting)
[...Section Deleted...]

Achievements

Best Win = IFBB
Mr. Olympia
Goldenglobeaward = New star of the year

Dark Side

No Property Found!

Figure 6.9: Informational aspects of Arnold Schwarzenegger

6.6 Application of CAF-SIAL for J.UCS Dataset

We have applied the above defined Concept Aggregation Framework for authors of J.UCS. The authors are located in LOD dataset and then authors' informational aspects are retrieved, aggregated, structured and presented nicely to users. The following sections provide details of this process.

6.6.1 JUCS-LOD system Architecture

The architecture design of the JUCS-CAFSIAL application is depicted in figure 6.10. The proposed system is divided into four modules named as Database and Triple Store Construction, URI Acquisition, Author URI Validation and Concept Aggregation Framework.

The database and triple construction part discusses the data acquisition, manipulation of J.UCS data set and the process of converting RDF personal dataset into local triple store. The URI acquisition module describes that how the URI of a J.UCS author is acquired from local triple stores and by remote semantic search services. Author's URI Validation module encompasses the heuristic writ-

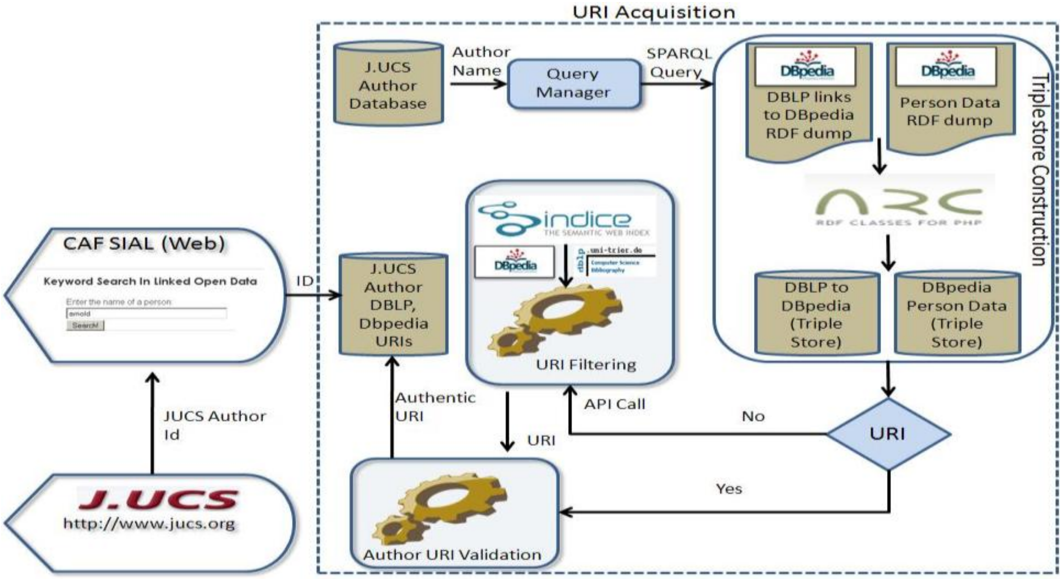


Figure 6.10: System architecture for CAF-SIAL-JUCS

ten to validate the URI. In the last section, Concept Aggregation Framework is presented.

(A) Database and Triple Store Construction

Three datasets were used, one from the Journal of Universal Computer Science (JUCS) and two from the DBpedia. Along with these datasets, web service of Sindice (a semantic search engine) was also utilized when the local search fails.

Description of each data set is given below.

J.UCS Dataset

The J.UCS dataset provides the list of the authors who have published their work in any of the Journal Issues. Author ID [Afzal et al 2007] maintained at JUCS server along with first, middle and last name of the respective author is tabulated in this dataset. In total 2593 authors from JUCS were used for this experiment.

Semantic Datasets

DBpedia

DBpedia is currently one of the most promising knowledge bases in LOD, 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 strong 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 includes a data set about persons.

Two RDF dumps about personal information (Persondata and Links to DBLP) were selected to find relevant information of J.UCS author. These dataset are freely available in RDF dumps for download. These RDF dump were converted into a local triple store by using ARC2. ARC2 utility star triple store configuration gives a facility for querying at statement level. The details of these semantic datasets are given below:

Persondata

This data set includes information about persons (date and place of birth etc.) extracted from the English and German Wikipedia, represented using the FOAF vocabulary. At the moment 29,498 URIs of persons stored in local triple store were matched to find respective URI of a JUCS author.

(B) URI Acquisition

Different techniques were piled up to locate the URI from local triple store and the Linked Data cloud. These techniques are listed below:

- Direct matching of JUCS authors with DBpedia Persondata dataset
- Direct matching of JUCS authors with Links to DBLP dataset
- Querying and Filtering of URI from Sindice

The details of these techniques can be found in implementation section.

(C) URI Validation

For making sure the authenticity of a DBpedia URI attained after URI acquisition process, set of heuristics were written by manual inspection of various auhtor's Dbpedia profiles. The detail of these heuristics can be seen in implementation section.

(D) Concept Aggregation

After the processes of URI acquisition and URI validation, the aggregation of important aspects is performed using Concept Aggregation Framework as discussed earlier in this chapter. The authors are represented in four informational aspects such as: 1) personal information, 2) professional information, 3) academics, and 4) published work. Presentation of these aspects is depicted in figure 6.11.

6.6.2 Implementation

The implementation of this application can be divided into two phases.
(A) Locating Author DBpedia URI
(B) Concept Aggregation and Presentation

(A) Locating Author DBpedia URI

In this implementation phase different steps were performed to find a URI of the respective JUCS author in DBpedia.

Pre-processing of Author's Dataset

Sometimes, the authors' names contain umlaute characters which need to be processed before matching them in LOD cloud. An automated scripted was written to remove such inconsistencies. Subsequently First, middle, and last names were concatenated to construct a full name for the matching.

DBpedia Person Data Direct Matching with JUCS Author Dataset

In the first step, complete name of the authors were matched with the DBpedia Persondata triple store. After this operation seven out of total authors were matched yielding in very low success rate.

DBPedia Links to DBLP Direct Matching

In second step Links to DBLP local triple store were matched for authors name resulting eight out of total JUCS authors.

Due to inconsistencies in local triple store and the outdate DBpedia dataset the result were not so attractive. To overcome these limitations, Sindice (Semantic Search Service), a vast and up-to-date indexing system, was used.

Sindice Search Service

A web service was written to call the API of Sindice with the formulated query. The process runs iteratively for every unfound JUCS author. In response, Sindice gives the list of the URI's which were further filtered out on the basis of DBpedia provenance. In the end direct matching on the name of author in the DBpedia URI list was performed to pick the exact URI of a resource.

After processing 337 DBPedia URI's out of entire J.UCS author list, a substantial improvement was noted. All this findings were stored in data table for further processing.

DBpedia URI Authentication

To find author's names from the retrieved results, a script was executed. However, by the manual inspection, it was found that there are some inconsistencies in the retrieved URIs:

- The URI of the author page exist (wrongly indexed by Sindice) but no information is present in the page, making this URI useless for processing.
- Lot of URI's which matched with exact name of the author but representing persons who are not associated with educations giving rise to ambiguities.

