# A Data Mining Approach for User Search Goals with Hybrid Evolutionary Approach.

[1]S.L.M.Deepthi, [2] R.G. Chaitanya Nath
[1]slmdeepthi8@gmail.com
[2]chaitu.rg89@gmail.com

*Abstract:* **Search engine optimization is an interesting research arena in the field of data search for retrieving users desired results at the stroke of their keywords. Understanding, satisfying and meeting the user search goal for a specific query is a complicated task, as there is surplus amount of related and unrelated data available over the Internet. Therefore, In this paper, we proposed an pragmatic model of search mechanism with FP (Frequent Patterns) Tree for finding frequent and sequential use of patterns (Urls) obtained from the Click through rate and evolutionary algorithms for prime results with efficient feedback sessions (based on query clicks) which are obtained and simulated from user click-through logs and can efficiently reflect the information needs of users thus reducing the un-relevant results for a specific search.**

*Keywords*: **User search goals, Feedback Sessions, FP growth, Genetic Algorithms.**

## 1. INTRODUCTION

In internet search applications, queries are submitted to search engines to represent the information needs of the users. However, sometimes they may not exactly represent users with the specific information that is needed since many ambiguous results may cover a vast topic and different users may want to get information on different aspects when they submit the same query.

Identifying user information needs is one of the fundamental and major issues in the development of web search engines. It makes it even more challenging as most of the web users submit short queries and try to get the intended results and often end up with no successful results for that particular query. Term suggestion is a basic kind of information retrieval method used by almost all search engines that attempts to suggest relevant terms of user queries to help formulate more effective queries automatically. If these suggested terms are highly related and well organized, it provides less active but more comprehensive aids for users. Traditional approaches to this technique are based on occurrence of key terms from retrieved documents that are highly ranked. Such approaches are said to be document based approaches. Comparative issues with these approaches are the documents which are relatively high ranked should not necessarily be the relevant result to the target search. The proposed another possible set of approach system can be of log-based approach dependent on click-through raw data obtained from the end users. It is a method of recording every user's click-through data that comprises different queries through the search engine and the end landing pages that the corresponding user visits or clicks out of all those obtained results from the search engine. There are few challenges in this approach as well as the end user usually will be compelled to the results with high ranked with most frequent hits which can ultimately again land the user in the most unwanted arena of his search. This results in most of the url's being associated to certain queries leaving the users to take some more time and precise search keywords and its associated methods to get the required result.

In previous works, feedback sessions are extracted along with respective pseudo documents and clustered based on their similarity. Cosine similarity is the measure used to compute similarity based on the frequency of keywords in the document chosen. K-means algorithm is proposed but we identified its drawback in prior specification of number of clusters and random selection of centroid which is not suitable for different density of objects[3].

An efficient pattern based technique for identifying the interesting patterns of clicked url's according the user request is proposed. Feedback session log maintains the user queries, session ids and url's. Initially our approach searches the session oriented results for input query and finds the frequent patterns with FP growth algorithm by constructing the FP tree. After the generation of the frequent patterns,

© 2014 IJAIR. All Rights Reserved 

patterns can be forwarded to evolutionary approach for extraction of optimal patterns from the FP tree generated patterns.

## 2. LITERATURE SURVEY

Several years of simulative research has been done in the field of SEO (Search Engine Optimization) by different researchers and proposed their own methods of approaches in which every method had its own advantages and advancements along with their drawbacks. Most of the search engines commonly work on the basis of score, relevant time stamps and click graphs of the queries. Linguistic comparisons of keywords, cache capturing and implementations for ideal performance and localization has been the latest trends of concepts that are being used in search engine technology developments.

"Hsiao-Tieh Pu" has proposed the traditional log based and term based approaches. The keyword that is matched is found by the term based approach with its synonyms from that of the log and the documents relevant are retrieved on the basic functionality of the frequency of the keyword or terms [1] and an Agglomerative graph that is based on clustering approach proposed by "Doug Beeferman" and "Adam Berger" over query log for clustering the relevant data [2].

File relevance score is simulated on the basis of term frequency depending on the number of occurrences of a particular keyword in a document and also on the inverse document frequency parameters, this approach mainly focuses on the frequency of the keywords and but not much with the time stamps of the document. Therefore, though the document is user relevant no priority is given to it.

Time stamp approaches functions with recent time of the documents uploaded with respect to the file relevance score. Clustering techniques provides better results as is combines similar objects on the basis of time stamp and file relevance scores of the documents [8].

Many search engines works with query clicks, which are based on earlier user's results or urls, the logs of urls along with the keywords are maintained by the server that extracts query oriented outputs for the user related and relevant results.

