Webology, Volume 5, Number 2, June, 2008

Information retrieval and machine learning: Supporting technologies for web mining research and practice

MPS Bhatia
Netaji Subhas Institute of Technology, University of Delhi, India. E-mail: mpsbatia (at) nsit.ac.in

Akshi Kumar Khalid
Netaji Subhas Institute of Technology, University of Delhi, India akshi.kumar (at) gmail.com

Received May 18, 2008; Accepted June 24, 2008

Abstract

With the enormous increase in recent years in the volume of information available on-line, and the consequent need for better techniques to access this information, there has been a strong resurgence of interest in Web Mining research. This paper expounds how research in Machine Learning (ML) and Information Retrieval (IR) will help develop applications that can drive the next generation of Web search with the key to support relevant search results by effectively and efficiently digging out user- centric information. This study attempts to probe the role of these two web mining supporting technologies (ML & IR). It reviews the web mining systems from the perspective of the Information Retrieval and illustrates how machine learning is likely to make substantial gains in web mining research and practice by developing standards and improving effectiveness.

Keywords

Web Information Retrieval, Machine Learning Paradigms, Web mining

Introduction

The World Wide Web is a huge, widely distributed, global source for information services, hyper-link information, access and usage information and web site content and organization. With the transformation of the Web into a ubiquitous tool for "e-activities" such as e-commerce, e-learning, e-government, e-science, its use has pervaded to the realms of day-to-day work, information retrieval and business management. Therefore, it is imperative to provide users with tools for efficient and effective resource and knowledge discovery. By all measures, the Web is enormous and growing at a staggering rate, which has made it increasingly intricate and crucial for both people and programs to have quick and accurate access to Web information and services. Search engines have played a key role in the World Wide Web's infrastructure as its scale and impact have escalated. Although search engines are important tools for knowledge discovery on the Web, they are far from perfect. The poor quality of retrieved results, handling a huge quantity of information, addressing subjective and time-varying search needs, finding fresh information and dealing with poor quality queries are commonly cited glitches.

The World Wide Web Impact: Opportunities and Challenges

The stakeholders could encounter, among others, the following problems when interacting with the Web (Kobayashi & Takeda, 2000):

a) The "Abundance" problem:
With the phenomenal growth of the Web, there is an ever increasing volume of data and information published in numerous web pages. According to worldwidewebsize.com, the indexed Web contains at least 27.56 billion pages (Sunday, 24 August, 2008).

b) Web Search results usually have low precision and recall:
For finding relevant information, the search service is generally a keyword-based, query-triggered process which results in problems of Low Precision (difficulty to find relevant information) and Low Recall (inability to index all information available on the web).

c) Limited query interface based on keyword-oriented search:
It is hard to extract useful knowledge out of information available because the search service used to find out specific information on the Web is retrieval-oriented, whereas to extract potentially useful knowledge out of it, is a data-mining oriented, data-triggered process.

d) Lack of Personalization of Information and Limited Customization to individual users:
Most knowledge on the Web is presented as natural-language text with occasional pictures and graphics. This is convenient for human users to read and view but difficult for computers to understand. It also limits the state of art search engines, since they cannot infer contextual meaning. For example the occurrence of word 'bat' refers to a bird or to a cricket bat. These factors uphold the inevitable creation of intelligent server and client-side systems that can effectively mine for knowledge both across the Internet and in particular web localities.

e) Heterogeneity:

- Information /data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.
- Much of the Web information is semi-structured due to the nested structure of HTML code.
- Much of the Web information is linked.
- The Web is noisy: A web page typically contains a mixture of many kinds of information, e.g., main contents, advertisement, navigational panels, copyright notices.

f) Dynamics:
The freedom for anyone to publish information on the Web at anytime and anywhere implies that information on the Web is constantly changing. It is a dynamic information environment whereas traditional systems are typically based on static document collection.

g) Duplication:
Several studies indicate that nearly 30% of the Web's content is duplicated, mainly due to mirroring.

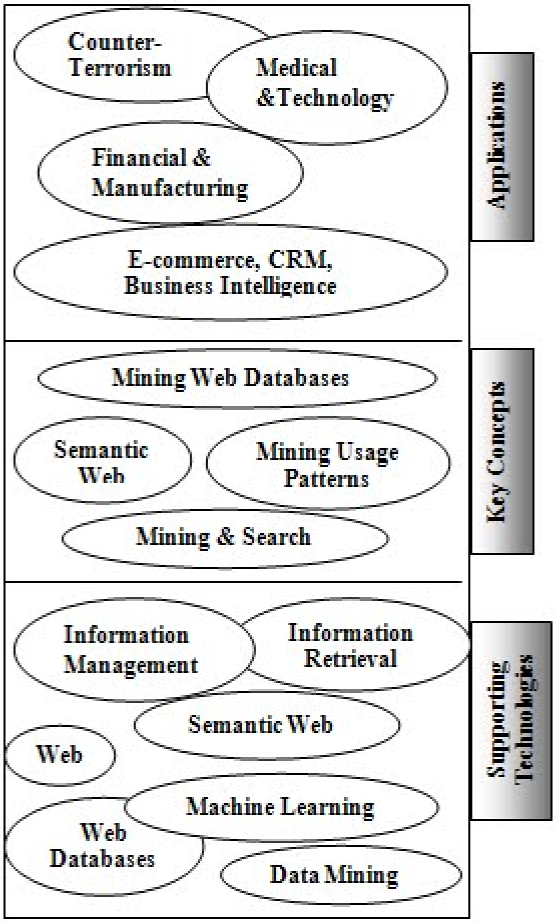
Despite its success as a preferred or de-facto source of information, the Web implicates a key challenge: How to target search & provide improved systems that retrieve the most relevant information available to satisfy user's need with accurate balance of novelty and pertinence.