To disambiguate authors, a set of heuristics were written as described below:

Heuristics Construction

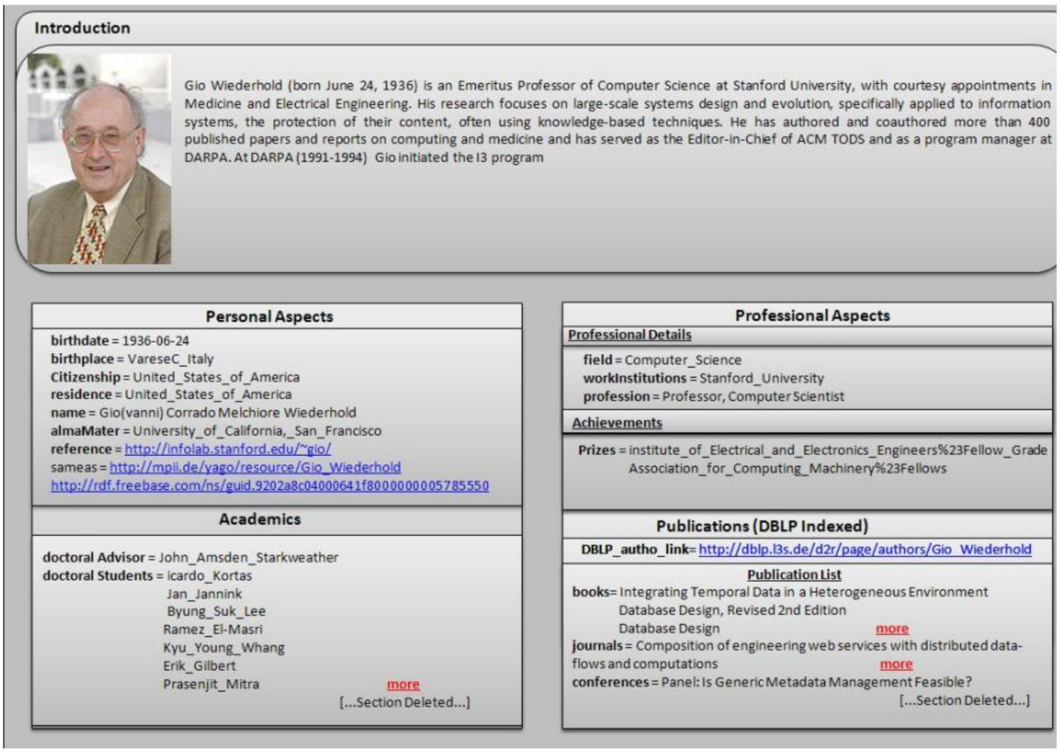
After manual inspection it was noted that there are certain kind of properties which can be exploited to disambiguate individuals. For example, SKOS categories and keyword which are being used to represent the persons belonging to education profession. Certain types of keywords are normally mentioned for educationist. An automated script was written to check the SKOS categories in the respective URI data. After applying this script on 337 authors, 66 URIs were selected.

These finding also highlight that there exists many profiles which are represented with the same name as of authors but representing other persons. Concrete steps needs to be taken by Dbpedia in assigning URI's to the persons having similar name. At the moment in DBpedia a Unique Id is assigned to person basis of his name which at certain point is good but if there are persons having similar name, then it becomes difficult to identify and disambiguate. Assigning a URI on the basis of profession can be a alternative step.

(B) Concept Aggregation and Presentation

The concept aggregation works in the same fashion as described earlier in the chapter. Here we show the strength of the system by using an example.

For example, a user queries on "Gio Widerhold", the Concept Aggregation Framework makes a conceptual representation of the available properties into different aspects of a person as shown in figure 6.11. From the Figure, it is obvious that a user will get instant information about the concerned person. A brief introduction of a person along with the picture is shown. Furthermore different



Introduction

Gio Wiederhold (born June 24, 1936) is an Emeritus Professor of Computer Science at Stanford University, with courtesy appointments in Medicine and Electrical Engineering. His research focuses on large-scale systems design and evolution, specifically applied to information systems, the protection of their content, often using knowledge-based techniques. He has authored and coauthored more than 400 published papers and reports on computing and medicine and has served as the Editor-in-Chief of ACM TODS and as a program manager at DARPA. At DARPA (1991-1994) Gio initiated the I3 program

Personal Aspects

birthdate = 1936-06-24
birthplace = VareseC_Italy
Citizenship = United_States_of_America
residence = United_States_of_America
name = Gio(vanni) Corrado Melchiore Wiederhold
almaMater = University_of_California,_San_Francisco
reference = http://infolab.stanford.edu/~gio/
sameas = http://mpii.de/yago/resource/Gio_Wiederhold
http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000005785550

Academics

doctoral Advisor = John_Amsden_Starkweather
doctoral Students = icardo_Kortas
Jan_Jannink
Byung_Suk_Lee
Ramez_El-Masri
Kyu_Young_Whang
Erik_Gilbert
Prasenjit_Mitra more
[...Section Deleted...]

Professional Aspects

Professional Details

field = Computer_Science
workInstitutions = Stanford_University
profession = Professor, Computer Scientist

Achievements

Prizes = institute_of_Electrical_and_Electronics_Engineers%23Fellow_Grade
Association_for_Computing_Machinery%23Fellows

Publications (DBLP Indexed)

DBLP_autho_link= http://dblp.l3s.de/d2r/page/authors/Gio_Wiederhold

Publication List

books= Integrating Temporal Data in a Heterogeneous Environment
Database Design, Revised 2nd Edition
Database Design more
journals = Composition of engineering web services with distributed data-flows and computations more
conferences = Panel: Is Generic Metadata Management Feasible?
[...Section Deleted...]

Figure 6.11: Informational aspects of a J.UCS author

informational aspects like personal, professional, academics, and publications are shown to the user. From this coherent view, user gets a first overall impression about a person and can follow any hyperlink to see further details.

Chapter 7

Summary and Outlook

This brief chapter elaborates the overview of the work, discuss key results and concludes the thesis. The future possible extensions to the work are also pointed out.

7.1 System Evaluations

The first version of Links into the Future system was made available to users of J.UCS in 2007. Since then the system is up and running and have been upgraded by adding expertise finding and linking the journal with socially maintained digital libraries. The user feedbacks show that the system is very useful in: knowledge discovery of research papers, suggesting experts in an area and leading to serendipitous discoveries of evolving concepts in social digital libraries. However, the system for linking J.UCS to semantic resources (Linked Data) is still in prototype use. The link between J.UCS and Linked Data resources is established based on the past research of CAFSIAL, the next section discusses the evaluation of the CAFSIAL system by comparing it with contemporary systems.

The system was evaluated with the help of user interviews. We collected data with the help of combining focus groups [Kitzinger 1995] and post-search interviews. We held two focus group sessions having 4-6 participants each (10 participants in total). All participants of both groups were Web users and had knowledge of basic Web search. One focus group comprised of users having experience in Semantic technologies while the other user group was naive Web users. Each group session was conducted by a skilled representative.

The selected application for evaluation in comparison with CAF-SIAL were Marble, Snorql, and Freebase. We conducted semi-structured interviews. Users were asked about comments, feedbacks, overall satisfaction, problems faced etc.

The interviews showed that it was very difficult to get the sought-after information from Marble and Snorql, because the users did not know the exact URI of the resource, and due to the difficulty of formulating Subject Predicate Object logic. On the other hand, Freebase and CAF-SIAL systems were easier to use. Although Freebase was comparatively better in terms of providing rich information and content organization, CAF-SIAL was useful most of the time to get the sought-after information. One negative aspect that was mentioned about CAF-SIAL was the fact that when the users searched for a person and then clicked on a particular property within the retrieved result set, they were again redirected to complex systems like DBpedia. Then again it became difficult to find required information from a long list of properties.

When comparing CAF-SIAL to Freebase, there are some noteworthy differences:

1. Dbpedia, which provides that basis for CAF-SIAL, is built around a controlled vocabulary (an ontology, actually), whereas freebase adopts the folksonomy approach in which people can add new categories much like tags (O'Reilly 2007).

2. Along with the semi-automatic approach of Freebase to collect and organize data in to their knowledge base, a group of editors is responsible to pre-check the organization and add new knowledge in a structured way manually. On the other hand, CAF-SIAL works on a set of heuristics. These heuristics were defined manually once by experts and can then be applied by the system to organize knowledge in an entirely automated way. The system makes a transition to exploit LOD resources in an autonomous way, which could provide significant help in navigating the ever-growing LOD cloud.

The system has been made publicly available at http://cafsial.opendatahub.org/. For continuous evaluations, users can give feedback at any time online. The submitted information is saved and we plan to extend the system by incorporating users' comments and feedback.

