Genetic Algorithms: an Inevitable Solution for Query Optimization in Web Mining – a Review

L. Melita
Department of Computer Science
Jimma University
Jimma, Ethiopia.
pmmelita@gmail.com

Gopinath Ganapathy
School of Computer Science and Engineering
Bharathidasan University,
Tiruchirappalli, India
gganapathy@gmail.com

Sebsibe Hailemariam
Department of Computer Science
Addis Ababa University, Ethiopia.
sebsibe2004@yahoo.com

Abstract — **Despite the wide range of applications in different areas, the world-urging factors like Internet Search Speed or the rate of data retrieval, has found its wherewithal from the concept of Evolutionary Algorithms. Due to the suitability of the nature of the algorithms, query optimization for web mining, has been implemented in terms of Genetic Algorithms. This paper shall review upon the characteristics of Genetic Algorithms, suitability to web search optimization, their impact on implementation aspects, challenges, success rate and suggested enhancements in further research.**

Keywords — **Genetic Algorithms, Query Optimization, Web Mining, Web Search.**

I. INTRODUCTION

For three decades, many mathematical programming methods have been developed to solve optimization problems. However, until now, there has not been a single totally efficient and robust method to cover all optimization problems that arise in the different engineering fields. The lack of a single method available to deal with multidimensional problems, including those with several goals to optimize, has generated the need to apply numerical processes for optimization. Despite numerous researches on optimization techniques, the typical ones are based on calculus, numerical methods, and random methods. Due to the shortcomings of the calculus based techniques and the numerical ones, random methods have increased their popularity [1].

The methods of random search otherwise known as Evolutionary Algorithms are based on the principles of natural selection of Darwin [2] and the genetic theory of natural selection of R.A. Fisher [3]. In general, all recursive approaches based on population, which use selection and random variation to generate new solutions, can be seen as evolutionary techniques. Genetic Algorithms (GA) is one such method that can handle global optimization problems effectively and this has been a subject of discussion by [4-7] respectively. The genetic algorithms were developed by Holland [8] and the most popular reference is by Goldberg [9] and a more recent one by Bäck [10].

Genetic Algorithms is a versatile tool that can be applied as a global optimization method to problems of web search, as they are easy to implement to non-differentiable functions and discrete search spaces. Genetic algorithm is an example of a search procedure that uses random selection for optimization of a function by means of the parameters space coding. Moreover, genetic algorithms have been proven successful for robust searches in complex spaces. Genetic algorithms are observed to be simple and extremely capable in their task of searching for objective improvement, though they are not just limited to the search space [1].

Turning our focus to Web mining, it is a rapidly emerging field since its inception in or around 1996, and various new methodologies are being developed both using classical and soft computing approaches concurrently. Considering the immense potential application of soft computing to web mining and more specifically the implementation of concepts like Fuzzy Logic, Neural Networks or Genetic Algorithms, it is likely to make more study on them [11].

This paper describes on the relevance of web mining, how web search turns out to be an optimization problem, applications of GA in web mining, suitability of GA to web search, present implementations, challenges and opportunities in GA implementations and amendable suggestions. We shall also analyse upon the effectiveness of Genetic Algorithms and its algorithmic structure, while performing a web search or web mining of data.

II. WEB MINING

Data mining refers to the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable

patterns in data, whereas Web mining is broadly defined as the discovery and analysis of useful information from the WWW. In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization's database. The type (e.g., text, image, audio, symbolic) of collected data may differ depending upon the location of the source and hence the techniques applied for retrieval may vary from data to data.

Since web is a vast collection of completely uncontrolled heterogeneous documents, Web mining techniques are used to automatically discover and extract information from Web documents and services [12]. Presently, this area of research is fairly huge, partly 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. Similar to Etzioni [12], many research articles [13] suggest decomposing Web mining into the following subtasks, namely:

1) Information Retrieval/ Resource finding: the task of retrieving intended Web documents.
2) Information Selection/ Extraction and pre-processing: automatically selecting and pre-processing specific information from retrieved Web resources.
3) Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
4) Analysis: validation and/or interpretation of the mined patterns.

