ISSN:0975-9646

Shamali K. Kherdikar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 4911-4915

A Novel Approach for Auto Classification and Grouping Similar User Query for Image Search

Shamali K. Kherdikar, Rajesh Kulkarni

TSSM's Bhivarabai Sawant College of Engineering & Research,
Narhe, Pune, India 41

***Abstract*— Image retrieval and re-ranking as per user image query has become the popular and effective of image retrieval techniques. Similar user query and click through log is important for the success of an image search engine. User search goal analysis will also enhance user experience of a search engine. Using this as a base and leveraging click logs we propose a new design in this paper. In this paper, we focus on designing a new machine learning approach for auto classification and grouping similar user queries for image search system to address a specific kind of image search. Our approach finds most relevant images for a user based on a given user query. Here, our focus is to evaluate the effective association between User Queries and Click through data and customizes search results according to each individual preferences/interests. We also present a ranking procedure to score the images that are retrieved using the proposed approach.**

***Keywords* - Search Engines, machine learning, Information Retrieval, Image search, auto classification, image search.**

I. INTRODUCTION

Search engine services are a popular means for information searching. They provide a simple and direct way of searching information for various resource types, not only textual resources, but also multimedia [1] [2] [3]. Most search engines present similar interfaces allowing people to: submit a query; receive a set of results; follow a link; explore the information space; and modify a query [4] [5] [6]. This process is generally repeated during interactive searching. The popular use of search engine services has led to many investigations of general search habits on the Web. Querying behaviour – query formulation and reformulation – has especially been an active area of research in information retrieval.

Digital image is nowadays the second most prevalent media in the Web only after text. Image search engines play an important role in enabling people to easily access to the desired images. A variety of search interfaces have been employed to let users submit the query in various forms, e.g., textual input, image input, and painting based input, to indicate the search goal. To facilitate image search, query formulation is required not only to be convenient and effective for users to indicate the search goal clearly, but also to be easily interpreted by image search engines. Therefore, recently more and more research attention has been paid on search interface design in developing image search engines.

In this paper, we focus on designing a new machine learning approach for auto classification and grouping similar user queries for image search system to address a specific kind of image search. Our approach finds most relevant images for a user based on a given query. Here, our focus is to evaluate the effective association between User Queries and Click through data and customizes search results according to each individual preferences/interests. We also present a ranking procedure to score the images that are retrieved using the proposed approach.

II. RELATED WORK AND EXISTING SYSTEM

Since last some years, the research on text based image search has been increased, but in fact, their works belong to query classification. Some works analyse the search results returned by the search engine directly to show different query aspects [6], [20]. However, query aspects without ranking procedure have limitations to improve search engine relevance. Some works take user feedback into account and analyse the different clicked URLs of a query in user click-through logs directly; nevertheless the number of different clicked URLs of a query may be not big enough to get ideal results

The above mentioned Image search engines provide an effortless route, but currently are limited by poor precision of the returned images and also restrictions on the total number of Images provided. While several studies reveal general characteristics of image searching based on transaction log data, little has been investigated concerning whether or not image searching behaviour, especially querying behaviour – query iterations and query length – differs based on a user‘s contextual aspects and different sources of collections on Web search engines. The existing methods for image searching and ranking suffer from the unreliability of the assumptions under which the initial text-based image search results. However, producing such results containing a large number of images gives more number of irrelevant images.

The existing methods for image searching and re-ranking suffer from the unreliability of the assumptions under which the initial text-based image search result. However, producing such results contains a large number of images and with more number of irrelevant images.

A. Text Based Image Retrieval

This one is very popular framework of image retrieval then was to first annotate the images by text and then use text-based database management systems (DBMS) to perform image retrieval. Many advances, such as data modeling, multidimensional indexing, and query evaluation, have been made along this research direction.

There are two disadvantages to use this image retrieval system, especially when the size of image collections is large (tens or hundreds of thousands). One is the vast amount of labor required in manual image annotation. The other difficulty, which is more essential, results from the rich content in the images and the subjectivity of human perception. That is, for the same image content different people may perceive it differently. The perception subjectivity and annotation impreciseness may cause unrecoverable mismatches in later retrieval processes.

B. Content Based Image Retrieval

The emergence of large-scale image collections, the two difficulties faced by the manual annotation approach became more and more acute. To overcome these difficulties, content-based image retrieval (CBIR) was proposed. That is, instead of being manually annotated by text-based key words, images would be indexed by their own visual content, such as color and texture. Since then, many techniques in this research direction have been developed and many image retrieval systems, both research and commercial, have been built. The advances in this research direction are mainly contributed by the computer vision community.

