Graz University of Technology

Doctoral Thesis

Aspects of Content Quality Management in Digital Libraries of Scholarly Publications

Muhammad Salman Khan

Institute for Information Systems and Computer Media,
Graz University of Technology, Austria

First Examiner:
Univ.-Prof. Dr.Dr.h.c.mult. Hermann Maurer
Graz University of Technology, Austria

Second Examiner:
Univ.-Prof. Dr. Denis Helic
Graz University of Technology, Austria

Graz, June 2011

STATUTORY DECLARATION

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

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

....

(date) (signature)

To my parents, Azra Naz and Anwer Ahmed Khan

Kurzfassung

Digitale Bibliotheken wissenschaftlicher Publikationen spielen eine entscheidende Rolle beim Erlangen und Verbreiten von Wissen. Im aktuellen digitalen Zeitalter sind Forscher auf diese Ressourcen beim Suchen, Nachschlagen und beim Vermitteln ihres Wissens angewiesen. Voraussetzung für den Erfolg dieses Systems ist die Schaffung eines Vertrauensverhältnisses der Nutzer in die Qualität der Inhalte in diese digitalen Bibliotheken.

Um gute Qualität zu gewährleisten, setzt man bei der Herausgabe wissenschaftlicher Arbeiten auf das traditionelle Peer-Review-System. Einige wissenschaftliche Zeitschriften führen bibliometrische und inhaltliche Analysen der eingereichten Manuskripte durch um sicherzustellen, dass diese auch ihren Vorgaben entsprechen und interessante Forschung vermitteln. Allerdings sind durch die immer größer werdende Menge an Informationen, den Umfang der Einreichungen und die Anforderungen der wissenschaftlichen Gemeinschaft herkömmliche Techniken nicht mehr ausreichend, um die Qualität der Inhalte in diesen wissenschaftlichen Medien zu gewährleisten. Innovative Ideen zur Unterstützung der Herausgabe wissenschaftlicher Arbeiten sind daher dringend notwendig, um die gewünschte Qualität sicherzustellen. In dieser Arbeit werden verschiedene neue Ansätze zur Lösung dieser Probleme behandelt.

Die Arbeit unterstreicht zuerst die Notwendigkeit, sich die neuen Entwicklungen im Web, das mittlerweile zu einer community-betriebenen Plattform wurde, zu Nutze zu machen, um die Anzahl der Leser, der Autoren und die Qualität der Publikationen elektronischer Zeitschriften zu erhöhen. Sie gibt ein erstes Beispiel für ein Web-Mashup (ein Web-2.0-Paradigma), das eine große collaborative Computing-Plattform als eine zeitgemäße Lösung für die Herausgabe von digitalen Zeitschriften bietet, und als Entscheidungshilfe für die Benutzer und die Herausgeber der Zeitschrift dient.

Traditionell werden von den Herausgebern wissenschaftlicher Medien bibliometrische und inhaltliche Analysen von Manuskripten mit Hilfe von Diagrammen und Tabellen durchgeführt. Diese Dissertation zeigt, wie dieser Prozess durch ein interaktives Visualisierungssystem erweitert werden kann, das solche Analysen besser unterstützt. Es bietet ein interaktives, einfach zu bedienendes Werkzeug, um verschiedene verborgene Muster in wissenschaftlichen Publikationen aufzudecken, und um somit die Herausgabe von wissenschaftlichen Zeitschriften zu unterstützen.

Peer Review wird bei wissenschaftlichen Arbeiten als Grundlage der Sicherung der Qualität von Manuskripten betrachtet. Allerdings können unterschiedlichste soziale und kognitive Interessenskonflikte (COI) zwischen Autoren und Begutachtern das Ergebnis der Gutachten entscheidend beeinflussen. Bestehende COI Erkennungssysteme basieren hauptsächlich auf Ko-Autoren Netzwerken, die wiederum aus bibliographischen Datenbanken erstellt werden, um mögliche COI Situationen zu identifizieren. Diese befassen sich jedoch nicht mit kognitiven COI Situationen. Diese Dissertation zeigt, wie man verschiedene Zitat-Netzwerke verwenden kann, um bestehende COI Nachweisverfahren verbessern und so potenzielle soziale UND kognitive COI Zusammenhänge zwischen Forschern sichtbar zu machen.

Mit Hilfe der Anzahl der Artikel Downloads kann der Impact Faktor eines Artikels im Vergleich mit späteren Zitierungszahlen gemessen werden. Das kann auch dazu verwendet werden, um zukünftige Zitierungszahlen eines Artikels zu berechnen. Unter Bedachtnahme der Bedeutung von Downloads, untersucht diese Arbeit verschiedene lokale und globale, einem Artikel zugeordnete Attribute, um die Anzahl künftiger Downloads vorherzusagen. Sollte die Anzahl der Einreichungen stark ansteigen, könnten solche Vorhersagen Herausgebern beim Einreichungsprozess dahingehend unterstützen, dass bereits eine erste Begutachtung durchgeführt bzw. eine Vorauswahl getroffen wird, bevor die Arbeiten einer strengeren Begutachtung unterzogen werden.

Die in dieser Arbeit vorgestellten Lösungen unterstützen die Herausgabe von wissenschaftlichen Arbeiten in den verschiedenen Phasen des Prozesses, d.h. von der Einreichung eines Artikels bis hin zur Analyse der gesammelten Veröffentlichungen.

Es wird erwartet, dass die hier vorgestellten Ideen Herausgeber dabei unterstützen können, die Qualität von Publikationen und Diensten für die wissenschaftliche Gemeinschaft zu verbessern.

Abstract

Digital libraries of scholarly publications play a vital role in the acquisition and dissemination of knowledge. In the current digital age, researchers rely heavily on these resources to search, consult, and communicate their knowledge. Establishing a relationship of trust of users on the quality of content in these digital libraries is at the heart of the success of this process. To ensure quality, the administration of scholarly communications relies on a traditional peer-review system. Some scholarly journals conduct scientometrics, bibliometrics, and content analysis of their manuscripts to ensure that it is aligned to its policies and is communicating quality research. However, with their growing information size, volume of submissions, and increasing demands of scholarly communities, conventional techniques for managing the quality of content in these scholarly mediums are becoming insufficient. There is an impending need to come up with innovative ideas in helping managers of scholarly communications to ensure quality. This thesis proposes various novel solutions to address these problems.

The thesis first highlights the need of harnessing the new developments on the web, which by now has turned into a community-driven platform, in expanding electronic journals' readership, authorship and quality of their publications. It demonstrates a pioneer example of a web mash-up (an emerging Web 2.0 paradigm), which provides a rich collaborative computing platform as a timely solution for the content management of a digital journal, and as a decision-making tool for the users and the administration of the journal.

Traditionally, the administration of scholarly communications conducts scientometrics and content analysis of manuscripts using static charts and tables. This dissertation demonstrates the extension of an interactive visualization system that can support such analysis at deeper level. It provides an interactive, easy to use solution to uncover various hidden patterns in scholarly publications to strengthen the internal administration of scholarly communications.

Peer review in scientific communications is considered as a basis to ensure quality of manuscripts. However, different kinds of social and cognitive conflict of interest (COI) situations between authors and reviewers can compromise the review decision. Existing COI detection systems primarily rely on co-authors networks extracted from bibliographic databases to identify potential COI situations. They do not deal with

cognitive COI situations. This dissertation demonstrates a novel idea of using different citations relationships as an improvement to existing COI detection techniques to spotlight potential social and cognitive COI situations between researchers.

Articles downloads provide a timely measure about the usage impact of an article as compared to citations. It can also be used to anticipate future citations of an article. By keeping in view the importance of downloads, this thesis explores various local and global attributes associated with an article to predict its future downloads. In cases, where volume of submissions is increasing rapidly, such predictions can assist the administration of scholarly journals to conduct an initial review or pre-selection of manuscripts before submitting it for more rigorous review.

The solutions presented in this thesis support the administration in various phases of scholarly communication process, i.e. from submission of an article to the scientometrics and content analysis of articles collection. It is expected that the ideas demonstrated in this thesis will help the administration in following their policies to increase the quality of content and services for the scholarly community.

Acknowledgement

First of all, I would like to thank Almighty Allah for providing me such an opportunity that has changed my life. I am thankful to Him for enabling me to think rationally. I remember all those things that appeared unachievable to me, but by the grace of Allah everything became easy for me.

I would like to thank my supervisor Prof. Hermann Maurer for giving me the opportunity to work with him. His working style and the way of addressing research questions inspired me throughout my studies. He is really an inspiration for young scientists. I am also thankful to Prof. Denis Helic for being the part of evaluating this thesis as second reader.

I am grateful to Prof. Narayanan Kulathuramaiyer for his support and encouragements during his stay in Graz and even when he left for Malaysia after completing his doctoral studies. His quick reviews, suggestions, and moral support enabled me to learn about the know-how of research.

I would like to pay gratitude to all my colleagues in IICM, specially Tanvir Afzal, Bial Zaka, and Mohammad Al-Smadi for encouragements and useful discussions. Thanks to Dana Kaiser and Helmut Leitner for giving me insights about J.UCS, and for providing help in testing and deploying mash-ups. I would also like to thank Maria-Luise Lampl for providing me help in every administrative work, and for making arrangements for conferences travels and their registration process. I really appreciate the technical assistance provided by Ingeborg Schinnerl, Walter Schinnerl, and Karlheinz Trummer.

Beside my working activities, I spend most of my time with my Pakistani friends in Graz. I am thankful to them for their time and encouragements specially, Syed Nadeem Ahsan, Shahzad Saleem, Shafiq Siraj, Tahir Mushtaq, Jadoon Khan, and Zahid Hussain Abro. I would also like to pay my gratitude to my hostel mates Konal Pal, Philip, and Angelo Spinola. Last but not least, thanks to Inayat Khan who discussed and provided valuable inputs for the classification experiments in the dissertation.

I really appreciate the Higher Education of Pakistan (HEC) for providing me the scholarship for doctoral studies. Similarly, I am also grateful to ÖAD for arranging my accommodation in Graz and providing every assistance related to my scholarship.

I would also like to acknowledge my officials in Virtual University of Pakistan who helped me in achieving the scholarship of HEC. I am thankful to Dr. Naveed A. Malik, Mr. Tajdar Alam, Dr. Sadaqat Mehdi, and Dr. Zahir Fikri for trusting in my abilities. I would also like to thank Rizwan for reviewing my proposal for research. Thanks to Miss Saba Toor, Ehtesham Dar, Mr. Anwar, Mr. Shahid, Shahid Iqbal, Zeeshan Akbar, and Ehsan Dar for encouraging me to pursue for my doctoral studies.

I am thankful to my parents who encouraged me to go abroad and achieve my dreams. Their continued prayers always helped me throughout my work. I am thankful to my siblings Faiqa, Farhan, Imran, and Bushra for taking care of everyone in the family. Last but not least, I would like to say "thank you" to my wife Tania, who besides her own research work took care of me and motivated my work.

Muhammad Salman Khan
Graz, June 2011

Contents

List of Figures

List of Tables

1 Introduction

This chapter provides a brief overview of digital libraries, their history, and various current research trends and challenges associated with modern digital libraries. Furthermore, it describes the objective of this thesis and provides an overview about its contributions for managing the quality of content in digital libraries of scholarly publications.

The chapter is divided in to six sections. The first section gives a brief overview of the history of digital libraries. The second section highlights various perceptions of different communities on digital libraries. The third section discusses developments related to digital libraries of electronic journals. Thereafter, the fourth section presents various current research trends and challenges related to the digital libraries. Based on the challenges associated with digital libraries, the scope of this dissertation is described in the fifth section. Finally, an outline of the dissertation and its contributions for managing the quality of content in the digital libraries of scholarly publications is presented in sixth section.

1.1 A Short History Related to Digital Libraries

Long before current online publishing technology, Vannevar Bush was the first one to envision a novel system that could help in efficient storage and faster dissemination of human knowledge when compared to traditional library system [Bush, 1945]. He called this system as *"memex"*. This system was never implemented [Hitchcock, 2002], but its description inspired many ideas to implement a kind of memex with current technology. Dr. Bush was originally appointed as a Director of the Office of Scientific Research and Development, and lead activities of nearly six thousand American scientists to use applications of science for human welfare [Bush, 1945]. In his article, *"As We May Think"* published in 1945 in *"The Atlantic Monthly"*, Dr. Bush highlighted the growing

gap between the amount of human knowledge and research material being produced and the potential of investigators to grasp and remember this knowledge. This is a problem that we still face today as *"Information overload"*. Some important points from this article are summarized here. According to Dr. Bush [Bush, 1945]:

"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. Today we make the record conventionally by writing and photography, followed by printing; but we also record on film, on wax disks, and on magnetic wires. Even if utterly new recording procedures do not appear, these present ones are certainly in the process of modification and extension."

Here the word *"record"* is considered as a piece of information and should not be confused with the record stored in databases or flat files [Krottmaier, 2002]. Dr. Bush further highlighted various technologies present at his time and their possible extensions that could help in this regard. He anticipated the potential to reduce the size of information, and to store it cost efficiently in micro films. The whole *"Encyclopedia Britannica could be reduced to the volume of a matchbox"*, said Dr. Bush. People would be able to create information using machines that would understand spoken instructions and would perform complex mathematical computations. All recorded information would be rapidly accessible from the repository through indexes but also more importantly by associations between information. Bush described that the main information access system might consist of a desk, keyboard, and levers. A small part of the system would be devoted for storage and remaining for other necessary operations. The user would be able to open a book or material by typing its code on the keyboard. The book or material could be read by micro film projections on the screens placed on top of aforementioned desk. The lever would let the user to open pages of material in forward or backward direction. To simulate paper-based reading, the material could be annotated. The most important part of his system was that a user could interlink or join related pieces of information. The numerous items when joined together form a trail, which could be stored, reviewed, traversed with the lever, and could be shared or inserted in another memex system. The inter-linked information thus form a new personalized book, which could be further extended with new knowledge of other people.

Another pioneer in the field of digital libraries is J. C. R. Licklider [Arms, 2000]. According to William Y. Arms, in 1960, Licklider in his book *"The Library of the Future"* described the work required to build the library of the future. William Y. Arms noted that although the computing power at that time was not strong enough, J. C. R. Licklider managed to anticipate the future of digital libraries, whose most components later proved to be true.

Theodor Holm Nelson also known as Ted Nelson is another visionary scholar who worked in the early days of digital libraries. He coined the term *"Hypertext"* in 1965, while proposing a new form of information and file structure (zippered lists and evolutionary list file) for handling information (using proposed file language PRIDE) [Neslon, 1965]. He provided many examples of its applications. According to his proposed system, the contents would be accessible through indexes. The files would not be arranged in a default hierarchal structure but would be flexible to be structured in any form and under different categories imposed by the users. Bush trails would be facilitated (later Nelson called these trails Transclusions [Nelson et al., 2007], a function that facilitates the placement of a chunk of destination document in the body of the other document containing the link of the aforementioned destination document [Helic, 2001]). Annotations might be attached to any information. The content of the file and their arrangements would be able to be updated. It would enable "dynamic outlining", which would allow an automatic update to one text sequence caused by changes in an other attached text. The system would be able to keep different revisions or versions of a document as long as needed, to facilitate later comparisons by the user. He defined hypertext as a non-sequential text, and anticipated that all information one day will be accessible for users from one large repository containing different pathways [Hitchcock, 2002]. Ted Nelson defined *"Hypertext"* as follows [Neslon, 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 content 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 substantiate this vision and implement the described features, a project called *"Xanadu"* was started in 1960 [Nelson, 1987, Hitchcock, 2002]. This project is still under development, a partial implementation of it, however, was released in 2007 with the name *"XanaduSpace"* [Xanadu, 2011].

The first working hypertext system called Hypertext Editing System (HES) was developed in 1967 by Ted Nelson, Andries van Dam, and students at Brown University in collaboration with International Business Machines (IBM) [Helic, 2001]. In a keynote address at Hypertext Conference in 1987, Andries van Dam described that HES system was shown and ported to different universities, and various sites where IBM had the largest customers [van Dam, 2011]. The speaker further pointed out that the system was later sold by IBM to the Apollo mission team to produce their documentation.

In 1968, Douglas C. Engelbart, another pioneer of the current publishing systems demonstrated an extensive system called NLS or *"oN Line System"* developed at Stanford Research Institute [Engelbart and English, 1968, Engelbart, 1975]. The authors in [Engelbart and English, 1968] reported that NLS system provided an environment in which knowledge workers could perform all of their central and everyday tasks such as creating, reading, searching, and manipulating data files, with an option for collaboration and sharing information with other people. It was also the first system with a mouse driven interface [van Dam, 2011].

Inspired by the main features of Douglas's NLS system, Andries Van Dam in 1968 worked again on another hypertext system called FRESS, a File Retrieval and Editing System [van Dam, 2011]. According to authors in [Yankelovich et al., 1985], the FRESS system was a text only system with multi-user support, but without concurrent updating. The authors further pointed out that the system provided two main options for linking, i.e., tags to open a new document in a window while staying in the current document and jumps to navigate from one document to other document. It also supported bidirectional links, visualization of text structure, and most importantly the undo feature for revoking changes [van Dam, 2011]. The users had the option to assign names to links and text for later references and searching [Yankelovich et al., 1985]. FRESS was used for almost two decades at Brown university for creating and disseminating documents [Yankelovich et al., 1985].

The first widely available commercial hypertext product was OWL's Guide [Nielsen, 1995, Moser, 1998]. Guide was basically a standalone window-based hypertext system for both authoring and browsing of content [Helic, 2001]. It was originally developed for UNIX and later successfully commercialized for Macintosh and IBM PCs [Nielsen, 1995, Moser, 1998].

In a presentation at Banff, Canada, Hermann Maurer described the pre-web activities in Europe [Maurer, 2007]. Some important points from this presentation and a contribution in [Maurer, 2001b] are summarized here. According to Hermann Maurer, Sam Fedida, an engineer at British Telecom implemented the first networked information and communication system at large scale in 1976. This system was called "PRESTEL" and later as "Videotex". It used telephone lines and TV sets (as a display device for input and output), which people usually have at their home, to connect to the computer networks for information and services. Additionally, it required a decoder to connect with the TV to display the transmitted information, and a modem for communication between analog telephone lines and decoder. The remote of the TV was used as a keypad input device. The Prestel systems was capable to display both appealing

textual and graphical information by using 16 colors and a large set of "mosaic" characters. Various versions of Prestel system were implemented in different countries such as, France, Germany, Canada, Japan, and Austria. It attracted millions of users in Europe. These systems were basically used for providing information such as, news, directories, time tables, events, routes, ordering tickets and books, and SMS like messages, etc. The information was able to be searched with hierarchical menus and links via numeric codes. Germany provided better graphics for their Videotex version known as "Bildschirmtext" or "BTX". France replaced the TV remote with alphabetic keyboard which resulted in a system known as "Teletel". Later, black and white screen and the keyboard were integrated and distributed to millions of household users in France. This integrated system was called as "Minitel". In Austria, Maurer and Posch [Maurer and Posch, 1982], considered Videotex not only as an information system but a network of compatible computers. They replaced the Videotex decoder with a working computer. At that time personal computers were not available, so they developed a Z80 based colored graphics computer called "**M**ehrzweck **U**niversell **P**rogrammierbarer **I**ntelligenter **D**ecoder (MUPID)" [Maurer, 1982, Maurer and Posch, 1982]. MUPID worked as a personal computer with a keyboard and an optional external storage device. It allowed word-processing, discussion forums, multi-person remote games, electronic networked encyclopedia (links via numbers), and downloading of softwares called "telesoftware" similar to "Java Applets" used today. In 1986, the MUPID system was deployed for networked learning by delivering 500 hours of "COSTOC" lessons [Maurer and Kaiser, 1986, Maurer, 2007]. Approximately 50,000 MUPIDS were produced in Austria from 1982 to 1989. The MUPID systems, however, could not compete with the marketing power of IBM PCs and BTX not with WWW.

The Aspen Movie Map is believed to be the first ever hypermedia system [Moser, 1998]. It was developed in 1978 at Massachusetts Institute of Technology by Andrew Lippman and his colleagues [Lippman, 1980, Helic, 2001, Moser, 1998]. According to the author in [Lippman, 1980], the system was basically a topographic application to simulate drive through an unknown place. Lippman further explained that the user of the system was presented with sequences of images taken through single frame cameras about a place or with their computer generated 2D and 3D animated replicas. The users of the system could navigate through streets of a place by using a touch-screen or joystick provided with the system. The system also interlinked short videos about the buildings or the locales in the environment, so that a user could stop at any place and explore the surroundings.

NoteCards, developed by Frank G. Halasz and his team [Halasz, 1988] at Xerox PARC was a widely used second generation hypermedia system. According to the

author in [Halasz, 1988], it was basically designed for authors, researchers, and other intellectual laborers to help them in developing and organizing their ideas in a systematic way. The author further explained the working of the system. Halasz showed that the system consisted of four main components, i.e., notecards, links, browsers, and fileboxes. The notecards were an electronic generalization of 3×5 paper notecard, containing notecard title and text or graphic material. These notecards could be displayed and edited by using standard Xerox Lisp windows. The links could be used to interconnect individual notecards. The browsers were specialized notecards that could display and edit the structural diagram of interconnected notecards. The fileboxes were designed to further organize the notecards under some categorization or hierarchical structure. The notecards system further provided simple keywords search facility.

In 1987, a major breakthrough for hypertext and hypermedia systems took place when Apple decided to bundle HyperCard free with their Macintosh systems [Moser, 1998]. According to Moser, this initiative opened for the first time an interactive multimedia system for authoring and browsing by the general public. The author further described that the HyperCard was basically a simple standalone hypertext system that not only provided the basic hypertext functionality but an easy to use scripting language called HyperTalk. According to the author, the HyperTalk could be attached to any HyperCard object such as buttons, graphics, or links. These scripts were automatically invoked when the object is accessed. The author further reported that with easy to use functionalities many users started to make HyperCard Stacks as hypermedia applications.

In the mid 1980s the early hypertext systems were started to be superseded by Hypermedia systems that can interconnect audio-visual material in addition to textual data [Hitchcock, 2002]. However, the lack of integration of hypertext applications made them restrictive for the daily work of knowledge workers [Hitchcock, 2002]. As a solution to this problem an open hypermedia system called *"Intermedia"* was implemented in 1985 at Brown University [Yankelovich et al., 1985, Hitchcock, 2002]. According to Yankelovich et al., the Intermedia provided the basic framework of providing the navigational links between multimedia documents created using different applications. A general approach of link service was further introduced by [Pearl, 1989], that can integrate wider range of applications than realized by Intermedia [Hitchcock, 2002]. Later, many systems were developed to incorporate this approach [Hitchcock, 2002].

Sophisticated hypertext systems started to appear in late 80s and early 90s [Hitchcock, 2002]. The Web, WAIS, Gopher, and Hyper-G primarily are the results of these initiatives. It is quite interesting that the Web [Berners-Lee et al., 1994], Gopher [McCahill and Anklesaria, 1995], and Hyper-G [Maurer, 1996] projects were launched

in the same year. The Web (WWW or World Wide Web) technology started as a personal information tracking system in a distributed project. The Web was developed by Tim Berners-Lee, Robert Cailliau and their colleagues at CERN, a European research center for high energy physics in Switzerland [Berners-Lee et al., 1994]. It is currently considered as one of the most widely used forms of communication. Its ability to easily address any object on the Internet made it expand rapidly [Hitchcock, 2002]. The popularity of the web was further emphasized by innovative browser interfaces provided by Mosaic, through which web documents can be viewed, and the simple utility of HTML (Hypertext Mark-up Language) through which main web content can be created [Hitchcock, 2002]. The Mosaic was developed by Mark Anderson and his team in 1993 and later commercialized as Netscape Navigator [Moser, 1998].

The authors in [Andrews et al., 1995] provided an overview about differences between WAIS, Gopher, WWW, and Hyper-G. Some important points from this article are also summarized here. The Wide Area Information Systems (WAIS) [Kahle et al., 1992] began in 1989 as a joint project of Thinking Machines, Apple Computers and Dow Jones to access Wall Street. As mentioned in [Andrews et al., 1995], WAIS was basically a search engine for the indexed databases that included relevance feedback from the user to refine the subsequent search. The authors further pointed out that it did not support any structure of the information content as well as associative link or hyperlinks between related documents. Gopher [McCahill and Anklesaria, 1995] started in 1991 as an information system for University of Minnesota. According to authors in [Andrews et al., 1995], it provided a menu like access to the information space. But in Gopher there was no built-in capability of search and relied on external search engines as add-ons. Similar to WAIS it did not support hypertext [Andrews et al., 1995]. The Hyper-G system was launched by Hermann Maurer and his team at Institute for Information Systems and Computer Media, Graz University of Technology, Austria in 1991 [Maurer, 1996]. Hyper-G introduced many features as a solution to various problems faced by the WWW, WAIS, and Gopher [Andrews et al., 1995]. The major design goals of Hyper-G included: proper structuring of documents by using collection of documents, and collections of collections for hierarchal navigation (not strictly hierarchal but as an acyclic directed graph, where two directories could have common subdirectory), hyperlinks management to remove the problem of dangling links in WWW, annotations to documents, attaching additional links in addition to originals to different media (documents, images, audio, film, and 3D scene), and even links to and from a section of audio or video, facility to see documents pointing to current documents (linking backward), and integrated efficient search across different collections [Pam and Vermeer, 1995]. According

to the authors in [Pam and Vermeer, 1995], the Hyper-G system also included a native client or viewer of documents called Harmony. Harmony was able to provide support to different forms of media, i.e., text, images, audio, video, click-able moving objects 3D scenes, PostScript, and could be configured to run external programs [Pam and Vermeer, 1995]. Later, Hyper-G was commercialized and renamed as Hyperwave [Maurer, 2007]. As Hyper-G or Hyperwave provided various features that are necessary for the working of practical digital libraries, it was chosen for the implementation of a digital journal, i.e., Journal of Universal Computer Science (J.UCS) [Krottmaier, 2002]. J.UCS is now in its 17^{th} year. As peer-reviewed open access journal that also appears in printed form it can be considered as the ancestor of all serious electronic journals today. With a current 5 years impact factor of 0.799 and some 85,000 readers it remains one of the influential journals in computer science. The work presented in Chapter 2, 3, and 5 are based on studies conducted on J.UCS. A brief overview about J.UCS can be found in Section 1.3 of this chapter.

After providing a brief overview about the ideas related to digital libraries, the next section describes different definitions and perspectives about the term digital libraries as perceived by different communities.

1.2 Digital Libraries

In literature, the phrases *"electronic library"* and *"digital library"* are used interchangeably. However, the authors in [Fox et al., 1995] noted a shift from *"electronic"* to *"digital"* as a term, perhaps due to the growing interest in digital networks, digital audio, and digital video in relation to electronic publishing. The authors further pointed out that digital library can have different meanings for different people. According to the authors, for a computer scientist, it is simply a distributed interlinked information (text-based, multimedia based information) space, generally of high value available in or outside an organization. For a person from library science, it might just refer to carrying out functions of a conventional library in a new way that can include, new information resources, organizational practices, methods of preservation, methods of interaction with customers, and novel ways of cataloging and classification of resources [Fox et al., 1995]. Those working in educational technology see it as a support for both formal and informal learning [Fox et al., 1995]. The people working on collaborative technologies see digital library as a mean of communicating, creating, and sharing new knowledge [Fox et al., 1995]. The author in [Fox, 1993] quoted about the problem in defining the term as follows:

"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"

William Y. Arms [Arms, 2000] categorized the people in two groups, who are much involved in the innovation of digital libraries. According to William Y. Arms, the first group includes computer science researchers and Internet developers. The other group includes information professionals such as publishers and wide range of information providers (indexing and abstracting services) [Arms, 2000]. Tefko Saracevic [Saracevic, 2001] called these groups research community and practice community respectively. Saracevic further pointed out that in research community the digital libraries has not yet defined. According to him, the closest definition that encompasses the approaches taken by research community is provided by Lesk [Lesk, 1997] as follows:

"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"

William Y. Arms [Arms, 2000] also provided a definition for digital library, which the author called as informal definition, and can be considered as a representative of research community perspective. He defined digital library as:

"a digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network."

In 1995, the Digital Library Foundation (DLF) was founded in United States as an organization of research libraries and various national institutions [Saracevic, 2001]. The organization represented practice community, and managed to provide the first working definition of digital libraries from the perspective of practical community [Saracevic, 2001].

"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."

Tefko Saracevic directed that the best definition that bridges the gap between the two communities was provided by Borgman [Borgman, 1999]:

1. *"Digital Libraries are a 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

2. *Digital libraries are constructed, collected, and organized, by (and for) a community of users, and their functional capabilities support the information needs and uses of that community"*

Recently DELOS [DELOS, 2011], a Network of Excellence on Digital Libraries provided a more comprehensive definition of the digital library in their Digital Library Reference Model [Candela et al., 2007], which provides the perspective of both research and practice community. DELOS is basically an initiative partially funded by the European Commission in the frame of the Information Society Technologies Program (IST) [DELOS, 2011]. It facilitates the coordination and integration of ongoing research in the field of digital libraries by the major research teams in Europe and other regions of the world. The main objective of DELOS is *"to develop the next generation of Digital Library technologies, based on sound comprehensive theories and frameworks for the life-cycle of Digital Library information"* [DELOS, 2011]. According to the DELOS reference model, a digital library is [Candela et al., 2007]:

"An organisation, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialised functionality on that content, of measurable quality and according to codified policies."

1.3 Digital Libraries of Electronic Journals

While early hypertext systems and Engelbart's NLS system were expanding slowly in the 60s and 70s, scientists at other places were trying to produce first electronic journals [Hitchcock, 2002]. The history of scientific journals goes back to at least 350 years, where for the first time two journals appeared simultaneously in London (Philosophical Transactions of the Royal Society of London) and Paris (Le Journal des Scavans) in 1665 [Guédon, 2001]. As observed by Don Schauder [Schauder, 1994], these initiatives were taken approximately 200 years after the invention of printing. The author in [Roes, 1994] identified main functions of a journal, i.e., communication and dissemination of information, quality control and archiving of contents. The author further emphasized that these functions should be fulfilled by any new medium

if it has to serve properly as a form of scholarly communication. Various studies showed that the conventional print information cycle is collapsing, due to increase in volume and cost of journals literature, and delays in communicating the published material [Odlyzko, 1994, Arms, 2000]. The author in [Odlyzko, 1994] observed that in many fields the literature is becoming doubled every 10 years after the World War II. As an example, the author further highlighted that in Mathematics, only 840 papers were published up till 1870, and since then it has risen up to 50,000 papers annually. On the other side, the technical barriers to replicate the information cycle of print journals by electronic means were disappearing [Odlyzko, 1994],[Arms, 2000]. The cost of electronic storage of contents was decreasing, personal computers, and computer networks were being deployed [Arms, 2000]. By late 80s several publishers and libraries became interested in delivering the electronic versions of their scientific journals [Hitchcock, 2002]. Although the earlier adoption of electronic journals was slow, over time it has grown rapidly. The author in [Odlyzko, 1994] observed a growth of 70% in the number of electronic journals, and predicted that electronic publications will become a dominant form of scholarly communication between 2000 and 2010. Today, the term *"electronic journal"* is used for a journal that maintains many properties of the conventional printed journals, but at the same time is created and delivered through online means [Arms, 2000]. According to William Y. Arms [Arms, 2000], the same term rather confusingly is also used for journals that are published online only and for the electronic versions of a primarily printed journal. The first electronic journal projects where initiated in 1976 at the New Jersey Institute of Technology by using the *"Electronic Information Exchange System"* (EIES) [Hitchcock, 2002, Sheridan et al., 1981, Turoff and Hiltz, 1982]. Four projects were mainly initiated which included, informal newsletter, an unrefereed or preprints journal, a refereed journal titled *"Mental Workload"* similar to conventional print journal model, and a highly structured inquiry-response system [Sheridan et al., 1981]. The system also facilitated the comments to an article by the readers along with authors' responses as soon as the article is published [Sheridan et al., 1981]. The authors in [Sheridan et al., 1981] discussed the causes of failure of the Mental Workload journal project. According to the authors, the reason for the failure include: lack of motivation of the academic community, complicated interfaces, different policies, and cost issues. After a failed attempt for a peer reviewed electronic journal, a similar project titled *"Blend"* was originated in UK between 1980 to 1984 [Shackel, 1991]. Learning from the EIES project, in this project, the contents were not published, but were merely archived [Hitchcock, 2002]. According to Hitchcock, the main emphasis of Blend was to support the review cycle of manuscripts from submission to editorial acceptance. The

articles then could be published elsewhere with an acknowledgment [Hitchcock, 2002]. By attaching comments and discussion forums to papers, and support for collaborative writing, the project is considered to be a herald for many currently available electronic journals [Hitchcock, 2002].

In 1994, the Journal of Universal Computer Science (J.UCS) was launched at Institute for Information Systems and Computer Media, Graz, Austria [Maurer and Schmaranz, 1994, Calude et al., 1994]. J.UCS is a high quality peer reviewed digital journal that covers all aspects of computer science [J.UCS, 2007]. J.UCS publishes at least 12 issues per year. Currently, it has more than 300 highly profiled editors across the world enabling a broad coverage of all aspects of Computer Science. At the end of each year, a volume of J.UCS is published as a printed copy and archived [Maurer and Schmaranz, 1994, Calude et al., 1994]. This printed version exactly matches the electronic version with the same pagination as the online edition. The electronic versions are thus static documents, frozen over time. As previously described, the content of J.UCS are hosted using a Hyperwave system. It inherently provides extended search to full-text, and annotations feature of its content. An annotation refers to a note or a comment about an existing publication informing readers about new research results or errors. J.UCS also implemented private, public, and group-bases annotations. These annotations are applied subject to an evaluation based on a refereeing process and are only added if deemed appropriate. The verified annotations make it possible to insert only objective comments and prevent the misuse of annotations for personal disputes. J.UCS publishes articles mainly in PDF and PostScript format. The articles can also be browsed by volumes, authors and ACM categories. Publications in J.UCS have a corresponding meta-data (XML) file (one for each paper) that contains all the information about that paper such as title, authors, institutions, cities, countries, keywords used in paper, area of research, date of publication, volume and issue number, etc. All the meta-data files about articles are stored as Hyperwave objects and can be retrieved through Hyperwave APIs. Inspired from the open access movement described in the next paragraph, J.UCS also started to provide open access to its content since 2007. Recently, many new features such as "links to the future", finding experts, linking with linked open data, and social tags recommendations have been implemented for J.UCS [Afzal, 2010]. A brief overview about the "links to the future" capability is described in a forthcoming section.

In 1999, the author in [LeJeune, 1999] first implicitly provided a suggestion for the open access of scholarly journals. By open, the author meant that the original papers can be free but the packaged versions may require payment [Hitchcock, 2002]. ArXiv and other early Internet journals started to offer free access to their contents

[Hitchcock, 2002]. According to the author in [Hitchcock, 2002], this free access became possible due to cost reduction in maintaining the electronic journals. The reduced cost can be met by source of funding other than payment by readers or authors. For example, ArXiv managed to get funding for its cost efficient publication system from US National Science Foundation [Hitchcock, 2002]. In case of J.UCS the publishing cost is met by a consortium of academic institutions. Although the open access movement has been criticized by many electronic publishing researchers, it is accepted by the scholarly community [Hitchcock, 2002].

Similar to open access movement to digital archives of papers, another initiative titled *"Open Archives Initiative"* was born in 1999 [Hitchcock, 2002]. The objective of this initiative was to standardize the metadata describing the contents of a digital library and the means of communicating this metadata to individual services to present the unified views of the collected data, even from heterogeneous digital libraries [Hitchcock, 2002].

1.4 Challenges Associated with Digital Libraries

According to a survey conducted in [Lyman et al., 2011], the world has created five exabytes of new information in 2002. According to the report approximately 92% of this information is stored on magnetic media (largely on hard disks), while only 0.01% on paper (increase from 0.003%, reported by same source in 2000). Although the report highlighted the growth in printed information on paper, the majority of this information belongs to office documents and postal emails, but not in the form of published titles such as: books, journals, and newspapers. The report further estimated that 800MB of information is being produced annually per person, which corresponds to 30 feet of books for a person in the paper form. The report observed an increase of 30% in new stored information from 1999 to 2002. Unfortunately, little of this information is available through digital library collections [Delos, 2001]. The facility to offer a global and coherent access to this information collection will provide an inspiring impact on almost every activity of our daily life [Delos, 2001].

The authors in [Marchionini and Maurer, 1995] highlighted the role of digital libraries for both formal and informal learning activities. The authors anticipated about the libraries of the future that will help both teachers and students to take advantage of the wide range of remotely accessible materials, and in communicating with other people to share resources and expertise to achieve common goals. The authors suggested that the libraries should take responsibility for the legal usage of information

resources. According to the authors, digital libraries may include bibliographic and catalog databases for effective information seeking by users. They pointed out that with the growth of digital libraries, remotely accessible instructions and support will be required to help in improving the information seeking skills of users. They argued that some prior structuring of information is necessary to improve search rather than post processing of results. The authors pointed out that different types of search (e.g., boolean, full-text, and approximate matching) within and across different servers, and their integration with graphical display will make the large digital libraries more convenient for working. The authors envisioned that different communities of interest might be supported by digital libraries which will help in offering more specialized courses across geographical boundaries, and even by the students themselves. The authors emphasized that the students and teachers must learn how to teach and learn with multimedia resources. The authors further suggested that digital libraries should provide easy to use powerful tools in finding, managing, and publishing information. According to the authors, they should also provide a mix of software and people to support reference assistance and question answering systems.

Similarly, based on the practical experiences of Hyper-G and J.UCS implementations, the author in [Maurer, 2001a] identified various problems and lack of functionalities provided by classical digital libraries. According to the author, these problems include: conversion of old material in search-able electronic form, full text search, metadata based search, automated generation of metadata, metadata standards to exchange digital material, grammatical (stop-words, stemming, synonyms) and semantic based search, search in multimedia (sound, picture, and film) based documents, compression and long term storage issues, copy and intellectual rights, and price charging mechanism. The author further shared valuable suggestions and personal experiences from various projects to exploit the real power of the net or WWW which include, annotations (private, group-based, and public) to a document, facility for collaborative activities (chats, discussion boards, and joint workspaces), links to the future, transclusions, version control, and active documents. The author coined the term *"links to the future"* to keep track of a certain research activity and described this situation as *"a contribution x written at some stage may become obsolete due to a new result y published later. In this case a link can be added to the older paper x to point to y to help readers, thus providing a "link to the future""*. The author proposed that a systematic implementation of this functionality can reduce the duplication of research and obviously the anxiety and frustration of authors. The author in [Afzal et al., 2007] has demonstrated a partial implementation of links to the future. The vision of active document is to guide the user to suitable documents [Heinrich and Maurer, 2000],

[Maurer, 2001a]. According to the author in [Maurer, 2001a], active document is basically a question answering system, where a reader can ask any question about a particular document, and the system can automatically answer the question based on previous answers, or can direct it to an expert for immediate response. The author further pointed out that a partial implementation of this system has already been done in Hyperwave [Heinrich and Maurer, 2000].

The authors in [Dreher et al., 2004] further elaborated the gap between the promises and the implemented functionalities in the existing digital libraries. The authors highlighted that it is not possible to store every content in a single library, therefore a portal-approach needs to be implemented to integrate different libraries and provide contents to the users regardless of its storage location. According to the authors, the portal based system must be adaptive to the interests and profiles of the users. The authors further proposed many features for implementation such as: multimedia annotations, active annotations (similar to active documents described above), attachment of links to the any content, personalized links, exploring contents via visualized knowledge maps or topic maps, links to online catalogs of conventional libraries, intelligent search mechanism providing results according to the context of the user, conceptual search, white lists of systems that must be integrated, and adaptive interfaces according to the expertise of the users.

The Semantic Web is an emerging successor of Web, it is expected to change the way scientific publishing is currently produced and shared [Berners-Lee and Hendler, 2011]. Some visionary ideas about how semantic web can impact the scientific publishing have been discussed in [Berners-Lee and Hendler, 2011]. Digital libraries of the future need to take the advantage of semantic web technologies to support (provide/consume) data *"to be shared and reused across application, enterprise, and community boundaries"* [Berners-Lee and Hendler, 2011] for creating valuable services. Similarly, with the advent of Web 2.0, a new concept of Science 2.0 is emerging which applies Web 2.0 technologies and practices to improve, enhance and speed-up the feed-back process of scholarly communications [Kieslinger and Lindstaedt, 2009]. According to [Waldrop, 2011] *"Science 2.0 generally refers to new practices of scientists who post raw experimental results, nascent theories, claims of discovery and draft papers on the Web for others to see and comment on"*. Although there are some criticism concerning Science 2.0 (theft of ideas and vandalism) as with open access, the discussion about Science 2.0 practices might be helpful to resolve these issues [Kieslinger and Lindstaedt, 2009].

In 2001, the DELOS Network of Excellence arranged a brainstorming workshop of various researchers to identify the future directions for the European research program in the field of digital libraries [Delos, 2001]. The goal was to clarify various social

and technical challenges associated with digital libraries research. In this meeting a conceptual framework for a digital library system was defined. They divided the framework in three basic components which are described below.

1. Contents Component: This component is mainly concerned with the creation, preservation, accessibility, and variety of the contents stored in a digital library [Delos, 2001]. They summarized the vision about contents as *"Creating high-quality, semantically rich, comprehensive information collections, usable for long periods of time"* [Delos, 2001]. They identified five main research areas under contents components that need to be addressed. These areas are:
   – Building information collection, which includes automatic acquisition of contents, creating meta information about primary contents, and organization of both primary and meta information for specific tasks and user groups.
   – Accessing and navigating information collection, which require efficient search algorithms, new structure of data, and query optimization, etc.
   – Facilitate different kinds of objects in a collection. Theses objects can include scientific data, simulation models, text, audio, video, images, etc.
   – Support for multilingual and multicultural collections, which require language translation techniques (as storing information in all language is not possible) and meta information about different culture.
   – Long term preservation of the information collection. This research topic includes development of techniques to migrate the collections to new environment, format, and platforms, so that they remain available to users at all time.

2. Management Component: This component is mainly concerned with the research areas that are associated with the architecture, management and administration of the information collection and their metadata [Delos, 2001]. The vision about management component was expressed as *"Developing self-sustainable and expandable DL systems, offering high-quality information and services"* [Delos, 2001]. The major research topics for this component were identified as follows:

   – Developing new forms of architecture that includes component-based and multi-tear architectures to replace conventional 3-tier architecture, which is inadequate to provide the expected functionalities implied by advances in other architecture related issues.
   – Support for open architecture, which means that overall functionality of a digital library can be partitioned in well-defined services or smaller independent services. This requires techniques to be developed that a new component can be

automatically plugged and configured in to the system.
– Facilitate the interoperability of metadata, development and operation of interoperable systems.
– The work is required to make the system scalable, which may include decentralized architectures and various performance indicators of the system.
– Research is required to make the system available at all time. This may require automatic triggers to compensate any failed component and replication of information to other sites.
– Management of sessions and work-flow.
– Work is required to guarantee the integrity of information collection and privacy of user actions. Similarly, the work is required to guarantee the access rights of the material.
– Much of the work is required to define the quality criteria for a digital library. There is a need to define metrics and develop techniques to measure the given quality criteria. The system must process requests based on quality criteria which may be imposed by the system or the user.
– The design, collection and organization of contents must be controlled by a system administrator. Similarly, the administrator must be capable of defining individual users and groups of users for the system. There is a need to develop various tools to help in administrating the digital library.