7.2 Conclusions and Future work

The dissertation explores different possibilities of linking resources from a digital journal to other relevant resources that are available locally or in various external repositories. The information is presented to user in context-aware environment. The work shows its importance in emerging information supply systems for digital journals. The system provides the most relevant resources to the users by observing their current focus. In this manner, users do not need to make deliberate efforts for searching contents and the information is made available in timely fashion. The resource discovery is based on new techniques, mainly of heuristic nature. The proposed techniques often outperform existing ones and are able to discover highly relevant resources more effectively. The thesis contributions can be structured into four areas such as: 1) Links into the Future, 2) linking digital journals with social bookmarking systems, 3) discovery and visualization of expertise, and 4) linking digital journals with semantic resources. The concluding remarks and future extensions for each area are listed below.

7.2.1 Links into the Future

The Links into the Future system recommends the most relevant resources to users, based on their local context. The developed techniques can discover relevant papers that might /would not otherwise be viewed. Two different approaches were used to find Links into the Future: a) Metadata extraction technique, b) Citation mining technique. The metadata extraction technique uses available metadata to create Links into the Future. The results are discussed in details in chapter 3. The employed heuristics and techniques were able to discover most relevant papers. For example, a search query on a generic search engine retrieves millions of generic hits for users, while the proposed techniques filters noisy and irrelevant papers and at the end, the user is left with only few links to the most relevant papers.

The developed citation mining technique can find citations that were missed by prevailing citation indexes. The technique uses a generic heuristic approach and works in a two-tier process, first disambiguate venues (journals, conferences, etc.) and then find the intended citation. This two-tier process helped the system to reduce the chances of errors significantly. Comparing a citation entry within a large citation index leads to wrong citation identification if authors make any mistake while citing a paper. However, first disambiguating venues and then focusing on only the papers published within the venue, gives a leverage for performing direct and partial match and lead to a high accuracy. This has been shown in details in chapter 3. This research not only develops a generic citation mining technique but also develop a better venue disambiguation technique. The experiments show that the technique was able to overcome limitations of existing citation indexes like Google Scholar, CiteSeer, and ISI Web of Knowledge. The technique was tested on two datasets such as 1) a data set from a digital journal and 2) a generic dataset provided by Cora. One interesting finding was that the system did not find any false positive citation. The process of finding citations is innovative and can be easily adopted by autonomous citation mining indexes.

In the future, we plan to extend the implementation of Links into the Future system for discovering papers from sources like DBLP and CiteSeer. As mentioned in chapter 3, the papers residing at external sources normally do not contain papers' categories, hence it would be a challenging task to develop such a system which can first find papers' topic and then discover Links into the Future. The CiteSeer provides access to full text search while in DBLP the full text is not available, only the metadata can be acquired and sometimes the abstract can be found. Therefore, the developed heuristics for the Web documents may work for CiteSeer but for DBLP, one need a better categorization system which may work fine by analyzing the available metadata and building a better categorization system.

We also envision the sentiment analysis of citations to discover the context of **"cited-by"** papers. In the Links into the Future system, this will be helpful to rank papers based on positive and negative sentiments. This will also help to filter the cited papers that cite only for providing the background studies rather than extending the research of cited papers. The cited papers with positive sentiments and with negative comments about the focused paper would then be easily filtered and ranked for the users.

7.2.2 Linking Digital Journals with Social Bookmarking

The next contribution of this dissertation is linking papers of a digital journal with resources available in social bookmarking. Initially, an exploratory case study was conducted to find the importance of tags in a scientific domain. It has been proved that there exist a positive correlation between the tags of a paper and its citations. This shows the involvement and trend of social community in bookmarking services for annotating the available resources. Furthermore, the tag terms reoccur in the titles of the citing papers. This shows that the papers get popularity in tagging applications immediately and the specified tags are used in future to write papers which normally cite the tagged paper. The rank predication based on co-author model can be further enhanced with the help of models based on tags analysis. All of these exercises show that the tags have a potential to measure the research popularity and are used in different ways, sometimes to represent the content and sometimes to represent the context of resources. Therefore, it would be very helpful to create a link between resources in a digital journal with the relevant resources in tagging applications. The dataset of a well known scientific social bookmarking system -such as CiteULike- was selected to be linked with journal's papers. The author's keywords were matched with the tags available in CiteULike leading the serendipitous discoveries of same or similar resources. The system was also able to find newly evolved concept from social bookmarking. For example in the case of "wiki" as an author's keyword, the system was able to find tags like "wikification, semantic wiki, citeulikewikis etc". By following any of the links, the user can view recent papers and other resources that were tagged and marked by the community. In this way, users can find the current trends related to the focused content and may discover newly evolved fields belonging to the focus of the current paper.

In the future, more analysis can be made on the CiteULike rescuers. One direction could be to find tags of the focused paper from CiteULike and compare them with authors' keywords. This will rank relevant resources from CiteUlike which belongs more closely to the content of the focused paper. There is a need to have a system which can distinguish tag terms representing content of the resource, context of the user, and future context of use. Other social bookmarking systems like delicious and Bibsonomy would also be a good dataset for concept and resource discoveries.

7.2.3 Discovery and Visualization of Expertise

The discovery of expertise is another area of contribution of the dissertation. In automatic approaches, the experts are usually discovered with the help of one metric. For example in academics, the experts are often characterized based on the number of publication alone. However, in the past, an effort was made to rank experts by manipulating the impact factor of venues along with the number of publications. The impact factor is in itself questionable and debatable as shown in chapter 5. To overcome the limitations, a multi-faceted approach was used which incorporates multiple evidences of expertise to construct an expert profile. In contrast of using the impact factor of venues, the paper's citations are measured which will effectively infer the expertise level of an individual. Overall, four metrics (weights) were used: 1) Number of publication, 2) citations received, 3) Experience of a person, and 4) Reviewing authorities and responsibilities. Based on the combination of all of these weights, the experts are discovered and ranked. This has reduced the limitations of existing systems and made it possible for users to find the most suitable and highly ranked experts easily. To visualize experts, an extended hyperbolic visualization technique was proposed and implemented. The ACM classification was shown on a hyperbolic tree while a ranked list of experts is shown along the selected node of the hyperbolic tree using spiral visualization. This visualization is useful especially for journal administration to assign new reviewers/editors. Furthermore, the experts associated with the topics of a paper are provided to users in timely fashion.

Although the current implementation of the expertise system is based on multiple experience record, the measures provided are, however, not absolute indicators of expertise as the discoveries are limited by the coverage of database of publications and expert profiles used. Ranking someone within a digital journal is not enough. There is a need to expand the data set in order to include most of publications and citations of an individual. For this task CiteSeer and DBLP could serve as a good sources.

7.2.4 Linking Digital Journals with Semantic Resources

The next contribution was to link resources of a digital journal with semantic resources. For this task, one of the most successful projects of Semantic Web named as Linked Data was selected. By exploring this data set, it was found that although the resources are semantically rich, there exist some issues such as 1) the non-triviality of finding the intended resource URI, and 2) the non-existence of such autonomous system which can discover the structure, and present information to users of scholarly e-community. To find an intended URI from a huge repository of Linked Data, an intelligent technique was proposed and developed, and to deal with the second identified problem, a Concept Aggregation Framework was developed. The latter framework aggregates, structures, and presents most relevant aspects of a person at one place. The details can be seen in chapter 6. Subsequently, the authors of a digital journal were connected with their profiles available in Linked Data. The system characterizes a person in four broad aspects like personal, professional, social, and dark side. The system is able to discover resources from Linked Data and can distinguish the retrieved resources into the above mentioned aspects.

Currently, the system utilizes the DBPedia dataset for discovering information. But in future, there is a need to explore other dataset related to authors/experts and thus their relationships need to be explored from Linked Data and other semantic resources. As the data set is semantically rich, it will be helpful in finding the intended information more accurately. The dataset of different publishers, FOAF profiles and social networks would lead to make a really beneficial system for the scholarly e-community.

The mentioned research is now either in productive or in prototype use for a digital journal known as Journal of Universal Computer Science. The users are provided with the required information in a timely way. For example, when a user is viewing a research paper, then the system recommends him Links into the Future (other papers discovered by metadata and autonomous citation mining), experts associated with the topics of the focused paper, emerging concepts and associated resources from socially maintained digital libraries, as well as authors/-experts profiles from Linked Data. Users' feedbacks have shown that the system is able to find required information timely.