Text Based Image Retrieval led to two disadvantages. First one is that a considerable level of human labor is required for manual annotation. The second is the annotation inaccuracy due to the subjectivity of human perception. The current CBIR systems suffer from the semantic gap. Though a user feedback is suggested as a remedy to this problem, it often leads to distraction in the search. To overcome these disadvantages we propose a novel interactive image retrieval system, to enhance the image retrieval accuracy as per the user expectation.

III. BACKGROUND

We propose to leverage click session info, that indicates high correlations among the clicked pictures in a very session in user click-through logs, and mix it with the clicked images' visual info for inferring user image-search goals. The click session info will function past users' implicit guidance for cluster the photo graphs; a lot of precise user search goals may be obtained.

A. Image Classification

Image classification is the process of grouping of similar types of image into a single unit i.e. called cluster of image. Content-based image classification is aimed at efficient classification of relevant images from large image databases based on automatically derived imagery features. These imagery features are typically extracted from shape, texture, color properties of query image and images in the database. Potential application includes digital libraries, commerce, Web searching, biomedicine, surveillance, geographic information systems and sensor systems, education, commerce, crime prevention, etc.

B. Grouping Similar User Query

Grouping Query is a process used to discover frequently asked questions or most popular topics on a search engine. Despite the fact that keywords are not always good descriptors of contents, most existing search engines still rely solely on the keywords contained in documents and queries to calculate their similarity. This is one of the main factors that affect the precision of the search engines. In many cases, the answers returned by search engines are not relevant to the user's information need, although they do contain the same keywords as the query.

The queries submitted by users are very different, however, and they are not always well-formed questions. In order to group queries, two related problems have to be solved: (1) How can human editors determine which questions/ queries are frequently raised by users? (2) How can a system judge if two questions/queries are similar?

The classic approach to information retrieval (IR) would suggest a similarity calculation between queries according to their keywords. However, this approach has some known drawbacks due to the limitations of keywords. In the case of queries, in particular, the keyword-based similarity calculation will be very inaccurate (with respect to semantic similarity) due to the short lengths of the queries.

C. Clickthrough Data

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.

The click-through information from the past users can provide good guidance about the semantic correlation among images. By mining the user click-through logs, we can obtain two kinds of information: the click content information i.e., the visual information of the clicked images and the click session information i.e., the correlation information among the images in a session. Commonly, a session in user click-through logs is a sequence of queries and a series of clicks by the user toward addressing a single information need. Whereas query logs are auto saved data of user activities on search engines servers. It consists of user identity attributes as Session ID, IP address, Time-stamp, Query string, Number of results on results page and Results page number. A relevance clickthrough data also saved consisting of clicked URL, associated query, position on results page and Time-stamp attributes in the log. The application used in client side can be modified to handle the query and clickthrough usage logs in the user side computer.

Clickthrough data in search engines can be thought of as triplets (q, r, c) consisting of the query q, the ranking r presented to the user, and the set c of links the user clicked on. Since every query corresponds to one triplet, the amount of data that is potentially available is virtually unlimited.

Clickthrough data can be recorded with little overhead and without compromising the functionality and usefulness of the search engine. In particular, compared to explicit user feedback, it does not add any overhead for the user. The query q and the returned ranking r can easily be recorded whenever the resulting ranking is displayed to the user. For recording the clicks, a simple proxy system can keep a log file. Usually, the clicked images in a session have high correlations. This correlation information provides hints on which images belong to the same search goal from the viewpoint of image semantics.

IV. PROPOSED SYSTEM

Image search engines apparently provide an effortless route, but currently are limited by poor precision of the returned images and also restrictions on the total number of Images provided. While several studies reveal general characteristics of image searching based on transaction log data, little has been investigated concerning whether or not image searching behavior, especially querying behavior – query iterations and query length – differs based on a user‘s contextual aspects and different sources of collections on Web search engines. The existing methods for image searching and ranking suffer from the unreliability of the assumptions under which the initial text-based image search results. However, producing such results containing a large number of images gives more number of irrelevant images.

Machine learning algorithms have received a wide attention recently to learn functions that can perform desired operations when trained on required amount of data. The previous history of the user, can we learn a model representing the user. This makes user modelling a perfect application for machine learning.

We proposed on designing a new machine learning approach for auto classification and grouping similar user queries for image search system to address a specific kind of image search. Our approach finds most relevant images for a user based on a given query. Here, our focus is to evaluate the effective association between User Queries and Clickthrough data and customizes search results according to each individual preferences/interests. We also present a ranking procedure to score the images that are retrieved using the proposed approach. This approach capture user search goals in image search by exploring images which are extracted by mining single sessions in user click-through logs to reflect user information needs. Moreover, we also propose a novel evaluation criterion to determine the number of user search goals for a query. Experimental results demonstrate the effectiveness of the proposed method

Figure-1 describes the proposed system architecture for Image Classification and Grouping based on User Query and Clickthrough Data process. The system architecture consists of four major components as Query Handler, Query Formulation, Event Handler and Result Handler, which implement the algorithm for Image Classification and Grouping based on User Query and Clickthrough Data. It has log repositories which stores user query logs and clickthrough data. A Semantic similarity-based Matching algorithm will be implemented for classification and Grouping the search image results.