3. Usage Component: The purpose of this component is to mainly focus in the delivery of contents to the users of the digital library in an effective and efficient manner [Delos, 2001]. Its vision can be described as *"Provide optimal user experience in Digital Library interactions, i.e, support users in accessing Digital Libraries and ensure that they obtain the desired information in the best possible way"* [Delos, 2001]. They identified the main research topics in this area as follows:

   – New paradigms or the integration of existing techniques need to be developed for integrated interaction with these collections. Novel interface are required to facilitate any particular task. There is a need to develop interface description languages to define the family of different user interfaces and generation of interfaces according to any specifications.
   – Techniques are required to visualize different kinds of information and meta-information for dynamic exploration.
   – Personalization and customization of interaction, which require work in explicit and implicit profiling of users. Similarly, the users can annotate any object, and

work is required for the efficient storage of these annotations and intelligent processing of requests containing these annotations.
– All the features can be extended from the personal level to community level. The users can create and share annotations, and can assign ratings and opinions to any information object to guide the community.
– Multilingual and multicultural dependent delivery of contents.
– Facilitate collaboration between users while visiting a digital library.
– The access to digital library should be universal by any person, at any place, and using any device.
– Facilitate persistent sessions across multiple devices and the contents, and services should be adaptable to devices.

1.5 Scope of the Dissertation

From the discussion of the previous sections, it can be concluded that the digital library research area is very broad. Different research groups all around the world are working in one or more sub research topics. It is certainly not possible to tackle all the challenges described above in a single dissertation. Therefore, we only focus on the administration research topic which is described in the previous section under the management component. Administrating digital libraries includes various tasks. However, in this thesis, we focus on only managing the quality of contents in the digital libraries of scholarly publications. Managing the quality of contents in digital libraries is again a very broad and dynamic field. In scholarly publishing systems, the quality of manuscripts is often judged by a rigorous peer review. There are various variations of the peer review system that have been proposed in literature. A detailed discussion about the strengths and weaknesses of these methods is provided in Chapter 4 of this thesis. Some ideas relating to managing the quality of contents has been described in [Arms, 2002]: such as editorial control, reputation system, volunteer reviews, and citations rank. According to the recently released DELOS reference manual [Candela et al., 2007], the quality parameter for a digital library encompasses various inter-related factors and their sub-components which include: generic quality (reputation, sustainability, performance, scalability, etc.), content quality (integrity, authenticity, trustworthiness, viability, etc.), policy quality (policy consistency and precision), functionality quality (usability, user satisfaction, availability, fault management, impact of service), user quality (user behavior and activeness), and architecture quality (ease of administration, log quality, load balancing performance, maintenance

performance, etc.). These quality parameters are not exhaustive and can be extended to model various quality facets based on individual needs [Candela et al., 2007]. In this dissertation, we explore various possibilities and solutions that can assist in the quality management of contents of scholarly communication systems that can help in improving their general standing and reputation.

1.6 Structure of the Dissertation

The dissertation is divided into four chapters in addition to this introductory one and a summary chapter. The majority of work presented in this thesis is based on previously published papers in peer-reviewed journals and conference proceedings. A separate list of these publications has been provided in an Appendix.

The Chapter 2 of this dissertation describes new developments of the web, and the necessity of digital journals to utilize this evolution to expand its readership, authorship, and quality of publications. With their growing information size, conventional techniques to manage the journal and supporting its authors and readers are becoming insufficient. In this context, we describe some concerns faced by the administration and users of a digital journal namely, Journal of Universal Computer Science (J.UCS). In this chapter, we explore the application of an innovative Web 2.0 concept to address these problems. More specifically, we explore the application of mash-ups for J.UCS. A mash-up for a digital journal can serve as a means of providing a rich collaborative platform to support diverse users' needs about high level access options to the e-collections of a journal, and as an administrative support tool to facilitate rapid expansion of the journal and to find potential cases of conflict of interest situations between authors and reviewers on the basis of their locations. In this chapter, we explore a pioneer example of such a mash-up for J.UCS by combining the J.UCS digital repository with Google Map APIs to address various users and administrative concerns. The chapter also describes the difficulty in ascertaining the location information of authors of the journal to locate their precise position on Google Map. It further reports on the potential of mash-ups as a means of community assisted content management system to verify the location information of authors, and updating other contents of the journal. The work presented in this chapter can serve as a model for other contemporary electronic journals. We believe that an extension of this work with more publications data from other sources such as: CiteSeer and DBLP will be very useful for many academic appraisal tasks such as promotions, finding experts, social behavioral patterns and conflict of interest detection in peer review system.

The Scientometrics and content analysis of scholarly publications has been a tradition of many electronic and printed journals to ensure quality and the journal's standing. Traditionally, these analysis are usually conducted using normal tables and pie charts. Recently, these kinds of analysis are also being supported by the field of information visualization. The Chapter 3 provides a brief overview about these visualization techniques and their limitations to support such analysis. It further describes applying an interactive visualization system that can help for deeper scientometrics and content analysis of digital journals. The adapted visualization system is an easy to use web application, based on animated 2D bubble charts and pie charts to handle geographical, temporal, and large kinds of categorical data. As a first case study and in continuation of the first chapter, we report our results after applying this technique to J.UCS to strengthen the internal administration of the journal. In the second study, we employ the visualization tool with a few improvements to five journals and two conferences in the field of e-learning to realize different research trends and patterns in the field of e-learning. The second study will allow novice and experienced educators, researchers and policy makers to understand what kind of different research areas exist in the field of e-learning, and to identify different research trends over the last six years using the visualization tool. Some results of the second study were also presented in a peer reviewed conference, and it received a best paper award. Similarly, the visualization tool was also presented as a possible solution of finding collaborators in the first workshop of Science 2.0 for technology enhanced learning.

The Chapter 4 of this dissertation deals with enhancing the peer review process of scientific communications. Peer review of manuscripts is considered as a basis in the advancement of any discipline. It ensures quality and reputation of scholarly communications. However, different kinds of conflict of interest (COI) situations can compromise the review decision. Current COI detection systems primarily rely on the co-authors network, inferred from the publicly available bibliographic databases as an implicit measure of social and collaborative relationships between researchers. However, different citations relationships have also been claimed to be indicative of various social and cognitive relationships between authors. This can be useful for improving existing COI detection techniques by highlighting those hidden relationships that can not be handled by traditional systems. To prove our hypothesis, in this chapter, we first present the potential of different citations relationships to highlight the existence or non-existence of social relationships between authors. In this context, we use basic citations relationships, i.e., co-citations, bibliographic coupling, inter-citations, and temporal information associated with these relationships as features to predict the social networks of our selected sample of authors. Our experiments show that

our defined features identified these social networks as best with 0.80 precision and 0.99 recall for non-sparse data and with 0.79 precision and 0.05 recall for sparse data. In the second part of the fourth chapter, we present citations as a measure of different cognitive COIs between researchers. We use co-citations, bibliographic coupling, and inter-citations to compute the cognitive relationships between any two authors. We discuss possibilities to assign weights to these cognitive relationships to reveal the strength of cognitive COIs. As a case study, we computed weighted cognitive relationships between the authors and reviewers of WWW2006 conference performance track. We found several cases where authors and reviewers do not have any apparent social relationships but they have strong cognitive relationship between each other. From this situation one might get the impression of the existence of cognitive COI between authors and reviewers. For practical applications, however, the severity of these cognitive COIs can only be identified by assigning contexts and sentiments to the cognitive relationships. In this context, we use and identify different contexts and sentiments that can be assigned to these cognitive relationships. In literature, researchers have always tried to assign contexts and sentiments to inter-citations and only to a single case of co-citations, i.e., "alternative or competitive work". In this chapter, we present a scheme based on existing theory to assign sentiments to even bibliographic coupling. Moreover, we use extended set of contexts and sentiments for co-citations. We also report our experiments for automated prediction of context and sentiments associated with any cognitive relationship for our WWW2006 authors and reviewers. As we used extended scheme for co-citations context and sentiments. We defined various features that can be used for the automated prediction of these contexts and sentiments. Finally, we assigned the context and sentiments to our selected authors and reviewers with very high cognitive relationships to reveal the severity of cognitive COI between them. Although in our reported results we did not find any severe case of cognitive COI, but we believe that such analysis might help in other situations.

Citations are usually considered as a unit of analysis in the field of bibliometrics and scientometrics to evaluate the scientific and intellectual impact of individuals, journals, nations, and research organizations [Garfield, 1970, Garfield, 1972, Oppenheim, 1997, Bormann and Daniel, 2008]. However, according to the literature, it represents a partial view of articles usage, as readers do not necessarily cite all papers they read. Similarly, articles can take some time to get citations, which makes them unsuitable to evaluate the impact of articles as soon as they are published. However, with the availability of online electronic journals a new criteria for evaluating the impact of articles is emerging (although at present it has not replaced the citations measure). This criteria is the download frequency of articles. The count of articles downloads has the

potential to implicitly provide a timely measure about articles usage as soon as they are published. Moreover, various studies have shown that there is a significant positive correlation between the number of articles downloads and their future citations. This suggests that download counts have the potential to anticipate in advance future citations for an article! By keeping in view the importance of downloads, the Chapter 5 of the dissertation presents various local and global attributes that are associated with a manuscript to determine its current value. More specifically, we explore the possibility to predict the download counts of a manuscript in the digital library of an electronic journal to reveal the current value of an article in terms of its future usage, and implicitly in terms of its expected future citations. In this chapter, we use articles from Journal of Universal Computer Science (J.UCS) for our detailed prediction experiments. In this context, we defined various novel features extracted from J.UCS articles and external bibliographic databases and evaluated their performance in predicating the future downloads of articles published in J.UCS. Moreover, we used only prior features which are available at the time of publishing an article. By using only prior information about articles, we can timely evaluate their future performance. Our experiments show that our selected features helped us in reducing the mean absolute error up to 13% as compared to the defined baseline error. In cases, where the volume of submissions is increasing rapidly, such analysis can help in facilitating an initial review or pre-selection of manuscripts by the administration before delegating it for more rigorous review by the experts in the field. Moreover, it can also help in identifying the factors that must be included in an article to increase its visibility or reading.

Finally, the dissertation closes with some conclusions and potential work for the future.

2

Applications of Mash-ups for a Digital Journal*

The World Wide Web has been experiencing a revolutionary growth due to numerous emerging tools, techniques and concepts. Digital journals thus need to transform themselves to cope with this evolution of the web. With their growing information size and access, conventional techniques for managing a journal and supporting authors and readers are becoming insufficient. Journals of the future need to provide innovative administrative tools in helping its managers to ensure quality. They also need to provide better facilities for assisting authors and readers in making decisions regarding their submission of papers and in providing novel navigational features for finding relevant publications and collaborators in particular areas of interest. In this chapter, we explore an innovative solution to address these problems by using emerging Web 2.0 technology. We explore the application of mash-ups for J.UCS - the Journal of Universal Computer Science. J.UCS can then serve as a model for contemporary electronic journals.

2.1 Web 2.0: A Note about the Terminology

The term Web 2.0 was coined by Tim O'Reilly [O'Reilly, 2007] to describe the revolutionary growth of the web by differentiating between old and new generations of websites. Although the term appears to represent a new version of the web, but it does not refer to its technological developments. It rather refers to the fundamental mind shift of people to actively participate and collaborate in the generation of new content, structures and services on the web [Davis, 2007]. Previously, the information on the

*The material presented in this chapter is based on a previously published paper [Khan et al., 2008]

web, by and large, was created and maintained by professionals, companies and organizations in form of personal homepages, e-commerce services, news, academic websites, etc [Kolbitsch and Maurer, 2006]. The lack of infrastructure, technical knowledge and simplified tools prevented many users to participate in the development of new web content [Lindhal and Blount, 2003]. However, the technological advancements of web in recent years have enabled the development of many novel applications. Now, web applications that are based on Web 2.0 concept allow users to do more things than just viewing the information or placing an order for a product. They provide an infrastructure and simple tools to encourage users to participate and collaborate in creating, storing, and disseminating information to add value to the website as they use it [Paul, 2005]. As a result, the web has turned into a powerful social computing platform supporting the extensive growth of e-communities by involving masses in creating and sharing contents [Kulathuramaiyer, 2007]. Many new community-driven services such as wikis, blogs, file sharing and podcasts are gaining influence rapidly [Kolbitsch and Maurer, 2006]. Blogs or weblogs for instance allow any user to become a publisher of content, sharing thoughts and information about any event or issue. Readers can write comments in their own blogs, resulting in a massive global interconnected community, facilitated through technologies such as permalinks, trackbacks and RSS [Efimova and de Moor, 2004].

Wikis in general are collaborative websites, where any user at any time can edit existing webpages and add new documents. The fundamental idea behind wikis is to engage a vast number of users to read and edit the contents of a document to complete or contrast it over a period of time [Kolbitsch and Maurer, 2006]. However, for these systems to spread among masses, proper marketing and users gratification are required [Szybalski, 2005]. Similarly, the pioneer users must generate sufficient material in the system to motivate prospective users to participate in adding value to the content [Kittur et al., 2007]. One of the largest wiki to date is Wikipedia, developed by Ward Cunnigham [Wikipedia, 2007]. The purpose of Wikipedia is to exploit the collective wisdom of masses in developing a concrete, free and comprehensive encyclopedia. Due to open editing nature, Wikipedia has also been widely criticized for various problems such as: vandalism, bias [Priedhorsky et al., 2007], quality assurance, and edit wars [Kolbitsch and Maurer, 2006]. Much of the users' roles have been defined under Wikipedia to develop valuable contents. These roles include readers, administrators, recent changes patrollers (for reverting vandalism), bureaucrats, stewards, reviewers, and editors, etc [Wikipedia, 2011a], [Wikipedia, 2011b]. Various other proposals have also been introduced to cope the problems. One example of such measures that has been implemented is the "featured articles" that undergoes a

thorough review process after being nominated by the community to meet the highest standards [Wikipedia, 2011a]. Another example of suggested measures is the reputation algorithm of authors which assigns a quality level to the article based on the reputation of the participating authors [Adler and de Alfaro, 2007]. Similarly, the authors in [Kolbitsch and Maurer, 2006] suggested a hierarchical wiki to ensure quality of content. Another wiki-based encyclopedia that is growing rapidly is Austria-Forum [Austria-Forum, 2011]. However, it is restricted to topics that are related to Austria. The Austria-Forum contains approximately 180,000 objects, which includes text, audio, video, and images, and is expected to contain more than a million entries by the year 2013 [Maurer and Mueller, 2011]. This forum also tries to overcome some problems faced by Wikipedia, e.g., editorial control (thus making it cite-able in scientific contributions), time-stamped entries, search by meta-data, and books that can be annotated and can be linked to other information within Austria-Forum and outside [Maurer and Mueller, 2011].

Podcasting, videocasting, and photocasting facilitates the blogging and sharing of audio, video, and photos by millions of users across the network [Kolbitsch and Maurer, 2006]. The topics covered by these broadcasting services ranges from music, entertainment, marketing, politics, and travel to education [Kolbitsch and Maurer, 2006]. Similarly, other community based applications such as Flickr, YouTube, Del.icio.us allow users to tag and share common interests, photos, video clips, and bookmarks.

The above mentioned community-driven paradigm has produced enormous contributions to the web with its socially generated contents. By engaging millions of users on the web, massive collaborative projects have been achieved, e.g., YouTube, Wikipedia, Flickr, and many more. There is an imminent need to provide mechanisms to harness this collective intelligence of individuals, and to design appropriate solutions to address various concerns, including novel services in the area of digital libraries [Kulathuramaiyer, 2007]. According to [Casey and Savastinuk, 2006], the ever growing, diverse and heterogeneous requirements of library users can be better satisfied by giving them participatory roles in enhancing library services and by allowing them to customize services according to their own needs. Various exciting Web 2.0 tools and services such as RSS, wikis, blogs, and tagging are now becoming part of library services in bridging the gap between users and information [Shri et al., 2010]. This chapter is also one such effort where we explore the applicability of an emerging Web 2.0 paradigm called mash-ups for a digital journal, namely the Journal of Universal Computer Science (J.UCS).

2.2 Web Mash-ups: One of the new Web 2.0 Paradigms

A mash-up is an emerging Web 2.0 paradigm that allows anyone to combine pre-existing data or information which is accessible through a public interface or API from sources like Amazon, Google, Yahoo, eBay, etc., in innovative ways, enabling people to access customized information that matters to them in a meaningful way. As blogs and wikis allow anyone to become an author, mash-ups facilitates rapid web applications development by allowing anyone to combine existing data or application functionalities from multiple sources in a single web environment [Kulathuramaiyer, 2007]. Thus, the creative energy of many people can be combined to address different users' requirements. The availability of various technologies at presentation (e.g., HTML, CSS, JavaScript, Ajax), data accessibility (e.g., SOAP, REST) and data handling (e.g., XML, KML) levels have made web an appropriate place for the development of such mash-ups. The providers of the APIs through which contents are accessible are usually referred to as mash-up enablers, while the mash-up builders who utilize the data provided by enablers are referred to as mash-up assemblers [Watt, 2010]. In [Kulathuramaiyer, 2007], the author has described the presence of different types of mash-ups such as mapping mash-ups, time-line mash-ups, meta-search mash-ups, image-based organization mash-ups, etc.

Programmableweb [Programmableweb, 2007] serves as an important resource for the categorization of information and analysis of an evolving collection of interesting and useful mash-ups. Although the lack of design tools puts a limitation to the development of mash-ups by ordinary end users, according to Programmableweb on average 3.45 mash-ups appear everyday. However, efforts are being made to fill this gap, and to facilitate non-programmers to make useful mash-ups with little manual effort [Yahoo Pipes, 2010, IBM Mashup Hub, 2010] and in fully automated way [Fischer et al., 2009].

The authors in [Hoyer and Fischer, 2008] broadly classified mash-ups in two categories, i.e., consumer and enterprise mash-ups, and provided a comprehensive market review about their development tools. During the last couple of years, the potential of mash-ups as a flexible and rapid solution to the heterogeneous and diverse needs of individuals has been realized in both corporate enterprises [Janner et al., 2009, Hoyer and Fischer, 2008] and e-learning domains [Chatti et al., 2009, Taraghi et al., 2009]. According to [Hodgins, 2008], various learning resources can be created using the time-line mash-ups (a mash-up that adds time or time-lines to other data). The availability of time-stamped encyclopedias like Austria-Forum provides a great opportunity to create such novel resources to assist in learning

process. For example, it can help in viewing the pictures taken at different dates to observe a change in a city or river, and different essays or points of views about a topic over the period of time [Maurer and Mueller, 2011]. Recently, a non-profit consortium called "Open Mashup Alliance" has evolved to promote the adoption of mash-ups solutions for enterprises by incorporating evolving enterprise mash-up standards, i.e., "Enterprise Mashup Markup Language (EMML)" [The Open Mashup Alliance, 2010]. A recent report from Business Wire has predicted that the usage of enterprise mash-ups is expected to grow upto $1.74 billion by 2013 [Business Wire, 2010].

According to Programmableweb, about 50% of the overall mash-ups are geographic (map based) mash-ups. Mibazar.com [Mibazaar, 2007] illustrates the variety of Google-map based mash-ups that provide information on education, history, top celebrities, shopping, events and much more.

One example of such a mash-up is used by the New York City Coalition against Hunger. It is an organization which is engaged in solving the hunger problem of the New York city. By using a combination of Google Maps' free application programming interface, geographic information from ArcWeb, and its own information about the locations of city's soup kitchens, the organization was able to build an online map of the city's charitable food providers' locations along with their contact information. As a result, the soup kitchens in the city have a fast and easy way to find each other, in order to coordinate work to solve a common problem [TechSoup, 2007].

A mash-up for a digital journal can serve as a means of providing rich collaborative platform to support diverse users' needs about high level access options to the e-collections of a journal, and as an administrative support tool to facilitate rapid expansion of the journal. In this chapter, we explore a pioneer example of such a mash-up for J.UCS by combining the J.UCS digital repository with Google Map APIs to address various users and administrative concerns described in the next section.

2.3 Major Concerns of J.UCS

The Journal of Universal Computer Science (J.UCS) is an open access, high quality, peer reviewed electronic journal having more than 1000 publications since 1994. J.UCS has been published in volumes (one per year) with at least 12 yearly issues. It has more than 300 high profile editors across the world enabling a broad coverage of all aspects of Computer Science [J.UCS, 2007]. At the end of each year, a volume of J.UCS is published as a printed copy and archived [Maurer and Schmaranz, 1994, Calude et al., 1994]. This printed version exactly matches the electronic version with

the same pagination as the online edition. The electronic versions are thus static documents, frozen over time. This practice of having print equivalent for an electronic journal has also been suggested by McElroy [McElroy, 2002].

Since its beginning, J.UCS has introduced and incorporated a number of novel ideas to support its readers, authors, editors and administration, e.g., the annotation feature. An annotation refers to a note or a comment about an existing publication informing readers about new research results or errors. Annotations in J.UCS are applied subject to an evaluation based on a refereeing process and are only added if deemed appropriate. Verified annotations make it possible to insert only objective comments and prevent the misuse of annotations for personal disputes. Other features of J.UCS include full-text search of its contents, fast access to published papers, and better contents management using Hyperwave [Maurer and Schmaranz, 1994]. Furthermore many new features and ideas are being explored to be incorporated into J.UCS, such as mash-ups [Kulathuramaiyer, 2007] and links to the future [Maurer, 2001a]. The original idea of links to the future and its design via the application of annotations together with other novel ideas were presented in [Maurer, 2001a]. Additional ideas about links to the future were then discussed in [Krottmaier, 2003] and [Afzal et al., 2007]. The "links to the future" ability maintains the static nature of papers while dynamically accumulating related works via annotations [Afzal et al., 2007].

Due to the growing information size, readership, authors and editors, the management of J.UCS has experienced some administrative and users concerns in ensuring its long-term sustainability. The major concerns addressed with regards to the users of the journal include concerns of management with respect to effective administration as well as potential concerns of authors and readers. We will first discuss the management concerns which emphasize the importance of employing mash-ups to support administrative tasks.

2.3.1 Management Concerns

Geographical Distribution of Papers

In expanding the publications of J.UCS it is necessary to gain insights and information about the current state of readership and accessibility of the journal. For example, there is a need to determine which locations (cities and institutes) are contributing less, more and which have stopped contributing. This information can be extremely useful for a number of reasons. For instance, it allows the determination of coverage patterns and popularity of certain topics. It can show places from where papers have

never been submitted or places from where papers were coming in the past but the stream of submissions from these locations has turned into a trickle.

Furthermore, the administration of J.UCS also wants to know the trends of publications in a particular research area or location. It could also help the management in the analysis and review of proposals for special issues of J.UCS.

Geographical Distribution and Coverage of Editors

In principal the administration of J.UCS would like to have an even distribution of editors across the world with a balanced distribution across topics in order to sustain the journal's claim of being a truly universal and international publication instrument. In particular, there are potential intrinsic distributions patterns that need to be uncovered in enhancing the image of journal. To illustrate this point we consider the following scenario; if all editors in a particular research area come from a specific region, it could lead to the perception about the journal being too restricted or even been seen to be biased (especially if there are political or religious aspects are involved between editors and prospective authors, for such examples see, [Triggle and Triggle, 2007] and [Godlee and Dickersin, 2003]). The management thus needs to find out about such patterns and to ensure the balanced representation of editors across the globe in each research area.

Determining Groups of Authors and Editors

J.UCS adopts a unique approach regarding the review process of submitted papers. When a submission is received, it is forwarded to all the members of the J.UCS editorial board (this policy, however, is subject to change in near future where the papers will be submitted only to the relevant editorial members in a given field of study). If at least three editors sign up for a review of the paper, the submission is accepted and the review process is initiated for the paper. If the required number of editors is not adequate to review it, the paper is forwarded again for the second time to all the editors. If still there are not enough editors signed up for its review, the managing editor then asks the author of the paper to nominate six editors (possibly outside of J.UCS). The managing editor then selects 2-3 editors from the nominated list of editors and asks them to proceed with reviews. The managing editor may also further request these specialists to become members of the editorial board of J.UCS: all editors have to be at least tenured Associate Professors or equivalent, and have to have a substantial publications record in reviewed journals. This approach distinctly differs from the approach employed in other journals which usually follow a different

procedure: papers submitted are sent for reviews directly to selected editors, based on their area of expertise.

Although J.UCS has this unique refereeing process with its own benefits (fast review process, reduced management burden for manager editorial), this approach may be subject to abuse or may lead to the grouping of editors and authors. Such a grouping may then affect the quality of the journal. In Section 2.8.3, we demonstrate a combined mash-up of authors and reviewers to avoid such location based conflict of interest situations.

Although we do not believe that such situations are currently happening, we highlight here the importance of dealing with this issue. As the journal expands, it will become even more difficult to ensure that such phenomena do not happen. This issue will also need to be taken into consideration by other journals even employing alternative reviewing procedures.

Selection of Special Issues

In order to expand its coverage and focus on upcoming issues, J.UCS also publishes special issues periodically. J.UCS maintains a well defined guideline for special issues to get accepted and published. For example, to ensure a high international standing, special issues with a large number of papers coming from the same country are not permitted.

2.3.2 Authors and Readers Concerns

Determining Collaborators

As in any particular discipline the scientific research depends on extensive social interactions or scientific collaborations [Kraut et al., 1988]. The readers or potential authors may then be interested to know the affiliations of other authors in a particular research area in order to collaborate with them.

Determining Distribution, Coverage and Expertise of Editors

The authors and readers may also be interested to know whether particular research areas have indeed high quality referees with an even geographical distribution. If editors distribution in a field are seen to be local or having an ethnic, religious or other forms of bias, potential authors may consider to be better off to submit papers elsewhere. This is exactly why the composition and distribution of the editorial team is one of the main concerns of the administration of J.UCS as mentioned earlier.

Navigation through Digital Publications

Navigating an electronic journal's contents is typically performed by browsing publications either via a directory listing of journal papers, searching facility (by author, keywords, title, full text etc.) for related papers, articles by topic, articles by authors, co-authors or other papers from the same authors, etc. Alternatively, readers or authors may require a visual representation of the publications for a specific region and research area. A visual display enables a quick overview with minimal efforts.

In an effort to address these issues in making J.UCS more manageable and to enable it to become a still more user centered digital journal, we explore a variety of solutions. In this chapter we explore the use of mapping mash-ups (Google Map based) as a novel tool to harness the power of community dynamics in expanding J.UCS.

In the next section we describe the mechanism we adopted in extracting the J.UCS meta-data about publications, and the techniques adopted to refine and resolve the location related information of authors.

2.4 Information Extraction

J.UCS documents collections (papers) and meta-data about these documents are stored and managed using the Hyperwave system [Hyperwave, 2007, Maurer, 1996]. The Hyperwave system is an information management system that allows organization of knowledge and information in a sophisticated way. Publications in J.UCS have a corresponding meta-data (XML) file (one for each paper) that contains all the information about that paper such as title, authors, institutions, cities, countries, keywords used in paper, area of research, date of publication, volume and issue number, etc. In the development of the mash-ups, meta-data about papers published was captured from the Hyperwave server using Hyperwave APIs and was represented in a relational database.

Preliminary experiments revealed that the location information of a large number of the authors was found to be missing, incorrect or not compliant to the standard format. In the next section we will discuss the approach employed to extract the correct location information.

2.4.1 Variations in Representation of Locations

The location data of authors first needed to be cleaned, as the data contained spelling mistakes (e.g., Unted Kingdom), variations in forms of data which includes abbreviations (e.g., US, U.S., USA, U.S.A, United States, United States of America; Zuerich, Zürich.)

In order to rectify and standardize the names of the countries and the cities, we employed the GeoBytes database [GeoBytes, 2007] (containing the cities and countries information from around the world) to verify the city and country information of authors. Locations that did not have any match in the GeoBytes database were then identified. We manually constructed a database of mismatched words and associated them with the intended city and country names. As this information should have actually been verified at the point of data entry, the mash-up system developed could then serve as a tool for verifying data acquisition in the future. This derived geo-data names database will be applied to verify non-standard or incorrect information of locations entered in the J.UCS repository.

2.4.2 Determining Unknown Location Information of Authors

The difficulty in ascertaining the exact city-location of institutions has been described in [Kulathuramaiyer, 2007]. The institution name and corresponding country information for all the authors were available in the meta-data files. However, a large number of authors did not specify the corresponding city information (where their institution was located). Although it may seem a trivial task to automatically place an institution in its rightful city location, this had proved to be a challenge. Firstly, we were not able to acquire a comprehensive publicly available database to automate the task of matching institution names directly to their respective cities. Furthermore there are also numerous institutions having multiple campuses or localities, e.g., University of California has campuses in different cities. Without the information about the city where an institution resides, it became a challenge to determine the geo-code for the institution (to place on a map).

We first tried to discover the location information by using Google MAP APIs and Search APIs. The information provided by Google MAP APIs was much too restrictive (even compared to the Google search APIs). The database associated with the MAP API tends to be a much smaller subset. Searching for the cities of the institutions through Google Search APIs was again complicated because it returned numerous results for a single query. It was also difficult to extract the name of the city from the

HTML pages returned, because some pages did not contain the name of any city or returned multiple city names.

An alternative way to identify the locations of authors was proposed in [Klerkx and Duval, 2007] by using the domain names of authors' email addresses. They employed a domain name look up service MaxMind [MaxMind, 2007] having 128,000 domain names registered in their database. This approach was again problematic for a number of reasons: it only works for institution-specific email addresses. It will not work for hotmail or gmail addresses. The database is also limited in terms of its coverage of institutions, e.g., pwr.wroc.pl, domain of the Wroclaw University of Technology in Poland, was not found. Additionally, location of authors belonging to an institution having multiple branches such as Microsoft, larger universities, virtual universities, could not be located precisely. This in a way has justified the need to build our own internal institution-city database.

We then devised a heuristic approach to extract the information of missing locations. The first effort was to construct an internal database in mapping of institution-city locations. This database is subsequently used in checking for city information of institutions whose location was not known. The proposed approach first parsed the institutions names to identify whether there was a city name found within the institution name. This could be done by looking for a string in the institution name that matches a city in either our internal database of cities or the database of cities and countries (including longitude and latitude) collected from GeoBytes. For example the institution name "Technical University of Graz" contains the name of the city in which it is situated, i.e., Graz. In this approach the country name was applied to disambiguate the cities names existing in more than one country.

In the original database, only 47% of the authors specified exact location information where city and country data was known. By applying the above mentioned approach, we were then able to map out the locations of 86% of the authors. Manual inspection revealed that the remaining 13% of the authors had stated country information but city information was still found missing and 1% of the authors had provided no information about both city and country.

Although the above mentioned approach had not solved our problem completely, the results have been satisfactory. It should be noted that using techniques that employ pattern matching alone, it is difficult even for a human to be sure of the precise location of an author, e.g., Washington Business School of North Virginia is not in Washington but it is located in Vienna, Virginia. In order to create a platform for users to provide feedback and verify the locations, we decided to employ a mash-up. This mash-up visualizes the location of each author together with his or her name and

institution name, in order to collect feedback about locations that are not confirmed or not precisely located. We have therefore resorted to a community centered approach to verify the contents of the journal.

2.5 Mash-up as a means of Community Assisted Content Management

The main interface for the mash-up with community feedback is shown in Figure 2.1. This mash-up allows users to view the distribution of authors or editors across the globe. The user also has the option to browse across volumes or topics. Each marker on the map represents the collection of either editors or authors from a particular location; its size represents the number of authors or editors. The authors with missing city information were visually positioned according to the geo-code of their respective country. It has to be noted that the location information is based only on the institution from which an author published a paper (This depends on the institutional information provided in published papers). Information of current affiliation of authors will however need to be maintained in a separate user profile database. A future work in this direction can explore user profile mapping to support a community of users further. The Figure 2.1 and Figure 2.2 illustrate the mash-up developed which also displays the statistics of authors or editors in each country. To assist the users, we distinguish between known cities (whose locations are confirmed) and unknown cities (whose locations are not confirmed). The user can click on any marker to view the details of each person in an information window referred to by a marker as shown in Figure 2.3.

The community (which includes the original authors and editors) can assist the administration of J.UCS in correctly relocating and profiling the attributes of any author or editor. The mash-up provides editorial support for data updating as shown in Figure 2.4. The data acquisition facility is further enhanced by allowing users to position the marker at a particular location for an author. This is useful in cases where Google Maps service failed to provide geo-code of a particular address such as in the case of “Tatsunokuchi, Japan”, which no longer exist as an independent town but now is a part of “Nomi” city after February 1, 2005 [Tatsunokuchi, 2010]. This capability of mash-up further facilitates to even capture the precise building information about an author and can be very useful in highlighting interesting patterns to answer various questions.

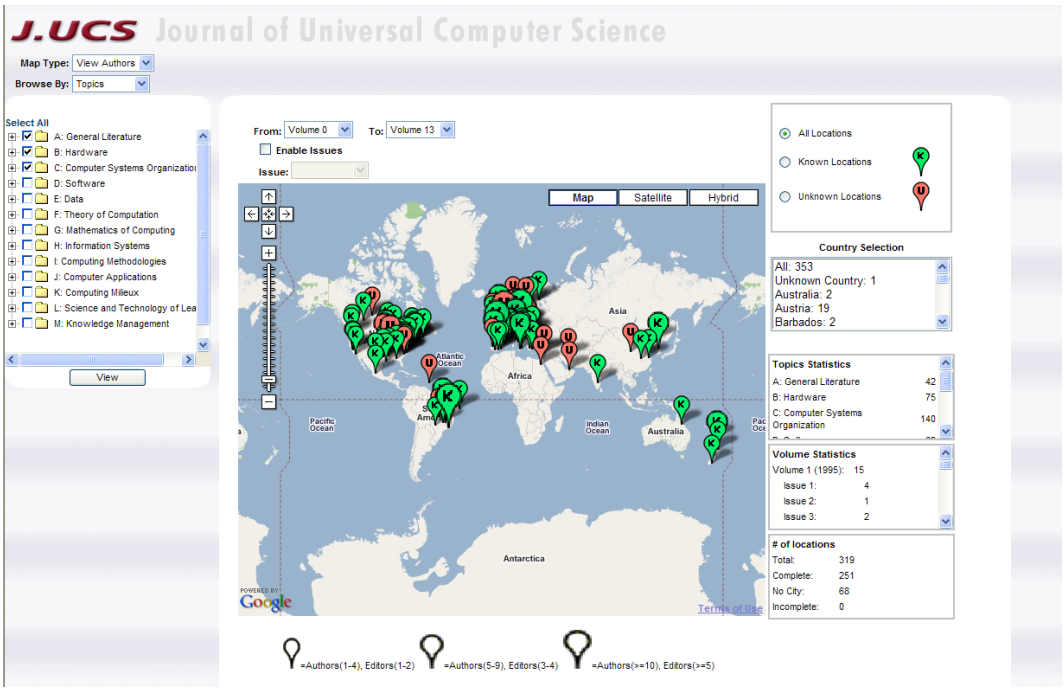


Figure 2.1 Main interface of the community feedback mash-up.

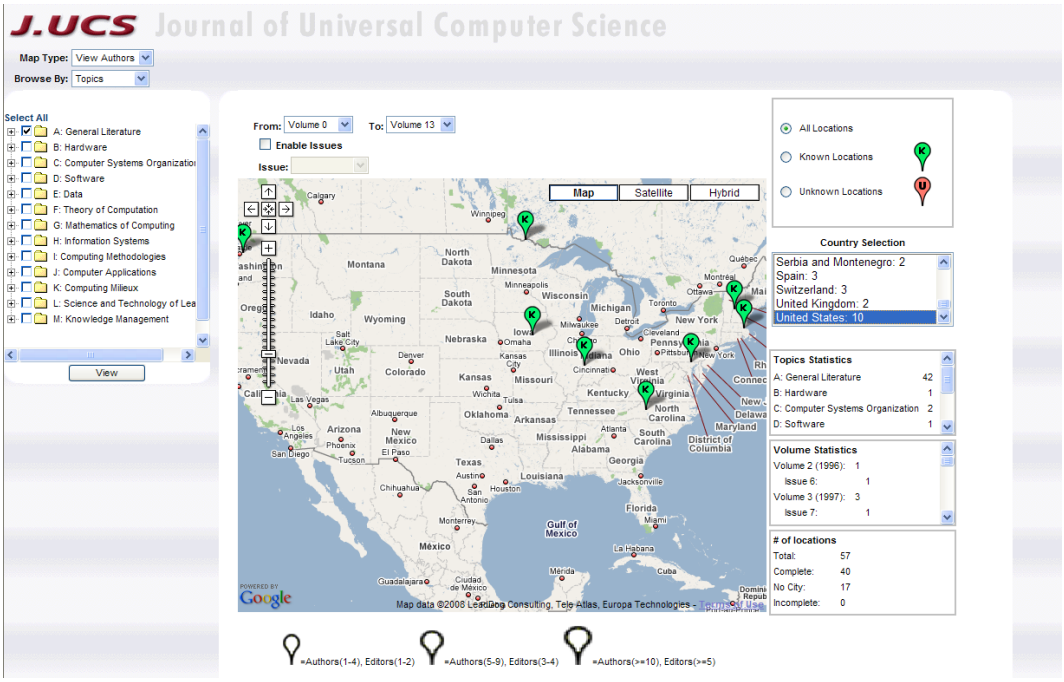


Figure 2.2 Distribution of authors filtered by country.

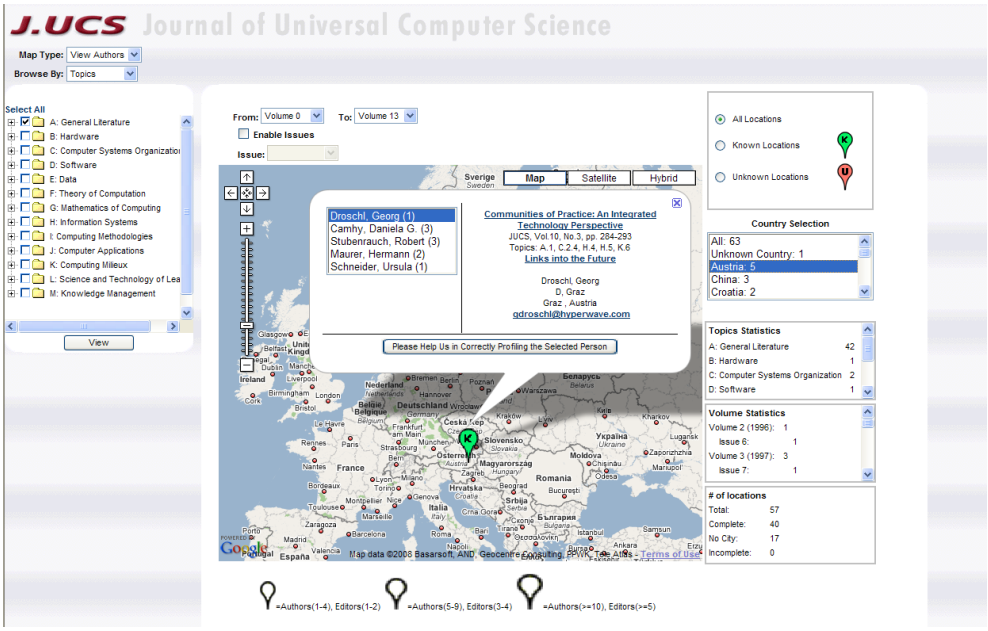


Figure 2.3 Information about authors.

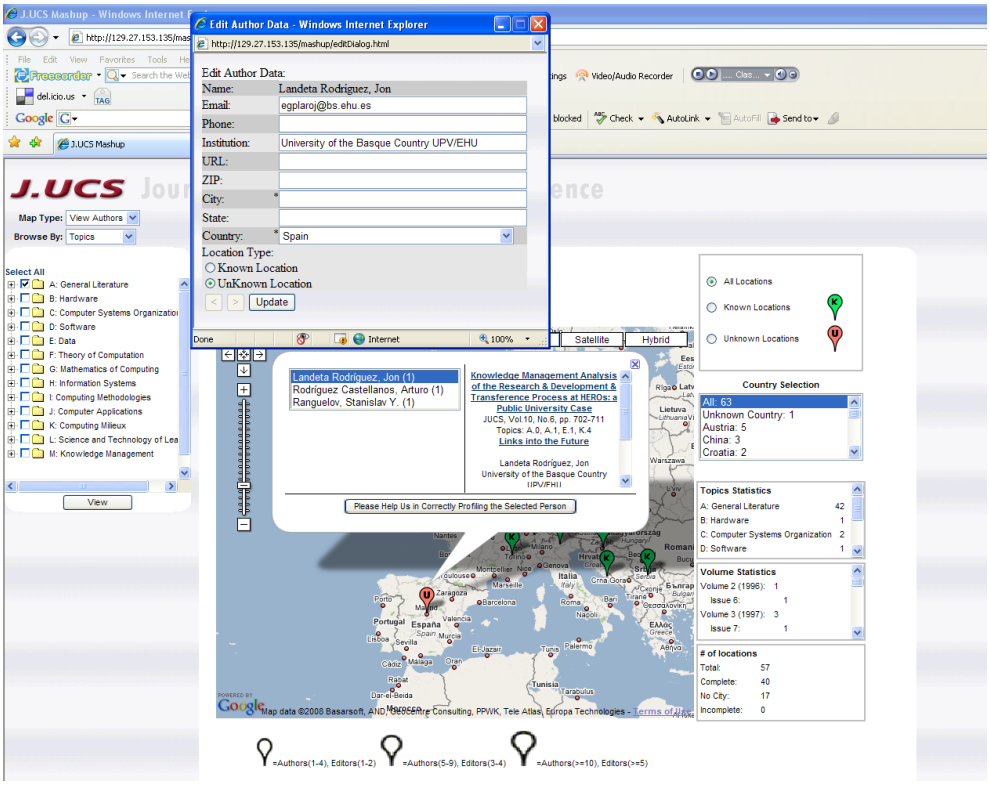


Figure 2.4 Updating of missing information.

As authors often tend to be in more than one location for example visiting scholars, job change, and there are some large universities, which span multiple locations (University of Arizona, Curtin University in Miri, Sydney as well as in Perth). The annotation tool provided for the community can be further enhanced to add new information about the profile of each author, and reflect these emerging historical realities using our mash-up to the general public (on author's wish due to privacy reasons) and to the administration of the journal for various administrative tasks. The links that have been used to show the linking between co-authors as shown in Figure 2.8 can also be used to reflect the movement of any particular author over the time.

The above mentioned steps that we followed to extract, standardize and augment J.UCS meta-data helped us in defining a generic architecture for J.UCS mash-up which is described in the next section.

2.6 System Architecture

In this section we describe a detailed architecture of the mash-up application as shown in Figure 2.5.

The meta-data information (XML files) about each publication in J.UCS is extracted by "Custom XML Files Generator" each time a new issue is published, and a separate customized XML file is generated. The "Repository Populator" application extracts the available information from the customized XML file with the help of XML parser to populate the "Mash-up Repository". The "GeoBytes Repository" provides the name of the cities, countries and their longitude and latitude information to standardize and augment the location information. The "Mash-up Application" is responsible to accept requests from users, extract the information from the "Mash-up Repository" and display it with the help of "Google Maps Service" using Google Maps APIs. The "Location Extractor Module" resolves the location of the authors (each time if there is a not-resolved location under the current selection of options), when the "Mash-up Application" is requested to show the authors distribution across the globe under the current selection of options in the main interface of the mash-up. The "Data Updating Module" takes the location and data updating information from the user and stores it back in the "Mash-up Repository". The same architecture is applicable for the profiles of editors in J.UCS.

