Web Mining for Self-Directed E-learning

Prasanna Desikan, Colin DeLong, Kalyan Beemanapalli, Amit Bose,
Jaideep Srivastava
University of Minnesota

Abstract

Self-directed e-learning focuses on the independent learner, one who engages in education at his own pace, free from curricular obligation. A number of tools, some purposefully and others serendipitously, have become key enablers of this learning paradigm. For example, tools such a Google Scholar, CiteSeer Research Index, etc. make it possible to do literature search without stepping out of one's room. Due to the same technologies which helped make self-directed e-learning possible in the first place, these tools are in danger of delivering diminishing returns as micro-learning, lifelong education, and continuous education become the norm in our Information Age. Web Mining, however, may potentially offer a solution to this issue. In this chapter, we investigate specific examples of self-directed e-learning and how their functionality and utility can be improved through the use of Web Mining technology, techniques, and practices. Our work demonstrates the usefulness of Web Mining as it applies to self-directed e-learning and the need to map implicit relationships in learner behaviour, usage, and context.

1 Introduction

The introduction of the World Wide Web has had a profound impact on education, reducing the necessity of a learner and teacher to share the same physical space, and creating an entirely new form of knowledge delivery. With an ever-increasing number of Internet users and websites, online learning, training, and online educational multimedia – all generally referred to as "e-learning" – are becoming increasingly prevalent [1]. Additionally, while some educational outlets have used e-learning to supplement existing brick-and-mortar instruction (using software such as Web-CT [2]), others have replaced traditional instruction all together and replaced it with e-learning, creating a Virtual

University [3]. The reasons for the increase in e-learning, and their sociological implications are almost as numerous as the systems available to enable e-learning [4].

There is, however, a common thread linking most of these systems, i.e. the user is given the ability to access expert information with some level of interaction. It is this level of interaction that is a key distinguishing feature when comparing different kinds of e-learning systems. For instance, the advantages of a web-enabled video feed may have only marginal utility over attending the lecture being taped; and may, in fact, be worse since face-to-face real time interaction is lost. Thus, e-learning must keep the people it's designed for in mind in order to be effective. Further, since individual needs differ, there is no reason why a single learning or teaching technique will work equally well for everyone.

Significant differences exist between students, such as their learning rate, personal interests, and a priori domain knowledge. If e-learning delivery can be brought into alignment with these individual traits, the learners' experience can be vastly improved over current models. As a specific example, material could be adapted to each student, or to a group of them (i.e., a class), who share some characteristics pertaining to the desired (target) knowledge or context [5].

However, designing systems that pre-determine all possible usage scenarios is not feasible. Additionally, it may not be practical or efficient in many situations because of the diverse and rapidly-changing requirements of learners. What is necessary is an informed e-learning system that continually "educates" itself about the requirements of its learners, while delivering material that is most appropriate for individual learners.

Towards this, Web mining techniques have been used to help identify usage behaviour characteristics not obvious by other methods. There has been extensive amount of work on Web usage-based mining [6], including various aspects such as proper data preparation and pre-processing [7], web usage-based recommendations for e-learning [8, 9, 10], and models to assist online e-learning assessment [11].

Web mining can also be used to aid a user by integrating the implicit information from multiple sources of Web data. At the simplest level, it can be a keyword-oriented search. However, learning is often aided by the inclusion of other kinds of data, such as a concept hierarchy and usage data. Often meta-information, such as authors, citations, and other expert-defined data, also help improve the learning process.

Given that Web Mining techniques can extract knowledge from the behaviour of past users to help future ones, these techniques have much to offer existing e-learning systems.

Focus of our study: In this chapter, we examine how e-learning systems can be improved using various Web Mining techniques, and provide example applications that help illustrate our claims. Given the broad scope of e-learning, we will focus on "*self-directed e-learning*", a facet of e-learning in which the learner is able to access a vast amount of expert-defined information, but is not necessarily subject to curricular constraints (i.e., semesters, grades, etc). Thus, the focus here is on improving the user experience in self-directed e-learning systems through the use of Web Mining techniques.

This chapter is organized as follows: In section 2 we provide the motivation for self-directed e-learning and present scenarios that describe its nature and significance. Section 3 presents some prominent self-directed e-learning applications that exist today in different domains, with a brief discussion on what they have to offer. In section 4 we discuss the gaps in existing technologies by presenting issues that have not yet been addressed for an efficient self-directed e-learning. An introduction to Web Mining, the state-of-the art research in the area and how it can be applied to overcome the existing gaps in technologies is discussed in section 5. We identify possible research directions to enable efficient self-directed e-learning in Section 6. Finally, in the last section we summarize the ideas discussed herein and provide conclusions.

2 Why Self-Directed E-Learning?

W. Edwards Deming once said, "Learning is not compulsory, but neither is survival". In our current context - a world increasingly driven by knowledge – this is an especially salient observation. In many ways, a successful life depends on what we have learned and our understanding of the constantly evolving world around us; and thus one must engage in continuous learning to remain relevant. However, the learning that we are talking about is not restricted to traditional notions of learning bound to a specific physical location, such as high schools, colleges, and universities. Rather it is an **ongoing process** whereby its extent and nature may vary from person to person. Some may do it for pleasure, some for the love of learning, and some out of plain necessity. Some may invest more time in order to gain a deep understanding while others make merely be looking for a cursory overview. As such, self-directed e-learning is well-suited to reconcile these differing learner preferences. Such a system enjoys the benefits of access to a wealth of information via the Internet, as well as knowledge of individual learning habits, so that the specific needs of each individual learner can be catered to.

- Consider, for example, the situation of a **stay-at-home mother**. The task of raising a child can require knowledge of many kinds. Right from monitoring the hourly activities of a new-born to disciplining a young school-going kid, there are numerous issues involved that needs to be handled with care. While attending to a doctor for regular check-ups is necessary, often moms are faced with situations that need expert advice in monitoring the growth and daily activities of their children. Typical problems faced by such a mom are two-fold -availability of information, and flexibility in time to access such information. With the advent of Internet and e-learning systems, these problems are no more a technical issue. Mothers can access information from Web sites that offer expert advice or information on child related issues or they can participate in forums where parents can exchange their ideas and experiences. This enables them to learn from the experts, as well as experiences of other parents, on the best way to approach the issues involved with their children. The infrastructure helps provide moms with multiple sources of information that can be accessed at their leisure or time of necessity, providing them a perfect platform to learn about the variety of aspects involved in child upbringing.

