www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 1 January 2016, Page No. 15446-15452

Review On An Approach To Infer User Search Goals

Manjushree V. Manwatkar [1]***, Mrs. Arti Waghmare*** [2]

[1]Department of Computer Engg., Dr.D.Y.Patil School of Engg. & technology, Lohgaon Savitribai Phule
Pune University Pune, India
manjushri.gajbhiye@gmail.com
&
[2]Department of Computer Engg., Dr.D.Y.Patil School of Engg. & technology,Lohgaon Savitribai Phule
Pune University Pune, India
arti.waghmare@dypic.in

Abstract: There have been recent interests in studying the goal behind a user's Web query so that this goal can be used to improve the quality of a search engine's results. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, main focus is on the survey of infer user search goal approaches in previous study. Additionally this paper projected a framework to search various user search areas for a query by clustering the feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Additionally, a novel technique is projected to generate pseudo-documents to better represent the feedback sessions for clustering.

Keywords: User search goals, pseudo-documents, feedback sessions, clustering**.**

1. Introduction

Web search engines attempt to satisfy needs of users' information by means of ranking sites with respect to queries. However the fact of web search is that it is usually a process of querying, learning, and reformulating. A series of interactions between user and search engine are often necessary to satisfy a single information need. In web search applications, queries are submitted to search engines to represent the knowledge needs of users. However, generally queries might not precisely represent users' specific information needs since many ambiguous queries could cover a broad topic and totally different users might want to get information on different aspects after they submit the same query. For instance, when the query "the sun" is submitted to a search engine, some users wish to find the homepage of a United Kingdom newspaper, whereas some others wish to find out the natural information of the sun. Therefore, it is necessary and potential to capture totally different user search goals in information retrieval. User search goals will describe because the information on totally different aspects of a query that user groups wish to get. Information need could be a user's specific need to get information to satisfy his/her need. User search goals can be considered because the clusters of knowledge need for a query. The inference and analysis of user search goals will have lots of benefits in up search engine relevance and user expertise. Some benefits are summarized as follows. First, web search results will restructure [13], [14] according to user search goals by grouping the search results with the same search goal; therefore, users with different search goals will easily find what they require. Second, user search goals represented by some keywords can be utilized in query recommendation [2], [4]; thus, the recommended queries will help users to create their queries more exactly. Third, the distributions of user search goals may be helpful in applications such as re-ranking web search results that contain totally different user search goals. Due to its usefulness, several works about user search goals analysis have been investigated. They will be summarized into three classes: query classification, search result reorganization, and session boundary detection. Within the first class, people attempt to infer user goals and intents by predefining some specific categories and performing query classification accordingly. Lee et al. [8] consider user goals as “Navigational” and “Informational” and categorize queries into these two categories. Li et al. [9] define query intents as “Product intent” and “Job intent” and they attempt to classify queries according to the defined intents. Different works specialize in tagging queries with some predefined concepts to enhance feature illustration of queries [12].

However, since what users care regarding varies a lot for different queries, finding appropriate predefined search goal categories is extremely difficult and impractical. Within the second category, people try to reorganize search results. Wang and Zhai [13] learn interesting aspects of queries by analyzing the clicked URLs directly from user click-through logs to arrange search results. However, this technique has limitations since the number of various clicked URLs of a query could also be small. Different works analyze the search results returned by the search engine when a query is submitted [14]. Since user feedback is not considered, several noisy search results that are not clicked by any users may be analyzed as well. Therefore, this kind of methods cannot infer user search goals exactly. Within the third category, people aim at detecting session boundaries. Jones and Klinkner [6] predict goal and mission boundaries to hierarchically segment query logs. However, their technique only identifies whether or not a pair of queries belongs to a similar goal or mission and does not care what the goal is in detail.

Web users have to go through the list and examine the titles and (short) snippets sequentially to identify their needed results. This can be a time-consuming task once multiple sub-topics of the given query are mixed together. An attainable solution to this drawback is to (online) cluster search results into various groups and to enable users to identify their needed group at a glance.

Clustering strategies do not need pre-defined categories as in classification methods. Thus, they are more accommodative for numerous queries. However, clustering strategies are more difficult than classification methods since their results are conducted in a totally unsupervised way. Moreover, most traditional clustering algorithms cannot be directly used for search result clustering, because of some practical difficulties.