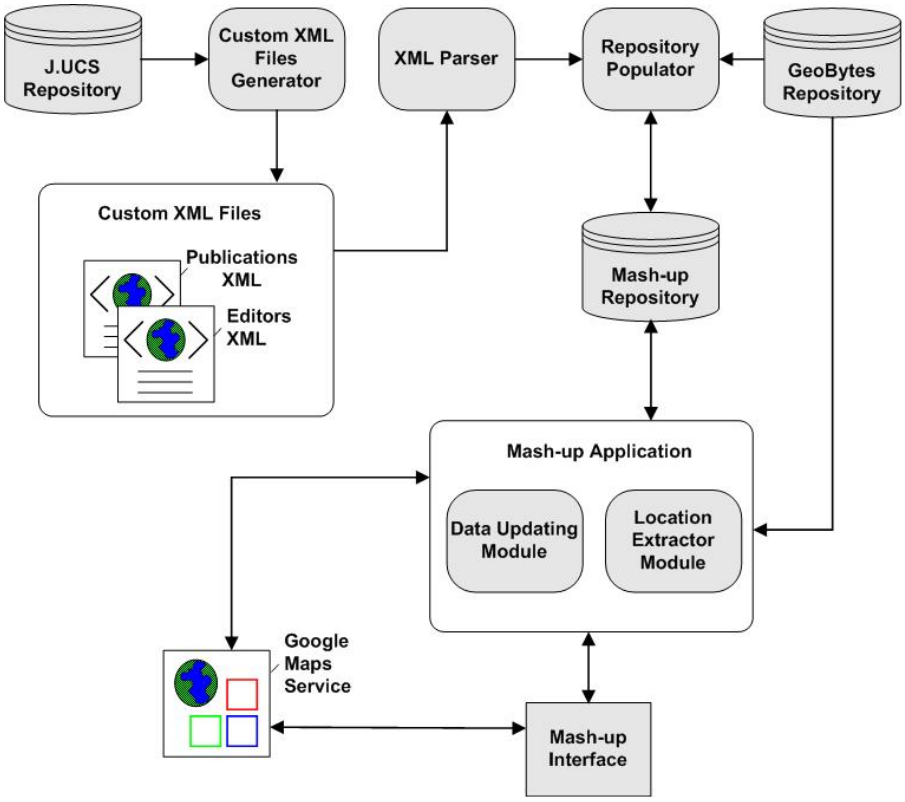


Figure 2.5 System Architecture.

2.7 Overview of Administrative Mash-up

The experiments carried out have helped to prepare the necessary data to be used for the development of decision making mash-up. However the over-reliance on community feedback results in the risk of human error or misinformation. We have thus incorporated an automated email alert facility to notify the administrator of the mash-up whenever an update is effectively made by any user. We have also incorporated an administrative mash-up to help the administrator to either accept or undo the changes made by users as shown in Figure 2.6. The administrator can also lock any information about any author or editor preventing users from modifying information that is known to be correct.

Based on our initial experiments, we have enabled the first administrative mash-up for an electronic journal. The features built into this preliminary mash-up have also been incorporated into the journal's website (http://www.jucs.org) for users to update their profiles and construct new profiles. Currently the administrative tasks are being performed by J.UCS office, however in future, these tasks can be further delegated

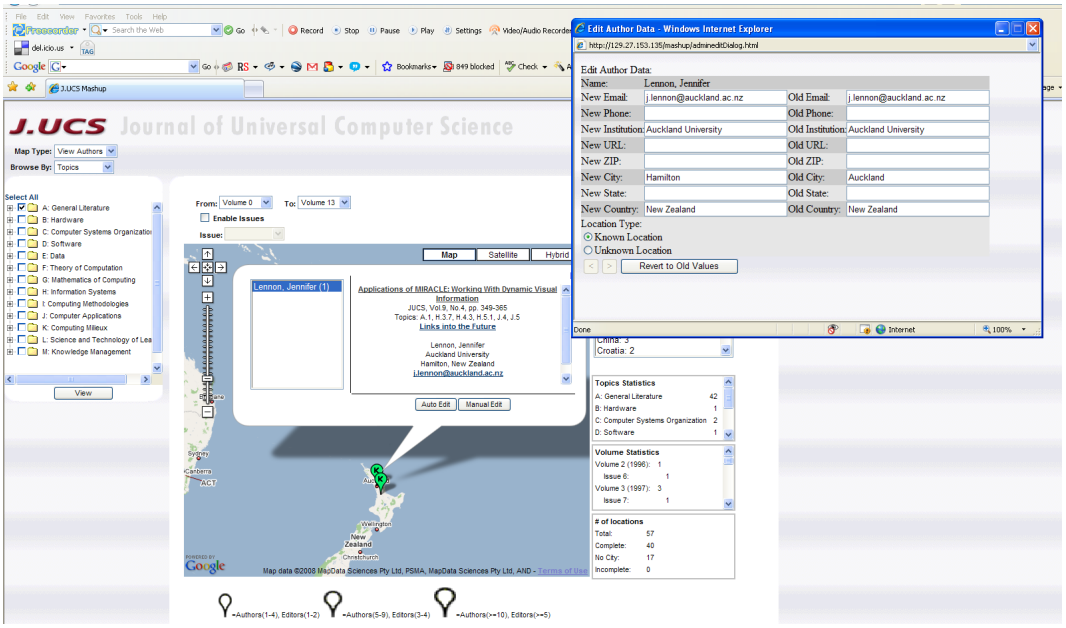


Figure 2.6 Undo of changes made by general user.

to the users of the journal through intelligent task routing [Cosley et al., 2007] and a hierarchal wiki [Kolbitsch and Maurer, 2006] with different levels of users' permissions.

The next section describes the resulting mash-up that can be utilized to address administration and user concerns as highlighted in Section 2.3.

2.8 Mash-up as a Decision Making Tool

2.8.1 Distribution of Publications

This interface in Figure 2.7 depicts the distribution of publications in J.UCS over the years across the globe. The user can select any volume, issue, paper and any topic or group of topics from the list of volumes and topics and is presented with a distribution of papers for a selection of volumes and topics.

The size of the marker indicates the number of publications at a particular location as opposed to the number of authors in user feedback mash-up (see Section 2.5), where the main concern is to profile the attributes of authors. Moreover it is also useful to show the number of papers at a particular location specially while comparing it with editors (see Section 2.8.3) instead of authors because the size of the markers representing the number of authors can not convey the number of papers written from

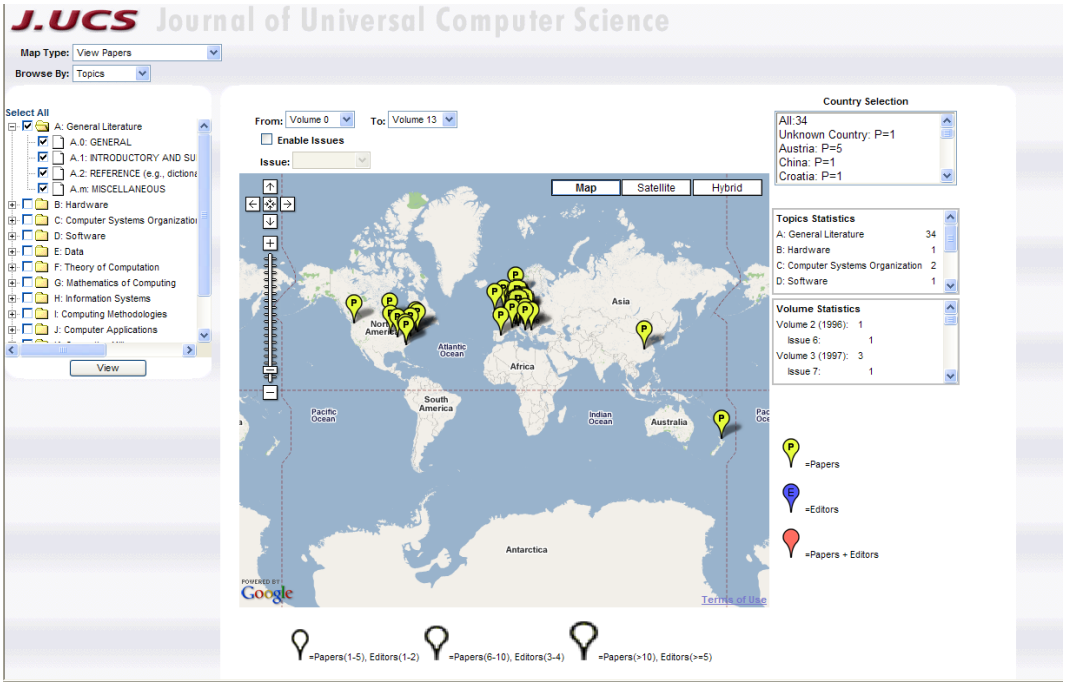


Figure 2.7 Interface for browsing by topics.

a particular location and hence can not be used to accurately highlight the possibility of bias.

In this application we have assigned markers for representing location information of the co-authors of all publications. Assigning markers to the locations of only the first authors will not provide a true picture of the distribution of publications across the globe. The visual markers as used in the current implementation do not correspond directly to the exact number of publications. Summarized information about papers is however provided in tabular forms. In our current implementation, a link has been provided to visualize co-author linking as shown in Figure 2.8.

This mash-up will help the authors, readers and management to know that publications in any particular research areas came from certain locations over a period of time. It will help the readers or researchers to navigate in a novel manner. The mash-up thus provides an additional visual entry point to the document collection of J.UCS (based on geographical locations of authors of papers). It can be useful in facilitating the development of a social network for scholarly community, by enabling researchers to find collaborators living nearby or across the world working in their field of interest.

It will also help the management to know from which region or institution the publications have never come or stopped coming over the years. As it can be seen in Figure 2.9, for the topic "Computer System and Organization" no papers have ever

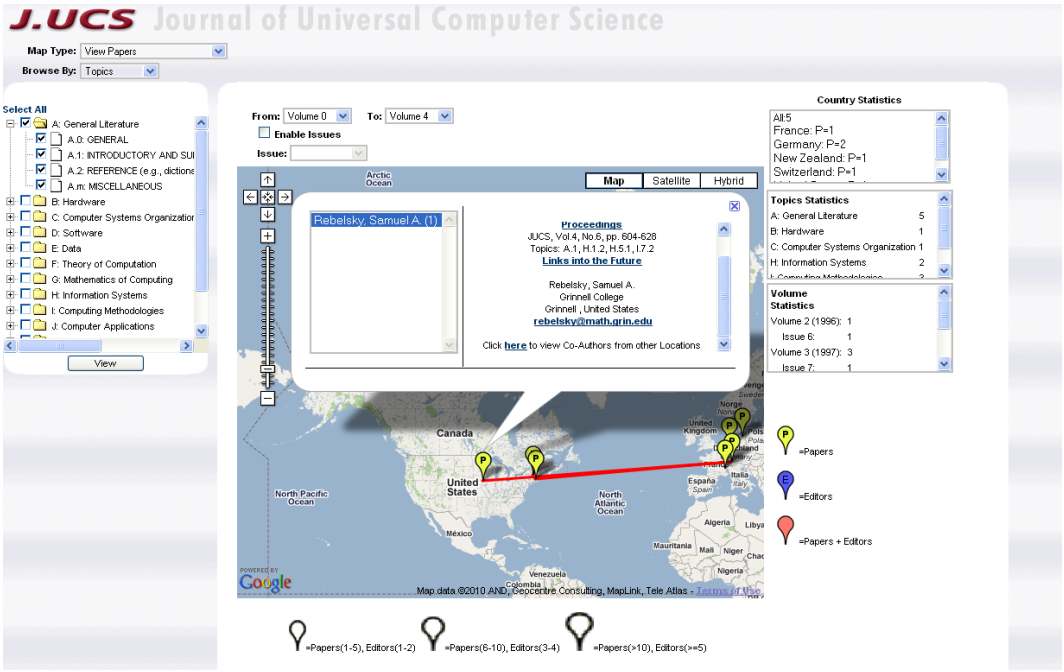


Figure 2.8 Co-Authors Linking.

been submitted from South Africa or Japan. There has only been one publication received from Australia so far. Figure 2.10 shows that for the topic "Software" the journal has published a single paper from Wollongong, Australia uptil volume 3 but has not accepted any newer publications.

A future work can also incorporate information about submission of papers to contrast with acceptance patterns as described in [Taylor, 2001]. The inclusion of information of submission together with time-line data may be used to indicate the efficiency of an established journal in its review process. The mash-up can also be applied in the selection of special issues to be published in the journal. As it can be seen in Figure 2.11, more than 90% of the publications in the special issue 9 of volume 12 came from a single country. In such a situation the managing editor is responsible to explicitly scrutinize the papers to ascertain the international standing of the papers.

As can be seen in Figure 2.7, by viewing the map at an abstract level some of the markers tend to overlap with each other, some locations are hidden behind the markers and some countries are too small to be visible in just a single view of the globe, e.g., European countries. To overcome this problem the users can zoom in to an appropriate level using the zooming and panning facility provided to get a more detailed view of the distribution of papers as shown in Figure 2.12.

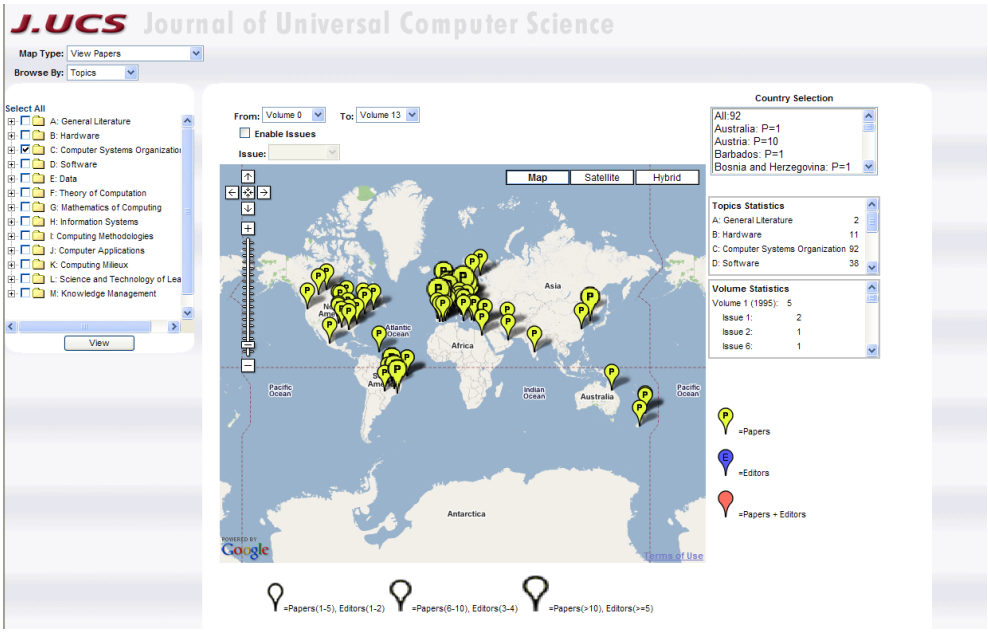


Figure 2.9 Distribution of publications for topic "Computer System Organization" (C) from volume 0 to volume 13.

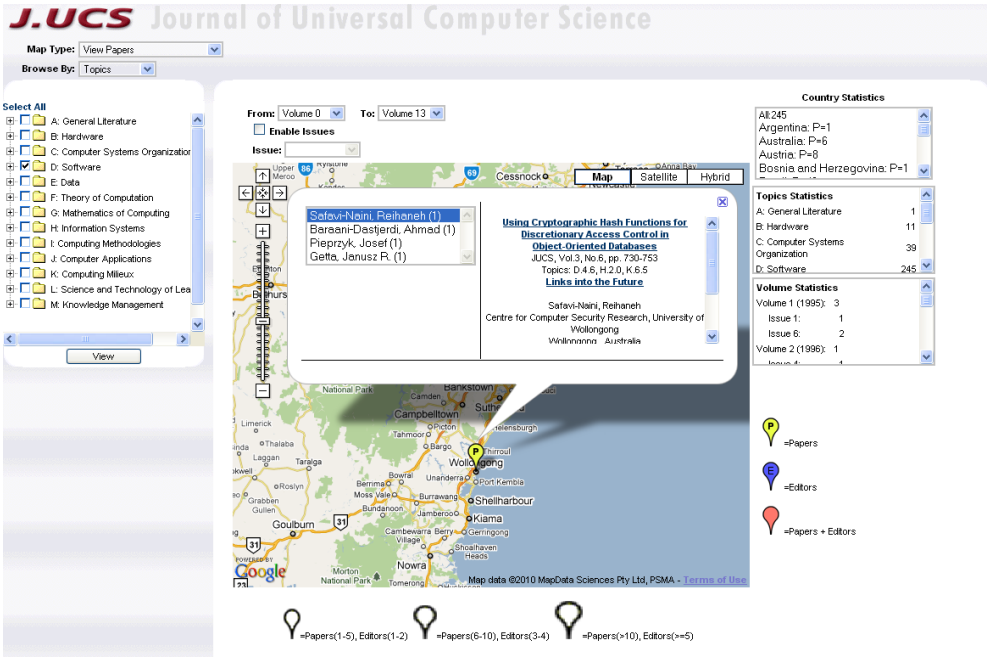


Figure 2.10 Sole publication from Wollongong, Australia for the topic "Software" (D) up till volume 13.

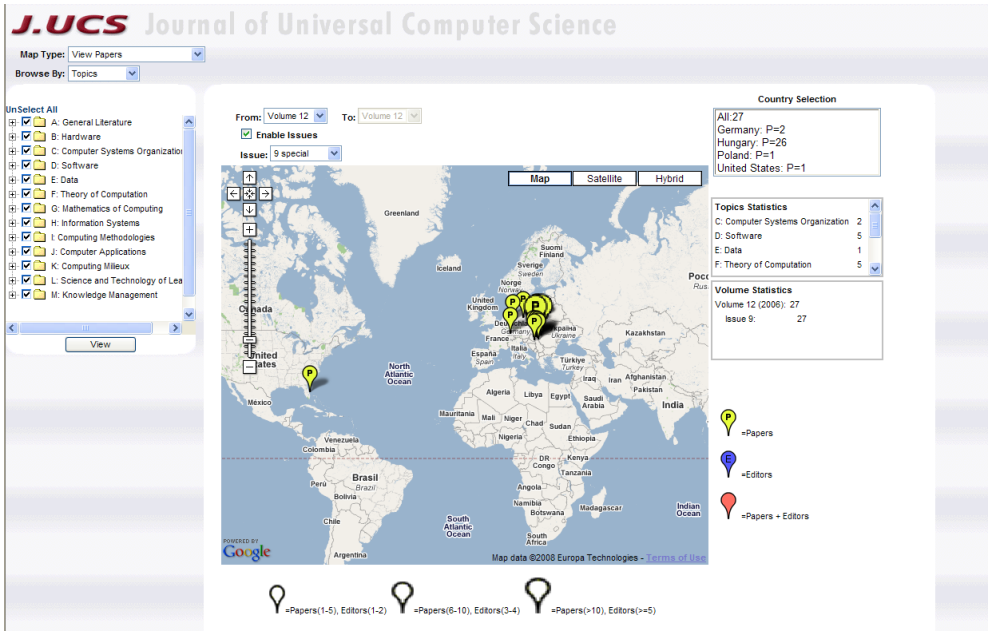


Figure 2.11 Distribution of all publications in special issue 9 of volume 12.

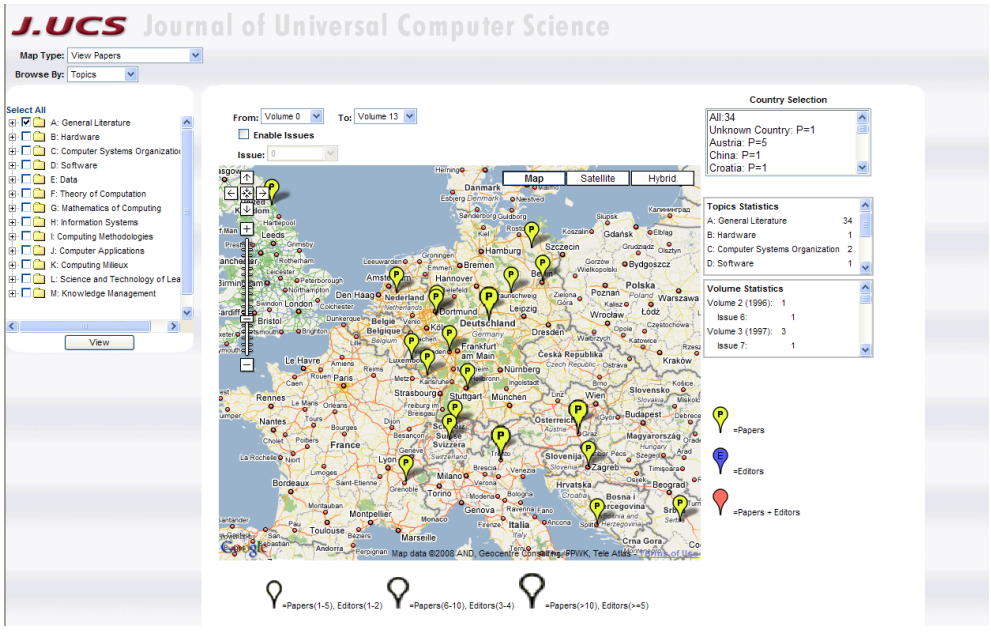


Figure 2.12 Distribution of papers across the globe at appropriate zoom level.

2.8.2 Distribution of Editors

Figure 2.13 depicts the topic wise distribution of editors across the globe. The user can select any topic or group of topics and is presented by the system, the distribution of editors across the globe under the current selection.

This interface helps the authors in making decisions regarding their submissions to J.UCS. The authors can view information about the editors in their area of research and their affiliation before deciding to submit a paper to J.UCS. It can also help the management to ensure the even distribution of editors across topics over the world. Figure 2.14 illustrates that, for the topic "Social Issues" (K.4.2) there are three editors from Germany and there is one editor from Australia. The management will then need to take appropriate action to expand the number of editors in such topic and thereby ensuring a uniform representation of editors across the globe in each research area.

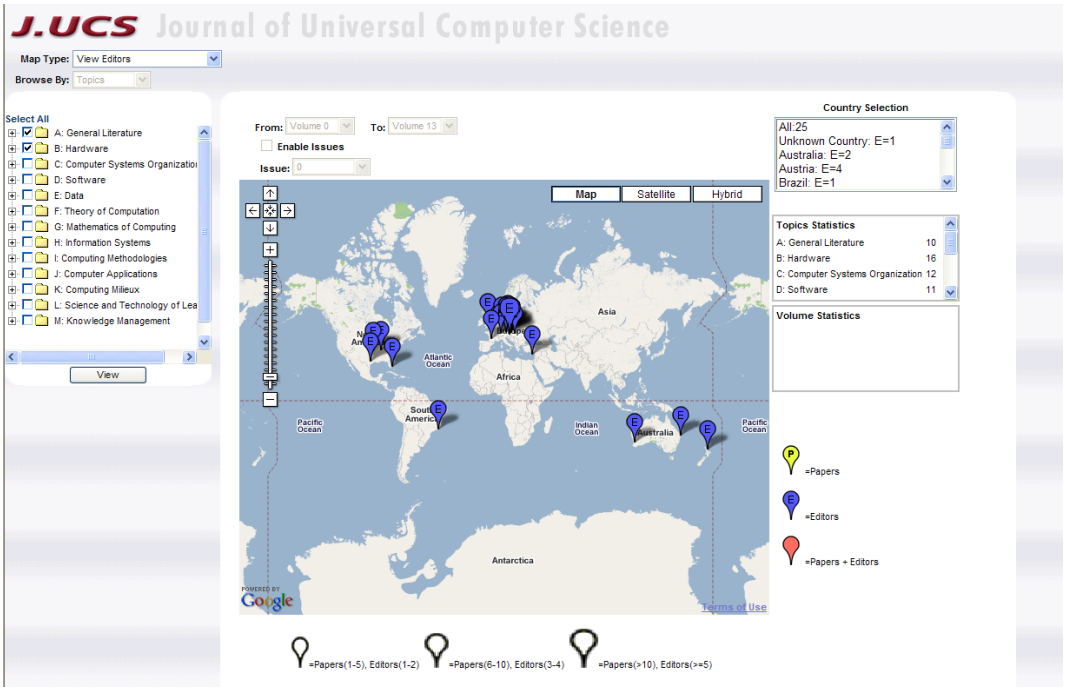


Figure 2.13 Distribution of editors across the globe for topics "General Literature" (A) and "Hardware" (B).

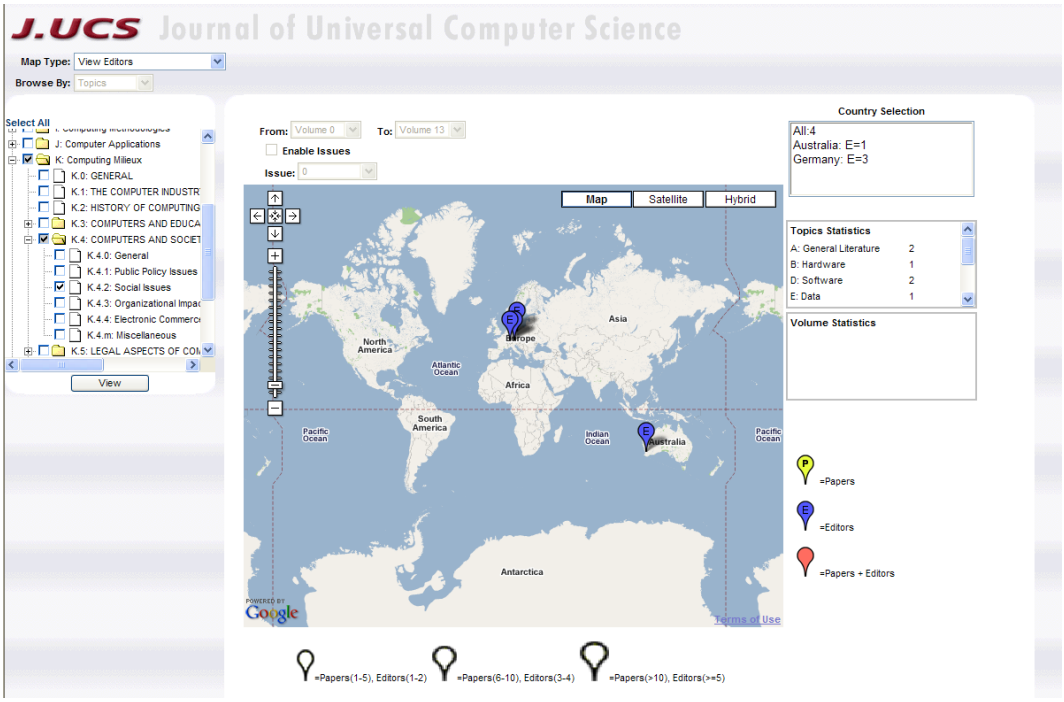


Figure 2.14 Editors distribution for topic "Social Issues" (K.4.2).

2.8.3 Distribution of Papers versus Editors

A comparative analysis of the distribution of authors and editors can also be visualized using mash-ups. The admin can select any range of volumes and topics and is presented by the system the distribution of papers versus editors across the globe as shown in Figure 2.15. A blank marker without any letter (P or E) represents the location where there are both editor and papers. The size of the blank marker represents the number of papers from that location.

By visualizing authors and editors together, this tool effectively enables the administration to become aware of a possible bias or location based conflict of interest situations in the review process. The size of the marker immediately highlights locations from which many number of papers have been accepted, with an indication of availability of editors from similar locations. This tool can help in minimizing the situations in which too many papers come from a particular research area or from the same location. Similarly, the quality of papers can be ensured by avoiding such conflict of interest situations and allocating reviewers from diverse locations.

This interface can also help the administration in making decision on expanding the coverage of the journal. As shown in Figure 2.16, for topic "General Literature" (A),

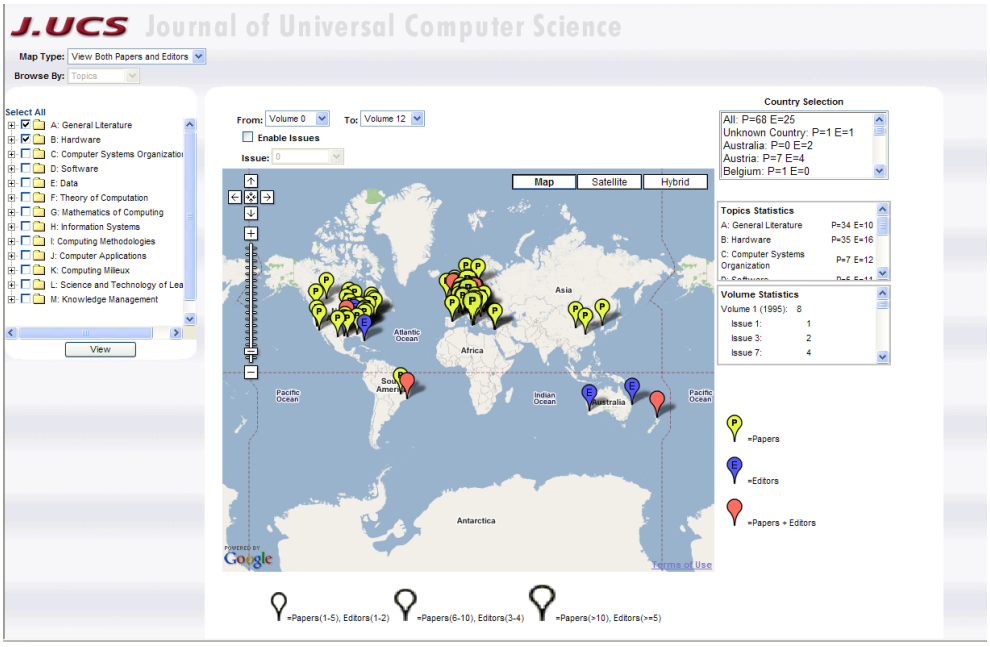


Figure 2.15 Distribution of authors vs. editors across the globe for the topics "General Literature" (A) and "Hardware" (B) from volume 0 to volume 12.

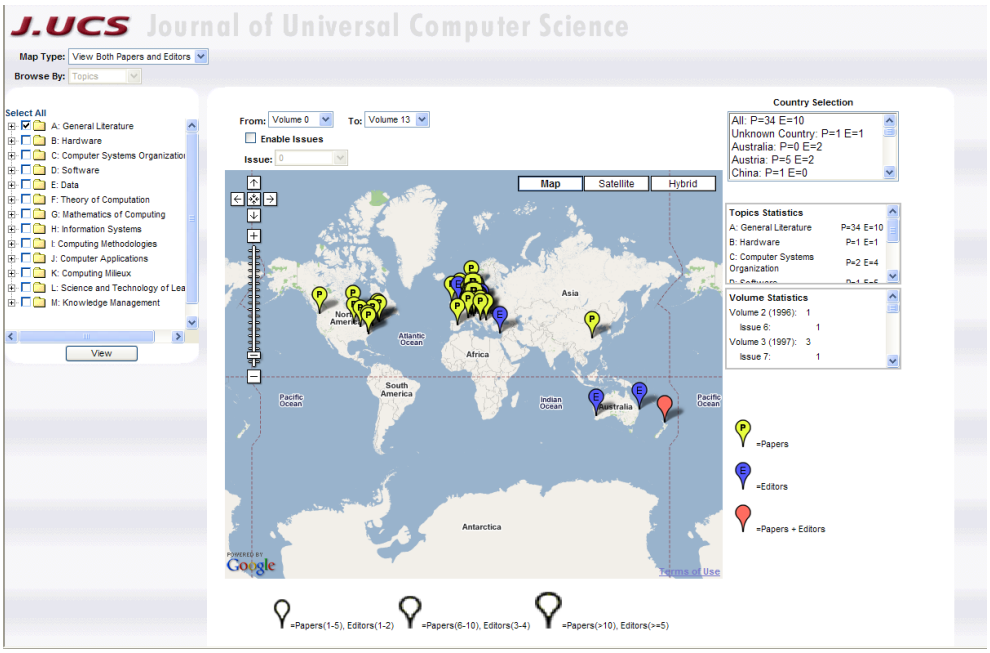


Figure 2.16 Distribution of authors vs. editors under the topic "General Literature" (A) for all volumes.

there are two editors from Australia but there is no publication from Australia under this topic. These editors could then be asked to assist in promoting the journal to their colleagues. Moreover, places in which multiple authors come from can serve as a basis for selection of new editors. This will also allow for a more even distribution of editors geographically.

2.9 Conclusions

In this chapter, we highlighted the need of utilizing the new developments of the web for expanding an electronic journals readership, authorship and quality of its publications. In this context, we explained some of the concerns faced by J.UCS management, its authors and readers. We demonstrated the employing a mash-up (an emerging web 2.0 paradigm) as a content management and decision-making tool for the users and the administration of the journal. Our experimentations conclude that mash-ups are a powerful application development platform that enabled us to address the concerns mentioned earlier. It is our belief that mash-ups will serve as an important paradigm shift for an electronic journals expansion. The work presented in this chapter can serve as a model for other contemporary electronic journals. Future work can further enhance the ideas presented in this chapter to develop multi-featured mash-ups by incorporating more publications and bibliographic data from different journals, publicly available digital repositories such as CiteSeer, DBLP and by searching publications using publicly available search APIs such as Google, Yahoo and other search engines. The implications of such a work will be very useful for many academic appraisal tasks such as promotions, finding experts, social behavioral patterns and conflict of interest detection in peer review system.

3

Extended Visualization for Scholarly Communications*

Scientometrics and content analysis of scholarly publications has been a tradition of many electronic and printed journals to ensure quality and the journal's standing. Traditionally, researchers conducted such analysis using normal tables and statistical charts. In this chapter, we apply an interactive visualization system that can help for a deeper scientometrics and content analysis of digital journals. The adapted visualization system is an easy to use web application, based on animated 2D bubble charts and pie charts to handle geographical, temporal, and large kinds of categorical data. To demonstrate the applications of this visualization technique, we conducted two separate studies. We first apply this technique to the Journal of Universal Computer Science (J.UCS) as an assistive tool to strengthen its internal administration. In the second study, we employ the visualization tool with few improvements to five journals and two conferences in the field of e-learning. The second study will also allow novice and experienced educators, researchers and policy makers to understand what kind of different research areas exist in the field of e-learning, and to identify different research trends over the last six years using the visualization tool.

3.1 Scientometrics and Content Analysis

In any academic discipline research publications represent the knowledge structure of that discipline. This knowledge structure reflects the history, research trends, social structure of researchers, networks of scholarly papers, experts, key papers, contributions and collaborations of institutions and regions. Much can be learned

*The material presented in this chapter is based on my previously published papers in [Khan et al., 2009, Maurer and Khan, 2010]

by analyzing the research contributions in a journal or conferences of any discipline about a given field of study [Taylor, 2001]. This practice of analyzing publications has been a tradition of many printed and electronic journals. During the last few decades many studies have been conducted to analyze the publications patterns by length of articles, citations, affiliations and geographical distribution of authors, contributions in different research areas, trends of research areas over time [Taylor, 2001, Menz, 2000, Marcouiller and Deller, 2001]. Many researchers have also put their efforts in analyzing the scientific knowledge of a given discipline in a broader picture by considering the publications from more than one academic journals [Tutarel, 2002, Hawkins, 2001]. These types of analysis are usually studied by the fields of bibliometrics, scientometrics, and content analysis. According to Pritchard, bibliometrics deals with the application of mathematical and statistical methods to books and other forms of communication to measure the production and consumption of their material. While scientometrics defined by Nalimov and Mulchenko is the application of these quantitative methods to science communications, to measure the scientific process [Glänzel, 2003]. On other hand, content analysis, a well-established area of research deals with the extraction of meaning from the text material [Borgman and Furner, 2002]. There are various benefits of such analysis; first it helps the administration of a journal or conferences to increase their quality by examining the coverage and impact of the venue. It helps in investigating whether the venue is aligned to its policies, providing valuable research contributions, and in comparing with other venues. Secondly, it can be used to evaluate individuals, organizations, nations, and groups. This in turn can be used to know the impact of decisions and policies made for allocating resources and funds, and proposing the future directions for the field. Moreover, it reduces the researchers' menial efforts to conduct their surveys themselves and shows them a broader picture of their field of interest [Börner et al., 2003].

3.2 Support from Information Visualization Domain

The different kinds of analysis that we discussed in previous section are further supported by the techniques developed in information visualization domain. Traditionally, researchers have tried to conduct such analysis for scholarly publications using tables and statistical charts. Interactive visualizations has been used by [Ahmed et al., 2004, Ke et al., 2004, In-SPIRE, 2007], to realize different patterns such as citations network of publications, number of papers over time and the correlated research areas in the publications published during the 8 years of InfoVis conferences.

In [Erten et al., 2004], authors used 100,000 unique ACM computer science papers and analyzed with the help of interactive node link graph the evolution of different research areas and the collaborative network of scientists in the field of computer science. In [Chen, 2005], interactive node link diagrams have been used along with a time line to visualize the co-citation network of papers in an effort to detect and visualize the emerging trends and transient patterns in the field of mass-extinction. All the systems mentioned above are good tools in understanding either the networks of papers, authors, and research areas, or how the research areas have emerged over the time. However, they do not demonstrate the change of interest in publications contributions and research areas across different regions. NetLens [Kang et al., 2007] applied a different approach where interactive bar charts, list view and multiple coordinated windows were used to analyze and compare trends over time among different research areas and locations, important authors, papers and institutions for CHI conferences. But the adopted visualization approach can handle a fewer number of research areas (categories) of the papers, which if increased can limit the users to compare trends of different research areas over the time. Moreover, the users can not compare the contributions from different locations with each other over the period of time (how different locations have progressed as compared to others). In [Havre et al., 2002, Gapminder, 2008], some general interactive visualization tools for trends anaylsis have been developed. In this chapter, we developed an easy to use visualization tool that will allow users to conduct a detailed scientometric and content analysis of scholarly communications.

3.3 Development of the Visualization Tool for J.UCS

As discussed in the previous chapter, the Journal of Universal Computer Science (J.UCS) is an open access, high quality, peer reviewed electronic journal having more than 1000 publications since 1994. The journal covers all aspects (all ACM categories and two additional categories, i.e., "Science and Technology of Learning" and "Knowledge Management") of computer science discipline. J.UCS is being published in volumes (one per year) with at least 12 yearly issues since 1994 [J.UCS, 2007].

The previous chapter also explored the usage of mash-ups, an emerging Web 2.0 technology to strengthen the internal administration and providing better facilities for the authors and readers of a digital journal. The proposed system was a good tool in highlighting the geographical coverage of publications and editors in any particular research area. We demonstrated its potential in identifying the bias groups of authors and editors in the review process of articles on the basis of their location. We further

showed its applicability for the selection of special issues according to the geographical policy of the journal, novel navigational features, and in determining research collaborators for authors and readers. However, the proposed system was limited in providing various features that can further help in increasing the quality of a digital journal. For example, the administration of a journal might be interested to know that how the authors' and institutions' contributions and research interests have changed across the globe or in a particular location over the period of time. This in turn can help in promoting the journal globally, and in finding the research areas that are becoming localized to a specific community. Similarly, the journal's managers might be concerned in determining the research areas that are evolving or declining over the period of time to assist in making decisions regarding the call and acceptance of special issues and acquiring reviewers accordingly. In this chapter, our work seeks to develop a visualization system than can help in addressing these potential requirements by providing a deeper analysis (scientometric and content analysis) of scholarly publications of a digital journal over time.

3.3.1 Visualization Tool Design Choices

By keeping in view the requirements mentioned in the previous sections, we conducted a detailed survey (highlighted in Section 3.2) of various available visualization tools. The motivation was to develop a user friendly, easy to understand trend analyzer that targets not only the experts but also general academic users. An appropriate choice was to use Gapminder [Gapminder, 2008], which can visualize geographical trends over time in the form of animated bubble charts. But the limitation of this tool is that it does not cater to categorical data, which in our case are ACM categories. On the basis of Gapminder a visualization tool was implemented that allows the user to select any predefined statistical choices along x-axis and y-axis of the bubble chart. The main interface of the visualization tool is shown in Figure 3.1. The user can select to view patterns in the publications published in regular issues, special issues or both. The user can also select to view patterns in the whole world as a single entity or across countries and regions. The results can also be filtered by selecting any topic or country from the list of countries and topics. A temporal slider has been provided for the users to scroll across different years. The user can also play the slider for automatic scroll across the years and can view the animated moving bubbles and pie chart to reveal different patterns. Each bubble on the chart represents a country, region or the whole world based on the user selection. The color and size of the bubbles represents the location and number of publications respectively. The axis of the animated bubble

chart contains various options in which the user might be interested, such as number of institutes, number of authors, number of papers, average length of papers, and average number of authors per paper. The pie chart represents the distribution of publications across topics for any particular country, region, or the whole world. As J.UCS has more than 400 topics which can span up to three levels, the information for sub-topics is provided only on user demand. Providing contents on user demand is very important for any good visualization tool as proposed in [Shneiderman, 1996]. Based on this guideline, the user can select the bubble of a particular location. The corresponding pie chart provides the distribution of papers for the first level topics available at the selected location. The user can select any sector in this first level pie chart to view the distribution of papers in a particular sub-topic at the second level, and subsequently can reach up to third level. One can further enhance the analysis capability of the visualization tool by toggling locations in the pie chart and topics in the bubble chart. In this case, the moving bubbles in the bubble chart will represent the evolution of topics over the period of time, and pie chart will represent the distribution of papers across locations for selected topics. In the initial prototype, it was proposed to use moving pie charts instead of a bubble for a location. The sectors of the pie chart were supposed to shrink or expand with the passage of time depending upon the number or ratio of papers in each topic. But the limitation of this animation is that it can generate too much cognitive load for the users to understand the trends because of

Figure 3.1 Main Interface.

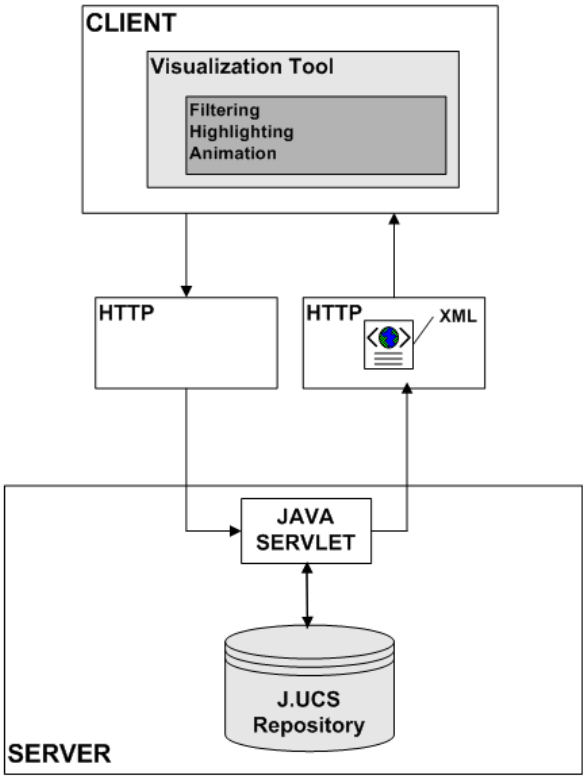


Figure 3.2 Application Architecture.

growing bubbles with expanding or decreasing sectors over the time. Therefore, bubbles representing a pie chart were replaced by simple bubbles representing the number of papers in a particular location. The user can explore the distribution of papers for each level of topics in a separate pie chart.

Application Architecture

The main architecture of the visualization application is shown in Figure 3.2. On the client side, the application is responsible for user interactions, animations and accepting requests from users and sending the HTTP requests to the server for the data. The server side is responsible to accept and process the request. The server extracts the required data from the database and sends it as XML file to the client. The client side of the application has been implemented using Adobe Flex for animated bubble chart and user interaction, and the server side has been implemented using Java Servlets. Google has also started to provide visualization APIs for the animated bubble charts, but these APIs were not mature enough to be used for user interaction.

3.3.2 Experimental Results

In this section, some interesting results are presented that can be obtained with this visualization tool. In order to facilitate an easy analysis and to better understand the results, the publications of J.UCS from 1994-2008 have been divided in three groups; each spanning 5 years, i.e., 1994-1998, 1999-2003, and 2004-2008 (so far papers for the first six months of year 2008 are available). In this study, we are considering mainly regular issues instead of special issues of J.UCS for our analysis because they represent a clear picture about various trends. The following sub-sections represent some interesting results with regards to three different views (world, regional, and countries).