Heasoo Hwang et al proposed query grouping mechanism that groups similar queries of the users that computes relevant group queries that are represented in terms of graphs known as Fusion Graphs. Fusion Graph is a combination of query reformulation and query click graph. In this approach, the query is compared with the set of matches which are previously accessed frequent queries[6][7].

Lee et al [13] proposed automatic identification of user search goals and stated that majority of queries have predictable goal and treated user goals as "Navigational" and "Informational" and had categorized search queries into these two classes. X-Wang et al proposed clustering of search results which organizes and allows a user to navigate into relevant documents very quickly. Li et al. [14] defined query intents as "Product intent" and "Job intent" and intended to classify queries in accordance to the defined intents.

## 3. METHODOLOGY

The proposed method is an effective mechanism to meet the satisfaction of the user search goals on the basis of user search history sessions that in turn would be based on click logs concerning to the user query.

Each feedback session of the query is initially extracted from the user click through logs and mapped to Pseudo-documents. An efficient evolutionary approach called as Genetic Algorithm is then integrated to the earlier pattern mining approaches for prime results. In this approach, crossover (that is combining existing patterns to generate a new pattern) and mutation (that alters the url of the pattern) are performed on mined patterns (sequence of urls) of the end users to obtain relevant results for the prime patterns in the mined patterns.

### 3.1 FP-Growth Algorithm

FP Growth is called as a Frequent Pattern Growth. It is an efficient scalable technique for mining frequent patterns in a database.

This technique requires two scans to the database.

**Step1**: To build a compact and firm data structure called FP Tree.

**Step 2**: To extract sets of frequent items precisely from the FP Tree.

### 3.2 Genetic Algorithm

© 2014 IJAIR. All Rights Reserved 

Here, we showcase both positive and negative association rules. Positive association rule considers all current items in the transaction. Negative association rule considers every possible item absent in the transaction but not all items within the transaction. The association rule represents the chromosome structure of the resultant chromosomes which are encoded in Genetic Algorithm

Genetic algorithm uses binary encoding and permutation encoding. For the current algorithm we used binary coding which consists of two bits 0 and 1. 0 is represented for an item that is absent and 1 is represented for an item that is present.

There are two methods in genetic algorithm known as Crossover and Mutation.

Crossover is a method of selection of a random gene with respect to the length of the chromosomes and swaps all the genes after that point. Mutation is a method that alters the contemporary solutions such that it adds stochasticity for better solutions in the search. This is where there is a chance that a bit within a chromosome can be flipped from either 0 to 1 or vice versa. After the completion of crossover and mutation, calculation of completeness, confidence factor and fitness of the chromosomes are needed .

Confidence Factor,

$$CF = TP / (TP + FN)$$

Completeness factor is used to simulate and compute the fitness function

$$Comp=TP/ (TP+FP)$$

TP = True Positives = Number of examples that satisfies item set A and item set B

FN = False Positives = Number of examples that satisfies item set A but not item set B

FP = False Negatives = Number of examples that would not satisfy item set A but satisfies item set B

TN = True Negatives = Number of examples neither satisfy item set A nor item set B.

On calculation of these values, we can find out the confidence factor and fitness function and accordingly if the fitness function value is greater than that of the minimum threshold, then that chromosome is optimized chromosome.

**Fitness=CF *Comp**

With more usage of data mining techniques and development of its tools much work is being focused on finding more negative patterns that can provide far more valuable information. Nevertheless, mining of negative association rules is a challenging task, as the matter of fact that there are fundamental differences between positive and negative association rule mining. Mining association rules is not full of reward until it can be utilized to improve decision-making process of an organization. We can hereby, find all valid positive and negative association rules in support using an FP-Growth Algorithm.

## 4. EXPERIMENTAL RESULTS

Fig1 illustrates the outcome and the duration submitted to the search engine of the data set for user's query. The input query here is education and the minimum duration is 2seconds. Therefore, it comes up with all the relevant feedback sessions available to satisfy the query along with the URL, unique id, session id and its duration. Feedback sessions are responsible for generation of the patterns.