The rest of paper is divided into some sections as follows: Section II gives the essential background. Section III addresses feedback session overview. Section IV introduces the mapping concept among feedback session and pseudo documents. Section V describes previous techniques used for clustering and finally section VI concludes the summary of paper.

2. Literature Review

Ricardo Baeza-Yates et al. [2] proposed a technique that, given a query submitted to a search engine, suggests a list of associated queries. The strategy projected is predicated on a query clustering method within which groups of semantically similar queries are known. The clustering method uses the content of historical preferences of users registered within the query log of the search engine.

In [7] R. Jones et al. presented novel document illustration model supported implicit user feedback obtained from search engine queries. Throughout this work they extend and formalize as query model an existing but not very well-known plan of query view for document illustration. Moreover, they produce a unique model supported frequent query patterns referred to as the query-set model.

R. Jones and K. L. Klinkner [6] introduced the notion of query substitution that is, generating novel query to exchange a user's original search query. Their technique uses modifications supported typical substitutions web searchers produce to their queries. During this approach the new query is strongly related to the initial query, containing terms closely related to all of the initial terms. This contrasts with query expansion through pseudo-relevance feedback that is expensive and should cause query drift.

X. Li et al. [9] investigated an orthogonal approach instead of enriching feature representation aiming at drastically increasing the amounts of training data by semi-supervised learning with click graphs. Specifically, they infer class memberships of unlabeled queries from those of labeled ones consistent with their proximities throughout a click graph. Moreover, they regularize the learning with click graphs by content-based classification to avoid propagating inaccurate labels. They demonstrate the effectiveness of algorithms in two completely different applications, product intent, and job intent classification.

S. M. Beitzel et al. [3] examines two previously unaddressed difficulties in query classification: pre vs. post-retrieval classification effectiveness and therefore the effect of training explicitly from classified queries vs. bridging a classifier trained victimization document taxonomy.

H. Cao et al. [4] proposed a context-aware query suggestion approach that represented is in two steps. Within the online model learning step, to deal with information sparseness, queries are summarized into concepts by clustering a click-through bipartite. Then, from session information, a concept sequence suffix tree is constructed because the query suggestion model. Within the online query suggestion step, a user's search context is captured by mapping the query sequence submitted by the user to a sequence of concepts.

T. Joachims et al. [5] presented a comprehensive study addressing the reliableness of implicit feedback for web search engines that combines detailed proof regarding the users' decision method as derived from eye tracking, with a comparison against specific relevance judgments.

Uichin Lee et al. [8] study whether and how can automate goal-identification method. They first present results from a human subject study that strongly indicates the feasibility of automatic query-goal identification. Then proposed two types of features for the goal identification task: user-click behavior and anchor-link distribution.

Xiao Li et al. [9] investigated a totally orthogonal approach like instead of enriching feature illustration, aiming at drastically increasing the amounts of training information by semi-supervised learning with click graphs. Specifically, they infer class memberships of unlabelled queries from those of labeled ones according to their proximities during a click graph. Moreover, regularize the learning with click graphs by content-based classification to avoid propagating incorrect labels.

M. Pasca and B. V. Durme [10] introduced a technique for extracting relevant attributes, or quantifiable properties, for numerous classes of objects. The strategy extracts attribute like the capital city and President for the class Country, or cost, manufacturer and side effects for the category Drug, while not relying on any expensive language resources or complicated process tools. In a departure from previous approaches to large-scale data extraction, and explore the role of web query logs, instead of web documents, as an alternate source of class attributes.

B. Poblete et al. [11] presented a document illustration model supported implicit user feedback obtained from search engine queries. The main objective of this model is to achieve higher results in non-supervised tasks, like clustering and labeling, through the incorporation of usage information obtained from search engine queries.

Dou Shen et al. [12] initially build a bridging classifier on an intermediate taxonomy in an offline mode. This classifier is then utilized in a web mode to map user queries to the target classes via the above intermediate taxonomy. a major innovation is that by leveraging the similarity distribution over the intermediate taxonomy, not need to retrain a new classifier for every new set of target categories, and thus the bridging classifier has to be trained just the once. Additionally, introduce category choice as a new technique for narrowing down the scope of the intermediate taxonomy supported which classify the queries. Category choice will improve each efficiency and effectiveness of the online classification. However, this method does not consist query clustering.

