Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Sciences and Engineering

Elixir Comp. Sci. & Engg. 60 (2013) 16356-16360

ISSN: 2229-712X

Application of data mining to improve the efficiency of a search engine

Geetha Mary A, Rohit Singh and Sudarshan Kumar
School of Computing Sciences and Engineering, VIT University, Vellore, India.

ARTICLE INFO

Article history:
Received: 1 May 2013;
Received in revised form:
24 June 2013;
Accepted: 9 July 2013;

Keywords
Search engine optimization,
Query clustering,
Web intelligence,
Concept clustering,
Web mining,
Multilayered web.

ABSTRACT

It is a significant challenge to search, comprehend and use the semi-structured HTML, XML, database-service-engine information stored on the web. This data is more complex and dynamic than commercial databases' data. Data mining has been applied to web-page ranking to supplement keyboard-based indexing. In this context, data mining improves the quality and efficiency of search results[1]. For the web to be at its best, we must improve its usability. Data mining can play a vital role in the development of intelligent web. It will make the web a more exhaustive, intuitive and intelligent, usable resource. The paper shows how data mining can be applied to discover and catalog important links and patterns that will make our web interactions directed and intelligent.

Introduction

An exhaustive and inimitable source for data mining techniques is provided by the web which is an enormous and vigorous repository of pages including unquantifiable hyperlinks and colossal amount of access and usage statistics. However, numerous problems must also be considered that obstruct useful resource and knowledge discovery:

i. The sheer structural and design intricacy of large databases associated with a web page is far telling in its nature compared to conventional repositories of text files. Several non-uniform structural layouts, styles of authoring and variety in content are intrinsic components of web pages. Also, lack of indexing complicates the task of efficiently searching data.

ii. The nature of information stored in the web database is highly volatile and dynamic in nature in terms of both, additions and updates as well as linkage framework and access records, e.g., News, Sports, Stock market web sites.

iii. The users of the internet differ greatly in their usage, interests and backgrounds which include wont of knowledge of structural complexities of information networks, heavy cost incurred by searches and the users also get swayed away by the huge amount of information the web database has to offer.

iv. The required, true and relevant search results are only present in a very small amount relative to the total information that the web has to offer or that which gets dug up when searches are made, thus slogging the entire process down due to unimportant data flooding.

The question remains – "How can a search find truly relevant, concentrated information and of distilled quality?"

As of today, there are primarily three ways in which a user can approach towards accessing web information:

i. Keyword-based: Uses keyword indexing or manual directories to search for documents.

ii. Deep querying: Used where a large catalog of information or data is stored and is hidden behind database query forms, otherwise inaccessible through non-dynamic URL links, e.g., amazon.com

iii. Random surfing of web pages through hyperlinks.

The above techniques have succeeded considerably and this fact goes on to indicate the great potential of the internet to become the most exhaustive information repository.

Concerns in Design

It has been quite a major concern in research to realize an intelligent Web and the achievement of this has overcoming of two intrinsic problems as its prerequisites:

i. The conventional methodologies for accessing the extremely voluminous data on the Web inherently adopt the keyword-based view at the level of abstraction.

ii. Replacement of prevailing crude access methodologies by more intuitive versions must be implemented at the service level.

Confines in Access

Data mining's role in the realization of intelligent Web will be very important despite the support for information searches by keyword, address and topic-based Web searches because the Web in its present form is unable to provide quality services which are also intelligent. In this regard, the factors under consideration which have steered the inspiration behind the research are:

i. Inefficient searches implementing the keyword-based algorithm, which are tainted with problems like

a. Returning an unreasonable and ineffective amount of information if the keywords belong to frequently popular categories like sports, etc.

b. Returning poor results in case of semantic overload and context-based discrepancies.

c. Missing highly relevant and effective results just because of absence of a particular keyword although that article is extremely informative in the context searched.

ii. Mediocre quality and effectiveness in deep Web-querying despite the presence of well-structured, well-designed and richly

Tele:
E-mail addresses: rohitrssingh@gmail.com

populated databases, due to the inability of web crawlers to query them. This ultimately leads to perpetually invisible but actually quite relevant information. To overcome this, we must integrate the heterogeneous databases having different individual query-supports that comprise the hugely informative Web.

iii. Manually constructed directories that are topic-oriented or type-based are highly appropriate as they represent a Web portion in an organized manner and also support searches based on semantics. This leads to efficient searching but such directories are costly, can provide only curtailed exhaustion and suffer in both scalability and adaptability.

