IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. II (Jan. 2014), PP 49-56
www.iosrjournals.org

Web Mining for an Academic Portal: The case of Al-Imam Muhammad Ibn Saud Islamic University

Hamad Ibrahim Alomran, PhD
Department of Information Management, College of Computer and Information Sciences
Al-Imam Muhammad Ibn Saud Islamic University
Riyadh, Saudi Arabia

Abstract: *This paper highlights the role of web mining in building a web retrieval system to extract hidden knowledge for the Semantic Web by using association rule mining techniques to find a frequent itemset for the Al-Imam University portal. An experimental method is used in this paper. The proposed system builds an information system in which web (data) mining in Semantic-Web technology is applied using the correlation base association rule algorithm. The steps those are necessary for building ontology for an academic portal by using the OntoStudio tool. Using the OntoPortal system will be a successful way to build a semantic portal using a large itemset as the result of web mining to build and classify ontologies for a semantic portal.*
Key words: *web mining, association rule mining, Semantic Web*

I. INTRODUCTION

Analysis of the data will provide universities with the information needed to provide effective feedback to all stakeholders in the system. The different kinds of web data that are available in universities are mined, which assists universities to assess the services by enabling them to provide more personalised services.

This mining also contributes to the search for potential value-added services. Web-usage mining is the process of extracting useful information from server logs (i.e. web-usage mining is the process of discovering what users are looking for on the internet). Though the web is rich with information, gathering and making sense of the data is difficult because the documents on the web are generally unorganised. The greatest challenge over the next several decades will be developing methods to extract effectively and efficiently machine-understandable and quotable information and knowledge layers, referred to as the Semantic Web, from unorganised, human-readable web data. [1][2]

This problem has motivated many specialists to search for new methods and techniques for information retrieval, and to discover new concepts to solve this issue. The aim is not retrieving available information on the web, but creating information and new knowledge using the available information and data, and this is achieved through creating links between information. This study endeavours to build an information system on the web using the association rule mining techniques and by discovering new knowledge that presents information through the designed portal. This leads to integrated research that assists the user to access the greatest amount of possible integrated and interrelated information [8]. The importance of this paper is presented from two perspectives.

The theoretical perspective is as follows. Being one of the few studies that deals with data mining in the Semantic Web, the researcher hopes that this paper will enrich cognitive stock in its field, and open the way for detailed future studies that deal with aspects that are not addressed in this paper, or dealt with in the literature.

The practical perspective is as follows. The practical importance of this paper lies in building an information system available on the web through which information and knowledge can be retrieved and new knowledge can be discovered through data-mining technology in the Semantic Web.

This paper aims to build an information system on the web based on data mining in the Semantic Web. The first step is to build an information system on the web that permits entering bibliographic data in all research work in the data base, including the name of the author, title of the paper, periodical name, number, month, year, page numbers and keywords. The researcher uses the experimental method, as it is suitable to the nature of the research. An information system is built where data mining in Semantic-Web technology is applied using the correlation base association rule algorithm. Initially, a knowledge database is built, accordingly building ontology for new knowledge using the association rule algorithm, and compositing a 'large itemset' equation, then classifying this, giving it weights, and adding it to the composed-knowledge databases.

This paper discusses the concept of the Semantic Web and examines the manner in which ontology is used in the application of the Semantic Web and demonstrates the steps that must be followed to create an ontology portal. An example of ontology will be provided at the end. The emphasis of this paper is on the construction of a semantic portal through the creation of ontological elements for an academic portal.

II. LITERATURE REVIEW

While many studies address the issues of data mining and the Semantic Web, few studies deal with the connecting these.

One of the most important studies in the field of data mining was conducted by Dhenakaran and Yasodha [16] who state that Semantic-Web mining aims to combine the development of two research areas: the Semantic Web and web mining.

