ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Using Feedback Sessions for Inferring User Search Goals

Sukanya S. Gawade[1], Gyankamal J. Chhajed[2]

ME student, Department of Computer Engineering, VPCOE, Baramati, Pune university, India [1]

Assistant Professor, Department of Computer Engineering, VPCOE, Baramati, Pune university, India [2]

Abstract: Identifying or inferring user's search goal from given query is a difficult job as search engines allow users to specify queries simply as a list of keywords which may refer to broad topics, to technical terminology, or even to proper nouns that can be used to guide the search process to the relevant collection of documents. Information needs of users are represented by queries submitted to search engines and different users have different search goals for a broad topic. Sometimes queries may not exactly represent the user's information needs due to the use of short queries with ambiguous terms.

Hence to get the best results it is necessary to capture different user search goals. These user goals are nothing but information on different aspects of a query that different users want to obtain. The judgment and analysis of user search goals can be improved by the relevant result obtained from search engine and user's feedback.

Here, feedback sessions are used to discover different user search goals based on series of both clicked and unclicked URL's. The pseudo-documents are generated to better represent feedback sessions which can reflect the information need of user. With this the original search results are restructured and to evaluate the performance of restructured search results, classified average precision (CAP) is used. This evaluation is used as feedback to select the optimal user search goals.

Keywords: AP (Average Precision), CAP (Classified Average Precision), SVM (Support Vector Machine), URL (Uniform Resource Locator), VAP (Voted AP).

I. INTRODUCTION

Web search engines attempt to satisfy user's information needs by ranking web pages with respect to queries. Web search is a process of querying, learning, and reformulating. A series of interactions between user and search engine can be necessary to satisfy a single information need.

For broad queries and topics different users have different ways of representations i.e. different users have different search goals. Sometimes user specific information needs may not be represented by queries since many ambiguous queries may cover a broad topic. Therefore, it is necessary to capture different user search goals. User search goals are information on different aspects of query that user want to obtain. Inference and analysis of user search goals have advantages such as restructure the web search results according to user search goals by grouping the search results with the same search goal, user search goals represented by some keywords can be utilized in query recommendation and distribution of user search goals.

There are three classes representing user search goals:

1. Query classification,
2. Search result reorganization,
3. Session boundary detection.

In first class, some specific classes are predefined and query classification is performed accordingly. User goals are classified into navigational and informational. For navigational, user has particular web page in mind but for informational user's does not have particular page in mind or intends to visit multiple pages. Some other methods used for defining queries as product intent and job intent. Next method defined is tagging queries with some predefined contents to improve feature representation of queries. Disadvantages of this classification are finding suitable predefined search goal class is difficult because what user cares about varies a lot for different queries.

In second class, people try to recognize search results. First method used is learning interesting aspects of queries by analyzing the clicked URLs directly from user click-through logs to organize search results. Limitation of this is number of clicked URL's may be small. Another method used is analyzing the search results returned by a search engine when a query is submitted. But disadvantage of this method is feedback is not taken into account so noisy results that are not clicked by user may be analysed.

In third class, aim is to detect session boundaries. This method predicts goal and mission boundaries to hierarchically segment queries logs. Limitation with this if it only identifies whether a pair of queries belong to same goal and does not care about the goal in detail.

Here, aim is to discover the number of different kinds of user search goals for a query and describing each goal with some keywords. For this purpose first approach is to Cluster the feedback sessions to infer user search goals. Feedback session contains both clicked and unclicked URL's and ends with the last URL that was clicked in a session. The distributions of different search goals can be obtained after feedback sessions are clustered. Then to reflect user information needs effectively map these feedback sessions to pseudo-documents. This is nothing but the optimization method to combine the enriched URL's in a feedback session. CAP(Classified average precision) is used to evaluate the performance of user search goal inference based on restructuring web search results. Using which we can determine number of user search goals for a query.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

II. LITERATURE SURVEY

A. Automatic identification of user goals:

U. Lee, Z. Liu, and J. Cho[2], proposed automatic identification of user search goals. They stated that majority of queries have a predictable goal. Taxonomy of query goals based on two types:

A.1. Navigational queries

In this type, user has a particular web page in mind and is primarily interested in visiting that web page. User may either have visited that site before, or just assumes such a site exists. Here, user's will only visit the correct sites.