Since our attention is towards the application of GA in Web Mining, we shall discuss the issues and challenges on GA. Genetic Algorithms has found its applications in many areas of Web mining like inducing classification rules [14], applying GA directly as a classifier [15], to optimize the prediction accuracy over raw classification [16], for obtaining extraction rules [17] or patterns [18], to determine the best initialization of clusters as well as the optimization of the initial parameters during clustering [19], to obtain fast and scalable multi-objective association rule mining technique [20] etc.

Among the above-said subtasks, the focus of the present review falls on the task of retrieving data from Web documents, which is otherwise predominantly known as "Web Search". As the web continues to increase in size, the relative coverage of web search engine is decreasing, and search tools that combine the results of multiple search engines are becoming more valuable. Moreover, the necessity of improving the web search is emergent due to retrieval of low quality data, problems with integrating data, problems due to lack of structure of data, high user requirement of ready-made and appropriate data, coverage and convergence to voluminous data from several servers etc. These reasons make it mandatory to use metasearch engines or search engines in retrieving quality pages more appropriate to the user request.

A metasearch engine searches the web by making requests to multiple search engines such as Alta vista, Yahoo, etc. Results of the individual search engines are combined into a single result set. The Advantage of metasearch engines includes a consistent interface to multiple engines and improved coverage. Performing genetic search shall improve the effect of metasearch engines [21]. Genetic search is characterized by the fact that a number (N) of potential solutions of an optimization problem simultaneously samples the search space.

We assume that it is possible to perform additional computation to the result from standard search engines, a point that is lacking to standard search engines. This may consist of instance in formulating a "richer" request to download the pages in order to well analyze their content [22], to propose a textual clustering of the results [23], and to perform additional search with a given strategy [24]. We deal in this paper with the last point and we make use of the optimality of genetic algorithms [8], with respect to finding the most interesting pages for the user.

III. WEB SEARCH AS AN OPTIMIZATION PROBLEM

When one is trying to look for documents or files (web search), one is searching the space of all possible solutions for the most appropriate one which fits in the given query request. Any search task has several components. One needs to define a search space, the abstract form or structure of all possible entities that are being searched through and an evaluation function, a precise way to evaluate any member of this search space and decide on its "quality" how good or useful a solution it is.

It can be seen that web search can also be formulated as a standard optimization problem similar to the problem of function optimization. Recent statistical studies have modelled the web as a graph in which the nodes are web pages and the edges are the links that exist between these pages [25, 26].

The WWW is an information environment made of a very large distributed database of heterogeneous documents, using a wide area network (WAN) and a client server protocol. The structure of this environment is that of a graph, where nodes (web pages) are connected by edges (hyperlinks). The typical strategy for accessing information on the WWW is to navigate across documents through hyperlinks, retrieving the information of interest along the way.

According to [21], the search space S of the optimization problem is the set of web pages and is structured with neighbourhood relationship V: $S \rightarrow Sk$ (since there may be k outgoing (or incoming) links, such that k is drawn from the power-law distribution [25]). We associate to S an evaluation or fitness function, which can numerically evaluate web pages. A search engine tries to output pages, which maximizes this function, and thus tries to solve the optimization problem. To scan S, optimization algorithms and search engines make use of search operator say creation operator that initialize points from S and operators that will modify existing points in the population.

Querying standard search engine performs the creation of individuals. In the web context, random creation operator to randomly generate IP addresses has already been studied [22], which may not be suitable for web search as it gives a lower chance for a valid web server at some instances. Many search

engines, either based on a metasearch or agents, use a heuristic that builds a solution from the description of the problem, by querying one or more index-based search engines and outputting the obtained links, thereby creating the initial generation of the population.

Web robots and more generally web agents [27, 28], explore the links found in pages or a standard heuristic in optimization such as hill climbing is directly adapted to the web, starting from a given page to explore its links and select the best one according to the presented fitness function F in order to define a new starting point.

IV. GENETIC AGORITHMS FOR WEB MINING

The Genetic Algorithm (GA) is an optimization algorithm which simulates biological evolution according to the principle of survival of the fittest. GAs are executed iteratively on a set of coded solutions (genes), called population, with three basic operators: selection, crossover and mutation. They use only the payoff (fitness function) information and probabilistic transition rules for moving to the next iteration.