Xuanhui Wang et al. [13] proposed a distinct strategy for partitioning search results that addresses these two deficiencies through imposing a user-oriented partitioning of the search results. Firstly, authors learn “interesting aspects" of comparable topics from search logs and organize search results based on these “interesting aspects". Secondly they generate additional meaningful cluster labels using past query words entered by users. But, this technique has limitations as the number of different clicked URLs of a query may be minor.

3. Feedback Session

Generally, a session for web search may be a series of sequential queries to satisfy one information want and a few clicked search results [6]. They target inferring user search goals for a specific query. Therefore, the single session containing just one query is introduced, that distinguishes from the standard session. Meanwhile, the feedback session relies on one session, though it will be extended to the complete session. The projected feedback session consists of each clicked and unclicked URLs and ends with the last URL that was clicked in a single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks.

Generally speaking, since users can scan the URLs one by one from top to down, considering that besides the three clicked URLs, the four unclicked have also been browsed and evaluated by the user and they should reasonably be a part of the user feedback. Inside the feedback session, the clicked URLs tell what users need and therefore the unclicked URLs reflect what users do not care about. It should be noted that the unclicked URLs when the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not. Every feedback session will tell what a user needs and what he/she does not care about. Moreover, there are many various feedback sessions in user click-through logs. Therefore, for inferring user search goals, it is additional efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

Laura Granka et al. [5] explored and measured strategies for how to automatically generate training examples for learning retrieval functions from determined user behavior. In contrast to explicit feedback, such implicit feedback has the advantage that it can be collected at much lower cost, in much larger quantities, and while not burden on the user of the retrieval system. However, implicit feedback is more difficult to interpret and potentially noisy. They analyze that types of implicit feedback will be reliably extracted from determined user behavior, specifically clickthrough knowledge in web search.

Authors [6] given a combination of four sorts of session feature to evaluate performance on the boundary detection drawback. They considered temporal features, Word and Character Edit Features, query Log Sequence Features, and web Search Features.

4. Map Feedback Sessions to Pseudo-Documents

Since feedback sessions vary a lot for various click-through and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. Some representation methodology is required to explain feedback sessions in a very additional efficient and coherent manner. There are several types of feature representations of feedback sessions like a preferred binary vector methodology. The binary vector is used to represent the feedback session, wherever “1” represents “clicked” and “0” represents “unclicked.” However, since different feedback sessions have different numbers of URLs, the binary vectors completely different feedback sessions might have different dimensions. Moreover, binary vector illustration is not informative enough to tell the contents of user search goals. Therefore, it's improper to use ways like the binary vectors and new ways are required to represent feedback sessions. For a query, users can typically have some vague keywords representing their interests in their minds. They use these keywords to see whether or not a document will satisfy their needs. However, though goal texts will reflect user information needs, they are latent and not expressed explicitly. Therefore, pseudo-documents are introducing as surrogates to approximate goal texts. Thus, pseudo-documents are used to infer user search goals.

However, though goal texts will reflect user info wants, they're latent and not expressed expressly. Therefore, pseudo-documents are introducing as surrogates to approximate goal texts. Thus, pseudo-documents is adapted infer user search goals. The construction of a pseudo-document includes two steps. They are defined in the following:

4.1 Representing the URLs in the feedback session

In the first step, enhance the URLs with extra textual contents by taking out the titles as well as snippets of the returned URLs appearing in the feedback session. In this manner, every URL in a feedback session is characterized by an insignificant text paragraph that contains its title and snippet. After that, some textual procedures are applied to those text paragraphs, for example transformations of all the letters to lowercases, stemming as well as removing stop words. Lastly, every URL’s title is represented by a Term Frequency (TF) vector and snippet is represented Inverse Document Frequency (IDF) vector.

4.2 Developing Pseudo-Document

In order to obtain the feature representation of a feedback session, an optimization method is proposed to combine both clicked and unclicked URLs in the feedback session. It is worth noting that people will also skip some URLs because they are too similar to the previous ones. In this situation, the "unclicked" URLs could wrongly reduce the weight of some terms in the pseudo-documents to some extent. However, method in [1] can address this problem. Let us analyse the problem from three situations.

Situation 1 (the ideal case): one term appears in all the clicked URLs and does not appear in any unclicked ones. In this case, people skip because the unclicked URLs do not contain this important term. The weight of the term in the pseudo-document will be set to the highest value.

