A literature review of web mining

Isaac Kamga
Department of Computer Science
University of Buea
Cameroon.

January 10 2011

Abstract

When Sir Tim Berners-Lee and Robert Cailliau wove the World Wide Web (WWW), it was ordained to be a pool of human knowledge which will allow collaborators in remote sites to share their ideas and all aspects of a common project(Wikipedia, 2010). This gigantic information resource is non-administered, resides on the Internet and is ever growing with over 11.5 billion web pages (as of January 2005) and over 109.5 million web sites operated (as of May 2009). Mankind rushed to the World Wide Web (W3) to quench its thirst for knowledge, albeit, met challenges such as its size, talk less of the complexity and informality of its data. These could be resolved by techniques pertaining to data mining. This paper provides a survey of the existing work done by scholars in the area of web mining. It also introduces the notion of data mining, the difficulties encountered when mining web data and summarizes the existing efforts made in this field.

Key words: **World Wide Web, Internet, data mining, web mining.**

Contents

1   Introduction

1.1   Background

Humanity's search for information has caused it to always extract interesting patterns from available data. Traditional statistical approaches to these derivations such as Bayes' Theorem (used in the 1700s) and regression analysis (used in the 1800s) are inadequate for the information age in which we live. Hence, automating this process of knowledge discovery is the only road ahead.

Han and Kamber (2006) defined data mining (or Knowledge Discovery in Databases, KDD) as extracting knowledge from large amounts of data. It is an interdisciplinary area of research which involves fields like databases, data warehouses, machine learning, information retrieval, statistics and probability. It involves tasks such as characterization, association, classification, clustering, outlier analysis and evolution analysis. Dealing with small amounts of data is generally referred to as data dredging. Data dredging (fishing) is the use of data mining techniques on sample sizes that are too small for statistical inferences to be made about the validity of any patterns discovered (Wikipedia,2010). Large amounts of data reside in large information repositories like databases, data marts, the World Wide Web.

In this review paper, focus is centered on data on the World Wide Web which is a collection of interconnected documents and other resources linked together by hyperlinks and Uniform Resource Locators (URLs) (Wikipedia, 2010). Web mining includes the discovery and analysis of data, documents and multimedia from the World Wide Web (Anthony Scime, 2004). It is essentially divided into three classes, which are, web structure mining, web content mining and web usage mining.

1.2   Research statement

The challenges faced by web mining have prompted research into scalable, effective and efficient methods of knowledge discovery on the World Wide Web. Albert Einstein said "the significant problems that we face today, cannot be solved at the same level of reasoning we were at when we created them".

1.3   Importance of study

In recent years, Web mining has attracted a great deal of attention in the information industry because of the following reasons;

1) The invention of the internet and the explosion in the size, complexity and popularity of web data.

2) The proliferation, ubiquity and increasing potency of information and communication technologies (ICTs) has increased data collection and storage.

3) The need for users to utilize automated tools to get desired information and break free from the data rich but information poor phenomenon.

1.4   Paper organisation

This paper is divided into 4 sections. Section 2 deals with the challenges encountered with searching web data. Section 3 gives a survey of the existing efforts in Web mining and Section 4 has the conclusion.

2   The challenges

2.1   Size of the WWW

The size of the WWW is in the order of hundreds of terabytes (Han and Kamber, 2006) and continues to grow at roughly a million pages per day (Chakrabarti, Kleinberg, Dom, Kumar and Raghavan, 1999). A good idea would be to build a data warehouse for this data and mine it. However, as of May 2009, over 109.5 million websites were operated (Wikipedia, 2010). Needless to say, it is impossible to replicate, store and integrate all this data.

2.2   The dynamism(flux) of the WWW

According to Sir Tim Berners-Lee, " The power of the web is in the Universality. Access by everyone regardless of disability is an essential aspect ". The accessibility given to users is not necessarily as read-only but for many users to create and update websites. Consequently, linkage information and access records frequently change.

2.3   Broad diversity of user communities

The internet currently connects about 50 million workstations and its user's community is still expanding rapidly (Han and Kamber, 2006). Users have different backgrounds, interests and usage purposes. Many users do not have any knowledge of the structure of the information network and may easily get bored waiting impatiently for a piece of information.