Web Mining Layered Framework

Web Mining tries to solve these issues that arise due to the Web phenomenon. It tries to overcome these problems by applying data mining techniques to content, (hyperlink) structure and usage of web resources (Kosala & Blockeel, 2000; Bin & Zhijing, 2003; Chakarbarti, 2003; Henzinger, 2004; Srivastava, Desikan & Kumar, 2002). In the following sections of the paper, we expound the research in machine learning and information retrieval and help develop and review web mining applications that can more effectively and efficiently utilize the Web of knowledge.

A Three-layered Web Mining Framework is illustrated in Figure 1, where, the Layer 1 is the *Supporting Web Mining Technologies Layer* that expounds the various supporting technologies that contribute to Web Mining and the Layer 2 & Layer 3 are *Web Mining Concepts Layer & Web Mining Applications Layer* respectively. In this paper we attempt to talk about the use of *Information Retrieval & Machine Learning Supporting Technologies of Web Mining* that drive the next generation of Web search with the key to support relevant search results by effectively and efficiently digging out user- centric information.

Figure 1. Web mining framework



From IR to Web IR

Web Information Retrieval (Web IR) can be defined as the application of theories and methodologies from IR to the World Wide Web. It is concerned with addressing the technological challenges facing Information Retrieval (IR) in the setting of WWW (Nunes, 2006). The characteristics of Web make the task of retrieving information from it quite different from the Pre- Web (traditional) information retrieval. The Web is seemingly unlimited source of information with users from cross-section of society seeking to find information to satisfy their information need. They require the Web to be accessible through effective and efficient information retrieval systems that deliver information need fulfillments through the retrieval of web content. The paper (Bhatia & Khalid, 2008a) talks about the Web IR paradigm, as a variant of classical IR, by illustrating its basics, the components, model categories, tools, tasks and the performance measures that quantify the quality of retrieval results.

Diversity and complexity of Web information, as well as its richness, call for approaches that reach beyond conventional IR (Kuhlen, 1991; Baeza-Yates, 1999). For many years, information retrieval research focused mainly on the problem of ad-hoc document retrieval, a topical search task that assumes all queries are meant to express a broad request for information on a topic identified by the query. This task is exemplified by the early TREC conferences, where the ad-hoc document retrieval track was prominent. In recent years, particularly since the popular advent of the World Wide Web and e-commerce, IR researchers have begun to expand their efforts to understand the nature of the information need that users express in their queries. The unprecedented growth of available data coupled with the vast number of available online activities has introduced a new wrinkle to the problem of search: it is now important to attempt to determine not only what the user is looking for, but also the task they are trying to accomplish and the method by which they would prefer to accomplish it. In addition, all users are not created equal; different users may use different terms to describe similar information needs; the concept of "what is relevant" to a user has only become more and more unclear as the Web has matured and more diverse data have become available. Because of this, it is of key interest to search services to discover sets of identifying features that an information retrieval system can use to associate a specific user query with a broader information need. Specifically, the operative challenges motivating researchers in Web IR include problems relating either to data quality or user satisfaction (Kobayashi & Takeda, 2000). The problems facing successful Web Information Retrieval are a combination of challenges that stem from traditional information retrieval and challenges characterized by the nature of the World Wide Web.

The ultimate challenge of Web IR research is to provide improved systems that retrieve the most relevant information available on the Web to better satisfy a user's information need. In an information retrieval scenario, the most common evaluation is retrieval effectiveness and the effect of indexing exhaustivity and term specificity on retrieval effectiveness that can be explained by two widely accepted measures: Precision and Recall.

