Research Paper

Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

A Novel Personalized Web Search with Offline Capability

Engineering

KEYWORDS : user feedbacks, uri-tracker, personalized web search, ranking, pseudo document processing, clustering

Roy T P MTech Scholar, Vedavyasa Engineering College, Calicut

Ginnu George Assistant professor, Vedavyasa Engineering college, Calicut

ABSTRACT

Nowadays Internet is widely used by users to satisfy various information needs. However, ambiguous query/topic submitted to search engine doesn't satisfy user information needs, because different users may have different information needs on diverse aspects upon submission of same query/topic to search engine. So discovering different user search goals becomes complicated. Also for mobile users, it is not possible to get internet in all time. The evaluation and depiction of user search goals can be very useful in improving search engine relevance and user knowledge. This paper proposes a novel approach for inferring user search goals by analyzing user query logs from various search engines. The proposed approach is used to discover different user search goals for a query by clustering the user feedback sessions. Also we integrate a new offline capability called URI-Tracker, which download all the relevant contents of the website. This helps user to access their needed information without any internet connectivity. Also URI-tracker keeps the indexing locally hence improves query processing time.

Introduction

In web based search applications, user submits the query to search engine to search efficient information. The information needs of different user may differ in various aspects of query information. This becomes difficult to achieve user information needs. Sometimes ambiguous queries may not exactly represented by users so it results in less understandable to search engine. To achieve the user specific information needs many ambiguous/uncertain queries may cover a broad topic and dissimilar users may want to get information on different aspects when they submit the same query. For example, when user submits a query "java" to search engine, some users are interested to know information about programming language and some users want to know information about island of Indonesia. Therefore, it is necessary to discover different user information search goals. User information need is to desire and obtain the information to satisfy the needs of each user. To satisfy the user information needs by considering the search goals with user given query, cluster the user information needs with different search goals. Because the interference and evaluation of user search goals with query might have a numeral of advantages in improving the search engine significance and user knowledge. So it is necessary to collect the different user goal and retrieve the efficient information on different aspects of a query. Capturing different user search goals related to information needs changes the normal query based information retrieval.

Evaluation and analysis of user search goals has many advantages as follows.

- Reorganize web search results according to user search goals by grouping search results with same information need. This can be useful to other users with different search goals to find easily what they want.
- Query recommendation by using user search goals depicted with some keywords. This can be helpful to other users to form their query more effective.
- Reranking web search results according to different user search goals.

User search goal analysis is important to optimize search engine and effective query results organization. When query is submitted to search engine, the returned web pages of search results are analyzed [1], [2]. Since it does not consider user feedback, many unuseful and noisy search results that are not clicked by user may be analyzed. This may degrade the search goals discovery. X. Wang and C-X. Zhai [3] learns interesting aspects of similar query/topic from web search logs which consists clicked web pages URLs and organize search results accordingly. Their approach may results in limitation, as the different clicked URLs for a query/topic may be small in number. There are many works [4], [5] which classify queries into some predefined specific classes and try to find out query intents and user goals. However, different queries have different search goals and finding precise, suitable predefined search goal classes may be difficult and sometimes impossible to categorize.

literature survey

Since many years, research in web log mining has been subject of interest. Many previous works has been investigated on problem of analyzing user query logs [6], [7], [8], [9], [10]. The information in query logs has been used in many different ways, such as to infer search query intents or user goals, to classify queries, to provide context during search, to facilitate personalization, to suggest query substitutes and to identify frequently asked questions (FAQs). Effective organization of search results is critical for improving utility and relevance of any search engine. Clustering search results is an effective way to organize search results which allows a user to navigate into relevant documents quickly. Generally all existing work [1], [11] perform clustering on a set of top ranked results to partition results into general clusters, which may contain different subtopics of the general query term. However, this clustering strategy has two deficiencies which make it not always work well. First, discovered clusters do not necessarily correspond to the interesting aspect of a topic from user-oriented perspective. Second, cluster labels are more general and not informative to identify appropriate clusters. Wang and Zhai [3] proposed approach to organize search results in user-oriented manner. They used search engines log to learn interesting aspects of similar queries and categorize search results into aspects learned. Cluster labels are generated from past query words entered by users. H-J Zeng et.al [1] proposed a query based method to cluster search results. For a given query, the rank list of documents return by a certain Web search engine, it first extracts and ranks most salient phrases as candidate cluster names, based on a regression model learned from pervious training data. Candidate clusters are formed by assigning documents to relevant salient phrases and the final cluster are generated by merging these candidate clusters. But this method only produces the result with higher level of the documents only and it doesn't make the results for all search based user goals.