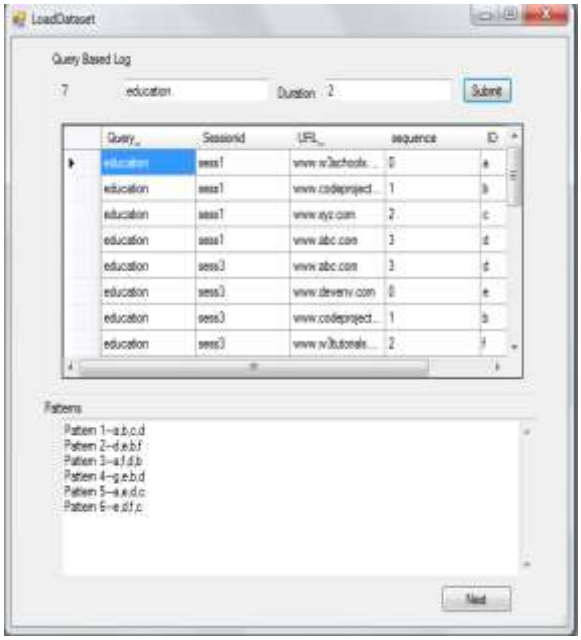


**Fig1: Data Set**

Fig2 illustrates the frequent patterns in the data sets from the available patterns. The output of this FP-Growth is fed as input to the Genetic Algorithm for optimized results.

© 2014 IJAIR. All Rights Reserved


**Fig2: Frequent Patterns**

Fig3 illustrates the result of operations after crossover and mutation on chromosomes. A genetic based research framework is used to discover optimal frequent patterns. This approach prunes the most frequent URL's.

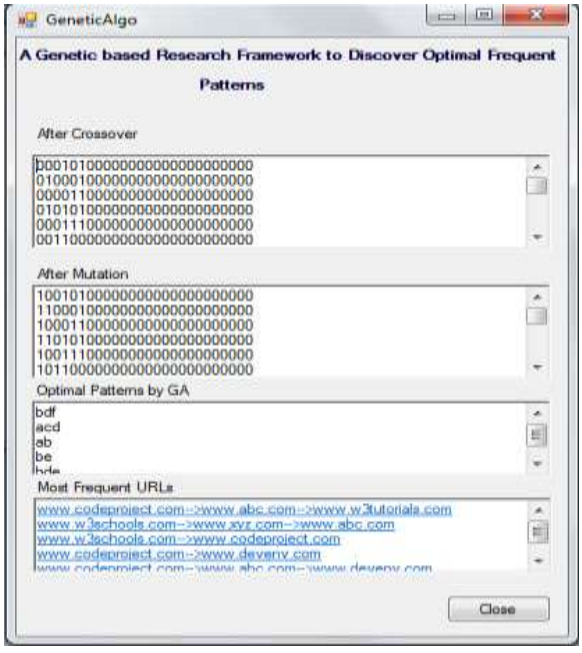


**Fig3: Optimized patterns after Genetic Algorithm.**

## 5. CONCLUSION

The proposed method emphasis on our result oriented research with effective and efficient pattern mining with genetic approach for fulfilling the search goals of the user. FP Growth Algorithm explores and finds out the frequent pattern of URL's in accordance with the query of the end user. Thus generated patterns are simulated and forwarded to the evolutionary approach for fitness computation after crossover and mutation operation over chromosomes.

## *6. REFERENCES*

**1)** Hsiao-Tieh Pu, Hsin-Chen Chiao, "Web Relevant Term Suggestion Using Log-based and Text based Approaches", December 2006.

**2)** Doug Beeferman, Adam Berger, "Agglomerative clustering of a search engine query log".

**3)** Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin , "A New Algorithm for Inferring User Search Goals with Feedback Sessions" IEEE Transactions on Knowledge and Data Engineering , Vol.25, No.3,March 2013.

**4)** Jiawei Han, Jian Pei, and Yiwen Yin, "Mining Frequent Patterns without Candidate Generation".

**5)** Melanie Mitchell, "An introduction to Genetic Algorithms".

**6)** Heasoo Hwang, Hady W. Lauw, Lise Getoor, and Alexandros Ntoulas, "Organizing User Search Histories" IEEE Transactions on Knowledge and Data Engineering, Vol.24 , No.5, May2012.

**7)** C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

**8)** Wisam Dakka, Luis Gravano, and Panagiotis, G. Ipeirotis , "Answering General Time-Sensitive Queries".

**9)** T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

**10)** T. Joachims, L. Granka, B. Pan "Accurately Interpreting Clickthrough Data as Implicit Feedback, "Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

**11)** R. Jones and K.L. Klinkner, "Beyond the Session Timeout:Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management(CIKM '08), pp. 699-708, 2008.

© 2014 IJAIR. All Rights Reserved 

**12)** R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," Proc. 15th Int'l Conf. World Wide Web (WWW '06),pp. 387-396, 2006.

**13)** U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05),pp. 391-400, 2005.

**14)** X. Li, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08),pp. 339-346, 2008.

© 2014 IJAIR. All Rights Reserved 