Situation 2 (the general case): one term appears in both the clicked URLs and a subset of the unclicked ones. In this case, some unclicked URLs are skipped because they are irrelevant and some are skipped because of duplication. The weight of the term will be reduced to some extent; however, it will not be set to zero. Therefore, skipping because of duplication does not affect too much in this case.

Situation 3 (the bad case): one term appears in both the clicked URLs and almost all the unclicked ones. In this case, people skip because of duplication.

However, when this case happens, both the clicked and the unclicked URLs are almost about one single subject and the term is no longer distinguishable. Therefore, even if people skip some unclicked URLs because of duplication, our method can still assign reasonable weight of the term in most cases.

5. Clustering

C. Hurtado et al. [2] presented a technique for suggesting associated queries supported a clustering method over information extracted from the query log. The presented framework avoids the issues of comparing and clustering sparse collection of vectors.

B. Poblete et al. proposed a unique document illustration model, principally for clustering and labeling, however that may even be used for classification. They formalize a query document model and introduce a new illustration supported frequent query patterns, referred to as the query-set document model. They analyze that each one of the query-based representations outperform the vector space model once clustering and labeling documents of a website [11].

Authors [14] reformalize the search result clustering drawback as a salient phrases ranking drawback. Therefore they convert an unsupervised clustering drawback to a supervised learning drawback. Though a supervised learning technique needs extra training information, it makes the performance of search result grouping significantly improve, and allows us to evaluate it accurately. Their projected technique is a lot of appropriate for web search results clustering because they emphasize the efficiency of identifying relevant clusters for web users. Many properties, moreover as many regression models, are planned to calculate salience score for salient phrase. However these strategies cannot infer user search goal effectively.

With the planned pseudo-documents, user search goals will infer. During this section, how to infer user search goals and depict them with some significant keywords is describe. The clustering method is predicated on a term-weight vector illustration of queries, obtained from the aggregation of the term-weight vectors of the clicked URL's for the query. Authors in [1] perform clustering pseudo-documents by K-means clustering which is straightforward and effective. Once clustering all the pseudo-documents, every cluster is thought of as one user search goal. The center point of a cluster is computed because the average of the vectors of all the pseudo-documents within the cluster. The terms with the highest values within the center points are used because the keywords to depict user search goals [1]. The analysis of user search goal reasoning may be a huge drawback, since user search goals are not predefined and there is no ground truth. Previous work has not planned an appropriate approach on this task. Moreover, since the optimum number of clusters is still not determined once inferring user search goals, feedback information is required to finally verify the most effective cluster number. Therefore, it is necessary to develop a metric to evaluate the performance of user search goal inference objectively. Considering that if user search goals are inferred properly, the search results can even be restructured properly, since restructuring web search results is one application of inferring user search goals. From another purpose of view, feedback sessions may also be viewed as a pre-clustering of the clicked URLs for a more economical clustering.

6. Conclusion

This paper presented a survey of strategies on inferring user search goal. First, introduce feedback sessions to be analysed to infer user search goals instead of search results or clicked URLs. Each the clicked URLs and therefore the unclicked ones before the last click are considered as user implicit feedbacks and taken under consideration to construct feedback sessions. Therefore, feedback sessions will reflect user information desires more efficiently. Second, map feedback sessions to pseudo documents to approximate goal texts in user minds. The pseudo-documents will enrich the URLs with additional textual contents as well as the titles and snippets. Based on these pseudo-documents, user search goals will then be discovered and represented with some keywords. Finally, approach to infer user search goals for a query by clustering is presented.

References

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, and Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", IEEE Transactions on Knowledge and Data Engg., vol. 25, no. 3, March 2013.

[2] 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.

[3] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[4] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[11] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 41-50, 2008.

[12] 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.

[13] 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.

[14] 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.

Author Profile

Manjushree V. Manwatkar has received the 'Bachelor of Engineering' degree in

‘Computer Science and Engineering’ stream from ‘Sipna college of Engineering and Technology, Sant Gadge Baba Amravati University’ in the year 2006. She is student of ME Computer Engineering in Dr. D. Y. Patil School of Engineering & technology, Lohgaon, Savitribai Phule Pune University. And her research interest includes Data mining and Information Retrieval.

Ms. Arti Waghmare received ME Computer Engineering from Thadomal Shahani College of Engineering, Mumbai University. During her ME Research project she worked on the Recommendation System and publishes several papers to address various issues in the area of information storage and retrieval.