Precision: The proportion of retrieved and relevant documents (A) to all the documents retrieved:

$$|\mathbf{Ra}| / |\mathbf{A}|$$

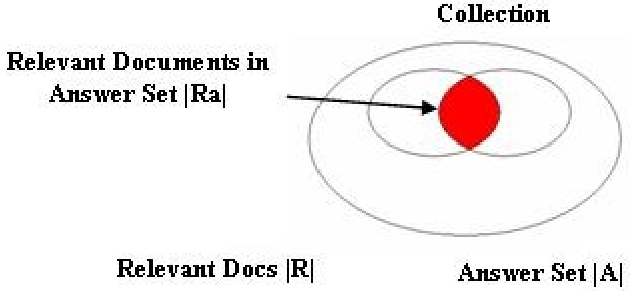
Recall: The fraction of the relevant documents (R) that is successfully retrieved:

$$|\mathbf{Ra}| / |\mathbf{R}|$$

A perfect Precision score of 1.0 means that every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved) and a perfect

Recall score of 1.0 means that all relevant documents were retrieved by the search (but says nothing about how many irrelevant documents were also retrieved) (Singhal, 2001).

Figure 2. Defining IR metrics



Existing search engines focus mainly on basic term-based techniques for general search, and do not attempt query understanding. Traditionally, a term in a given document is considered to be significant if it occurs multiple times within that document. This observation is commonly referred to as Term Frequency (*tf*) (Luhn, 1957). His study was based on the fact that the authors of documents typically emphasize a topic or concept by repeatedly using the same words. Since then, most information retrieval approaches (Kobayashi & Takeda, 2000; Baeza-Yates, 1999) have adopted *tf* (or variations of it) as a benchmark for indicating term significance or relevance within a given document. In particular, it is normally combined with inverse document frequency (*idf*) to form the *tf-idf* measure (Salton & Yang, 1973). Even with the emergence of Web Information Retrieval, *tf* still continues to be a standard measure of term significance within a document. There are several examples for content-based web information retrieval systems (Anh & Moffat, 2003; Craswell et al., 2003; Robertson & Walker, 1999; Yu et al., 2003) that assess term significance using *tf*. But as is the case for many potentially relevant documents, *tf* is not always the best or most useful indicator of term significance or relevancy. Quite often, there are relevant documents that contain only a single or a few occurrences of a particular term. Consequently, through *tf* these terms will rarely be considered significant, and thus never contribute impressively to the rank score of the potentially relevant document they appear within. This is especially the case when infrequently occurring terms appear in large documents containing hundreds or even thousands of terms.

In the query-centric approaches to retrieval (Plachouris et al., 2003; Plachouris & Ounis, 2002) queries can be classified to aid in the choice of retrieval strategy. Kang & Kim (2003) classify queries as either pertaining to topic relevance, homepage finding or service task and use this classification as a basis of dynamically combining multiple evidences in different ways to improve retrieval. Plachouris & Ounis (2002) use WordNet in a concept-based probabilistic approach to information retrieval where queries are biased according to their calculated scope. In their work, scope is an indication of generality or specificity of a query and is used as a factor of uncertainty in Dempster-Shafer's theory of evidence.

Context-based retrieval approaches aim to provide a more complete retrieval process by incorporating contextual information into the retrieval process. The use of context in information retrieval is not a new idea. Jing & Tzoukermann (1999) use context as a basis of measuring the semantic distances between words. Billhardt, Borrajo and Maojo (2002) propose a context-based vector space model for information retrieval. The WEBSOM (Honkela et al., 1997) system is an example of another way in which context has been used for information retrieval. It uses a two level Kohonen's self-organizing map approach to group words and documents of contextual similarity. Context in WEBSOM is limited to the terms that occur direct either sides of the term in question. IntelliZap (Finkelstein et al., 2001) is a context-based web search engine that requires the user to select a keyword in the context of some text. The approach makes effective use of the contextual information in the immediate vicinity of the keywords selected, so that retrieval precision can be improved.

Inquirus (Glover et al., 2001; Lawrence & Giles, 1998) is another web search engine that uses contextual information to improve search results. A user must specify some contextual information, considered as preferences, pertaining to the query. This context (preferences) provides a high-level description of the users information need and ultimately control the search strategy used by the system.

Kleinberg (1999) illustrates how hyperlink information in web pages can be used for web search when using a set of retrieved documents. An approach that also uses the characteristics of link information from a set of retrieved documents for topic distillation is presented by Amitay et al. (2002).

PageRank is hyperlink-based retrieval algorithm that calculates document scores by considering the entire hyperlink connected graph represented by all the links in the entire document collection (Brin & Page, 1998). The model described in (Zakos & Verma, 2006), uses traditional query expansion to determine context of query. Another closely related work on context-driven ranking on the Web (Jonathan, 2006), implicitly deduce context using three different algorithms. Finally (Pickens & Farlance, 2006), offers Term context model as a new tool for accessing term presence in a document. Bhatia & Khalid (2007; 2008b) offer Web retrieval system architecture with a novel re-ranking algorithm to effectively refine the ranking of web search results for locating the most useful documents.

An Overview of Machine Learning

Since 1940's, many knowledge-based systems have been built that acquire knowledge manually from human experts, which is very time-consuming and labor-intensive. To address this problem, Machine Learning algorithms have been developed to acquire knowledge automatically from examples or source data. It is defined as "any process by which a system improves its performance" (Simon, 1983).