iv. Absence of semantic queries that are tried by the developers to be replaced or substituted only with keyword-based algorithms with certain additional features like “match the exact phrase” or “match all the words” or “match only a single word” etc. but even then these don’t make up for the inefficiency caused due to the lack of these queries.

v. Feedback of users’ activities and usage statistics help enormously in observing collective behavior to identify the authentic, authoritative and high-quality web pages. However, as human activities are extremely temporal, the links in the Web must also be dynamic enough to support the flow.

vi. There is limited support for multidimensional analysis and data mining due to the predominantly keyword-based searching and consequently, curb the possibilities of operations like running queries to list data mining centers and then focus on those having high number of papers and analyze their changes.

The above confines have been a great incentive for researchers to develop ways to efficiently, effectively and accurately mine and discover internet resources.

Tasks in Mining for Efficient Knowledge Discovery

Further discussed are the problems in research that must be solved in order to implement effective development of intelligent Web.

Dredging search-engine data

An index-based search algorithm works by crawling the pages, arranging them in indices, constructing and storing enormous indices based on keywords that assist in finding the web pages with the specified keywords. Hence by using highly specific and conditional keywords, one can quickly and easily locate relevant results. [1]

But there are some concerning limitations inherent in the above mentioned algorithm:

i. A specific topic may contain innumerable document entries containing the keyword(s) which may lead to returning of a large number of results out of which most are quite irrelevant or poor in quality.

ii. There might me many documents which don’t contain the keywords that specifically define the mentioned or searched topic. This may lead to omission or rejection of search results that might have been otherwise useful and quite relevant to the user.

As a result, we propose to integrate data mining techniques with the search engine in order to improve the quality of search results. As a first step in this direction, we can proceed by dilating the keyword set to accommodate their synonyms and contextual partners also. For example, a search for the term “Jaguar” can include keywords to be searched for as “Jaguar car”, “Jaguar animal” or a search for “data mining” may include “data dredging” or “knowledge discovery” etc. This will give a larger result set of documents which can be further filtered for highly relevant results and then provided to the user finally[10]. An analysis of web linkages and dynamics thus provides a strong foundation for high quality knowledge discovery.

Web linkage framework analysis

The results returned to the user must not only be topic-specific but also, authoritative and of high quality. A direct measure of the authoritativeness of web pages can be determined by analyzing web page linkage frameworks.[4] Hyperlinks in web pages comprise a large amount of hidden human comments and annotations that prove to be helpful in automatically deducing the authority. For example, if a web page author creates a hyperlink on his/her page pointing to another page it can be considered as an endorsement or a certificate of authenticity of that page. In this way, the recursive and collective framework of hyperlinks can lead us to authoritative web pages. But there are some problems posed in this notion as mentioned below: [4]

i. Not every hyperlink can be considered as an endorsement as it might have been created for other purposes like advertising, navigation, etc.

ii. Generally, the owner of a web page will not choose to put a hyperlink to his/her competitor’s web page on his/her page despite its authenticity.

iii. Lack of detailing in web pages that are highly authoritative.

Here comes an important concept called “Hubs”. This term refers to a web page or a set of web pages that contain links to authorities. Although slim, it does provide links to relevant sites about a given topic. A hub can comprise either the links that are recommended on individual pages or catalogued by a third-party site. Thus, it helps to find authoritative web pages in that if a web page is pointed to by many hubs then it is quite authoritative and if a good hub points to web pages, those web pages are authoritative. This will help in mining of authoritative web pages and high-quality knowledge discovery. A direct consequence and practical example of identification of hubs and authoritative web pages are “PageRank”[1] and “HITS”[3] algorithms.

Automated categorization of documents on the Web

Automatic classification, in comparison to human readers, is highly desirable due to less cost and improved speed, when categorizing web documents. Conventionally, the categorization algorithms incorporate the use of training sets viz. a viz. positive and negative examples and assigning each document a category based on fixed categories and documents[5]. For example, the taxonomy of Rediffmail and its documents can be used as the training and test sets to realize a classifier which can then be used to classify other web documents.

Satisfactory results can be obtained by using conventional classification techniques like Bayesian classification, decision-trees, association rules, etc. to classify web pages. The semantic information associated with hyperlinks can be used to achieve even better results than the previously mentioned classification algorithms.

However, ingenuous use of terms in the neighbourhood of a document’s hyperlinks may lead to poor accuracy due to the presence of noisy and irrelevant data in pages which are a part of back linking, e.g., advertisements, etc.[5,6]

Generally, automatic classification is based only on positive and not negative data sets to avoid easy exclusion of relevant data.

