International Journal of Computer Science and Informatics

Volume 1 | Issue 4 Article 10

April 2012

Automatic Query Reformulation Using Classification in Web Searching

Ralla Suresh
Pulipati Prasad Engineering College,Khammam, A.P, rella.suresh@gmail.com

Kurmachalam Ajay Kumar
Sree Kavitha Engineering College, Khammam, AP., ajaykumarcvr502@gmail.com

Automatic Query Reformulation Using Classification in Web Searching

Ralla Suresh[1], Kurmachalam Ajay kumar[2] & Ramesh Dharavath[3]

[1]Rayalaseema University, Andhra Pradesh, India. [2]Sree Kavitha Engineering College, Khammam, AP,India
[3]Pulipati Prasad Institute of Technology & Sciencs, Khammam, AP, India.
E-mail : rella.suresh@gmail.com[1], ajaykumarcvr502@gmail.com[2]

Abstract **-** The search query is a set of words or phrases a user enters when looking for information on a specific topic or subject. Formulating a search query is a challenging task for most of users because they are required to express their anomalous states of knowledge. In the query reformulation stage, users modify their initial queries and submit new ones that more accurately reflect their information needs. Classification and Prediction are two forms of query analysis that can be used to extract models describing frequently used query classes or to predict the reformulation for the query.. An ideal solution might be that the system automatically generates a concise and informative summary for each perspective of the query. In our approach a model is constructed by analyzing queries described by the frequency of search, query is assumed to belong to a predefined class.

Key words **-** *Query, formulating, anomalous, classification, reformulation.*

I. INTRODUCTION

The popularization of the Internet does not only mean the increase of its users but also change the usage of it. In the beginning, the Internet was mainly used for research or business but now people have started to use the Internet for their daily life. Among various internet services the Web is one of the most popular, and most of internet users access it frequently. We can obtain information on various topics of daily life such as shopping, eating, healthcare, education etc. The expansion of the number of web users and the amount of information stored in the web has raised many new problems in information retrieval. The most common way to find information in the web is using a web search engine. Web search engines have common appearances with conventional information retrieval systems. The user forms query words to express his information needs and the system returns documents that are estimated to be relevant.

Web query classification/categorization is a problem in information age. The task is to assign a web search query to one or more predefined categories, based on its topics. The importance of query classification is underscored by many services provided by Web search. A direct application is to provide better search result pages for users with interests of different categories. For example, the users issuing a Web query "*apple*" might expect to see Web pages related to the fruit apple, or they may prefer to see products or news related to the computer company. Online advertisement services can rely on the query classification results to promote different products more accurately. Search result pages can be grouped according to the categories predicted by a query classification algorithm. However, the computation of query classification is non-trivial. Different from the document classification tasks, queries submitted by Web search users are usually short and ambiguous; also the meanings of the queries are evolving over time.

II. RELATED STUDIES

The Web has become an indispensable aspect in the lives of many people, and search engines are the main portal to the Web. Search engines are "the tool" for accessing the information, Internet sites, and services on the Web that many people use on a daily basis. Beyond their popularity, how are people using these Web search engines? How can we determine what these people are seeking? What task, goal, need, or intent are they trying to address with their Web searching? Web search engines can help people find the resources they are looking for by more clearly identifying the searcher's intent behind the query.

In this paper, we classify user searcher based on intent in terms of the type of content specified and operationalize these classifications with defining characteristics. We implement this operationalized classification in an application that automatically classifies queries from a search engine transaction log. We discuss how this model can be used to improve Web search engines.

III. RESEARCH OBJECTIVES

We apply domain-specific approach, in which we made special-purpose search systems for each domain. By targeting the specific purpose, we can reduce the difficulty caused by heterogeneity of the web and ambiguity of user's needs. Of course, many domain-specific web search system have been developed so far. However, existing methods for building domain-specific search systems require specialized facilities or human expertise knowledge and it is hard to apply same method for developing the system in other domains.