Machine Learning Paradigms

Following are the five major Machine Learning paradigms (see Table 1).

Table 1. Various Machine Learning Paradigms

Probabilistic Models	Symbolic Learning and Rule Induction	Neural Networks	Analytic Learning and Fuzzy Logic	Evolution-based models
The use of probabilistic models was one of the earliest attempts to perform machine learning. The most popular example: Bayesian method, Originating in	Symbolic learning can be classified as rote learning, learning by being told, learning by analogy, learning from examples, and learning from discovery (Cohen & Feigenbaum,	A Neural Network is a graph of many active nodes (neurons) that are connected with each other by weighted links (synapses). Knowledge is learned and remembered by a network of	Evolution-based algorithms rely on analogies to natural processes and Darwinian survival of the fittest. There are three categories of evolution-based algorithms: genetic	The Analytic Learning represents knowledge as logical rules, and performs reasoning on such rules to search for proofs which can be compiled into more complex rules to solve

pattern recognition research (Duda & Hart, 1973) A Bayesian model stores the probability of each class, the probability of each feature, each feature given each class, based on the training data, to classify new instances according to these probabilities (Langley et al., 1992) A variation of the Bayesian model, NaÃ¯ve Bayesian model has been widely used in various applications in different domains (Fisher, 1987; Kononenko, 1993).	1982; Carbonell et al., 1983). Learning from examples is implemented by applying an algorithm that attempts to induce a general concept description that best describes the different classes of the training examples. ID3 decision-tree building algorithm (Quinlan, 1983), and its variations such as C4.5 (Quinlan, 1993)	interconnected neurons, weighted synapses, and threshold logic units (Rumelhart et al., 1986a; 1986b; Lippmann, 1987). Based on training examples, learning algorithms can be used to adjust the connection weights in the network so that it can predict or classify unknown examples correctly (Belew, 1989; Kwok, 1989; Chen & Ng, 1995).	algorithms, evolution strategies, and evolutionary programming. Among these, Genetic Algorithms are most popular and have been successfully applied to various optimization problems. They were developed based on the principle of genetics (Goldberg, 1989; Michalewicz, 1992).	similar problems. Fuzzy systems and logic have been applied for imprecision and approximate reasoning by allowing the values of False or True to operate over the range of real numbers from 0 to 1 (Zadeh, 1965).

Machine Learning Algorithms

Machine learning algorithms are organized into a taxonomy, based on the desired outcome of the algorithm. In general, Machine learning algorithms can be classified as:

- **Supervised learning**: Training examples contain input/output pair patterns. Learn how to predict the output values of new examples.
- **Unsupervised learning**: Training examples contain only the input patterns and no explicit target output. The learning algorithm needs to generalize from the input patterns to discover the output values.
- **Reinforcement learning**: in which the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- **Transduction**: similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and test inputs which are available while training.
- **Learning to learn**: in which the algorithm learns its own inductive bias based on previous experience.

Machine Learning for Information Retrieval Â¬

Learning techniques had been applied in Information Retrieval (IR) applications long before the recent advances of the Web. In this section, we will briefly survey some of the research in this area, covering the use of machine learning in:

- Information extraction
- Relevance feedback
- Information filtering
- Text classification and Text clustering.

Information Extraction

Information extraction (IE) refers to the techniques designed to identify useful information from text documents automatically. It has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed.

Table 2. Differences between IR and IE

Information Retrieval (IR)	Information Extraction (IE)
Aims to select relevant documents, i.e., it finds documents.	Aims to extract relevant facts from the document, i.e., it extracts information.
Views text as a bag of unordered words.	Interested in structure and representation of the document.

Thus IE works at a finer granularity level than IR does on documents and makes information retrieval more precise. Applications of IE include the summarization of documents in well defined subject areas and automatic generation of databases from text. IE includes the following subtasks:

Named entity recognition (NER)
It refers to recognizing relevant entities in text. It is the automatic identification from text documents of the names of entities of interest.

Relation extraction
Linking recognized entities having particular relevant relations.
The Machine learning-based entity extraction systems rely on algorithms based on Neural networks, Decision tree (Baluja et al., 1999), Hidden Markov Model (Miller et al., 1998), Entropy maximization (Borthwick et al., 1998), rather than human-created rules to extract knowledge or identify patterns from texts.

Relevance Feedback

Relevance Feedback helps users conduct searches iteratively and reformulate search queries based on evaluation of previously retrieved documents. Using relevance feedback, a model can learn the common characteristics of a set of relevant documents in order to estimate the probability of relevance for the remaining documents (Fuhr & Buckley, 1991). Various Machine Learning algorithms, such as genetic algorithms, ID3, and simulated annealing, have been used in relevance feedback applications (Kraft et al., 1995; 1997; Chen et al., 1998).

Information filtering

Information filtering techniques try to learn about users' interests based on their evaluations and actions, and then to use this information to analyze new documents. Many personalization and collaborative systems have been implemented as software agents to help users in information systems (Maes, 1994).

