ISSN No. 0976-5697

Volume 2, No. 1, Jan-Feb 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Search Engine Technique Using New Ranking Algorithm for Web Mining

Nymphea Saraf Sandhu*
Department of C.S.E
Samrat Ashok Technological Institute
Vidisha, (M.P.), India,
nymphea.saraf@yahoo.com

Yogendra Kumar Jain
Department of C.S.E
Samrat Ashok Technological Institute
Vidisha, (M.P.), India,
ykjain_p@yahoo.co.in

Abstract:Nowadays, there are a huge amount of resources on the Web, which raises a serious problem of accurate search. This is because data in HTML files is useful in some contexts but meaningless under other conditions. In addition, HTML cannot provide description of data encapsulated in it. In many situations a flat list of ten search results is not enough, and the users might desire to have a larger number of search results grouped on-the-fly in folders of similar topics. In addition, the folders should be annotated with meaningful labels for rapid identification of the desired group of results. In other situations, users may have different search goals even when they express them with the same query. In this thesis the research results should be personalized according to the users' on-line activities. There are also situations where users might desire to access fresh information. In these cases, traditional link analysis could not be suitable. In fact, it is possible that there is not enough time to have many links pointing to a recently produced piece of information. In order to address this necessity, we discuss the algorithmic and numerical ideas behind a new ranking algorithm suitable for ranking fresh type of information, such as news articles.

Keywords: Web mining, New Ranking Algorithm, Cluster.

I. INTRODUCTION

In order to find useful information, two paradigms are well-established in traditional Information Retrieval. Searching is a discovery paradigm which is useful for a user who knows precisely what to look for, while browsing is a paradigm useful for a user who is either unfamiliar with the content of the data collection or who has casual knowledge of the jargon used in a particular discipline. Browsing and searching complement each other and they are most effective when used together [2].

The goal of a modern Web search engine is to retrieve documents considered "relevant" to a user query from a given collection. Nowadays, a user query is modeled as a set of keywords extracted from a large dictionary of words; a document is typically a Web page, pdf, postscript, doc file, or whatever file that can be parsed into a set of tokens.

Global search engines serve as de facto Internet portals, local search engines are embedded in numerous individual Web sites, and browsing is the most common activity on the Web, due to the hyper-linked structure that provides access to a large quantity of information in a restricted space. In addition to traditional Information Retrieval issues, we may identify at least five specific obstacles which Web searching and browsing must overcome. The first key difficulty in solving the above retrieval problem relies on the characterization of the adjective "relevant". Modern search engines have spent a lot of effort in ranking Web objects, providing valuable access to the information contained on the Internet. A breakthrough piece of technology has been introduced by adopting social network theories and link analysis, such as Page rank and Hits, largely adopted by modern search engines. Nevertheless, in many situations traditional link analysis is not the perfect solution because the problem of ranking search results is inherently complex. One aspect of this complexity derives from the different types of queries submitted to search engines. Narrow topic queries are queries for which very few resources exist on the Web, and which present a "needle in the haystack" challenge for search engines. On the other hand, broad topic queries pertain to topics for which there is an abundance of information on the Web, sometimes as many as millions of

relevant resources (with varying degrees of relevance). Broad topic queries are the most difficult to solve since the vast majority of users show poor patience: they commonly browse through the first ten results (i.e. one screen) hoping to find there the "right" document for their query. In order to find relevant results, another aspect to take into consideration is the spamming phenomenon. With the pervasive use of the Web, it is crucial for business sites to be ranked highly by the major search engines. There are quite a few companies who sell this kind of expertise (also known as "search engine optimization") and actively research ranking algorithms and heuristics of search engines, and know how many keywords to place (and where) in a Web page so as to improve the page's ranking (which has a direct impact on the page's visibility). An interesting book to read on the subject is. Search engine optimization is a legitimate activity, while spamming is a malicious one. Search engines and spammers are engaged in an endless fight, from one side, to improve their ranking algorithms and, from the other side, to exploit them. Unfortunately, it is very difficult to distinguish between optimization and spamming. For instance, one of the difficulties of the fight against spam is that there are perfectly legitimate optimizations (e.g. using synonyms as keywords) that might trigger anti-spam defenses. The second key difficulty is related to the huge quantity of information available. It goes without saying that the quality of the search engines is influenced by the completeness and freshness of the index which should have few outdated pages and broken hyper-links. Unfortunately, the explosion of digital information available on the Web is making it impossible, on the one hand, to index the whole