2.4   Complexity of web pages

Although web pages may appear colourful and fancy, they are very complicated. Taken as a whole the set of web pages lacks a unifying structure and shows far more authoring style and content variation than seen in traditional text-document collections (Chakrabarti *et al*, 1999). Although the WWW is considered a huge digital library, there is no index by category or by title, author, cover page, table of contents and so on. It can be very difficult to search desired information in such a library.

2.5   Relevance of web pages

When querying search engines, a topic of any breadth may easily return hundreds of thousands of web documents. Users may need only a few of these. It is said that 99% of the web information is useless to 99% of Web users. Thus a user is interested in a ten-thousandth portion of the WWW. The lack of proper topic distillation techniques in index-based search engines like altavista and Yahoo! ; and the usual problems of text search (synonymy, polysemy and context sensitivity) become very obvious.

2.6   Definitions

The work done by scholars to solve these problems has segregated web mining into 3 phases;

Web content mining: Application of data mining techniques to unstructured or semi-structured data, usually Hypertext Markup Language (HTML) documents (Fürnkranz,2007).

Web structure mining: Use of the hyperlink structure of the Web as an (additional) information source (Fürnkranz,2007).

Web usage mining: Application of data mining techniques to discover usage patterns from web data, in order to understand and better serve the needs of web based applications (Srivastava, Cooley, Deshpande and Tan, 2000).

3 Literature survey of existing efforts in web mining

3.1 Web Structure mining (WSM)

A key feature which distinguishes hypertext mining from Warehouse mining and much of information retrieval is the analysis of the social organization on the WWW. Social network theory is concerned with properties related to connectivity and distances in graphs with applications in espionage, citation indexing, epidemiology etc. [27] Social networks exist between academics by co-authoring, advising, serving on committees; between people making phone calls and transmitting infections and between web pages by hyper linking to other web pages. The web is an example of a social network where each web page is a node and a hyperlink on the page is an edge between two nodes. Applying social network analysis to this huge graph radically changed information retrieval and categorization of web pages.

3.1.1 Information retrieval

While information retrieval is concerned with representing content in a form that can easily be accessed by its users with information needs, web structure mining goes beyond that to exploit the social organization that exists on the web. A major breakthrough in web structure mining occurred when scholars wanted search engines to go beyond retrieving relevant web pages to satisfying user queries with authoritative web pages. Authoritative here means to locate web pages which provide the best source of information on a given topic (Chakrabarti *et al*, 1999). Thus for any given broad search topic, it requires to locate not only a set of relevant pages but also those relevant pages of the highest quality.

3.1.1.1. Hypertext induced topic search (HITS) algorithm.

Kleinberg (1999) claimed that the creation of hyperlinks on the web is for authority conferral, thus by mining the collective judgment contained in the set of such endorsements, a rich understanding of relevance and popularity of web pages can be gained. The HITS algorithm does not preprocess the web but depends on another search engine which retrieves relevant pages that accompany a broad search query while HITS aims at chooses the most authoritative of these pages that a user needs. Thus a query to HITS is forwarded to a supporting search engine such as Alta Vista which retrieves a sub graph (**root set**) of the web whose pages match the query. For each page p in the sub graph, we add all pages it points to and all pages that point to it. This augments the root set to a much larger set of pages called the **base set**. HITS iteratively assigns two measures to all the pages in the base set namely the authority weight Ap and the hub weight Hp.

$$A_p \sum_{q \to p} A_q \qquad (1)$$

$$H_q \sum_{q \to p} H_q \qquad (2)$$

and normalizes both of these sums to 1. The A and H scores converge respectively to the measure of a page being an authority and the measure of a page being a hub. While HITS is enlightening, it has its own flaws namely; contamination and topic hijacking. Although relying extensively on hyperlinks can lead to encouraging results, HITS may encounter some difficulties by ignoring textual contexts for example HITS sometimes drifts when hubs contain multiple topics, it may cause 'topic hijacking' where many pages from a single website point to the same single popular site giving the site too large a share of the authority weight. Such problems can be overcome by replacing the above sums with weighted sums hence scaling down the weights of multiple links from within the same site (Bharat & Henzinger, 1998). The main drawback of HITS Algorithm is that the hubs and authority weights must

be computed iteratively from the query result which does not meet the real time constraints of an online search engine. Systems based on HITS include Clever.

The implementation of a similar idea in the Google search engine resulted in a major breakthrough which made it a major player in search engine technology.