World View

This view reflects all publications in J.UCS as a single entity. Figure 3.3 demonstrates the evolution of J.UCS with regards to the number of publications, institutions, and authors for the regular issues.

As it can be seen in Figure 3.3 that up to 1998 the total number of publications, authors, and institutions in J.UCS were 130, 206, and 200 respectively. It can be observed that there is a consistent decline in publications for the time period 1999-2003 (90) and 2004-2008 (84), whereas authors and institutions first declined for the time period 1999-2003 (140 authors, 133 institutions) and then started to increase in 2004-2008 (177, 156). These statistics also reflect the inclusion of new authors and institutions in the journal instead of being occupied by some groups of authors. Figure 3.4 demonstrates the distribution of publications across different research areas. As it can be seen that the two top most research areas are "Theory of Computation" (1994-1998: 27; 1999-2003: 32; 2004-2008: 22) and "Information Systems" (1994-1998: 28; 1999-2003: 20; 2004-2008; 28). The same view can also be used to visualize research areas that have started to diminish or grow. For example, up to 1998 there was only one publication in the research area "Computer Applications". Then there is a sudden rise in publications (11) from 1999-2003 and again a decline in publications (2) for the period 2004-2008. A similar phenomenon happened with "General Literature" where the journal has accepted 3 publications up to 1998 and only 1 publication up to 2008.

The user has the choice to view the similar trends in the sub-categories of any research area by clicking any top level category on the animated pie chart. The Figure 3.5 reflects that "Software Engineering" (a sub-category of "Software") was not a dominant research area when compared to "Programming Languages" up to 1998, but it started to evolve from 1999-2008 and is now the most dominant research area in its category.

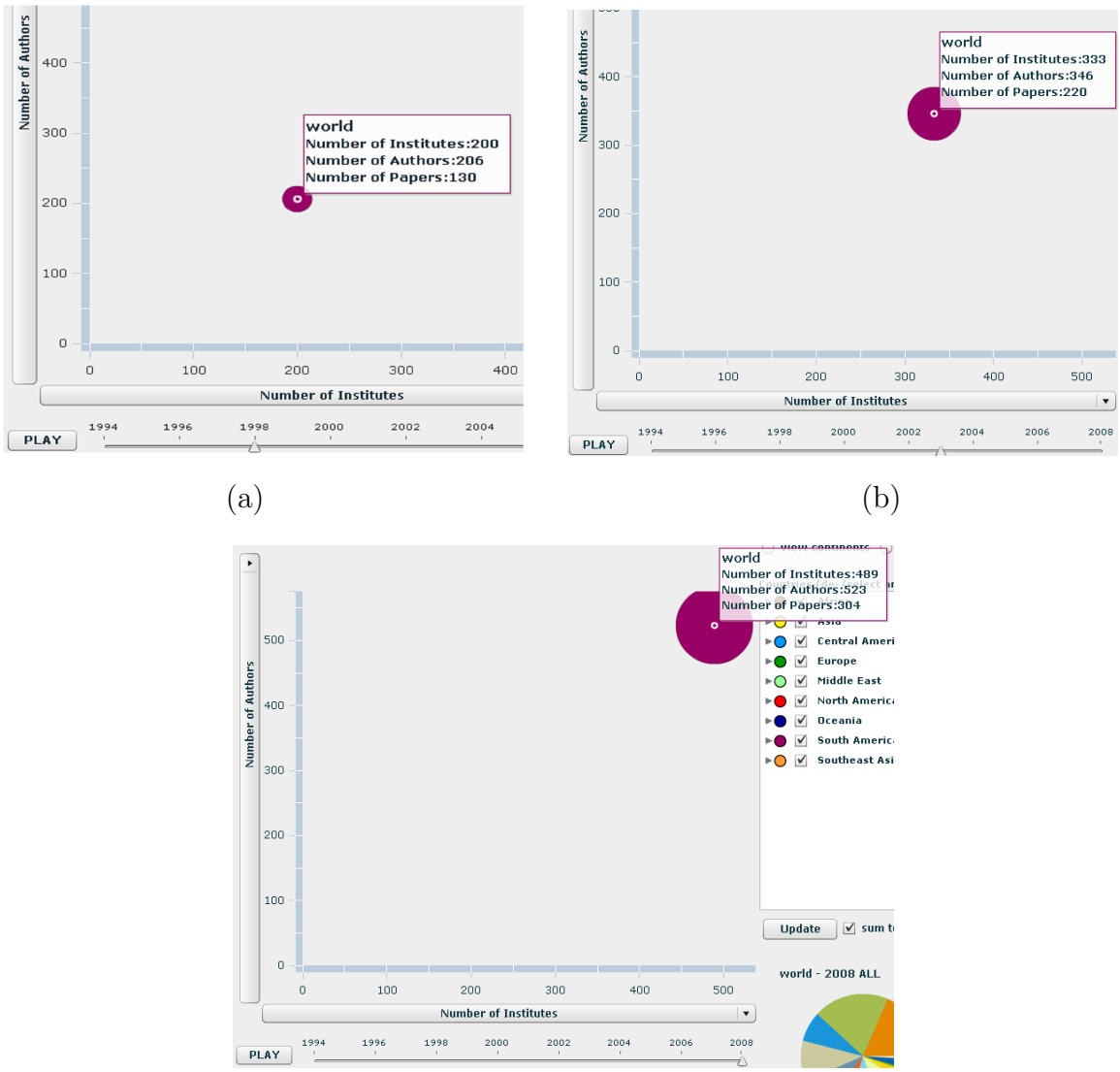


Figure 3.3 World View: (a) Publications up to 1998; (b) Publications up to 2003; (c) Publications up to 2008.

The emergence of "Software Engineering" can be further validated by considering both special and regular submissions to the journal as shown in Figure 3.6. Moreover, one can also additionally investigate that the current growth of "Software Engineering" is a global phenomenon or is localized to a particular group of authors, institutions, regions or few countries by using the Regional and Countries view which are explained in the following subsections. Such analysis of research areas is necessary as it gives an overview to the new researchers about the emergent or hot research areas of their field. Moreover, it helps the administration of the journal to acquire reviewers for

each research area accordingly, for the call and acceptance of special issues and to ensure that the coverage of the journal in each research area remains global instead to a particular locality.

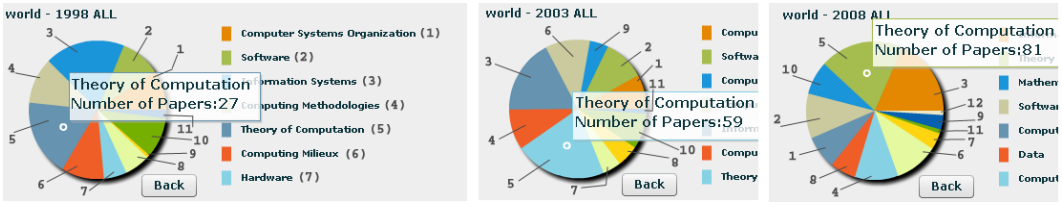


Figure 3.4 Distribution of Publications for the Level-1 Categories.

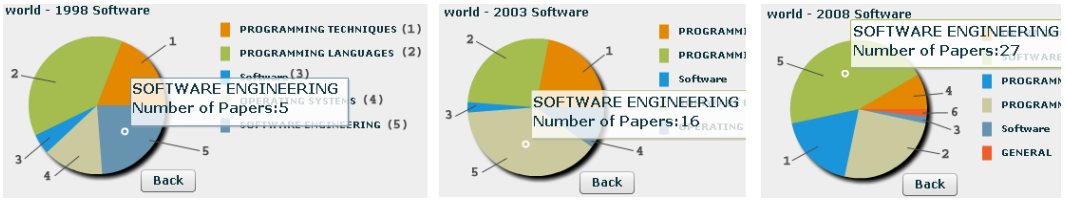


Figure 3.5 The Evolution of "Software Engineering" under the Category "Software" for Regular Issues.

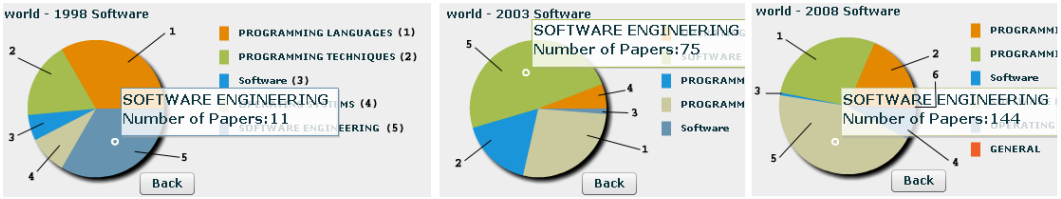


Figure 3.6 The Evolution of "Software Engineering" under the Category "Software" for both Regular and Special Issues.

The user can also filter the countries from the list of countries to visualize the impact of a single country or combined effect of different countries with the passage of time.

Regional View

This view demonstrates the distribution of publications across different regions. It can help the users to understand how different regions have evolved with the passage of time, which region occupies the journal and which region is active or passive as a whole or in any particular category.

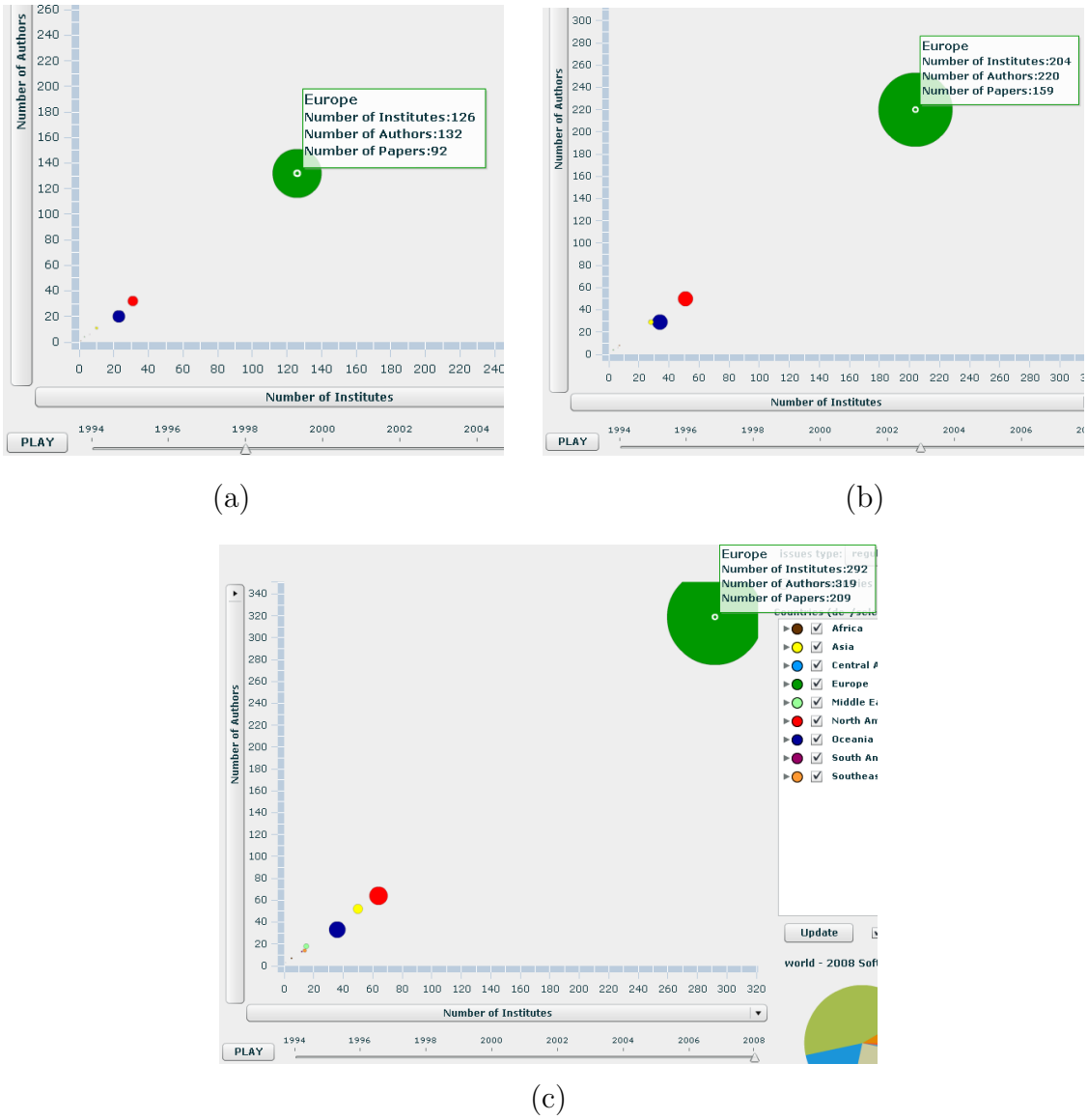


Figure 3.7 Regional View: (a) Publications up to 1998; (b) Publications up to 2003; (c) Publications up to 2008

The Figure 3.7 demonstrates that European countries remain as the main source of publications in the journal for all the time periods, but there is consistent decline of publications (1994-1998: 92; 1999-2003: 67; 2004-2008: 50) with the passage of time from Europe. Further analysis revealed that the Asian countries are consistently contributing more publications (1994-1998: 6, 1999-2003: 6, 2004-2008: 10) in the journal. One more important benefit of this view is that a region can be compared

with any single country or group of countries from other regions by using the filtering option from the list of countries.

Countries View

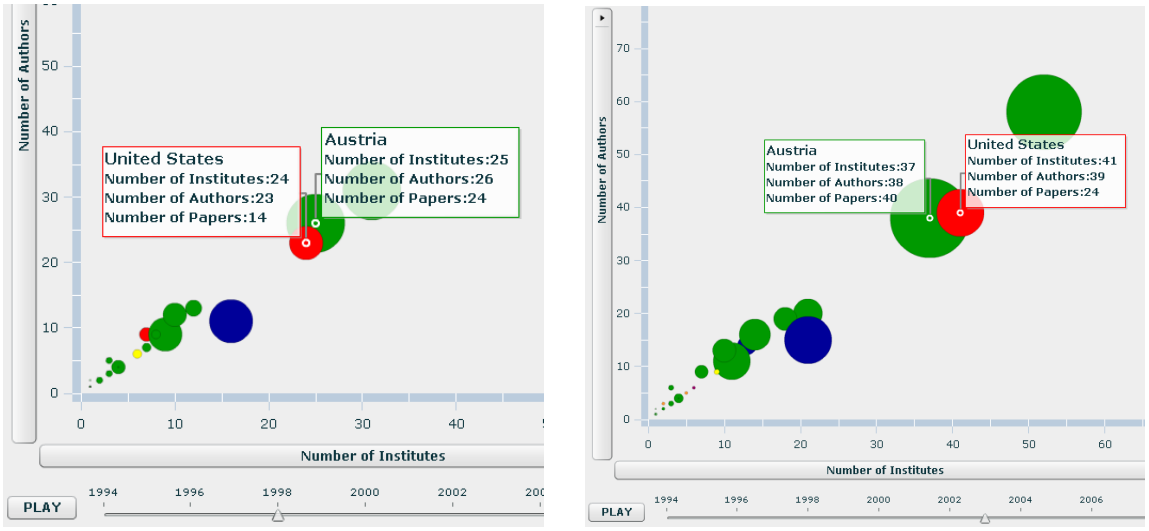
This view further provides more insight into publications patterns. It enables the users to understand the participation of each country in the journal, when a country started to contribute, when it stopped to contribute, who is contributing more or less, which country is strong or passive in any particular research area, how different research areas have evolved in each country.

Figure 3.8 reflects that most of the publications in J.UCS have been contributed by Austria (1994-1998: 24; 1999-2003: 16; 2004-2008: 12) followed by Germany. Further analysis revealed that most of the authors (1994-1998: 31; 1999-2003: 27; 2004-2008: 13) and institutions (1994-1998: 31; 1999-2003: 21; 2004-2008: 11) participations are from Germany. Interestingly Finland and New Zealand were contributing frequently in the journal for the first two time periods, but each of them has contributed only one publication from 2003 to 2008.

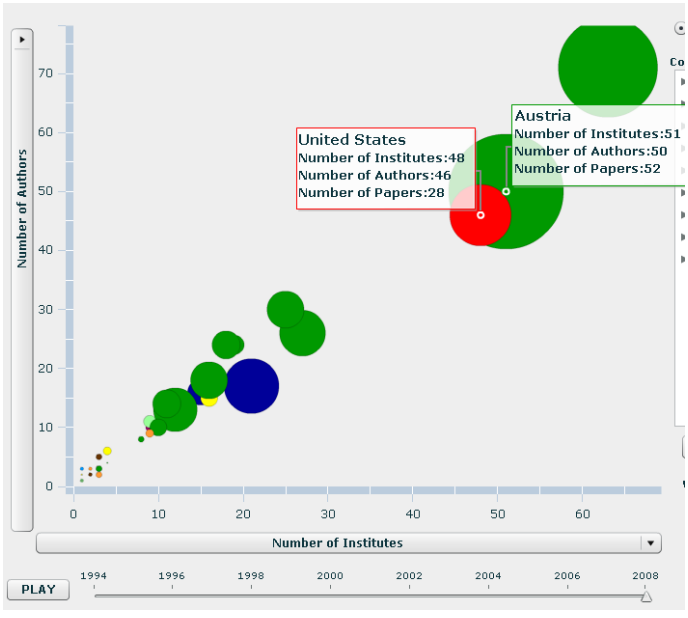
The administration of the journal in this case can take action to encourage researchers in these locations to submit their papers in the journal. The same view can also be used to see which country is contributing none, low or many publications in any particular research area or group of research areas. As it can be seen in Figure 3.9 that for the topic "Multimedia Information Systems", Austria has been the major contributor. Moreover, there is no contribution from Asian countries in this category.

3.4 Limitations of the Visualization Tool

In a co-occurrent study by Robertson et al. [Robertson et al., 2008], the authors conducted a detailed usability study of Gapminder. They highlighted two main limitations of this visualization. One obvious limitation of this visualization is the lack of trend line between the continuous moving states of any bubble. This in turn makes the data analysis more difficult. The authors suggested to use fading bubbles from most transparent (earliest states) to most opaque (latest states), in addition to the fading traces lines connecting the sequence of bubbles to visualize a sequence of flow. They highlighted that clutter is another problem that can make the analysis more difficult and error-prone by increasing the number of bubbles in the visualization tool. The authors proposed to use small multiple views, each containing a single bubble to avoid clutter produced by the increased number of bubbles and their fading traces. However,



(a) (b)



(c)

Figure 3.8 Countries View: (a) Publications up to 1998; (b) Publications up to 2003; (c) Publications up to 2008

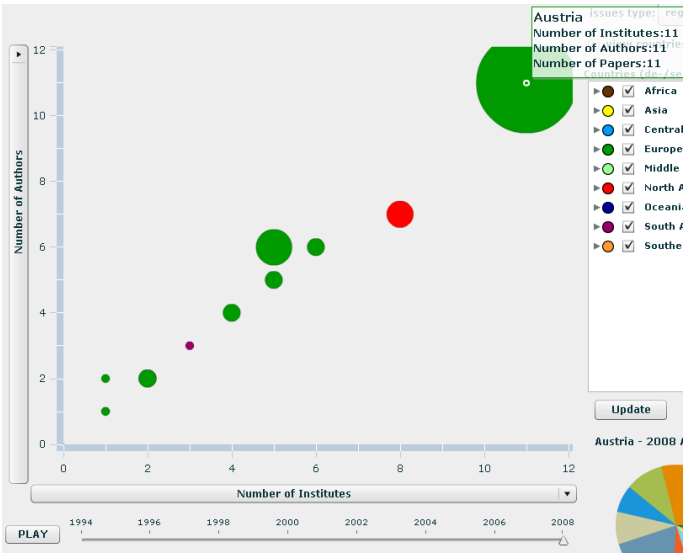


Figure 3.9 Distribution of Publications among Different Countries for the Category "Multimedia Information Systems".

according to the authors, the user in this case has to scan all small multiple views to compare trends and answer various questions.

In our updated tool, we tried to tackle these problems by providing a separate tab containing traditional but interactive line charts for the deeper and error-free analysis of data. The line charts can visualize the trends across regions as well as different topics. We found that these charts are very helpful in finding various patterns during our analysis. The next section describes another case study that utilizes the updated visualization tool for the scientometric and content analysis of selected journals and conferences in the field of e-learning.

3.5 Case Study in the Field of E-Learning

The significance of analyzing trends in distance education or e-learning has also been realized by many researchers. The author in [Masood, 2004] provided a content analysis of ten years of an e-learning journal by classifying the publications in meaningful categories. In [Lee et al., 2004], the authors uncovered the hidden trends patterns by examining four well-known distance education journals by analyzing papers published from 1997 to 2002. The authors in [Berg and Mrozowski, 2000] also analyzed 890 publications in distance education. In [Shih et al., 2007], the authors considered five e-learning journals for their analysis. All these studies provided useful information about overall research themes, methods, trends and important papers. Our work

is also along these lines, where we seek to uncover different research trends in the field of e-learning using our visualization tool. Recently, [Ochoa et al., 2009] gave the scientometric and bibliometric analysis of research publications published in the last ten years of Ed-Media conferences. The authors exposed the most prolific authors, countries, collaborations, and citations networks by using only a single venue.

This section presents a scientometric and content analysis of the studies published in five Social Science Citation Index (SSCI) e-learning journals, i.e., "Computers & Education", "British Journal of Educational Technology", "Educational Technology Research and Development", "Innovations in Education and Teaching International", "Journal of Computer Assisted Learning" and two conferences from e-learning domain, i.e., "Educational Multimedia, Hypermedia & Telecommunications (Ed-Media)" and "IEEE International Conference on Advanced Learning Technologies (ICALT)". Although journals and conferences are clearly different publication mediums, but considering only journals would put a limitation on the studies due to time lag problem of their rigorous review process. This problem can be minimized by considering publications from conferences for those research trends that are not yet appeared in journal articles. In this study, we have also carried out a comparative study between these two mediums.

The aim of this study is to allow novice and experienced educators, researchers and policy makers in the field of e-learning to understand what kind of different research areas exist and to identify different research trends over the last six years using our internally built interactive visualization tool.

3.5.1 Method

Selection of Journals and Extraction of Data

In this study we have used papers published by five major journals indexed by SSCI and two well-known conferences from 2003 to 2008 to realize the research trends in the field of e-learning. These journals and conferences are chosen in order to make our dataset more comprehensive, and for their long publishing history in the area of e-learning. Each paper published in these journals and conferences has a well formatted html page in their respective digital libraries, containing all the information about that paper such as: paper title, authors' names, affiliations, countries, year of publication, volume, and issue numbers. However, it should be noted that the IEEEXplore digital library for ICALT only contains the affiliation/country information of first authors. Therefore, the papers or meta-data for this venue will not be considered for any geographical trends

in our analysis. We parsed these html pages using regular expressions to extract meta-data about each paper. The meta-data was stored in a relational database to be used by our visualization tool. This study only analyzed 7,759 original research articles. Other types of articles such as book reviews, editorials have been excluded. The number of papers from each venue has been listed in Table 3.1.

Table 3.1 Venues and their respective number of papers used in this study.

Venue	Number of papers
Computers & Education (CE)	546
British Journal of Educational Technology (BJET)	333
Educational Technology Research and Development (ETRD)	155
Innovations in Education and Teaching International (IETI)	200
Journal of Computer Assisted Learning (JCAL)	246
Educational Multimedia, Hypermedia & Telecommunications (Ed-Media)	4,607
IEEE International Conference on Advanced Learning Technologies (ICALT)	1,672

Analysis and Normalization of Meta-data

The country information of the authors needed to be cleaned and standardized as the data contained various representations of the same location (e.g., USA, United States). In order to standardize them to a unique name, we compared the countries data with GeoBytes database [GeoBytes, 2007] containing the names of all countries and cities across the globe. The countries that had no match with the GeoBytes database were identified and corrected manually. There were 526 authors out of 14,721 (excluding ICALT) whose country information was found missing. In order to deal with this problem, we first parsed the names of the institutes to locate the name of the country. In cases where this method failed, we searched for the same institute name for other authors to look for the existence of country information. By using the above-mentioned approaches, we succeeded in finding the countries information of 382 authors. Moreover, there were 78 authors whose institution information was not available.

Authors and Institutions Names Disambiguation

In many cases different authors or institutions can have the same name, and same author or institution can have different names representations. For accuracy of results there is a need to disambiguate the institutions and authors names. In our case, we have applied a simple text similarity algorithm called n-grams found in [Shannon, 1951] to disambiguate the institutions names. In order to get better results, advance techniques also exist that include dictionary and matching rules [Yang et al., 2008]. However, in many cases human intervention is also required to resolve this problem.

The authors' names were disambiguated using the approaches found in [Aleman-Meza et al., 2008], but without adapted reference reconciliation algorithm. Two authors are reconciled according to the criteria shown in Table 3.2. The thresholds have been selected to minimize the number of false positives and true negatives; however, it is practically very difficult to ensure 100 percent accuracy. As compared to previous studies, the numbers of co-authors in common are kept up to two because our gathered bibliographic database is very small in comparison to other bibliographic databases such as DBLP or CiteSeer.

Table 3.2 Authors attributes and similarity thresholds.

Attributes	**Similarity threshold**
Authors with exact similar names	
Institutions names	≥0.8
Countries	1
Authors with dissimilar names	
Authors names	≥0.8
Institutions names	≥0.8
Countries	1
Co-authors in common	≥2

Identification of Topics Based on Concepts Clusters

In order to identify the research topics in the field of e-learning, we used paper titles and abstracts, and employed open-source [doc2mat, 2009] utility to convert the documents in a vector space format. The term-document matrix generated by this tool was again given as an input to another open-source tool named gCLUTO [Ramussen and Karypis, 2008] for non-overlapped clustering of documents. We used gCLUTO because it is an open-source software and has user-friendly interface. It uses

various similarity detection algorithms, allows any number of clusters of documents to generate, and provides percentage of most frequent terms or concepts within a cluster. By keeping in view the number of papers and to generate meaningful clusters of considerable size, we identified 150 clusters of documents, each representing a concept. Then the author of this thesis and another PhD student working in the field of e-learning partially inspected each cluster manually and assigned a theme to each cluster based on the documents it contained and the most frequent terms in a cluster. We further classified the clusters in more meaningful categories by following the classification system developed by [Masood, 2004]. This classification has been done to the best of our knowledge and understanding. The description of 14 main categories and corresponding 150 concepts are given in Table 3.3. The concepts or terms have been rounded off using standard porter stemming algorithm employed by doc2mat utility.

As gCLUTO use non-overlapped clustering of documents to identify the concepts clusters, however, originally a document can exist in more than one class. In order to deal with this problem a simple heuristic rule based classifier was implemented to place any particular document in more than one class. In order to assign a document to a particular class, the classifier uses the existence/non-existence of most frequent terms identified by gCLUTO.

3.5.2 Results and discussion

In this section, some interesting results have been presented that can be obtained from the visualization tool for e-learning domain (based on selected journals and conferences). The main interface of the extended visualization tool is shown in Figure 3.10. By keeping in view the limitations of the visualization tool as mentioned in Section 3.4. In current implementation, the trends about locations and research areas can be further explored with the help of line charts in the "Report" tab. Moreover, by clicking on any bubble, a list of authors appears which provides the affiliation information and respective publications of authors corresponding to the selected bubble. In this way, it is possible to find an expert based on total number of publications for any particular topic and location. The user can also click any author in this list to view the author's performance over the period of time in terms of number of publications as shown in Figure 3.11. We also extended our levels of analysis beyond world, continents, and countries by including institutions view to realize research trends at even the institutions level.

In order to better understand and facilitate an easy analysis of the results, we have divided the publications into three groups each spanning to two years, i.e., 2003-2004,

Table 3.3 Concepts and categories used in the study.

Categories	Abv.	Concepts/terms
Instructional/educational tech.	IET	ICT, elearn/elearning, innov/innovation, technolog/ technology, chang/technology change.
Instructional process variables	IPV	Inquir/inquiry, feedback.
Instructional process elements	IPE	Ontolog/ontology, navig/navigation, search, mine/data mining.
Teaching/learning perspectives	TLP	Constructivist.
Instructional methods	IM	Patchwork, wiki, robot, Annot/annotation, script, pbl/problem based learning, blog, game (appeared two times), scaffold, dialogu/dialogue, map/concept map, forum, CSCL, simul/simulation, blend/blended learning, self/self-regulation, team, messag/messages, synchron/synchronous, knowledge, discuss, social, cognit/cognition, collabor/collaboration, webquest, commun/online communication, task/task solving.
Delivery systems/media formats	DS	Cellular, podcast, eportfolio/portfolio (two times), phone, cyber, devic/device, textbook, handheld/handheld device, mobil, semant/semantic web, tutor/intelligent tutor, laptop, LM/LMS, hypermedia, video, webct, access, websit/website, stream/streaming, diagram, multimedia, anim/animation, distanc/distance learning, materi/material, text, internet, softwar/software, web, media, cours/course, digit/digital library, online, comput/computer.
Instructional development	ID	Museum, music, pattern, agent/intelligent agent, IM/IMS, metadata, style/learning style, scorm, health, graph, repository, quality, english, lesson, scenario, speech, adapt, busi/business, trust, laboratory, case, lectur/lecture, space/ virtual space, visual, network, service/service architecture, exercise, statist/statistics, program, imag/image, user, instruct, train, resource, field, curriculum, model, manag/manage, contextu/contextual.
Production variables	PV	Object/reusable object.
Learner outcomes	LO	Creative, write, read, skill/thinking skill.
Learner variables	LV	Doctor, parent, emot/emotion, nurs/Nurse, pupil,mathemat/mathematics, children, women, elementary, school, institute, learner.
Learning environment	LE	VLE/virtual learning environment, virtual.
Evaluation	EV	Peer, usabl/usability, test, assess.
Culture	CU	Culture, language, global, chinese, european.
Teacher variable	TV	Mentor, preservice (appeared two times), belief,profession, faculty, instructor, staff, teacher.

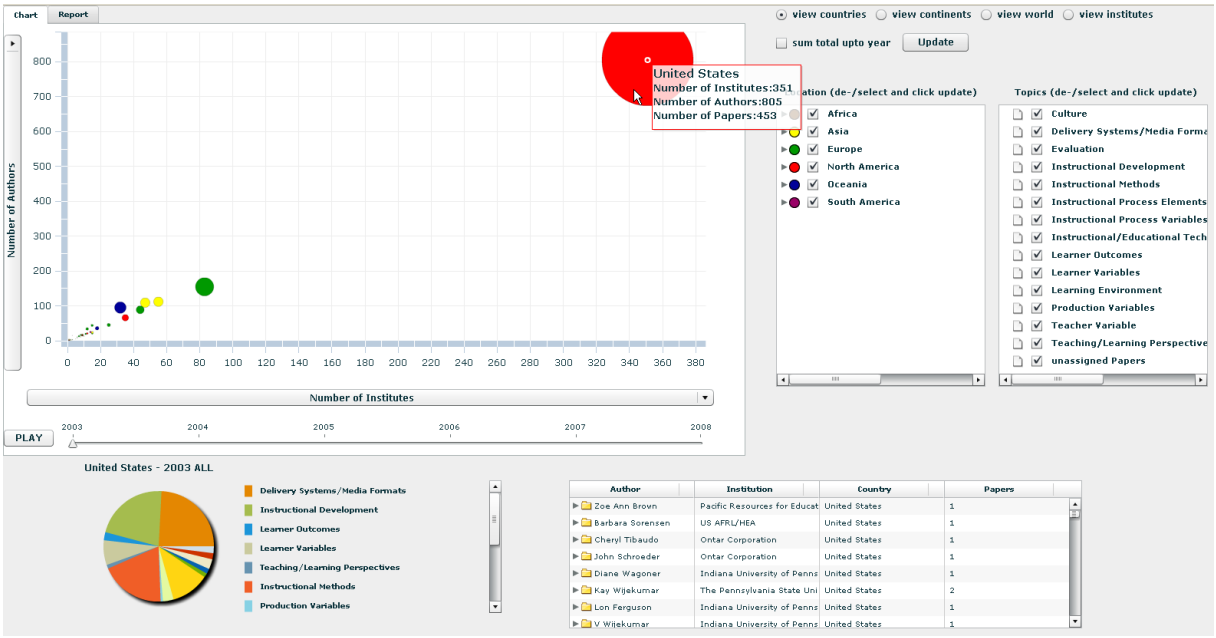


Figure 3.10 Extended main interface.

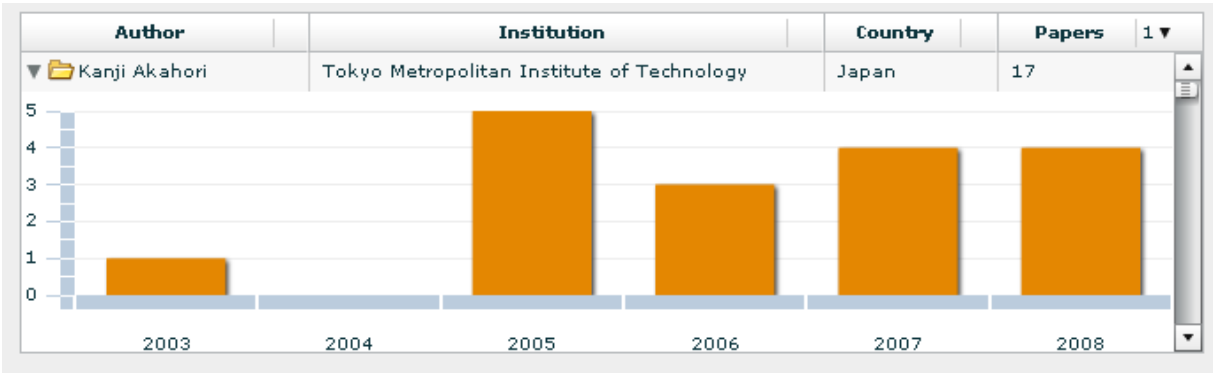


Figure 3.11 Performance of an author.

2005-2006, and 2007-2008 inclusive. The following sub-sections represent some interesting results based on four different kinds of views, i.e., world, continents, countries, and institutions.

World view

From the experiments, it was observed that the total number of papers, authors, and institutions up to 2004 were 2,077, 4,157, and 1,751, respectively (excluding ICALT). An overall decline in them was seen for the time period 2005-2006 (1,713 papers, 3,353 authors, 1,109 institutions) and later increase was witnessed during the time period

2007-2008 (2,297 papers, 4,588 authors, 1,631 institutions). The decline for the time period 2005-2006 was due to the reduction in contributions to Ed-Media conferences, JCAL and IETI. However, there is a consistent rise through all the periods for other journals. These results also revealed the inclusion of new authors and institutions in the field instead of being occupied by some groups of authors. Apart from the visualization tool, we observed that there were only 57 out of 12,098 authors and 169 out of 4,491 institutions who contributed both in Ed-Media conferences and journals articles. The average number of authors for Ed-Media conferences is 2.39, whereas for journals it is 2.50.

The Figures 3.12 and 3.13 represents the distribution of papers across different research areas and their trends over the period of time respectively. It is clear from these figures that the top five research areas in order of rank are DS (2003-2004: 1,944; 2005-2006: 1,779; 2007-2008: 2,168), ID (2003-2004: 1,827; 2005-2006: 1,771; 2007-2008: 2,036), IM (2003-2004: 1,554; 2005-2006: 1,484; 2007-2008:1,879), TV (2003-2004: 888; 2005-2006: 760; 2007-2008: 1,020), and LV (2003-2004: 604; 2005-2006: 517; 2007-2008: 762). In Figure 3.13, the percentage of research areas for each time period has been used to reveal the trends instead of raw count of number of papers because the total number of papers is not normally distributed over the years. The same pattern was found for both journals and conferences individually. These results are slightly contrary to the findings by [Ely et al., 1992] and [Klein, 1997], where instructional processes and ID were the hottest topics. Moreover, it confirms the findings of [Masood, 2004] where DS, ID, and IM were the top most researched areas, respectively. Despite the consistent decline of DS it is still the most studied research area and constitutes more than 75 percent contributions for each time period. A consistent rise in IM has been observed in this analysis. It appears that contributions in this research

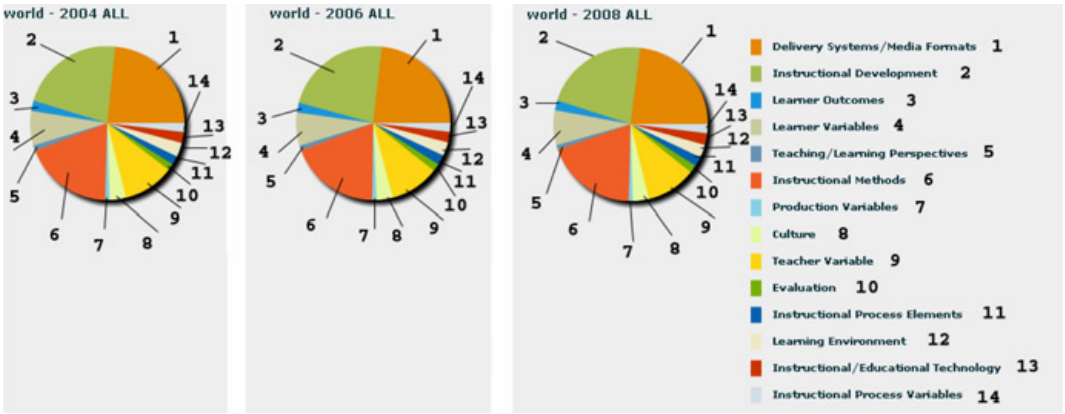


Figure 3.12 Distribution of papers across research areas.

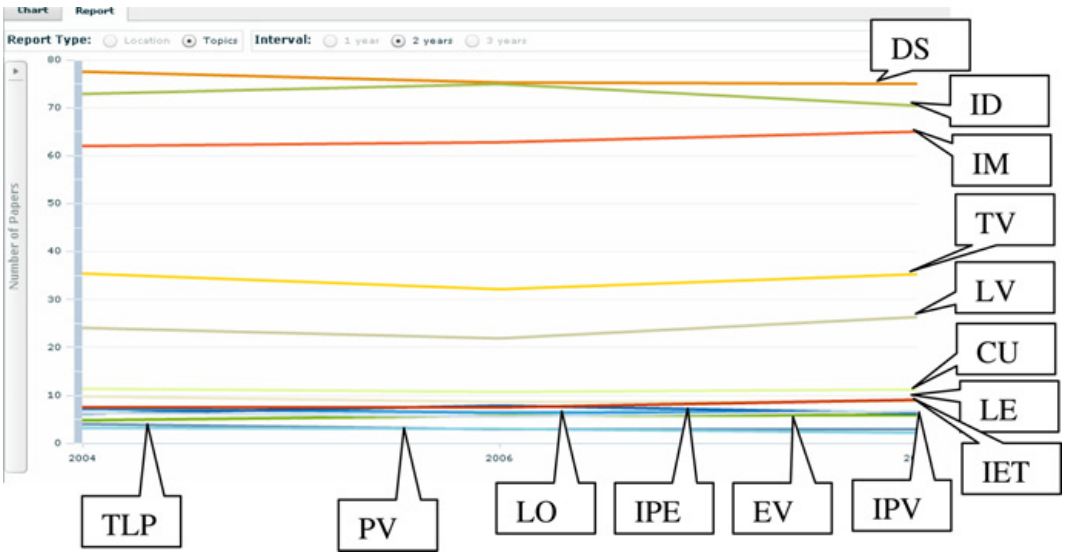


Figure 3.13 Trends of research areas across the world.

area will become equal to DS and ID in near future. The research areas TV, LV, and CU constitute less than 40, 30, and 15 percent of the total contributions, respectively, for all the time periods. All other research areas constitute less than or equal to 10 percent of all the contributions individually.

Further investigations revealed that the research area EV is consistently growing, whereas PV and TLP are declining for all the time periods. The research topics in ID and IPE first gained popularity for the time period 2005-2006 and then got declined for the time period 2007-2008. The research areas CU, IPV, IET, LO, LV, LE, and TV declined for the first time period, i.e., 2005-2006, but received growth in contributions for the time period 2007-2008. In journal papers, LV is getting consistent rise, whereas TLP is declining for all the time periods. The topics covered by CU, EV, IPE, LE, and PV received growth in the number of publications in the first time period, but the contributions got declined for the time period 2007-2008. The rest of the research areas first declined, but got growth for the time period 2007-2008. In conferences, the number of contributions in IM is increasing while for DS and PV the contributions are decreasing for all the time periods. The research areas EV, IPE, IET, and ID gained growth in contributions for the time period 2005-2006, but got declined for the second time period. All other research areas first declined and then got raised for the time period 2007-2008, except LO which remained more or less consistent or a slight decline for all the time periods.

As in any academic discipline, the research activity extensively depends on social interactions and scientific collaborations. The user can view the most prolific authors in

Author	Institution	Country	Papers 1 ▼
Kanji Akahori	Tokyo Metropolitan Institute of Technology	Japan	17
Ron Oliver	Edith Cowan University	Australia	16
Joe Luca	Edith Cowan University	Australia	15
Catherine McLoughlin	Australian Catholic University	Australia	15
Lori Lockyer	University of Wollongong	Australia	13
Barry Harper	University of Wollongong	Australia	12
Yoneo Yano	Tokushima University	Japan	11
Gavin Sim	University of Cenral Lancashire	United Kingdom	11

Figure 3.14 Top authors in Ed-Media across the world.

Author	Institution	Country	Papers 1 ▼
Ya-Ting C. Yang	Institute of Education and Centre for Teacher E	Taiwan	4
Philip Barker	University of Teesside	United Kingdom	4
Michael J. Jacobson	National Institute of Education, Learning Scienc	Singapore	3
Jeroen J. G. van Merriënboer	Open University of the Netherlands, Educationa	Netherlands	3
Paul van Schaik	University of Teesside	United Kingdom	3
Peter van Rosmalen	Open University of the Netherlands	Netherlands	3
Francis Brouns	Open University of the Netherlands	Netherlands	3
Peter B. Sloep	Open University of the Netherlands	Netherlands	3

Figure 3.15 Top authors in journals across the world.

their field by sorting the list of authors for any particular location from the visualization tool. Figure 3.14 and Figure 3.15 provides the top authors with highest number of publications in Ed-Media and journals, respectively. It was observed that none of the top ten authors from both venues were top researchers in other.

Regional View

It was observed from the experimentations that North American (2003-2004: 874; 2005-2006: 692; 2007-2008: 774) countries which include Barbados, Canada, Costa Rica, Cuba, Puerto Rico, Trinidad and Tobago, USA remained the main contributor followed by Europe (2003-2004: 679; 2005-2006: 535, 1,214; 2007-2008: 850), Asia (2003-2004: 338; 2005-2006: 377; 2007-2008: 550), Oceania (2003-2004: 187; 2005-2006: 137; 2007-2008: 177), Africa (2003-2004: 51; 2005-2006: 43; 2007-2008: 36), and South America (2003-2004: 29; 2005-2006: 20; 2007-2008: 48). However, Europe has the highest number of contributions in 2007-2008. Moreover, despite the decline of the contributions for the time period 2005-2006, Asia's contributions have increased for all the time periods. The growth of Asia as a big player in distance education has also been predicted by McIsaac and Gunawardene [McIsaac and Gunawardene, 1996]. This analysis also confirms it. Furthermore, it was observed that only small number

of papers have been contributed from Africa and South America as compared to other continents for the entire time period. This analysis reflects that still more efforts and initiatives are required to promote the applications of educational technologies in developing countries of Africa and South America. According to Latchem [Latchem, 2006], many national, international, and non-profit agencies are putting their efforts to promote education and knowledge in these developing countries by exploiting the power of educational technologies. In terms of number of institutions, Europe (2003-2004: 632; 2005-2006: 356; 2007-2008: 653) leads other regions, except for 2005-2006 where North America (2003-2004: 611; 2005-2006: 386; 2007-2008: 442) had the largest number of contributing institutions. Moreover, Europe (2003-2004: 1,430; 2005-2006: 1,051; 2007-2008: 1,802) was behind North America (2003-2004: 1,510; 2005-2006: 1,241; 2007-2008: 1,247) in number of authors up to 2005-2006, but it succeeded North America for the time period 2007-2008.