- Another common example is of a **graduate student** searching for interesting topics of research in her field. She starts out with a survey of literature pertaining to her interests and continues on to newsgroups, discussion forums, Internet search engines (like Google) and digital libraries (such as those provided by the IEEE). At the outset, she may have only a vague idea of what she is actually looking for. What she requires is a facilitator that recognizes her interests and introduces online material that can spark further research. A self-directed e-learning system can be such a facilitator: nominal input from the individual can produce a rich set of information while bringing their understanding into balance with the depth of the material being researched.

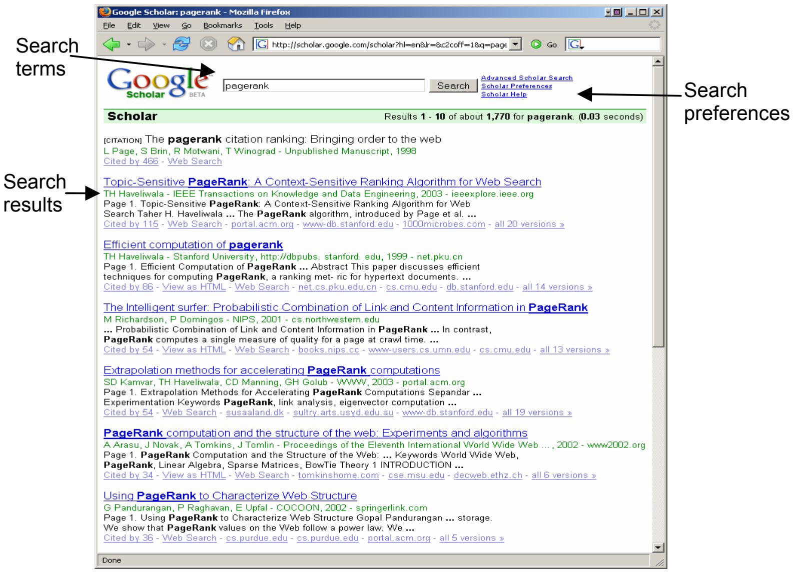
The examples above highlight some important points about learning today. It is a continuous process and the knowledge sought can vary from individual to individual in form, content, depth, and purpose. Learning for the most part is a highly focused endeavour, but flexible in its method, and directed by the individual's preferences. *Micro-learning* would be a more appropriate term for this. On the other hand material required for learning is *readily available*, thanks to the ubiquitous Internet. A seemingly limitless amount of information of all kinds is available through millions of web-sites and web-pages. We are also equipped with *tools and technologies* that allow us to access this data anywhere, anytime, within a few mouse clicks.

Even so, leveraging this enormous body of information to learn effectively and efficiently remains a challenging task. *Individualization* is a key requirement for this, and we can improve the knowledge-acquiring experience by learning from previous instances. In this way, mining data from *past learning experiences* can provide useful insights into human learning methodology, which in turn will go a long way in providing that personal touch to the learning experience. Data mining - and web mining in particular - can provide immense opportunities to realize the true potential of self-directed e-learning.

3 Web-Based Self-Directed E-Learning Applications

A number of examples of self-directed e-learning exist. We focus on some of the most notable ones and give a short overview of what each has to offer.

3.1 Google Scholar



http://scholar.google.com/scholar?hl=en&lr=&c2coff=1&q=page

pagerank

Results **1** - **10** of about **1,770** for **pagerank**. (**0.03** seconds)

[CITATION] The **pagerank** citation ranking: Bringing order to the web
L Page, S Brin, R Motwani, T Winograd - Unpublished Manuscript, 1998
Cited by 466 - Web Search

Topic-Sensitive **PageRank**: A Context-Sensitive Ranking Algorithm for Web Search
TH Haveliwala - IEEE Transactions on Knowledge and Data Engineering, 2003 - ieeexplore.ieee.org
Page 1. Topic-Sensitive **PageRank**: A Context-Sensitive Ranking Algorithm for Web Search Taher H. Haveliwala **...** The **PageRank** algorithm, introduced by Page et al. **...**
Cited by 115 - Web Search - portal.acm.org - www-db.stanford.edu - 1000microbes.com - all 20 versions »

Efficient computation of **pagerank**
TH Haveliwala - Stanford University, http://dbpubs. stanford. edu, 1999 - net.pku.cn
Page 1. Efficient Computation of **PageRank ...** Abstract This paper discusses efficient techniques for computing **PageRank**, a ranking met- ric for hypertext documents. **...**
Cited by 86 - View as HTML - Web Search - net.cs.pku.edu.cn - cs.cmu.edu - db.stanford.edu - all 14 versions »

The Intelligent surfer: Probabilistic Combination of Link and Content Information in **PageRank**
M Richardson, P Domingos - NIPS, 2001 - cs.northwestern.edu
... Probabilistic Combination of Link and Content Information in **PageRank ...** In contrast, **PageRank** computes a single measure of quality for a page at crawl time. **...**
Cited by 54 - View as HTML - Web Search - books.nips.cc - www-users.cs.umn.edu - cs.cmu.edu - all 13 versions »

Extrapolation methods for accelerating **PageRank** computations
SD Kamvar, TH Haveliwala, CD Manning, GH Golub - WWW, 2003 - portal.acm.org
Page 1. Extrapolation Methods for Accelerating **PageRank** Computations Sepandar **...** Experimentation Keywords **PageRank**, link analysis, eigenvector computation **...**
Cited by 54 - Web Search - susaaland.dk - sultry.arts.usyd.edu.au - www-db.stanford.edu - all 19 versions »

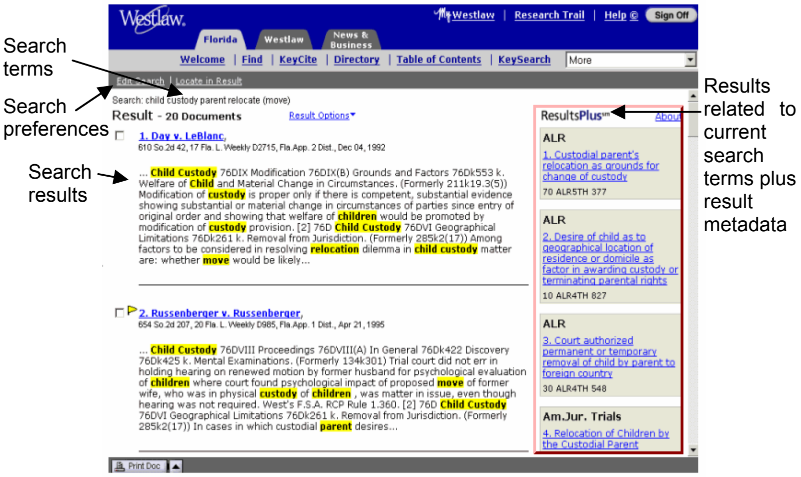
PageRank computation and the structure of the web: Experiments and algorithms
A Arasu, J Novak, A Tomkins, J Tomlin - Proceedings of the Eleventh International World Wide Web ... , 2002 - www2002.org
Page 1. **PageRank** Computation and the Structure of the Web: **...** Keywords World Wide Web, **PageRank**, Linear Algebra, Sparse Matrices, BowTie Theory 1 INTRODUCTION **...**
Cited by 34 - View as HTML - Web Search - tomkinshome.com - cse.msu.edu - decweb.ethz.ch - all 6 versions »