Mining of semantic constructs and contents

Different web pages might have different semantic structures, e.g., a department's page vs. a professor page. Firstly, identification of relevant extractable structure is required using either manual or programmatic methods of induction from a set of document examples. Secondly, this information can be used to implement automatic extraction from information database. Also, it will lead to clearer and deeper analysis of database contents.[7]

Mining the volatile changes in Web pages

Identification, recognition and mining of the volatile and dynamic changes in web pages can be done based on three categories – content, structure and patterns of access[2]. Storage of the web data in chronological order helps to detect changes easily. But the exhaustive and massive nature of the information stored on the Web makes it difficult to maintain records in an orderly fashion. Thus, we resort to the mining of accesses to ensure better results delivered to the user. This can be implemented using mining of the logs or history of records in the web. The homogeneity in the patterns of access in terms of IP address, the website address and the time of the access can help us to identify the most requested web pages, the most frequent time of access and the most frequent users and the data populated in such a way can be mined for extracting knowledge using OLAP operations[9]. This algorithm can prove to be very beneficial for fields like business, market research and customer service support[8].

Incorporating Multidimensional Web

Just like we apply mining techniques like clustering and outlier analysis to data in data warehouses, we apply these techniques to sets of web pages treating them like data. The web pages are grouped into clusters if they are tightly bound in terms of data, access patterns or structure and following this, we create descriptors for semantic analysis and record-keeping of these web pages. These descriptors are semantic in nature and can be used for the construction of evolving and dynamic web directories for the realization and implementation of data in multiple layers and dimensions. After this, the next step is to apply abstraction to the various layers such that every layer is a further abstracted version of the preceding layer but preserving the relevant, special and characterizing features which would facilitate easy and simple yet accurate search results and in less time since the search would proceed from only the highly abstracted layers to the deeper ones.

Other Methods for Mining and Efficiency Improvement

Methods like building personalized search service and web servers based on the user's web navigation history proves to be of great accuracy. This can be done by profiling using the user's history and then building the database accordingly.

Based on Query Clustering Algorithms

Query clustering is a technique for discovering similar queries on a search engine. Also it is a class of techniques aiming at grouping users' semantically related, not syntactically related queries in a query repository, which were accumulated with the interactions between users and the system[11,12,13,14,17].

In graph-based clustering, a query-page bipartite graph is first constructed with one set of the nodes corresponding to the set of submitted queries, and the other corresponding to the sets of clicked pages. If a user clicks on a page, a link between the query and the page is created on the bipartite graph. After obtaining the bipartite graph, an agglomerative clustering algorithm is used to discover similar queries and similar pages. During the clustering process, the algorithm iteratively combines the two most similar queries into one query node, then the two most similar pages into one page node, and the process of alternative combination of queries and pages is repeated until a termination condition is satisfied.

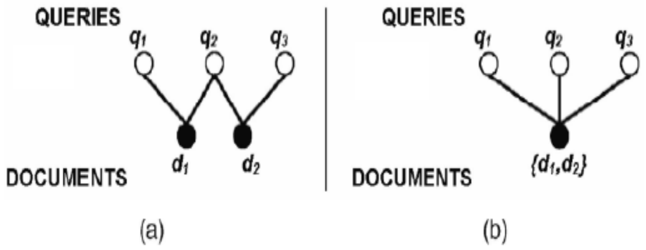


Fig. 1

Queries q1 and q3 seem unrelated before document clustering (Fig 1.a). (Fig 1.b) After document clustering, queries q1 and q3 are then related to each other because they are both linked to the document cluster (d1; d2).

Concept Based Clustering Algorithm – This technique is composed of two steps: 1) Bipartite graph construction using the extracted concepts and 2) agglomerative clustering using the bipartite graph constructed in step 1.

Based on naïve Bayesian classification

A naive Bayesian classifier simply apply Bayes' theorem on the context classification of each search, with a strong assumption that the words included in the searches are independent to each other. In the beginning, we get two samples from the real life searches in order to create the training dataset. Then a detailed filtering process of the naive Bayes classification will be applied on our sample phrases for testing[15].