In the Ed-Media conference, all the continents followed the same pattern as mentioned above in terms of number of contributions. In journals, Europe remained the main contributor followed by North America, Asia, Oceania, Africa, and South America. Amazingly, Asia's (2003-2004: 73; 2005-2006: 77; 2007-2008: 179) number of contributions has grown more than North America (2003-2004: 89; 2005-2006: 105; 2007-2008: 173) for the time period 2007-2008. Furthermore, it was observed that DS, ID, and IM are the most prominent research areas in each continent. The dominance of these research areas in each continent also prevails for both Ed-Media and journals.

In North America, it was observed that the research areas IM, CU and LE are growing and DS is declining consistently. The topics covered by PV, IPE, EV received rise in publications for the time period 2005-2006, but got decline in contributions for the second time period. Furthermore, the rest of the research areas declined for the first time period, but gained growth for the time period 2007-2008. In Ed-Media conference, the research areas IM and LE are getting popularity while the contributions in IPV are decreasing for all the time periods. The research areas DS, EV, ID, IPE, IET, and PV gained increase in contributions for the first time period, but contributions got declined for the time period 2007-2008. All other remaining research areas declined for the time period 2005-2006 and then got growth for 2007-2008. In journals, TLP is continuously going downward for all the time periods whereas CU, IPE, IET, and PV gained growth in contributions for the first time period, but a decrease in contributions for 2007-2008.

In Europe, the research publications in CU, EV, IM, IPV, LV, LE, and TV are continuously rising up, but for DS and ID the contributions are reducing for all the time periods. The topics covered by IPE, PV gained popularity for 2005-2006, but the

contributions for them declined for the second time period. In Ed-Media conference, the research areas EV, IM, IPV, IET, LV, and TV are consistently growing whereas the research areas ID and LO are declining for all the time periods. The research areas IPE and PV followed the same pattern as mentioned above. The number of papers in the rest of the research areas decreased for the first time period, but got increased for the time period 2007-2008, except the TLP, where publications declined for the first time period and gained the same number of contributions in the second time period. In journals, CU, IPE, LV, and TLP are consistently rising up whereas EV, IM, and PV are facing reduction in number of publications for all the time periods. The research area DS and LE gained popularity for the time period 2005-2006, but the contributions in them declined for the time period 2007-2008.

In Asia, the research areas EV, ID, IET, and LV are growing while CU, DS, IPE, and PV are declining for all the time periods. The research areas IM and TLP grew up for the first time period, but got reductions in publications for the second time period. All other research areas faced drop in publications for the time period 2005-2006, but a gain in 2007-2008. In Ed-Media conference, EV, IET, and TLP are consistently gaining more publications while CU, DS, IPE, and PV are facing drop in contributions for all the time periods. The research areas ID and IM gained growth in the first time period, but a decline in the second time period. In journals, EV and TV are consistently rising up while LE is declining for all the time periods. The research topics covered by CU and TLP got rise in publications for the time period 2005-2006, but a drop in contribution for 2007-2008.

In Oceania, the research areas ID, IM, PV and LO gained popularity for all the time periods. The research topics covered by EV, IPE, and TLP gained growth in publications for the first time period, but a drop in contributions in 2007-2008. In Ed-Media conference, the research areas CU, IM, LO and PV got consistent decline for all the time periods. The number of contribution in EV, ID, IPE, and TLP increased for the time period 2005-2006, but faced a reduction in publications for the second time period. In journals, the research areas CU, ID, and IET are continuously gaining rise while IM, IPE, IPV, and TLP are falling for all the time periods. The number of papers for DS, EV, and LE grew up for the first time period, but the contributions dropped down in the second time period. Moreover, LE got no contribution up till 2004; similarly PV got no publication up till 2006. The research area IPE got no publication for the time period 2007-2008.

In Africa, the research areas IPE and TV are continuously gaining growth while TLP, IPV, CU, and LO are facing drop in publications for all the time periods. The topics covered by DS, EV, ID, and LE gained increasing number of publications for the

time period 2005-2006, but a reduction in publications for the second time period. The research area PV received only one contribution. In Ed-Media conference, the number of contributions in IPE, IPV, and TV are consistently rising up while CU, LE, and TLP are facing reduction in number of papers for all the time periods. The research areas DS, EV, ID, LO, and PV gained growth in the first time period, but a decline in the second time period. Moreover, the research areas IPE and PV got no contribution up till 2004. The research area PV again got no contributions for 2007-2008 like LO. In journals, the number of papers in CU and IM are dropping down for all the time periods. The research areas TLP and LE gained rise in publications for the first time period, but got reduction in the second time period. Moreover, LE, TLP, and PV got no contribution up till 2004 while EV and LO got publications only in 2003-2004. Similarly, the research area IET got no papers for 2005-2006 while CU, IPV, PV got no publications for the time period 2007-2008.

In South America, the research areas ID, CU and IPE are getting consistent growth in publications while LO, LE, and TLP are declining for all the time periods. The research areas PV and EV gained popularity in the first time period, but then declined. The research area EV got no publications up till 2004. Similarly, IPV and IET got no publications for the time period 2005-2006. In Ed-Media conference, the research topics covered by CU, ID, and IPE gained continuous rise in contributions whereas LE faced reduction in publications. The research areas LO, PV, TV, and TLP gained growth in first time period while a decline in second time period. The research area EV got no contribution till 2004 and got equal number of publications for rest of the periods. The research areas IPV and IET got no contributions for 2005-2006 similarly PV got no contribution for 2007-2008. In journals, ID is rising up while IM rose for the first time period and then declined for the time period 2007-2008. The research area EV got no publications at all. The research areas CU and TLP got no contributions for 2007-2008 while the research areas LE, PV, IET, IPV, and IPE got no papers till 2006.

Countries View

Figure 3.16 demonstrates the countries view. It was observed that USA (2003-2004: 758; 2005-2006: 532; 2007-2008: 591) remain the most contributing country for all the time periods followed by UK (2003-2004: 208; 2005-2006:161; 2007-2008: 274), Taiwan (2003-2004: 106; 2005-2006: 156; 2007-2008: 190), Canada (2003-2004: 99; 2005-2006: 152; 2007-2008: 172), Australia (2003-2004: 152; 2005-2006: 107; 2007-2008: 153), and Japan (2003-2004: 104; 2005-2006: 100; 2007-2008: 150). Interestingly, Canada was

behind Germany, Japan, and Australia up till 2003-2004, but for the rest of the periods it contributed more and succeeded other countries. A similar phenomenon occurred where Taiwan started to publish more than Australia after the time period 2003-2004.

In Ed-Media conference, the USA (2003-2004: 684; 2005-2006: 440; 2007-2008: 451) remain the most contributing country for all the time periods followed by Canada (2003-2004: 87; 2005-2006: 140; 2007-2008: 139), Japan (2003-2004: 104; 2005-2006: 96; 2007-2008: 144), Australia (2003-2004: 136; 2005-2006: 83; 2007-2008: 113), UK (2003-2004: 94; 2005-2006: 64; 2007-2008: 131), and Taiwan (2003-2004: 69; 2005-2006: 119; 2007-2008: 99), respectively.

In journals, the UK (2003-2004: 114; 2005-2006: 97; 2007-2008: 143) remained the main contributor followed by USA (2003-2004: 74; 2005-2006: 92; 2007-2008: 437),

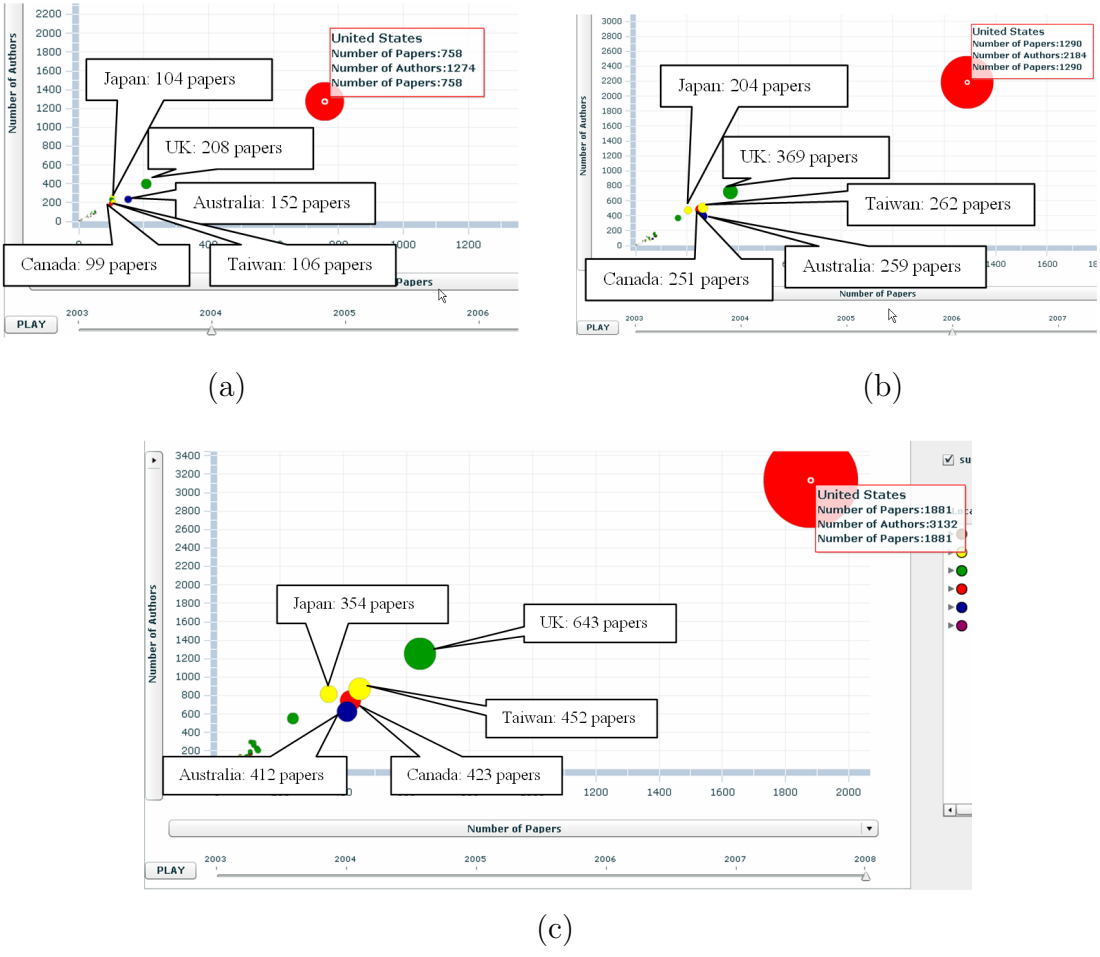


Figure 3.16 Countries View for e-learning journals and Ed-Media conference: (a) Publications up to 2004; (b) Publications up to 2006; (c) Publications up to 2008.

Taiwan (2003-2004: 37; 2005-2006: 37; 2007-2008: 91), Australia (2003-2004: 16; 2005-2006: 24; 2007-2008: 40), Netherlands (2003-2004: 24; 2005-2006: 20; 2007-2008: 29), and Canada (2003-2004: 12; 2005-2006: 12; 2007-2008: 33).

In USA, UK, Canada, Taiwan, and Australia, the DS, ID, and IM are the top most researched areas. Moreover, all these countries follow the same pattern in terms of top research areas in Ed-Media conference and journals.

Institutions View

It was observed that "University of Wollongong, Australia" (2003-2004: 24; 2005-2006: 15; 2007-2008: 17) is continuously contributing publications in the field followed by "Edith Cowan University, Australia" (2003-2004: 21; 2005-2006: 12; 2007-2008: 14), "National Taiwan Normal University, Taiwan" (2003-2004: 11; 2005-2006: 19; 2007-2008: 10), "Teachers College, Columbia, USA" (2003-2004: 11; 2005-2006: 11; 2007-2008: 11), and "University of Alberta, USA" (2003-2004: 19; 2005-2006: 7; 2007-2008: 4), respectively. Moreover, it was observed that Asian institutions are competing with Western institutions and taking active part in the field of e-learning. Furthermore, it was observed that in Ed-Media conference the same institutions are again the most contributing institutions in the field. However, in journals, "Anglia Polytechnic University, UK" (2003-2004: 8; 2005-2006: 7; 2007-2008: 1), and "Institute of Educational Technology, The Open University, UK" (2003-2004: 5; 2005-2006: 5; 2007-2008: 5) are the top contributing institutes.

3.6 Conclusions

Scientometrics and content analysis has been a tradition of many electronic and printed journals to ensure quality and journal's standing. Much can be learned about a field of study using such analysis. It can be used to know the impact of decisions and policies made for allocating resources and funds, and proposing the future directions for the field. Moreover, it reduces the researchers' menial efforts to conduct their surveys themselves and shows them a broader picture of their field of interest. Traditionally, such analysis has been conducted using normal tables and statistical charts. Recently, the researchers in the field of information visualization have also proposed many approaches to support such analysis. In this chapter, our work is also along these lines where we adapted a simple visualization technique (based on Gapminder) to conduct scientometrics and content anaylsis of scholarly communications. We first apply this technique to the Journal of Universal Computer Science (J.UCS) as an assistive tool

to strengthen its internal administration. In the second study, we employ the visualization tool with few improvements to five journals and two conferences in the field of e-learning. The second study will also allow novice and experienced educators, researchers and policy makers to understand what kind of different research areas exist in the field of e-learning, and to identify different research trends over the last six years using the visualization tool. Our experimentations conclude that the adopted visualization system is a powerful tool in determining the impact, coverage and the status of the journal at deeper level. A detailed usability study of the adapted visualization technique can be found in [Robertson et al., 2008].

4

Exploring Citations for Conflict of Interest Detection in Peer Review System*

Peer review in scientific communications plays an important role in the advancement of any given field of study. However, different sorts of conflict of interest (COI) situations between authors and reviewers can compromise the review decision. Current COI detection systems primarily rely on co-authors networks, inferred from publicly available bibliographic databases as an implicit measure of collaborative and social relationships between researchers. However, different citations relationships have also been claimed to be indicative of various social and cognitive relationships between authors. This can be useful for improving existing COI detection techniques by highlighting those hidden relationships that can not be handled by traditional systems. To prove this hypothesis, in this chapter, we first present the potential of different citations relationships to highlight the existence or non-existence of social relationships between authors. In this context, we used basic citations relationships, i.e., co-citations, bibliographic coupling, inter-citations, and temporal information associated with these relationships as features to predict the social networks of our selected sample of authors. Our experiments shows that our defined features identified these social networks as best as with 0.80 precision and 0.99 recall for non-sparse data and with 0.79 precision and 0.05 recall for sparse data. In the second part of this chapter, we used different citations relationships as a potential indicator of different types of cognitive COIs between researchers. We discuss possibilities to assign weights to these cognitive relationships to reveal the strength of cognitive COIs. As a case study, we assigned these weighted

*The material presented in this chapter is partially based on my previously published papers in [Khan, 2010, Khan, 2011]

cognitive relationships to our selected authors and reviewers from WWW2006 conference performance track. We found various cases where authors and reviewers do not have any apparent social relationships but they are strongly associated to each other through cognitive relationships. These strong relationships might give an impression of cognitive COI between authors and reviewers. To highlight the severity of these possible COIs, we described different contexts and sentiments that can be assigned to these cognitive relationships. In literature, researchers have always tried to assign contexts and sentiments to inter-citations and only to a single case of co-citations, i.e., "alternative or competitive work". In this chapter, we present a scheme based on existing theory to assign sentiments to even bibliographic coupling. Moreover, we use extended set of contexts and sentiments for co-citations. In this chapter, we also report our experiments for automated prediction of context and sentiments associated with any cognitive relationship for our WWW2006 authors and reviewers. As we used extended scheme for co-citations context and sentiments, we defined various features that can be used for the automated prediction of these contexts and sentiments. Finally, we assigned the context and sentiments to our selected authors and reviewers with very high cognitive relationships to reveal the severity of cognitive COI between them. Although in our reported results we did not find any severe case of cognitive COI, we believe that such analysis might help in other situations.

4.1 The Peer Review System

The peer review of manuscripts in journals and conferences is considered as a basis for the advancement of any discipline. The long history of peer review systems for academic journals goes back to at least 17th century [Kundzewicz and Koutsoyiannis, 2005]. The objective reviews of experts and knowledgeable researchers ensure the quality of the paper to be published and serve to set the standards in a particular field. Although the peer review has been criticized for many reasons such as: lack of objective measures, breach in secrecy, conflict of interest and delays in review time, it is widely accepted among scientific community because people seek some form of assurance that the published reports are authentic [Rennie, 1993]. Other forms of scholarly communications such as pre-print repositories also exist. However, without explicit and authentic certification the credibility of the work is primarily judged by readers themselves which is an extra burden for the community [Rodriguez et al., 2006]. In the literature various types of peer review models have been proposed to overcome these deficiencies. These models broadly vary from complete blind review to

full open reviews [Kundzewicz and Koutsoyiannis, 2005]. According to the authors in [Kundzewicz and Koutsoyiannis, 2005], the most widely used option adopted among scholarly communities is half blind review, where names of the authors are known to reviewers, but the names of the reviewers are unknown to authors. The authors further pointed out that although this system has proved to be workable, it is prone to some problems that include: subjectivity, bias, abuse, frauds and misconduct. The authors claimed that open peer review tries to overcome some problems of half blind review such as: bias and abuse by declaring names of both authors and reviewers. However, the reviewers in most of the cases are reluctant to expose their identity due to various reasons, e.g., writing negative about a manuscript written by someone in power or friend/colleague, to protect reputation in cases where inadequate or superficial reviews have been done due to time constraint or uninteresting topic [Kundzewicz and Koutsoyiannis, 2005]. In a study conducted by Dolan [Dolan, 2001] for *Aquatic Microbial Ecology journal*, the author found that 54% of the reviewers prefer anonymity while only 8% were ready to expose their identity. Another peer review model consisting of complete blind or double blind review is believed to fix the problems of bias and discrimination by hiding the names of both authors and reviewers from each other [Kundzewicz and Koutsoyiannis, 2005]. However, this method is costly and difficult to implement, and by removing some lines about the identity and affiliation of authors from the manuscript is not sufficient [Kundzewicz and Koutsoyiannis, 2005]. The authorship of a paper in some cases can be guessed by hidden information in terms of self-citations or sentences about previously published work, which can not always be removed from the manuscript [Kundzewicz and Koutsoyiannis, 2005]. In some cases, the authors and reviewers are working on the same problem and know each other in advance. These scenarios can be exemplified by a real life experiment conducted for the "British Medical Journal", where the reviewers were able to identify anonymous authors of manuscripts in 42% of the cases [van Rooyen et al., 1998]. With the advent of World Wide Web, a new concept of interactive journals is emerging [Pöschl, 2004]. The interactive journals employ two step procedure where in first step the submitted manuscript is discussed in an open forum by the community. After a thorough discussion and number of revisions the manuscript is refined rigorously and in the next step the manuscript is submitted for the standard peer review system. By engaging a large number of community members, this system can greatly reduce the reviewers' workload and can provide diverse evaluations for author. However, this system has the tendency to overwhelm authors with too many superficial and redundant reviews [Rodriguez et al., 2006]. Furthermore, the researchers sometimes show

unwillingness to deal with such pre-prints that have not passed quality control yet [Kundzewicz and Koutsoyiannis, 2005].

4.2 Conflict of Interest in Peer Review System

In any peer review system, reviewers' identification has always remained a challenging task to review a manuscript. The editors and conferences organizers usually rely on their personal knowledge, literature search and professional networks to select appropriate reviewers for submissions [Rockwell, 2010]. The expertise of the reviewer in the relevant field is the most important selection criteria [Rockwell, 2010]. A number of handful algorithms [Dumais and Nielsen, 1992, Basu et al., 2001, Yarowsky and Florian, 1999] in literature have also been proposed to automate reviewers' identification. These algorithms usually rely on matching referees' research interests and contents of the submission. Recently, authors in [Rodriguez and Bollen, 2008] introduced a robust algorithm that utilizes the co-authors networks in references of a manuscript and proposes potential reviewers by assigning each of them a context-sensitive weight. During the peer review process, the reviewers sometimes are presented by an awkward situation known as "conflict of interest" that might compromise the objectivity of review [Rockwell, 2010].

The Conflict of Interest (COI) can be broadly defined as *"a situation in which personal interests could compromise, or could have the appearance of compromising, the ability of an individual to carry out professional duties objectively"* [Biaggioni, 1993]. The presence of COI between authors and reviewers in the context of peer review can influence the decision of a reviewer. In the literature many types of COIs between an author and a reviewer have been identified which can be broadly classified in two categories, i.e., Social and Cognitive. However, the boundary between these categories is blurred and not always neatly separable. The social COI situations impose some degree of acquaintanceship between authors and reviewers such as: same affiliation, collaborators, colleagues, friends, family members, financial relationships, employer and employee, people in power, and even disliked people [Rockwell, 2010]. The cognitive COI on other hand depends upon the cognitive contents of the reviewer while reviewing a manuscript. A strong personal, ethnic, religious belief can effect the evaluation of a manuscript [Rockwell, 2010]. Similarly, researchers in some cases promote their own field and give favor to work that confirms their hypothesis or theory and may decline any competitive work.

4.3 COI Detection Approaches

The COI detection problem is usually addressed manually on the basis of declarations from the reviewers or authors. The process of currently available automated COI detection systems usually involves social network analysis of authors and reviewers. These social networks are typically derived from the collaborative information of authors, which is explicitly available in the form of co-author, co-editor and co-affiliation relationships in publicly available bibliographic databases. For example, the system introduced by [Papagelis et al., 2005] uses the suffix of email addresses in addition to previous co-authorship relations inferred from DBLP (Digital Bibliography & Library Project) as a measure to determine potential COIs. Similarly, the authors in [Aleman-Meza et al., 2008] integrated social networks of researchers from DBLP and FOAF (friend of a friend) documents by using ontologies to disambiguate authors, and developed an algorithm for the detection of possible COIs. But the problem with these automated approaches is that they consider only certain COI situations such as: co-authors and co-affiliations and ignores other types of COIs. Moreover, they are based on a limited portion of co-authors inferred from publicly available databases as all papers from a particular author are not necessarily indexed by these databases. Some social networking websites, e.g., LinkedIn.com, MySpace.com, Facebook.com can also provide implicit or explicit social information of people to detect COIs, but the integration and privacy concerns of these sites puts a limitation to utilize this enriched opportunity [Aleman-Meza et al., 2008]. The authors in [Matsuo et al., 2006, Mori et al., 2006] introduced various automated and semi-automated approaches to extract social networks of academic researchers by querying the web. These methods are not feasible for large number of entities pairs due to the cost associated with text analysis of large number of web pages. Although the link analysis on a network of homepages is another possibility that can be utilized to predict the communities of people and the context of their relationships [Adamic and Adar, 2003], but finding people homepages is challenging and it is not necessary that every person has a homepage and that it contains links to other people [Li and Wu, 2008]. However, some bibliographic digital libraries such as CiteSeer [CiteSeer, 2009] often present other attributes of a particular author that can be explored for COI detection. One of the most interesting components is the citation relationship.

4.4 Citations Theory

Citations were first used as a unit of analysis in the field of bibliometrics and scientometrics to evaluate the performance of individuals, journals, departments, research laboratories and nations [Garfield, 1970, Garfield, 1972, Oppenheim, 1997, Bormann and Daniel, 2008]. Although some researchers believe the applicability of citations counts as an implicit measure of intellectual and scientific impact, there are several studies that doubt its use. This is due to the dependence of citations counts on various factors such as: time, field, journal, article type, language, and availability. [Bormann and Daniel, 2008]. However, the central problem in using citations counts is due to its lack of capability to highlight the intentions and motives of the citers [Cronin, 1982]. According to this camp of researchers, the use of citations counts as a measure of scientific impact is only applicable if the citing author has really used the cited document and citation is truly depicting its significance and quality [Bormann and Daniel, 2008].

Authors often cite each other due to various reasons such as: related work, competitive work, extension of previous work, or disclaiming others' work to name a few. The earliest work listing the motivation of citers was published by Garfield in 1962 [Garfield, 1962]. The motive behind citations has always remained debatable between researchers. The citations between authors are usually considered to be representative of intellectual influence [Baldi, 1998, Kurtz et al., 2004]. However, the authors in [Cronin and Shaw, 2002, Johnson and Oppenheim, 2007] found that the repetitive citations can also highlight various social acquaintanceships between authors. This might be due to the fact that scientist with similar interests or subject specialty usually collaborate, communicate and support each other to work towards important goals in a particular discipline, one output of which is inter-citation [White et al., 2004]. In this context, the notion of "invisible college" is really important where scientists (even geographically distant) gather together to achieve specific tasks by using both formal and informal communications [Zuccala, 2006]. With the advent of new technologies and concepts for instant communications such as: blogs, wikis, file sharing, instant messaging, emails, open access initiatives, these invisible colleges are increasingly emerging. Cronin [Cronin, 2005] emphasized the social dimension of citations motive as follows:

"there is a battery of social and psychological reasons for citing, which may have as much to do with, for instance, rhetorical gamesmanship (persuading the reader of one's viewpoint through selective under- or over-citation) or strategic coat-tailing (citing friends, immediate colleagues or celebrity authors) as with the topical appropriateness or semantic suitability of the citations themselves".

Half a century ago, Kessler [Kessler, 1963] and Small [Small, 1973] introduced the notion of bibliographic coupling and documents co-citation as a measure to determine documents with similar topics. In [White and Griffith, 1981], the authors introduced a new technique called authors co-citations to understand the intellectual structure of a discipline by grouping co-cited authors together, who work on similar themes as seen by citers. Recently the authors in [Zhao and Strotmann, 2008] studied author's bibliographic coupling as a complementary approach of author's co-citations to reveal the current internal structure of a discipline by grouping authors thematically. The authors' co-citation studies have also been claimed to be representative of social relationships between pairs of authors [Rowlands, 1999], while authors' bibliographic coupling until now has only been studied from the perspective of cognitive distance [Zhao and Strotmann, 2008].

In the context of COI detection, one can conclude from the discussion of this section that different citations relationships between authors have the capability to highlight the possibility of both cognitive and social biases in peer review system.

4.5 Citations as Predictor of Socio-Cognitive Relationships

The citations and social relationships of authors often overlap to some extent usually due to socio-cognitive ties between authors [White et al., 2004]. This overlap can be depicted by a hypothetical Venn diagram as shown in Figure 4.1. The socio-cognitive is a special term used by White [White et al., 2004] to describe the relationship between any two authors, where both authors have intellectual as well as some kind of social relationship with each other. Co-authors, colleagues, student/mentor and editors/contributors are few examples of socio-cognitive ties.

This section follows this direction and explores to discover any pattern in citations relationships that can act as a predictor to identify socio-cognitive relationships. The current investigation is limited to two types of socio-cognitive relationships, i.e., co-authors and co-affiliation/collegial relationships. The results of this study in turn can help in improving existing COI detection approaches by exploiting citations as an additional or alternative means to determine socio-cognitive relationships between authors and reviewers.

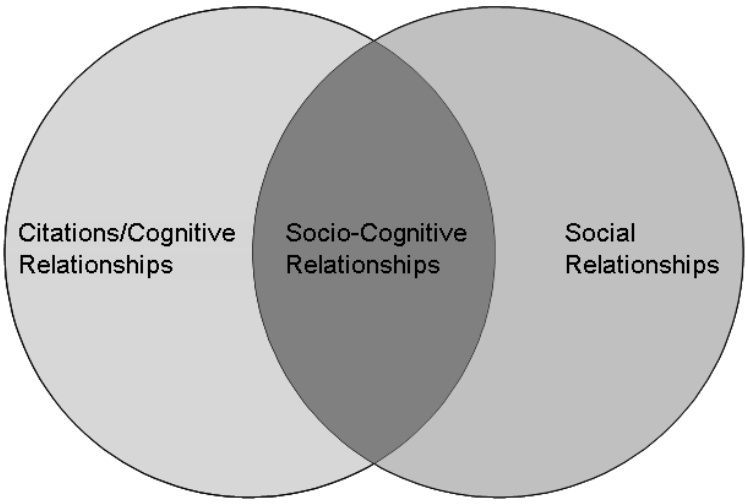


Figure 4.1 Structure of social, citations/cognitive and socio-cognitive relationships.

4.5.1 Design of the Study

Citations and Socio-Cognitive Measures

In this study, different citations measures have been used, i.e., co-cited, co-cites and cross- cites. Theses measures will be referred as basic citations measures in the rest of this study. The details about these measures are as follows:

- Co-Cited. The co-cited is the frequency that two authors have been cited together in literature, independent of the contents of the cited documents.
- Co-Cites. The co-cites is the number of times that two authors cite together one or more documents. It is similar to bibliographic coupling [Kessler, 1963], but instead of documents, authors have been taken as a unit of analysis.
- Cross-Cites. The cross-cites as its name implies represents the asymmetric number of citations that any particular author has given to any other author. There are two kinds of cross-cites relations that have been used in this study, i.e., from "primary author" to "secondary author" and vice versa. The primary authors are those randomly selected authors for whom various citations and socio-cognitive relationships have been computed. The secondary authors represent those authors that have any citations relationships with primary authors. Further details about both primary and secondary authors can be found in the forth coming sub-sections.

Two kinds of socio-cognitive relationships have been considered in this study, i.e., co-authors and co- affiliation. The details about these relationship are as follows:

- Co-Affiliation. The co-affiliation relationship symbolizes whether any two authors have ever been associated with the same organization or institution.
- Co-Authors. The co-authors relationship is further categorized in two categories, i.e., direct co-authors and indirect co-authors. The direct co-authors relationship represents whether any two authors have ever published a paper together. The indirect co-authors relationship on other hand represents the existence of any common collaborator/co-author between two authors.

These socio-cognitive relationships will be used as ground truth for the classification experiments in Section 4.5.2.

Selection of Datasets

In order to determine citations and socio-cognitive relationships, a free publicly available bibliographic data about publications has been used from CiteSeer as the primary input for the experiments. CiteSeer contains approximately 700,000 papers from computer and information science disciplines. It contains both inward (cited) and outward (citing) citations information, but only for those papers that are indexed in CiteSeer. There were only 337,118 unique papers (approx. 48%) that have outward citations and 196,134 unique papers (approx. 28%) having inward citations. CiteSeer also indexes the affiliations and location information of authors. We further noticed that several papers have duplicated copies in CiteSeer, for the same year. We removed these duplicate copies based on the corresponding authors' names information, resulting in approximately 550,000 papers. Similarly, we further normalized the papers references by removing the duplication of referenced papers for any citing paper. This resulted in only one reference "to" a paper "by" a particular paper. We performed this step because it is time consuming to ensure that the duplicated references were due to the data entry mistake or due to the multiple referenced sentences to a paper by the citing paper.

In order to conduct the experiments where most of the citations, coauthors and affiliation information are available, 20 random authors were selected based on the following criteria: authors must have minimum of 10 papers, 10 co-authors, 10 inward citations, 10 outward citations, and at least one affiliation information. These authors will be referred as primary authors in the rest of this study. As peer reviewers are usually experts in a given domain, it is expected that they can easily meet these

Table 4.1 List of randomly selected primary authors for experiments.

Sr. No.	Name	Co-Authors	Papers	Inward Citations	Outward Citations
1	Micha Sharir	64	188	1234	949
2	Marc Moonen	69	24	271	333
3	Wim H. Hesselink	24	37	46	48
4	Rainer Lienhart	35	35	126	83
5	Franz Baader	58	141	125	804
6	Peter Bro Miltersen	50	74	242	187
7	Minyue Fu	42	58	45	69
8	Panos Constantopoulos	32	116	272	543
9	Jian Shen	21	31	48	41
10	Prabhakar Raghavan	95	191	1721	542
11	Sanjoy Baruah	33	56	135	323
12	M. Tamer	44	102	265	282
13	Tapas Kanungo	42	61	167	184
14	Ljubomir Josifovski	16	17	43	63
2	Ellen W. Zegura	42	100	1053	407
16	Eyal Kushilevitz	44	120	718	823
17	Jennifer Seberry	67	160	310	268
18	Remzi H. Arpaci-dusseau	25	54	79	579
19	Ferenc A. Jolesz	24	63	223	136
20	B. R. Badrinath	49	93	1411	540

criteria. The Table 4.1 shows these primary authors and their corresponding selection attributes.

Citations and Socio-Cognitive Measures Calculation

In the first step, the papers that belong to randomly selected authors were separated from CiteSeer. Next, all the authors having any citations relationship with primary authors were determined. These authors will be referred as secondary authors in the rest of this study. The frequency of citations relationships of primary authors with secondary authors, i.e., co-cited, co-cites, cross-cites from primary to secondary author (cross-cites$_{ptos}$) and cross-cites from secondary to primary authors (cross-cites$_{stop}$) were computed. The numbers of secondary authors having any citation relationship with primary authors are summarized in Table 4.2.

Table 4.2 Number of authors having any citations relationship with primary authors.

Co-Cited	Co-Cites	Cross-Cites$_{ptos}$	Cross-Cites$_{stop}$	Total unique secondary authors
53,570	124,163	4,880	8,282	158,728

Table 4.3 Number of authors having both citations and socio-cognitive relationships with primary authors.

Citations and direct co-authors	Citations and affiliation	Citations and indirect co-authors	Total unique authors
1,116	2,651	11,643	12,843

In the next step, the secondary authors that also have any socio-cognitive (co-affiliation, direct co-authors, indirect co-authors) relationship with primary authors were determined. The affiliations information of primary and secondary authors was matched using Q-Gram [Ukkonen, 1992] string distance measure with a threshold of 0.90, which was chosen empirically. In order to increase the accuracy of the affiliation names matching, stop words and keywords such as: "university", "college", "school", "institute", "department" were avoided in determining similarities. As CiteSeer indexes only limited papers, the additional co-authors information has been extracted from DBLP, which contains approximately 1,940,000 bibliographic records from computer science discipline. In order to retain only original articles, the titles that correspond to "proceedings", "symposiums", "home page" and "workshops" were removed from DBLP. Moreover, DBLP contains very little citations (8232 outwards and 21,391 inwards) and affiliation information of authors, which are not included in the experiments. The number of secondary authors having both citations and socio-cognitive relationships are shown in Table 4.3.

From the various calculated citations and socio-cognitive measures, it was noticed that the probability of the existence of socio-cognitive relationship increases with the increase in the strength of citations relationships as shown in Figure 4.2. The probability even approaches more than 90 percent in the case of co-cited and cross-citations, which is quite encouraging for the development of a predictor based on citations relationships to highlight socio-cognitive relationships.

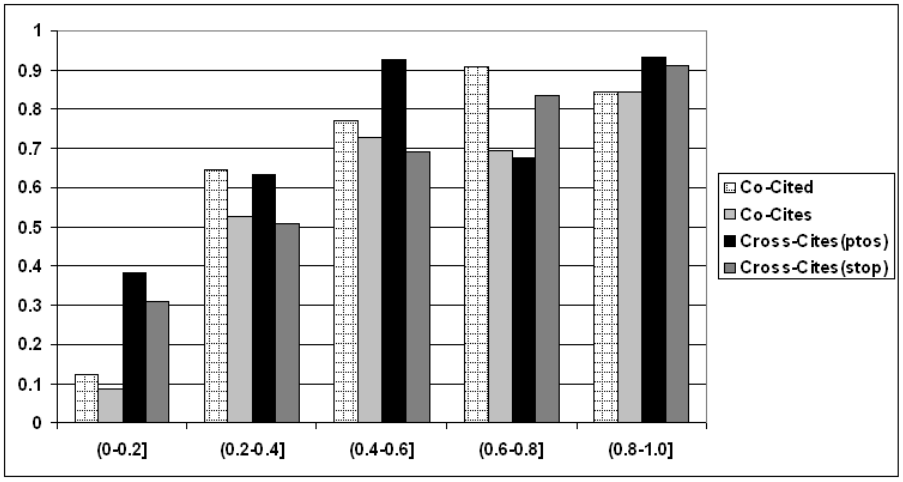


Figure 4.2 Probability of socio-cognitive relationships. X-axis: normalized citations counts, Y-axis: probability.

4.5.2 Experimental Results

For the different citations measures that were computed from the corpus, decision tree (J-48) and Support Vector Machines (SVM) classifiers were trained and tested using WEKA [Weka, 2009] to predict the existence or non-existence of socio-cognitive relationships. The decision tree was chosen because of its strong capability to classify instances by branching at different values of the features. Similarly, SVM which is based on statistical learning theory has received considerable attention these days and has shown promising results in many classification problems [Chapter 5 of CS445 in Yale, 2005]. In our experimentations, we used nonlinear SVM, which basically transforms the input features in a high dimensional space via kernel trick and creates a maximum-margin hyper-plane between them to differentiate the instances of different classes. We used Radial Basis Function (RBF) kernel for SVM and LIBSVM [Chang and Lin, 2010] library for SVM implementations which is also available as WEKA plug-in. The citations features belonging to each primary author were normalized ranging from 0 to 1 using the formula, i.e., X_{new}=(X - X_{min})/(X_{max} - X_{min}). There are also other normalization methods used in literature such as correlation, cosine similarity between two authors' citations relationships vectors. However, these approaches were adopted for limited number of authors' pairs and can be very costly in terms of computations for the current study. The target class or ground truth values in each classification experiment were given in the form of binaries, where class "yes" and class "no" represents the existence and non-existence of any socio-cognitive relationship respectively. In each classification experiment 10-fold cross validation were

used in WEKA. The final classification results obtained were evaluated using Precision, Recall and F-Measure, where precision can be defined as the proportion of instances which truly belong to class x among all those instances that are classified as class x. Similarly, recall is the proportion of instances that are classified as class x among all those instances that truly belong to class x. The F-Measure is simply a combined measure of precision and recall that can be calculated by the formula, i.e., (2*recall×precision)/(recall+precision). The purpose of F-Measure is to obtain a single measure to characterize the overall performance of a classifier for a particular class.

It was observed that the distribution of classes "yes" and "no" in this classification experiment are extremely unbalanced. Only 8% of total citations relationships have instances for class "yes". The input citations features are also observed to be sparse. The citations features are dense for approximately 10% of total overlapped socio-cognitive relationships. Due to the sparsity and lack of balanced dataset, it was decided to mainly focus in the training and testing of the classifiers for dense dataset where all citations features are available, and continue with the analysis of unbalanced and sparse dataset.

The Table 4.4 summarizes the performance of decision tree and SVM classifiers for class "yes" and class "no". It can be observed from the table that both classifiers performed adequately in terms of precision, recall and F-Measure for class "yes". However, the results of both classifiers are not satisfactory for class "no". It can be further noticed that the decision tree performed relatively better than SVM for both classes. The classifiers were also evaluated individually for direct co-authors and authors with similar affiliations, but none of them was found to be strong enough in terms of precision, recall and F-Measure. The results obtained for indirect co-authors were not too much different from the ones presented in Table 4.4. The possible reason for such results is due to the major proportion of indirect co-authors in collective socio-cognitive measures and substantial overlap with direct co-authors and authors with similar affiliations.

Table 4.4 Precision, recall and F-Measure for class "yes" and class "no" using basic citations measures.

Decision Tree				**Support Vector Machine**			
Precision	Recall	F-Measure	Class	Precision	Recall	F-Measure	Class
0.79	0.92	0.85	yes	0.79	0.86	0.82	yes
0.49	0.22	0.31	no	0.38	0.27	0.31	no

Extending Citations Features

After analyzing results from the experiments in the previous section, it was decided to include more citations based measures. An interesting set of measures associated with citations relationships is temporal information. It is expected that academics inter-cite, co-cite or get co-cited with social acquaintances in relatively shorter period of time after publishing a paper. Similarly, the raw count of unique papers that interconnect two authors through any citations relationships may also provide useful information. It is expected that social acquaintances are usually interconnected through more than one paper via any citation relationship. Based on these assumptions two extended sets of citations measures were defined that can be evaluated for classification in combination with basic citations measures.

The first group of measures is based on temporal information of citations. The details about these measures are as follows:

- Co-Cited Average Time. It is the average difference in the publication years of co-cited papers. However, it must be noted that if a particular paper A from one author is co-cited with more than one papers B_n of the other author. Then a paper B_i with minimum publication year will be selected for computing the difference with paper A. This measure was calculated for both primary authors and secondary authors resulting in two separate measures.

- Co-Cites Average Time. It is the average difference in the publication years of papers that co-cites together. If a particular paper A from one author co-cites with more than one papers B_n of the other author. Then a paper B_i with minimum publication year will be selected for computing the difference with paper A. This measure was calculated for both primary authors and secondary authors resulting in two different measures.

- Cross-Cite Average Time. It is the average of number of years when any author cites any paper of the other author for the first time. Similar to the basic citations relationships, this measure has been calculated from "primary author" to "secondary author" and vice versa, resulting in two separate measures.

The second group of measures is based on the unique papers that interconnect any two authors through any citation relationship. The details about these measures are as follows:

- Unique Papers Co-Cited. It is the number of unique papers of any author that has been co-cited with the papers of other author. This measure was calculated

for both "primary authors" and "secondary authors" resulting in two different measures.

- Unique Papers Co-Cites. It is the number of unique papers of any author that co-cites with the papers of other author. This measure was also calculated for both "primary authors" and "secondary authors" resulting in two separate measures.
- Unique Papers Cross-Cites. It is the number of unique papers of any author that cites the papers of other author. This measure has also been calculated for both "primary authors" and "secondary authors". Similar to the basic citations relationships, this measure has been calculated from "primary author" to "secondary author" and vice versa resulting in four different measures.

Table 4.5 Precision, recall and F-Measure for class "yes" and class "no" using basic and temporal citations measures.

Decision Tree				**Support Vector Machine**			
Precision	Recall	F-Measure	Class	Precision	Recall	F-Measure	Class
0.80	0.92	0.86	yes	0.80	0.99	0.88	yes
0.54	0.27	0.36	no	0.86	0.24	0.38	no

Table 4.6 Precision, recall and F-Measure for class "yes" and class "no" using basic and unique papers measures.

Decision Tree				**Support Vector Machine**			
Precision	Recall	F-Measure	Class	Precision	Recall	F-Measure	Class
0.81	0.89	0.85	yes	0.80	0.98	0.88	yes
0.51	0.34	0.41	no	0.81	0.21	0.34	no

The Table 4.5 and Table 4.6 summarize the performance of classifiers for both above mentioned groups in combination with basic citations measures. It can be observed from these tables that the performance of class "no" has significantly improved for SVM classifier. The classifier was able to identify instances of class "no" with more than 0.80 precision in both cases. However, the classifier was able to identify class "no" instances with 0.24 and 0.21 recall for temporal and unique papers based measures respectively. Similarly, the results for class "yes" in each case have also increased in

terms of recall (0.98-0.99) in the case of SVM. Furthermore, it can be observed that temporal information performed relatively better than unique papers based measures in terms of precision and recall for class "no". The decision tree on other hand again did not perform adequately for class "no" in terms of precision, recall and F-measure. The classifiers were also evaluated by combining all basic and extended citations measures as shown in Table 4.7. However, it did not result in any significant improvement for both decision tree and SVM classifiers. The performance of classes even declined as compared to the results of temporal based citations measures in case of SVM classifier.

In summary, although our classifiers were not able to identify all the cases for class "no", but they performed sufficiently for class "yes" and in terms of precision for class "no". After obtaining some considerable classification results as observed in Tables 4.5 to 4.7 for SVM classifier. We decided to train and test the SVM classifier for our complete dataset (unbalanced and sparse) with all citations features (basic and extended). The results of the classifications are summarized in Table 4.8. As it can be observed from the table that the classifier performed adequately for the instances of class "no" with 0.92 precision and 0.99 recall. This might be due to the extremely unbalanced class priors as mentioned earlier. Furthermore, it can be observed that the classifier was able to identify instances of class "yes" with only 0.05 recall, but with 0.79 precision.