Click-through data

In web search environment, there are many abundant queries and user clicks. User clicks represent implicit relevance feedback. In this framework, user clicks are recorded in user click-through data. User uses click-through data stored in user logs to simulate user experience in web search. In general, when query is issued, the user usually scans links to documents in a result list from first to last. Clearly, the user clicks on the links to the documents that look relevant of informed choice and skips other documents. Therefore, the proposed approach utilize user click as relevance judgments to evaluate search precision since click-through data can be collected at low cost, it is possible to

do large scale evaluation under this framework.

1) Feedback sessions: Feedback sessions are considered as users' implicit feedback. In general, a session for web search is a sequence of consecutive queries to satisfy single information and some clicked results. But to infer user search intents/goals for a particular query, single session is considered. Single session corresponds to only one query, which differs from conservative session. The proposed feedback session consists of both clicked and unclicked URLs for a particular query in a single session and ends with last clicked URL. This shows that before last clicked URL, all the URLs are scanned and evaluated by user.

Therefore, all clicked URLs and unclicked URLs before last click are considered as user feedbacks. In each feedback session clicked URL (visited link) tells users information need and unclicked URL (unvisited link) tells what users do not want. This visited link is called as positive feedback and unvisited link is called as negative feedback. There are large numbers of diverse feedback sessions in user Click-through log. So it is efficient to examine feedback sessions for inferring user search goals than to examine clicked URLs or search results directly.

Building pseudo-documents

As URLs alone are not informative enough to tell intended meaning of a submitted query. To obtain rich information, we enrich each URL with additional text content by extracting the titles and snippets of URLs appearing in feedback session. Thus, each URL in feedback session is represented by small textual content which contains its title and snippet. Then some text pre-processing is done on those textual contents, such as transforming all letters to lowercase, eliminating stop words (frequent words) and word stemming by using porter algorithm [15]. Lastly, TF-IDF [16] vector of URL's titles and snippets are formed respectively as,

$$\mathbf{T}_{u_i} = [t_{w_1}, t_{w_2}, \ldots, t_{w_n}]^T, \quad \mathbf{S}_{u_i} = [s_{w_1}, s_{w_2}, \ldots, s_{w_n}]^T, \tag{1}$$

where $\mathbf{T}_{ui}$ and $\mathbf{S}_{ui}$ are TF-IDF vectors of URL's title and snippet, respectively. The u_i is i^{th} URL in feedback session and w_j is the j^{th} term in the enriched URL. The t_{wj} and s_{wj} denotes j^{th} term in the URL's title and snippet respectively. Feature representation $\mathbf{F}_{ui}$, of i^{th} enriched URL is weighted sum of $\mathbf{T}_{ui}$ and $\mathbf{S}_{ui.}$

$$\mathbf{F}_{u_i} = \omega_t \mathbf{T}_{u_i} + \omega_s \mathbf{S}_{u_i} = [f_{w_1}, f_{w_2}, \ldots, f_{w_n}]^T, \tag{2}$$

where w_t and w_s are weights of title and snippet respectively. Each term of $\mathbf{F}_{ui}$, denotes importance of term in i^{th} URL.

In order to obtain feature representation of a feedback session, optimization method is used to merge feature representations of each clicked and unclicked enriched URLs in the feedback session. Let $\mathbf{F}_{fs}$ be feature representation of a feedback session, $f_{w_{c_i}}(w)$ and $f_{w_{\bar{c}_i}}(w)$ are feature representation of clicked and unclicked URLs respectively and $f_{fs}(w)$ is value for term w. $\mathbf{F}_{fs}$ should be such that sum of distance between $\mathbf{F}_{fs}$ and each $\mathbf{F}_{w_{c_i}}$ is minimized and sum of distance between $\mathbf{F}_{fs}$ and each $\mathbf{F}_{w_{\bar{c}_i}}$ is maximized.

$$\mathbf{F}_{fs} = [f_{fs}(w_1), f_{fs}(w_2), \ldots f_{fs}(w_n)]^T \tag{3}$$

Each feedback session is represented by $\mathbf{F}_{fs}$. This is nothing but pseudo-document which is used for discovering user intents or search goals. These pseudo-documents contain what user requires and what do not, which is used to learn interesting aspects of a query.