IV. RESEARCH DESIGN

Query enrichment is a key step because our goal is to classify short and ambiguous queries without any additional descriptions about these queries. After this step, two kinds of information for each query are collected. One is the list of Web pages related to the target query. The other is the set of categories corresponding to the related pages

For research question one, we qualitatively analyzed samples of queries from Web search engine transaction logs [3, 5]. In order to identify characteristics for each query category. For the analysis, we selected random samples of queries and manually classified them in one of three categories (information, navigational, and transactional) as define in [2]. We then derived characteristics for each category that would serve to define the queries in that category. This was an iterative process with multiple rounds of "query selection – classification –characteristics refinement".

From the previous two approaches, we can build various classifiers independently which can classify the input queries into the target categories. These approaches are based on different mechanisms and can be complementary to each other. Previous works have shown that a proper combination of different base classifiers can improve the final classification performance

1. Input a query to obtain a list of results containing snippets, collection names, and keywords.
2. Randomly sample search results (snippets) from the retrieved collection to extract topic distribution probabilities.
3. Compute the document-topic probability distribution for each candidate result.
4. Calculate the topic-keyword and topic-collection probability distributions.
5. Visualize the topics by showing top words, keywords, and collections covered by each topic through an interactive user interface, which will allow users to view the topic constructs and easily reconstruct their queries.

The following summarizes our algorithm for extracting keyword spices.

i) Generate input keywords according to some estimate of distribution p(k) and collect web pages that contain keyword k and classify them into positive and negative examples by hand.

ii) Split the examples into two disjoint subsets, the training set, $D_{training}$, for generating the initial decision tree and the validation set, $V_{alidation}$, for pruning.

iii) Make the initial decision tree from $D_{training}$ using an information gain measure without any pruning technique.

iv) Convert the learned tree into a set of positive rules by creating one rule for each path from the root node to each leaf node: this classifies positive examples.

v) **for each** rule r **do**

Repeat Remove the precondition keyword that results in the Maximum increase in the harmonic mean

$$F_r = \frac{2}{\frac{1}{R_{validation}} + \frac{1}{P_{validation}}}$$

of precision measure $P_{validation} = I\, D_{rule} \cap D\, r_{ecipe}\, I / I\, D_{rule}\, I$and recall measure $R_{validation} = I\, D_{rule} \cap D_{recipe}\, I / I\, D_{recipe}\, I$ Where D_{recipe} is the set of relevant documents classified by human and D_{rule} is the set of documents which r classify relevant in $D_{validation}$.

Until there is no keyword that can be removed without decreasing the harmonic mean $F_{r.}$

End Make a disjunctive normal form of Boolean expression h by making a disjunction of all preconditions of the positive rules.

vi) **Repeat** Remove the conjunctive component from the disjunctive normal form h that results in the maximum increase in the harmonic mean

$$F_s = \frac{2}{\frac{1}{R_{validation}} + \frac{1}{P_{validation}}}$$

of precision measure $P_{validation} = | D_{rule} \cap D_{recipe} | / | D_{rule} |$ and recall measure $R_{validation} = | D_{rule} \cap D_{recipe} | / | D_{recipe} |$ where D_{recipe} is the set of relevant documents classified by human and D_{spice} is the set of documents which h classify relevant in $D_{validation}$.

Until there is no conjunction that can be removed without decreasing the harmonic mean F s.

Return h

V. RESULTS

To evaluate the performance of our keyword spice method, we conducted realistic tests in the cooking domain with external commercial search engines.

For the experiment we chose the keywords of (*khali mirchi* (pepper) and *Adhrakh* (ginger) and *dhaniya*(coriander)) which were not used to generate the keyword spices and used the major Indian web search engine "123khoj.com" and evaluated the keyword spices "(ingredients AND NOT speciality AND NOTgoods). Then we forwarded the queries containing only keywords and the queries with the keyword spices to "123khoj.com" and compared the

Fig 5.1 Recorded precision values for queries results.