Using **PageRank** to Characterize Web Structure
G Pandurangan, P Raghavan, E Upfal - COCOON, 2002 - springerlink.com
Page 1. Using **PageRank** to Characterize Web Structure Gopal Pandurangan **...** storage. We show that **PageRank** values on the Web follow a power law. We **...**
Cited by 36 - Web Search - cs.purdue.edu - cs.purdue.edu - portal.acm.org - all 5 versions »

Figure 1: Google Scholar.

Google Scholar [12] is essentially a search engine for academic publications that are available on-line; each publication linked to others by way of citations. Both natural language and Boolean searches are possible. The searches themselves can be customized to query certain authors, publications, or within a year range (e.g., 1999 – 2003). The search results are similar to Google's regular search, and are generally in order of descending citation count (because Google's result-ranking algorithm, PageRank, relies on "backlinks" – citations, in this case – to influence result authority).

3.2 WestLaw


Westlaw.
My Westlaw | Research Trail | Help | Sign Off
Florida
Westlaw
News & Business
Welcome | Find | KeyCite | Directory | Table of Contents | KeySearch
More
Edit Search | Locate in Result
Result Options
About
Print Doc

Search: child custody parent relocate (move)

Result - 20 Documents

□ 1. Day v. LeBlanc,
610 So.2d 42, 17 Fla. L. Weekly D2715, Fla.App. 2 Dist., Dec 04, 1992

... Child Custody 76DIX Modification 76DIX(B) Grounds and Factors 76Dk553 k. Welfare of Child and Material Change in Circumstances. (Formerly 211k19.3(5)) Modification of custody is proper only if there is competent, substantial evidence showing substantial or material change in circumstances of parties since entry of original order and showing that welfare of children would be promoted by modification of custody provision. [2] 76D Child Custody 76DVI Geographical Limitations 76Dk261 k. Removal from Jurisdiction. (Formerly 285k2(17)) Among factors to be considered in resolving relocation dilemma in child custody matter are: whether move would be likely...

□ 2. Russenberger v. Russenberger,
654 So.2d 207, 20 Fla. L. Weekly D985, Fla.App. 1 Dist., Apr 21, 1995

... Child Custody 76DVIII Proceedings 76DVIII(A) In General 76Dk422 Discovery 76Dk425 k. Mental Examinations. (Formerly 134k301) Trial court did not err in holding hearing on renewed motion by former husband for psychological evaluation of children where court found psychological impact of proposed move of former wife, who was in physical custody of children , was matter in issue, even though hearing was not required. West's F.S.A. RCP Rule 1.360. [2] 76D Child Custody 76DVI Geographical Limitations 76Dk261 k. Removal from Jurisdiction. (Formerly 285k2(17)) In cases in which custodial parent desires...

ResultsPlus℠

ALR
1. Custodial parent's relocation as grounds for change of custody
70 ALR5TH 377

ALR
2. Desire of child as to geographical location of residence or domicile as factor in awarding custody or terminating parental rights
10 ALR4TH 827

ALR
3. Court authorized permanent or temporary removal of child by parent to foreign country
30 ALR4TH 548

Am.Jur. Trials
4. Relocation of Children by the Custodial Parent

Search terms

Search preferences

Search results

Results related to current search terms plus result metadata

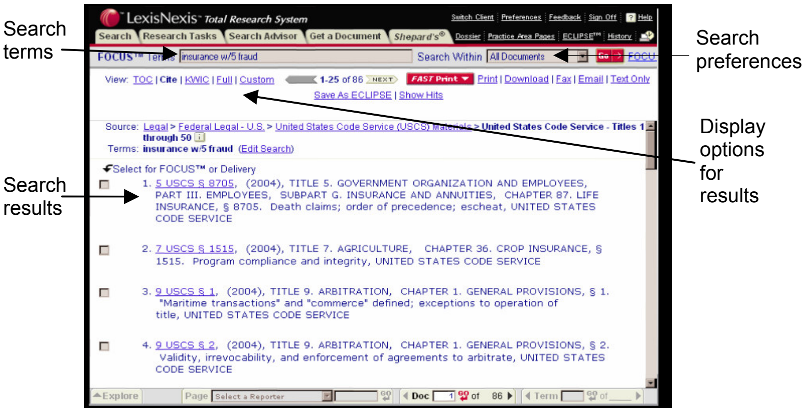
Figure 2: WestLaw: search system for law documents.

Westlaw [13] is a sophisticated search and retrieval system for legal documents and other ancillary material available in Thompson West's proprietary database. Similar to Google Scholar, Westlaw offers a natural language search as well as Boolean search options, called "Terms and Connectors". Searches are extremely customizable: by date range, document database (e.g., state, federal, circuit court, etc), and type of document (e.g., cases and statutes). The results look similar to those found in other search engines – albeit more legal text-heavy – and also include a sidebar called "ResultsPlus". ResultsPlus does background searches for relevant documents using the current set of search keywords as well as metadata contained within the set of returned results. The ResultsPlus list can be customized to search specific document archives, helping pare down the results to those of interest to the user.

3.3 LexisNexis

LexisNexis [14] is a collection of information search and retrieval tools coupled with a vast library of available documents, including those related to law, academic, law enforcement, news, market intelligence, government, and insurance. As is the case with Westlaw, LexisNexis searches offer a wide array of search options, including natural language and Boolean searches, results customizable by document archive, similar result documents, and selected text.

LexisNexis offers additional capabilities in its display of results, of which there are 5 options, namely Cite (title/headline, author, source, and date), Show Hits (displays keyword matches in results), KWIC (displays "Key Words In Context", similar to Show Hits), Full (displays full text of documents in results), and Custom (displays selected portions of documents).