set of existing Web pages and, on the other hand, to restrict the number of relevant documents to return to the user, since this number grows in proportion to the Web size. The first difficulty can be partially alleviated by the use of meta-search engines which exploit a pool of individual search engines to collect a larger set of relevant answers for a given user query. However the use of multiple sources, each one exploiting its own ranking policy, makes the retrieval of relevant documents even harder to deal with because of the necessity to merge multiple ranked lists into one unique ranked list. The third key difficulty is that the relevance of a document is a subjective and time-varying concept. In fact, the same set of keywords may abstract different user needs that may also vary over time according to the context in which the user is formulating his/her own query. As an example, for the query "search engine" a researcher may be interested in finding scientific papers, whereas a student may be interested into easy-to-read descriptions; nonetheless, the same researcher might be interested in "popular works" in the case that (s)he has to prepare a presentation for a high-school. The fourth key difficulty is that users often aim to find fresh information. This is particularly true in the presence of unanticipated events such as a "Tsunami", or the death of a celebrity, or a terrorist attack. In this case, interest is focused on news articles which cannot simply be ranked by adopting link analysis techniques. In fact, when a news article is posted, it is a fresh type of information with almost no hyper-link pointing to it.

Clustering is a process which receives a set of documents as input, and groups them based on their similarity. It is distinct from classification, in which an a priori taxonomy of categories is available beforehand, and where the documents are placed in their proper category. Clustering, in contrast, is a process where the categories are part of the (discovered) output, rather than part of the input. Clustering is a useful post-processing technique applied in the presence of broad queries. In fact, when no pre-imposed classification scheme is available (like in the heterogenous Web), automatic clustering is crucial for organizing a huge amount of answers in a collection of browsable and dynamically organized hierarchies. The browsable nature of the hierarchy helps in identifying time-varying interests for a given subject. This process has also another important advantage: it can improve the search experience by labeling the clusters with meaningful sentences. These meaningful sentences are an interesting alternative to the flat list of search results currently returned by the most important search engines. In short, clustering can act as a booster which helps the user in solving on-the-fly the polysemy and synonymy problems and extracts hidden knowledge from retrieved texts. Browsing can be used to discover topics to search for, and search results can be organized for a more refined browsing session. The first goal of my MTech Thesis, has been to investigate the use of Web page clustering as an innovative WebIR tool which can help users to search the Web.

Ranking is the process which estimates the quality of a set of results retrieved by a search engine. Traditional IR has developed boolean, probabilistic, or vector-space models, aiming to rank the documents based on the content of the collection. Modern WebIR exploits the link structure of the Web. Note that links provide a positive critical assessment of aWeb page's content which originates from outside of the control of the page's author (as opposed to assessments based on Web page's textual content, which is completely under the control of Web page's author). This makes the information extracted from informative links less vulnerable to manipulative techniques such as spamming. It goes without saying that one can use the large quantity of academic publications about Web Ranking available in literature. Nevertheless, as we already pointed out, many aspects remain to be investigated. The other goal of my MTech Thesis is to study and design new methodologies for fast Web rank computation and to integrate link analysis with models suitable for ranking fresh information [1] and [3]. Clustering and Ranking are WebIR tools linked by a mutual reinforcement relationship. In one direction, a good ranking strategy can provide a valuable base of information for clustering in a dynamically organized hierarchy. In the opposite direction, a good cluster strategy can provide a valuable base for altering the rank of the retrieved answer set, by emphasizing hidden knowledge meanings or solving synonymy and polysemy problems which are not captured by traditional text or link based analysis. In addition, clustering algorithms can be used in order to extract, on the user behalf, a knowledge which goes behind the traditional flat list of about ten results.

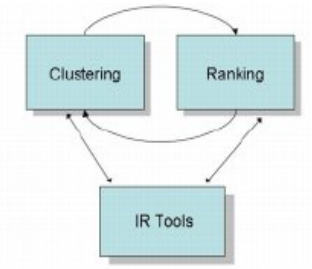

Figure.1 Smart Webir Tools Exploit The Mutual Reinforcement Relation Between Clustering And Ranking Methodologies.