3.1.1.2 The Page Rank Algorithm.

Brin and Page (1998) suggested the use of the probability that a page is visited by a random web surfer on the web as a key factor for ranking search results. If a surfer wonders on the net for infinite time, following a random link out of a page with probability 1- p and jumps to a random web page with probability p, then different pages would be visited at different rates. Popular pages with many in-links would tend to be visited more often. They approximated this probability with the so-called Page Rank formula which is again computed iteratively:

$$PageRank(v) = \frac{p}{N} + \sum_{u \to v} \frac{PageRank(u)}{Out - degree(u)} \tag{3}$$

where '**->**' this means 'links to' and **N** is the total number of nodes in the web graph.

The first term of this sum models the behavior that a surfer gets bored with probability p where p is typically set to 0.15 and jumps to a randomly selected page in the entire set of N pages. The second term uniformly distributes the current Page Rank of a page to all its successor pages. Thus a page receives a high page rank if it is linked to many pages which in turn have a high page rank and/or only few successor pages.

The main advantage of the Page Rank over the HITS Algorithm is that it can be computed off-line that is it can be pre-computed for all pages in the index of a search engine. Thus Google can be potentially as fast as any relevant ranking search engine.

The main draw back of the Page Rank algorithm is that it discriminates against new websites and favors already established sites.

3.1.1.3 Contamination.

Adding text information to link information, HITS graph expansion from the root set to the base set some times lead to topic contamination or drift. This is partly because in the HITS algorithm all edges in the graph have the same importance. Contamination can be reduced by recognizing that hyperlinks that contain keywords near the anchor text are more relevant for this query than other hyperlinks. Systems such as Clever incorporate such query-dependent modifications of edge weights (Chakrabarti *et al*, 1999).

Bharat, Broder, Henzinger, Kumar, and Venkatasubramanian (1998) invented another method to integrate textual content and hence reduce contamination of the base set. They modeled each page according to a "vector space" model. During the graph expansion step , instead of including all nodes at distance 1 from the preliminary query result, they pruned the graph expansion at nodes whose corresponding term vectors are outliers with respect to the set of vectors retrieved from the search engine. They described several heuristics which cut down the query time severely by using the connectivity server.

3.1.2   Automatic classification.

In data mining terms, classification may be defined as finding models that aggregate or segregate data to predict the class of objects whose class label is unknown. With respect to web mining, each web page is assigned to a class label from a set of predefined topic categories. Various scientists have addressed this problem by merging the text of the predecessor pages with the text of the page to be classified or by keeping a separate feature set for the predecessor pages. Chakrabarti, Dom and Indyk (1998) evaluated 2 variants; one that simply appended the text of the neighbouring pages to the text of the target page and one

that used two different sets of features. Results were not good and the scientists concluded that the text from the neighbors is too noisy to help classification.

Other authors explicitly encode the relational structure of the WWW in first-order logic. Craven *et al* (1998) used a variant of Foil Algorithm (Quinlan, 1990) to learn classification schemes that incorporate features from neighbouring pages. Slattery and Mitchell (2000) improved the Foil algorithm by integrating it with HITS while Craven and Slattery (2001) combined Foil with a Naïve Bayes classifier to obtain better results.

Despite the efforts of the above authors, the aforementioned approaches still present the following shortcomings;

- Features of predecessor pages should not be considered as shown by Chakrabarti *et al*(1998a).

- The entire text of the predecessor page may not be relevant.

- Not all pages have relevant meta-information.

Later, Fürnkranz (2001) introduced hyperlinks ensembles for classification of web pages to resolve the above problems. This he achieved by training a classifier that classifies hyperlinks according to the class of pages they point to based on the words that occur in their link neighbourhood (base case, anchor text). Hence each page is assigned different predictions for class membership which are combined to a final prediction by voting procedures.

Hyperlink ensembles outperformed a conventional full-text classification in the WebKB domain as shown by Craven *et al* (2000). Later, different voting schemes were compared. Features such as headings preceding the current paragraph and anchor texts were investigated and turned out to be important.

3.2   Web Usage mining (WUM)

Web usage mining is the application of data mining techniques to discover usage patterns from web data in order to understand and better serve the needs of web based applications (Srivastava *et al*, 1999). It has seen a rapid increase in interest because of its application potentials in areas like e-commerce, web system design, system performance improvement, adaptive web sites and personalization.

Web usage data describes the pattern of usage of web pages and includes attributes such as user IP address, page references and the time stamp of user access. Data may be collected from different sources such as servers, clients and proxies.