Zhou, Hui and Fong [21] state that with the explosive growth of information on the web, it has become more difficult to access relevant information from the web. One possible approach to solve this problem is web personalisation. In the Semantic Web, user access behaviour models can be shared as ontology. Agent software can then utilise it to provide personalised services such as recommendation and search.

Chakravarthy [3] proposes research on the manner in which Semantic-Web technologies can be used to mine the web for information extraction. Chakravarthy [3] also examines the manner in which new unsupervised processes can assist in extracting precise and useful information from semantic data, thus reducing the problem of information overload. Chakravarthy [3] points out that the Semantic Web adds structure to the meaningful content of web pages; hence, information is given a well-defined meaning, which is both human readable as well as processed by machine.

III. DATA MINING AND KDD

Data mining is primarily used today by companies with a strong consumer focus (e.g. retail, financial, communication, and marketing organisations). It enables companies to determine relationships among 'internal' factors such as price, product positioning, or staff skills, and 'external' factors such as economic indicators, competition, and customer demographics. In addition, it enables companies to determine the effect of these factors on sales, customer satisfaction, and corporate profits. Data mining also enables companies to 'drill down' into summary information to view detailed transactional data. [5]

Thearling [17] agrees with this concept, and believes that the extraction of hidden predictive information from large databases is a powerful new technology with great potential to assist companies to focus on the most important information in their data warehouses. Data-mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data-mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Hegland [9] also agrees with this concept, and believes that data mining is the search for relationships and global patterns that exist in large databases but are hidden among vast amounts of data (e.g. the relationship between patient data and their medical diagnosis). Hegland [9] also believes that data mining is an attempt to innovate a simplified model of the complex world described in the database. Therefore, it can be said that data mining is a method by which to deal with the massive amounts of information that exist on the web, and a method to assist in finding useful information faster.

The Gartner Group [7] also agrees with the concept of data mining. The Gartner Group [7] believes that data mining is a process of discovering new meaningful links, patterns and trends through scrutinising and auditing vast amounts of data stored in data warehouses, using techniques of recognising patterns in addition to statistical and mathematical methods and techniques.

IV. ROLE OF ASSOCIATION RULE MINING

a. Web Mining

Web mining is an application of data mining, and is considered an information process technique over the World Wide Web. It aims to merge and integrate information gathered by traditional data-mining methodologies and techniques with information gathered over the World Wide Web. It is worth noting here that the term 'mining' can be defined as extracting something useful or valuable from a baser substance, for example, mining gold from the earth. [10][15]

Franklin [2][6] defines web mining as the use of data-mining techniques to automatically discover and extract information from web documents and services. This area of research is so prevalent today in part due to the interests of various research communities, the tremendous growth of information sources available on the web and the recent interest in e-commerce.

b. Areas of Web Mining

Three areas of web mining are commonly distinguished as follows:

- Content mining,
- Structure mining,
- Usage mining.

In this paper, usage mining is considered along with the technical support from association rule mining. Association rule mining, one of the most important and well-researched techniques of data mining, was first introduced in the work of Agarwal, Aggarwal and Prasad [1]. It is used to discover association rules that satisfy the predefined minimum support and confidence from a given database. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, and inventory control. Various association mining techniques and algorithms will be briefly introduced and compared later in this paper.

In this paper, the proposed system using the Apriori Algorithm proposed by Agrawal and Aggarwal. [2] I = {i1, i2, …, im}: a set of items:

i. Transaction t:

Transaction data: a set of documents

- A text document data set (each document is treated as a 'bag' of keywords)
  - doc1: Student, Teach, school
  - doc2: Student, School
  - doc3: Teach, School, city, Game
  - doc4: E-course, Teacher
  - doc5: Colleges, Departments
  - doc6: Research centre, Professors.

ii. Rules:

- T a set of items, and $t \subseteq I$
- Transaction database T: a set of transactions T = {t1, t2, …, tn}
- A transaction t contains X, a set of items (itemset) in I, if $X \subseteq t$
- An association rule is an implication of the form: $X \rightarrow Y$, where X, $Y \subset I$, and $X \cap Y = \varnothing$
- An itemset is a set of items.
  - e.g. X = {milk, bread, cereal} is an itemset
- A k-itemset is an itemset with k items.
  - e.g. {milk, bread, cereal} is a 3-item itemset

iii. Support and confidence:

- Support: the rule holds with support sup in T (the transaction data set) if sup% of transactions contain $X \cup Y$.
  - $sup = Pr(X \cup Y)$
- Confidence: the rule holds in T with confidence conf if conf% of transactions that contain X also contain Y.
  - $conf = Pr(Y \mid X)$
- An association rule is a pattern that states when X occurs, Y occurs with certain probability.
- Support count: the support count of an itemset X, denoted by X count, in a data set T is the number of transactions in T that contain X. Assume T has n transactions.
- Then,

$$support = \frac{(X \cup Y).count}{n} \quad (1)$$

$$confidence = \frac{(X \cup Y).count}{X.count} (2)$$

iv. Goal:

Find all rules that satisfy the user-specified minimum support (minsup) and minimum confidence (minconf).

c. Association Rule with Web-usage Mining

Web-usage mining is the automatic discovery of user-access patterns from web servers.

i. Direct Association in the Web

Direct association rules represent regularities discovered from a large data set. The problem of mining association rules is to extract rules that are strong enough and have the support and confidence value greater than given thresholds. For example, the following figure demonstrates the direct relation of two web pages. In Figure 1, di and dj are two web pages that are directly associated with each other [18].

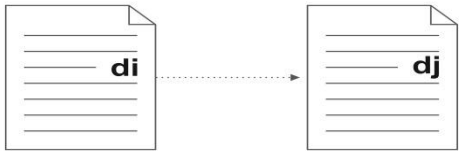

Figure 1: Direct Association in the Web

ii. Indirect Association in the Web

Let di be an independent web page (document) and D a website content (the web-page domain) that consists of independent web pages di € D. The support function is used only to exclude weak rules (i.e. only rules that exceed the level of the minimum direct support 'supmin' are considered for recommendation). The support expresses the popularity of a given rule among all others. A direct confidence function con (di-^dj) denotes with which belief the page dj may be recommended to a user while watching the page di. The direct confidence factor is the conditional probability P(dj ^ di) that a session containing the page di also contains the page dj:

con (di→dj) = P(dj |di) = nij/ni (3)

Where nij is the number of sessions with both di and dj, ni stands for the number of sessions that contain a:

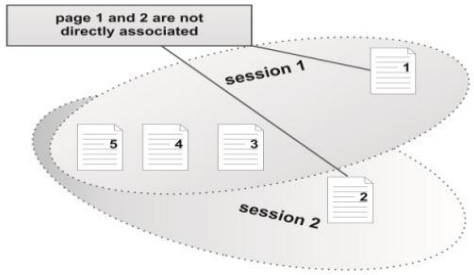

Figure 2: Indirect Association in the Web

The above Figure 2 demonstrates the indirect correlation of page one and two. They are indirectly associated with mediator pages; in the first web session, the first page comes with third, fourth and fifth pages. In the next session, the second page comes with these mediators. Such associations are known as indirect association in the web. [19]

iii. Partial Indirect Association Rules

Partial indirect association rule di—>dj, dk is the indirect implication from di to dj with respect to dk, for which two direct association rules exist: di—> dk and dk—>dj; sup(dk—>dj)>supmin, con(dk^dj)>conmin; where di, dj, dk€D; di^dj^dk. The page dk, in the partial indirect association rule di—> dj, dk, is referred to as the transitive page. There may be many transitive pages dk for a given pair of pages di, dj and as a result, many partial indirect association rules di—>dj, dk. Each indirect association rule is described by partial indirect confidence, conP(di—>dj, dk), as follows:

conP (di → dj, dk) = con(di → dk) con(dk → dj) (4)

Pages di, dj in di—> dj, dk do not need to have any common sessions, but we respect only 'good' direct associations to ensure that indirect associations are based on sensible grounds. From questionable or uncertain direct knowledge we should not derive reasonable indirect knowledge.