II. BACKGROUND

A. Web Information Retrieval

Information Retrieval (IR) is not a recent discipline. In the 1960's, Gerard Salton developed SMART, an experimental information retrieval system. SMART has been a test–bed for algorithms which perform automatic indexing and retrieval of full–text documents. A lot of theoretical models from natural language processing, statistical text analysis, word–stemming, stop lists, and information theory has been experimented in the system. He showed that the traditional task of IR was to retrieve the most "relevant" set of documents from a collection of documents, for a given query. He also assumed, as usual in traditional IR, that the collection was controlled, in the sense that no document was created for spamming – created with the intent of being selected for un–related queries, relatively small and almost never changing. In 1995 everything changed with the creation of the Web. Web objects are heterogenous since they can be of different types: Web pages, audio, video, news articles, usenet, blogs, to name a few. Web objects are un–controlled collections, in the sense that billions of authors create them independently and, very often, they create them for spamming. In addition, Web

objects are the largest collection of information ever created by humans, and this collection changes continuously when new objects are created and old ones removed. In order to adapt to this changed scenario, a new discipline has been created: Web Information Retrieval. It uses some concepts of traditional IR, and introduces many innovative ones.

B. Web Snippet Clusterin

Cluster analysis has been successfully exploited in statistics, numerical analysis, machine learning and in other fields. The term "Clustering" denotes a wide range of methodologies for identifying hidden common structures in large sets of objects. A cluster is a group of objects whose members are more similar to each other than the members of other clusters. In this case, we say that intracluster similarity is high and intercluster similarity is low. Clustering methods are classified according to four aspects.

a. The Structure: This could be flat (there is no relationship between different clusters), hierarchical (clusters are organized in a tree), or overlapping (objects can be members of more than one cluster).

B. The Indexing Unit: Documents are represented by means of a set of words, a so-called bag of words representation, or by means of sentences where the order of words is taken into account.

c. The Duration: The clustering is either carried out on top of a persistent collection of documents or on top of documents which exist for a very short period, like the set of search results given as an answer to a particular query submitted to a search engine. Several authors call this ephemeral clustering.

d. The Algorithm: It is used to generate the clusters, and could be divisive (starting from a set of objects and splitting them into subsets, possibly overlapping) or agglomerative (starting from individual objects and merging them into clusters).Until a few years ago, persistent clustering was considered the "default" clustering technique, "in normal circumstances, the cluster structure is generated only once, and cluster maintenance can be carried out at relatively infrequent intervals". The ephemeral clustering process organizes the documents in groups, which will survive just for the current session. Nowadays, ephemeral clustering is used by several search and meta-search engines to organize their results in fast browsable groups. Surprisingly, ephemeral clustering has been less studied than persistent clustering in literature.

In Web Snippet Clustering, the original Web pages are not given and the clustering process receives an abstract (called "the snippet" of the page) as its input. Snippet clustering was introduced in a primitive form by Northernlight and then made popular by Vivisimo. The problem consists of clustering the results returned by a (meta-) search engine into a hierarchy of folders which are labeled with variable-length sentences. The labels should capture the "theme" of the query results contained in their associated folders. This labeled hierarchy of folders offers a complementary view to the ranked list of results returned by current search engines. Users can exploit this view navigating the folder hierarchy driven by their search needs, with the goal of extracting information from the folder labels, reformulating another query, or narrowing the set of relevant results. This navigational approach is especially useful for informative, polysemous and poor queries [1], [2] and [4].

C.Web Ranking

Social network theory is concerned with features related to connectivity and distances in graphs, which can be applied to diverse fields such as epidemiology, human relationships and citation indexing, to name just a few. Social network theory is successfully used to study the Web graph, GWeb. Due to its size, modern search engines have spent a lot of effort to rank Web objects and to provide valuable access to the information contained in the Internet [7] and [8].

D.Personalized Web ranking

Using a personalized Web ranking algorithm, the Web pages returned as search results for a user U1 are not the same set of pages returned to the user U2, even when U1 and U2 submit the same query. The goal is to provide search results which vary according to different behaviour, interest or tastes implicitly or explicitly expressed by the users. As an example, one user can be interested to the helicopter "apache", another in the native Americans "apache" population, and yet another in the "apache" web server. The industrial scenario: This consists of on a beta version by Google. In addition, Google allows the users to search also their own Web-search history by offering some additional information about the frequency and the last visit of each search result. A similar service is offered by Yahoo and A9.com. Another interesting proposal is Eurekster which relies on patented learning search technology and patent pending processes that link search algorithms to social networks. However, all of these approaches either need to maintain up-to-date user profiles, possibly defined on a restricted (tiny) set of alternatives, or they require an explicit login which allows the underlying search engine to track the user behaviour. The scientific scenario: The need of search personalization is due to the different types of queries submitted to the search engines. In fact, literature offers many studies on different types of queries submitted to search engines by the users. The author creates a taxonomy of intents that express different search goals. A query can either be informational, as in traditional IR, or transactional, to express an "intent to perform some web-mediated activity", or navigational, to express the need to reach a web site.