3.2.1   Types of collection sources.

3.2.1.1. Server - level collection.

A server-class machine registers a web log entry for each page consisting of the requested URL, the user IP address and time stamp. The data recorded in the server logs reflects the access of a website by multiple users. Thus popular websites such as www.google.com and www.facebook.com may register records in the order of hundreds of terabytes daily. However, site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the web environment.

3.2.1.2. Client- level collection.

Client-side data can be implemented by using a remote agent (such as Java Applets or Javascripts) or by modifying the source code of an existing browser such as Mozilla Firefox to enhance data collection capabilities. Client-side collection is more advantageous than server-side collection as it solves the problem of caching and session identification. However it may be difficult convincing users to utilize modified browsers on a daily basis except by the use of incentives.

3.2.1.3. Proxy-level collection.

A web proxy acts as an intermediate level of caching between client browsers and web servers. Proxy caching can be used to reduce the loading time of a web page experienced by users as well as the network traffic load at the server and client-sides. This may help characterize the browsing behavior of a group of anonymous users sharing a common proxy server.

3.2.2   Some definitions.

The data from these three collection sites may identify data abstractions such as user, server sessions, episodes, click streams and page views. For convention sake, the World Wide Web Consortium Web Characterization Activity (WCA) defines the above data abstractions below as:

1. A user is a single individual that is accessing files from one or more servers through a browser.

2. A page view consists of every file that contributes to the display on a user's browser at one time.

3. A click stream is a sequential series of page view requests.

4. A user session is the click stream of page views for a single user across the entire web.

5. A server session (or visit) is a set of page views in a user session for a particular website.

6. An episode is a semantically meaningful subset of a user or server session.

There are three main tasks for performing web usage mining namely preprocessing, pattern discovery and pattern analysis.

3.2.3   Preprocessing

This consists of converting the usage, content and structural information contained in the available data sources into the data abstractions necessary for pattern discovery.

3.2.3.1. Usage preprocessing.

This is arguably the most difficult task in the web usage mining process due to the incompleteness of the available data. Some of the typically encountered problems are;

- Single IP address/multiple server sessions. Internet service providers (ISPs) typically have a pool of proxy servers that users access the web through. A single proxy server may have several users accessing the website, potentially over the same time period.

- Multiple IP address/ single server session: Some ISPs randomly assign each request from a user to one of several IP addresses. In this case, a single server session can have multiple IP addresses.

- Multiple IP address/Single user: A user that accesses the web from different machines would have different IP addresses from session to session making repeat visits from the same user impossible to track.

- Multiple Agent/ Single user: Again, a user that uses more than one browser even on the same machine would appear as multiple users. Assuming that each user has been identified, (using cookies, logins etc), a click stream for each user is divided into sessions. This data is then cleaned, condensed and transformed in order to retrieve and analyze significant and useful information.

3.2.3.2 Content preprocessing.

Content preprocessing consists of converting the text, image, scripts and other files such as multimedia into forms that are useful for the web usage mining process. This consists of the following contents mining such as classification or clustering. While applying data mining to the content of website is an interesting area of web research in its own right, in the context of web usage mining, the content of a site can be used to filter the input to or output from the pattern discovery algorithms. For example, results of a classification algorithm could be used to limit the discovered patterns to those containing page views about a certain class of products. Also, page views can be classified according to their intended use. Page views can be intended to convey information to the user, gather information from the user, allow navigation, or a combination of the above.

Dynamic page views present more of a challenge. Content servers that employ personalization techniques or draw upon databases to construct page views may be capable of forming more page views than can be preprocessed.

3.2.3.3. Structure preprocessing.

The structure of a site is created by hyperlinks between page views. The structure can be obtained and preprocessed in the same manner as the content of a site.

3.2.4   Pattern discovery.

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. This section describes the kind of mining activities that have been applied to the web usage domain.

3.2.4.1. Statistics.

Statistical techniques are the most common methods to extract knowledge about visitors to a website. By analyzing the session file, one can perform different kinds of descriptive statistical analysis (mean, median mode etc) on variables such as page views, viewing time and length of a navigational part. Despite the fact that such methods lack the depth of analysis, the knowledge derived could be used to improve system performance, system security, site modification and support marketing decision.

3.2.4.2. Association.

Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. Association rule discovery may reveal a correlation between users who visited a page containing electronic products to those who access a page about sporting equipment. Also, the presence or absence of such rules can help web designers restructure their websites.