Table 4.7 Precision, recall and F-Measure for class "yes" and class "no" using all citations measures.

Decision Tree				**Support Vector Machine**			
Precision	Recall	F-Measure	Class	Precision	Recall	F-Measure	Class
0.82	0.91	0.86	yes	0.80	0.98	0.88	yes
0.57	0.36	0.44	no	0.84	0.24	0.37	no

Table 4.8 Precision, recall and F-Measure for class "yes" and class "no" using all citations measures.

Support Vector Machine			
Precision	Recall	F-Measure	Class
0.79	0.05	0.09	yes
0.92	0.99	0.95	no

Apart from our original hypothesis, we also used similar venues and journal titles information, text similarity of paper titles and abstracts (we used cosine vector [Salton and McGill, 1983] model for text similarity), location (city and country) in addition to citations information for our classification experiments, but the results did not provide any significant improvements. Similarly, we also experimented to classify the instances of direct co-authors and indirect co-authors from other instances based on their collaboration strengths as used in [Aleman-Meza et al., 2008], but that did not have significant results for discussion.

From these experiments, it can be concluded that the possibility of using citations to automate the process of potential socio-cognitive relationship detection, one can only identify some proportion of possible cases with considerable precision. However, there are many other social relationships such as: friends, allies, regular correspondents, sought advices that are not considered in this study might further improve the results.

4.6 Citations as a Measure of Cognitive Distance

4.6.1 Selection of Dataset

As we discussed in Section 4.4 different citations relationships can be indicative of both social and cognitive ties between authors. This section is an effort to explore the applicability of citations as a potential indicator of cognitive conflict of interest in peer review system. In order to demonstrate and analyze the effectiveness of using citations as a potential indicator of cognitive distance, we used the subset of authors and reviewers from the WWW2006 conference's performance track. We used the same CiteSeer database as mentioned in Section 4.5.1 to compute the frequency of different citations relationships, i.e., co-cited, co-cites and inter-citations for both authors and reviewers. To further understand the applicability of citations based cognitive distance measures, we also computed the co-authors network of reviewers up to two degrees, i.e., direct co-authors and indirect co-authors (co-authors of direct co-authors) from CiteSeer and DBLP.

4.6.2 Weighting Citations Relationships for Cognitive Distance

Traditionally, in authors' co-citations and bibliographic coupling, the strength of cognitive relationships has always been computed using the Pearson product-moment correlation coefficient between authors' pairs. However, the authors in [Ahlgren et al., 2003] highlighted the disadvantages of this approach by demonstrating the effects of adding

zeros in raw co-citation counts matrix with both hypothetical and real life data. They found that the correlation coefficient value between a pair of authors decreases with the inclusion of those authors in the matrix that do not have been co-cited with both authors. They recommended researchers to choose an appropriate association measure depending on the nature of the problem under investigation. Similarly, in the context of the COI detection, the association measures like correlation coefficient, Salton's cosine [Salton and McGill, 1983] and Jaccard [Leydesdorff, 2008] measure between authors and reviewers may not be feasible. The reason behind this rational is that the similarity score of an author and reviewer will be low if both are even co-cited together frequently, but simultaneously co-cited with a complete or partial disjoint set of other authors. This can be explained with a simple hypothetical example in Table 4.9, where Ai represents an author and R1 represents a reviewer. The results of the different similarity measures between an author A1 and reviewer R1 can be summarized in Table 4.10, which appears to be very low even with a high co-citation rate between A1 and R1.

Based on the results in Table 4.10, it was decided to use standard normalization formula, i.e., X_{new}=(X - X_{min})/(X_{max} - X_{min}) to compute the cognitive distance between authors and reviewers. The adopted approach has the capability to assign an appropriate score to the cognitive similarity between authors and reviewers in relation to other authors. This can be confirmed by the same hypothetical example in Table 4.9. The cognitive similarity of A1 with R1 for this particular example is equal to 1 and

Table 4.9 Hypothetical raw citation relationship matrix (5 authors and 1 reviewer in the sample).

	A1	A2	A3	A4	R1	A5
A1	-	2	0	2	55	12
R1	55	0	6	10	-	0

Table 4.10 Similarity counts.

Similarity Measure	**Similarity Score**
Pearson correlation	-0.55
Cosine Similarity	0.01
Jaccard Index (We used Tanimoto as data is non-binary [Tanimoto, 1957])	0.003

vice-versa. Moreover, it was observed that the normalized similarity score from only reviewer's side might be sufficient. Because it is the reviewer who has to make the final decision, and normalizing any type of citation relationship in this way can depict how close the author is working in domain of the reviewer in comparison with other authors.

The Table 4.11 summarizes the results of assigning normalized citations counts between our selected reviewers and authors of WWW2006 along with the type of the citations relationships. It can be noticed from this table that there are significant cases where reviewers and authors do not have any visible social relationships in terms of co-authors network, but have strong intellectual ties. For example in the case of "Alec Wolman" and "Balachander Krishnamurthy", the reviewer is citing the author a number of times, but apparently do not have any social tie. This may imply that the reviewer is already aware of the author's work and influenced with his research methods and materials. Similarly, in the case of "Michael Rabinovich" and "Craig E. Wills", the author and reviewer appear to be working in a close research area due to high bibliographic coupling between them and substantial citations for reviewer's work from the author. Additionally, they have not collaborated with each other in terms of publications, but they are inter-connected with each other through a common collaborator. Another interesting case is about "Alec Wolman" and "Amin Vahdat" where the author and the reviewer have never published a paper together, but they are citing each other at a significant rate, implying that they know each others work in advance. Finally, the cases where cognitive similarity is not very significant can be ignored.

Although Table 4.11 has highlighted various cases of cognitive similarity between authors and reviewers, but an analysis of the citations context by an expert or an automated system can further elaborate the meanings associated with these citations relationships. This in turn can help in identifying the severity of the possible conflict of interest between authors and reviewers. The next section discusses in detail possible citations contexts and abstract classes of sentiments that can be assigned to our identified citations relationships. It also reports on our experiments for the automated classification of these citations contexts. Finally, we discuss some results after assigning these citations contexts to our WWW2006 authors and reviewers who have significant frequency of citations relationships between each other as mentioned in Table 4.11.

4.6.3 Related Work for Citations Context Identification

In the literature, there are number of studies that describe the reasons why an author has cited other author. One of the earliest work in this direction was done by Garfield

Table 4.11 Normalized citations relationships count between authors and reviewers of WWW2006 performance track.

Authors	Revi. Reviewers	Martin F. Arlitt				Jeffrey S. Chase				Michael Rabinovich				Oliver Spatscheck				Maarten Van Steen				Alec Wolman			
		cd	cs	co	ci	cd	cs	co	ci	cd	cs	co	ci	cd	cs	co	ci	cd	cs	co	ci	cd	cs	co	ci
Balachander Krishnamurthy		0.18	0.08	0	0.2	0.01	0.01	0	0	0.45	0.26	0.45	0.16	0.04	0.006	0.2	0	0.01	0.02	0.02	0	0.06	0.24	0.37	0
Craig E. Wills		0.07	0.09	0	0.1	0.01	0.02	0	0	0.28	0.41	0.11	0.38	0.02	0.10	0	0	0.004	0.02	0.007	0	0.03	0.37	0.37	0
Tracy Kimbrel		0.01	0.03	0	0	0.05	0.09	0.14	0.02	0.02	0.002	0	0	0.004	0	0	0	0	0	0	0	0.01	0.003	0	0
Giovanni Pacifici		0	0	0	0	0	0	0	0	0.003	0.003	0	0	0.004	0	0	0	0	0	0	0	0	0	0	0
Mike Spreitzer		0.003	0	0	0	0.002	0.005	0.04	0	0.006	0.01	0	0	0	0	0	0	0.01	0	0.015	0	0	0	0	0
Patrick Reynolds		0	0.01	0	0.1	0	0.02	0	0.02	0	0.02	0	0	0	0	0	0	0	0.01	0	0	0.04	0.07	0.18	0
Amin Vahdat		0.09	0.11	0	0.3	0.10	0.61	0.43	0.30	0.08	0.33	0.04	0.04	0.07	0.19	0	0	0.04	0.12	0.02	0.02	0.13	0.56	0.31	0.25

cd: Co-Cited, cs: Co-Cites, co: Citations from reviewer, ci: Citations from author, dark gray cell: Direct co-authors, light gray cell: Indirect co-authors

[Garfield, 1962]. Garfield in his paper [Garfield, 1962], formally listed in passing, fifteen reasons for citing, but it is said to be the foundation of various citations classifications schemes developed later [Radoulov, 2008]. The first formal classification of citations was done by Moravcsik and Murugesan [Moravcsik and Poovanalingam, 1975], [Radoulov, 2008]. Their classification scheme contains four main categories with more than one sub-categories in each main category. This classification was done by using 702 citations used in 30 articles published from 1968 to 1972 in *Physical Review*. Later, various authors [Chubin and Moitra, 1975, Spiegel-Rosing, 1977, Oppenheim and Renn, 1978, Teufel et al., 2006a] developed and modified existing classification schemes depending upon their research hypothesis [Radoulov, 2008]. Similar to defining the classification schemes for citations, much of the efforts has also been done in the automated classification of citations contexts. Garzone [Garzone and Mercer, 2000], Nanba and Okumura [Nanba and Okumura, 1999] defined rule based schemes to automatically classify the citations [Radoulov, 2008]. Although, their classifiers work satisfactory, but defining such parsing rules is difficult and requires an expert knowledge in linguistic domain [Radoulov, 2008]. Similarly, another rule based classification system was developed by Pham and Hofmann [Pham and Hofmann, 2003], which is similar to decision trees [Radoulov, 2008]. The advantage of their system is that it does not require any knowledge engineer from linguistics, but relies on the knowledge of the domain expert in defining the rules for each node in the tree [Radoulov, 2008]. The authors showed that their system outperforms the methodologies of both Garzone and Nanba [Radoulov, 2008].

Teufel et al. [Teufel et al., 2006b] were the first to use machine learning techniques for the classification of citations [Radoulov, 2008]. They selected a subset of articles from a corpus of 360 conference articles for citations annotations by three annotators, according to the guidelines defined from another subset of articles. Despite the complexity and the number of citations categories, they found a significantly high inter-annotator agreement. They further identified number of features to be used by the IBk (k-nearest neighbor) algorithm for automated classification. These features include: 1762 cue phrases identified from 80 articles, two main agent types (author of current paper, and other people) modeled by 185 patterns, 20 manually acquired verb clusters, verb tense, modality, location of the citation in the article, section and paragraph, 892 additional cue phrases identified by annotators and self citations. The training and testing for citations classification was performed on 2829 citations instances extracted from 116 separate articles and achieved substantially significant results. In an other article by Teufel and Moens [Teufel and Moens, 2000], the authors noted a common rhetorical pattern in the introduction section of the articles. It was observed that the

articles usually start with sentences describing the general background of the research then some long sentences discussing the specific related work in a neutral way. This discussion is usually followed by sentences mentioning some limitations of the prior work and then the contribution of the current work. This observed syntactic structure has been utilized by many researchers to classify citations. Recently, Angrosh et al. [Angrosh et al., 2010] also used this rhetorical pattern to classify the citations sentences and even sentences adjacent to these citations with significantly high accuracy that appeared in the related work sections of 50 articles.

4.6.4 Citations Relationships Context Identification and Classification Experiments

In order to determine and demonstrate the automated classification of contexts associated with citations relationships between our WWW2006 authors and reviewers, we downloaded only those articles of reviewers and authors which are listed in our CiteSeer database, and had been utilized to determine cognitive distances in the previous section. The total downloaded articles were 472, some papers were not available online. The downloaded files were first converted into XML format. There were 57 papers that were scanned and could not be converted into XML. We then wrote small scripts to extract the citations sentences from these files using regular expressions. Our routines located the names of the cited authors in the references list and extracted the sentences containing those references. The typical references include [1], ABC et al. or (ABC, 2008), etc. For bibliographic coupling scenario, we also matched the cited paper titles to extract only those references which have been cited by any two author and reviewer associated through bibliographic coupling. As a result, we found 137 unique inter-citations sentences, 1006 unique citations instances for bibliographic coupling, and 51 unique co-cited instances. The whole parsing process was challenging because of typing errors. Also, in some cases the XML conversion was not in parsable form. Similarly, there were a few cases where cited author's name was mistakenly not mentioned in the references section. As we mentioned earlier in Section 4.5.1 we removed the duplication of references and each paper now contains only one citation for a particular paper. However, in some cases we found more citations sentences for the same reference in a paper when compared to CiteSeer database. However, for the computation of final results described in Section 4.6.5, we normalized the count of the additionally found sentences to unit one.

Classification Schemes for Citations Relationships

For our experiments, we used a modified version of the citations classification scheme of Teufel et al.[Teufel et al., 2006a]. One category, i.e., "strength" has been taken from [Angrosh et al., 2010]. We preferred this scheme because it is easy to operationalize without any explicit knowledge of the domain and can provide enough information for our COI application. For simplification, we decided to classify the citations only on the basis of context of the sentences that contain the citations. However, one can go further to locate pronouns and abbreviations of authors names and theories in other sentences, which is technically not possible for all the cases [Teufel et al., 2006a]. Similarly, the context of the citation can be identified at a paragraph level or at an article level. The detail of our adopted classification scheme is summarized in Table 4.12. Unlike previous work, we treated co-citations as a separate classification problem from inter-citations. This is due to the fact that sometimes a sentence can contain more than one citation, and it is important to discover the purpose of these citations and their inter-relationship with each other. For example, consider the sentence *"Emerging technologies such as PlanetLab [19] and ScriptRoute [22] may help enable these more detailed measurements"* [Vahdat et al., 2010]. In the case of inter-citations, the authors of the article are describing the strength of the cited work, but on the other hand in case of co-citation, both cited works appears to be similar. Similarly, previous studies used only one form of co-citation relationship, i.e., competing works or alternative works usually occurring in consecutive sentences. However, in our experiments we

Table 4.12 Inter-citations classification scheme.

Class	Description
Similar	Author's work is similar to the cited work.
Supports/Confirm	Author's work supports or confirm the cited work.
Strength	Author's work describes the strength of the cited work.
Weak	Author's work describes the shortcomings of the cited work.
Motivated/Extends	Author's work is motivated by the cited work.
Contrast	Author's work is in contrast/comparison with the cited work.
Uses	Author's work uses/modifies/adapts the cited work.
Neutral	Cited work is described in a neutral way, or enough textual information is not available.

consider only those co-citations which exist in a single sentence. Moreover, we show that co-citations can be classified in more categories similar to inter-citations scheme. As we mentioned earlier that we found only 51 co-citations sentences. We then decided to use the citations sentences from our inter-citations and bibliographic coupling corpus for defining co-citations context classification scheme and their automated classification experiments. In this collection, we found 233 unique instances of co-citations sentences. After a detailed analysis of this co-citations data, we used the scheme listed in Table 4.13.

Table 4.13 Co-citations classification scheme.

Class	Description
Similar	Co-Cited works are similar.
Uses	One work uses other work.
Motivated/Extends	One work extends or motivated by other work.
Contrast	One work is in contrast with other.
Neutral	Enough textual information is not available.

Results of Classification Experiments

We manually annotated all the citations and co-citations according to the defined classification schemes. The distribution of citations sentences among the citations context classes is summarized in Table 4.14. In defining the features for automated classification experiment, we followed the set used by Angrosh et al. [Angrosh et al., 2010]. We extracted cue words and phrases from each sentence and grouped them in to generalized categories as described in [Angrosh et al., 2010]. These categories include background terms, subject of inquiry terms, outcome terms, strength terms, shortcoming terms, subjective pronouns, words of stress, alternate approach terms, result terms and contrasting terms. However, after analyzing citations and depending upon our own classification scheme, we defined six more categories that are summarized in Table 4.15. We identified a total of 556 cue words. The distribution of these cue words in each generalized categories is listed in Table 4.16.

In our experiments, we used Hidden Naive Bayes (HNB) algorithm [Zhang et al., 2005] for citations classification. We used the presence and absence (binary) of generalized categories as input features for the HNB classifier. We choose HNB because some input features were observed to be conditionally dependent on

Table 4.14 Percentage distribution of citations sentences among citations context classes.

Neutral	Uses	Contrast	Motivated/ Extends	Weak	Strength	Supports/ Confirm	Similar
68.48%	11.07%	1.33%	1.05%	6.2%	8.59%	1.33%	2.19%

Table 4.15 Additional generalized categories of terms.

Category	Examples	Description
Usage terms	uses, adopt, utilize	terms describing usage of anything.
Confirming terms	confirm, consistent with	terms confirming other work.
Example terms	example, like, such as	terms used to give a list of examples.
Similarity terms	similar, likewise	terms used to show similarity between two works.
Motivation terms	motivated, inspired by	terms used to show motivation.
Extension terms	extends, extension	terms describing extension of previous work.

each other. The results of the classification for inter-citations sentences and sentences used in bibliographic coupling are listed in Table 4.17.

As it can be observed from Table 4.17, by following a simple approach, we can achieve considerable results for citations classification. None of the class has F-Measure below 0.65. The F-Measure in case of classes "uses", "similar" and "neutral" is above 0.80. The citations classes can further be grouped in a more abstract scheme of sentiments as mentioned in [Teufel et al., 2006a]. According to this scheme, the classes, i.e., similar, uses, motivated/extends, supports/confirm and strength can be grouped as positive class, while contrast and weak classes can be grouped as negative class. The classification results for the sentiments based generalization scheme is summarized in Table 4.18. Although, by grouping citations classes in sentiments the F-measure for the negative is only 0.66, it is quite significant for positive classes, i.e. 0.85. The precision, recall and F-measure remained same for neutral class. As in conflict of interest situations both

Table 4.16 Frequency of terms in each generalized terms categories.

Category	Number of cue words
Background terms	47
Alternative approach terms	5
Confirming terms	5
Contrasting terms	20
Example terms	25
Extension terms	6
Motivation terms	3
Outcome terms	33
Result terms	11
Shortcoming terms	26
Similarity terms	15
Subject of inquiry terms	232
Subjective pronouns terms	12
Strength terms	35
Usage terms	54
Words of stress terms	27

Table 4.17 Classification results of citations context for inter-citations.

Precision	Recall	F-Measure	Class
0.81	0.85	0.83	uses
0.75	0.64	0.69	contrast
0.83	0.87	0.85	similar
0.87	0.63	0.73	motivated/extends
0.87	0.63	0.73	supports/confirm
0.68	0.66	0.67	weak
0.77	0.73	0.75	strength
0.92	0.93	0.93	neutral

positive (e.g., similar or confirming work) and negative (e.g., competitive or criticizing work) sentiments are important. We can further combine these sentiments in another abstract scheme. More specifically, we can combine positive and negative sentiments as polarity class and can separate their sentences from neutral class. The experimental

Table 4.18 Classification results of generalized citations sentiments for inter-citations.

Precision	Recall	F-Measure	Class
0.85	0.86	0.85	positive
0.72	0.61	0.66	negative
0.92	0.93	0.93	neutral

Table 4.19 Classification results of abstract level citations polarity for inter-citations.

Precision	Recall	F-Measure	Class
0.86	0.84	0.85	polarity
0.93	0.94	0.93	neutral

Table 4.20 Percentage distribution of co-citations sentences among co-citations context classes.

Neutral	Similar	Uses	Contrast	Motivated/Extends
24.6%	63.2%	8.22%	3.46%	0.43%

results of this classification are presented in Table 4.19. It can be observed from Table 4.19 that the classification accuracy in this case is quite significant for both classes: it is 0.85 for polarity class and 0.93 for neutral class.

In case of co-citations, the distribution of co-citation sentences among identified co-citations classes is summarized in Table 4.20. We found only one example of motivated/extends category, which we ignored for our classification experiments. However, it can be used for generalized scheme of sentiments.

For our co-citations classification experiment, we first transformed co-citations sentences in simplified versions. We replaced each citation by a reserve word, e.g., "RESERVE_WORD". We found that citations occurring consecutively and separated by either ",", "and", "or", "or by", "and by", ", noun" or combinations of these can be considered as similar work. We considered these patterns and citations as a single unit and replaced them with a single reserve word. For example, the sentence *"Krishnamurthy and Arlitt [16] and Krishnamurthy and Wills [19] examine accesses*

to many Web sites." [Bent et al., 2004] can be transformed in a simple sentence as *"RESERVE_WORD examine accesses to many Web sites."*. We simplified sentences because it made the features extraction process easier (which will be explained later), and furthermore, we found that most of the simplified sentences with a single reserve word belong to the "similar" category (47.94% of total similar category) and few to the neutral category (12.2% of total neutral category). We used this property as a binary feature for our classifier training and testing. We also used the same generalized cue words categories as mentioned earlier. However, for the co-citation classification experiment, we marked usage and contrasting terms as present if they exist in between of any two reserve words. This approach was adopted after reviewing the usage of these terms in the co-citations annotated as "uses" and "contrast". We further defined a binary feature on the basis of two coordinating conjunctions, i.e., "and", "or" present between two reserve words, and found it helpful in the co-citations classification experiments. We also identified 25 cue words and some patterns that can be helpful in separating neutral co-citations from other categories. Some examples of these cue words includes: "broad efforts", "variety of tasks", "several", "other domains", etc. The examples of some patterns include: "for RESERVE_WORD *any sequence of words* for RESERVE_WORD", "the RESERVE_WORD *any sequence of words* the RESERVE_WORD", "RESERVE_WORD on RESERVE_WORD", "within RESERVE_WORD", "via RESERVE_WORD", etc. We used these cue words and patterns as a single binary feature for co-citations classification experiment. The results of the classification experiment are outlined in Table 4.21. However, it must be noted that in a co-citation sentence, there can be more than two citations. In our experiments, we classified the relationship between only those co-citations in a sentence that have the features or patterns as mentioned earlier.

It can be observed from Table 4.21 that the F-Measure in case of "similar" and "contrast" classes is more than 0.80. The F-measure for "uses" class is 0.69 with the precision 0.75 and recall 0.63. In case of "neutral" class, although F-Measure is 0.63, but the precision is 0.77. This implies that we can identify only some proportion of "neutral" class, but with considerable precision. Similar to inter-citations, the co-citations classes can also be grouped in abstract classes of sentiments. The classification results for sentiments classes are summarized in Table 4.22. It can be observed from Table 4.22 that the precision of neutral class in this case has reached 0.88. The F-measure for negative class in this case is 0.71 with 0.67 precision and 0.75 recall. The F-measure for positive class has reached 0.91 F-measure with 0.85 precision and 0.97 recall. Similarly, the classification results of the polarity and neutral class for co-citations are listed in Table 4.23. It can be observed from Table 4.23 that by combining

the positive and negative sentiments classes under polarity class, the F-measure for neutral class has increased to 0.67 with 0.86 precision. The F-measure for polarity class in this case is 0.92 with 0.87 precision and 0.97 recall.

Table 4.21 Classification results of co-citations contexts.

Precision	Recall	F-Measure	Class
0.83	0.94	0.88	similar
0.78	0.88	0.82	contrast
0.75	0.63	0.69	uses
0.77	0.53	0.63	neutral

Table 4.22 Classification results of generalized co-citations sentiments.

Precision	Recall	F-Measure	Class
0.85	0.97	0.91	positive
0.67	0.75	0.71	negative
0.88	0.51	0.64	neutral

Table 4.23 Classification results of abstract level co-citations polarity.

Precision	Recall	F-Measure	Class
0.87	0.97	0.92	polarity
0.86	0.54	0.67	neutral

In above experiments, we talked about the annotation and automated classification of contexts and sentiments between two authors on the basis of inter-citations and co-citations. In case of bibliographic coupling, one can use the context classification similar to inter-citations, and can use this information to know the relationship between two authors. However, to determine sentiments for bibliographic coupling relationships, we can use the concept of "birds of a feather flocks together". This concepts has been widely investigated in the field of psychology. The researchers found similarity of personality, physical appearance, race, values, demographies and even cognitive similarity as a major driving force for decision making [Murnieks et al., 2007]. As the citations can be classified as positive, negative, or neutral, any two authors with similar

sentiments towards a third author can be grouped together and can be assigned positive sentiments for each other. The only exception to this scheme is for "uses" and "similar" classes. If for example, an author A has "uses" relationship with a third author C, and another author B has "similar" relationship with the same author C, the relationship or sentiment in this case is not clear between author A and author B. In this case they can be assigned neutral sentiments for each other. Similarly, any two authors with opposite sentiments for a particular author can be assigned negative sentiments for each other. However, if both or either one author has neutral sentiments then neutral sentiments can be assigned between them. These rules are summarized in Table 4.24.

Table 4.24 Sentiments assignment scheme for bibliographic coupling.

Author's sentiment	Reviewer's sentiment	Bibliographic sentiment
positive	positive	positive
positive	negative	negative
negative	positive	negative
negative	negative	positive
neutral	negative/postive/neutral	neutral
negative/postive/neutral	neutral	neutral

4.6.5 Results after Assigning Contexts to Citations Relationships

After the detailed discussion about identification of contexts associated with citations relationships and the possibility of their automated classifications, we present the results after assigning these contexts and sentiments to our WWW2006 authors and reviewers. The Table 4.25 list some sample results about the presence and absence of polarity between the authors and reviewers for their citations relationships. We ignored normalized citations counts below 0.2 and considered them insignificant for further discussion. However, the journals' editors and conferences' managers can vary these thresholds depending upon the availability of reviewers. As we mentioned earlier, during the citations extraction process, in some cases we found more citations sentences for the same reference in a paper which were counted as one in CiteSeer. In this scenario, we assigned each additional citation sentences a proper weight on the basis of the total citations listed in CiteSeer for that reference in a paper. For example, if we found two citations sentences for a reference, we would assign a weight of 0.5 to each citation sentence. The sum of these weights is similar to the count for this

reference listed in CiteSeer. Such normalization was necessary otherwise the final normalized citations counts or cognitive similarity presented in Table 4.11 and reproduced in Table 4.25 can be distorted. The Table 4.25 also lists the proportion of normalized citations counts that we were able to extract from the pdf files in comparison to the actual ones listed in CiteSeer. The extraction process, however, can be further enhanced to discover complete information about these citations relationships. It can be observed from Table 4.25 that the presence of polarity among most of the citations relationships is not at a very critical level. The only interested case for further discussion is about "Alec Wolman", where the reviewer is citing to authors with the possibility of some sentiments with reasonable normalized citations counts. We can further elaborate the context associated with these polar relationships. In case of "Alec Wolman" and "Amin Vahdat" the reviewer is positively associated with author with 0.16 normalized citations count. These positive sentiments are due to 0.09 normalized citations counts for using the work of reviewer and 0.06 for the similarity of work. In case of "Craig E. Wills", the reviewer "Alec Wolman" is negatively associated to author with 0.12 normalized citations counts. These negative sentiments are due to the identification of weaknesses in the work of author by reviewer. In the case of "Alec Wolman" and "Balachander Krishnamurthy", the reviewer is associated to author with 0.1 normalized counts for positive sentiments and 0.05 for negative sentiments. These positive and negative sentiments are due to the description of the strength and weakness of the cited work by the reviewer respectively.

Table 4.25 Polarity relationships between authors and reviewers of WWW2006 performance track.

Authors	Reviewers	Sentiments	Martin F. Arlitt	Jeffrey S. Chase			Michael Rabinovich				Oliver Spatscheck	Alec Wolman		
			ci	cs	co	ci	cd	cs	co	ci	co	cs	co	ci
Balachander Krishnamurthy		Polarity	0				0.05	0.01	0.07		0.03	0.01	0.14	
		Neutral	0.2				0.01	0.08	0.29		0.08	0.1	0.23	
		Found/ Total	0.2/0.2				0.06/0.45	0.09/0.26	0.36/0.45		0.11/0.2	0.11/0.24	0.37/0.37	
Craig E. Wills		Polarity					0.02	0.01		0.04		0.02	0.12	
		Neutral					0.01	0.09		0.15		0.12	0.25	
		Found/ Total					0.03/0.28	0.10/0.41		0.19/0.38		0.14/0.37	0.37/0.37	
Amin Vahdat		Polarity	0.1	0.01	0.03	0.02		0.01				0.04	0.16	0.03
		Neutral	0	0.12	0.04	0.15		0.09				0.2	0.09	0.1
		Found/ Total	0.1/0.3	0.13/0.61	0.07/0.43	0.17/0.30		0.10/0.33				0.24/0.56	0.25/0.31	0.13/0.25

dark gray cell: Direct co-authors, light gray cell: Indirect co-authors

4.7 Conclusions

In this chapter, we discussed the problem of conflict of interest (COI) situations in peer review system for scholarly communications. In this context, we described different kinds of COIs that can exist between an author and a reviewer. We categorized these COIs in two broad categories, i.e., Social COIs and Cognitive COIs. We further identified current approaches that are primarily based on social network analysis of authors that are implicitly available in the form of co-authors networks in digital bibliographic databases. We also mentioned the limitations of extracting social networks from social networking websites, authors' homepages and querying the web.

With a brief review of citations theory, we highlighted that different citations relationships can be an indicator of both social and cognitive relationships between researchers. This in turn can be helpful in improving existing COI detection approaches as an additional or alternative means to identify possible social and cognitive bias in peer review system. We investigated this more closely, and performed some experiments to predict the existence of social relationships from citations relationships. We found that a certain proportion of social relationships can be predicted using citations relationships with considerable accuracy. Similarly, we performed an experiment on the authors and reviewers of the WWW2006 conference performance track, and described the potential of citations relationships as an indicator of cognitive distance between these authors and reviewers. We described different contexts and sentiments that can be assigned to these cognitive relationships. We conducted some experiments to highlight the possibility of automated prediction of these context and sentiments. These contexts and sentiments in turn can help in spotlighting the possible severity of cognitive COIs between authors and reviewers. Although we did not find a very severe case of cognitive COI for our selected authors and reviewers, we believe that such analysis might be helpful in other cases.

As outcome of our research if someone is interested in the COI detection of researchers in a certain restricted area, we suggest to proceed the following steps. Obviously, these steps can only be taken when an author and a reviewer are not co-authors in a current article under review.

1. First try to get the previous collaborative information of authors which is available in the form of co-authors networks from the available bibliographic databases or by searching previous papers using APIs of Google or Yahoo search engines. This is the most straight forward way to highlight the social COI situations. A prototype implementation of such a system has been already demonstrated in

[Aleman-Meza et al., 2008]. The domain names of email addresses as mentioned in [Papagelis et al., 2005] can be helpful. Similarly, the acknowledgment sections of papers from authors and reviewers can also be parsed using available natural language processing tools to extract common named entities, and cases where either authors, reviewer, or both have acknowledged to each other in previous contributions.

2. If there is no social relationship has been detected in the first step, try to train and test the model as described in our experiments to highlight social relationships from the citations relationships. However, it requires a large number of training data to get a generic model which might be applicable to most of the cases.

3. If the above steps do not highlight the possibility of social COI or the extracted results are too weak to conclude anything, try to extract cognitive COIs between authors and reviewers through citations relationships as described in our experiments to support the final decision. It must be noted that the research to fully automate the identification of contexts and sentiments associated with citations is still in progress. However, currently we can use semi-automated approaches to achieve this goal as described in our experiments. Obviously, it requires a large number of training samples to achieve an approximately generic model for contexts and sentiments predictions. As a solution to this problem, the author in [Radoulov, 2008] suggested researchers to share their training data with each other or in a central repository to create a big corpus to train and test prediction models. Finally, we suggest that the administration of scholarly communications can either use only the existence of polarity between researchers to consider it as a case of potential cognitive COI or it can go to further levels of sentiments and contexts depending upon their policies and the availability of reviewers.

5

Estimating Articles Downloads for a Digital Journal

In scholarly communications, citations are usually used to evaluate the impact of articles and journals. However, according to the literature, it represents a partial view of articles usage, as readers do not necessarily cite all papers they read. Similarly, articles can take some time to get citations. This makes it impossible to evaluate the impact of articles shortly after they are published. However, with the availability of online electronic journals, a new criteria for evaluating the impact of articles is emerging. This criteria is the download counts of articles. Articles downloads have the potential to implicitly provide a timely measure about articles usage. However, this is one facet of articles downloads. Various studies have also shown that their is a significant positive correlation between articles downloads and their future citations. This suggests that download counts have the potential to anticipate in advance about the future citations for an article. By keeping in view the importance of articles downloads this chapter presents various local and global attributes that are associated with a manuscript to determine its current value. More specifically, we explore the possibility to predict the download counts of a manuscript in the digital library of an electronic journal and implicitly its expected future citations count. In this context, we used Journal of Universal Computer Science (J.UCS) for our detailed prediction experiments. We defined various novel features extracted from J.UCS articles and external bibliographic databases and evaluated their performance in predicating the future downloads of articles published in J.UCS. Moreover, we used only prior features which are available at the time of publishing an article. By using only prior information about articles, we can timely evaluate their future performance. Experiments showed that our selected features helped us in reducing the mean absolute error up to 13% relative to the defined baseline error.

5.1 Importance of Downloads Based Usage Impact

Citations of manuscripts in peer reviewed journals and conferences is a well established unit of analysis in the field of scientometrics and bibliometrics to estimate the intellectual and scientific impact of people, journals, institutions, and nations [Garfield, 1970, Garfield, 1972, Oppenheim, 1997, Bormann and Daniel, 2008]. It explicitly provides a mean to evaluate the performance and quality of articles published in journals or conferences. As discussed in the previous chapter, although some researchers believe the applicability of citations count as an explicit measure of intellectual impact, there are several studies that doubt its use. This is mainly due to the dependence of citations on various factors which includes: motive of citations (e.g., perfunctory and negative citations), field, journal, article type, language, availability, and most importantly, time delays [Bormann and Daniel, 2008]. These time delays between an article acceptance, publication, reading, and citing by other authors and get published their work can take a lot of time (even years depending on field), which eventually leads to delay in evaluating the impact of the article [Brody et al., 2006]. There is a need to identify a different measure that can evaluate the importance of an article soon after its publication.

Before the usage of citations to evaluate the performance of a journal, librarians used to rely on the usage data and surveys from users to acquire books and journals [Gorraiz and Gumpenberger, 2010]. It is a difficult task to count physical usage and conduct annoying surveys [Gorraiz and Gumpenberger, 2010]. However, with the availability of articles accessible through online digital libraries of electronic journals, a new criteria for evaluating their impact is emerging. This criteria is the download counts of articles, which can be easily logged on the server hosting the contents. It provides an implicit measure about the number of times an article has been used or read by users. This usage based measure can provide valuable information to different stake holders involved in scholarly publications system. The librarians for instant can rely on this information for the acquisition of new journals and books. The publishers can evaluate the effectiveness of their offers or deals for subscriptions to their collection of periodicals (e.g., [Nicholas et al., 2003]). The editors can monitor whether their journals are aligned to the editorial policies and attracting wide range of readers. Similarly, the prospective authors can decide which journal is most suitable to attract more readers in their field and in increasing their chance to get cited. Though citations also provide an implicit measure about the usage of an article, it represents only partial statistics [Brody et al., 2006]. The download counts can possibly provide a more complete measure of an article's usage.

Although downloading an article does not necessarily mean that it will be read, it is generally accepted as a most widely used measure of articles usage [Nicholas et al., 2003]. Alternatively, there are many other usage metrics that range from simple hit counts to site penetration metric [Nicholas et al., 2003]. COUNTER (Counting Online Usage of Networked Electronic Resources) is an international consortium that is working in the direction to standardize the recording and exchange of reliable, consistent, and compatible usage data about online resources which includes journals, literary databases, books, and reference works [COUNTER, 2011]. The authors in [Shepherd, 2007, Bollen et al., 2008] emphasized the potential of this usage based metric to provide additional insights and timely evaluations about the performance of scholarly journals. The authors in [Gorraiz and Gumpenberger, 2010] proposed the usage based impact factor for scholarly journals which they believed to be comparable to the journal citation reports produced by Thomson Reuters. However, it requires cooperation from publishers to share various usage based statistics in a standardized way. The standards developed at COUNTER are an ideal opportunity to pursue research in this direction [Shepherd, 2007].

5.2 Relationship between Articles Downloads and Citations

There are various studies that found a significant correlation between the download counts of an article and its later citations count. One of the earliest studies that found a positive correlation between citations and hit counts (full text articles and HTML version) was conducted by Perneger [Perneger, 2004]. The author found a correlation of 0.50 for the 153 papers published in BMJ journal. Similarly, the author in [Moed, 2005] found a correlation of 0.11 between citation counts and early downloads (three months downloads) and 0.35 for later downloads. The author noticed that as the published articles grows older their downloads distribution become statistically more similar to their citation counts distribution. The author further observed that citation counts and download counts of an article effect each other, and are associated to different phases of relevant information gathering and processing for the production of later articles by the scientific community. Later, the authors in [Brody et al., 2006] also conducted a detailed study to find the relationship between citation counts and download counts for the articles published in arXiv pre-prints archive hosted at UK mirror. They found positive correlations of 0.462, 0.347, 0.477, and 0.330 between citations and number of

downloads for the articles in the field of physics, mathematics, astrophysics, and condensed matter, respectively. Similarly, the author in [Watson, 2009] found a correlation of 0.74 between overall citations and downloads of articles published in the journal of vision. The author further highlighted that the correlation for any single year was as high as 0.80 and mostly above 0.60. In an another article [Chu and Krichel, 2007], the authors found that the article downloads in RePEc digital library for economics is correlated with Google Scholar citations and SSCI (Social Sciences Citation Index) with values 0.61 and 0.54, respectively.

However, this is just one facet of download counts that they have the capability to predict the future citations [Brody et al., 2006]. The other part of downloads is that they provide an estimate about the usage of an article, which is not necessarily reflected in citations [Brody et al., 2006]. This is why Tim Brody [Brody et al., 2006] termed citation counts as "citation impact", while download counts as "usage impact".

5.3 Limitations of Downloads Based Usage Impact

In an article by Jamali et al. [Jamali et al., 2005], the authors highlighted the advantages and limitations of log files based analysis, which are also applicable to the download counts based studies. Some important points from [Jamali et al., 2005] are also reproduced here in the context of articles downloads. One of the biggest problems is the difficulty in identifying users. The identification of users is necessary as it can spotlight how many times an article has been used by each individual, and how many people have actually accessed the article. The factors such as: proxy servers, dynamic IP addresses, anonymous browsing, and firewalls make the user identification difficult. Some browsers provide cache facility to store the contents locally on the user's machine to increase the performance for browsing. Similarly, pages can be cached at proxy servers and large regional caches. These factors can greatly effect the measurement of actual download counts of an article. However, despite the shortcomings with download counts based usage matrices, a majority of authors, librarians, and publishers acknowledge its importance in evaluating the value of journals [Shepherd, 2007] and articles [Rowlands and Nichols, 2005, Banks and Dellavalle, 2008].

By keeping in view the importance of downloads, this chapter explores the possibility to predict the download counts of articles in the digital libraries of electronic journals. It will help to evaluate the current importance of an article, its expected readability and implicitly the future citations. In cases where the volume of submissions is increasing rapidly, such analysis can also help in facilitating an initial review or pre-selection of

manuscripts by the administration before delegating it for more rigorous review by the experts in the field. Moreover, it can also help in identifying the factors that must be included in an article to increase its visibility or reading.

5.4 Predicting Articles Download Counts

In the literature, various studies tries to predict the future citation counts based on features associated with an article [Boyack and Klavans, 2005, Fu and Aliferis, 2008, Castillo et al., 2007] and its early citation counts [Castillo et al., 2007]. A comprehensive overview about the efforts in predicting future citations can be found in [Davis, 2010]. However, it must be noted that in contrast, research on predicting the future download counts of articles has been much neglected. In 2003, KDD Cup was organized in conjunction with 9th ACM SIGKDD conference. It arranged four separate tasks for competition including citations and downloads prediction tasks. The organizing committee released the relevant bibliographic and downloads data-sets about the articles published in "High Energy Particle Physics" discipline from the ArXiv repository for the competition. The participants had to predict the download counts of each 150 most downloaded articles in the first two months of their publications in April 2000, March 2001, and February 2002. These predictions were needed to be made on the basis of two months downloaded data from the papers published in the months of February and March of 2000, February and April of 2001 and March and April of 2002. Although the winners of the cup provided significant results for downloads prediction, in their extended technical report they argued to explore various other features directly associated with an article or available through external bibliographic databases for future studies [Brank and Leskovec, 2003]. The study presented in this chapter continues research in this direction and explores various features for downloads predictions. However, in contrast to previous work, we consider predicting the download counts of articles published in a digital journal instead of a pre-print archive. We used articles from Journal of Universal Computer Science (J.UCS) for our experiments. Moreover, we tried to predict the download counts of all articles as opposed to predicting the downloads of only the top most downloaded articles. The current study explores various attributes that are directly associated with an article to predict its future download counts. It also explores two publicly available bibliographic databases to provide different prior information which is directly or indirectly associated with an article, and is not usually documented in a digital journal. We do not use any posterior

information linked to the downloaded articles. Such analysis can provide the importance of an article even at the time of its submission. In KDD Cup, the bibliographic information was restricted to only ArXiv repository, and competitors were allowed to use posterior information about the downloaded articles (e.g., citations to downloaded articles). Finally, in addition to publicly available bibliographic databases, we also explore tagging information before the publication date of downloaded articles from CiteULike repository, in assisting to predict the download counts. CiteULike is a free service that allows users to search, store, organize, and share scholarly publications [CiteULike, 2011]. A previous study has shown a significant correlation between the citations and the number of times a paper is tagged [Saeed et al., 2010]. In this study, we explore whether this social tagging information of papers can provide valuable information for the prediction of future downloads of an article.

5.5 Design of the Study

5.5.1 Selection of data-sets

In this work, we used Journal of Universal Computer Science (J.UCS) articles as a primary input for our experiments. J.UCS is a high quality peer reviewed digital journal that covers all aspects of computer science discipline. We selected J.UCS because of our accessibility to its log files for the extraction of downloads data about articles. Moreover, it maintains structured meta-data about each article. This meta-data includes, paper title, abstracts, authors names, authors affiliations, locations, ACM categories for each article, and dates of submission, acceptance and publication. A detailed description about the extraction of meta-data from J.UCS repository has already been described in Chapter 2 of this dissertation. The articles in J.UCS are published in PDF, PostScript and sometimes in HTML format. For our experiments, we selected 547 articles published in J.UCS from January 2006 to January 2010. We extracted the download counts for these articles from J.UCS log files by counting the requests for their PDF, PostScript and HTML files during the first six months after their publication date. Details about the number of articles in each year are described in Table 5.1. While measuring the download counts, we ignored failed requests, and requests from various crawlers and bots. We also ignored the hits to the abstracts of these articles because we assumed it as equivalent to non-reading. There were five articles whose first six months download counts were more than 1000. These articles might be very exceptional [Brank and Leskovec, 2003]. We considered them as outliers and did not consider them for our further experiments as recommended

in [Brank and Leskovec, 2003]. The download distribution of the remaining articles is described in Figure 5.1.

We further used ParsCit software [ParsCit, 2011] to extract various features of J.UCS articles that are not maintained in J.UCS repository. The extracted features include, cited authors, cited papers titles and date of publication, citations to journals, citations to conferences, and information about the structure of the whole article (headings, sub-headings, figures, equations, and tables). For the extraction of these features, PDF files of each article were manually downloaded from J.UCS website.

For our external bibliographic information, we used the data-sets from CiteSeer and DBLP. The recent update of CiteSeer contains approximately 1,473,409 papers. CiteSeer also contains citations information about articles. We found that approximately 60% papers contain inward citations while 74% papers contain outward citation. CiteSeer does not contain the exact publication dates of articles. It rather stores their publication years. We also noticed that CiteSeer contains duplicate copy of papers

Table 5.1 Number of selected articles from each year.

