ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 423- 426 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

A Genetic Based Pattern Mining approach for Search engine optimization

[1]K.Satya Srivalli, [2]V. Vidya Sagar
[1]Final M.Tech Student, [2]Associate Professor
[1,2]Dept of Computer Science and Engineering
[1,2]Pydah College of Engineering, Visakhapatnam, AP, India

Abstract: Generation of query oriented relevant information from search engine is always an interesting research issue in the field of information retrieval because of billions of relevant and irrelevant information available over internet. Satisfying the user search goal is a complex task when searching for a user specific query because of billions of related and unrelated data available over the network. In this proposed approach we are proposing an empirical model of search mechanism with FP Tree for finding frequent set of patterns (sequence of urls) and evolutionary algorithm for optimal results from efficient feedback sessions (based on query clicks) that are constructed from user click-through logs and can efficiently reflect the information needs of the users.

INTRODUCTION

Various approaches have been proposed by authors from years of research in the field of search engine optimization. Every research work have its own pros and cons. Some of the most commonly used search engines works based on relevance score, time stamps and query click graph. Latest technology of search engine follows basic concepts of semantic comparison of keywords, localization and cache implementations for optimal performance.

A simple term based and log based approach proposed by "Hsiao-Tieh Pu" finds the matched keywords and its synonyms from the log and retrieves the relevant documents based on frequency of the keyword or terms [1] and an Agglomerative graph based clustering approach proposed by "Doug Beeferman" and "Adam Berger" over query log to cluster the relevant data [2]. File relevance score is computed based on the term frequency (number of occurrences of a keyword in a single document) and inverse document frequency parameters. This approach concentrates on frequency of keyword but not with time stamps of the document. Therefore there is no priority for newly updated documents even though it is user relevant.

Time stamp based approaches works with recent time of the uploaded document along with the file relevance score. Clustering based techniques gives good results by combining the similar type of objects based on the time stamp and file relevance scores between the documents [8].

Some of the search engines work with query clicks based on previous user clicks or urls. Server maintains the log of urls with respect to keywords and mines the query oriented results based on user interests. Heasoo Hwang and others proposed the query grouping mechanism to group the similar queries based on relevance and computes relevance with Query Fusion graph. It is the combination of query reformulation (in this approach initially it compares the query that matches with the previously accessed frequent queries to check if they match) and query click graph [6][7].

Lee et al. [13] consider user goals as "Navigational" and "Informational" and categorize queries into these two classes. Li et al. [14] defined query intents as "Product intent" and "Job intent" and then tried to classify these queries according to the defined intents.

RELATED WORK

In previous works, feedback sessions of query are extracted from user click through logs along with the respective pseudo documents. These documents can be clustered based on the similarity between the documents. Cosine similarity is the measure to compute the similarity between documents based on frequency of keywords in document. They proposed k means algorithm for clustering of documents based on the similarity between documents. The main drawback with k-means clustering is prior specification of number of clusters (K value), random selection of the centroid and is not suitable for different density of objects [3].

We are proposing pattern mining approach instead of clustering approach for frequent use and relevant use of urls with respect to user query. Apriori algorithm is one of the simple frequent pattern mining algorithm, but the disadvantage of this algorithm is the generation of candidate set for every frequent item set generation and another major issue is multiple database scans [4]. FP Growth algorithm generates frequent item sets without candidate set generation.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 423- 426 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

An efficient pattern based technique for identifying the interesting patterns of clicked urls that the user request is proposed. Feedback session log maintains the user queries, session ids and urls. Initially our approach searches the session oriented results for input query and find the frequent patterns with FP growth algorithm by constructing the FP tree. After the generation of the frequent patterns, patterns can be forwarded to evolutionary approach for extraction of optimal patterns from FP tree generated patterns.

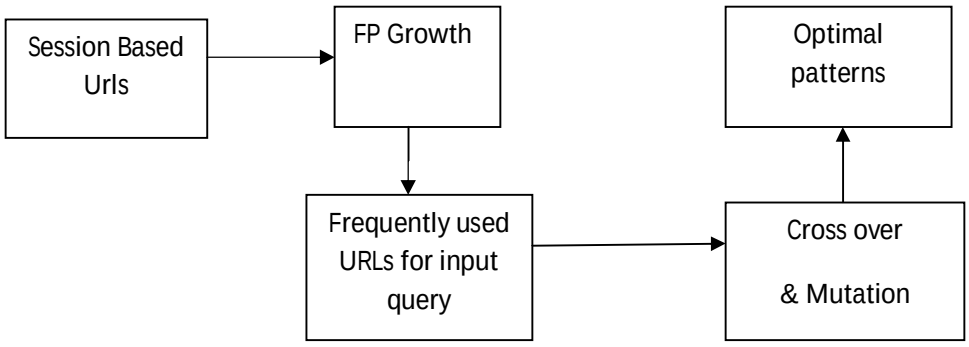
We integrate an efficient evolutionary approach (Genetic algorithm) to the previous pattern mining approach for optimal results. In this approach we apply cross over (i.e. combines the existing patterns to generate a new pattern) and mutation (i.e alter the genes (url) of the pattern) on mined patterns (sequence of urls) of user interesting results for optimal patterns in mined patterns [5].