3.2.4.3. Clustering.

It is a technique to group together a set of items having similar characteristics. In web usage mining, there are types of interesting clusters to be discovered, namely usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentations in e-commerce applications or provide personalized web contents to the users. On the other hand, clustering of pages would discover groups of pages having related contents. This is useful for internet search engines and web assistance providers.

3.2.4.4. Classification.

It is a task of mapping a data item (a target) into a predefined class. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, support vector machines etc. For example classification on server logs may lead to the discovery of interesting rules such as: 30% of users who placed an online order in /product/music are in the 18-25 age group and live on the west coast.

3.2.4.5. Sequential patterns.

The technique of sequential patterns discovery finds intersession patterns such as the presence of a set of items is followed by another item in a time ordered set of sessions or episodes. By using this approach, web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups.

3.2.5   Pattern analysis.

Pattern analysis is the last step in the overall web usage mining process. Pattern analysis aims at filtering out uninteresting rules or patterns from the set found in the pattern discovery phase. The most common form of pattern analysis consists of a knowledge query mechanism such Structured Query Language (SQL). Another method is to load usage data into a data cube in order to perform an Online Application Processing (OLAP) analysis. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.
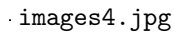
Figure 1: World-wide web photo

One of the most interesting application areas of web usage mining is personalization. Personalization: web usage mining is greatly applied to areas of personalization due to the paradigm shift from mass communication to mass customization. Knowledge discovered from web logs can be used to make a web site more responsive to the unique needs of each user. It covers areas like recommender systems, customization and adaptive websites.

Spiliopoulou (2000) provided a rationale for why web log data should be mined. To retain users, a website should effectively provide them with the contents they need in the most optimized manner. Spiliopoulou describes a process by which mining for navigational patterns may be used to gain insight into a website usage and optimality with respect to its current user population.

Cingil, Dogac and Azgin (2000) described the need for interoperability when mining the web and how W3C standards such as extensible markup language (XML), resource description framework (RDF), and platform for privacy preferences (P3P), can be used to achieve personalization applications.

Perkowitz and Etzioni (2000) addressed personalization as a process that adapts an internet site through the automated generation of index pages for the website. The article explores adaptive websites (sites that automatically improve their organization and presentation by learning from visitor access patterns). Adaptive websites are easily navigable.

3.3   Web Content mining (WCM)

Web Content Mining may be defined as the application of data mining techniques to un-structured or semi structured data, usually HTML documents. The content of a web page is the data the page was designed to convey to the users usually text and graphics. The data stored on the WWW is mostly text in a semi structured format, thus some areas of study that pivot web content mining like information retrieval and database systems have been developing in parallel for many years. A typical information retrieval problem is to locate relevant documents based on user input, such as keywords e.g. documents, lack of structure of documents and the notion of relevance. Two basic measures for assessing the quality of text retrieval are given below:

• Precision: This is the percentage of retrieved documents that are relevant to the query.

$$\frac{|Relevant \cap Retrieved|}{|Retrieved|} \tag{4}$$

• Recall: This is the percentage of documents that are relevant to the query and were retrieved.

$$\frac{|Relevant \cap Retrieved|}{|Relevant|} \tag{5}$$

Most information retrieval systems support keyword based and/or similarity based re-trieval. In keyword-based information retrieval, a document is represented by a string which can be identified by a set of keywords provided by a user. A good information retrieval system should consider synonyms when answering such queries. Keyword-based retrieval is a simple model that can encounter two major difficulties.

1.The synonymy problem: A keyword such as 'Harvard' may not appear anywhere in Harvard University's website even though the page is very closely related to Harvard University.

2.The Polysemy problem: The same keyword such as 'mining' may mean different things in different contexts.

In Similarity-based retrieval, similar documents are retrieved based on a set of common keywords. The output of such retrieval should be based on the degree of relevance.

In information retrieval systems, a stop list is associated with a set of documents. A stop list is a set of words deemed irrelevant e.g. *The, a* and *of* and,may vary from document to document e.g. databases could be a key word in a newspaper but a stop word amongst research papers in a database system conference. A group of different words may share the same word stem. That is they are syntactic variants of each other.