Year	Articles
2006	95
2007	119
2008	179
2009	143
2010	11

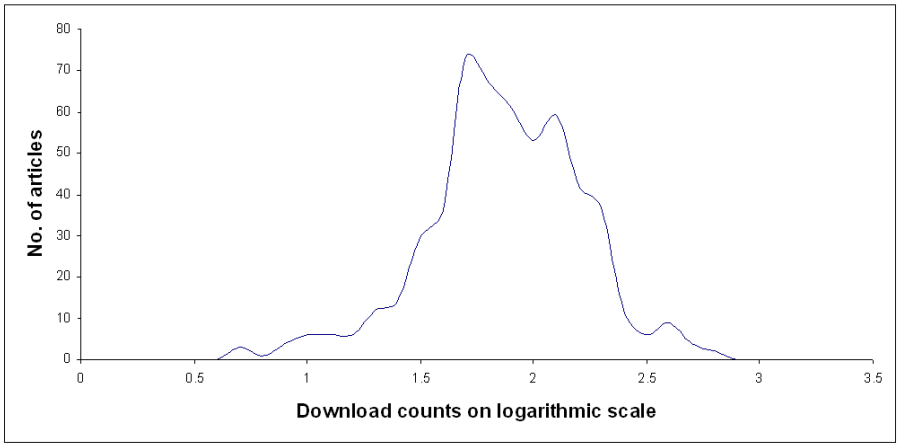


Figure 5.1 The distribution of articles across the logarithmic scale of their first six months download counts.

published in the same year. We removed these duplicates on the basis of authors information that resulted in approximately 1,388,460 papers in CiteSeer. We further found that approximately 649,061 articles contained the full dates of publications, but we assume that these dates represent the entry dates of these articles in the CiteSeer database. We ignored these articles for features extraction. We normalized the papers references by removing the duplication of referenced papers for any citing paper. This resulted in only one reference "to" a paper "by" a paper. We performed this step because it is time consuming to ensure that the duplicated references were due to the data entry mistake or due to the multiple referenced sentences to a paper by the citing paper. Moreover, we used the same DBLP data-set as used in the previous chapter. However, it must be noted that in this dataset the publications are available up till 2009 inclusive (downloaded in June, 2009), and it also stores only the publication year of the articles.

Additionally, we also used the CiteULike tagging data for our experiments. CiteULike is a service that helps to store, organize, and share scholarly publications [CiteULike, 2011]. The meta-data about tags for each article is made available to download for research purposes on request to CiteULike service team. However, it must be noted that this meta-data does not contain the title of articles and names of authors. It rather contains unique identifiers of users and papers, tags, and the date on which the tags were assigned to these papers by the users.

5.5.2 Selection of Features

We explored a number of features that are associated within an article or available in external bibliographic databases. However, as mentioned previously, we extracted only those features which are usually available at the time of publishing an article. We also used some feature identified in [Brank and Leskovec, 2003]. These features are authors names, their affiliations, keywords in abstracts and titles as features vector, their length, number of authors, and year of inclusion. We wrote various routines to compute all features from the above mentioned J.UCS repository, parsed articles from ParsCit, CiteSeer, DBLP, and CiteULike. We mainly divided these features in six groups which are described in the following sub-sections along with the arguments for their inclusion. Please note that these features are quite large in numbers. However, in our experiments described in Section 5.5.3, we try to reduce these features gradually based on their usefulness in predicting download counts.

G1– Features Locally Available in an Article

These features represent the information which is readily available in an article published in J.UCS. Although most of the features mentioned in this group are not apparently visible to readers while previewing an article on the journal's website, we wanted to evaluate their role in the decision of downloading the article. This is due to the fact that scholarly communications is a social process, so it is hypothesized that a well written article is recommended by people to others in their field. For our experiments, we used the following list of features.

– *Number of Authors* (NA) represents the number of authors that have written the downloaded article. It is interesting to know that whether the number of authors in an article has any impact on the download frequency of articles or if it is irrelevant.

– *Number of Sections* (NS) is the number of sections in the downloaded article.
– *Number of Sub-Sections* (NSS) is the number of sub-sections in the downloaded article.

We selected NS and NSS as features because they have the tendency to highlight how well an article is structured. It can play an important role in predicting downloads frequency.

– *Number of Figures* (NF) is the number of figures in the downloaded article.
– *Number of Equations* (NE) is the number of equations in the downloaded article.
– *Number of Tables* (NT) represents the number of tables in the downloaded article.

We selected NF and NT as features because they have the potential to highlight how well an article has been illustrated. It might be that a well illustrated article has more impact on its number of downloads. Similarly, we chose NE as a feature to distinguish between a purely theoretical paper and an experimental paper which is supported mathematically and statistically.

– *Number of Pages* (NP) is the number of pages in the downloaded article. The length of the article is important as it can highlight whether users tend to read short articles or long articles.

– *Number of Words in Title* (NWT) represents the number of words used in the title of the downloaded article.
– *Number of Words in Abstract* (NWA) represents the number of words used in the abstract of the downloaded article.

The features NWT and NWA are relevant as the readers usually see this abstract information before deciding to download an article.

– *Number of Keywords by Authors* (NKA) represents the number of keywords phrases

mentioned in the downloaded article. This information is easily available in the preview of the article on the journal's website, and helps in summarizing the concepts discussed in an article. Its analysis for prediction will also help in deciding that whether identification of too many keywords increases the chances of downloads or it has minimal effect on readers.

– *Number of ACM Categories by Authors* (NACA) represents the number of ACM categories mentioned in the downloaded article.
– *Number of Second Level ACM Categories by Authors* (NSLACA) represents the number of second level ACM categories mentioned in the downloaded article.
– *Number of Third Level ACM Categories by Authors* (NTLACA) represents the number of third level ACM categories mentioned in the downloaded article.

The ACM categories are also available to the readers in the preview of the article, and depict the breadth of the topics discussed in the article. It will be interesting to know if readers download articles covering a wide range of topics or are interested in articles that are more focused on a single topic. Moreover, such analysis will also help in deciding that whether the description of categories up to a finer granularity is important or if it has minimal effects on readers. We did not consider first level categories because there was only one paper containing a first level category. It will be of no significance in cross-validation training and testing reported in Section 5.5.3.

– *Number of Distinct First Level ACM Categories in the Article* (NDFLACA) represents the number of first level ACM categories mentioned in the downloaded article. This is an extension of the above mentioned feature. It also represents the range of different topics covered by an article.

– *Article in Special Issue* (ASI) is a binary feature representing whether the article was published in special issue (represented by 1) or regular issue (represented by 0). This feature highlights user's preference in terms of downloading the extended versions of papers from conferences or the regular papers.

– *Number of Days for Acceptance* (NDA) represents the number of acceptance days from submission date to acceptance date for only regular papers. This information is not available for papers published in a special issue. We assume that a paper which is accepted in a shorter time period has more quality, and hence has a better chance to get downloaded more frequently.

– *Article Published in First Quarter* (APQ) is a group of four binary features that represent quarter of the year in which the article is published. We used these feature to find whether the timings of publishing an article has any impact on its chances to get downloaded.

– *Authors from Academic Institutions* (AAI) represents the number of authors from

academic institutions in the downloaded article.
– *Authors from Non-Academic Institutions* (ANAI) represents the number of authors from non-academic institutions in the downloaded article.

We selected AAI and ANAI to find that whether the affiliations type has any impact on article downloads, or it is irrelevant.

– *Number of Institutions* (NI) represents the number of institutions involved in publishing the downloaded article.
– *Number of Countries* (NC) represents the number of countries involved in publishing the downloaded article.

We chose NI and NC to find the impact of variety of institutions and countries in downloading an article.

– *Article in HTML* (AH) is a binary feature which represents that whether the article was also published in HTML (represented as 1) or only in PDF and PostScript format (represented by 0). It might be that readers prefer to open the full articles in HTML format rather than downloading the articles in PDF or PostScript. This is due to the fact that readers usually have browsers to read HTML articles, while reading the PDF or PostScript file require an additional viewer software.

– *Year of Publication* (YP) is a group of binary features representing the year in which the downloaded article was published. We used these features to find the effect of publication years on article downloads.

– *First Level ACM Categories* (FLAC) is a group of feature representing the first level ACM categories. Each category contains the number of times it has been mentioned by the author of the downloaded article.
– *Second Level ACM Categories* (SLAC) is a group of feature representing the second level ACM categories. Each category contains the number of times it has been mentioned by the author of the downloaded article.
– *Third Level ACM Categories* (TLAC) is a group of feature representing the third level ACM categories. Each category contains the number of times it has been mentioned by the author of the downloaded article.

We used these features to find the role of each ACM level on the downloads frequency of articles.

– *Authors Names* (AN) is not really a single feature. It is a group of features, where each feature represents an author name. We found that 98 authors out of 1575 authors from downloaded articles have two or more papers.
– *Countries Names* (CN) is not really a single feature. It is a group of features, where each feature represents a country name. We found that 54 countries from downloaded

articles have two or more papers.
– *Institutions Names* (IN) is not really a single feature. It is a group of features, where each feature represents an institution name. We found that 294 institutions from downloaded articles have two or more papers.

We included AN, CN, and IN to find out that whether the names of authors, countries, and institutions have any impact on the downloads frequency of articles.

– *Terms Categories Describing Current Work* (TCDCW) is a group of features that represents the type of terms used in the abstract for describing the work done in an article. It will be interesting to find out whether the choice of words, or the type of research conducted in the article has any impact on its download frequency. The detail of these features and the number of terms in each feature is described in Table 5.2. All the features described in Table 5.2 were manually extracted from the 542 articles used in the current study.

Table 5.2 Categories of terms

Category	**Description**	**No. of terms**
Improvements terms	terms describing improvements, e.g., improve, improvements	4
Evaluation terms	terms describing evaluations conducted in the current article, e.g., evaluate, tested, validate	17
Description terms	terms describing current work, e.g., aim, assuming, attempts, chose, provide	125
Comparison terms	terms describing comparisons, e.g., better, better than, contrast	12
Survey terms	terms describing whether the current article is a survey paper or original research contribution, e.g., overview, review, survey	11
Projects terms	terms describing that the current article is an outcome of an ongoing project which includes project and projects	2
Novelty terms	terms describing the novelty of current work, e.g., innovative, novel, new	6
Case studies terms	terms implying that the current work conducts a case study, e.g., case study, user study, interviews	9
Extension terms	terms describing that the current work is an extension of previous works, e.g., extend, extension	4
Solution terms	terms describing the proposed solution in the article, e.g., addressed, prove, propose, solve, present	37

Continued on next page

Table 5.2 – *Continued from previous page*

Category	Description	No. of terms
Analysis terms	terms describing the analysis done in the current work, e.g., investigates, explore, analysis, analyze	13
Usage terms	terms describing the usage of any previous work, e.g., adapt, used, utilized, modified	14
Outcome terms	terms describing the outcome of the current work, e.g., result, concludes, find out	7
Implementation terms	terms describing implementations in the current work, e.g., applied, deploy, implemented	34

G2– Features Locally Available in the Digital Journal

These features include those information that are not directly available in the article, but in other articles published before in the same journal.

– *Number of Papers in JUCS* (NPJ) is the total number of papers published earlier in J.UCS by the authors of the current article. We chose this feature to find out that whether publishing multiple articles in a single journal increases the chances of readability or it has no effect in the downloads of later articles.

– *Authors as Reviewers in JUCS* (ARJ) represents the number of authors, who are also reviewers in J.UCS. As reviewers in J.UCS are experts in the their field. It is expected that the articles written by reviewers have a better chance to get downloaded by readers.

– *Papers in Same First level Categories in JUCS* (PSFCJ) is the number of papers published earlier in the same first level categories.
– *Papers in Same Second level Categories in JUCS* (PSSCJ) is the number of papers published earlier in the same second level categories.
– *Papers in Same Third level Categories in JUCS* (PSTCJ) is the number of papers published earlier in the same third level categories.
– *Article Complete Keywords in JUCS* (AKCJ) is the number of times the keywords phrases of the downloaded article are used in earlier published papers in J.UCS.
– *Article Keywords in Parts in JUCS* (AKPJ) is the number of times the keywords extracted from keywords phrases are used in earlier published papers in J.UCS.
– *Article Title Keywords in JUCS* (ATKJ) is the number of times the title keywords

of the downloaded article are used in earlier published papers in J.UCS.
– *Article Abstract Keywords in JUCS* (AAKJ) is the number of times the abstract keywords of the downloaded article are used in earlier published papers in J.UCS.

We selected PSFCJ, PSSCJ, PSTCJ, AKCJ, AKPJ, ATKJ, and AAKJ to find the impact of availability of similar articles in the journal on the downloads of later articles.

–*All Keywords* (AK) is not really a single feature, but it represents all keywords used in title, abstract, keywords phrases, and keywords phrases in parts as features. Here, keywords in parts means that we break the keywords phrases into individual parts. For example, if there is a keyword "visual analytic", then it was divided in "visual" and "analytic".

G3– Features Related to Co-Authors Networks

These features basically represent the co-authors networks of author/authors of the article in J.UCS. All the features mentioned in this group were extracted from CiteSeer and DBLP using exact name matches. Alternatively, one can also use partial matching of author names using string similarity algorithms. The co-authors networks of J.UCS author/authors were calculated from the papers in DBLP and CiteSeer that had been published earlier than their respective article in J.UCS. We assume that the articles written by author/authors with a large co-authors network gets more downloads.

– *Number of Co-Authors in CiteSeer* (NCAC) represents the number of direct co-authors of the author/authors of the downloaded article extracted from CiteSeer.
– *Number of Co-Authors in DBLP* (NCAD) represents the number of direct co-authors of the author/authors of the downloaded article extracted from DBLP.
– *Number of Second Level Co-Authors in CiteSeer* (NSLCAC) represents the number of second level co-authors of the author/authors of the downloaded article extracted from CiteSeer.
– *Number of Second Level Co-Authors in DBLP* (NSLCAD) represents the number of second level co-authors of the author/authors of the downloaded article extracted from DBLP.

Here, the direct authors are those authors who have ever written a paper with the author/authors of the article. Whereas second level authors are the co-authors

which are connected with the authors/authors of the article through a common collaborator (direct co-authors).

G4– Features Related to Reputation of Authors

These features represent the reputation of author/authors before publishing the article in J.UCS in terms of citations received and the quantity of research output. Similar to features in G3, exact matching of authors names were used to calculate the features in this group. We chose these features to find out that whether these factors has any impact on articles downloads or readers download papers irrespective of the reputation of authors.

– *Number of Earlier Papers* (NEPC) represents the number of papers written earlier by author/authors of the article extracted from CiteSeer.

– *Number of Earlier Papers* (NEPD) represents the number of papers written earlier by author/authors of the article extracted from DBLP.

– *Number of Citing Papers to Authors* (NCPA) represents the number of citing papers to the author/authors of the downloaded paper extracted from CiteSeer.

– *Number of Citing Authors to Authors* (NCAA) represents the number of citing authors to the author/authors of the downloaded paper extracted from CiteSeer.

G5– Features Related to Quality of References used in Current Article

These features represent the quality of references used in the article. The cited names in JUCS articles are usually written by last name preceded by the first letter of first name. We used this information to extract authors names from CiteSeer and DBLP and calculated different features presented in this group. As many people can have similar last names and first letter of the first name, it is expected that a large number of false positive names were also extracted. The method of extraction can be improved, but we ignored this issue at the moment.

– *Number of Papers by Cited Authors in CiteSeer* (NPCAC) represents the number of papers published by the cited authors in CiteSeer.

– *Number of Papers by Cited Authors in DBLP* (NPCAD) represents the number of papers published by the cited authors in DBLP.

– *Number of Citations to Cited Authors* (NCCA) represents the number of citations received by the cited authors in the downloaded article.

– *Number of Citations to References* (NCR) represents the number of citations received by the references mentioned in the current article.

We chose NPCA and NCCA to find out whether the reputation of cited authors in terms of number of publications and citations has any impact on the article downloads. It is assumed that the choice of referenced papers written by popular authors attracts more readers. Similarly, NCR represents the popularity of the referenced material. This feature basically represents the quality of current article. It might be that a quality written article has a high number of popular references leading to more downloads.

– *Number of Citations* (NC) represents the number of citations in the downloaded article.
– *Number of Citations to Journals* (NCJ) represents the number of citations to journals in the downloaded article.
– *Number of Citations to Conferences* (NCC) represents the number of citations to conferences in the downloaded article.

We used NC, NCJ, and NCC features as they have the tendency to highlight the breadth of the literature covered in the article. Moreover, the references to journals are generally considered as a choice of good articles to cite. We assume that an article containing many references to journals may prompt the users to download an article, and to find many quality articles for further readings in their field. Similarly, there are many conferences that provide valuable research in their research area.

– *Average Age of Citations* (AAC) is the average of citations used in the downloaded article. This feature is selected to check if the users download the article which cites recent publications or old material.
–*Average Number of Papers by Cited Authors in CiteSeer* (ANPCAC) is the average number of papers written by cited authors and indexed in CiteSeer.
–*Average Number of Papers by Cited Authors in CiteSeer* (ANPCAD) is the average number of papers written by cited authors and indexed in DBLP.

G6– Features Related to the Popularity of Similar Literature available in External Bibliographic Databases

This group of features represents earlier material, available community, and popularity of the research area with respect to the downloaded article. We believe such global information plays a vital role in encouraging readers to download an article.

– *Number of Authors in DBLP with same Keywords* (NADK) represents the number of authors in DBLP using any of the keywords phrases used in the downloaded article.

– *Number of Authors in DBLP with same Keywords in Parts* (NADKP) represents the number of authors in DBLP using any of the keywords extracted from keywords phrases used in the downloaded article.

– *Number of Authors in CiteSeer with same Keywords* (NACK) represents the number of authors in CiteSeer using any of the keywords phrases used in the downloaded article.

– *Number of Authors in CiteSeer with same Keywords in Parts* (NACKP) represents the number of authors in CiteSeer using any of the keywords extracted from keywords phrases used in the downloaded article.

– *Number of Papers in DBLP with same Keywords* (NPDK) represents the number of papers in DBLP using any of the keywords phrases used in the downloaded article.

– *Number of Papers in DBLP with same Keywords in Parts* (NPDKP) represents the number of papers in DBLP using any of the keywords extracted from keywords phrases used in the downloaded article.

– *Number of Papers in CiteSeer with same Keywords* (NPCK) represents the number of papers in CiteSeer using any of the keywords phrases used in the downloaded article.

– *Number of Papers in CiteSeer with same Keywords* (NPCKP) represents the number of papers in CiteSeer using any of the keywords extracted from keywords phrases used in the downloaded article.

– *Number of Authors in DBLP with same Keywords in Title* (NADKT) represents the number of authors in DBLP using any of the keywords used in the title of the downloaded article.

– *Number of Authors in CiteSeer with same Keywords in Title* (NACKT) represents the number of authors in CiteSeer using any of the keywords used in the title of the downloaded article.

– *Number of Papers in DBLP with same Keywords in Title* (NPDKT) represents the number of papers in DBLP using any of the keywords used in the title of downloaded article.

– *Number of Papers in CiteSeer with same Keywords in Title* (NPCKT) represents the number of papers in CiteSeer using any of the keywords used in the title of the downloaded article.

– *Number of Authors in DBLP with same Keywords in Abstract* (NADKA) represents the number of authors in DBLP using any of the keywords used in the abstract of the downloaded article.

– *Number of Authors in CiteSeer with same Keywords in Abstract* (NACKA) represents the number of authors in CiteSeer using any of the keywords used in the abstract of the downloaded article.

– *Number of Papers in DBLP with same Keywords in Abstract* (NPDKA) represents the number of papers in DBLP using any of the keywords used in the abstract of downloaded article.

– *Number of Papers in CiteSeer with same Keywords in Abstract* (NPCKA) represents the number of papers in CiteSeer using any of the keywords used in the abstract of the downloaded article.

– *Number of Papers Tagged with Keywords* (NPTK) represents the number of distinct papers tagged in CiteULike using the keywords phrases used in the downloaded article.

– *Number of Papers Tagged with Keywords in Parts* (NPTKP) represents the number of distinct papers tagged in CiteULike using the keywords extracted from keywords phrases used in the downloaded article.

– *Number of Users Tagging with Keywords* (NUTK) represents the number of distinct users tagging in CiteULike using the keywords phrases used in the downloaded article.

– *Number of Users Tagging with Keywords* (NUTKP) represents the number of distinct users tagging in CiteULike using the keywords extracted from the keywords phrases used in the downloaded article.

– *Number of Times Tagged with Keywords* (NTTK) represents the number of times tagging was done in CiteULike using the keywords used in the downloaded article.

– *Number of Times Tagged with Keywords* (NTTKP) represents the number of times tagging was done in CiteULike using the keywords extracted from keywords phrases used in the downloaded article.

– *Number of Papers Tagged with Title Keywords* (NPTTK) represents the number of distinct papers tagged in CiteULike using the keywords used in the title of the downloaded article.

– *Number of Users Tagging with Title Keywords* (NUTTK) represents the number of distinct users tagging in CiteULike using the keywords used in the title of the downloaded article.

– *Number of Times Tagged with Title Keywords* (NTTTK) represents the number

of times tagging was done in CiteULike using the keywords used in the title of the downloaded article.

– *Number of Papers Tagged with Abstract Keywords* (NPTAK) represents the number of distinct papers tagged in CiteULike using the keywords used in the abstract of the downloaded article.

– *Number of Users Tagging with Abstract Keywords* (NUTAK) represents the number of distinct users tagging in CiteULike using the keywords used in the abstract of the downloaded article.

– *Number of Times Tagged with Abstract Keywords* (NTTAK) represents the number of times tagging was done in CiteULike using the keywords used in the abstract of the downloaded article.

– *All Keywords in DBLP* (AKD) is not a single feature but represents all keywords, i.e., keywords phrases, keywords extracted from keywords phrases, title keywords, and abstract keywords as features.

– *All Keywords in CiteSeer* (AKC) is not a single feature but represents all keywords i.e., keywords phrases, keywords extracted from keywords phrases, title keywords, and abstract keywords as features.

– *All Keywords Tagged* (AKT) is not a single feature but represents all keywords i.e., keywords phrases, keywords extracted from keywords phrases, title keywords, and abstract keywords as features.

Each of these features represents the frequency of papers in DBLP, CiteSeer, and CiteULike containing the corresponding keyword.

5.5.3 Experimental Results

For the various features described in the previous section, SVM linear regression was used to train and test the predictors. We used WEKA [Weka, 2009] and LIBSVM library for SVM implementations to train and test predictors. All the features except binary features were normalized between 0 and 1 using the formula, i.e., X_{new}=(X - X_{min})/(X_{max} - X_{min}). In each experiment, we used 10-fold cross validation in WEKA. The final regression results were evaluated using Mean Absolute Error (MAE). The MAE is the average difference between the actual value and the predicted value. This can be represented by the formula, i.e., |(actual value) - (predicted values)|/(total number of values). The actual values in our experiments are the download counts that need to be predicted by the trained model. We further used a ground truth value or baseline error to evaluate the performance of predictions by comparing its value with the MAE. The ground truth value in our case is obtained by taking the averaged

difference between the actual downloads and the median download counts as described in [Brank and Leskovec, 2003]. The median download counts in our experiments is 66, which leads to the ground truth value to 49.50. We also tried to use average download counts in calculating the baseline error. However, in this case the baseline error is 53.07 which is higher than baseline error obtained through median download counts. Therefore, selecting baseline error based on median download counts is much better in obtaining a more accurate predictor. In evaluating the performance of different predictors in each group, we kept the cost parameter of SVM regression model as 1. However, this parameter was adjusted to 0.7 to obtain the final predictor. In our experiments, we evaluated each group performance individually. We followed the scheme of "leave one out" to evaluate the performance of different combination of features. In this policy if by excluding any feature the MAE gets reduced, then we consider that feature as irrelevant and ignore it in further experiments. The Tables 5.3 to 5.8 summarize the results about the performance of different combinations of features in each group.

It can be observed from the results summarized in Tables from 5.3 to 5.8 that different features included in groups G1, G2, G5, and G6 helped in reducing the MAE. The minimal error rate for these groups is smaller than the baseline error, i.e. 44.87, 47.03, 49.36, and 46.99 for groups G1, G2, G5, and G6, respectively. In the case of groups 3 and 4 some features reduced the MAE, but their minimal error rate is higher than the baseline error, i.e., 49.60 and 49.62 respectively. This implies that the co-authors networks and reputation of authors are not really useful in predicting the downloads of an article. Although we can not generalize it for other journals and publications archives, but it is at least true in the case of J.UCS and the way we are calculating these features. Further validity of these results, however, needs more experiments for other authors and publications archives. One can also think of new ways to apply these features. For example, authors usually work in different research areas during their life. One important direction might be to calculate their co-authors networks and reputation in a particular research area for articles downloads prediction.

Table 5.3 point out some interesting results for group G1 features. It can be observed from Table 5.3 that MAE for all features in group G1 is 44.99. However, step by step exclusion of some features resulted in reducing the MAE up to 44.87. The exclusion of features, i.e., FLAC, NI, NSLACA, NKA, NWT, NF, and NA helped in reducing the MAE. The exclusion of features, i.e., NDA, NDFLACA, NWA, NP, NT, NSS, and NS did not effect the MAE. Therefore, we also considered them as irrelevant and ignored them in calculating MAE in next steps. It can also be observed from Table 5.3 that the major contributing feature in reducing the error in group 1 is YP (years of

Table 5.3 Predictions using Group–1 features.

Features	MAE
All	**44.99**
All-AN	45.13
All-IN	45.14
All-CN	45.12
All-TLAC	45.12
All-SLAC	45.11
G1R1=>All-FLAC	**44.96**
G1R1-TCDCW	45.03
G1R1-YP	46.92
G1R1-AH	45.09
G1R1-NC	44.98
G1R2=>G1R1-NI	**44.94**
G1R2-ANAI	44.97
G1R2-AAI	44.95
G1R2-APQ	45.97
G1R3=>G1R2-NDA	**44.94**
G1R3-ASI	45.02
G1R4=>G1R3-NDFLACA	**44.94**
G1R4-NTLACA	44.96
G1R5=>G1R4-NSLACA	**44.93**
G1R5-NACA	44.96
G1R6=>G1R5-NKA	**44.92**
G1R7=>G1R6-NWA	**44.92**
G1R8=>G1R7-NWT	**44.91**
G1R9=>G1R8-NP	**44.91**
G1R10=>G1R9-NT	**44.91**
G1R10-NE	45.02
G1R11=>G1R10-NF	**44.89**
G1R12=>G1R11-NSS	**44.89**
G1R13=>G1R12-NS	**44.89**
G1R14=>G1R13-NA	**44.87**

publications). The exclusion of YP features leads the MAE up to 46.92. It implies that the downloads of articles are somehow more related to the years in which they are published. Similarly, the yearly quarters represented by APQ are also important in estimating future downloads, whose exclusion increases the MAE to 45.97. The rest of the features in comparison are contributing little in reducing the MAE.

In case of group G2, the results are presented in Table 5.4, the feature AK is most

Table 5.4 Predictions using Group–2 features.

Features	MAE
All	**47.11**
G2R1=>All-AAKJ	**47.10**
G2R1-ATKJ	47.11
G2R2=>G2R1-AKPJ	**47.09**
G2R3=>G2R2-AKCJ	**47.08**
G2R4=>G2R3-PSTCJ	**47.07**
G2R5=>G2R4-PSSCJ	**47.06**
G2R6=>G2R5-PSFCJ	**47.04**
G2R6-ARJ	47.05
G2R7=>G2R6-NPJ	**47.03**
G2R7-AK	49.50

Table 5.5 Predictions using Group–3 features.

Features	MAE
All	49.63
G3R1=>All-NSLCAD	**49.63**
G3R1-NSLCAC	49.65
G3R2=>G3R1-NCAD	**49.61**
G3R3=>G3R2-NCAC	**49.60**

Table 5.6 Predictions using Group–4 features.

Features	MAE
All	49.63
G4R1=>All-NEPC	**49.62**
G4R2=>G4R1-NEPD	**49.62**
G4R3=>G4R2-NCPA	**49.62**

Table 5.7 Predictions using Group–5 features.

Features	MAE
All	**49.39**
All-ANPCAD	49.44
All-ANPCAC	49.40
All-AAC	49.51
All-NCC	49.48
G5R1=>All-NCJ	**49.39**
G5R1-NC	49.45
G5R2=>G5R1-NCR	**49.38**
G5R3=>G5R2-NCCA	**49.37**
G5R3-NPCAD	49.38
G5R4=>G5R3-NPCAC	**49.36**

effective. The exclusion of AK increased the error up to 49.50 which is equivalent to baseline error. The exclusion of features, i.e., AAKJ, AKPJ, AKCJ, PSTCJ, PSSCJ, PSFCJ, and NPJ slightly helped in reducing the error. Finally, only three features, i.e., ATKJ, AK, and ARJ helped in reducing the error up to 47.03.

The results using features in group G5 shown in Table 5.7 helped in reducing the error up to 49.36, which is slightly smaller than baseline error. The exclusion of features, i.e., NCJ, NCR, NCCA, and NPCAC helped in reducing the error. The results for group G5 implies that the quality of references used in articles have minimal effects on articles downloads.

In case of group G6, the results presented in Table 5.8 show that the only effective feature is AKD, which alone reduced the MAE up to 46.99. The feature AKC also reduced the error up to 47.30. However, this error is larger than MAE of AKD. The combination of both AKD and AKC increases the error. It means although they represent the same features but calculated from different sources do not complement each other. As it can be further observed from Table 5.8 the tagging information from CiteULike in any calculated feature is contributing little to reduce the MAE.

Finally, we combined all those combinations of features from group G1, G2, G5, and G6 which helped in reducing the MAE. The prediction results from the combination of these features are listed in Table 5.9. It can be observed from this table that combining these features reduced the error to 43.73. We further tried different values of cost parameter for SVM regression model to reduce the MAE. The cost value 0.7 received the best results which reduced the MAE to 43.04. These results suggest that

Table 5.8 Predictions using Group–6 features.

Features	MAE
All	**48.61**
G6R1=>All-NADK	**48.61**
G6R2=>G6R1-NADKP	**48.61**
G6R3=>G6R2-NACK	**48.60**
G6R4=>G6R3-NACKP	**48.60**
G6R5=>G6R4-NPDK	**48.59**
G6R6=>G6R5-NPDKP	**48.59**
G6R7=>G6R6-NPCK	**48.58**
G6R8=>G6R7-NPCKP	**48.57**
G6R9=>G6R8-NADKT	**48.57**
G6R10=>G6R9-NACKT	**48.57**
G6R11=>G6R10-NPDKT	**48.57**
G6R12=>G6R11-NPCKT	**48.56**
G6R13=>G6R12-NADKA	**48.56**
G6R14=>G6R13-NACKA	**48.55**
G6R15=>G6R14-NPDKA	**48.55**
G6R16=>G6R15-NPCKA	**48.54**
G6R17=>G6R16-NPTK	**48.54**
G6R18=>G6R17-NPTKP	**48.54**
G6R19=>G6R18-NUTK	**48.54**
G6R20=>G6R19-NUTKP	**48.53**
G6R21=>G6R20-NTTK	**48.53**
G6R22=>G6R21-NTTKP	**48.52**
G6R23=>G6R22-NPTTK	**48.52**
G6R24=>G6R23-NUTTK	**48.52**
G6R25=>G6R24-NTTTK	**48.52**
G6R26=>G6R25-NPTAK	**48.51**
G6R27=>G6R26-NUTAK	**48.51**
G6R28=>G6R27-NTTAK	**48.51**
G6R29=>G6R28-AKT	**48.22**
G6R30=>G6R29-AKC	**46.99**
G6R31=>G6R29- AKD	47.30

Table 5.9 Predictions using all effective features.

Features	MAE	Percentage Improvement
Ground Truth	49.50	–
All features	**43.73 (cost=1)**	**11.65%**
All features	**43.04 (cost=0.7)**	**13.05%**

we succeed in reducing the MAE from its ground truth by 13%. As a comparison, the results reported in [Brank and Leskovec, 2003] for ArXiv reduced the error by 23% compared their ground truth value. Although there is an absolute difference of 10 between our results and the results reported in [Brank and Leskovec, 2003], it must be noted that the referenced study tried to predict the top most downloaded papers from ArXiv, while we made predictions for all papers. Moreover, we did not use any posterior information such as citations to downloaded articles for predictions. As noted in [Brank and Leskovec, 2003], this is clearly not an easy task. We need to think of other novel features to further improve articles downloads predictions. One important direction might be to use the downloads patterns of similar articles published earlier in the same journal. In our future work, we will investigate this information to improve prediction results.

5.6 Conclusions

In this chapter, we highlighted the importance of downloads based usage impact of articles. Articles download counts have the potential to implicitly provide a more complete and timely measure about articles usage impact than citations based measures. However, this is only one facet of articles downloads. Various studies have also shown that their is a significant positive correlation between articles download counts and their future citations. This suggests that download counts have the potential to anticipate in advance the future citations for an article. By keeping in view the importance of articles downloads, in this chapter, we used various attributes that are locally and globally associated with an article to predict its future downloads, to reveal its current value and implicitly its expected future citations. In this context, we used Journal of Universal Computer Science (J.UCS) for our detailed prediction experiments. We defined various novel features extracted from J.UCS articles and external bibliographic databases and evaluated their performance in predicting the future downloads of articles published in J.UCS. As compared to a previous study, we tried to predict the

download counts of all articles rather than top downloaded articles. Moreover, we used only features which are available at the time of publishing an article. By using only prior information about articles we can timely evaluate their future performance. Experiments showed that our selected features helped us in reducing the mean absolute error by 13% when compared to the defined baseline error.

6

Summary and Outlook

In this thesis, we provided an overview about the importance of digital libraries, their history and some commonly used definitions regarding digital library terminology. We also highlighted some important events in the developments of electronic journals' digital libraries. In addition to that, we provided a detailed overview about the key issues and challenges associated with the development of a truly usable digital library. A detailed discussion about these issues highlighted that digital libraries research area is quite large and contains various sup-topics. It is certainly not possible to tackle all these issues in a single thesis. Therefore, we decided to work on topics related to the quality assurance of content in digital libraries of scholarly publications. Managing the quality of contents in digital libraries is again very broad and dynamic. In scholarly publishing systems, the quality of manuscripts is usually determined by a rigorous peer review process. Some journals conduct scientometrics, bibliometrics, and content analysis to ensure that the journal is aligned to its policies and is producing valuable research contributions. However, with growing information size, expanding scholarly communities, increase in submissions of articles, and growing burden on reviewers of these journals, conventional techniques to ensure the quality of content are becoming insufficient. There is an imminent need to discover innovative administrative tools to address these problems. By keeping in view the limitations of current systems, this thesis presents various novel solutions and possibilities that can assist in the quality management of content in the digital libraries of scholarly publications. This in turn can help in improving the general standing and reputation of scholarly communications mediums. The main contributions presented in this thesis can be divided in four parts which are summarized below along with their important findings. Similarly, lessons learned and potential future works that can be done to extend the ideas presented in this thesis are also described.

– In the first contribution, we highlighted the current developments of the WWW,

which has turned the web in to a socially driven platform. In this context, we outlined various community driven initiatives which is producing an enormous amount of information on the web. The existing literature emphasize that there is an emergent need to harness this collective intelligence of people to create novel solutions to address various problems including challenges faced in the area of digital libraries. Besides other community driven initiatives, we provided a brief overview about web mash-up (an emerging Web 2.0 paradigm) that allows anyone to combine existing data or information through publicly available APIs like Amazon, Google, Yahoo, eBay, etc. Such data from various sources can be combined in innovative ways to provide novel solutions to various problems. We emphasized that a mash-up for a digital journal can serve as an important platform to address users' diverse requirements and as an administrative support tool to facilitate rapid expansion of the journal. In this contribution, we demonstrated a pioneer example of such a mash-up for J.UCS by combining J.UCS metadata with Google Map APIs to solve various users and administrative concerns. Our experimentations concluded that mash-ups are a strong computing platform that helped us to resolve the issues faced by the users and administration of the journal. The online version of this system is available at www.jucs.org. During development of this mash-up we found that location (city, country) information of various authors in J.UCS is not complete. Without this information the position of authors or papers can not be precisely located on Google Maps. Although we applied heuristics to extract this information from the affiliation information of authors, the results still needed to be verified, as heuristics in some cases can extract wrong information. As a solution, we further used a community assisted geographical mash-up in confirming the geographical locations of authors and for updating other contents of the journal.

In the future, this work can be extended by creating multi-feature mash-up by combining data from more than one bibliographic databases. Such mash-up can be used for various academic appraisals such as: promotion, finding collaborators and experts, and conflict of interest detection. Similarly, there are various bibliographic databases such as CiteSeer that contain incomplete information about authors, publications, affiliations, and other meta-data. Moreover, they contain ambiguous authors' names. By ambiguous, we mean that two different authors can have similar name. Similarly, an author name or affiliation can have different representations which might be due to the limitations of the automated process used for the extraction of this information from articles, or by the mistake of data entry, and even by the author himself. Without disambiguated information, it is difficult in evaluating the accurate performance of authors and institutions. Although there are automated methods to resolve these issues, but they are not 100% accurate and requires human intervention to confirm the

findings. A community driven mash-up along with wikis (with administrative control), and the facility of intelligent task routing can really help in this regard. The availability of such a mash-up will allow people to add, update, and even disambiguate authors' names, affiliations, and departments.

– The second contribution of this thesis deals with enhancing the scientometrics and content analysis of scholarly publication in digital journals. Traditionally, these analyses were conducted using static tables and statistical charts. However, with the developments in the field of information visualization, such analyses have started to be supported by interactive visualizations techniques. In this context, we provided a brief overview about these techniques. We identified limitations of these techniques that may restrict the deeper scientometrics and content analysis. After conducting a survey of these techniques, we selected and applied an extended visualization technique which is primarily used for general data analysis, to help us for a deeper scientometrics and content analysis of digital journals. The adapted visualization system is an easy to use system that allowed us to find hidden patterns in the publications of J.UCS to strengthen its internal administration. We also employed this visualization system with few improvements to some selected journals and conferences in the field of e-learning to realize hidden research trends for experienced educators, researchers and policy makers working in the field of e-learning. Our experiments conducted on J.UCS conclude that the adopted visualization system is a powerful tool in determining the impact, coverage and the status of the journal at deeper level. Moreover, it is an effective tool to find hidden patterns in any given field of study.

In the future, we propose a detailed comparative usability study of different visualization techniques with our adapted visualization technique to find users preferences in different scenarios.

– In the third contribution, we provided an overview about the peer-review system of scholarly publications. We discuss the strengths and weaknesses of different forms of the peer review system. We highlight that different forms of conflict of interest (COI) situations can compromise the review decisions. We mainly categorized these COIs in Social COIs and Cognitive COIs. We described various automated systems and existing techniques that can be used to identify these COI situations. However, each of this system deals with social COI detections. According to literature, different citations relationships can be an indicator of both social and cognitive relationships between researchers. This can be useful for improving existing systems by highlighting the possibility of both social and cognitive COIs. To prove this hypothesis, in this contribution, we used basic citations relationships, i.e., co-citations, bibliographic coupling, and inter-citations along with their temporal information to identify social

relationships between our selected set of authors. Our experiments showed that our defined features succeeded in identifying a proportion of these social relationships with considerable precision. In the second part of our study, we tried to determine cognitive COIs between researchers using the same basic citations relationships. We discussed the ways to reveal the strength of cognitive COI between researchers by assigning weights to these citations relationships between them. However, we emphasized that the real severity of cognitive COIs can only be determined by assigning context and sentiments to these citations relationships. To demonstrate the potential of citations relationships to highlight possible cognitive COIs, we performed some experiments on the authors and reviewers of the WWW2006 conference's performance track. Based on the citations data about these authors and reviewers, we demonstrated how context and sentiments can be assigned to different citations relationships. In case of inter-citations, various studies in past have developed the context and sentiments classification scheme. We also used one of them in our experiments. However, based on an established theory, in this study, we also proposed a possible way to assign sentiments to even bibliographic coupling. Similarly, in previous studies, researchers have always focused in determining the context of only one type of co-citation relationship. In this work, we extended it to include more context types, based on the context types of inter-citations. We also performed experiments to highlight the possibility for the automated prediction of the context and sentiments associated with basic citations relationships. Although we did not find any severe case of cognitive COI for WWW2006 authors and reviewers, we believe that such analysis might be helpful in other cases. In a nutshell, we can conclude on the basis of above mentioned studies that citations are more than a source of evaluating intellectual impact. Rather, a careful modeling of citations can effectively lead to a powerful COI detection system. The results reported in this dissertation are first steps in this direction.

For future work, we suggest that although our experiments identified a certain proportion of social relationships from citations with considerable precision, there is a still need to apply our identified features in the prediction of the social networks of other authors to confirm the results. It is expected that the inclusion of other social relationships such as: friends, allies, regular correspondents, sought advices might further improve the results in future. However, the collection of this information is not easy. In case of cognitive COIs detection, there is a need to acquire the COIs declarations information from the administration of journals or conferences and tally this information with the cognitive COIs detected through our proposed approach to support our arguments more firmly.

– In the fourth contribution, we described the importance of downloads based usage

impact. The downloads of articles has the potential to timely foretell the impact of an article on scholarly community as compared to citations. In addition to that, it can also help in anticipating the future citations of an article. By keeping in view the importance of downloads, we tried to predict the future download counts for the articles published in J.UCS. We explored various attributes that are locally or globally associated with an article in our prediction experiments. We found that our selected features succeeded in reducing the prediction error up to 13% compared to the defined baseline error. Moreover, we did not use any posterior attributes. We selected only information which was available before publishing an article. By using only such prior features, we can anticipate the current value of an article even at the time of submission. In cases, where the administration of scholarly journals is facing expanding numbers of submissions, such predictions can be used as a criterion for pre-selecting an article before submitting it for a more rigorous review. It will help in saving the time of reviewers and journal's administrative tasks. Moreover, well written articles might not be overlooked under the volume of increasing submissions.

Although in our prediction results, we succeeded in reducing the error rate from its ground truth, we need more improvements to make it applicable for practical applications. For the future, we suggest to explore more additional features that can help in articles download counts predictions. One important direction might be to use the downloads information of other articles published before in the same digital journal in the same research area.

In conclusion, the current thesis provided many ideas and demonstrated their practical applications in enhancing the quality of content in the digital libraries of scholarly publications. We believe that the research presented in this thesis will support the administration of scholarly communications to maintain their reputation and general standing by delivering quality of services and content for scholarly communities.

Appendix: List of Publications of the Author

Journal Publications:

- Khan, M. S. (2011). Exploring Citations For Conflict Of Interest Detection In Peer-review System. (To appear in) International Journal of Computer Information Systems and Industrial Management Applications, 3.
- Khan, M. S. (2011). Estimating Articles Downloads For A Digital Journal. (To be submitted in) Journal of Digital Information (JoDI).
- Maurer, H. and Khan, M. S. (2010). Research Trends In The Field Of E-learning From 2003 To 2008: A Scientometric And Content Analysis For Selected Journals And Conferences Using Visualization. Interactive Technology and Smart Education, 7(1): 5-18.
- Khan, M. S., Kulathuramaiyer, N., and Maurer, H. (2008). Applications Of Mash-ups For A Digital Journal. Journal of Universal Computer Science, 14(10): 1695-1716.

Conference Publications:

- Khan, M. S. (2010). Can Citations Predict Socio-cognitive Relationships In Peer Review System? In IADIS, European Conference on Data Mining, 19-26.
- Khan, M. S., Maurer, H.(2009). Discovering Trends In The Field Of E-Learning From 2003 To 2008 Using Visualization. In IADIS International Conference e-Learning 2009, 88-97*.
- Khan, M. S., Ebner, M., Maurer, H.(2009). Trends Discovery In The Field Of E-Learning With Visualization. In Proceedings of 21st ED-Media Conference, World Conference on Educational Multimedia, Hypermedia and Telecommunications, 4408-4413.
- Khan, M. S., Ebner, M., Taraghi, B.(2009). Visualizing Research Patterns In The Field Of E-Learning. In Science 2.0 for TEL Workshop, see e.g. `http://www.telearn.org/warehouse/Salman_2009_TelSci2.0_(002199v1).pdf`

*This paper received "Best Paper" award at the conference.