Search terms
Search preferences
Display options for results
Search results
LexisNexis Total Research System
Switch Client | Preferences | Feedback | Sign Off | Help
Search
Research Tasks
Search Advisor
Get a Document
Shepard's®
Dossier | Practice Area Pages | ECLIPSE™ | History
Go
FOCU
View: TOC | Cite | KWIC | Full | Custom
NEXT
FAST Print
Print | Download | Fax | Email | Text Only
Save As ECLIPSE | Show Hits
Select for FOCUS™ or Delivery
Explore
Page
Select a Reporter
Doc
Term

FOCUS™ Terms: insurance w/5 fraud

Search Within: All Documents

1-25 of 86

Source: Legal > Federal Legal - U.S. > United States Code Service (USCS) Materials > United States Code Service - Titles 1 through 50

Terms: insurance w/5 fraud (Edit Search)

□ 1. 5 USCS § 8705, (2004), TITLE 5. GOVERNMENT ORGANIZATION AND EMPLOYEES, PART III. EMPLOYEES, SUBPART G. INSURANCE AND ANNUITIES, CHAPTER 87. LIFE INSURANCE, § 8705. Death claims; order of precedence; escheat, UNITED STATES CODE SERVICE

□ 2. 7 USCS § 1515, (2004), TITLE 7. AGRICULTURE, CHAPTER 36. CROP INSURANCE, § 1515. Program compliance and integrity, UNITED STATES CODE SERVICE

□ 3. 9 USCS § 1, (2004), TITLE 9. ARBITRATION, CHAPTER 1. GENERAL PROVISIONS, § 1. "Maritime transactions" and "commerce" defined; exceptions to operation of title, UNITED STATES CODE SERVICE

□ 4. 9 USCS § 2, (2004), TITLE 9. ARBITRATION, CHAPTER 1. GENERAL PROVISIONS, § 2. Validity, irrevocability, and enforcement of agreements to arbitrate, UNITED STATES CODE SERVICE

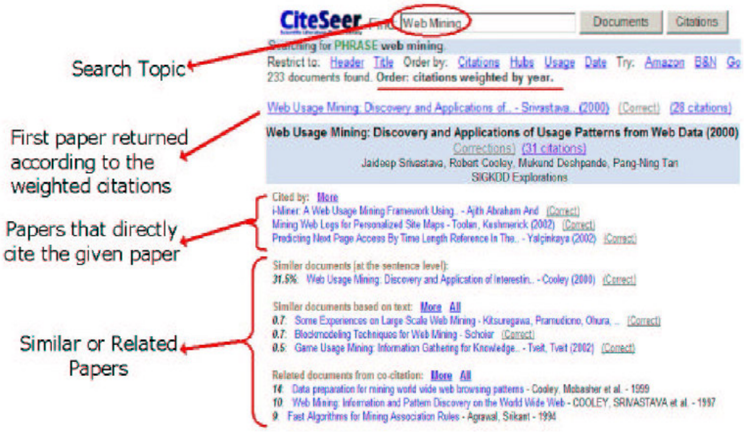
Doc 1 of 86

Figure 3: LexisNexis

3.4 CiteSeer

CiteSeer [15] is one of the most popular online bibliographic indices related to Computer Science. The key contribution of the CiteSeer repository is the ``Autonomous Citation Indexing'' (ACI) [16]. Citation indexing makes it possible to extract information about related articles. Automating such a process eliminates significant human effort and makes the search more effective and efficient. The key concepts are depicted in Figure 4.

Information about citations and their context is stored for each of these documents. The full text of the document is stored in several different formats. Information about documents that are similar at a sentence-level (percentage of sentences that match between the documents), at a text level, or related due to co-citation are also given. Citation statistics for documents are computed that enable the user to look at the most cited or popular documents in the related field. They also maintain a directory for computer science related papers, to make search based on categories easier. These documents are ordered by the number of citations. For a learner interested in documents related to computer science, this provides a very good repository and meta-level information for self-directed learning.


CiteSeer
Find:
Documents
Citations
Search Topic
First paper returned according to the weighted citations
Papers that directly cite the given paper
Similar or Related Papers

Web Mining

Searching for PHRASE web mining.

Restrict to: Header Title Order by: Citations Hubs Usage Date Try: Amazon B&N Go

233 documents found. Order: citations weighted by year.

Web Usage Mining: Discovery and Applications of.. - Srivastava.. (2000) (Correct) (28 citations)

Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000)
Corrections) (31 citations)
Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan
SIGKDD Explorations

Cited by: More
i-Miner: A Web Usage Mining Framework Using.. - Ajith Abraham And (Correct)
Mining Web Logs for Personalized Site Maps - Toolan, Kushmerick (2002) (Correct)
Predicting Next Page Access By Time Length Reference In The.. - Yalçinkaya (2002) (Correct)

Similar documents (at the sentence level):
31.5%: Web Usage Mining: Discovery and Application of Interestin.. - Cooley (2000) (Correct)

Similar documents based on text: More All
0.7: Some Experiences on Large Scale Web Mining - Kitsuregawa, Pramudiono, Ohura, .. (Correct)
0.7: Blockmodeling Techniques for Web Mining - Scholar (Correct)
0.5: Game Usage Mining: Information Gathering for Knowledge.. - Tveit, Tveit (2002) (Correct)

Related documents from co-citation: More All
14: Data preparation for mining world wide web browsing patterns - Cooley, Mobasher et al. - 1999
10: Web Mining: Information and Pattern Discovery on the World Wide Web - COOLEY, SRIVASTAVA et al. - 1997
9: Fast Algorithms for Mining Association Rules - Agrawal, Srikant - 1994

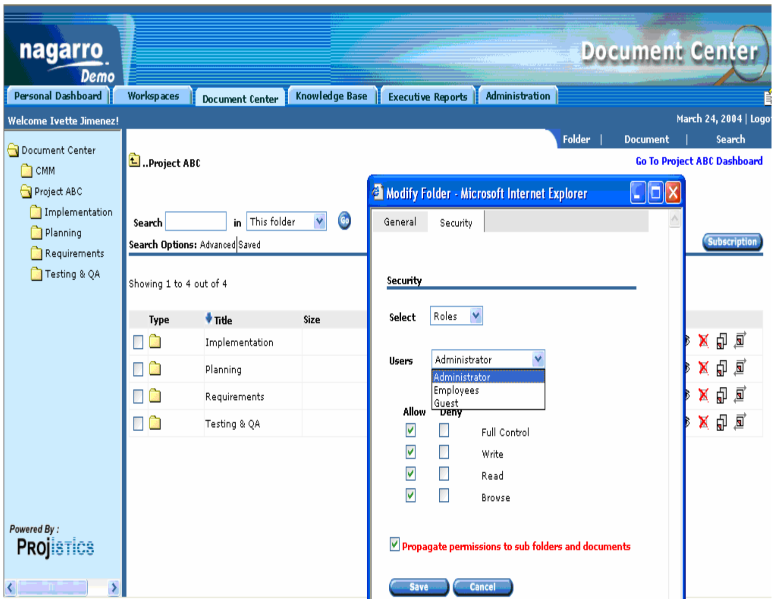
Figure 4: CiteSeer.