Bibliography

[**About Google Scholar 2009**] About Google Scholar, http://scholar.google.at/intl/en/scholar/about.html

[**ACM-CCS, 1998**] ACM (1998). ACM Computing Classification System.

[**Ackerman 1993**] Ackerman, M.S. (1993). Answer Garden: A Tool for Growing Organizational Memory. Ph.D. dissertation, Massachusetts Institute of Technology, 1993.

[**Afzal and Abulaish 2007**] Afzal, M. T., Abulaish, M. (2007). Ontological Representation for Links into the Future. In: Proceedings of International Conference on Convergence Information Technology, pp. 1832-1837, Gyeongju, Korea, 21-23, Nov. 2007.

[**Afzal et al 2007**] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2007). Creating Links into the Future, Journal of Universal Computer Science, 13 (9), pp. 1234-1245, 2007.

[**Afzal et al 2008**] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2008). Expertise Finding for an Electronic Journal, In: Proceedings of International Conference on Knowledge Management and Knowledge Technologies, pp. 436-440, Graz, Austria, 3-5, Sep. 2008.

[**Afzal et al 2009**] Afzal, M. T., Latif, A., Ussaeed, 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., Maurer, H., Balke, W. T., Kulathuramaiyer, N. (2009), Improving Citation Mining, In: Proceedings of International Conference on Networked Digital Technologies, pp. 116-121, Ostrava, Czech Republic, 28-31, Jul. 2009.

[**Afzal et al 2009b**] 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.

[**Afzal 2009a**] Afzal, M. T. (2009). Discovering Links into the Future on the Web, In: Proceedings of Fifth International Conference on Web Information Systems and Technologies, pp. 123-129, Lisbon, Portugal, 23-26, Mar. 2009.

[**Afzal 2009b**] Afzal, M. T. (2009). Information Supply of Related Papers from the Web for Scholarly e-Community, Accepted in Lecture Notes in Business Information Processing, Issue number, pp., 2009.

[**Afzal 2009c**] Afzal, M. T. (2009). Applying Ontological Framework for Finding Links into the Future from Web, In: Proceedings of International Conference on Semantic Systems, pp. 656-662, Graz, Austria, 2-4, Sep. 2009.

[**Afzal 2010**] Afzal, M. T. (2010). Context Aware Discovery and Visualization of Experts for Scholarly e-Community, accepted in Journal of Universal Computer Science, 2010

[**Afzal et al 2010a**] Afzal, M. T., Helic, D., Trattner, C.(2010). Context Aware Discovery and Visualization of Relevant and Evolving Concepts from Social Bookmarking for Scholarly e-Community, accepted in Journal of Universal Computer Science, 2010.

[**Afzal et al 2010b**] Afzal, M. T., Latif, A., Helic, D., Tochtermann, K., Maurer, H.(2010). Discovery and Construction of Authors' Profiles from Linked Data (A case study of Open Digital Journal). to be submitted to: WWW2010 worksop 'Linked Data on the Web(LDOW)'. Raleigh, North Carolina. 27, Apr. 2010.

[**Agichtein and Ganti 2004**] Agichtein, E., Ganti, V. (2004). Mining reference tables for automatic text segmentation. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 20-29, Seattle, WA, USA, 22-25, Aug. 2004.

[**Alani et al 2003**] Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., Shadbolt, N. (2003). Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems, 18 (1), pp. 14-21, 2003.

[**Aleman-Meza et al 2008**] Aleman-Meza, B., Decker, S.L., Cameron, D., Arpinar, I.B. (2008). Association Analytics for Network Connectivity in a Bibliographic and Expertise Dataset. In: Book of Cardoso, J., Lytras, M. D. Semantic Web Engineering in the Knowledge Society, IGI Global ISBN: 978-1-60566-112-4, 2008.

[**Anjewierden et al 2005**] Anjewierden, A., de Hoog, R., Brussee, R., Efimova, L. (2005). Detecting knowledge flows in weblogs. In: Proceedings of 13th International Conference on Conceptual Structures, Kassel, Germany, 18-22, Jul. 2005.

[**Andreasen et al 2004**] Andreasen, T., Jensen, P. A., Nilsson, J. F., Paggio, P., Pedersen, B. S., Thomsen, H. E. (2004). Content-based Text Querying with Ontological Descriptors, Data Knowledge Engineering, 48(2), pp. 199-219, 2004.

[**ARIADNE 2009**] ARIADNE Project, Electronic Library of Learning Modules (2009). http://www.ariadne-eu.org/index.php.

[**Arms 2000**] Arms, W. (2000). Digital Libraries. The MIT Press. ISBN-13: 978-0262011808.

[**Association of Research Library 1995**] Definition and Purposes of a Digital Library Association of Research Libraries, 1995. http://sunsite.berkeley.edu/ARL/definition.html.

[**ATIS 2000**] Alliance for Telecommunications Industry Solutions. Telecom Glossary 2000.

[**Auer et al 2009**] Auer, S., Bizer, C., Idehen, K. (2009). DBpedia Knowledge Base, http://dbpedia.org.

[**Auer et al 2007**] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z (2007). DBpedia: A Nucleus for a Web of Open Data. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 722-735, Busan, Korea, 11-15, Nov. 2007.

[**Balog et al 2006**] Balog, K., Azzopardi, L., de Rijke, M. (2006). Formal Models for Expert Finding in Enterprise Corpora. Special Interest Group on Information Retrieval, Seattle, pp. 43-50. Washington, 6-11, Aug. 2006.

[**Berners-Lee 1998**] Berners-Lee, T. (1998). Semantic Web Road Map, W3C Design Issues, http://www.w3.org/DesignIssues/Semantic.html.

[**Berners-Lee 2006**] Berners-Lee, T. (2006). Linked Data Design Issues. http://www.w3.org/DesignIssues/LinkedData.html.

[**Berners-Lee and Fischetti 1999**] Berners-Lee, T., Fischetti, M. (1999). Weaving the Web, Harper, San Francisco, 1999.

[**Berners-Lee et al 2006**] 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, pp. 06-22, Athens, Georgia, USA, 6, Nov. 2006.

[**Berners-Lee et al 2001**] Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. Scientific American, 279. 2001.

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

[**Bettencourt et al 2006**] Bettencourt, L. M. A., Castillo-Ch´avez, C., Kaiser, D., Wojick, D.E. (2006). Population Modeling of the Emergence and Development of Scientific Fields, Report for the Office of Scientific and Technical Information, 4, Oct. 2006.

[**Bontas et al 2005**] Bontas, E. P., Mochol, M., Tolksdorf, R. (2005). Case Studies on Ontology Reuse, In: Proceedings of International Conference on Knowledge Management and Knowledge Technologies, pp. 436-440, Graz, Austria, June 29 - July 1, 2005.

[**Borgman 1999**] Borgman, C. L. (1999). What are Digital Libraries? Competing Visions, Information Processing and Management, 35(3), pp. 227-243, 1999.

[**Borkar et al 2001**] Borkar, V., Deshmukh, K., Sarawagi, S. (2001). Automatic segmentation of text into structured records. In: Proceedings of the ACM SIGMOD Conference, pp. 175-186, San Francisco, California, USA, 26-29, Aug. 2001.

[**Boykin 2005**] Boykin, J. (2005). Google Top 10 choices for search results, http://www.jimboykin.com/googles-top-10-choices-for-search-results/.

[**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 and Page 1998**] Brin, S., Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 30, pp. 107-117, 1998.

[**Broder 2006**] Broder, A. (2006). The Future of Web Search: From Information Retrieval to Information Supply, In Lecture Notes in Computer Science, 4032, pp. 362, 2006.

[**Bush 1945**] Bush, V. (1945). As We May Think. The Atlantic Monthly, 176(1), pp. 101-108, 1945. http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm.

[**Calude et al 1994**] Calude, C., Maure, H., Salomaa, A. (1994). Journal of Universal Computer Science, 0 (0), pp. 109-116, 1994.

[**Cameron 2007**] Cameron, D. L. (2007). SEMEF: A Taxonomy-based Discovery of Experts, Expertise and Collaboration Networks. MS thesis, The University of Georgia, ATHENS, GA 2007.