The first step to model this problem as a GA problem, is determining the chromosome, GA operators, and fitness function. The process is better described in Figure 1. An optimizer cost model includes cost functions to predict the cost of operators, and formulas to evaluate the sizes of results. Cost functions can be expressed with respect to either the total time, or the response time [29]-[36]. The total time is the sum of all times and the response time is the elapsed time from the initiation to the completion of the query.

Algorithm 1 explains the optimization process during web search. An individual in the population is a web page that can be numerically evaluated with a fitness function. Initially, the first individuals are mostly generated with a heuristic creation operator which queries standard search engines to obtain pages. Then, the individuals can be selected/deleted according to their fitness, and can give birth to offspring with selection/crossover operators.

The crossover operator with probability of crossover Pc, is performed by selecting two parent individuals (web pages) from the population. It chooses one crossover position within the page randomly and exchanges the links after that position between both individuals (web pages). From an intuitive point of view, the behavior of this search algorithm can range from a meta-search engine (with Pc=0) which only analyzes/evaluates the results of standard search engines, to a search engine which explores in parallel as many local links as possible (Pc=1) with the help of selective pressure to guide the search through the links.

Here the fitness function depends upon the link quality and hence the page quality, where the link quality is given by the mean number of occurrences of the input strings in the given link and the page quality is determined from the total number of links per page. Our approach models the problem at a level that is closer to the fitness landscape. The GA search does not optimize the parameters of searching agents but rather directly deals with points in the search space.

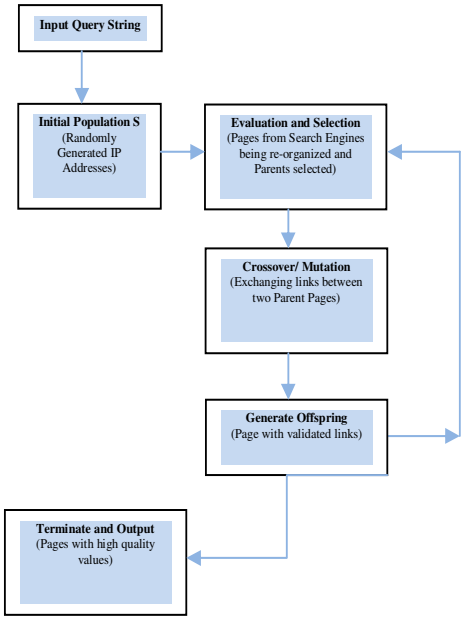

Figure 1: Web Search as a GA Optimization Problem

Algorithm 1

Input: Query String to Search

Output: Pages with links related to query string

1. **Initialize** Population with Randomly generated IP Addresses or links from standard search engines or web servers.
2. **Evaluate** the Page quality by identifying the quality of each link retrieved (weigh them).
3. **Select** the Parent Pages from different Servers.
4. **Crossover** the Pages by exchanging links, in order to generate offspring with improved page quality.
5. **Repeat** from step 2 until a complete output list of pages are ready or the iteration size reaches the maximum value.
6. **Terminate** and sort the pages based on page quality in descending order and output the resulting pages.

V. DISCUSSION

Algorithm 1 can be implemented for a meta-search engine or a search engine exploring many local links since it controls the intensity with which pages are explored. Other applications of GAs to the problem centered on the web are for instance [27, 28, 37- 40].

GAs are appropriate for web mining according to [41], since the space of all possible solutions is very large, the solution space is high-dimensional, the space is deceptive, the problem contains nonlinearities and constraints and there is no known analytical way to solve the problem.

As far as observed, the literature explaining the use of Genetic Algorithms to web mining especially for Information Retrieval seems to be even poorer than that of Fuzzy Logic and Neural Networks [11]. However, Genetic Algorithms are used effectively mainly in search, optimization, and description. Moreover, among the Web mining techniques, Genetic Algorithms could also be effectively used to create index terms for the Web search services.