a.Extensions to Pagerank: Generalized the Pagerank algorithm to compute flow values for the edges of the Web graph, and a TrafficRank value for each page. An interesting line of research aims to combine Pagerank with temporal information. On the Web, the temporal information for outgoing links is under the control of source pages and is, therefore, susceptible to "cheating". On the other hand, the incoming links reflect the attention a Web page has attracted and seem to be more democratic in their nature, they are also less susceptible to cheating. Among those incoming links, the link emanating from the random surfer's current position can be picked out and treated in a special way. These observations suggest that the probability of the random surfer choosing y when leaving his current page x is a combination of many factors: the freshness f(y) of the target page y, the freshness f(x, y) of the link from x to y, and the average freshness of all incoming links of y. In a similar way, the random jump probability of a target page y is a (weighted) combination of the freshness of y, the

activity of y, the average freshness of the incoming links of y, and the average activity of the pages that link to y. All these considerations lead to a modified version of Pagerank which takes in account temporal information. In addition, there are many Pagerank modifications which consider graphs with different levels of granularity (HostRank, Pagerank on host instead of Web pages), or with different link weight assignments (internal, external, etc.).

Recently, the research community has devoted increasing attention to reduce the computational time needed by Web ranking algorithms. In particular, many techniques have been proposed to speed up the Pagerank algorithm. This interest is motivated by three dominant factors: (1) the Web graph has huge dimensions and it is subject to dramatic updates in terms of nodes and links - therefore the Pagerank assignment tends to become obsolete very soon; (2) many Pagerank vectors need to be computed when adopting strategies for collusion detection (3) many different Pagerank values could be computed for addressing the need of personalized ranking. State-of-the-art approaches for accelerating Pagerank have gone in at least six different directions [1] and [5].

In addition, there is a more general need, since Pagerank has also become a useful paradigm of computation in many Web search algorithms, such as spam detection or trust networks, where the input graphs have different levels of granularity (HostRank) or different link weight assignments (internal, external, etc.). For each algorithm, the critical computation is a Pagerank-like vector of interest. Thus, methods to accelerate and parallelize these kinds of algorithms are very important [6] and [12].

III. PROPOSED TECHNIQUE

The proposed ranking algorithm is obtained by gradually introducing a number of constraints in order to match the requested requirements and is validated by intuitive limit cases:

• It allows the ranking of both the news articles and the news sources which produce information. The algorithm exploits a mutual relationship between news information and news sources: The value of a news article is proportional to how many times it has been replicated by other sources, and the value of a news source is proportional to the output of weighted news items.

• It takes into account the time when the articles have been produced and it dynamically models the importance of the articles just at the moment of time when they are provided as answer to a user query. This condition is modeled using a decay function which is translationally invariant, such as the exponential decay function. Note that other more complicated decay functions can be plugged into our schema, provided that they are translationally invariant.

• It exploits the similarities between articles for creating "virtual links" between pieces of news which share common topics. The intuition is that the more the articles discuss a topic areas, the more important the topic is.

• It acknowledges the importance of news sources that produce many "breaking news" items by giving them a-posteriori rank bonus. The intuition is that a source should be privileged if it is either breaking a story or it is following the story quickly. Moreover, it could potentially privilege news articles produced in the past which have many similar pieces of news produced after them. It should be pointed out that this last feature has not been exploited, since we have decided to privilege the fresher news articles. In a future work, we plan to plug this feature into a model that can exploit it in order to counter-balance the decay of the article importance.In *PageRank* algorithm, the rank of each page is defined as the weighted sum of ranks of all pages having back links or incoming links to the page. Then, a page has a high rank if it has more back links to this page have higher ranks. These two properties are true for *DistanceRank* also. A page having many incoming links should have low distance and if pages pointing to this page have low distance then this page should have a low distance. The above point is clarified using the following definition.

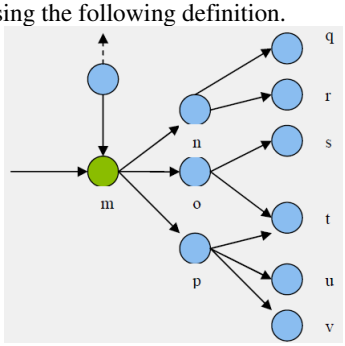

Figure.2 A Sample Graph

*Definition 1***.** If page *a* points to page *b* then the weight of link between *a* and *b* is equal to Log10O(*a*) where O(*a*) shows *a*'s out degree or outgoing links.