[**Cameron et al 2007**] Cameron, D., Aleman-Meza, B., Decker, S. L., Arpinar, I. B. (2007). SEMEF: A Taxonomy based Discovery of Experts, Expertise and Collaboration Networks. University of Georgia, LSDIS Lab, Technical Report, July 2007.

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

[**Candela et al 2006**] Candela, L., Castelli, D., Fuhr, N., Ioannidis, Y., Klas, C.-P., Pagano, P., Ross, S., Saidis, C., Schek, H.-J., Schuldt, H., Springmann, M. (2006). Current Digital Library Systems: User Requirements vs Provided Functionality, Deliverable D1.4.1, Mar. 2006, retrieved on 22, Oct. 2009.

[**Candela et al 2007**] Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D. Agosti, M., Dobreva, M., Katifori, V., Schuldt, H. (2007). The DELOS Digital Library Reference Model - Foundations for Digital Libraries. Version 0.98, Feb. 2007, retrieved on 22, Oct. 2009 from http://www.delos.info/files/pdf/ReferenceModel/DELOSDLReferenceModel0.98.pdf.

[**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.

[**Carmody et al 1969**] Carmody, S., Gross, W., Nelson, T. H., Rice, D., Dam, V. A. (1969). A Hypertext Editing System for the I360, Center for Computer Information Sciences, Brown University, April, 1969, Providence, Rhode Island, File Number HES360-0, Form AVD-6903-0, pages 26-27.

[**Catarci et al 2007**] Catarci, T., F, Levialdi., M, S., Batini, C. (2007). Visual Query Systems for Databases: A Survey. Journal of Visual Languages and Computing, 8(2), pp. 215-260, 2007.

[**Chakrabarti 2004**] Chakrabarti, S. (2004). Breaking through the Syntax Barrier: Searching with Entities and Relations. In: Proceedings of Principles and Practice of Knowledge Discovery in Databases, pp. 9-16, Berlin, Heidelberg, Germany, 20-24, Sep. 2004.

[**Chen et al 2007**] Chen, C., Maceachren, A., Tomaszewski, B., MacEachren, A. (2007). Tracing conceptual and geospatial diffusion of knowledge, Lecture Notes in Computer Science, 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, pp. 1101-1102, Beijing, China, 21-25 Apr. 2008.

[**Chimezie 2009**] Chimezie. (2009). Dereferencing a URI to RDF., http://esw.w3.org/topic/DereferenceURI.

[**Chirita et al 2006**] Chirita, Paul-A., Firan, C.S., Nejdl, W. (2006). Pushing Task Relevant Web Links down to the Desktop. In: Proceedings of 8th ACM International Workshop on Web Information and Data Management. Arlington, Virginia, USA, 10, Nov. 2006.

[**CiteSeer 2009**] CiteSeer - Citation index, http://www.citeulike.org/.

[**Citeulike 2009**] CiteUlike - Social bookmarking service, www.citeulike.org.

[**Cortez et al 2007**] Cortez, E., da Silva, A.S., Goncalves, M.A., Mesquita, F., de Moura, E.S. (2007). FLUX-CIM:Flexible Unsupervised Extraction of Citation Metadata. In: Proceedings of Joint Conference on Digital Libraries, pp. 215-224, Vancouver, British Columbia, Canada, 18-23, June. 2007.

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

[**Crow and Shadbolt 2001**] Crow, L., Shadbolt, N. (2001). Extracting Focused Knowledge from the Semantic Web. International Journal of Human-Computer Studies, 54(1), pp. 155-184. 2001.

[**Cugini et al 1996**] Cugini, J. V., Piatko, C. D., 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, pp. 213-220, Rockville, Maryland, USA, 12-16, Nov. 1996.

[**Dakka and Ipeirotis 2008**] Dakka, W., Ipeirotis, P. (2008). Automatic extraction of useful facet hierarchies from text databases. In: Proceedings of 24th International Conference on Data Engineering, pp. 466-475, Cancun, Mexico, 7-12, Apr. 2008.

[**Dakka et al 2006**] Dakka, W., Dayal, R., Ipeirotis, P. (2006). Automatic discovery of useful facet terms. In: Proceedings of SIGIR Workshop on Faceted Search, ACM Press, Seattle, Washington, USA, 2006.

[**Day 2008**] Day, M. (2008). Institutional Repositories and Research Assessment, Supporting Study No. 4, UKOLN, www.rdn.ac.uk/projects/eprints-uk/docs/studies/rae/rae-study.pdf.

[**Day et al 2007**] Day, M., Tsai, R.T., Sung, C., Hsieh, C., Lee, C., Wu, S., Wu, K., Ong, C., Hsu, W. (2007). Reference Metadata Extraction using a Hierarchical Knowledge Representation Framework. Decision Support Systems, 43, pp. 152-167, 2007.

[**DBLP 2009**] Digital Bibliography and Library Project, http://www.informatik.uni-trier.de/ley/db/index.html.

[**DELOS 2001**] DELOS Brainstorming Report. Digital Libraries: Future Directions for a European Research Programme, San Cassiano, Alta Badia, Italy. 13-15, Jun. 2001.

[**DELOS 2009**] Network of Excellence on Digital Libraries (2009).

[**DeRose and Dam 1999**] DeRose, S., Dam, V. A. (1999). The Lost Books of Hypertext. In Markup Technology, Vol. 1, Issue 1 - Winter 1999. Cambridge: MIT Press.

[**Diederich et al 2007**] Diederich, J., Balke, W-T., Thaden, U. (2007). Demonstrating the Semantic GrowBag: Automatically Creating Topic Facets for FacetedDBLP, In: Proceedings of JCDL'07, pp. 17-22, Vancouver, British Columbia, Canada, Jun. 2007.

[**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, pp. 47-62, Taipei, Taiwan, 8-9, Nov. 1999.

[**Ding et al 2004**] Ding, L., Finin, T., Joshi, A., Pan, R., S. Cost, R., Peng, Y., Reddivari, P., C. Doshi, V., Sachs, J. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. In: Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, pp. 652 - 659, Washington, D.C., USA, 8-13, Nov. 2004.

[**DLF 1998**] Digital Library Foundation (1998). A Working definition of Digital Library, retrieved on 22, Oct. 2009 from http://www.diglib.org/about/dldefinition.htm.

[**Dorogovtsev and Mendes 2002**] Dorogovtsev, S.N., Mendes, J.F.F. (2002). Evolution of Networks, Advances in Physics, 51, pp. 1079-1187, 2002.

[**Dreher et al 2004**] Dreher, H., Krottmaier, H., Maurer, H. (2004). What we Expect from Digital Libraries, Journal of Universal Computer Science, 10 (9), pp. 1110-1122, 2004.

[**Dublin Core Metadata Initiative 2001**] Dublin Core Metadata Initiative (2001). http://www.dublincore.org.

[**Duval 2000a**] Duval, E. (2000). Ariadne: share and reuse requires interoperable metadata, In: Proceedings of Workshop on Metadata for Multimedia Information - Dublin Core, Brussel, 25, Jan. 2000.

[**Duval 2000b**] Duval E. (2000). Learning Object Metadata, In: Proceedings of VDAB, Brussel, 18, Apr. 2000.

[**Engelbart 1968**] Engelbart, D.C. (1968). oNLine System Demo, 9, Dec. 1968. http://sloan.stanford.edu/MouseSite/1968Demo.html.

[**Feldstein 2009**] Feldstein, A. (2009). Search ranking factors. http://www.seomoz.org/article/search-ranking-factors.

[**Fensel et al 2001**] Fensel, D., Horrocks, I., Harmelen, F. van, McGuinness, D. L. Patel-Schneider, P. (2001). OIL: Ontology Infrastructure to Enable the Semantic Web, IEEE Intelligent Systems, 16(2), pp. 38-45, 2001.

[**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, Pharmacotherapy 2006,26:759-67. doi:10.1592/phco.26.6.759

[**Fox 1993**] Fox, E. (1993). Source Book on Digital Libraries, Version 1.0, 6, Dec. 1993. Covering a series of NSF Invitational Workshops and Related Information, ftp://fox.cs.vt.edu/pub/DigitalLibrary/DLSB.pdf.