Text classification and Clustering

Text classification is the classification of textual documents into predefined categories (supervised learning). For example, Support Vector Machine (SVM), a statistical method that tries to find a hyper plane that best separates two classes. Text clustering groups documents into non-predefined categories which dynamically defined based on their similarities (unsupervised learning). Kohonen's Self-Organizing Map (SOM), a type of neural network that produces a 2-dimensional grid representation for n-dimensional features, has been widely applied in IR (Lin et al., 1991; Kohonen, 1995). Machine learning is the basis of most text classification and clustering applications.

Web Mining

Web mining refers to the use of data mining techniques to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from web documents and services. It is about making implicit or "hidden" knowledge into explicit. The digital revolution and the phenomenal growth of the Web have lead to the generation and storage of huge amounts of data, prompting the need for intelligent analysis methodologies to discover useful knowledge from it. Due to the heterogeneous, semi-structured, distributed, time-varying and multi-dimensional facets of web data, automated discovery of targeted knowledge is a challenging task. It calls for novel methods that draw from a wide range of patent areas of data mining, machine learning, information retrieval, natural language processing, Multimedia, and Statistics. In this article, we will provide a review of the field from the perspectives of machine learning and information retrieval and how they have been applied in web mining systems. Machine learning is the basis for most data mining and text mining techniques and information retrieval research has largely influenced the research directions of web mining applications.

From Data Mining to Web Mining

Two notable and active areas of current research are *data mining* and the *World Wide Web*. An expected alliance of the two areas, sometimes referred to as web mining, has been the focus of several recent research projects and papers. Nevertheless, web mining has many unique characteristics compared with data mining. First, the source of web mining is web documents. We consider the use of the Web as a middleware in mining database and the mining of logs and user profiles on the Web server still belong to the category of traditional data mining. Second, the Web is a directed-graph consists of document nodes and hyperlinks. Therefore, the pattern identified can be possibly about the content of documents or about the structure of the Web. Moreover, the Web documents are semi-structured or non-structured with little machine-readable semantic while the source of data mining is confined to the structural data in database. As a result, some traditional data mining methods are not applicable to web mining (Mahanta, 2008).

The distinct axes in the Web Mining Research

The diversity of information on the Web leads to the variety of web mining, defined in the following three categories (Kosala & Blockeel, 2000; Bin & Zhijing, 2003; Chakarbarti, 2003; Henzinger, 2004; Srivastava, Desikan & Kumar, 2002):

1. Web Usage Mining

- DEFINITION: Web Usage Mining refers to discovering user access patterns from Web usage logs. It focuses on using data mining techniques to analyze search logs to find interesting patterns.
- EXAMPLE: Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, any other interaction data.
- APPLICATIONS: Learning user profiles, Identifying associate terms, and web traffic control, etc.

2. Web Structure Mining

- DEFINITION: Web Structure Mining refers to inferring useful knowledge from the structure of hyperlinks (in-links and out-links). It studies the model underlying the link structures of the Web.
- APPLICATIONS: Web page ranking in Google, etc.

Web structure mining can be further divided into external structure (hyperlinks between web pages) mining, internal structure (of a web page) mining and URL mining.

3. Web Content Mining

- DEFINITION: Web Content Mining refers to extracting use information and knowledge from content in web pages. It is the discovery of useful information from web contents, including text, images, audio, video, etc. It can be divided into text mining (including text file, HTML document, etc.) and multimedia mining.
- APPLICATIONS: Document categorization, Sentiment classification, etc.

The Road Ahead: Open Problems and Solutions

Web Mining Open problems

Nevertheless, web content is not always easy to use. Due to the unstructured and semi-structured nature of web pages and design idiosyncrasy of web sites, it is a challenging task to organize and manage content from the Web. Thus, web Mining continues to play an ever expanding and inevitable role. There are several major challenges for web mining research:

- First, most of web documents are in HTML format and contain many markup tags, mainly used for formatting.
- Second, while traditional IR systems often contain structured and well-written documents, this is NOT the case on the Web.
- Third, while most documents in traditional IR systems tend to remain static over time, web pages are much more dynamic.
- Fourth, web pages are hyperlinked to each other, and it is through hyperlink that the author of a web page cites other web pages.
- Last, the size of the Web is larger than traditional data sources or document collections by several orders of magnitude.

Machine Learning as a Solution

Machine Learning is likely to make substantial gains in web mining research and practice. We exploit few perspectives:

Web Content Mining. Web content mining is mainly based on research in information retrieval and text mining, such as information extraction, text classification and clustering, and information visualization. However, it also includes some new applications, such as knowledge discovery on the Web. Some important web content mining techniques and applications are reviewed in following subsections:

1. Text Mining for Web Documents

Text mining for Web documents can be considered a sub-field of web content mining. Information extraction techniques have been applied to web HTML documents. For example, Chang & Lui (2001) used a PAT tree to construct automatically a set of rules for information extraction. Text clustering algorithms also have been applied to web applications. For example, Chen et al. (2001; 2002) used a combination of noun phrasing and SOM to cluster the search results of search agents that collect web pages by meta-searching popular search engines.