- Khan, M. S., Afzal, M. T., Kulathuramaiyer, N., and Maurer, H. (2009). Extended Visualization For A Digital Journal. In Fifth International Conference on Web Information Systems and Technologies, Lisbon, Portugal, 385-388.

Bibliography

[Adamic and Adar, 2003] Adamic, L. and Adar, E. (2003). Friends And Neighbors On The Web. *Social Networks*, 25(3):211–230.

[Adler and de Alfaro, 2007] Adler, B. T. and de Alfaro, L. (2007). A Content-Driven Reputation System for the Wikipedia. In *International World Wide Web Conference*, pages 261–270, Banff, Alberta, Canada.

[Afzal et al., 2007] Afzal, M., N., K., and Maurer, H. (2007). Creating Links To The Future. *Journal of Universal Computer Science*, 13, 9:1234–1245.

[Afzal, 2010] Afzal, M. T. (2010). *Context Aware Information Discovery For Scholarly E-community.* PhD thesis, Institute for Information Systems and Computer Media, Graz University of Technology, Graz, Austria.

[Ahlgren et al., 2003] Ahlgren, P., Jarneving, B., and Rousseau, R. (2003). Requirements For A Cocitation Similarity Measure, With Special Reference To Pearson's Correlation Coefficient. *Journal of the American Society for Information Science and Technology,*, 54(6):550–560.

[Ahmed et al., 2004] Ahmed, A., Dwyer, T., Murray, C., Song, L., and Wu, Y. X. (2004). WilmaScope Graph Visualization. In *InfoVis'04, IEEE Symposium on Information Visualization*.

[Aleman-Meza et al., 2008] Aleman-Meza, B., Nagrajan, M., Ding, L., Sheth, A., Arpinar, I. B., Joshi, A., and Finin, T. (2008). Scalable Semantic Analytics On Social Networks For Addressing The Problem Of Conflict Of Interest Detection. *ACM Transaction on the Web*, 2(1):7:1–7:29.

[Andrews et al., 1995] Andrews, K., Kappe, F., and Maurer, H. (1995). The Hyper-G Network Information System. *Journal of Universal Computer Science*, 1(4):206–220.

[Angrosh et al., 2010] Angrosh, M. A., Cranefield, S., and Stanger, N. (2010). Context Identification Of Sentences In Related Work Sections Using A Conditional Random Field: Towards Intelligent Digital Libraries. In *Joint Conferene on Digital Libraries (JCDL)*, pages 293–302, Queensland, Australia.

[Arms, 2000] Arms, W. Y. (2000). *Digital Libraries.* MIT Press. Available from: `http://www.cs.cornell.edu/wya/DigLib/`.

[Arms, 2002] Arms, W. Y. (2002). What Are The Alternatives To Peer Review? Quality Control In Scholarly Publishing On The Web. *Journal of Electronic Publishing*, 8(1).

[Austria-Forum, 2011] Austria-Forum (2011). Available from: `www.austria-lexikon.at`.

[Baldi, 1998] Baldi, S. (1998). Normative Versus Social Constructivist Processes In The Allocation Of Citations: A Network-analytic Model. *American Sociological Review*, 63:829–846.

[Banks and Dellavalle, 2008] Banks, M. A. and Dellavalle, R. (2008). Emerging Alternatives To The Impact Factor. *OCLC Systems & Services: International digital library perspectives*, 24(3):167–173.

[Basu et al., 2001] Basu, C., Hirsh, H., Cohen, W., and Nevill-Manning, C. (2001). Technical Paper Recommendation: A Study In Combining Multiple Information Sources. *Journal of Artificial Intelligence Research*, 14:231–252.

[Bent et al., 2004] Bent, L., Rabinovich, M., Voelker, G. M., and Xiao, Z. (2004). Characterization Of A Large Web Site Population With Implications For Content Delivery. In *WWW2004*, New York, USA.

[Berg and Mrozowski, 2000] Berg, Z. and Mrozowski, S. (2000). Review Of Research In Distance Education 1990-1999. *The American Journal of Distance Education*, 15(3):5–19.

[Berners-Lee et al., 1994] Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., and Secret, A. (1994). The World Wide Web. *Communications of the ACM*, 37(8):76–82.

[Berners-Lee and Hendler, 2011] Berners-Lee, T. and Hendler, J. (2011). Scientific Publishing On The Semantic Web. Available from: `http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm`.

[Biaggioni, 1993] Biaggioni, I. (1993). Conflict-of-interest Guidelines: An Argument For Disclosure. *Pharmacy and Therapeutics*, 322.

[Bollen et al., 2008] Bollen, J., Van de Sompel, H., and Rodriguez, M. A. (2008). Towards Usage-based Impact Metrics: First Results From The Mesur Project. In *Joint Conference on Digital Libraries*, Pittsburgh, Pennsylvania, USA.

[Borgman, 1999] Borgman, C. L. (1999). What Are Digital Libraries? Competing Visions. *Information Processing and Management*, 35:227–243.

[Borgman and Furner, 2002] Borgman, C. L. and Furner, J. (2002). Scholarly Communication And Bibliometrics, Annual Review Of Information Science And Technology. *Information Today*, 36:3–72.

[Bormann and Daniel, 2008] Bormann, L. and Daniel, H. (2008). What Do Citations Counts Measure? A Review Of Studies On Citing Behavior. *Journal of Documentation*, 64(1):45–80.

[Boyack and Klavans, 2005] Boyack, K. W. and Klavans, R. (2005). Predicting The Importance Of Current Papers. In *International Conference of the International Society for Scientometrics and Informetrics*, Stolkhol, Sweden.

[Brank and Leskovec, 2003] Brank, J. and Leskovec, J. (2003). The Download Estimation Task On KDD Cup 2003. Technical report, Jozef Stefan Institute, Jamova 39, Ljubljana, Slovenia. Available from: `ai.ijs.si/kddcup03/kddcup03.pdf`.

[Brody et al., 2006] Brody, T., Harnad, S., and Carr, L. (2006). Earlier Web Usage Statistics As Predictors Of Later Citation Impact. *Journal of The American Society For Information Science And Technology*, 57(8):1060–1072.

[Bush, 1945] Bush, V. (1945). As We May Think. *The Atlantic Monthly*. Available from: `http://web.mit.edu/STS.035/www/PDFs/think.pdf`.

[Business Wire, 2010] Business Wire (2010). Available from: `http://www.businesswire.com/portal/site/google/?ndmViewId=news_view&newsId=20090921005858&newsLang=en`.

[Börner et al., 2003] Börner, K., Chen, C., and Bayak, K. W. (2003). *Visualizing Knowledge Domains*. Annual Review of Information Science and Technology, Inc/American Society for Information Science and Technology.

[Calude et al., 1994] Calude, C., Maurer, H., and Salomaa, A. (1994). Journal Of Universal Computer Science. *Journal of Universal Computer Science*, 0, 0:109–116.

[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., and Schuldt, H. (2007). The DELOS Digital Library Reference Model. Available from: `http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel%_0.98.pdf`.

[Casey and Savastinuk, 2006] Casey, M. E. and Savastinuk, L. C. (2006). Library 2.0: Service For The Next-generation Library. Available from: `http://www.libraryjournal.com/article/CA6365200.html`.

[Castillo et al., 2007] Castillo, C., Donato, D., and Gionis, A. (2007). Estimating Number Of Citations Using Author Reputation. In *SPIRE 2007, LNCS 4726*, pages 107–117.

[Chang and Lin, 2010] Chang, C.-C. and Lin, C.-J. (2010). Libsvm – A Library For Support Vector Machines. Available from: `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[Chapter 5 of CS445 in Yale, 2005] Chapter 5 of CS445 in Yale (2005). Classification: Alternative Techniques. Available from: `http://zoo.cs.yale.edu/classes/cs445/misc/chap5_other_classification.pdf`.

[Chatti et al., 2009] Chatti, M. A., Jarke, M., Wang, Z., and Specht, M. (2009). SMashup Personal Learning Environment. In *Mash-up Personal Learning Enironments (MUPPLE)*.

[Chen, 2005] Chen, C. (2005). Citespace Ii: Detecting And Visualizing Emerging Trends And Transient Patterns In Scientific Literature. *Journal of the American Society for Information Science and Technology*, 57 (3):359–77.

[Chu and Krichel, 2007] Chu, H. and Krichel, T. (2007). Downloads Vs. Citations: Relationships, Contributing Factors And Beyond. In *11th International Conference on Scientometrics and Informetrics*, Madrid, Spain.

[Chubin and Moitra, 1975] Chubin, D. E. and Moitra, S. D. (1975). Content Analysis Of References: Adjunct Or Alternative To Citation Counting? *Social Studies of Science*, 5:423–441.

[CiteSeer, 2009] CiteSeer (2009). Citeseer: Scientific Literature Digital Library And Search Engine. Available from: `http://citeseer.ist.psu.edu/`.

[CiteULike, 2011] CiteULike (2011). Citeulike. Available from: `http://www.citeulike.org/`.

[Cosley et al., 2007] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2007). SuggestBot: Using Intelligent Task Routing To Help People Find Work In Wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*.

[COUNTER, 2011] COUNTER (2011). Counter: Counting Online Usage Of Networked Electronic Resources. Available from: `http://www.projectcounter.org/`.

[Cronin, 1982] Cronin, B. (1982). Norms And Functions In Citation- The View Of Journal Editors And Referees In Psychology. *Social Science Information Studies*, 2:65–78.

[Cronin, 2005] Cronin, B. (2005). A hundred million acts of whimsy? *Current Science*, 89(9):1505–1509.

[Cronin and Shaw, 2002] Cronin, B. and Shaw, D. (2002). Identity-creator And Image-makers: Using Citation Analysis And Thick Description To Put Authors In Their Place. *Scientometrics*, 54(1):31–49.

[Davis, 2007] Davis, I. (2007). Talis, Web 2.0 And All That. Available from: `http://internetalchemy.org/2005/07/talis-web-20-and-all-that`.

[Davis, 2010] Davis, P. M. (2010). *Access, Readership, Citations: A Randomized Controlled Trial Of Scientific Journal Publishing.* PhD thesis, Faculty of the Graduate School of Cornell University.

[Delos, 2001] Delos (2001). Digital Libraries: Future Directions For A European Research Programme. Technical report, Alta Badia, Italy. Available from: `http://delos-noe.isti.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf`.

[DELOS, 2011] DELOS (2011). Delos Network Of Excellence. Available from: `http://www.delos.info/`.

[doc2mat, 2009] doc2mat (2009). doc2mat. Available from: `http://glaros.dtc.umn.edu/gkhome/files/fs/sw/cluto/doc2mat.html`.

[Dolan, 2001] Dolan, J. R. (2001). Contribution 3. In Inter-research Forum. Available from: `http://www.int-res.com/discussion-forums/meps-discussion-forum-2/#Dolan`.

[Dreher et al., 2004] Dreher, H., Krottmaier, H., and Maurer, H. (2004). What We Expect From Digital Libraries. *Journal of Universal Computer Science*, 10(9):1110–1122.

[Dumais and Nielsen, 1992] Dumais, S. T. and Nielsen, J. (1992). Automating The Assignment Of Submitted Manuscripts To Reviewers. *Research and Development in Information Retrieval*, page 233–244.

[Efimova and de Moor, 2004] Efimova, L. and de Moor, A. (2004). Beyond Personal Webpublishing: An Exploratory Study of Conversational Blogging Practices. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*.

[Ely et al., 1992] Ely, D., Foley, A., Freeman, W., and Scheel, N. (1992). Trends In Educational Technology. *Educational Media and Technology Yearbook*, 18:1–29.

[Engelbart, 1975] Engelbart, D. C. (1975). NLS Teleconferencing Features: The Journal, And Shared-screen Telephoning. In *CompCon75 Conference (IEEE Catalog No. 75CH0988-6C)*, pages 173–176.

[Engelbart and English, 1968] Engelbart, D. C. and English, W. K. (1968). A Research Center For Augmenting Human Intellect. In *Fall Joint Computer Conference*, volume 33, pages 395–410, San Francisco, CA.

[Erten et al., 2004] Erten, C., Harding, P. J., Kobourov, S. G., and Wampler, K. (2004). Exploring The Computing Literature Using Temporal Graph Visualization. Technical report, University of Arizona.

[Fischer et al., 2009] Fischer, T., Bkalov, F., König-Reis, B., Nauerz, A., and Welsch, M. (2009). An Evolutionary Algorithm for Auomatic Composition of Information-gathering Web Services in Mashups. In *Seventh IEEE European Conference on Web Services*.

[Fox, 1993] Fox, E. A. (1993). *Source Book On Digital Libraries.* National Science Foundation. Available from: `http://fox.cs.vt.edu/DigitalLibrary/DLSB.pdf`.

[Fox et al., 1995] Fox, E. A., Akscyn, R. M., Furuta, R. K., and Leggett, J. L. (1995). Digital Libraries. *Communications of the ACM*, 38(4):23–28.

[Fu and Aliferis, 2008] Fu, L. D. and Aliferis, C. (2008). Models For Predicting And Explaining Citation Count Of Biomedical Articles. In *AMIA Symposium*, pages 222–226.

[Gapminder, 2008] Gapminder (2008). Available from: `http://www.gapminder.org/world`.

[Garfield, 1962] Garfield, E. (1962). Can Citation Indexing Be Automated. *Essays of an Information Scientist*, 1:84–90.

[Garfield, 1970] Garfield, E. (1970). Citation Indexing For Studying Science. *Nature*, 227:669–671.

[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*, 178:471–490.

[Garzone and Mercer, 2000] Garzone, M. and Mercer, R. E. (2000). Towards An Automated Citation Classifier. In *13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, Springer-Verlag, pages 337–346.

[GeoBytes, 2007] GeoBytes (2007). Available from: `http://www.geobytes.com/FreeServices.htm`.

[Glänzel, 2003] Glänzel, W. (2003). *Bibliometric As A Research Field: A Course On Theory And Application Of Bibliometric Indicators.* Available from: `www.norslis.net/2004/BibModuleKUL.pdf`.

[Godlee and Dickersin, 2003] Godlee, F. and Dickersin, K. (2003). *Bias, subjectivity, chance, and conflict of interest in editorial decisions, Chapter 6 in Peer Review in Health Sciences.* BMJ Books, London, 2 edition.

[Gorraiz and Gumpenberger, 2010] Gorraiz, J. and Gumpenberger, C. (2010). Going Beyond Citations: SERUM — A New Tool Provided By A Network Of Libraries. *Liber Quarterly*, 20(1):80–93.

[Guédon, 2001] Guédon, J. (2001). In Oldenburg's Long Shadow : Librarians, Research Scientists, Publishers, And The Control Of Scientific Publishing, 2001. In Creating The Digital Future : Association Of Research Libraries. In *138th Annual Meeting*, Toronto, Ontario (Canada). Association of Research Libraries.

[Halasz, 1988] Halasz, F. G. (1988). Reflections On The Notecards: Seven Issues For The Next Generation Of Hypermedia Systems. *Communications of the ACM*, 31(7):836–852.

[Havre et al., 2002] Havre, S., Hetzler, E., Whitney, P., and Nowells, L. (2002). ThemeRiver: Visualizing Thematic Changes In Large Documents Collections. *IEEE Transaction on Visualization and Computer Graphics*, 8(1):9–20.

[Hawkins, 2001] Hawkins, D. T. (2001). Bibliometrics Of Electronic Journals In Information Science. *Information Research*, 7, 1. Available from: `http://InformationR.net/ir/7-1/paper120.html`.

[Heinrich and Maurer, 2000] Heinrich, E. and Maurer, H. (2000). Active documents: Concept, implementation and applications. *Journal of Universal Computer Science*, 6(12):1197–1202.

[Helic, 2001] Helic, D. (2001). *Aspects of Semantic Data Modeling in Hypermedia Systems.* PhD thesis, Institute for Information Systems and Computer Media, Graz University of Technology.

[Hitchcock, 2002] Hitchcock, S. M. (2002). *Perspectives In Electronic Publishing: Experiments With A New Electronic Journal Model.* PhD thesis, University of Southampton.

[Hodgins, 2008] Hodgins, W. (2008). The Snowflake Effect: The Future Of Mashups And Learning. Technical report, British Educational Communications and Technology Agency (BECTA). Available from: `http://dera.ioe.ac.uk/1504/1/becta_2009_emergingtechnologies_mashups_hodgins_report.pdf`.

[Hoyer and Fischer, 2008] Hoyer, V. and Fischer, M. (2008). Market Overview Of Enterprise Mashup Tools. In *ICSOC, Lecture Notes in Computer Science*.

[Hyperwave, 2007] Hyperwave (2007). Available from: `http://www.hyperwave.com/e/`.

[IBM Mashup Hub, 2010] IBM Mashup Hub (2010). Available from: `http://services.alphaworks.ibm.com/graduated/mashuphub.html`.

[In-SPIRE, 2007] In-SPIRE (2007). Available from: `http://in-spire.pnl.gov/`.

[Jamali et al., 2005] Jamali, H. R., Nicholas, D., and Huntington, P. (2005). The Use And Users Of Scholarly E-journals: A Review Of Log Analysis Studies. *Aslib Proceedings: New Information Perspectives*, 57(6):554–571.

[Janner et al., 2009] Janner, T., Siebeck, R., Schroth, C., and Hoyer, V. (2009). Patterns For Enterprise Mashups In B2B Collaborations To Foster Lightweight Composition And End User Development. In *IEEE International Conference on Web Services*.

[Johnson and Oppenheim, 2007] Johnson, B. and Oppenheim, C. (2007). How Socially Connected Are Citers To Those That They Cite. *Journal of Documentation*, 63(5):609–637.

[J.UCS, 2007] J.UCS (2007). Journal Of Universal Computer Science. Available from: `http//www.jucs.org`.

[Kahle et al., 1992] Kahle, B., Morris, H., Davis, F., Tiene, K., Hart, C., and Palmer, R. (1992). Wide Area Information Servers: An Executive Information System For Unstructured Files. *Internet Research*, 2(1):59 – 68.

[Kang et al., 2007] Kang, H., Plaisant, C., Lee, B., and Bederson, B. B. (2007). NetLens: Iterative Exploration Of Content-actor Network Data. *Information Visualization*, 6, 1:18–31.

[Ke et al., 2004] Ke, W., Börner, K., and Vishwanath, L. (2004). Major Information Visualization Authors, Papers And Topics In The ACM Library. In *IEEE Symposium on Information Visualization*.

[Kessler, 1963] Kessler, M. M. (1963). Bibliographic Coupling Between Scientific Papers. *American Documentation*, 14(1):10–25.

[Khan, 2010] Khan, M. S. (2010). Can Citations Predict Socio-cognitive Relationships In Peer Review System? In *IADIS, European Conference on Data Mining*.

[Khan, 2011] Khan, M. S. (2011). Exploring citations for conflict of interest detection in peer-review system. *(To appear in) International Journal of Computer Information Systems and Industrial Management Applications*, 3.

[Khan et al., 2009] Khan, M. S., Afzal, M. T., Kulathuramaiyer, N., and Maurer, H. (2009). Extended Visualization For A Digital Journal. In *Fifth International Conference on Web Information Systems and Technologies*, Lisbon, Portugal.

[Khan et al., 2008] Khan, M. S., Kulathuramaiyer, N., and Maurer, H. (2008). Applications of Mash-ups for a Digital Journal. *Journal of Universal Computer Science*, 14:1695–1716.

[Kieslinger and Lindstaedt, 2009] Kieslinger, B. and Lindstaedt, S. (2009). Science 2.0 Practices In The Field Of Technology Enhanced Learning. In *Science2.0 for TEL Workshop. ECTEL 2009*, Nice, France.

[Kittur et al., 2007] Kittur, A., Chi, E., Pendleton, A., Suh, B., and Mytkowicz, T. (2007). Power Of The Few Vs. Wisdom Of The Crowd: Wikipedia And The Rise Of The Bourgeoisie. In *CHI 2007*.

[Klein, 1997] Klein, J. (1997). Etr& D-development: An Analysis Of Content And Survey Of Future Direction. *Educational Technology Research and Development*, 45(3):57–62.

[Klerkx and Duval, 2007] Klerkx, J. and Duval, E. (2007). GlobeMash: A Mashup For Accessing GLOBE. In *Proc. 7th International conference on Knowledge Management (I-Know), Graz, Austria*.

[Kolbitsch and Maurer, 2006] Kolbitsch, J. and Maurer, H. (2006). The Transformation Of The Web: How Emerging Communities Shape The Information We Consume. *Journal of Universal Computer Science*, 12, 2:187–213.

[Kraut et al., 1988] Kraut, R., Egido, C., and Galegher, J. (1988). Patterns Of Contact And Communication In Scientific Research Collaboration. In *Proc. ACM Conference on Computer-Supported Cooperative Work, Portland, Oregon, USA*.

[Krottmaier, 2002] Krottmaier, H. (2002). *Aspects of Modern Electronic Publishing Systems*. PhD thesis, Institute for Information Processing and Computer Supported new Media.

[Krottmaier, 2003] Krottmaier, H. (2003). Links To The Future. *Journal of Digital Information Management*, 1, 1:3–8.

[Kulathuramaiyer, 2007] Kulathuramaiyer, N. (2007). Mashups: Emerging Application Development Paradigm For A Digital Journal. *Journal of Universal Computer Science*, 13, 4:531–542.

[Kundzewicz and Koutsoyiannis, 2005] Kundzewicz, Z. W. and Koutsoyiannis, D. (2005). Editorial-the Peer-review System: Prospects And Challenges. *Hydrological Sciences Journal*, 50(4):577–590.

[Kurtz et al., 2004] Kurtz, M. J., Eichorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., Martimbeau, N., and Elwell, B. (2004). The Bibliometric Properties Of Article Readership Information. *The Journal of the American Society for Information Science and Technology*, 56:111–128.

[Latchem, 2006] Latchem, C. (2006). Editorial: A Content Analysis Of The British Journal Of Educational Technology. *British Journal of Educational Technology*, 37(4):503–511.

[Lee et al., 2004] Lee, Y., Driscoll, M., and Nelson, D. (2004). The Past, Present, And Future Of Research In Distance Education: Results Of A Content Analysis. *The American Journal of Distance Education*, 18(4):225–241.

[LeJeune, 1999] LeJeune, L. (1999). Who Owns What? *Journal of Electronic Publishing*, 4(3).

[Lesk, 1997] Lesk, M. (1997). *Practical Digital Libraries: Books, Bytes, And Bucks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Leydesdorff, 2008] Leydesdorff, L. (2008). On The Normalization And Visualization Of Author Co-citation Data: Salton's Cosine Versus The Jaccard Index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85.

[Li and Wu, 2008] Li, Q. and Wu, Y. B. (2008). People Search: Searching People Sharing Similar Interests From The Web. *Journal of the American Society for Information Science and Technology*, 59(1):111–125.

[Lindhal and Blount, 2003] Lindhal, C. and Blount, E. (2003). Weblogs: Simplifying Web Publishing. *Computer*, 36, 11:114–116.

[Lippman, 1980] Lippman, A. (1980). Movie-Map: An Application Of The Optical Videodisc To Cinputer Graphics. In *7th annual conference on Computer graphics and interactive techniques*, pages 32–42.

[Lyman et al., 2011] Lyman, P., Varian, H. R., Swearingen, K., Charles, P., Good, N., Jordan, L. L., and Pal, J. (2011). How Much Information? 2003. Available from: `http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf`.

[Marchionini and Maurer, 1995] Marchionini, G. and Maurer, H. (1995). The Roles Of Digital Libraries In Teaching And Learning. *Communications of the ACM*, 38(4):67–75.

[Marcouiller and Deller, 2001] Marcouiller, D. W. and Deller, S. C. (2001). Thirty Years Of Academic Publishing In Regional Studies: A Content Analysis Of MCRSA's Scholarly Output. *JRAP*, 31, 2:33–43.

[Masood, 2004] Masood, M. (2004). A Ten-year Analysis: Trends In Traditional Educational Technology Literature. *Malaysian Online Journal of Instructional Technology*, 1(2):73–91.

[Matsuo et al., 2006] Matsuo, Y., Mori, J., and Hamasaki, M. (2006). POLYPHONET: An Advanced Social Network Extraction System From The Web. In *International World Wide Web Conference*, Edinburgh, Scotland.

[Maurer, 1982] Maurer, H. (1982). Will MUPID Revolutionize Austria's Videotex? In *Videotex 82*, pages 187–198, New York.

[Maurer, 1996] Maurer, H. (1996). *Hyperwave- The Next Generation Web Solution.* Addison Wesley Longman Pub. Co.

[Maurer, 2001a] Maurer, H. (2001a). Beyond Classical Digital Libraries. In *Proc. of Conference on Global Digital Library Development (Ding-chi Chen, ed.), Tsinghua University Press, Beijing.*

[Maurer, 2001b] Maurer, H. (2001b). Videotex. Encyclopedia of Computers and Computer History, Vol. 2 (Ed.: R. Rojas),. Available from: `http://www.iicm.tugraz.at/Ressourcen/Papers/videotex.doc`.

[Maurer, 2007] Maurer, H. (2007). Mass-market E-commerce In The 1980s. Available from: `http://www2007.org/webhistory.php`.

[Maurer and Kaiser, 1986] Maurer, H. and Kaiser, D. (1986). How To Develop A COSTOC Course. Technical Report 229, IIG.

[Maurer and Khan, 2010] Maurer, H. and Khan, M. S. (2010). Research Trends In The Field Of E-learning From 2003 To 2008: A Scientometric And Content Analysis For Selected Journals And Conferences Using Visualization. *Interactive Technology and Smart Education*, 7 (1):5–18.

[Maurer and Mueller, 2011] Maurer, H. and Mueller, H. (2011). How To Use The Web's Information Flood For Teaching. In *ED-MEDIA*, Lissabon, Portugal.

[Maurer and Posch, 1982] Maurer, H. and Posch, R. (1982). MUPID - An Austrian Contribution To Videotex. Technical report, Institute for Information Processing (IIG), Graz, Austria.

[Maurer and Schmaranz, 1994] Maurer, H. and Schmaranz, K. (1994). J.UCS - The Next Generation In Electronic Journal Publishing. *Journal of Universal Computer Science*, 0, 0:117–126.

[MaxMind, 2007] MaxMind (2007). Available from: `http://www.maxmind.com/`.

[McCahill and Anklesaria, 1995] McCahill, M. P. and Anklesaria, F. X. (1995). Evolution Of Internet Gopher. *Journal of Universal Computer Science*, 1(4):235–246.

[McElroy, 2002] McElroy, E. (2002). Dos And Don'ts For Electronic Journal Management: Some Advice To Publishers. *Learned Publishing*, 15, 2:125–128.

[McIsaac and Gunawardene, 1996] McIsaac, S. and Gunawardene, C. (1996). *Distance Education, Handbook Of Research For Educational Communications And Technology: A Project Of The Association For Educational Communications And Technology.* Simon & Schuster Macmillan, New York.

[Menz, 2000] Menz, H. B. (2000). The First Ten Year Of The Foot: A Retrospective Analysis Of Publication Patterns, 1991-2000. *Foot*, 11(3):113–118.

[Mibazaar, 2007] Mibazaar (2007). Available from: `http://www.mibazaar.com/`.

[Moed, 2005] Moed, H. F. (2005). Statistical Relationships Between Downloads And Citations At The Level Of Individual Documents Within A Single Journal. *Journal of The American Society For Information Science And Technology*, 56(10):1088–1097.

[Moravcsik and Poovanalingam, 1975] Moravcsik, M. J. and Poovanalingam, M. (1975). Some Results On The Function And Quality Of Citations. *Social Studies of Science*, 5(1):86–92.

[Mori et al., 2006] Mori, J., Tsujishita, T., Matsuo, Y., and Ishizuka, M. (2006). Extracting Relations In Social Networks From The Web Using Similarity Between Collective Contexts. In *Lecture Notes in Computer Science*, pages 487–500.

[Moser, 1998] Moser, M. (1998). Web Based Training Systems And Document Annotation – Implementations For Hyperwave. Master's thesis, Institute for Information Processing and Computer Supported New Media, Graz University of Technology.

[Murnieks et al., 2007] Murnieks, C. Y., Haynie, J. M., Wiltbank, R., and Harting, T. (2007). I Like How You Think: The Role Of Cognitive Similarity As A Decision Bias. In *Annual meeting of the Academy of Management, Philadelphia, PA.*

[Nanba and Okumura, 1999] Nanba, H. and Okumura, M. (1999). Classification Of Research Papers Using Citation Links And Citation Types: Towards Automatic Review Article Generation. In *IJCAI*, pages 926–931.

[Nelson, 1987] Nelson, T. H. (1987). *Literary Machines*.

[Nelson et al., 2007] Nelson, T. H., Smith, R. A., and Mallicoat, M. (2007). Back to the Future: Hypertext the Way It Used to Be. In *Eight ACM Conference on Hypertext and Hypermedia*, pages 227–227, Manchester, United Kingdom.

[Neslon, 1965] Neslon, T. H. (1965). A File Structure for the Complex, the Changing, and the Indeterminate. In *ACM 20th National Conference*, pages 84–100.

[Nicholas et al., 2003] Nicholas, D., Huntington, P., and Watkinson, A. (2003). Digital Journals, Big Deals And Online Searching Behavious: A Pilot Study. *Aslib Proceedings*, 55(1/2):84–109.

[Nielsen, 1995] Nielsen, J. (1995). *Multimedia And Hypertext. The Internet And Beyond.* Academic Press, Boston.

[Ochoa et al., 2009] Ochoa, X., Me´ndez, G., and Duval, E. (2009). Who We Are: Analysis Of 10 Years Of The Ed-media Conference. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 189–200.

[Odlyzko, 1994] Odlyzko, A. M. (1994). Tragic Loss Or Good Riddance? The Impending Demise Of Traditional Scholarly Journals. *Journal of Universal Computer Science*, 0(0):54–108.

[Oppenheim, 1997] Oppenheim, C. (1997). The Correlation Between Citations Counts And The 1992 Research Assessment Exercise Ratings For British Research In Genetics, Anatomy And Archaeology. *Journal of Documentation*, 53(5):477–487.

[Oppenheim and Renn, 1978] Oppenheim, C. and Renn, S. P. (1978). Highly Cited Old Papers And The Reasons Why They Continue To Be Cited. *Journal of the American Society for Information Science*, 29(5):227–231.

[O'Reilly, 2007] O'Reilly, T. (2007). What Is Web 2.0 Design Patterns And Business Models For The Next Generation Of Software. Available from: `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`.

[Pam and Vermeer, 1995] Pam, A. and Vermeer, A. (1995). A Comparison of WWW and Hyper-G. *Journal of Universal Computer Science*, 1(11):744–750.

[Papagelis et al., 2005] Papagelis, M., Plexousakis, D., and Nikolaou, P. N. (2005). Confious: Managing The Electronic Submission And Reviewing Process Of Scientific Conferences. In *Proceedings of the 6th International Conference on Web Information Systems Engineering*, New York.

[ParsCit, 2011] ParsCit (2011). Parscit. Available from: `http://aye.comp.nus.edu.sg/parsCit/`.

[Paul, 2005] Paul, G. (2005). Web2.0. Available from: `http://www.paulgraham.com/web20.html`.

[Pearl, 1989] Pearl, A. (1989). Sun's Link Service: A Protocol For Open Linking. In *Second ACM Conference on Hypertext*, pages 137–146.

[Perneger, 2004] Perneger, T. V. (2004). Relation Between Online "hit Counts" And Subsequent Citations: Prospective Study Of Research Papers In The BMJ. *BMJ*, 329:546–547.

[Pham and Hofmann, 2003] Pham, S. B. and Hofmann, A. (2003). A New Approach For Scientific Citation Classification Using Cue Phrases. In *Australian Joint Conference in Artificial Intelligence*, Perth, Australia.

[Priedhorsky et al., 2007] Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, Destroying, And Restoring Value In Wikipedia. In *GROUP*, pages 259–268, Sanibel Island, Florida, USA.

[Programmableweb, 2007] Programmableweb (2007). Available from: `http://www.programmableweb.com/`.

[Pöschl, 2004] Pöschl, U. (2004). Interactive Journal Concept For Improved Scientific Publishing And Quality Assurance. *Learned Publishing*, 17(2):105–113.

[Radoulov, 2008] Radoulov, R. (2008). Exploring Automatic Citation Classification. Master's thesis, University of Waterloo.

[Ramussen and Karypis, 2008] Ramussen, M. and Karypis, G. (2008). gCLUTO – An Interactive Clustering, Visualization, and Analysis System, CSE/UMN Technical Report: TR# 04-021. Technical report. Available from: `www-users.cs.umn.edu/~mrasmus/gcluto/doc/gcluto-1.2/report.pdf`.

[Rennie, 1993] Rennie, D. (1993). More Peering Into Editorial Peer Review. *Journal of the American Medical Association*, 270(23):2856–2858.

[Robertson et al., 2008] Robertson, G., Fernandez, R., Fisher, D., Lee, B., and Stasko, J. (2008). Effectiveness Of Animation In Trend Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14 (6):1325 –1332.

[Rockwell, 2010] Rockwell, S. (2010). Ethics Of Peer Review: A Guide For Manuscript Reviewers. Available from: `http://radonc.yale.edu/pdf/Ethical_Issues_in_Peer_Review.pdf`.

[Rodriguez and Bollen, 2008] Rodriguez, M. A. and Bollen, J. (2008). An Algorithm to Determine Peer-Reviewers. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 319–328.

[Rodriguez et al., 2006] Rodriguez, M. A., Bollen, J., and Van de Sompel, H. (2006). The Convergence Of Digital Libraries And The Peer-review Process. *Journal of Information Science*, 32(2):149–159.

[Roes, 1994] Roes, H. (1994). Electrnic Journals: A Survey Of The Literature And The Net. *Journal of Information Networking*, 2(3):169–186.

[Rowlands, 1999] Rowlands, I. (1999). Patterns Of Author Cocitation In Information Policy: Evidence Of Social, Collaborative And Cognitive Structure. *Scientometrics*, 44(3):533–546.

[Rowlands and Nichols, 2005] Rowlands, I. and Nichols, D. (2005). New Journal Publishing Models: An International Survey Of Senior Researchers. Available from: `www.ucl.ac.uk/ciber/ciber_2005_survey_final.pdf`.

[Saeed et al., 2010] Saeed, A., Afzal, M., Latif, A., and Tochtermann, K. (2010). Disseminating Knowledge through Tags: Recommending Tags for Scientific Resources. *Journal of IT in Asia*, 3:25–36.

[Salton and McGill, 1983] Salton, G. and McGill, M. (1983). *Introduction To Modern Information Retrieval*. McGraw-Hill, Auckland, New Zealand.

[Saracevic, 2001] Saracevic, T. (2001). Digital Library Evaluation: Toward An Evolution Of Concepts. *Library Trends*, 49(3):350–369.

[Schauder, 1994] Schauder, D. (1994). Electronic Publishing Of Professional Articles: Attitudes Of Academics And Implications For The Scholarly Communication Industry. *Journal of the American Society for Information Science*, 48(2):73–100.

[Shackel, 1991] Shackel, B. (1991). Blend-9: Overview and Appraisal. *British Library Research Paper No. 82*.

[Shannon, 1951] Shannon, C. (1951). Prediction And Entropy Of Printed English. *Bell System Technical Journal*, 30:50–64.

[Shepherd, 2007] Shepherd, P. T. (2007). The Feasibility Of Developing And Implementing Journal Usage Factors: A Research Project Sponsored By UKSG. *Serials*, 20(2):117–123.

[Sheridan et al., 1981] Sheridan, T., Senders, J., Moray, N., Stoklosa, J., Guillaume, J., and Makepeace, D. (1981). Experimentation With A Multi-disciplinary Teleconference And Electronic Journal On Mental Workload. Technical report, National Science Foundation, Division of Science Information Access Improvement.

[Shih et al., 2007] Shih, M., Fang, J., and Tsai, C. (2007). Research And Trends In The Field Of E-learning From 2001 To 2005: A Content Analysis Of Cognitive Studies In Selected Journals. *Computers & Education*, 15(2):955–967.

[Shneiderman, 1996] Shneiderman, B. (1996). The Eyes Have It: A Task By Data Type Taxonomy For Information Visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343.

[Shri et al., 2010] Shri, R., Sanjay, K., Nitin, P., and Alan, H. (2010). Acceptance And Usage Of Web 2.0 Services In Libraries: A Survey. In *ETTLIS*.

[Small, 1973] Small, H. G. (1973). Co-citation In The Scientific Literature: A New Measure Of Relationship Between Two Documents. *Journal of the American Society for Information Science*, 24(4):265–269.

[Spiegel-Rosing, 1977] Spiegel-Rosing, I. (1977). Social Science Studies: Bibliometric And Content Analysis. *Studies of Science*, 7(1):97–113.

[Szybalski, 2005] Szybalski, A. (2005). Why It's Not A Wiki World (yet). Available from: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.4709&rep=rep1&type=pdf`.

[Tanimoto, 1957] Tanimoto, T. T. (1957). Technical report, Internal report: IBM Technical Report Series.

[Taraghi et al., 2009] Taraghi, B., Ebner, M., and Schaffert, S. (2009). Personal Learning Environments For Higher Education: A Mashup Based Widget Concept. In *Mash-up Personal Learning Enironments (MUPPLE)*.

[Tatsunokuchi, 2010] Tatsunokuchi (2010). Available from: `http://en.wikipedia.org/wiki/Tatsunokuchi,_Ishikawa`.

[Taylor, 2001] Taylor, E. W. (2001). Adult Education Quarterly From 1989 To 1999: A Content Analysis of All Submissions. *Adult Education Quarterly*, 51, 4:322–340.

[TechSoup, 2007] TechSoup (2007). Mashups: An Easy, Free Way To Create Custom Web Apps. Available from: `http://www.techsoup.org/learningcenter/webbuilding/page5788.cfm?cg=searchterms&sg=mashups`.

[Teufel and Moens, 2000] Teufel, S. and Moens, M. (2000). What's Yours And What's Mine: Determining Intellectual Attribution In Scientific Text. In *Joint SIGDAT Conference on Emprical Methods in NLP*.

[Teufel et al., 2006a] Teufel, S., Siddharthan, A., and Tidhar, D. (2006a). An Annotation Scheme For Citation Function. In *7th SIGdial Workshop on Discourse and Dialogue*, pages 80–87, Sydney, Australia. Assosiation for Computational Linguistics.

[Teufel et al., 2006b] Teufel, S., Siddharthan, A., and Tidhar, D. (2006b). Automatic Classification Of Citation Function. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 103–110, Sydney, Australia. Assosiation for Computational Linguistics.

[The Open Mashup Alliance, 2010] The Open Mashup Alliance (2010). Available from: `http://www.openmashup.org/`.

[Triggle and Triggle, 2007] Triggle, C. R. and Triggle, D. J. (2007). What Is The Future Of Peer Review? Why Is There Fraud In Science? Is Plagiarism Out Of Control? Why Do Scientists Do Bad Things? Is It All A Case Of:"all That Is Necessary For The Triumph Of Evil Is That Good Men Do Nothing?". *Vascular Health and Risk Management*, 3(1):39–53.

[Turoff and Hiltz, 1982] Turoff, M. and Hiltz, S. R. (1982). The Electronic Journal: A Progress Report. *Journal of the American Society for Information Science*, 33(4):195–202. Available from: `http://web.njit.edu/~turoff/Papers/ElectronicJournal.html`.

[Tutarel, 2002] Tutarel, O. (2002). Geographical Distribution Of Publications In The Field Of Medical Education. *BMC Medical Education*, 2(3).

[Ukkonen, 1992] Ukkonen, E. (1992). Approximate String-matching With Q-grams And Maximal Matches. *Theoretical Computer Science*, 92:191–21. Available from: `http://www.cs.helsinki.fi/u/ukkonen/TCS92.pdf`.

[Vahdat et al., 2010] Vahdat, A., Chase, J., and Dahlin, M. (2010). The Perfect Storm: Reliability Benchmarking For Global-scale Services. Available from: `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.9699`.

[van Dam, 2011] van Dam, A. (2011). Memex And Beyond Web Site, Hypertext '87 Keynote Address. Available from: `http://www.cs.brown.edu/memex/HT_87_Keynote_Address.html`.

[van Rooyen et al., 1998] van Rooyen, S., Godlee, F., Evans, S., Smith, R., and Black, N. (1998). Effect Of Blinding And Unmasking On The Quality Of Peer Review: A Randomized Trial. *Jounal of American Medical Association*, 280(3):234–237.

[Waldrop, 2011] Waldrop, M. M. (2011). Science 2.0 – Is Open Access Science The Future? Available from: `http://www.scientificamerican.com/article.cfm?id=science-2-point-0`.

[Watson, 2009] Watson, A. B. (2009). Comparing Citations And Downloads For Individual Articles. *Journal of Vision*, 9(4):1–4.

[Watt, 2010] Watt, S. (2010). Mashups – The Evolution Of The SOA, Part 2: Situational Applications And The Mashup Ecosystem. Available from: `http://www.ibm.com/developerworks/webservices/library/ws-soa-mashups2/`.

[Weka, 2009] Weka (2009). Data Mining Software In Java. Available from: `http://www.cs.waikato.ac.nz/ml/weka/`.

[White and Griffith, 1981] White, H. D. and Griffith, B. C. (1981). Author Cocitation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science*, 32(3):163–171.

[White et al., 2004] White, H. D., Wellman, B., and Nazer, N. (2004). Does Citation Reflect Social Structure? Longitudinal Evidence From The 'Globenet' Interdisciplinary Research Group. *Journal of the American Society for Information Science and Technology*, 55(2):111–126.

[Wikipedia, 2007] Wikipedia (2007). Available from: `http://www.wikipedia.org/`.

[Wikipedia, 2011a] Wikipedia (2011a). Wikipedia:Featured articles. Available from: `http://en.wikipedia.org/wiki/Wikipedia:FA`.

[Wikipedia, 2011b] Wikipedia (2011b). Wikipedia:user Access Levels. Available from: `http://en.wikipedia.org/wiki/Wikipedia:User_access_levels`.

[Xanadu, 2011] Xanadu (2011). Project xanadu. Available from: `http://www.xanadu.com/`.

[Yahoo Pipes, 2010] Yahoo Pipes (2010). Available from: `http://pipes.yahoo.com/pipes/`.

[Yang et al., 2008] Yang, L., Morris, S., and Barden, E. (2008). Mapping Institutions And Their Weak Ties In A Specialty: A Case Study Of Cystic Fibrosis Body Composition Research. *Scientometrics*, 79(2):421–434.

[Yankelovich et al., 1985] Yankelovich, N., Meyrowitz, N., and van Dam, A. (1985). Reading And Writing The Electronic Book. *IEEE Computer*, 18(10):15–30.

[Yarowsky and Florian, 1999] Yarowsky, D. and Florian, R. (1999). Taking The Load Off The Conference Chairs: Towards A Digital Paper-routing Assistant. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very-Large Corpora*.

[Zhang et al., 2005] Zhang, H., Jiang, L., and Su, J. (2005). Hidden Naive Bayes. In *The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence*, page 919–924.

[Zhao and Strotmann, 2008] Zhao, D. and Strotmann, A. (2008). Evolution Of Research Activity And Intellectual Influences In Information Sciences 1996-2005: Introducing Author Bibliographic-coupling Analysis. *Journal of the American Society for Information Science and Technology*, 59(13):2070–2086.

[Zuccala, 2006] Zuccala, A. (2006). Modeling The Invisible College. *Journal of the American Society for Information Science and Technology*, 57(2):152–168.