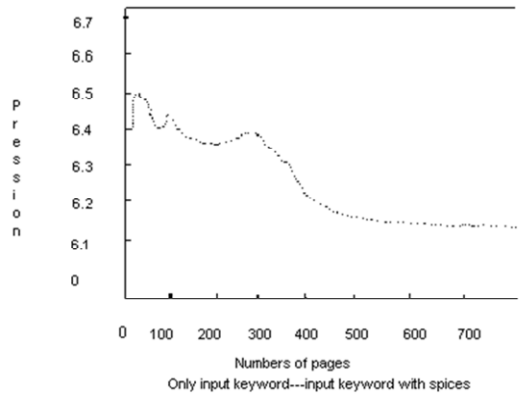

Fig 5.1 (a) : Query "pepper" forwarded to 123khoj.com

Figure 5.1 compares the precision values for the queries containing only keywords and the queries with keyword spices for the three input keywords.We checked up to the top 100 pages as ranked by the search engine 123khoj.com. In general, as the number of pages viewed increases, the precision with query-only input decreases or stays low, while the precision of queries with keyword spices stays high. Precision is higher than 97% for all queries

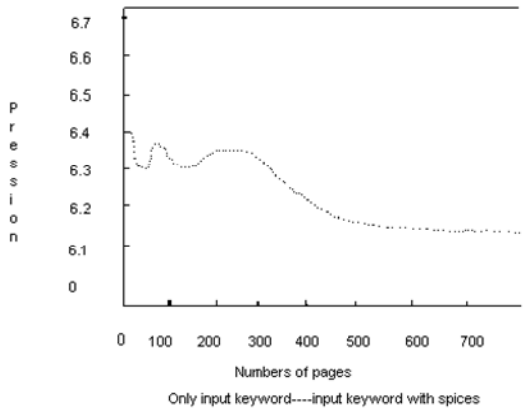

Fig. 5.1 (b) : Query"ginger" forwarded to 123khoj.com

VI. CONCLUSION

In order for Web search engines to continue to improve, they must leverage an increased knowledge of user behavior, especially efforts to understand the underlying intent of the searchers. The results of this research demonstrate the ability to implement of an approach for automatically classifying queries. Our approach does not depend on external content and can be implemented in real time. This makes it a viable solution for Web search engines to classify user intent based on the type of content desired. Additionally, the larger data set provides more accurate percentages of user intent classification than smaller mostly manual studies. The higher percentage of information queries indicates that users view search engines primarily as information retrieval tools rather than instruments of navigation or commerce. Future work involves both queries and sessions in order to identify more granular classifications of user intent (i.e. sub-categorizations of *informational*, *navigations*, and *transactional*). More targeted Web results to the underlying user content need will increase performance of future Web search engines.

REFERENCES

[1] Baeza-Yates, R., Calder ´on-Benavides, L. and Gonz´alez-Caro, C. The Intention Behind Web Queries. In the Proceedings of string processing and information retrieval (spire 2006**)**. Glasgow, Scotland, 98-109, 2006.

[2] Broder, A. A Taxonomy of Web Search. In the proceedings of SIGIR Forum.36, 2, 3-10, 2002.

[3] Jansen, B. J. and Spink, A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management. 42, 1, 248-263, 2005.

[4] Jansen, B. J., Spink, A., Blakely, C. and Koshman, S. forthcoming. Web Searcher Interaction with the ogpile.comMeta-Search Engine. Journal of the American Society for Information Science and Technology.

[5] Jansen, B. J., Spink, A. and Saracevic, T. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. Information Processing & Management. 36, 2, 207-227, 2000.

[6] Belkin, N., Oddy, R. and Brooks, H. Ask for information retrieval, parts 1 & 2. Journal of Documentation, 38, 2 1982), 61-71, 145-164.
He, D., Göker, A. and Harper, D. J. Combining evidence for automatic Web session identification. Information Processing & Management, 38, 5 (September 2002), 727-742.

[7] Jansen, B. J. and Pooch, U. Web user studies: A review and framework for future work. Journal of the American Society of Information Science and Technology, 52, 3 2001), 235-246.

[8] Meadow, C. T., Hewett, T. T. and Aversa, E. A computer intermediary for interactive database searching ii: Evaluation. Journal of the American Society for Information Science, 331982), 357-364.

◈◈◈