TRACKING SEARCH CONTEXT: A META HEURISTIC USER SEARCH SESSION EXPLORATION APPROACH

***Sahithi Tummala, **Ravi Tene**

**Dept. of Information Technology*

VNR Vignana Jyothi Institute of Eng. and Technology

Hyderabad, India

***Dept. of Information Technology*

VNR Vignana Jyothi Institute of Eng. and Technology

Hyderabad, India

ABSTRACT

This research application suggests elevating a search query log evaluation by enchanting into account the conceptual attributes of query provisions. We first illustrate a strategy for removing a conceptual description of a search query log after which reveal how we will utilize it to comparatively extract outcomes without ambiguity. The notion relation is consisting of set provisions consistent collectively concept relations as well as of a function to determine the concept spacing between provisions, which even more called as TF (term frequency). We then drain TF with feedback sessions, which we construct upon click using logs. Furthermore, a query terminology clustering algorithm is used on the log description to extract user pursuits.

Keywords: *component; pseudo-documents; clustering; TF-IDF and ranking*

INTRODUCTION

In online search programs, when user gives up the query to the search engine. The data requires of various user may vary in numerous features of query data. This gets difficult to attain user data specifications. Now-a-days, great deal of data is out there on the internet; query search has become an essential tool for internet users to achieve desired information. But, it becomes terribly troublesome task to induce precise information that user wish. Often uncertain concerns may not exclusively display by users so it leads to less reasonable to the search engine.

The utility of a search engine is littered with multiple factors. Whereas the first issue is that the soundness of the beneath lying retrieval model and ranking perform, the way to organize and present search results is additionally a really vital issue which will have an effect on the utility of a search engine considerably.Understanding user access patterns won't solely improve response from internet however conjointly lead to higher marketing decisions (e.g., by putting advertisements in often visited web pages, or by providing higher customer/user clustering and behavior analysis). Capturing user search patterns in such distributed info environments is termed as web usage mining.

Typically, internet users submit a brief query consisting of some words to look for response in search engines. These queries will be short and ambiguous. The way to interpret these queries in terms of groups of target classes has qualified as a serious analysis issue. To obtain user specified data needs numerous ambiguous or uncertain requests can incorporate a broad concept and different users should get data on various features when they specify the equivalent query.

For instance, when a user submits a query "java" to the search engine, many users are considering to know data about programming language as well as many users learn information about the island of Indonesia. Hence, it is required to discover various user data search objectives. User data require is to choose and get the data to fulfill the requirements of every user.

To fulfill the user data needs by taking the search plans with the user provided the query. First we will consider the obtained results for the user entered input. We will gather the related information which search engine responses to the query. Input will consist of URL's, Title and Snippet in a text format. Pre-processing is done is implemented to the text. Later, we will cluster the keywords based on the user input. Extracted results will be grouped into clusters. Finally, we will rank the data and display the results. By doing the above method we will successfully re-structure the web search results. As the interference as well as evaluation of user search objectives with query may contain an amount of benefits in enhancing the search engine significance as well as user understanding. Hence it is essential to accumulate various user goals as well as obtain the effective data on various features of a query.

The remaining of the report is arranged as follows: Section 2 describes the related work. Section 3 covers the proposed approach. In section 4, we describe proposed approach and it's implementation along with experimental work carried out by and finally section 5 describes the conclusion.

RELATED WORK

Mining procedures can be used to significant search query logs to extract data for user pursuits [1], [2], [3], [4], [5], [6] and [7]. This is specifically an essential step for the layout of true user centric programs in which user search actions are determined and considered into account. Recently, a lot research happens to be done in the area of search query log evaluation. Currently, researchers have mainly centered on statistical techniques for getting knowledge. These suggestions are not appropriate to issues regarding the semantics of the information like the identity of users search concerns.