2. Intelligent Web Spiders

Web Spiders, have been defined as "software programs that traverse the Web by following hypertext links and retrieving web documents by HTTP protocol" (Cheong, 1996). They can be used to build the databases of search engines (e.g., Pinkerton, 1994), perform personal search (e.g., Chau et al., 2001), archive web sites or even the whole Web (e.g., Kahle, 1997) and collect web statistics (e.g., Broder et al., 2000). Intelligent Web Spiders: some spiders that use more advanced algorithms during the search process have been developed. For example, the Itsy Bitsy Spider searches the Web using a best-first search and a genetic algorithm approach (Chen et al., 1998).

3. Multilingual Web Mining

In order to extract non-English knowledge from the Web, web mining systems have to deal with issues in language-specific text processing. The base algorithms behind most machine learning systems are language-independent. Most algorithms, e.g., *text classification and clustering*, need only to take a set of features (a vector of keywords) for the learning process. However, the algorithms usually depend on some phrase segmentation and extraction programs to generate a set of features or keywords to represent web documents. Other learning algorithms such as *information extraction* and *entity extraction* also have to be tailored for different languages.

4. Web Visualization

Web visualization tools have been used to help users maintain a "big picture" of the retrieval results from search engines, web sites, a subset of the Web, or even the entire Web. The most well known example of using the tree-metaphor for web browsing is the hyperbolic tree developed by Xerox PARC (Lamping & Rao, 1996). In these visualization systems, machine learning techniques are often used to determine how web pages should be placed in the 2-D or 3-D space. One example is the SOM algorithm described earlier (Chen et al., 1996).

5. The Semantic Web

Semantic web technology (Berners-Lee et al., 2001) tries to add metadata to describe data and information on the Web. Machine learning can play three roles in the Semantic Web:

- First, machine learning can be used to automatically create the markup or metadata for existing unstructured textual documents on the Web.
- Second, machine learning techniques can be used to create, merge, update, and maintain ontologies.
- Third, machine learning can understand and perform reasoning on the metadata provided by the Semantic Web in order to extract knowledge from the Web more effectively.

Web Structure Mining. Web link structure has been widely used to infer important web pages information. Web structure mining has been largely influenced by research in:

- Social network analysis,
- Citation analysis,
- Web usage data, web mining systems can discover useful knowledge about a system's usage characteristics and the users' interests which has various applications, web site design and evaluation,
- Personalization and collaboration in web-based systems,
- Marketing and decision support (e.g., Chen & Cooper, 2001; Marchionini, 2002).

Conclusion and Future Directions

Extracting user-centric information from the world's largest repository - the Web efficiently and effectively is becoming increasingly imperative. This article reviews how machine learning techniques can be applied to web mining. Major limitations of web mining research are lack of suitable test collections that can be reused by researchers and difficulty to collect web usage data across different web sites. Most web mining applications have been assessed in this paper. Although the activities are still in their early stages and should continue to develop as the Web evolves. Future research directions include:

- Multimedia data mining: a picture is worth a thousand words;
- Multilingual knowledge extraction: web page translations;
- Wireless Web: WML and HDML;
- The Hidden Web: forms, dynamically generated web pages; and
- Semantic Web

References

- Amitay, E., Carmel, D., Darlow, A., Lempel, R., & Soffer, A. (2002). Topic distillation with knowledge agents. *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002)*, Gaithersburg, MD, 2002.), Gaithersburg, Maryland, USA, 500; part 251, 263-272.
- Anh, V., & Moffat, A. (2003). Robust and web retrieval document-centric integral impacts. *Proceedings of the 12th Text Retrieval Conference (TREC-12)* , Gaithersburg, USA, 726-731.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison Wesley, New York.
- Baluja, S., Mittal, V., & Sukthankar, R. (1999) Applying machine learning for high performance named-entity extraction. *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, 1999, 365-378.
- Belew, R.K. (1989). Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva*l, 11-20.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web.*Scientific American*, 284(5), 35-43.
- Bhatia, MPS, & Khalid, A.K. (2007), Contextual proximity based term-weighting for improved web information retrieval. *Proceedings of KSEM 2007, Lecture notes of AI-4798*, Springer, 267-278.
- Bhatia, MPS, & Khalid, A.K. (2008a), A primer on the Web information retrieval paradigm. *Journal of Theoretical and Applied Information Technology*, 4(7), 657-662.

