Enhanced Image Search by Deducing User Search Goals using Query Logs

**Publication History**
Received: 23 January 2015  
Accepted: 3 March 2015  
Published: 16 April 2015

**Citation**
Enhanced Image Search by Deducing User Search Goals using Query Logs

P. Devapriya ¹, T. Poongothai ²

¹M.Tech (IT), Department of IT, K.S.R. College of Engineering, Tiruchengode, India
²Associate Professor, Department of IT, K.S.R. College of Engineering, Tiruchengode, India

Abstract - In web search applications, queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. To define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user's particular desire to obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. To discover the number of diverse user search goals for a query and depicting each goal with some keywords. First, a novel approach is proposed to infer user image search goals for a query by clustering feedback sessions. The feedback session is defined as the series of both clicked and un clicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Then a novel optimization method is proposed to map the feedback sessions to pseudo-documents which can efficiently reflect user information needs. Finally, these pseudo-documents are clustered to infer user image search goals and weighing factor is used to enhance web search.

Keywords: Feedback Session, Logs, Pseudo-documents, Image restructuring

I. INTRODUCTION

In Web search applications, users submit queries (i.e., some keywords) to search engines to represent their search goals. However, in many cases, queries may not exactly represent what they want since the keywords may be polysemous or cover a broad topic and users tend to formulate short queries rather than to take the trouble of constructing long and carefully stated ones. Besides, even for the same query, users may have different search goals. We find that users have different search goals for the same query due to the following three reasons. Inferring user search goals is very important in improving search-engine relevance and user experience. Normally, the captured user image-search goals can be utilized in many applications. For example, we can take user image-search goals as visual query suggestions to help users reformulate their queries during image search. Besides, we can also categorize search results for image search according to the inferred user image-search goals to make it easier for users to browse. Furthermore, we can also diversify and re-rank the results retrieved for a query in image search with the discovered user image-search goals. Thus, inferring user image-search goals is one of the key techniques in improving users’ search experience. However, although there has been much research for text search, few methods were proposed to infer user search goals in image search. Some works try to discover user image-search goals based on textual information (e.g., external texts including the file name of the image file, the URL of the image, the title of the web page that contains that image and the surrounding texts in image search results and the tags given by users. However, since external texts are not always reliable (i.e., not guaranteed to precisely describe the image contents) and tags are not always available (i.e., the images may not have corresponding tags that need to be intentionally created by users), these textual information based methods still have limitations. It should be possible to infer user image-search goals with the visual information of images (i.e., image features) since different image-search goals usually have particular visual patterns to be distinguished from each other. However, since there are semantic gaps between the existing image features and the image semantics, inferring user image-search goals by visual information is still a big challenge. Therefore, in this paper, we propose to introduce additional information sources to help narrow these semantic gaps.

II. LITERATURE SURVEY

U. Lee, Z. Liu, and J. Cho [1], 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:

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 will only visit the correct sites.

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 keywords 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.

X. Wang and C.-X Zhai[2], 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.

H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma [3], 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.

R. Jones and K.L. Klinkner [4], defined session boundaries and automatic hierarchical segmentation of 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.

Wangmeng Zuo, Lei Zhang and Chunwei Song [5] uses natural image statistics in denoising image and propose a texture enhanced image denoising method by enforcing the gradient histogram of the denoised image to be close to a reference gradient histogram of the original image. This idea has to be considered for effecting searching and also Lai-Kuan Wong; Kok-Lim Low [6] proposed 5-fold cross-validation (5-CV) classification accuracy for enhancing image quality there by improving classifications in web sites.

Jayaratne, Kithangodage Lakshman [7] proposed an image retrieval system that captures semantics of an image through effective use of its associated text and use integrated system architecture for keyword-based retrieval with low-level image features to enhance retrieval of images on the web. The system was based on an enhanced image representation that exploits the vast power of image semantics from the text associated with the images and higher-level semantic categories based on low-level image features of the images. The user-interface was designed to allow the user to communicate keywords based query and semantic categories to the image retrieval system. They confirmed that the integration of text associated with an image and low-level image features will lead to efficient retrieval system for content-based indexing of images on the web and will in fact substantially enhance the image searching capabilities on the web. [7] shows highly efficient search methods in retrieving image.

III. MODULE DESCRIPTION

User Query Module

User query is one of the inputs to be entered by the user. The user should enter the query from which he / she require particular image. When the user enters the query, the query is passed on to further modules for it to be processed.

User Click Logs Module

The process of user click logs is maintaining the information about previous user clicking on an image. The click rate measure the amount of times an image is clicked versus the amount of times it’s viewed.

Feedback sessions Module

Feedback sessions are considered as user’s 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 image search 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 images for a particular query in a single session and ends with last clicked image. This shows that before last clicked image, all the images are scanned and evaluated by user.