Definition 2. The distance between two pages *a* and *b* is the weight of the shortest path (the path with the minimum value) m from *a* to *b*. This is called *logarithmic distance* and is denoted as d_{ab}.

For example, in above figure, the weight of out-links or outgoing links in pages *m*, *n*, *o* and *p* is equal to log(3), log(2), log(2) and log(3) respectively and the distance between *m* and *t* is equal to log(3) + log(2) if the path *m*—*o*—*t* was the shortest path between *m* and *t*. The distance between *m* and *v* is log(3) + log(3) as shown in figure even though both *t* and *v* are in the same link level from *m* (two clicks) but *t* is closer to *m*.

*Definition 3***.** If *dab* shows the distance between two pages *a* and *b* as Definition 2, then *db* denotes the average distance of page *b* and is defined as the following where *V* shows number of web pages:

$$d_b = k/V \qquad \text{where} \quad k = \sum_{a=1}^{v} l \text{ and } l = d_{ab}$$

In this definition, we used an average click instead of the classical distance definition. The weight of each link is equal to log(O(*a*)). If there is no path between *a* and *b*, then *dab* will be set a big value. In this method after the distance computation, pages are sorted in the ascending order and pages with smaller average distances will have high ranking. This method is dependent on the out degree or out going links of nodes in the web graph like other algorithms. Apart from that it also follows the web graph like the random-surfer model used in *PageRank* in that each output link of page *a* is selected with probability 1/O(*a*). That is rank's effect of *a* on page *b* as the inverse product of the out-

degrees of pages in the logarithmic shortest path between *a* and *b*.

A. Clustering technique

The naive clustering used for the first set of tests set σ_{ij} = 1 if ni and nj are the same (i.e. they are mirrored). In our news collection, these cases are very limited. Hence, by using these values of σ_{ij} the news sources' ranks are highly correlated with the simple counting of the posted news articles. A more significant indication of this may be obtained by taking a continuous measure of the lexical similarity between the abstracts of the news posting. These abstracts are directly extracted by the index of the news engine itself. In our current implementation, the news abstract is represented by using the canonical "bag of words" representation and the abstracts are filtered out by a list of stop words. The lexical similarity is then expressed as a function of the common words shared by news abstracts. We point out that dealing with a continuous similarity measure produces a full matrix $\sum$, whose dimensions increases over time, although fortunately the decay rule allows us to consider just the most recently produced part of the matrix, maintaining its size as proportional to the news flow (t, c), and therefore satisfying the Requirement (R). Our research pointed out that the relationship between Clustering and Ranking produces mutual benefits even for news ranking. The above theoretical results show that the better the rank provided by the news sources is, the better the clustering of news articles by similar topics. Besides they show that the better the clustering built on-the-fly by our system is, the better the results of the ranking algorithms.

IV. RESULTS

The ever-growing size of the Web graph and the ubiquitous Pagerank-based ranking algorithms simply that, in the future, the value and the importance of using fast methods for Web ranking will increase. Moreover, the increasing interest in personalized Pagerank justifies the effort required in "pre-processing" the Web graph matrix, so that the many Pagerank vectors needed may be computed more rapidly. The results have a more general application, since Pagerank has also become a useful paradigm of computation in many Web search algorithms, such as spam detection or trust networks, where the input graphs have different levels of granularity (HostRank) or different link weight assignments (internal, external, etc.). Our best result achieves a 65% reduction in Mflops and a 92% reduction in terms of seconds required, compared to the Power method commonly used to compute the Pagerank. As a result, our solving algorithm requires almost a tenth of the time and boost the quality of search results. They have been largely used separately, but their relationship has never been investigated deeply before in literature, as we have done in this Thesis. We believe that clustering is essential in transforming the search experience into a "personalized navigational search experience". The traditional paradigm for personalizing search results is to observe users while surfing the Web and their habits, and infer from those the profile of the user. The personalized navigational search paradigm should be explored further by integrating text hierarchical clustering with an analysis of the query and click logs.

considerably less than half of the Mflops used by the Power method. In view of our results, the approach to speeding up Pagerank computation appears to be much more positive, especially when dealing with personalized Pagerank.