Consequently, it was assumed, rules di—>dk and dk—> dj must be 'strong' enough so that con(di—>dk) and con(dk—>dj) exceed conmin [19].

V. EFFECT OF PAGE-RANKING ALGORITHMS

The World Wide Web is a large repository of interlinked hypertext documents accessed via the internet. The web may contain text, images, video, and other multimedia data. The user navigates through these using hyperlinks. Search engines give millions of results and apply web-mining techniques to order the results. The sorted order of search results is obtained by applying special algorithms referred to as page-ranking algorithms.

There are two page-ranking algorithms: PageRank and Weighted PageRank. These are the most commonly used algorithms in web-structure mining. The algorithms measure the importance of the pages by analysing the number of inlinked and outlinked pages [21].

a. PageRank

PageRank is a numeric value that represents the importance of a page is on the web. PageRank is Google's method of measuring a page's 'importance'. When all other factors such as title tag and keywords are considered, Google uses PageRank to adjust results so that more 'important' pages move up in the results page of a user's search-result display. Google believes that when a page links to another page, it is effectively casting a vote for the other page. Google calculates a page's importance from the votes cast for it. The importance of each vote is considered when a page's PageRank is calculated. This is important because it is one of the factors that determine a page's ranking in the search results. It is not the only factor that Google uses to rank pages, but it is important. [14]

The order of ranking in Google works as follows:

- Find all pages matching the keywords of the search.
- Adjust the results by PageRank scores.

b. WPCR

The Weighted Page Content Rank Algorithm (WPCR) is proposed by Thearling [17] as a page-ranking algorithm that is used to provide a sorted order to web pages returned by a search engine in response to a user query. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web-structure-mining techniques as well as web-content-mining techniques. Web-structure mining is used to calculate the importance of the page, and web-content mining is used to discover the relevance of a page. In this context, the term 'importance' means the popularity of the page (i.e. how many pages are pointing to or are referred to by this particular page). It can be calculated based on the number of inlinks and outlinks of the page. In this context, 'relevance' means matching of the page with the query. If a page is maximally matched to the query, it becomes more relevant.

As stated earlier, WPCR is a numerical value used to represent the ranking of a web page. It can be seen in the formula to calculate the WPCR of a page U (equation 6):

$$PR(u) = (1-d) + d \sum_{v \varepsilon B(u)} PR(V) W^{in}(U,V) W^{out}(U,V) * (Cw + Pw) \quad (6)$$

Where PR (U) = page rank of page U, B (U) = set of all pages referring to page U. D = damping factor, which can be set between 0 and 1. $W_{in}(U,V)$ = inweight of link (U,V), W_{out} (U,V) = outweight of link (U, V), Cw = content weight of page U, Pw = probability weight of page U.

The various steps of the algorithm are explained below in detail. [4][12]

i. Algorithm: WPCR Calculator

Input Page P, Inlink and Outlink Weights of all backlinks of P, Query Q, d (damping factor). *Output:* Rank score:

- Step 1: relevance calculation:
  a) find all meaningful word strings of Q (say N)
  b) find whether the N strings are occurring in P or not
  c) Z = sum of frequencies of all N strings
  d) S = set of the maximum possible strings occurring in P
  e) X = sum of frequencies of strings in S
  f) content weight (CTF) = X/2
  g) C = number of query terms in P
  h) D = number of all query terms of Q while ignoring stop words
  i) Probability weight (PW) = C/D.
- Step 2: rank calculation:
  a) find all backlinks of P (say set B)
  b) PR(P) = (1-d) + d

$$\left[\sum_{V \varepsilon B} PR(V) W^{in}(P,V) W^{out}(P,V) \right] (CW + PW) \quad (7)$$

c) Output. PR (P) (i.e. the rank score).