Based on naïve Bayesian classification with detection and correction

This algorithm is a combination of Bayesian classification and an algorithm that defines three types of word misspelling - typographic, cognitive and phonetic errors and also classifies the words according to certain categories.[15]

Statistical Importance of Search Engine Optimization [16]

- Marketing Sherpa reports distribution lead to a 2,000% increase in blog traffic and a 40% increase in revenue.
- 70-80% of users ignore the paid ads, focusing on the organic results.
- 75% of users never scroll past the first page of search results.
- Search and e-mail are the top two internet activities.
- Companies that blog have 434% more indexed pages and companies with more indexed pages get far more leads.
- 81% of businesses consider their blogs to be an important asset to their businesses.
- A study by Outbrain shows that search is the #1 driver of traffic to content sites, beating social media by more than 300%.
- 79% of search engine users say they always/frequently click on the natural search results. In contrast, 80% of search engine users say they occasionally/rarely/never click on the sponsored search results.
- 42% of search users click the top-ranking link. 8% click the second-ranking link, and the click-through rate (CTR) continues to drop thereof.
- 80% of unsuccessful searches are followed with keyword refinement.

- 41% of searches unsuccessful after the first page choose to refine their keyword search phrase or their chosen search engine.

Experimental Results and Analysis

- Based on Naïve Bayesian Classification:

Word	Frequency in Search 1	Frequency in Search 2
The	185	930
Vehicle	21	58
Wagon	39	19

Table 1 - Training Data Set 1 (for theory, refer section 2.5.2)

Feature	Appearances in Search1	Appearances in Search2
A	165	1235
Advised	12	42
As	2	579
Chance	45	35
Clarins	1	6
Exercise	6	39
For	378	1829
Free	253	137
Fun	59	9
Girlfriend	26	8
Have	291	2008
Her	38	118
I	9	1435
Just	207	253
Much	126	270
Now	221	337
Paying	26	10
Receive	171	98
Regularly	9	87
Take	142	287
Tell	76	89
The	185	930
Time	212	446
To	389	1948
Too	56	141
Trial	26	13
Vehicle	21	58
Wagon	39	19
You	391	786
Your	332	450

Table 2 – Training Data Set 2 (for theory, refer section 2.5.2)

Given were the two samples from the real life searches in order to create the training dataset (See Table 1 and Table 2). Then a detailed filtering process of the naive Bayes classification will be applied on our sample phrases for testing which gives the new table of training data (Table 3) as shown below:

- **Bayesian Classification with Detection and Correction and Classification Algorithm:**

Classifier	Economy	History	Family	Islam	Sport	Health
Only Bayes	66.2	76.4	84.1	86.4	71.9	64
Bayes with D&C	74.6	80.3	89.6	90.2	75.1	69.6

Above is the table consisting of percentage of classification of the searches using the two variants of Bayesian classification. The detection and correction algorithm outperformed the Bayes algorithm by about 10%, without checking misspelling errors accuracy is 68.85%, while the average accuracy for the classification system with misspellings detection and correction is 71.77%. Overall improvement of 4.92% (for theory, refer section 2.5.3)

Feature	In Search1	In Search2	$P(w_{ij}S)$	$P(w_{ij}H)$
A	165	1235	0.3819444	0.5691244
Advised	12	42	0.0277778	0.0193548
As	2	579	0.0046296	0.2668203
Chance	45	35	0.1041667	0.0161290
Clarins	1	6	0.0023148	0.0027650
Exercise	6	39	0.0138889	0.0179724
For	378	1829	0.8750000	0.8428571
Free	253	137	0.5856481	0.0631336
Fun	59	9	0.1365741	0.0041475
Girlfriend	26	8	0.0601852	0.0036866
Have	291	2008	0.6736111	0.9253456
Her	38	118	0.0879630	0.0543779
I	9	1435	0.0208333	0.6612903
Just	207	253	0.4791667	0.1165899
Much	126	270	0.2916667	0.1244240
Now	221	337	0.5115741	0.1552995
Paying	26	10	0.0601852	0.0046083
Receive	171	98	0.3958333	0.0451613
Regularly	9	87	0.0208333	0.0400922
Take	142	287	0.3287037	0.1322581
Tell	76	89	0.1759259	0.0410138
The	185	930	0.4282407	0.4285714
Time	212	446	0.4907407	0.2055300
To	389	1948	0.9004630	0.8976959
Too	56	141	0.1296296	0.0649770
Trial	26	13	0.0601852	0.0059908
Vehicle	21	58	0.0486111	0.0267281
Wagon	39	19	0.0902778	0.0087558
You	391	786	0.9050926	0.3622120
Your	332	450	0.7685185	0.2073733

Table 3 - New table of training data (for theory, refer section 2.5.2)