[**Garfield 1955**] Garfield, E. (1955). Citation Indexes for Science. Science, 122, pp.108-111, 1995.

[**Garfield 1964**] Garfield, E. (1964). Can Citation Indexing be Automated. In: Symposium proceedings of Statistical Association Methods for Mechanized Documentation, pp. 189-192, 15, Dec. 1964.

[**Garfield 1972**] Garfield. E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. Science, 178, pp. 471-479. 1972.

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

[**Giles et al 1998**] Giles, C.L., Bollacker, K.D., Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. In: Proceedings of Third ACM Conference on Digital Libraries, pp. 89-98, Pittsburgh, Pennsylvania, United States. 23-26, Jun. 1998.

[**Glänzel and Thijs 2004**] Glänzel, W., Thijs, B. (2004). Does co-authorship inflate the share of self-citations? Scientometrics, 61 (3), pp. 395-404, 2004.

[**Goldfinch et al 2003**] Goldfinch, S., Dale , T., Derouen, K. Jr. (2003) 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.

[**Google API 2009**] Google Search API, http://code.google.com/apis/soapsearch/reference.html.

[**GoogleBlog 2008**] Google Blob, We knew the web was big... (2008).http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[**Graham 2001**] Graham, G. (2001). Cutting Through the Electronic Hype. 100-day Dialogue, Topic 3, What's Next For Publishing: Terminal Decline or Golden Age?, 2001.

[**Gruber 1993**] Gruber, T. R. (1993). A translation approach to portable ontology specification, Knowledge Acquisition, Special issue: Current issues in knowledge modeling , 5 (2), pp. 199-220. 1993.

[**Hall et al 2008**] Hall, R., Sutton, C., McCallum, A. (2008). Unsupervised Deduplication using Cross-field Dependencies. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 310-317. Las Vegas, Nevada, USA, 24 - 27, Aug. 2008.

[**Handschuh et al 2002**] Handschuh, S., Staab, S., Ciravegna. F. (2002). SCREAM - Semi-automatic Creation of Metadata, Lecture Notes In Computer Science, 2473, pp. 358-372, 2002.

[**Harth et al 2007**] Harth, A., Umbrich, J., Hogan, A., Decker. S. (2007). YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 211-224, Busan, Korea, 11-15, Nov. 2007.

[**Hecht et al 1998**] Hecht, F., Hecht, B. K., Sandberg, A.A. (1998). The Journal Impact Factor: A Misnamed, Misleading, Misused Measure, Cancer Genet Cytogenet, Elsevier Science Inc. 04, pp. 77-71, 1998.

[**Heinrich and Maurer 2000**] Heinrich, E., Maurer, H. (2000). Active Documents: Concept, Implementation and Application, Journal of Universal Computer Science, 6(12), pp. 1197-1202, 2000.

[**Hepp et al 2006**] Hepp, M., Siorpaes, K., Bachlechner, D. (2007). Harvesting Wiki Consensus Using Wikipedia Entries as Vocabulary for Knowledge Management, IEEE Internet Computing. 11(5), pp. 54-65. 2007.

[**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, pp. 272-285, Athens, Georgia, USA, 5-9, Nov. 2006.

[**Hirsch 2005**] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output, PNAS 102 (46), pp. 16569-16572. 2005.

[**Hitchcock 2002**] Hitchcock, S. M. (2002). Perspectives in Electronic Publishing: Experiments with a New Electronic Journal Model. Ph.D. dissertation, University of Southampton, 2002.

[**Hotho et al 2006**] Hotho, A., Jaschke, R., Schmitz1, C., Stumme, G. (2006). Information Reterival in Folksonomies: Search and Ranking, LECTURE NOTES IN COMPUTER SCIENCE, 4011, pp.411-426, 2006.

[**Hyperbolic Package 2009**] Hyperbolic Tree Library,
http://hypertree.cvs.sourceforge.net/viewvc/hypertree/hypertree/.

[**Hyperwave 2009**] Enterprise content management solution,
http://hyperwave.com/e/.

[**Hu and Jaffe 2003**] Hu, A.G.Z., Jaffe, A.B. (2003). Patent citations and international knowledge flow: the cases of Korea and Taiwan. International Journal of Industrial Organization, 21, pp. 849-880, 2003.

[**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), 2008.

[**Jacsó 2008**] Jacsó, P. (2008). Reference Reviews,
http://www.gale.cengage.com/reference/peter/200708/SpringerLink.htm

[**J.UCS 2009**] Journal of Universal Computer Science, http://www.jucs.org.

[**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 the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 295-308, Busan, Korea, 11-15, Nov. 2007.

[**Kitzinger 1995**] Kitzinger, J. (1995). Qualitative research. Introducing focus groups. British Medical Journal, 311, pages 299-302.

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

[**Krottmaier 2003**] Krottmaier, H. (2003). Links to the Future, Journal of Digital Information Management ,1 (1), pp. 3-8, 2003.

[**Krulwich 1995**] Krulwich, B., Burkey, C. (1995). ContactFinder: 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, pp. 85-91.

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

[**Kobilarov and Dickinson 2008**] Kobilarov, G., Dickinson, I. (2008). Humboldt: Exploring Linked Data. In: Proceedings of Linked Data on the Web Workshop, Beijing, China, 22, Apr. 2008.

[**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, pp13-14, Marina del Rey, California, USA, 2-4, Nov. 1994.

[**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), pp. 33-55. 1996.

[**Lamping et al 1995**] Lamping, J., Rao, R., Pirolli, P. (1995). A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In: Proceeding of the SIGCHI Conference on Human Factors in Computing Systems, pp. 401-408, Denver, Colorado, 7 - 11, May. 1995.

[**Latif et al 2009**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). CAF-SIAL: Concept aggregation framework for structuring informational aspects of linked open data, In: Proceedings of International

Conference on Networked Digital Technologies, pp. 100-105, Ostrava, Czech Republic, 28-31, Jul. 2009.

[**Latif et al 2009a**] Latif, A., Hoefler, P., Stocker, A., Ussaeed, A., Wagner, C. (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: Proceedings of International Conference on Semantic Systems, pp. 568-576, Graz, Austria, 2-4, Sep. 2009.

[**Latif et al 2009b**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009).Harvesting Pertinent Resources from Linked Data, accepted in Journal of Digital Information Management.

[**Latif et al 2009c**] Latif, A., Tanvir, M.T., Hoefler, P., UsSaeed, A., Tochtermann, K.(2009) "Translating Keywords into URIS", accepted in the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 24-26 Nov. 2009.

[**Liddle et al 2003**] Liddle, S. W., Hewett, K. A., Embley, D. W. (2003). An Integrated Ontology Development Environment for Data Extraction. In: Proceedings of 2nd International Conference on Information Systems Technology and its Applications (ISTA'03), pp. 21-33.

[**Liew and Foo 2001**] Liew, C.L., Foo, S. (2001). Electronic Documents: What Lies Ahead?, In: Proceedings of 4th International Conference on Asian Digital Libraries, pp 88-105, Banglore, India, 10-12, Dec. 2001.

[**Lipsman 2009**] Lipsman, A. (2009). ComScore Releases April 2009 U.S Search Engine Rankings, 2009.

[**Liu and Dew 2004**] Liu, P., Dew, P. (2004). Using Semantic Web Technologies to Improve Expertise Matching within Academia. In: Proceedings of the 2nd International Conference on Knowledge Management, pp. 370-378, Graz, Austria, June 30- July 2, 2004.

[**Lyman and Varian 2000**] Lyman, P., Varian, H. R. (2000). How Much Information, retrieved on 22, Oct. 2009 from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

[**MacGarvie 2005**] MacGarvie, M. (2005). The determinants of international knowledge diffusion as measured by patent citations, Economics Letters, 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 pp.152.

[**Mathes 2004**] Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication through Shared Metadata, Technical Report, 2004.,

[**Marchionini and Maurer 1995**] Marchionini, G., Maurer, H. (1995). The roles of digital libraries in teaching and learning, Communication of the ACM, vol. 38, No. 4, pp. 67-75, 1995.