Based on the experimental results, it has been observed that Genetic and randomized algorithms [30], [42] do not generally produce an optimal access plan for query optimization. But in exchange they are superior to dynamic programming in terms of running time. Experiments have shown that it is possible to reach very similar results with both genetic and randomized algorithms depending on the chosen parameters. Still, the genetic algorithm has in some cases proved to be slightly superior to randomized algorithms. There is no doubt that dynamic programming method always gives us optimal solution, however, the time and space complexity of the GA-based optimization is much less than the other algorithms.

VI. GA-BASED SEARCH IMPLEMENTATIONS – CHALLENGES AND OPPORTUNITIES

A GA-based search to find other relevant homepages, given some user-supplied homepages, has been implemented in G-Search [43]. In [44], Boughanem et al. developed a query reformulation technique using Genetic Algorithms, in which a GA generates several queries that explore different areas of the document space and determines the optimal one. Yang et al. [45] presented an evolutionary algorithm for query optimization by reweighing the document indexing without query expansion. Kraft et al. [46] apply genetic programming in order to improve weighted Boolean query formulation.

The web keeps creating new challenges to different component tasks of web mining as the amount of information on the web is increasing and changing rapidly without any control. The difficulties due to the inherent subjectivity, imprecision, and uncertainty related to user queries may be encountered during retrieval.

Query processing in search engines is performed as a simple keyword matching. This does not take into account the context and relevance of queries with respect to documents, while these are important for efficient machine learning. The current search engines have no deductive capability too. Current query processing techniques follow the principle of hard rejection while determining the relevance of a retrieved document with respect to a query. This is not correct since relevance, itself, is a "gradual" property of the documents [47], not a crisp one.

Let us consider here again the case of the popular search engine Google [48]. It computes the rank of a page using the damping factor, which has pages pointing to it, and is the number of outgoing links from page. Note that it takes into consideration only the popularity of a page (reputation of incoming links) and richness of information content (number of outgoing links) and does not take care of other important factors like, User preference, Validity and Interestingness of the user.

It is desirable, for convenience, to get the pages ranked with respect to "relevance" to user queries. The scheme for determining page ranks should incorporate 1) weights given to various parameters of the hit like location, proximity, and frequency; 2) weight given to reputation of a source, i.e., a link from yahoo.com should carry a much higher weight than a link from any other not so popular site; and 3) ranks relative to the user. Current IR systems are not able to index all the documents present on the web and this leads to the problem of "low recall."

In a GA-implementation, it shall be more effective if a local search is performed for the population of pages than just evaluating with the fitness function in order that time consumption for query optimization shall be considerably reduced when voluminous pages of data has been incrementally included. The search engine users' sequential search behavior suggested by Zhiyong Zhang et. al [49] can be utilized to form the basis for the search engine's own query refinement process, which can be exploited to learn useful information that helps generate related queries. This method can be combined with a traditional text or content based similarity method to compensate for the shortness of query sessions and sparsity of real query log data.

The GA-implemented search engines and even others can be improvised if the representation of hyperlinks is enhanced and rather than making simple retrieval of raw data from various web servers, it shall be beneficial for the metasearch engines if more other technical aspects are followed to the documents uploaded to be helpful for effective optimization. The suggestions are listed here (i) inclusion of annotations to the web documents (ii) indexing web pages incrementally (iii) inclusion of page rank based on the user preference, validity, interestingness, richness of the content and the popularity of the page respectively.

VII. CONCLUSION

As far as web search is required, optimizing the method of data retrieval is also to be considered. Although developing a robust cost metric is elusive, building extensible enumeration architecture is a significant undertaking. Despite many years of work, significant open problems remain, like search difficulty due to rapid increase in web information, simple key-word matching query processing, ranking pages with respect to user preference, validity, interestingness too in addition to page popularity and content relevance. However, it is necessary to

note that improving GA by local search on the population of pages shall considerably reduce the search time consumption.

REFERENCES

[1] A. Rangel-Merino, J. L. López-Bonilla, R. Linares y Miranda, Optimization Method based on Genetic Algorithms, Apeiron, Vol. 12, No. 4, October 2005.

[2] C. Darwin, The Origin of the Species, Cambridge, Ma., Harvard University Press, 1967.

[3] R.A. Fisher, The Genetical Theory of Natural Selection. Clarendon press, Oxford 1930.