A.2. Informational queries

These are the queries where user does not have a particular page in mind or intends to visit multiple pages to learn about the topic. User is exploring WebPages that provide background knowledge about a particular query topic. Users click on multiple results because they do not assume a particular website to be single correct answer.

Here, two features are used for the prediction of user goal:

1. Past user-click behavior:

If a query is navigational, users will primarily click on the result that the user has in mind. Therefore, by Observing the past user-click behavior on the query, we can identify the goal.

2. Anchor-link distribution:

If users associate particular query with a particular website then most of the links that contain the anchor will point to that particular website. Hence by observing the destinations of the links with the query keyword as the anchor, we can identify the potential goal of the query.

Limitations:

User queries are taken from the CS department that may show technical bias and are well crafted. In short, queries given by CS students are potentially work related. So, if we consider user queries by general people characteristics observed may not be true.

B. Web query classification

D. Shen, J. Sun, Q. Yang, and Z. Chen[3], published a work on classifying web queries into a set of target categories where the queries are very short and there are no training data. Here, intermediate taxonomy is used to train classifiers bridging and target categories so that there is no need to collect training data. Classifier bridging is used to map user queries to target categories.

Classification approaches:

B.1. Classification by exact matching

Two categories defined here are intermediate taxonomy and target taxonomy. One or more terms in each node along the path in the target category appear along the path corresponding to the matched intermediate category. For example, the intermediate category contains "Computers\Hardware\Storage" and target category contains "Computers\Hardware". We can directly map intermediate category to target category since both appears along the path "Computers\Hardware\Storage". In this approach, for each intermediate category we can detect whether it is mapped to target categories according to the matching approaches. It produces low recall because many search result pages no intermediate categories.

B.2. Classification by SVM

In this technique, it first constructs training data for target queries based on mapping functions between categories. If an intermediate category is mapped to a target category then the web pages are mapped into train SVM classifiers for the target categories. For each web query classify the query using SVM classifiers. This can improve the recall of classification result.

B.3. Classifiers by bridges

It connects the target taxonomy and queries by taking an intermediate taxonomy as bridge. The intermediate taxonomy may contain enormous categories and some of them are irrelevant to the query classification task corresponding with the predefined target taxonomy. Therefore, to reduce the computation complexity, we should perform "Category Selection".

C. Reorganizing search results

X. Wang and C.-X Zhai[4], proposed clustering of search results which organizes it and allows a user to navigate into relevant documents quickly. This approach organizes search results learned from search engine logs. Steps of this approach are as follows:

Given a query,

1. Get its related information from search engine logs. Working set is formed by using this information.
2. Learn the aspects from information in the working set. These aspects correspond to users interests.
3. Each aspect is labeled with representative query.
4. Categorize and organize the search results of the input query according to the aspects.

First we will find related past queries in our preprocessed history data collection. Next learn the aspects by clustering. And finally categorize the search results using categorization algorithm.

D. Clustering web search results

H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma[5], researched on reformalizing the clustering problem. This approach consists of four steps:

1. Search result fetching
2. Document parsing and phrase property calculation
3. Salient phrase ranking
4. Post-processing.

Given a query and ranked list of search results. Firstly, the whole list of titles and snippets is parsed, extracts all possible phrases from the contents and calculates several properties for each phrase such as document frequencies, phrase frequencies. Then the regression model is applied to combine these properties into a single salience score. Phrases are ranked according to salience score and the top ranked phrases are taken as salient phrases. In post processing, filter out the pure stop words

Disadvantages:

Feedbacks are not considered. So, noisy results that are not clicked by user may be analysed.