ii. Weight Calculation

The Win (v,u) and Wom (v,u) are the pre-processed weights. Both are simply inputted to the algorithm. Win(v,u) is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v and is given in equation (8).

$$W^{in}(U, V) = \frac{Iu}{\sum_{P \varepsilon R(V)} Ip} \quad (8)$$

Where Iu = number of inlinks of page u, L = number of inlinks of page p, R(v) = reference page text of page v. The Wout(v,u) is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v given in equation (9).

$$W^{out}(U,V) = \frac{Ou}{\sum_{P \varepsilon R(V)} Op} \quad (9)$$

Where Ou = number of outlinks of page u, Op = number of outlinks of page p.

iii. Relevance Calculation

The relevance calculator calculates the relevance of a page on the fly in terms of two factors: one represents the probability of the query in the page and the other gives the maximum matching of the query to the page [12].

The probability weight is the probability of the query terms in the web page. This factor is the ratio of the query terms present in the document and the total number of terms in the query (after ignoring factors such as stop words). The formula is given in equation (10):

Probability weight (PW) = Y/N (10)

Where Yi = number of query terms in i ith document, N = total number of terms in query. The content weight is the weight of content of the web page with respect to the query terms. This factor is the ratio of the sum of the frequencies of the highest possible query strings in order and sum of frequencies of all query strings in order. The maximum possible strings are selected in such a manner that all such strings represent a different logical combination of words. The formula for content weight is provided in equation (11):

Content weight (Cwi) = Xi/M (11)

Where Xi = sum of cardinalities of the highest possible query strings in order, M = sum of cardinalities of all possible meaningful query strings in order.

c. Comparison between PageRank and Weighted PageRank and Proposed WPCR

Table 1 presents a comparison between the two algorithms based on page rank and is weighted on how to use the links and the numbers of the links and the type of the mining using.

Table 1: Comparison between Two Algorithms

SN	Page rank/weighted page rank	WPCR
1	They rely only on links	They rely links as well as content
2	Page rank/weighted page rank is the structure mining only	WPCR is the combination of structure mining and content mining
3	Minimum determination of the relevancy of the paces to the given query page rank/weighted page rank algorithms provide important information about a given query	Maximum determination of the relevancy of the pages to the given query. WPCR algorithms provide important information and relevance about a given query

VI. EXPERIMENTAL SETUP AND TEST ENVIRONMENT

This paper has proposed to build an information-retrieval system for an academic portal by using the WPCR to order web results by finding large itemsets to cluster the results as frequent items.

A series of experiments has been conducted to discover the direct and indirect association rules from real web data. The data used for the experiments came from web log files of Al-Imam Muhammad Ibn Saud Islamic University (http://www.imamu.edu.sa). Each time the site is accessed, it is considered as a transaction and the web pages are considered as items. The page set and user sessions are unordered and without repetitions, turn navigational sequences (paths) into sets. Additionally, user sessions may be filtered to omit sessions that are too short or too long, as they are not sufficiently representative. This experiment used a combination of the Apriori Algorithm and the In-Direct Association Rules Miner (IDARM) Algorithm to obtain the complete indirect rules. This is coded in Java. Aprori is an influential algorithm for mining frequent itemsets for Boolean association rules. It employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+l)-itemsets. First the set of frequent 1-itemsets is found. This set is denoted LI, then LI is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The

discovery of each Lk requires one full scan of the database. The problem of mining association rules can be classified into two major sub-problems:

- Discovering all frequent itemsets for the academic portal for professors and colleges
- Using the frequent itemsets to generate strong rules, the professors' interests and research areas and the links in which they have interest.

Once all frequent itemsets from transactions in a database D have been found, it is straightforward to generate strong association rules from them, where strong association rules satisfy both a minimum support and minimum confidence threshold. IDARM was introduced to discover complete indirect association rules di—> dj and their complete indirect confidence con(di—> dj) from the set of direct rules di—>dj and proper input direct rules (i.e., those that exceed supmin and conmin are previously extracted using one of the well-known Apriori mining algorithms). The IDARM Algorithm constitutes the second stage in the recommendation process based on association rules.