[4] Carlos D. Toledo, Genetic Algorithms for the numerical solutions of variational problems without analytic trial functions, arXiv:Physics/0506188, pp. 1-3, June 2005.

[5] J. Holland, Genetic Algorithms, Sci. Am. pp.114-116, 1992.

[6] T. Bäck and H. P. Schwefel, An Overview of Evolutionary Algorithms, Evolutionary Comput. 1: pp. 1-23, 1993.

[7] Allen B. Tucker (Jr.), The Computer Science and Engineering Handbook, CRC Press, USA, pp. 557-571, 1997.

[8] J.H. Holland, Adaptive in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975.

[9] D.E. Goldberg, Genetic Algorithms, in Search, Optimization & Machine Learning. Addison Wesley, 1997.

[10] T. Bäck, Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, N.Y., 1996.

[11]Sankar K. Pal, Varun Talwar and Pabitra Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, IEEE Transactions On Neural Networks, Vol. 13, No. 5, September 2002.

[12] O. Etzioni. The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65–68, 1996.

[13] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, SIGKDD Explorations, Volume 2, Issue 1, July 2000.

[14] Saleh Mesbah Elkaffas, Ahmed A. Toony, Applications of Genetic Programming in Data Mining, World Academy of Science, Engineering and Technology 17 2006.

[15] Bandyopadhyay, S., and Muthy, C.A., Pattern Classification Using Genetic Algorithms, Pattern Recognition Letters, (1995).Vol. 16, pp. 801-808.

[16] Behrouz Minaei-Bidgoli, William F. Punch III, Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System, Genetic Algorithms Research and Applications Group (GARAGe), Department of Computer Science & Engineering, Michigan State University

[17] Kyung-shik Shin, Kyoung-jae Kim, Ingoo Han, Financial Data Mining Using Genetic Algorithms Technique: Application to KOSPI 200, Graduate School of Management, Korea Advanced Institute of Science and Technology.

[18] Susan M. Bridges, Rayford B. Vaughn, Fuzzy Data Mining And Genetic Algorithms Applied To Intrusion Detection, National Information Systems Security Conference (NISSC), October 16-19, 2000, Baltimore, MD.

[19] J.F. Jimenez, F.J. Cuevas, J.M. Carpio, Genetic Algorithms applied to Clustering Problem and Data Mining, Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization, Beijing, China, September 15-17, 2007.

[20] S. Dehuri, A. K. Jagadev, A. Ghosh, R. Mall, Multi-objective Genetic Algorithm for Association Rule Mining Using a Homogeneous Dedicated Cluster of Workstations, American Journal of Applied Sciences 3 (11): 2086-2095, 2006, ISSN 1546-9239, 2006 Science Publications.

[21] M. H. Marghny, A. F. Ali, Web Mining Based On Genetic Algorithm, AIML 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt.

[22] Lawrence S. and Giles C.L., 1999b, Text and image meta-search on the web, International Conference on Parallel and Distributed Processing Techniques and Application, 1999.

[23] Zamir O. and Etzioni O, 2000, Grouper: a dynamic clustering interface to web search results, Proceedings of the Ninth International Worldwide Web Conference, Elsevier, 2000.

[24] F.Picarougne, N.Monmarche, A.Oliver, G.Venturini, Search of information on the Internet by evolutionary algorithm, 2002.

[25] Albert R, Jeong H., BarabasiA.-L. 1999, Diameter of the Worldwide Web. Nature, 401:130-131, 1999.

[26] Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S. State R., TomkinsA. And Wiener J. 2000. Graph structure in the Web, Proceedings of the Ninth International Worldwide Web Conference, Elsevier, 2000.

[27] Menczer F, Belew R.K., Willuhn W. Artificial life applied to adaptive information agents. Spring Symposium on Information Gathering from distributed, HeterogeneousDatabases, AAIPress, 1995.

[28] MoukasA, Amalthea. Iinformation discovery and filtering using a multiagent-evolving ecosystem. Applied Artificial Intelligence, 11(5):437-457, 1997.