3.5 Knowledge Management Systems

Knowledge Management is defined as preserving actively and systematically the knowledge that is available in an organization. The interest in knowledge management in organizations has seen a recent surge. Companies emphasize the importance of relationship between knowledge and learning. Choenni, Walker, Bakker, and Baets [17], discuss the significance of an e-learning environment in knowledge management and the challenges in its implementation. The e-learning is self-directed as the people in an organization search the knowledge base to find the required information.

In the industry many knowledge management solutions or are software available. Here we will discuss one of them to highlight the present trend and the types of features they provide. Figure 5 shows the snapshot of a knowledge management and document management software called Projistics [18].

The main features this Knowledge Management system provides are document management, knowledge bases, persistent discussion threads, check-in/check-out functionality, extensive audit trail and change history maintenance, approval routing, and configurable workflows. Though the software suite is an excellent e-learning based environment, it does not serve personalized content to its users.



Security

Select: Roles

Users: Administrator (Administrator, Employees, Guest)

Allow	Deny	
☑	□	Full Control
☑	□	Write
☑	□	Read
☑	□	Browse

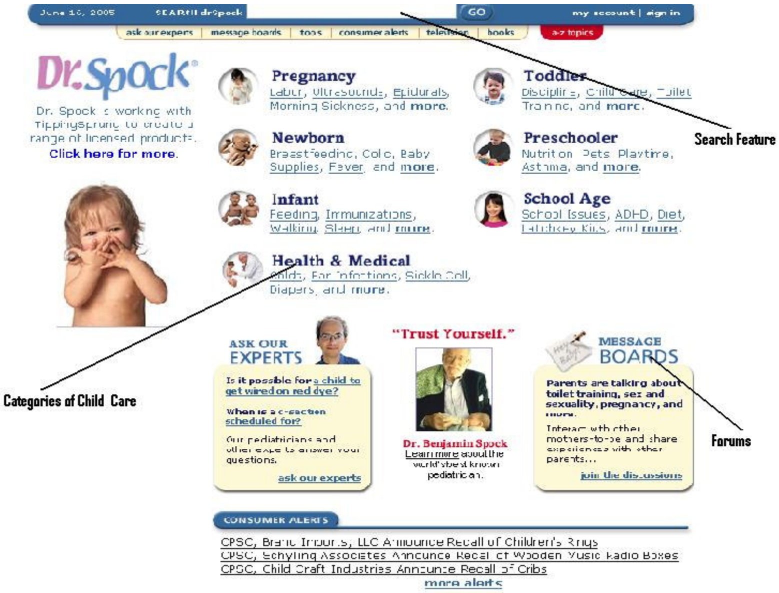
☑ Propagate permissions to sub folders and documents

Figure 5: Projistics – KMS.

3.6 Dr.Spock's Child Care

Figure 6 shows the web interface provided by 'The Dr.Spock Company' [19]. It is a leading parenting media and merchandising company that provides parents with latest expert advice, information, and inspiration on raising children. The company embodies the strength and identity of world-renowned paediatrician Dr. Benjamin Spock, providing parents with the latest expert content from today's leading authorities in parenting and children's health.

The search feature is a useful tool that enables parents to find information they are looking for. It also has discussion forums where parents can post questions that will either be answered by experts in the field or others parents. The information on the website is methodically classified for easy retrieval.



Dr. Spock is working with TippingSprung to create a range of licensed products. Click here for more.

Pregnancy
Labor, Ultrasounds, Epidurals, Morning Sickness, and **more**.

Toddler
Discipline, Child Care, Toilet Training, and **more**.

Newborn
Breastfeeding, Colic, Baby Supplies, Fever, and **more**.

Preschooler
Nutrition, Pets, Playtime, Asthma, and **more**.

Infant
Feeding, Immunizations, Walking, Sleep, and **more**.

School Age
School Issues, ADHD, Diet, Latchkey Kids, and **more**.

Health & Medical
Colds, Ear Infections, Sickle Cell, Diapers, and more.

ASK OUR EXPERTS

Is it possible for a child to get wired on red dye?

When is a c-section scheduled for?

Our pediatricians and other experts answer your questions.

ask our experts

"Trust Yourself."

Dr. Benjamin Spock
Learn more about the world's best known pediatrician.

MESSAGE BOARDS

Parents are talking about toilet training, sex and sexuality, pregnancy, and more.

Interact with other mothers-to-be and share experiences with other parents...

join the discussions

CONSUMER ALERTS

CPSC, Brand Imports, LLC Announce Recall of Children's Rings
CPSC, Schylling Associates Announce Recall of Wooden Music Radio Boxes
CPSC, Child Craft Industries Announce Recall of Cribs
more alerts

Figure 6: Dr.Spock's Child Care.

4 Gaps in Existing Technology

The aforementioned example applications, while maintaining high levels of quality in provided content and demonstrated utility through their widespread use, have issues that remain unsolved. E-learning systems today focus on the technology aspect with apparently lesser efforts spent on developing a system that can be tailored and adapted to individual learners. A brief discussion on shortcomings in current systems follows.

4.1 Lack of Community Collaboration

Consider the following scenario: there's a stack of documents on a table that many people are seated at. Occasionally, someone will grab one or two documents, perhaps more, and then leave the table. Other people arrive, sit at the table, and do the same. At no time does anyone converse with each other, even if two people have looked at the same document. They do not discuss what one may have located that the other did not, despite the potential of identifying additional relevant information.

All of the previous examples share this approach. Each user of the respective services poses queries, browses results, and learns in isolation from other users. In essence, there is no community aspect to self-directed e-learning, where likeminded users can contribute to each other’s research, as is the case in a real classroom.