To manage this concern, "Zheng Lu et al", instructed "A new Algorithm for Inferring User Search Goals with Feedback Sessions" [8]. Initially, this model presents feedback sessions are evaluated to generalize user search objectives rather than search outcomes otherwise clicked URL's. The clicked as well as the un-clicked URL's before the final click are regarded as customer implicit feedback also considered to report them for constructing feedback sessions. Hence, feedback sessions will exhibit user knowledge needs additional effectively. Secondly, it routes feedback sessions toward pseudo data for estimating objective texts in user minds. The pseudo-documents will enhance the URL's

with further textual information like the titles as well as snippets. In compliance with the pseudo-documents, user search goals are able to be intended and presented with certain keywords. Lastly a novel principle, CAP is developed to estimate the potency of user search goal illation.

PROPOSED APPROACH

With the desire accumulated from "A New Algorithm for Inferring User Search Goals with Feedback Sessions" [11]. Our aim is to locate search results by particulars learned from search engine log file. Given a user input, the procedure of our approach is:

1. Get user input connected information from computer program logs. All the information forms an operation set.
2. Learn feature from the knowledge acquired within the operating set. These feature correlate with the user's interests stated in the user input. Every verity is tagged with a specific query.
3. Grade and group the search results of the user input compatible with the specifications learned above.

Let's see an elaborated description of every step with the assistance of pictorial description.

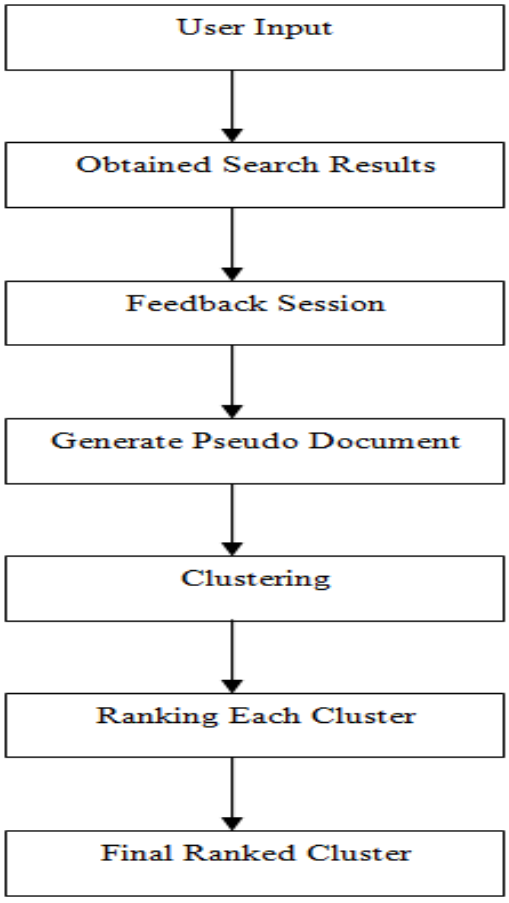

Fig.1. Framework of Proposed Method

A. *User Input*

The user needs to translate his information need into the search engine. Those inputs are treated as queries submitted by users to search engine to represent the data desires of user.

B. *Obtained Search Results*

The user input is maintained as a log and the results will be produced based on the keywords. Portray each goal accompanied with some keywords automatically. And store those along with their queries will be saved. We record the obtained search result for user input in a text document (nearly 100 obtained results). Both clicked and un-clicked URL's are recorded.

- **Text Paragraphs**

  Each and every URL's is enriched with excess information (textual content) that consists of title and snippet.

  1) **Title:** It is the textual content tagged with URL's and snippet on the result page. It describes the main theme of the web page in other words it outlines the content on the web page. The main purpose of it is to give a sight of the web page with just one sentence.
  2) **Snippet:** It describes the content by displaying some wage keywords. These are user input specific, and so they are always changing. In short, we can stay snippet can be said as keywords to describe the content of the web page.