The following steps are performed to obtain the result:

- Step (l) first the log data were cleansed of http://www.imamu.edu.sa. After removing one-page sessions and too long one. Initially, there were number of daily visits (7,940) and daily page views (33,349); this means the percentage of visitors is 77.83%. After cleansing and with our considerations only 50 visits are taken. For example, 212.138.117.22 [07 December 2013:05:04]
- Step (2) the search pre-processed uniform resource locators (URLs) are converted into the required input format of the implemented algorithm. One more module is developed to obtain the proper input of the above-implemented algorithm. The output of this module will be in matrix format where number of rows denoted as web pages and number of columns are taken as number of visits. This is also coded in Java. The output is as follows:
  - 1000
  - 0100
- Step (3) the matrix form dataset is now fed to IDARM to obtain the indirect rules of web data. Minimum support and minimum confidence are defined in con fig. text file. These are 10% and 0.2%, respectively. To ensure that indirect rules were formed only from strong direct ones, conmin and supmin thresholds were introduced. All the related details are given in the following Table 2.

Table 2: Number of Transactions

Number of transactions	Number of web pages	Supmin	Conmin	Number of direct rules	Number of indirect rules
500	~50	10%	0.2%	245	60

After developing this ontology, the next step is to convert it to a knowledge base. This is achieved by entering the information of academic staff such as research papers and teaching materials to the ontology.

The steps of entering information into the ontology are repeated until completing all information saved. During this step, the proposed system uses the association rule to generate the loge itemset and save it into the knowledge base. The owl file output from this process is the targeted knowledge base. This knowledge base can now be used in any Semantic-Web application that is developed.

a. Significance of Indirect Associations

The indirect associations are as significant as the direct associations of web data and these rules are being used in site modification, system improvement, web personalisation and many other applications. The indirect rules play a very significant role in web-usage mining. Through indirect rules, we can analyse users' behaviour and track their navigation path. Then we can modify our website according to the users' convenience (e.g. distance →datasheet).

This is a strong indirect rule. The user must access five to six links between two web pages. After discovering their indirect association, we can modify our website according to indirect correlation. This will reduce the access time for users. As such, indirect rules play a significant role in web mining.

b. Intermediate Outputs

The search engine generates some intermediate outputs, which can be stored in some appropriate data structures. These outputs can be represented in the form of tables and are used to calculate the importance and relevance of the web page.

The first output, shown in Table 2 is generated by the map-generator component. A module in this component provides information about the inlinks and outlinks of all web pages. The weight calculator pre-processes this information as demonstrated in Table 3.

Table 3: Output of Web-Map Generator

Document ID/URL	Inlinks	Outlinks

The definitions of the terms in Table 2 are as follows:

- 'Document ID' represents the URL of the web page—this column contains the URLs of all pages on the web.
- 'Inlinks' of a document are the numbers of documents referring that document.
- 'Outlinks' of a document are the numbers of documents referred to by that document.

The weight calculator calculates the inweight and outweights of a page and determines its importance as shown in Table 4.

Table 4: Output Obtained After Matching Query from Index

Document ID	Query string	Frequency	Selection

The definitions of the terms in Table 4 are as follows:

- 'Document ID' represents the URL of the web page—this column contains the URLs of all pages on the web.
- 'Query string' refers to a meaningful ordered sequence of words of the query fired by the user.
- 'Frequency' refers to the number of occurrences of query strings in document returned by a search engine.
- 'Selection': this column represents whether the query string is selected for consideration in the particular document. Its value is either 0 or 1.