[29] M. Tamer Özsu, Patrick Valduriez, “Principles of Distributed Database Systems, Second Edition”, Prentice Hall, ISBN 0-13-659707-6, 1999

[30] Kristina Zelenay,“Query Optimization”, ETH Zürich, Seminar Algorithmen für Datenbanksysteme, June 2005

[31] Yannis E. Ioannidis and Youngkyung Cha Kang, “Randomized Algorithms for Optimizing Large Join Queries”

[32] Michael Steinbrunn, Guido Moerkotte, Alfons Kemper, “Heuristic and Randomized Optimization for the Join Ordering Problem”,

The VLDB Journal - The International Journal on Very Large Data Bases, Volume 6, Issue 3 (August 1997), Pages: 191-208, ISSN:1066-8888

[33] D. Kossman, "The state of the art in distributed query processing" (ACM Computing Surveys, ISSN:0360-0300, 2000, Volume 32 , Issue 4, December 2000, Pages: 422 - 469

[34] P. Griffiths Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, T. G. Price, "Access path selection in a relational database management system", Morgan Kaufmann Series In Data Management Systems, Readings in database systems (3rd ed.), Pages: 141 – 152, 1998, ISBN:1- 55860-523-1

[35] Stocker, Kossman, Braumandl, Kemper , "Integrating semi join reducers into state of the art query processors", (ICDE 2001)

[36] Stefano Ceri, Giuseppe Pelagatti, "Distributed Databases: Principles and Systems", Mcgraw-Hill, ISBN-10: 0070108293, ISBN-13: 978- 0070108295, 1984.

[37] Fan W., Gordon M.D., Pathak P. Automatic generation of a matching function by genetic programming for effective information retrieval, Proceeding of the 1999AmericasConference on Information Systems, pp49-51.

[38] Monmarché N., Nocent G., Slimane M. and Venturini G. Imagine: a tool for generating HTML style sheets with an interactive genetic algorithm based on genes frequencies. 1999IEEE International Conference on Systems, Man, and Cybernetics (SMC'99), Interactive Evolutionary Computation session, October 12-15, 1999, Tokyo, Japan.

[39] Morgan J.J. and Kilgour A.C. Personalizing information retrieval using evolutionary modeling, Proceedings of Poly Model Applications of Artificial Intelligence, ed by A.O. Moscardini and P. Smit h, 142-149, 1996.

[40] Sheth B.D. A learning approach to personalized information filtering,. Master's thesis, Department of Electrical Engineering and Computer Science, MIT, 1994.

[41] Yannis E. Ioannidis and Youngkyung Cha Kang, "Randomized Algorithms for Optimizing Large Join Queries".

[42] Michael Levin, Design and Implementation of Genetic Algorithms for Solving Problems in the Biomedical Sciences, Genetics Department, Harvard Medical School, 200 Longwood Ave., Boston.

[43] F. Crestani and G. Pasi, Eds., Soft Computing in Information Retrieval: Techniques and Application. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50.

[44] M. Boughanem, C. Chrisment, J. Mothe, C. S. Dupuy, and L. Tamine, "Connectionist and genetic approaches for information retrieval," in Soft Computing in Information Retrieval: Techniques and Applications, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50, pp. 102–121.

[45] J. J. Yang and R. Korfhage, "Query Modification Using Genetic Algorithms in Vector Space Models," Dept. IS, Univ. Pittsburgh, Pittsburgh, PA, TRLIS045/1 592 001, 1992.

[46] D. H. Kraft, F. E. Petry, B. P. Buckles, and T. Sadasivan, "The use of genetic programming to build queries for information retrieval," Proc.IEEE Symp. Evol. Comput., 1994.

[47] C. V. Negotia, "On the notion of relevance in information retrieval," Kybernetes, vol. 2, no. 3, pp. 161–165, 1973.

[48] S. Brin and L. Page, "The anatomy of a large scale hypertextual web search engine," in Proc. 8th Int. WWW Conf., Brisbane, Australia, Apr. 1998, pp. 107–117.

[49] Zhiyong Zhang, Olfa Nasraoui, **"**Mining search engine query logs for social filtering-based query recommendation", Applied Soft Computing, Volume 8, Issue 4, September 2008, Pages 1326-1334.