3.3.1   Modeling text.

In order to model text, a machine representation of world knowledge is built and thus must involve a natural language grammar. We discuss such models in the following section below. Given a set of D documents and a set of T terms, each document could be modeled as a vector V in a T-dimensional space R^t. The jth coordinate of V is a number that measures the association of the jth term with respect to the given document. Vj is set at 0 if document does not contain the term and Vj is set at 1 as long as the Jth term occurs in the document, or let Vj be the term frequency (number of occurrences of the term Tj in the document) or the relative term frequency (term frequency versus total number of occurrences of all terms in the document). Metrics such as the cosine measure have been proposed to measure document similarity. If v1 and v2 are two document vectors, the cosine similarity is

$$Sim(v_1, v_2) = \frac{v_1.v_2}{|v_1||v_2|} \tag{6}$$

Where$v_1.v_2$= the vector product and $|v|^2 = v.v$.

The similarity metrics for documents may be used to construct similarity -based indices on

such documents. Text-based queries are represented as vectors and their closest neighbors are selected from the document collection.

For any document database with D documents and T terms, we have that D and T is usually large leading to inefficient computations. Also, the sparsity of vectors made it difficult to detect and exploit relationships among terms. To overcome these problems, a latent semantic indexing method was developed that effectively reduced the size of a frequency table for analysis.

3.3.1.1. Latent semantic indexing method.

Deerwester et al (1990) developed the Latent semantic indexing method for document similarity analysis which used a well known technique in matrix theory called the Singular Value Decomposition (SVD) to reduce the size of the term frequency matrix and hence resolve the above conflicts. This consists of the following basic steps:

1. By creating a T x D term frequency matrix, F.

2. Computing singular value decompositions of F by splitting F into three smaller matrices U, S and V where U and V are orthogonal,S is a diagonal matrix of size K x K and is a reduced version of F.

3. Replacing the original vector of each document in D by a new one that excludes the terms eliminated during SVD.

4. Storing the set of all vectors and creating indices for them using advanced multidimensional indexing techniques.

The transformed document vectors can be used to compare the similarity between two documents or find the top N matches for a query.

3.3.1.2. Inverted index.

An inverted index is an index structure that maintains two hash indexed or B+ - tree indexed tables: document_ table and term_table, where

• document_ table consists of a set of document records, each containing two fields: doc_id and posting_list, where posting_list is a list of terms that occur in a document, sorted according to some relevance measure.

• term_table consists of a set of term records, each containing two fields: term_id and posting_list where posting_list specifies a list of document identifiers in which the term appears.

With such organization, it is easy to answer queries like "find all the documents associated with a given set of terms" Although inverted indices are widely used in industry and easy to implement, they are still not satisfactory at handling the synonymy and polysemy problems. Also, the posting_list could be rather long making the storage requirements very large.

3.3.1.2. Signature files.

Tsichritzis and Christodoulakis (1983) developed the use of signature files. A signature file is a file that stores a signature record for each document in the database with each having a fixed size of b bits representing terms. A simple encoding scheme goes as follows:

Each bit per document signature is initialized to 0. A bit is set to 1 if the term it represents appears in the document thus, a signature S_1 matches a signature S_3 and matches another signature S_2 if each bit that is set in signature S_2 is also set in S_1 and S_3 since there are usually more terms than available bits, there may be multiple terms mapped into the same bit. Such many-to-one mappings make the search expensive since a document that matches a signature of a query does not necessarily contain the set of key words of the query

3.3.2   Keyword based Association analysis.

Association analysis is the discovery of association rules showing attribute values that occur frequently together in a given set of data. It collects sets of keywords or terms and finds the correlation relationship among them.

Feldman and Hirsh (1998) studied methods for association rule mining in text databases and found that the text data is first preprocessed by parsing, stemming, removing stop words and then invoking association mining algorithms. Each document can be viewed as a transaction while a set of keywords in a document can be considered as a set of items in the transaction. That is the database is in the format (*documents_id, a_set_of_key_words*). Thus the problem of keyword association mining in databases is reduced to item association mining where many interesting solutions have been developed. It can be observed that a set of frequently occurring consecutive or closely related keywords may vary from a term or a phrase. Association analysis helps detect compound association that is domain dependent phrases e.g. [Cameroon, President, Paul, Biya] or non-compound associations such as [CFA Francs, Shares, Douala exchange, State, Securities]. Term recognition and term level association mining enjoys certain advantages such as;

Texts and phrases that are automatically tagged hence not needing human effort.

The numbers of irrelevant document returns are reduced.

Apart from keyword-based association analysis, document classification analysis is an important task in text mining. However these techniques have already been discussed under WSM above.