VII. CONCLUSION

Semantic-Web mining is relatively new sub-field of data mining. It has a vast scope for investigation, considering the availability of tonnes of unstructured data on the World Wide Web. Applying association rule mining is yet another area to be explored. This paper integrates association rule mining with web-usage data and demonstrates the effect of direct association rules and page-ranking algorithms for the chosen domain. Integrating the indirect association rules and partial association rules and improving the page-ranking algorithms are a few of the challenges that remain for future research. In addition, it supports the e-library to obtain a higher ranking among universities that allow it to compete successfully with other leading library websites.

REFERENCES

[1] Agarwal, R. C., Aggarwal, C. C. & Prasad V. V. V. 2000. A tree projection algorithm for generation of frequent itemsets. Retrieved 8 January 2014, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.9466&rep=rep1&type=pdf

[2] Agrawal, D. & Aggarwal, C. C., On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (pp. 247–255), 2001. New York, NY: Association for Computing Machinery.

[3] Chakravarthy, A. Mining the Semantic Web. Workshop at the 13th European Conference on Machine Learning (Ecml'02)/6th European Conference on Principles and Practice of Knowledge Discovery in Databases (Pkdd'02), *4*(2), 2005.

[4] Berendt, B., Hotho, A., Stumme, G, Semantic Web mining and the representation, analysis, and evolution of web space. In *Proceedings of RAWS 2005 Workshop* (pp. 1–16). Prague-Točná, Czech Republic.

[5] Chen, M.-S., Jan, J., Yu, P. S., Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883, 1996.

[6] Franklin. 2014, Retrieved from www.ntu.edu.sg/home/ascpfu/future.html

[7] Gartner Group. Magic quadrant for customer data mining. Retrieved from http://dml.cs.byv.edu/~cgc/docs/mldm_tools/Reading/Gartner2007.pdf

[8] Han, J., Kamber, M., *Data mining: concepts and techniques*. San Francisco, CA: Morgan-Kaufmann Academic Press, 2001.

[9] Hegland, M., Algorithms for association rules. *Lecture Notes in Computer Science*, *2600*, 226–234, 2003.

[10] Jiang, Q., Web usage mining: Process and application. Presentation for CSE 8331, 2003.

[11] Kazienko, P., Mining indirect association rules for web recommendation. *International Journal of Applied Mathematics and Computer Science, 19* (1), 165–186, 2009.

[12] Kazienko, P. & Kuzminska, K. 2005. The Influence of indirect association rules on recommendation ranking lists. In *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications* (pp. 482–487). Los Alamitos, CA: IEEE.

[13] Sharma, P., Tyagi, D., Bhadana, P., Weighted page content rank for ordering web search result. *International Journal of Engineering Science and Technology*, *2*(12), 7301–7310, 2010.

[14] Jain, R., Purohit, G. N., Page ranking algorithms for web mining. *International Journal of Computer Applications*, *13*(5), 0975–8887, 2011.

[15] Rouse, M., Web mining. SearchCRM. 2005.

[16] Dhenakaran, S. S., Yasodha, S., Semantic web mining—a critical review. *International Journal of Computer Science and Information Technologies*, *2*(5), 2258–2261, 2011.

[17] Thearling, K. 2014. An introduction to data mining: Discovering hidden value in your data warehouse. Retrieved from: http://www.thearling.com/text/dmwhite/dmwhite.htm

[18] Wan, Q. & An, A., An efficient approach to mining indirect, associations. *Journal of Intelligent Information Systems*, *27*(2), 135–158. 2006.

[19] Wan, Q. & An, A., Efficient indirect association, discovery using compact transaction databases. In *Proceedings of the IEEE International Conference on Granular Computing, GrC'06*. Los Alamitos, CA: IEEE, 2006.

[20] Xing, W. & Ghorbani, A., Weighted PageRank Algorithm. In *Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04)*. Los Alamitos, CA: IEEE, 2004.

[21] Zhou, B., Hui, S. & Fong, A. 2014. Web usage mining for Semantic Web personalization. Retrieved from http://www.virtualvinodh.com/download/IT_Papers/Semantic%20Web%20Minnig/Web%20usage%20mining%20for%20semantic%20web%20personalization.pdf