Experiments And Results

In this section, we will show experiments of our proposed algorithm. The data set that we used is based on the click-through logs from a commercial search engine collected over a period of two months, including totally 2,300 different queries, 2.5 million single sessions and 2.93 million clicks. On average, each query has 1,087 single sessions and 1,274 clicks. However, these queries are chosen randomly and they have totally different click numbers. Excluding those queries with less than five different clicked URLs, we still have 1,720 queries. Before using the data sets, some preprocesses are implemented to the click-through logs including enriching URLs and term processing.

In our approach, we have two parameters to be fixed: K in K-means clustering and γ in [12]. When clustering feedback sessions of a query, we try five different K (1,2,….,5) in K-means clustering. Then, we restructure the search results according to the inferred user search goal and evaluate the performance by CAP, respectively. At last, we select K with the highest CAP.

TABLE 1
Abstracted Keywords Used to Depict User Search Goals for Some Ambiguous Queries

Query	Four keywords to depict user search goals
earth	google, map, wikipedia, planet
	planet, solar, system, nineplanet
	nasa, science, gov, nineplanet
graffiti	art, wall, writing, free
	game, yahoo, art, play
india	map, city, region, information
	travel, information, welcome, land
lamborghini	car, history, company, overview
	new, auto, picture, vehicle
	club, oica, worldwide, lamborghiniclub
sex on the beach	photo, vh1, gallery, cocktail
	recipe, vodka, cocktail, drink
	demeter, fragrance, cocktail, perfume
the sun	news, photo, information, newspaper
	star, earth, solar, sunspot

Before computing CAP, we need to determine γ in [12]. We select 20 queries and empirically decide the number of user search goals of these queries. Then, we cluster the feedback sessions and restructure the search results with inferred user search goals. We tune the parameter γ to make CAP the highest when K in K-means accord with what we expected for most queries. Based on the above process, the optimal γ is from 0.6 to 0.8 for the 20 queries. The mean and the variance of the optimal γ are 0.697 and 0.005, respectively. Thus, we set γ to be 0.7. Moreover, we use another 20 queries to compute CAP with the optimal γ (0.7) and the result shows that it is proper to set γ to be 0.7.

In the following, we will first give intuitive result of discovering user goals to show that our approach can depict user search goals properly with some meaningful words. Then, we will give the comparison between our method and the other two methods in restructuring web search results.

We infer user search goals for a query by clustering its feedback sessions. User search goals are represented by the center points of different clusters. Since each dimension of the feature vector of a center point indicates the importance of the corresponding term, we choose those keywords with the highest values in the feature vector to depict the content of one user search goal.

Table 1 gives some examples of depicting user search goals with four keywords that have the highest values in those feature vectors. From these examples, we can get intuitive results of our search goal inference. Taking the query "lamborghini" as an example, since CAP of the restructured search results is the highest when (K=3), there are totally three clusters (i.e., three

Volume : 3 | Issue : 5 | May 2014 • ISSN No 2277 - 8179

lines) corresponding to "lamborghini" and each cluster is represented by four keywords. From the keywords "car, history, company, overview," we can find that this part of users are interested in the history of Lamborghini. From the keywords "new, auto, picture, vehicle," we can see that other users want to retrieve the pictures of new Lamborghini cars. From the keywords "club, oica, worldwide, Lamborghiniclub," we can find that the rest of the users are interested in a Lamborghini club. We can find that the inferred user search goals of the other queries are also meaningful. This confirms that our approach can infer user search goals properly and depict them with some keywords meaningfully

Some of the screen shots of the project is shown below:

Fig. 3: Search Page

Fig. 4: Search Results

Conclusions

The proposed system can be used to improve discovery of user search goals for a query by clustering user feedback sessions represented by pseudo-documents. Using proposed system, the inferred user search goals/intents can be used to restructure web search results. So, users can find exact information needed as they want very efficiently. The discovered clusters can also be used to assist users in web search. Also there is a URI-Tracker which stores site information so users can access them without bothering about net connectivity. It also provides local URI indexing, which fastens the process of query execution.

REFERENCE

[1] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to | Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004. | | [2] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI'00), pp. 145-152, 2000 | | [3] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007 | | [4] 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. | | [5] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for Web Query Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 131-138, 2006 | | [6] 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. | | [7] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000. | | [8] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001 | | [9] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges for Web Query Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 131-138, 2006. | | [10] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004 | | [11] O. Zamir and O. Etzioni.Grouper: A dynamic clustering interface to web search results.Computer Networks, 31(11-16), pp.1361-1374, 1999 | | [12] 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. | | [13] 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 | | [14] 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. | | [15] Porter, M.An algorithm for suffix stripping. Program, Vol. 14(3), pp. 130-137, 1980. | | [16] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999 |