Despite the fact that each person’s learning requirements may be different from others, there are often wide areas of overlap between individuals that can be mutually beneficial. Similarity in learning needs define *functional communities* of learners. These are virtual communities with fairly vague and overlapping definitions. Moreover, such communities are dynamic in the sense that needs of learners may change over time. Satisfying such communities is a difficult task, and yet it is conceivable to develop systems that learn from some in order to help the others. Most systems of today do barely little to tap into the colossal amounts of usage information already available.

4.2 Time Management

For many users, time is a precious commodity, pressing them to accomplish as much as possible as quickly as is possible. Current self-directed e-learning systems impose several requirements on their users, assuming a relatively equal distribution of a priori knowledge about the subject matter and the capability of always being able to properly formulate the “right question”. All of us have, at some point or other, struggled to find what we were looking for on the web, essentially because we couldn’t examine the right sources or pose the right queries. This imposition detracts from the experiences of users, especially the newer ones, who may spend inordinate amounts of time looking for documents that have very specific wordings. Compound this with the ever-increasing size of document repositories and querying can become difficult even for those who are well versed in the domain’s knowledge.

4.3 Not Self-improving

Perhaps this statement is too obvious, but it deserves mentioning anyway: the vast majority of search result lists are not 100% precise. Were they so, any query for documents would return a perfect list of relevant results. Current self-directed e-learning systems, while providing powerful sets of tools for querying, still operate from a self-contained idea of document relevancy, whereby the outcomes the results are meant to produce are not mapped back to their model of what makes a document relevant in the first place. Logs of usage of online documents, for example, constitute an implicit feedback from users about the relevance of these documents in different contexts. The challenge lies in extracting and deciphering user feedback from these massive repositories of data.

4.4 Implicit Relationships Not Mapped

Another opportunity for improvement is in the area of inter-document relationships. Certainly, one document is related to another if it is explicitly defined through a citation, which can be considered a “link”. However, it is also the case that one document can be related to another implicitly, even if they are not linked by citation. Search results try to map this through textual similarity, where documents are related if they share some of the same words or phrases. In general, this works fairly well, but is prone to failure when such words or phrases can be mapped to multiple contexts. It is in this, the identification of user context, where results can be distilled to match those of the user's intent. In self-directed e-learning systems – and search engines more broadly – this is a problem yet unsolved.

Deducing inter-document relationships is also a severe issue with present knowledge-management systems. Knowledge management systems of today hardly mine data for new knowledge in the form of hitherto undiscovered relationships or trends connecting employees, management, or other stakeholders, their activities, and leveraging this information effectively throughout an organization.

5 Web Mining

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. A panel organized at ICTAI 1997 [20] asked the question "Is there anything distinct about Web mining (compared to data mining in general)?" While no definitive conclusions were reached then, the tremendous attention on Web mining in the past five years, and a number of significant ideas that have been developed, have answered this question in the affirmative in a big way. In addition, a fairly stable community of researchers interested in the area has been formed, largely through the successful series of WebKDD workshops, which have been held annually in conjunction with the ACM SIGKDD Conference since 1999 and the Web Analytics workshops, which have been held in conjunction with the SIAM data mining conference. Kosala and Blockeel [21] provide a good survey of the research in the field till the end of 1999.

Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks [22]. Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [23]. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers that have addressed the issue. In this paper we define the data-centric view of Web mining, which is defined as,

"Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data."

The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. And its application in e-learning has also found its utility. In the following sub-sections, we will describe the taxonomy of Web Mining Research and applicability of Web Mining to E-learning.

5.1 Web Mining Taxonomy

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories and a figure depicting the taxonomy is shown in Figure

Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images, in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.

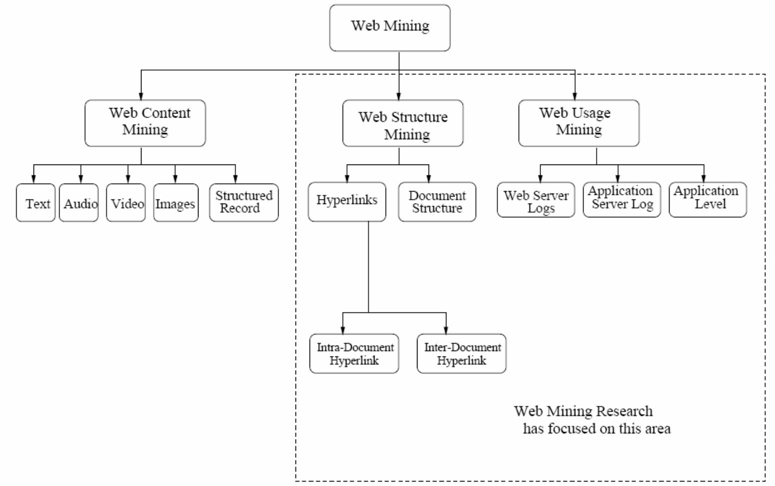


Figure 7: Web Mining Taxonomy.

Web Structure Mining: The structure of a typical Web graph consists of Web pages as nodes~, and hyperlinks as edges connecting related pages. Web

Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.

- **Hyperlinks:** A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis, of which [24] provides an up-to-date survey.

- **Document Structure:** In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications \cite{srivastava00web}. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

- **Web Server Data:** The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

- **Application Server Data:** Commercial application servers, e.g. Weblogic, etc. have significant features in the framework to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

- **Application Level Data:** Finally, new kinds of events can always be defined in an application, and logging can be turned on for them - generating histories of these specially defined events.

5.2 Web Mining Research – State of the Art

The interest of the research community and the rapid growth of the work in this area have resulted in good surveys over the past years that presents updated work and points to directions of further work [21,23,25]. Research on Web Content Mining has focused on issues such as extracting information from structured and unstructured data and integration of information from the various sources of content. Earlier work on Web Content mining can be found in Kosala's work [21]. Web Content mining together with other kinds of Web data can be used for application such as Web Page Categorization, Topic Distillation. Liu and Chang [26], in their work, have presented some of the key issues in Web Content Mining that has captured the attention of research community. Research in Web structure mining has focused primarily on hyperlink analysis and has found its utility in a variety of applications. A survey on hyperlink analysis techniques and a methodology to pursue research has been proposed by Desikan et al. [24]. Among these techniques, PageRank [27], developed by Google founders, is the most popular metric for ranking hypertext documents according to their importance. The key idea is that a page has high rank if many highly ranked pages point it to. So the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined. Oztekin et al [28], proposed Usage Aware PageRank incorporating usage statistics into framework of PageRank model. The other popular metric is *hub* and *authority* scores. From a graph theoretic point of view, *hubs* and *authorities* can be interpreted as 'fans' and 'centers' in a bipartite core of a Web graph. The hub and authority scores for a page are not based on a formula for a single page, but are computed for a set of pages related to a topic using an iterative procedure called HITS algorithm [29].