- Bhatia, MPS, & Khalid, A.K. (2008b). The context-driven generation of web search. *Proceedings of CISTM 2008*, 281-287.
- Billhardt, H., Borrajo, D., & Maojo, V. (2002). A context vector model for information retrieval. *Journal of American Society on Information Science Technology*, 53(3), 236-249.
- Bin, W., & Zhijing, L. (2003). Web mining research. *Proceedings of 5th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03)* .
- Borthwick, A., Sterling, J., Agichtein, E., & Grishman, R. (1998). NYU: Description of the MENE named entity system as used in MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC- 7)* .
- Brin, S., & Page, L. (1998).The anatomy of a large-scale hyper-textual web search engine. *Proceedings of the 7th WWW Conference*, 107-117, Brisbane, Australia.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the Web. *Proceedings of the 9th International World Wide Web Conference*.
- Carbonell, J.G., Michalski, R.S., & Mitchell, T.M. (1983). An overview of machine learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 3-23). Pa10 Alto,CA Tioga.
- Chakarbarti S. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann Publisher, San Francisco, CA.
- Chang, C.H., & Lui, S.C. (2001). IEPAD: Information extraction based on pattern discovery. *Proceedings of the 10th World Wide Web Conference*.
- Chau, M., Zeng, D., & Chen, H. (2001). Personalized spiders for Web search and analysis. *Proceedings of the 1st ACM-IEEE Joint Conference on Digital Libraries*, 79-87.
- Chen, H. (2001). *Knowledge management systems: A text mining perspective*. Tucson, AZ: University of Arizona., http://ai.bpa.arizona.edu
- Chen, H., Chau, M., & Zeng, D. (2002). CI spider: A tool for competitive intelligence on the Web. *Decision Support Systems*, 34(l), 1-17.
- Chen, H.M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 52, 888-904.
- Chen, H., & Ng, T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic brand and bound search vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science*, 46, 348-369.
- Chen, H., Schuffels, C., & Orwig, R. (1996). Internet categorization and search: A machine learning approach. *Journal of Visual Communication and Image Representation*, 7(1), 88-102.
- Chen, H., Shankaranarayanan, G., Iyer, A., & She, L. (1998). A machine learning approach to inductive query by examples: An experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing. *Journal of the American Society for Information Science*, 49, 693-705.
- Cheong, F.C. (1996). *Internet agents: Spiders, wanderers, brokers, and bots*. Indianapolis, IN: New Riders Publishing.
- Cohen, P.R., & Feigenbaum, E.A. (1982). *The handbook of artificial intelligence* (Vol. 3). Reading, MA: Addison-Wesley.
- Craswell, N., Hawking, D., Upstill, T., McLean, A., Wilkinson, R., & Wu, M. (2003). TREC 12 Web and interactive tracks at CSIRO. *Proceedings of the 12th Text Retrieval Conference (TREC-12)*, Gaithersburg, USA, 193-203.
- Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: Wiley.

- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *Proceedings of the 10th International World Wide Web Conference*, 406-414.
- Fisher, D.H.(1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2, 139-172.
- Fuhr, N., & Buckley, C. (1991). A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9, 223-248.
- Glover, E., Lawrence, S., Gordon, M., Birmingham, W., & Lee Giles, C. (2001). Web search â€" your way. *Communication ACM*, 44(12), 97-102.
- Goldberg, D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.
- Henzinger, M. (2004). The Past, Present, and Future of Web Search Engines. *Proceedings of 31st International Colloquium, ICALP 2004*, Finland.
- Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1997). WEBSOM â€" self-organizing maps of document collections. *Proceedings of WSOM_97 (Workshop on Self-Organizing Maps)*, Espoo, Finland, 310-315.
- Jing, H., & Tzoukermann, E. (1999). Information retrieval based on context distance and morphology. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 90-96.
- Jonathan, S. (2006). Context driven ranking for the Web. http://infolab.stanford.edu/~jonsid/
- Kahle, B. (1997). Preserving the Internet. *Scientific American*, 276(6), 82-83.
- Kang, I., & Kim, G. (2003). Query type classification for web document retrieval. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 64-71.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment, *Journal of the ACM*, 46(5), 604-632.
- Kobayashi, M. & Takeda, K. (2000). Information retrieval on the Web. *ACM Computing Surveys*, 32 (2).
- Kohonen, T. (1995). *Self-organizing maps*. Berlin, Germany: Springer-Verlag.
- Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7, 317-337.
- Kosala, R. & Blockeel, H. (2000). Web mining research: A survey, *SIGKDD Explorations*, 2(1), 1-15.
- Kraft, D.H., Petry, F.E., Buckles, B.P., & Sadasivan, T. (1995). Applying genetic algorithms to information retrieval systems via relevance feedback. In: Bosc, P. & Kacprzyk, J. (Eds.), *Fuzziness in database management systems* (pp. 330-344). Heidelberg, Germany: Physica-Verlag.
- Kraft, D.H., Petry, F.E., Buckles, B.P., & Sadasivan, T. (1997). Genetic algorithms for query optimization in information retrieval: Relevance feedback. In: Sanchez, E., Shibata, T., & Zadeh, L.A. (Eds.), *Genetic algorithms and fuzzy logic systems* (pp. 155-173). Singapore: World Scientific.
- Kuhlen, R. (1991). Information and pragmatic value-adding: Language games and information science. *Computers and the Humanities*, 25, 93-101.
- Kwok, K.L. (1989). A neural network for probabilistic information retrieval. *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21-30.
- Lamping, J., & Rao, R. (1996). Visualizing large trees using the hyperbolic browser. *Proceedings of the ACM CHI '96 Conference on Human Factors in Computing Systems*, 388-389.
- Langley, F., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the 10th National Conference on Artificial Intelligence*, 223-228.
- Lawrence, S., & Giles, C. (1999). Context & page analysis for improved web search. *IEEE Internet Computing*, 2(4), 38-46.

- Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 262-269.
- Lippmann, R.P. (1987). An introduction to computing with neural networks. *IEEE Acoustics Speech and Signal Processing Magazine*, 4, 4-22.
- Luhn, H. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research. Development*, 1(4), 309-317.
- Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37(7), 3140.
- Mahanta, A.N. (2008). Web mining: Application of data mining. *Proceedings of NCKM-2008*, pp. 111-116. http://annamalaiuniversity.ac.in/download/conference/NCKM-2008/paper19.pdf
- Marchionini, G. (2002). Co-evolution of user and organizational interfaces: A longitudinal case study of WWW dissemination of national statistics. *Journal of the American Society for Information Science and Technology*, 53, 1192-1209.
- Michalewicz, Z. (1992). *Genetic algorithms + data structures = evolution programs*. Berlin, Germany: Springer-Verlag.
- Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., & the Annotation Group BBN. (1998). Description of the SIFT system as used for MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.
- Nunes, S. (2006). *State of the art in web information retrieval*. Technical Report, FEUP.
- Pickens, J., & Farlance, A.M. (2006). Term context models for information retrieval, *ACM CIKM'06*.
- Pinkerton, B. (1994). Finding what people want: Experiences with the Web Crawler. *Proceedings of the 2nd International World Wide Web Conference*.
- Plachouris, V., Cacheda, F., Ounis, Iadh, & van Rijsbergen, C. (2003). University of Glasgow at the Web track: Dynamic application of hyperlink analysis using the Query Scope. *Proceedings of the 12th Text Retrieval Conference (TREC-12)* , Gaithersburg, USA, 636-642.
- Plachouris, V., Ounis, I. (2002). Query-biased combination of evidence on the Web. *Workshop on Mathematical/Formal Methods in Information Retrieval, ACM SIGIR Conference*, 105-121.
- Quinlan, J.R. (1983). Learning efficient classification procedures and their application to chess end games. In: Michalski, R.S., Carbonell, J.G. & Mitchell, T. M. (Eds.), *Machine learning: An artificial intelligence approach* (pp. 463-482). Palo Alto, CA: Tioga.
- Quinlan, J.R. (1993). *C4.5: Programs for machine learning*. Los Altos, CA: Morgan Kaufmann.
- Robertson, S. (1991). On term selection for query expansion. *Journal of Documentation*, 46 (4), 359 - 364.
- Robertson, S., & Walker, S. (1999). Okapi/ Keenbow at TREC-8. *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, Gaithersburg, USA, 151-161.
- Rumelhart, D.E., Hinton, G.E., & McClelland, J. L. (1986a). A general framework for parallel distributed processing, Rumelhart, D.E., McClelland, J.L. & the PDP Research Group (Eds.), *Parallel distributed processing* (pp. 45-76). Cambridge, MA: The MIT Press.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986b). Learning internal representations by error propagation. Rumelhart, D.E., McClelland, J.L. & the PDP Research Group (Eds.), *Parallel distributed processing* (pp. 31S362). Cambridge, MA: MIT Press.
- Salton, G., & Yang, C. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4), 351-372.

- Simon, H.A. (1983). Why Should Machine Learn? In: Michalski, R.S., Carbonell, J., & Mitchell, T.M. (Eds.), *Machine learning: An artificial intelligence approach* (pp. 25-38). Palo Alto, CA Tioga Press.
- Srivastava, J., Desikan, P., & Kumar, V. (2002). Web Mining- Accomplishments and Future directions. *Proceedings of National Science Foundation Workshop on Next Generation Data Mining (NGDM'02)*, Baltimore, Maryland.
- Yu, S., Cai, D., Wen, J., & Ma, W. (2003). Improving pseudo-relevance feedback in web information retrieval using web page segmentation. *Proceedings of the 12th International Word Wide Web Conference*.
- Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8, 338-353.
- Zakos, J., & Verma, B. (2006). A novel context matching based technique for web document retrieval. *World Wide Web*, Springer, 9 (4), 485-503.

Bibliographic information of this paper for citing:

Bhatia, MPS and Khalid, Akshi Kumar (2008). "Information retrieval and machine learning: Supporting technologies for web mining research and practice." *Webology*, **5**(2), Article 55. Available at: http://www.webology.org/2008/v5n2/a55.html

Alert us when: New articles cite this article