C. *Feedback Sessions*

Feedback session is maintained by grouping set of URL's of one session of a specific user input. It resides of each clicked as well as unclicked URL's. It focuses on user's input. Since it represents each clicked and unclicked URL's its unsuitable to use feedback sessions directly for clustering.

D. *Generate Pseudo Document*

It consists of search results of user input in a very precise manner. It consists of URL's along with their Title and Snippet. In such resemble each and every URL's in the feedback session is portrayed by a text in form of a paragraph that includes of the URL's its title and snippet. Those are stored in a text file.

Text pre-processing is done on the text paragraph.

- **Pre-Processing**

  Pre-processing includes following steps

  1) **Transforming Text:** Converts tokens to lowercase, removes punctuations/numbers.

2) **Eliminate Stop Words**: Stop words are very common and rarely provide useful information. So we remove stop words which will reduce dimensionality.
3) **Stemming**: In many cases words, need to be stemmed to retrieve the radicals. Convert the terms into their stemmed form remove normalizes verb tenses, plurals and different word forms. The output of stemming will be the basic elements.

E. Clustering

Clustering is an important process to improve web search results. It is the method of dividing information into categories or clusters. Things in the similar cluster have similar properties in many ways. Similar components are grouped based on following factors. The term with the highest values is treated as the center point for the cluster. The center point is the keyword which is used to depict the user search goal. The process is based on Term frequency weight for each keyword.

F. Ranking Each Cluster

For one user input we will get many numbers of clusters. Based on the distance, rank and similarity factor for key terms. We will rank the cluster so that the cluster with maximum weight will be on top. To make certain that the maximum weighted cluster and its grouped components will be displayed on top of web search results. So that the search results are rearranged with most viewed and related result on top of the web page.

G. Final Ranked Cluster

In this section we will rank the cluster and combine them so they can be grouped with similar results in one place and ranking will be done to every cluster for obtaining the best results. Those will be displayed on the top of the web page.

PROPOSED APPROACH AND IT'S IMPLEMENTATION

In the above section-3, theoretical description is explained. In this section we will see the mathematical and step-by-step approach for the above mentioned method.

Here we will take the input as "The Sun" and we will record obtained 100 results from search engine. Future, we implement above mentioned mathematical methods on the recorded text using JAVA in Netbeans IDE.

Step.1: First we submit user input as "The Sun" and extract top 100 results.

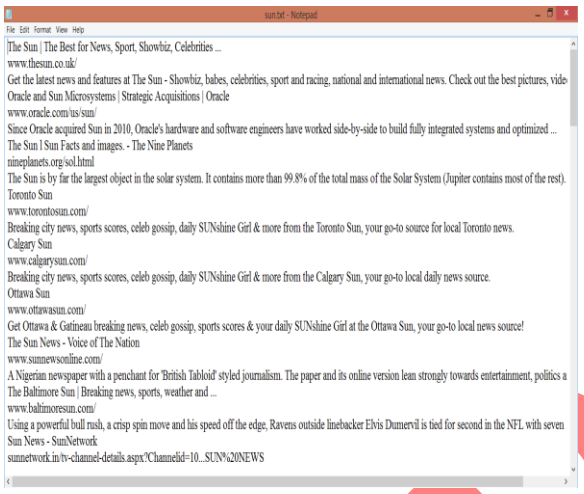

Fig.2. Recorded results for the input "The Sun"

Step.2: URL's, Title, Snippet are extracted and pre-processing is applied. On completion of pre-processing we will get keywords.

Step.3: We will evaluate TF and IDF.

Step.4: Calculate TF*IDF value for each keyword.

Step.5: Calculate threshold.

Step.6: Generate frequent candidate item sets and mark frequent item sets.

Step.7: Cluster the obtained keyword based on their value.

Step.8: Evaluate cosine similarity for every pair of words.

Step.9: Compute similarity factor.

Step.10: Rank each document by multiplying TF-IDF result for every keyword with similarity factor.