Web usage data has captured attention due to its nature of bringing user's perspective of the Web as opposed to creator's perspective. Understanding user profiles and user navigation patterns for better adaptive web sites and predicting user access patterns has evoked interest to the research and the business community. Methods for pre-processing the user log data and to separate web page references into those made for navigational purposes and those made for content purposes have been developed [7]. Perkowitz and Etzioni introduced the concept of adaptive web in their work [30]. Since then, Markov models have been used extensively to predict user behaviour [31, 32, 33]. An extensive updated survey on Web usage mining can be found in [34].

5.3 Web Mining applicable to E-learning

We have described in the previous subsections how information can be extracted from different kinds of Web data. From a direct perspective, information extraction can be viewed as a form of learning. Web mining techniques has been effectively used in search engine technologies to retrieve the most relevant and significant pages. However, the contribution of Web mining has not been restricted to such explicitly available information such as page content. It must be noted that learning is often aided with inclusion of other kinds of data such as concept hierarchy on which a Web structure is based or usage information. These kinds of data do not directly reflect the information in the page but help in building the context and circumstances in which such information is sought.

Web usage mining techniques as discussed earlier can be used to discover user navigation patterns. The user in our case is the self-directed learner. The creator of the Web pages would represent the expert who has designed the Web site to represent a series of notes. However, it is the usage information that actually reflects how a user is navigating or learning from the Web site. Such usage information can not only serve as a useful feedback to the experts about the learners approach, but can also suggest to learners from the 'navigation experience' of other user's on what they found useful. Initial work on analysing Web logs to discover patterns and associations between Web pages visited provided the right direction for such kind of analysis, but did not especially address the issue of e-learning. These kinds of analysis can be done either offline or online, or integrating both. A natural extension to such analysis was to develop recommender systems based on offline [9] as well as an integrated approach.

Web mining techniques coupled with integrated meta-information such as author info, download info, and other additional info explicitly defined by a domain expert helps to improve the learning process. Given a large, knowledge-dense website and a non-expert user seeking information, recommending relevant content becomes a significant challenge. Web mining has also been shown as a useful tool for providing expert-driven recommendations to non-experts, helping them understand what they NEED to know, as opposed to what is popular among other users. Another different dimension of Web Mining has focussed on modelling user navigation behaviour. The popular techniques are based on the first order Markov model where the user is modelled as a random surfer. Other models include a Markov chain model and reducing the 'randomness' factor by introducing a *bias* either based on the past usage patterns [28] or due to natural clustering of documents [35].

6 Future Directions of Research

Looking ahead, much can be done through the use of web mining to improve self-directed e-learning applications such that their function is in closer alignment with the expectations of their users. In particular, web mining is capable of realizing relationships that more accurately map what is known in a target domain, enabling web mining-enhanced self-directed e-learning systems to be brought into epistemological balance with other more traditional methods of learning, as in classroom education.

Several methods can be brought to bear in the realm of self-directed e-learning systems that could significantly enhance the overall user experience:

6.1 Usage Rules

One of the key components of most recommender systems, such as the one used by Amazon.com, is the mining of usage patterns from server logs [36, 37]. As an indicator of how a website is used, usage data can also be used to identify relevant documents based on the browsing and querying experiences of other users. These documents can then be presented to the user as they browse results in the form of a sidebar of recommendations.

In essence, this is akin to having the users, who are all sitting at the same table with the same set of documents, converse with each other. Not only do usage-based recommendations connect users to other documents of interest, they can also help maximize research time by reducing the number of 'guess-work' queries.

6.2 Keyword Clustering: The Conceptual Thesaurus

Searching can be tedious work if one is unsure of how to formulate the "question", or in this case, the query. Wouldn't it be better if the e-learning system could figure out what one might be searching for automatically? Often one has to search for something multiple times using different keywords/phrases till you found the right combination.

By clustering query keywords together into conceptually similar groups, we can suggest similar search terms/phrases or return results from closely-related keyword clusters [38]. This could easily be combined with usage rules to display a sidebar of results from related searches, as well as some suggestions for related "next" searches that are connected to the current set of keywords implicitly.

6.3 Recommendation Mining

The mining of recommendations themselves – either in the search results or sidebar results - could be used to help pare down or expand sidebar results and/or search keyword recommendations by building a model of result relevancy given a user context. For instance, if one searches for "link analysis" and the sidebar has a result no one ever clicks on, this may help identify a low confidence recommendation, which one may not want to include in the future.

6.4 Smart Results – model of relevance

The most popular model of user navigation that has been used to rank Web documents of a search query has been a random surfer model. This is based on the assumption that each user who issues the query is at the same level of knowledge and wants to explore the topic to the same depth of knowledge. However, in a real world scenario that is not entirely true. Not all users have in depth knowledge about all topics. The goal of a novice browsing a topic is more likely to get an overview of the field as opposed to an expert, who is aware of what the field is about and would like to explore in depth. Existing systems do not incorporate this kind of model. Though there are systems that take into account user profile to a certain extent and current navigation sessions, the relevance of Web pages to each user is not taken into account.

6.5 Intelligent Knowledge Management Systems

Web logs are the clear indicators of users' browsing behaviour. Web Mining can be applied to these logs to extract valuable information regarding the user interest, his generic profile, etc. Based on this extracted knowledge, when the user accesses the system in the future, personalized recommendations can be made. Tang and McCalla [39] explain how this task can be achieved. We present an example scenario with recommendations to illustrate the type of recommendations possible in an E-learning Knowledge Management system.

Example Scenario: A user submits the query "Inventory Management" to the system. When the user re-visits the system, the following recommendations can be made:

- Last week you searched for "Inventory Management". A new paper has been published in a journal related to this topic. Would you like to look at it?
- Based on your previous visits we have discovered that, you might be interested in the following topics.
  - Demand Forecasting
  - Supply-Chain Management
  - Warehouse Resources
- Users have answered a question posted by you in the forum. Here is a link to the answer. Would you also like to look at the answers given by these users to other questions?

Such recommendations will enable the user of a knowledge management system, to find a broader range of relevant information.