[**Marlow et al 2006**] Marlow, C., Naaman, M., Boyd, M., Davis, M., HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp.31-40, Odense, Denmark, 22-25, Aug. 2006.

[**Maurer 1982**] Maurer, H. (1982). Videotex makes dramatic breakthrough, In: Proceedings of Viewdata, pp. 135-143, London, UK, 1982.

[**Maurer 1996**] Maurer, H. (1996). Hyper-G now Hyperwave - The Next Generation Web Solution. Addison-Wesley.

[**Maurer 2009**] Maurer, H. (2009). Before the Internet, invited speech at Ars Electronica, Linz/Austria, 06, Sep. 2009.

[**Maurer and Posch 1982**] Maurer, H., Posch, K. (1982). MUPID: An Austrian contribution to videotext, Technical Report F 87, IIG Graz (1982).

[**McCahill and Anklesaria 1995**] McCahill, M. P., Anklesaria, F. X. (1995). Evolution of Internet Gopher, Journal of Universal Computer Science, 1(4), pp. 235-246, Nov. 1994.

[**Maurer and Tomek 1990**] Maurer, H., Tomek, I. (1990). Some aspects of Hypermedia Systems and their treatment in Hyper-G, Wirtschaftsinformatik 32 (1990), pp- 187-196, 1990.

[**Maurer 2001**] Maurer, H. (2001). Beyond Digital Libraries. Global Digtial Library Development in the New Millenium, In: Proceedings of NIT Conference, pp.165-173, Beijing, 2001.

[**Maurer and Tochtermann 2002**] Maurer, H., Tochtermann, K. (2002). On a New Powerful Model for Knowledge Management and its Applications, Journal of Universal Computer Science, 8(1), pp. 85-96, 2002.

[**Maurseth and Verspagen 2002**] Maurseth, P. B., Verspagen, B. (2002). Knowledge Spillovers in Europe: A Patent Citations Analysis, Scandinavian Journal of Economics, 104 (4), pp. 531-545, 2002.

[**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, pp. 39-44, Graz, Austria, 17-21, Jun. 1995.

[**MeSH 2009**] Medical Subject Headings (MeSH), http://www.nlm.nih.gov/mesh/.

[**Michlmayr and Cayzer 2007**] Michlmayr, E., Cayzer, S. (2007). Learning User Profiles from Tagging Data and Leveraging them for Personalized Information Access, In: Proceedings of 16th International World Wide Web Conference, Banff, Alberta, Canada, 8-12, May. 2007.

[**Mika 2005**] Mika, P. (2005). Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Proceedings of 4th International Semantic Web Conference, pp. 5-15, Galway, Ireland, 6-10, Nov. 2005.

[**MSN API 2009**] MSN Live API, http://msdn2.microsoft.com/enus/library/bb266180.aspx.

[**Mockus and Herbsleb 2002**] Mockus, A., Herbsleb, J. A. (2002). Expertise Browser: A Quantitative Approach to Identifying Expertise. In: Proceedings of International Conference on Software Engineering, pp. 503-512, Orlando, Florida, 19-25, May. 2002.

[**Natalya and Deborah 2001**] Natalya, F. N., Deborah, L. M. (2001). Ontology Development 101: A Guide to creating your First ontology Recommendations,

[**Nature Editorial 2005**] Nature Editorial (2005). Not-so-deep impact, Nature 435, pp. 1003-1004, 23, Jun. 2005.

[**Nelson 1965**] Nelson, T. H. (1965). Complex information processing: a file structure for the complex, the changing and the indeterminate. In: Proceedings of the twentieth national conference, pp. 84-100, Cleveland, Ohio, United States, 1965.

[**Nelson et al 2007**] Nelson, T.H., Smith, R. A., Mallicoat, M. (2007). Back to the Future: Hypertext the Way It Used to Be. In: Proceedings of the HyperText, 2007, pp. 227-228, Manchester, United Kingdom.

[**Odlyzko 1994**] Odlyzko, A. M. (1995). Tragic Loss or Good Riddance? The Impending Demise of Traditional Scholarly Journals, Journal of Universal Computer Science, 0(0), pp-3-53, 1994.

[**Odlyzko 1998**] Odlyzko, A. M. (1998). The economics of electronic journals, Journal of Electronic Publishing, 4(1), Sep. 1998.,

[**Open Office 2009**] Open Office, a DOC to PDF converter, http://www.openoffice.org/.

[**O'Reilly 2007**] O'Reilly, T. (2007). Freebase Will Prove Addictive. O'Reilly Radar. http://radar.oreilly.com/archives/2007/03/freebasewillp1.html.

[**Park and Park 2006**] Park, G., Park, Y. (2006). On the measurement of patent stock as knowledge indicators, Technological Forecasting and Social Change, 73 (7), pp. 793-812, 2006.

[**PDFBox 2009**] PDFBox, a PDF to text converter, http://www.pdfbox.org/

[**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, pp. 111-136, MIT-Press, Cambridge MA.

[**PLoS Medicine Editors 2006**] The PLoS Medicine Editors. (2006). The impact factor game. It is time to find a better way to assess the scientific literature. PLoS Med.3:e291.

[**PLoS Editorial 2006**] PLoS Medicine Editors (2006). The impact factor game. It is time to find a better way to assess the scientific literature. PLoS Medicine. 3, pp. 707-708, 2006.

[**Popescul and Ungar**] Popescul, A., and 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, pp. 282, 2008.

[**Price 2008**] Price, G. (2008). Google Scholar Documentation and Large PDF Files, http://blog.searchenginewatch.com/blog/041201-105511.

[**PubMed 2009**] PubMed, http://www.ncbi.nih.gov/entrez/query.fcgi.

[**Puntschart and Tochtermann 2006**] Puntschart, I., Tochtermann, K. (2006). Online-Communities and the un-importance of e-Moderators. In: Proceedings of Networked Learning, Lancaster, UK, Apr. 2006.

[**Ratprasartporn and Ozsoyoglu 2007**] Ratprasartporn, K., Ozsoyoglu, G. (2007). Finding Related Papers in Literature Digital Libraries. ECDL: Lecture Notes In Computer Science, 4675, pp. 271-284, 2007.

[**Rhodes and Maes 2000**] Rhodes, B.J., Maes, P. (2000). Just-in-time information retrieval agents, IBM System, 39 (3-4), pp. 685-704, 2000.

[**Roberts et al 2007**] Roberts, R.J., Varmus, H.E., Ashburner, M., Brown, P.O., Eisen, M.B., Khosla, C., Kirschner, M., Nusse, R., Scott, M., Wold, B. (2001). Building A GenBank of the Published Literature. Science, 291 (5512), pp. 2318-2319. http://www.sciencemag.org/cgi/content/full/291/5512/2318a.

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

[**Russell et al 2008**] Russell, A., Smart, P. R., Braines, D. and Shadbolt, N. R. (2008). NITELIGHT: A Graphical Tool for Semantic Query Construction. In: Proceedings of Semantic Web User Interaction Workshop, Florence, Italy, 5, Apr. 2008.

[**Samur and Daniel 2009**] Samur, C. F., Daniel, S. (2009). Explorator: a tool for exploring RDF data through direct manipulation. In: Proceedings of Linked Data on the Web Workshop (LDOW). Madrid, Spain. 20, Apr. 2008.

[**Saracevic 2000**] Saracevic, T. (2000). Digital Library Evaluation: Towards an Evolution of Concepts. Library Trends, 49(3), pp. 350-369. http://comminfo.rutgers.edu/ tefko/LibraryTrends2000.pdf.

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

[**Scharnhorst and Wouters 2006**] Scharnhorst, A., Wouters, P. (2006). Web Indicators - a new Generation of S T Indicators, International journal of scientometrics, Informmetrics and Bibliometrics, 10 (1), 2006.

[**Schmaranz 1998**] Schmaranz, K. (1998). Aspects of Electronic Publishing in Hypermedia Systems, Ph.D. dissertation, Graz University of Technology, 1998.

[**Seglen 1997**] Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research, BMJ, 314(7079), pp. 497.