E. Session boundaries

R. Jones and K.L. Klinkner[6], defined session boundaries and automatic hierarchical segmentation of

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

search topics. In this approach, analysis of typical timeouts used to divide query streams into sessions and the hierarchical analysis of user search tasks into short-term goal and long-term missions is done.

Timeout is nothing but elapsed time of 30 minutes between queries which signifies that the user has discontinued searching. Here, combination of diverse set of syntactic, temporal, query log and web search features can predict mission boundaries and goals. Hence, best approach to clustering queries within the same goal may build on first identifying the boundaries then matching subsequent queries to existing segments.

Disadvantages:

It only identifies whether a pair of queries belong to the same goal or mission but does not care about what the goal is in detail.

III. SYSTEM OVERVIEW

There are four modules in this system like capturing feedback sessions, building pseudo-documents, clustering pseudo-documents, restructuring based on web search results.

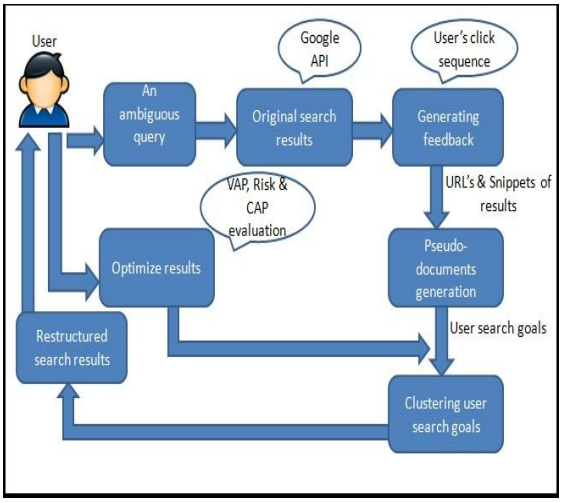

Fig. 1. *System architecture*

A. Mathematical model

Given an ambiguous query, q. When the user submits query search results are obtained on the basis of that query, say

S={s1,s2,s3,s4,…,sn}

First, user will click on some of the results,say{s1,s4,s5} and the click sequence obtained from this is,{s1=1,s4=2,s5=3}. So, the clicked sequence of results is as follows,

{s1=1,s2=0,s3=0,s4=2,s5=3,….sn=0}

One feedback session contains URL's till the last clicked URL. These feedback sessions are represented by, {fs1,fs2,…fsn}. Map these feedback sessions to pseudo-documents to find out the user goals. so, pseudo-documents are created as,{pd1,pd2,..pdn}. Finally, cluster these pseudo-documents to find out similarity,

{pd1=sg1,sg2,…sgn|pd2=sg1,sg2,…sgn|…|pdn=sg1,sg2, ..sgn}

Similarity computation,

simi,j=cos(Ffsi,Ffsj)

Where, Ffs is the feature representation of feedback session. After clustering all the pseudo-documents, each cluster is considered as one user search goal. Evaluation based on web search results[1],

$$AP = \frac{1}{N^+}\left\{\sum_{r=1}^{N} rel(r)\frac{R_r}{r}\right\} \quad (1)$$

N+ is number of relevant documents

r is rank

N is total number of retrieved documents

rel() binary function on the relevance of given rank

Rr is number of relevant retrieved documents

VAP(voted AP) is the AP of the class with more clicks as votes. Here URL's in the single session are restructured into two classes, bold-faced and unbold-faced. VAP is still unsatisfactory. So, there should be a risk to avoid classifying search results into too many classes[1].

$$Risk = \frac{\sum_{i,j=1}^{m}(i<j)^{d_{ij}}}{C_m^2} \quad (2)$$

This calculates normalized number of clicked URL pairs that are not in same class. Here, m is number of clicked URL's[1].

$$C_m^2 = \frac{m(m-1)}{2} \quad (3)$$

is the total number of clicked URL pairs[1].