Based on Query Clustering Algorithms

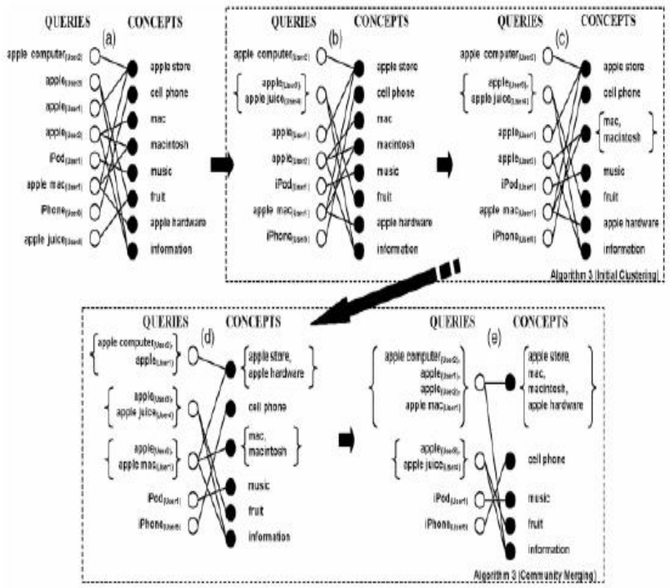
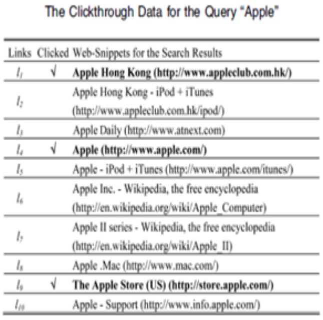


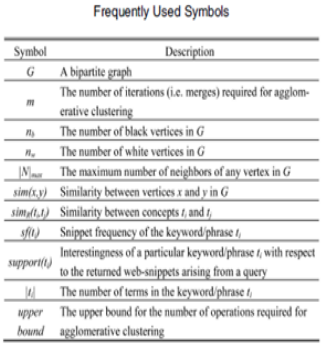
Fig. 2 - Performing personalized concept-based clustering algorithm on a small set of clickthrough data. Starting from top left: (a) the original bipartite graph. (b), (c) initial clustering. (d), (e) Community merging. (For theory, refer section 2.5.1)



The Clickthrough Data for the Query "Apple"

Links	Clicked	Web-Snippets for the Search Results
l_1	√	**Apple Hong Kong (http://www.appleclub.com.hk/)**
l_2		Apple Hong Kong - iPod + iTunes (http://www.appleclub.com.hk/ipod/)
l_3		Apple Daily (http://www.atnext.com)
l_4	√	**Apple (http://www.apple.com/)**
l_5		Apple - iPod + iTunes (http://www.apple.com/itunes/)
l_6		Apple Inc. - Wikipedia, the free encyclopedia (http://en.wikipedia.org/wiki/Apple_Computer)
l_7		Apple II series - Wikipedia, the free encyclopedia (http://en.wikipedia.org/wiki/Apple_II)
l_8		Apple .Mac (http://www.mac.com/)
l_9	√	**The Apple Store (US) (http://store.apple.com/)**
l_{10}		Apple - Support (http://www.info.apple.com/)

Table 4



Frequently Used Symbols

Symbol	Description
G	A bipartite graph
m	The number of iterations (i.e. merges) required for agglomerative clustering
n_b	The number of black vertices in G
n_w	The number of white vertices in G
$\|N\|_{max}$	The maximum number of neighbors of any vertex in G
$sim(x,y)$	Similarity between vertices x and y in G
$sim_R(t_i,t_j)$	Similarity between concepts t_i and t_j
$sf(t_i)$	Snippet frequency of the keyword/phrase t_i
$support(t_i)$	Interestingness of a particular keyword/phrase t_i with respect to the returned web-snippets arising from a query
$\|t_i\|$	The number of terms in the keyword/phrase t_i
upper bound	The upper bound for the number of operations required for agglomerative clustering

Table 5

The above two tables (Table 4 and Table 5) show the clickthrough data for the word "Apple" with respect to user choices and the corresponding symbols and their descriptions of the bipartite graph generated for the clickthrough query data (for theory, refer section 2.5.1).