[**Shneiderman 2002**] Shneiderman, B. (2002). Inventing Discovery Tools: Combining Information Visualization with Data Mining, Information Visualization, 1(1), pp. 5-12, mar. 2002.

[**Small 1973**] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, pp. 265-269. 1973.

[**Snoussi et al 2002**] Snoussi, H., Magnin, L., Nie, J. -Y. (2002). Toward an Ontology-based Web Data Extraction. In: Proceedings of 15th Canadian Conference on Artificial Intelligence. Calgary, Alberta, Canada, 26, May. 2006.

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

[**Speretta and Gauch 2005**] Speretta, M., Gauch, S. (2005). Personalized Search Based on User Search Histories. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, France, 19-22, Sep. 2005.

[**Spoerri 2004**] Spoerri, A. (2004). RankSpiral: Toward Enhancing Search Results Visualizations. In: Posters Compendium, IEEE Symposium on Information Visualization, pp. 39-40, Austin, Texas, USA, 10-12, Oct. 2004.

[**Suchanek et al 2007**] Suchanek, F. M., Kasneci, G., Weikum, G. (2007). Yago: A Core of Semantic Knowledge- Unifying WordNet and Wikipedia. In: Proceedings of 16th International World Wide Web Conference, pp. 697-706, Banff, Alberta, Canada, 8-12, May. 2007.

[**Sun and Giles 2007**] Sun Y., Giles C. L. (2007). Popularity Weighted Ranking for Academic Digital Libraries, In: Proceedings of 29th European Conference on Information Retrieval Research, pp. 605-612, Rome, Italy, 5-7, Apr. 2007.

[**Tho et al 2007**] Tho, Q.T., Hui, S.C., Fong, A.C.M. (2007). A Citation Based Document Retrieval System for Finding Research Expertise, Elsevier: Information Processing and Management, Issue 43, pp. 248-264, 1, Jan. 2007.

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

[**Tummarello et al 2007**] Tummarello, G., Delbru, R. Oren, E (2007). Sindice.com: Weaving the open linked data. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 545-560, Busan, Korea, 11-15, Nov. 2007.

[**Turtle Graphics 2009**] Turtle Graphics,
http://www.gkrueger.com/java/aufgaben/loesung/TurtleGraphics.java

[**UsSaeed 2007**] Ussaeed, A., Stocker, A., Hoefler, P., Tochtermann, K. (2007). Learning with the Web 2.0: The Encyclopedia of Life. In: Proceedings of

Interactive Conference on Computer Aided Learning, Villach, Austria, 20, Jan. 2009.

[**UsSaeed 2008a**] Us Saeed, A., Afzal, A., Latif, A., Stocker, A., Tochtermann, K. (2008). Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study. In: Proceedings of International Conference on Convergence and Hybrid Information Technology, pp. 605 - 610, Busan, Korea, Nov. 11-13, 2008.

[**UsSaeed 2008b**] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers, In: Proceedings of IEEE International Mutitopic Conference, pp. 392-397, Karachi, Pakistan, Dec. 23-24, 2008.

[**Vargas-Vera et al 2001**] Vargas-Vera, M., Motta, E., Domingue, J., Buckinham, S. S., Mattia, L. (2001). Knowledge Extraction by using an Ontology Based Annotation Tool, In: Siegfried Handschuh, Rose Dieng, Steffen Staab (Eds), Proceedings Workshop on Knowledge Markup Semantic Annotation, held in association with the First International Conference on Knowledge Capture, pp. 5-12, Victoria, Canada, 2001.

[**Wu et al 2006**] Wu, H., Zubair, M., Maly, K. (2006). Harvesting Social Knowledge from Folksonomies. In: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp.111-114, Odense, Denmark, 22-25, Aug. 2006.

[**Xanadu 2009**] http://www.xanadu.net/.

[**Yahoo API 2009**] Yahoo Search API, http://www.programmableweb.com/api/yahoo-search.

[**Yankelovich et al 1987**] Yankelovich, N., Smith K. E., Garrett, L. N., Meyrowitz, N. (1987). Issues in Designing a Hypermedia Document System: The Intermedia Case Study, in Learning Tomorrow: Journal of the Apple Education Advisory Council, n3, pp. 35-87, 1987.

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

Appendices

Appendix A

List of Publications

The work covered by this thesis led to following publications:

[Afzal and Abulaish 2007] Afzal, M. T., Abulaish, M. (2007). Ontological Representation for Links into the Future. In: Proceedings of International Conference on Convergence Information Technology, pp. 1832-1837, Gyeongju, Korea, 21-23, Nov. 2007.

[Afzal et al 2007] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2007). Creating Links into the Future, Journal of Universal Computer Science, 13 (9), pp. 1234-1245, 2007.

[Afzal et al 2008] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2008). Expertise Finding for an Electronic Journal, In: Proceedings of International Conference on Knowledge Management and Knowledge Technologies, pp. 436-440, Graz, Austria, 3-5, Sep. 2008.

[UsSaeed 2008a] Us Saeed, A., Afzal, A., Latif, A., Stocker, A., Tochtermann, K. (2008). Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study. In: Proceedings of International Conference on Convergence and Hybrid Information Technology, pp. 605 - 610, Busan, Korea, Nov. 11-13, 2008.

[UsSaeed 2008b] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers, In: Proceedings of IEEE International Mutitopic Conference, pp. 392-397, Karachi, Pakistan, Dec. 23-24, 2008.

[Khan et al 2009] Khan, M. S., Afzal, M. T. Kulathuramaiyer, N., Maurer, H. (2009). Extended Visualization for a Digital Journal, In: Proceedings of the Fifth International Conference on Web Information Systems and Technologies, pp. 385-388, Lisbon, Portugal, March 23-26, 2009.

[Afzal et al 2009] Afzal, M. T., Latif, A., Ussaeed, 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., Maurer, H., Balke, W. T., Kulathuramaiyer, N. (2009), Improving Citation Mining, In: Proceedings of International Conference on Networked Digital Technologies, pp. 116-121, Ostrava, Czech Republic, 28-31, Jul. 2009.

[Afzal et al 2009b] 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.

[Afzal 2009a]Afzal, M. T. (2009). Discovering Links into the Future on the Web, In: Proceedings of Fifth International Conference on Web Information Systems and Technologies, pp. 123-129, Lisbon, Portugal, 23-26, Mar. 2009.

[Afzal 2009b] Afzal, M. T. (2009). Information Supply of Related Papers from the Web for Scholarly e-Community, Accepted in Lecture Notes in Business Information Processing, Issue number, pp., 2009.

[Afzal 2009c] Afzal, M. T. (2009). Applying Ontological Framework for Finding Links into the Future from Web, In: Proceedings of International Conference on Semantic Systems, pp. 656-662, Graz, Austria, 2-4, Sep. 2009.

[Latif et al 2009] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). CAF-SIAL: Concept aggregation framework for structuring informational aspects of linked open data, In: Proceedings of International Conference on Networked Digital Technologies, pp. 100-105, Ostrava, Czech Republic, 28-31, Jul. 2009.

[Latif et al 2009b] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). Harvesting Pertinent Resources from Linked Data, accepted in Journal of Digital Information Management.

[Latif et al 2009c] Latif, A., Tanvir, M.T., Hoefler, P., UsSaeed, A., Tochtermann, K.(2009). "Translating Keywords into URIS", accepted in the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 24-26 Nov. 2009.

[Afzal 2010] Afzal, M. T. (2010). Context Aware Discovery and Visualization of Experts for Scholarly e-Community, accepted in Journal of Universal Computer Science, 2010.

[Afzal et al 2010a] Afzal, M. T., Helic, D., Trattner, C.(2010). Context Aware Discovery and Visualization of Relevant and Evolving Concepts from Social Bookmarking for Scholarly e-Community, accepted in Journal of Universal Computer Science, 2010.

[Afzal et al 2010b] Afzal, M. T., Latif, A., Helic, D., Tochtermann, K., Maurer, H.(2010). Discovery and Construction of Authors' Profiles from Linked Data (A case study of Open Digital Journal). to be submitted to: WWW2010 worksop 'Linked Data on the Web(LDOW)'. Raleigh, North Carolina. 27, Apr. 2010.

List of Figures

List of Tables