CAP is extension of VAP as,

$$CAP = VAP \times (1 - risk)^r \quad (4)$$

CAP selects the AP of the class that user is interested in and takes the risk of wrong classification into account. r is used to adjust the influence of risk on CAP.

B. Capturing feedback sessions

Sessions for a web search is a series of successive queries to satisfy a single information need and some clicked search results. Here, feedback session consists of both clicked and unclicked URL's and ends with the last URL that was clicked in a single session. Clicked URL's state what users require and unclicked URL's reflect what users do not care about. For inferring user search goals it is more efficient to analyze the feedback sessions than to analyze search results or clicked URL's directly because there are different feedback sessions in user click-through logs.

It is unsuitable to directly use feedback sessions for inferring user search goals, because they vary a lot for different click-through logs and queries. We can represent feedback sessions by binary vector method. In this method, 0 represents unclicked URL's in click sequence and 1 represents clicked URL's. But, binary vector representations are not informative enough. So, we used pseudo-documents to infer user search goals. Users have some unclear words for representing their interests. They use these keywords to determine whether a document can satisfy their needs. These keywords are known as "goal texts". Goal texts can reflect user information needs, they are hidden and not expressed explicitly. So, pseudo-documents are used as surrogates to approximate goal texts.

C. Building pseudo-documents

This includes two steps

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

- Representing the URL's in feedback session.

Each URL's title and snippet are represented by term frequency-inverse document frequency as below,

Tui=[tw1,tw2,...,twn]T

Sui=[sw1,sw2,...,swn]T

Where Tui and Sui are TF-IDF vectors of the URL's title snippet. ui means ith URL in the feedback session. Wj(j=1,2,..,n) is jth term appearing in the enriched URL.

Fui=wtTui + wsSui =[fw1,fw2,..,fwn]T

Here, Fui is feature representation of ith URL in feedback session. Wt and Ws are weights of title and snippet. Here title should be more significant than snippets. So, the weight of title should be higher.

- Forming pseudo-documents based on URL representations:

Here, an optimization method is used to combine both clicked and unclicked URL's in the feedback sessions. Let Ffs be the feature representation of feedback sessions and ffs(w) be the value for term w. Fucm(m=1,2,..,M) and Fucl(l=1,2,..,l) be the representation of clicked and unclicked URL's in the feedback sessions. Fucm(w) and Fucl(w) are the values of term w in vectors. Obtain such a Ffs that sum of distances between Ffs and each Fucm is minimized and sum of distances between Ffs and each Fucl is maximized. Optimization on each dimension is obtained as follows[1]:

$$F_{fs} = \left[f_{fs}(w_1), f_{fs}(w_2), .., f_{fs}(w_n)\right]$$

$$f_{fs}(w) = arg \min_{f_{fs}(w)} \left\{ \sum_{M} \left[f_{fs}(w) - f_{uc_m}(w)\right]^2 - \lambda \sum_{L} \left[f_{fs}(w) - f_{\overline{uc}_l}(w)\right]^2 \right\}, f_{fs}(w) \in I_c$$

Lamda is balancing and unclicked URL's. when lamda is 0, unclicked URL's are not taken into account.

D. Clustering pseudo-documents

Similarity between two pseudo-documents is computed as cosine score of Ffsi and Ffsj as follows[1]:

$$Sim_{i,j} = \cos\left(F_{fs_i}, F_{fs_j}\right) = \frac{F_{fs_i} . F_{fs_j}}{\left|F_{fs_i}\right| \left|F_{fs_i}\right|} \quad (5)$$

Pseudo-documents are clustered by using k-means clustering algorithm. And the optimal value will be determined through the evaluation. After clustering of all pseudo-documents, each cluster is considered as one user search goal. Centre point is computed as average of all vectors as[1],

$$F_{center\ i} = \frac{\sum_{k=1}^{C_i} F_{fs_k}}{C_i}, (F_{fs_k} \subset cluster\ i) \quad (6)$$

Fcenter is ith cluster centre. Ci is number of pseudo-documents in ith cluster[1].