7 Conclusion

Current trends are clear indicators that online learning is gaining in importance. Self-directed e-learning provides the right mix of technology and individualization that can enhance the learning experience. We have illustrated numerous examples of self-directed e-learning systems, yet there exist wide gaps in current technology that hinder the potential of e-learning. Some of these shortcomings were discussed. Web-mining techniques have been immensely successful in a variety of application domains, and this leads us to believe that web mining techniques will enable us to overcome the limitations in current e-learning systems. Towards this end, we have provided a glimpse of what web mining is today, and outlined research areas in the field that have the potential to improve the efficacy of self-directed e-learning.

8 Acknowledgements

We would like to thank Data Analysis and Management Research group at the Department of Computer Science for providing valuable suggestions. This work was supported by AHPCRC contract number DAAD19-01-2-0014. The content

of the work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred.

References

[1] E-learning, Derek Stockely, http://derekstockley.com.au/eindex2b.html

[2] WebCT, http://www.webct.com

[3] Groeneboer, T. C. C. & Stockley, D., Virtual-u: A collaborative model for online learning environments. Second International Conference on Computer Support for Collaborative Learning, Toronto, Canada, December 1997.

[4] Naber, L. & Köhle, M., If e-Learning is the Answer, what was the Problem? Proceedings of AusWeb 2002

[5] Carchiolo, V., Longheu, A. & Malgeri, M., Adaptive Formative Paths in a Web-based Learning Environment. Educational Technology & Society, 5 (4), 2002.

[6] Srivastava, J., Cooley, R., Deshpande, M. & Tan, P., Web usage mining: Discovery and applications of usage patterns form web data. SIGKDD Explorations, 1(2), January 2000.

[7] Cooley, R., Mobasher, B. & Srivastava, J., Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1(1), pp. 5–32, 1999.

[8] Zaïane, O. R., Web Usage Mining for a Better Web-Based Learning Environment. Proc. of Conference on Advanced Technology for Education, pp 60-64, Banff, Alberta, June 27-28, 2001.

[9] Zaïane, O. R., Building a Recommender Agent for e-Learning Systems. Proc. of the 7th International Conference on Computers in Education (ICCE 2002). pp 55-59, Auckland, New Zealand, December 3 - 6, 2002.

[10] Cooley, R., Tan, P. N. & Srivastava, J., Websift: the Web site information filter system. Proceedings of the Web KDD 1999

[11] Ling Guo, Xin Xiang, YuanChun Shi, Use Web Usage Mining to Assist Background Online E-Learning Assessment. 4th IEEE ICALT, 2004.

[12] Google Scholar, http://scholar.google.com

[13] WestLaw, http://training.west.thomson.com/index.asp

[14] LexisNexis, http://www.lexisnexis.com /elearning/lexis/researchtool/ searchtool.asp

[15] CiteSeer, http://citeseer.ist.psu.edu/cs

[16] Lawrence, S., Giles, C. L. & Bollacker K., Digital libraries and autonomous citation indexing. IEEE Computer, 32(6), pp. 67-71, 1999.

[17] Choenni S, Walker R, Bakker R, and Baets W: E-learning as a Vehicle for Knowledge Management, In: Proceedings of the 14th Conf. on Applications of Prolog, INAP 2001 / 2001

[18] Projistics, http://www.projistics.com/

[19] Dr. Spock Company, http://www.drspock.com/

[20] Srivastava, J. & Mobasher, B., Panel discussion on “Web Mining: Hype or Reality?” ICTAI 1997.

[21] Kosala, R. & Blockeel, H., Web mining research: A survey. SIGKDD Explorations, 2(1), pp. 1 - 15, 2000.

[22] Etzioni, O., The World Wide Web: Quagmire or Gold Mine? Communications of the ACM, 39(11), pp. 65-68, November 1996.

[23] Cooley, R., Mobasher, B. & Srivastava, J., Web mining: information and pattern discovery on the world wide web. 9th IEEE ICTAI 1997.

[24] Desikan, P., Srivastava, J., Kumar, V. & Tan, P. N., Hyperlink Analysis: Techniques and Applications. Technical Report 2002-0152, Army High Performance Computing and Research Center, 2002.

[25] Srivastava, J., Desikan, P. & Kumar, V., Web Mining - Concepts, Applications and Research Directions (Book Chapter). Data Mining: Next Generation Challenges and Future Directions, MIT/AAAI 2004.

[26] Liu, B. & Chang, K.C.C., Editorial: Special Issue on Web Content Mining. SIGKDD Explorations special issue on Web Content Mining, Dec, 2004.

[27] Page, L., Brin, S., Motwani, R. & Winograd, T., The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies, January 1998.

[28] Oztekin, B.U., Ertoz, L. & Kumar, V., Usage Aware PageRank, World Wide Web Conference, 2003.

[29] Kleinberg, J.M., Authoritative Sources in Hyperlinked Environment. 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 668-667, 1998.

[30] Perkowitz, M. & Etzioni, O., Adaptive Web sites: an AI challenge. IJCAI97

[31] Pirolli, P., Pitkow,J.E., "Distribution of Surfer's Path Through the World Wide Web: Empirical Characterization." World Wide Web 1:1-17, 1999.

[32] Sarukkai, R.R., Link Prediction and Path Analysis using Markov Chains. Proc. of the 9th World Wide Web Conference, 1999.

[33] Zhu, J., Hong, J. & Hughes, J.G., Using Markov Chains for Link Prediction in Adaptive Web Sites. Proc. of ACM SIGWEB Hypertext, 2002.

[34] Mobasher, B., Web Usage Mining and Personalization. Practical Handbook of Internet Computing, ed. M.P. Singh, CRC Press, 2005.

[35] Padmanabhan, D., Desikan, P., Srivastava, J. & Riaz, K., WICER: A Weighted Inter-Cluster Edge Ranking for Clustered Graphs. Proc. of 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005).

[36] WebCT Mobasher, B., Cooley, R. & Srivastava, J., Automatic personalization based on web usage mining. Communications of the ACM, 43(8), pp. 142-151, 2000.

[37] Berendt, B., Hotho, A. & Stumme, G., Towards Semantic Web Mining. Proc of the International Semantic Web Conference, pp. 264-278, Sardinia, Italy. June 2002.

[38] Hodge, V. J. & Austin, J., Hierarchical Wordclustering - Automatic Thesaurus Generation, Neurocomputing, 48(1–4), pp. 819–846, 2002.

[39] Tang T.Y, McCalla G, Smart Recommendation for an Evloving E-learning System, Proc. 11th Int'l Conf. on Artificial Intelligence in Education AIED'2003), pp.699-710, 2003