Please see a picyure below.
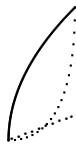
4   Conclusion

In attempting to review the efforts made by the scientific community to solve data mining problems on the World Wide Web, the solutions to web mining problems have begun motivating and will continue to greatly motivate business and most especially e-commerce. Web mining being a relatively new field, will continue to arouse interest among researchers.

5   References

[1] Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann Publishers.

[2] Data mining (2010).In *Wikipedia, the free encyclopedia.* Retrieved August 17, 2010, from `http://en.wikipedia.org/wiki/Data_Mining`

[3] Scime, A. (2005). *Web mining: Applications and techniques.* Hershey, PA: Idea Group Publishing.

[4] Chakrabarti, S., Kleinberg, J., Gibson, D., Kumar, R., & Raghavan, P. (1999). Mining the Web's Link structure. *IEEE Computer*, August 1999.

[5] Fürnkranz, J., (2007). Web structure Mining: exploiting the Graph structure of the World Wide Web.

[6] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web Usage Mining: Discovery and Application of Usage Patterns from Web Data, *SIGKDD Explorations*, ACM.

[7] Social Networks (2010). In *Wikipedia, the free encyclopedia.* Retrieved August 20,2010, from `http://en.wikipedia.org/wiki/Social_networks`.

[8] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. ACM-SIAM *Symposium on Discrete Algorithms.*

[9] Bharat, K. & Henzinger, M. (1998). Improved Algorithms for topic distillation in a hyperlinked environment. 21st ACM SIGIR *Conference on Research and Development in Information Retrieval.*

[10] Brin, S. & Page, L. (1998). The anatomy of a Large-scale hypertextual web search engine. Proceedings of the 7th World Wide Web Conference (WWW7).

[11] Bharat, K., Broder A., Henzinger, M., Kumar, P., & Venkatasubramanian, S. (1998). The connectivity Server: Fast access to linkage information on the Web. 7th World Wide Web Conference, Brisbane, Australia.

[12] Chakrabarti, S., Dom, B., & Indyk, P. (1998). Automatic resource complication by analyzing hyperlink structure and associated text. Proceedings of the 7th International World Wide Web Conference, Brisbane Australia.

[13] Quilan, J.R. (1990). Learning logical definitions from relations. Machine Learning,5(3).

[14] Craven, M., DiPasquo, D., Freitag, A., McCallum, T., Mitchell, K., Nigam, K., & Slattery S. (2000). Learning to construct Knowledge bases from the World Wide Web. Artificial Intelligence Archive,118(1-2).

[15] Craven, M. & Slattery, S. (2001). Relational Learning with statistical predicate invention: Better models for hypertexts. Machine learning,43(1-2).

[16] Fürnkranz J (1999). Exploiting Structural information for text classification on the WWW. Proceedings of he 3rd International Symposium (IDA-99), Amsterdam, Netherlands.

[17] Fürnkranz, J. (2001). Hyperlink ensembles: A case study in hypertext classification. Information Fusion,3(4).

[18] Cohen, E., Krishnamurthy, B., & Rexford (1998). Improving end-to-end performance of the web using servers' volumes and proxy filters. The ACM SIGCOMM '98 Conference on Applications, Techniques, Architecture and Protocols for Computer Communication.

[19] Pirolli, P., Piktow, J., & Rao, R. (1996). Silk from a Sow's ear: Extracting Usable Structures from the Web. ACM SIGCHI '96, Vancouver.

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

[21] Spiliopoulou, M. (2000). Web Usage mining for Web site evaluation. Communications of the ACM, 43(8).

[22] Cingil, I., Dogac, A. & Azgin, A. (2000). A Broader Approach to Personalization. Communications of the ACM.

[23] Perkowitz, M. & Etzioni, O. (2000). Adaptive web sites. Communications of the ACM.

[24] Deerwester, S., Dumais, S., Furnas, G., Laundauer, R., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science,1(1).

[25] Tsichritzis, D., & Christodoulakis, S. (1983). Message files. ACM Transactions on Information Systems, 1(1).

[26] Feldman, R. & Hirsh, H. (1998). Finding Associations in Collections of Text. In Machine learning and Data Mining: Methods and Applications (pp.223-240). New York: John Wiley & Sons.

[27] Chakrabarti, S. (2000), Data Mining for hypertext: A tutorial survey. ACM SIGKDD explorations, 1(2).