E. Restructuring based on web search results

Restructuring web search results is an application of inferring user search goals. Inferred user search goals are represented by the vectors and each URL's feature representation is calculated. Then, we can categorize each URL into cluster. This is performed by choosing smallest distance between URL vector and user search goal vectors and the user search goals are restructured. Possible evaluation criteria is Average precision(AP). It evaluates according to user implicit feedbacks. It is computed at the point of each relevant document in ranked sequence.

IV. RESULT ANALYSIS

System relies on the feedback of user. Feedback are then converted into pseudo-documents which represents the keywords from the documents. After that the pseudo-documents are clustered using the k-means clustering algorithm. Results are evaluated using Risk, VAP and CAP. Table 1.1 shows the keywords depiction of different queries. Those are nothing but user search goals.

Snapshots:

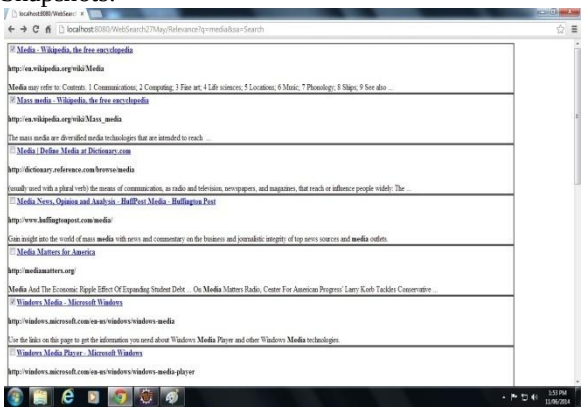

Fig. 2. *Snapshot of original results*

Fig. 3. *Snapshot of restructured results*

Table 1.1 Keyword depiction of different queries

Query	Keywords used to depict user search goals
Taj	India
	Mahal
	Taj
Nasm	Netwide Assembler
	Personal trainer institute, American council
	Wikipedia
Apple	Apple, Wkikpedia
	News
	Official
Vastu	Android apps google play
	Maharshi, Architecture
	Vastu

Table 1.2 shows evaluation of queries such as mean average VAP, risk factor and CAP.

Table 1.2 Query Evaluations

Query	Mean average VAP	Risk	CAP
Nasm	0.705	0.6	0.611
Vastu	0.333	0.2	0.632
Taj	0.444	0.66	0.551

V. CONCLUSION

Proposed approach is used to infer user search goals by clustering the feedback sessions. Feedback sessions consist of both clicked and unclicked URL's before the last click is considered as users implicit feedback. Then feedback sessions are mapped to pseudo-documents to approximate goal texts in users mind. These documents enrich URL's with additional contents including titles and snippets. Based on these documents search goals can be depicted with some keyword. Finally, to evaluate the performance of the user search goals CAP is used. By using this method users can find what they want easily.

REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[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] 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.

[8] 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.

[9] 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.

[10] 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.

BIOGRAPHIES

Sukanya S. Gawade, Received the Bachelor degree (B.E.) in Computer engineering in 2012 from SVPM COE, Malegaon(Bk). She is now pursuing Master's degree in Computer Engineering at VidyaPratishthan's College of Engineering, Baramati. Her current research interests include Data mining & information retrieval.

Gyankamal J. Chhajed, Obtained Engineering degree (B.E.) in Computer Science and Engineering in the year 1991-95 from S.G.G.S.I.E.T, Nanded and Postgraduate degree (M.Tech.) in Computer Engineering from College of Engineering, Pune (COEP) in the year 2005-2007. She is approved Undergraduate and Postgraduate teacher of Pune university and having about 17 yrs. of experience. Her area of interest includes stegnography and watermarking, specially for black and white images, Data mining & information retrieval. She is life member of ISTE, IACSIT. She authored a book & has 18 publications at the national, international level for conferences and journals.
Email:gjchhajed@gmail.com