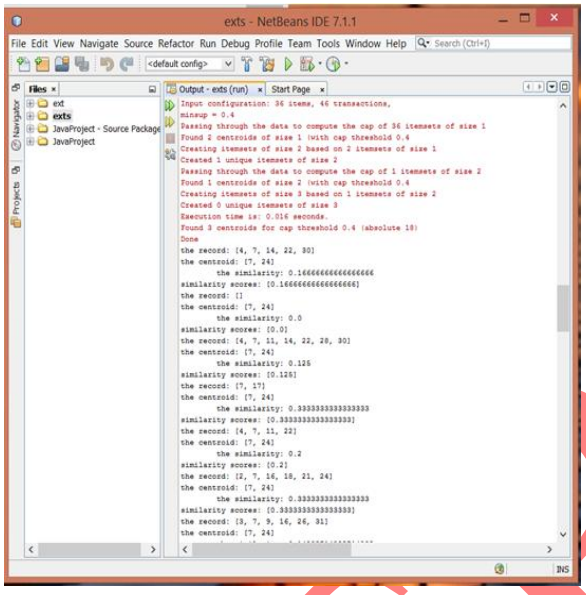

Fig.3. Output Screen Showing Clustering and Ranking

Step.11: Finally, we will get the sorted document for user input based on their ranked values we will get the desired output.

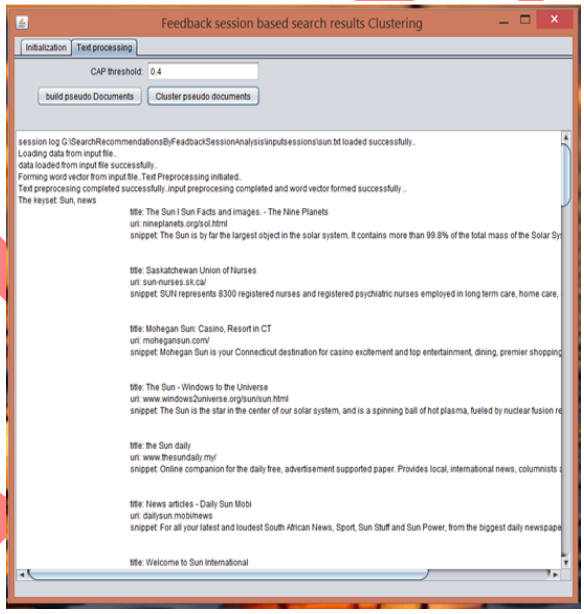

Fig.4. Output Screen Showing Final Ranked Cluster

CONCLUSION

Right here in this document we suggested a novel statistical design to determine the user search goals by feedback sessions, which is an improvised adaptation of the model developed in [11]. In our

design, rather utilizing each feedback sessions, we processed these by confirming context sensitivity around search queries as well as feedback sessions. The developed design is the motivation of the process discovered in [10]. The empirical outcomes showing that examining the context sensitivity result in augment the cluster correctness as well as to cluster

REFERENCES

[1] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining,J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[2] 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.

[3] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "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.

[4] 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.

[5] 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.

[6] 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.

[7] 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 Zheng Lu; Hongyuan Zha; Xiaokang Yang; Weiyao Lin; Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," Knowledge and Data Engineering, IEEE Transactions on , vol.25, no.3, pp.502,513, March 2013 doi: 10.1109/TKDE.2011.248.

AUTHORS PROFILE

T.Sahithi received her B.Tech degree in Computer Science and Engineering from VITS Karimnagar in 2013. She is pursuing M.Tech in Computer Networks And Information Security from VNR VJIET Hyderabad. Her research interests include data mining

RAVI TENE received his B. Tech degree from JNTU Hyderabad and M. Tech degree from JNTU Hyderabad. He is an Assistant professor in the Department of Information Technology VNR VJIET. His research interests include Image processing and data mining.