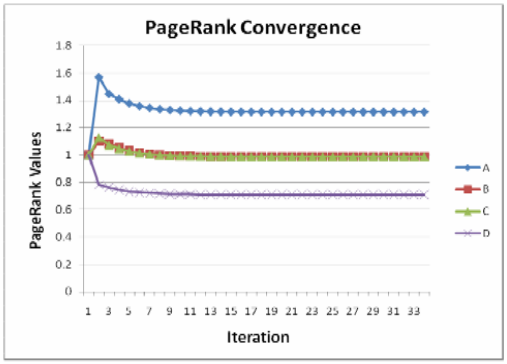

Figure.3 Page Rank Convergence Chart.

V. CONCLUSION AND FUTURE WORK

The Web has become "the place" for accessing any type of information. There are billions of Web pages every day new content is produced. Therefore, the use of search engines is becoming a primary Internet activity, and search engines have developed increasingly clever ranking algorithms in order to constantly improve their quality. Nevertheless, there are still many open research areas of tremendous interest where the quality of search results can be improved. We have shown that clustering and ranking are WebIR tools linked by a mutual reinforcement relationship and that their joint use might further

VI. REFERENCES

[1] Sweah Liang Yong, Markus Hagenbuchner and Ah Chung Tsoi, "Ranking Web Pages using Machine Learning Approaches", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[2] L.A. Barroso, J. Dean, and U. Holzle. Web search for a planet: the google cluster architecture. In IEEE Micro, pages 22–28, 2003, March.

[3] K. Bharat, A. Z. Broder, J. Dean, and M. R. Henzinger. A comparison of techniques to find mirrored hosts on the WWW. IEEE Data Engineering Bulletin, 23(4):21–26, 2000.

[4] K. Bharat, B.W. Chang, M. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between Web sites. In Proceedings of the IEEE International Conference on Data Mining, pages 51–58, San Jose, California, U.S., 2001.

[5] R. Cilibrasi and P. M. B. Vit´anyi. Clustering by compression. IEEE Transactions on Information Theory, 51(4):1523–1545, 2005.

[6] P. Ferragina and A. Gulli. The anatomy of a hierarchical clustering engine for web-page, news and book snippets. In Proceedings of the 4th IEEE

International Conference on Data Mining, pages 395–398, Brighton, UK, 2004.

[7] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. In Journal of IEEE Transactions on Knowledge and Data Engineering, pages 515–528, 2003.

[8] C. Aggarwal, C. Procopiuc, J. Wolf, P. Yu, and J. Park. Fast algorithms for projected clustering. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 61–72, Philadelphia, U.S., 1999.

[9] C. C. Aggarwal, J. Han, J. Wang, and P. Yu. Clustream: A framework for clustering evolving data streams. In Proceedings of the 29th International Conference on Very Large Data Bases, pages 81–92, Berlin, Germany, 2003.

[10] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining *applications. In Proceedings of ACM SIGMOD international* conference on Management of data, pages 94–105, Seattle, Washington, U.S., 1998.

[11] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 207–216, Washington, D.C., U.S, 1993.

A. Anagnostopoulos, A. Z. Broder, and D. Carmel. Sampling search engine results. In Proceedings of 14th International World Wide Web Conference, pages 245–256, Chiba, Japan,2005.

.

VII. AUTHORS BIBLIOGRAPHY

Dr. Yogendra Kumar Jain presently working as head of the department, Computer Science & Engineering at Samrat Ashok Technological Institute Vidisha M.P India. The degree of B.E. (Hons) secured in E&I from SATI Vidisha in 1991, M.E. (Hons) in Digital Tech. & Instrumentation from SGSITS, DAVV Indore (M.P), India in 1999. The Ph. D. degree has been awarded from Rajiv Gandhi Technical University, Bhopal (M.P.) India in 2010. Research Interest includes Image Processing, Image compression, Network Security, Watermarking, Data Mining. Published more than 40 Research papers in various Journals/Conferences, which include 10 research papers in International Journals. **Tel**: +91-7592-250408,**E-mail**:ykjain_p@yahoo.co.in

Ms Nymphea Saraf Sandhu is presently studying in the final semester of M.Tech, Computer Science & Engineering at Samrat Ashok Technological Institute Vidisha, M.P, India. Has obtained degree of B.E. secured in Computers from Pune University in 2002. PGDBA obtained in the field of Operations Management from Symbiosis CDL. Research Interest includes Web Search Engine Techniques, Artificial Intelligence, MIS and Data Mining. **Tel**:+91-9977212333
E-mail:nymphea.saraf@yahoo.com