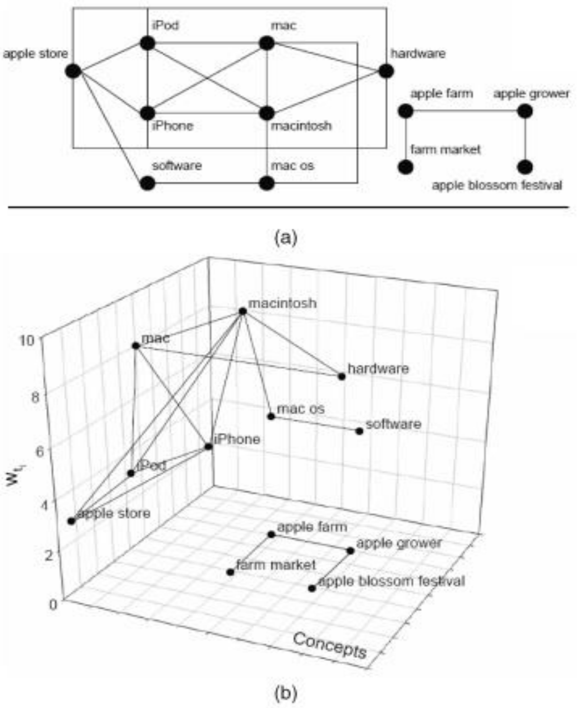


Fig. 3. (a) A concept relationship graph for the query "apple" derived without incorporating user clickthroughs. (b) A concept preference profile constructed using the user clickthroughs and the concept relationship graph in (a). w_{ti} is the interestingness of the concept t_i to the user. More clicks on a concept gradually increase the interestingness of the concept w_{ti}

Conclusion

Data mining for improving the efficiency of search engines will prove to be an effective impetus for research in SEO technology. It will be one that facilitates exhaustive and comprehensive usage of the data available in large databases or on the web. However, it is essential that we overcome the limitations and the challenges in this field that have previously been discussed in order to make a friendlier and more efficient search engine delivering high quality results. Also, we must incorporate the various mining techniques with improved algorithms to ensure even better knowledge dredging.

References

[1] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Proc. 7th Int'l World Wide Web Conf. (WWW7), ACM Press, New York, 1998, pp. 107-117.

[2] J. Srivastava, Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, vol. 1, no. 2, 2000, pp. 12-23.

[3] J. Wen, J. Nie, and H. Zhang, "Query Clustering Using User Logs," ACM Trans. Information Systems, vol. 20, no. 1, pp. 59-81,2002.

[4] S. Chakrabarti Byron E. Dom_ David Gibson, Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, "Mining the Web's Link Structure," Computer, Aug. 1999, pp. 60-67.

[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, Reading, Mass., 1999.

[6] J. Han and M. Kambert, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[7] V. R. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic Segmentation of Text into Structured Records," Proc. ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD 2001), ACM Press, New York, 2001, pp. 175-186.

[8] S. Chaudhuri and U. Dayal, "An Overview of DataWarehousing and OLAP Technology," SIGMOD Record, vol. 26, no. 1, 1997, pp. 65-74.

[9] M. Perkowitz and O. Etzioni, "Adaptive Web-Sites," Comm. ACM, vol. 43, no. 8, 2000, pp. 152-158.

[10]K. Yu et al., "Instance Selection Techniques for Memory-Based Collaborative Filtering," Proc. SIAM Int'l Conf. Data Mining (SIAM 02), ACM Press, New York, 2002, pp. 59-74.

[11] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD, 1993.

[12] B. Goethals and M. Zaki, "Frequent Itemset Mining Implementations," Proc. ICDM Workshop Frequent Itemset Mining Implementations (FIMI), 2003.

[13] B. Koester, "Conceptual Knowledge Retrieval with FooCA:ImprovingWeb Search Engine Results with Contexts and Concept Hierarchies," Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM), 2006.

[14] J. Wen, J. Nie, and H. Zhang, "Query Clustering Using User Logs," ACM Trans. Information Systems, vol. 20, no. 1, pp. 59-81, 2002.

[15] Abdullah Mamoun Hattab and Abdulameer Khalaf Hussein, "Arabic Content Classification System Using statistical Bayes classifier With Words Detection and Correction, "World of Computer Science and Information Technology Journal (WCSIT), vol. 2, no. 6, 193-196, 2012.

[16] www.intraspin.com/news/10-statistics-that-demonstrate-the-value-of-seo

[17] Kenneth Wai-Ting Leung, Wilfred Ng, and Dik Lun Lee, "Personalized Concept-Based Clustering of Search Engine Queries, "IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, November 2008