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

The seven URLs consist of three clicked URLs and four unclicked URLs in this example. Generally speaking, since users will scan the URLs one by one from top to down, we can consider that besides the three clicked URLs, the four unclicked ones in the rectangular box 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 require and the unclicked URLs reflect what users do not care about. It should be noted that the unclicked URLs after the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not. Each feedback session can tell what a user requires and what he/she does not care about. Moreover, there are plenty of diverse feedback sessions in user click-through logs. Therefore, for inferring user search goals, it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

Building pseudo-documents Module

An image alone is 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 preprocessing 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.

The URLs returned by the search engine when the query “the sun” is submitted, and “0” represents “unclicked” in the click sequence. The binary vector [0110001] can be used to represent the feedback session, where “1” represents “clicked” and “0” represents “unclicked.” However, since different feedback sessions have different numbers of URLs, the binary vectors of different feedback sessions may have different dimensions. Moreover, binary vector representation is not informative enough to tell the contents of user search goals. Therefore, it is improper to use methods such as the binary vectors and new methods are needed to represent feedback sessions.

In order to cluster pseudo-documents with k-means, the important factor is to define the distance measure between two data points and defining the number of clusters. There are two variations of distance measures, one is derived from cosine similarity and the other is derived from Jacard similarity coefficient.

K-means algorithm is used to cluster pseudo-documents because of its simplicity and effectiveness. K-means clustering results in good quality performance for document clustering. As a prior number of user search goals for a query are unknown so we have chosen arbitrary value for k initially. Then, perform clustering on these five different values. The optimal value for k is determined by evaluation criterion. After clustering all pseudo-documents, each cluster denotes user search goal i.e. intention of user. Centroid of a cluster is calculated by taking average of all the vectors of the pseudo-documents in the cluster. User search goals/intents depicted with the terms with highest values in the center points of each cluster. These depicted keywords can be used to suggest more meaningful and precise query.

**K-means algorithm:**

- Let $X=\{x_1,x_2,x_3,\ldots,x_n\}$ be the set of data points and $V=\{v_1,v_2,\ldots,v_c\}$ be the set of centers.
- 1. Randomly select ‘c’ cluster centers.
- 2. Calculate the distance between each data point and cluster centers.
- 3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
- 4. Recalculate the new cluster center using:

$$V_=-\left(1/c_i\right)\sum_{j=1}^{c_i}x_i$$

Where, $c_i$ represents the number of data points in $i^{th}$ cluster.
- 5. Recalculate the distance between each data point and new obtained cluster centers.
- 6. If no data point was reassigned then stop otherwise repeat from step 3.

Restructuring image search Module

Web search results are reorganized on the basis of discovered user search goals/intents. As inferred user search goals are depicted with vectors and feature representation of each image in search result is calculated. Then categorize each image into a cluster centered with user search goals by selecting smallest distance between user search goal vectors and image vectors. The images chosen by the users are collected and the similarity score has to be found according to the query words and image attribute content. The similarity score is calculated by for each image by summing the matching score of all index terms that match query terms. This phase represents image bya vector taking on a value based on the presence or absence of the word indexed within a global word list. Queries are represented in a similar manner. Determining whether an image is relevant for a given query involves computing a similarity measure between the image and query. This is here done by cosine of the angle between the two vectors.
The weight vector for document \(d\) is

\[
v_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T, \text{ where}
\]

\[
w_{t,d} = t_\text{f}_{t,d} \cdot \log \frac{|D|}{|\{d' \in D | t \in d'\}|}
\]

and

- \(t_\text{f}_{t,d}\) is term frequency of term \(t\) in document \(d\) (a local parameter)
- \(\log \frac{|D|}{|\{d' \in D | t \in d'\}|}\) is inverse document frequency (a global parameter). \(|L|\) is the total number of documents in the document set; \(\{d' \in L | t \in d'\}\) is the number of documents containing the term \(t\).

Using the cosine the similarity between document \(d_j\) and query \(q\) can be calculated as:

\[
sim(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^{n} d_{ij} q_i}{\sqrt{\sum_{i=1}^{n} d_{ij}^2} \sqrt{\sum_{i=1}^{n} q_i^2}}
\]

IV. EXPECTED RESULT AND DISCUSSION

Restructuring the search results is an application of inferring user search goals. The technique to restructure the search results to infer the user goals are introduced. Then the evaluation based on restructuring search results is also described.

The inferred user search goals are represented by the vectors and the feature representation of each image in search results can be computed. Here the categorization by choosing the smallest distance between the images of the vector and the user search goal vectors. By this way, the search results can be restructured according to the inferred user image search goals.

V. CONCLUSION

To concentrate on designing a new machine learning method for auto classification and grouping similar user queries for image search system to address a specific kind of image search. This approach, search most relevant images for a user query. Here focused to examine the effective association between User Queries and Click through data log and updates search results according to each individual preferences/interests. The leverage clicks session information and combines it with image visual information to infer user image-search goals. Click session information can serve as the implicit guidance of the past users to help clustering. Based on this framework, propose two strategies to combine image visual information with click session information. The weighing factor also provides effective analysing of images thereby evaluating the computational value easier.

ACKNOWLEDGEMENT

The author would like to thank the respective Head of the Institution, Head of the Department and Faculty members for giving valuable advices and providing technical support.

REFERENCES