PROPOSED WORK

In this approach proposed an efficient mechanism to satisfy the user search goals based on the session based user search history that are constructed from user click logs with respect to the user query. All the feedback sessions of a query are first extracted from user click-through logs and are mapped to pseudo-documents.

Individual session based results involves a unique session id, query and respective urls. Consider each individual url as event and sequence of urls as patterns for mining of urls that are most frequently accessed by the users. For the mining of urls we are using FP tree algorithm for finding the frequent patterns of urls visited by the previous users.

Architecture



Optimal URL pattern mining with FP Tree and Genetic Approach

Initially we consider set of session based urls from feedback session log for an input query passed by the new user and patterns by the set of session based individual urls. Patterns are forwarded to the FP growth algorithm for generation of frequent patterns or set of frequently visited urls with respect to user query.

This pattern mining approach involves in two phases. One is FP tree construction; there it maintains the two passes over dataset and frequent item set generation is done by traversing through FP tree. Sequential steps of FP growth algorithm are as follows :

Initially scans the dataset and finds minimum threshold value for each item and discards infrequent items.

1. Frequent items can be sorted based on decreasing order of their support or threshold.
2. Every node in tree maintains a counter for items.
3. Fixed order is used, so that the paths can overlap when transactions share items (when they have the same prefix). In this case, counters are incremented.
4. Pointers are maintained between the nodes containing the same item, creating singly linked lists.
5. Frequent item sets are extracted from the FP-Tree.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 423- 426 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

Genetic Approach

Genetic algorithm is the one of the evolutionary approach for finding the solution for the problems like NP hard problem. Genetic algorithm mainly works with selection, cross over and mutation operations over chromosomes. In the current scenario we are considering the pattern (sequence of visited urls) as chromosome.

Cross over: Cross over is one of the genetic operations between two chromosomes that is done by interchanging the genes of one parent to other parent chromosomes. It generates a child chromosome or offspring and evaluates the fitness of the chromosome.

Mutation: Mutation is the genetic operation within a chromosome by environmental changes. In our current scenario offspring or child chromosome can be generated by interchanging the genes (URL) of the chromosome itself.

For extraction of optimal patterns from the patterns generated by the fp growth algorithm we use the following approach as an example

After finding the frequent item sets from fp growth algorithm, to that association rules we have to apply GA.

Consider an example. Initially chromosomes can be constructed as follows:

Itemset abcde here a,b,c,d and e are urls

total items are 5. So initially 00000

from the item set we construct chromosome : 11001

after constructing all item sets in this manner we apply GA on those initial chromosomes.

GA steps:

Crossover: Take any position in the chromosome example 1 and 2 positions

ex: -- c1:11001

c2:00111

Swap 1 and 2 positions of the two chromosomes.

Results: c1:00001

c2:11111

Mutation: interchange or flip any bit from any position.

ex:-- c2:11111

flip position 5: 11110

Apply these steps for all chromosomes then we apply fitness function on the these chromosomes.

For a rule: Suppose A and B are two item sets where A and B contains some items present or its negation. It is known that higher the values of TP and TN and lower the values of FP and FN, the better is the rule.
Confidence Factor, CF = TP / (TP + FN)
We also introduce another factor completeness measure for computing the fitness function. Comp=TP/ (TP+FP), Fitness=CF *Comp
The fitness function shows that how much we nearer to generate the rule.
According to this, min confidence is very small i.e. less than 0.1.
If (fitness function > min confidence) set B = B U {x $\rightarrow$y}

The labels in each quadrant of the matrix have the following meaning:
TP = True Positives = Number of examples satisfying item set A and item set B.
FP = False Positives = Number of examples satisfying item set A but not item set B.
FN = False Negatives = Number of examples not satisfying item set A but satisfying item set B.
TN = True Negatives = Number of examples not satisfying item set A nor item set B.

CONCLUSION

In this paper, a genetic based pattern mining approach has been proposed to satisfy the user search goals. First, FP growth algorithm finds the frequent pattern of urls with respect to user query and then the generated patterns are forwarded to evolutionary approach for the computation of fitness after cross over and mutation operation are performed over chromosomes.

REFERENCES

[1] Web Relevant Term Suggestion Using Log-based and Text based Approaches. Hsiao-Tieh Pu, Hsin-Chen Chiao.

[2] Agglomerative clustering of a search engine query log Doug Beeferman, Adam Berger.

[3] A New Algorithm for Inferring User Search Goals with Feedback Sessions Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng.

[4] Mining Frequent Patterns without Candidate Generation Jiawei Han, Jian Pei, and Yiwen Yin.
[5] An introduction to Genetic Algorithms Melanie Mitchell.

[6] Organizing User Search Histories Heasoo Hwang, Hady W. Lauw, Lise Getoor, and Alexandros Ntoulas.

[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] Answering General Time-Sensitive Queries Wisam Dakka, Luis Gravano, and Panagiotis G. Ipeirotis, Member, IEEE.

[9] T. Joachims, "Optimizing search Engines Using Click through 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. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Click through 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.

[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.