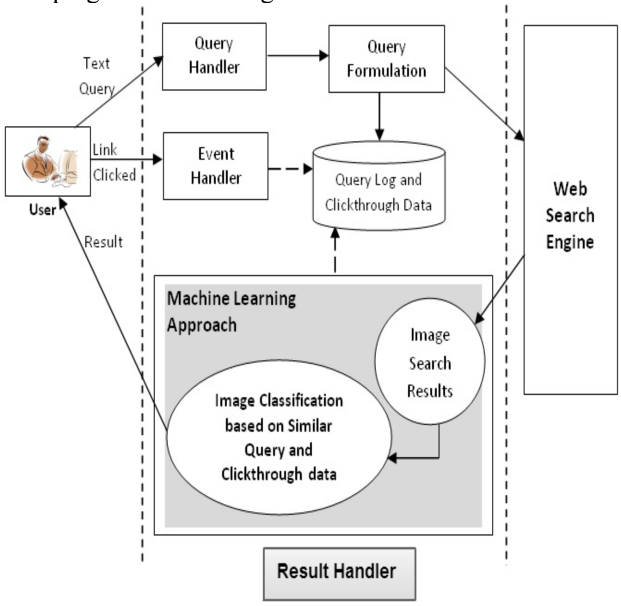

Fig-1: **Proposed System Architecture**

Input:
User input Query *(Q)*, and
Clickthrough data *(C)* from the database
Output: Semantically Associated and Cluster Results (S_R)
In relevance to the user queries
Begin
Create an empty cluster vector as E_C
Create keywords, K from Q.
For each keyword of (K) ***do***
Select $K(i) \rightarrow k_w$
For each click through data in (C) ***do***
Select $C(i) \rightarrow C_w$
If Compute Association $(k_w \in C_w) == true$
If Cluster E_C does not contain $k_w == true$
Add C_w to Cluster E_C
End if
End if
End for
End for
Create an empty object vector as S_R
For each objet in Cluster E_C ***do***
Object Count $(O_C) \rightarrow 0$
Select $E_C(i) \rightarrow O_w$
For each clickthrough data in (C) ***do***
Select $C(i) \rightarrow C_w$
If Compute $(O_w \in C_w) == true$
$O_C = O_C + 1$
End if
End for
Update $S_R(i) \rightarrow O_C$
End for
End

Fig-2: Semantic Similarity Algorithm

In general, when a user pose a query, the user usually navigates the entire result links list from top to bottom in a page. User generally clicks one or more result link that looks appropriate and relevance and skips those links which are not relevant. Effective information retrieval is achieved when a precise personalization approach perform re-ranking of the relevant links and place it in higher in results list.

Input:
Cluster Results (S_R)
Output:
Re-Ranked Results (R_{Rank})
Begin
Create an empty Object_Rank vector as, R_{Rank}
For each record in S_R ***do***
Select $S_R(i)$ -> $FR_{(i)}$
Intial_Top_Rank (T_{Rank}) -> $FR_{(i)}$
For each record in S_R ***do***
$k = i+1$
Select $S_R(k)$ -> $NR_{(k)}$
If $NR_{(k)} > FR_{(i)}$
$T_{Rank} = NR_{(k)}$
End if
End for
Update R_{Rank} -> T_{Rank}
Remove T_{Rank} object from S_R
End for
End

Fig-3: Re-ranking Algorithm

Therefore, we utilize user clicks as relevance decision measure to evaluate the searching accuracy. Since clickthrough data can collect straightforward with less effort, it is possible to do required behavior and interest evaluation implementing this framework. Moreover, clickthrough data shows the actual real world distribution of user search interest queries, and searching scenarios. Therefore, using clickthrough data makes a closer real time personalization requirement cases in compare user feedback survey.

V. METHODOLOGY

When a query submitted by a user it received by the request handler. The submitted query might not be in appropriate structure for submitting to a search engine. Query Handler process this query to filter and prepare the keywords and phrases which are submitted to the search engine, and at the same time it update the Query log database. On retrieval of image search result from search engines Result Handler filter the duplicate result and organize the image as per the search engine ranking. The organize result under goes re-ranking process with support of Query log and clickthrough data recorded for this user query.

Execution of classification and grouping process reorder the organized data as per the user passed interest and a relevant personalized result presented to the user. It may possible the presented result may not be so relevant to user needs. To make it more precise each result link bounds with a click event. On clicking a particular link on the page event handler listen and records the clicked link data to click through database. The continuous of these activities improve the relevancy as click through data against a query increases for a user.

VI. SYSTEM MODULES

The framework consists of below modules:

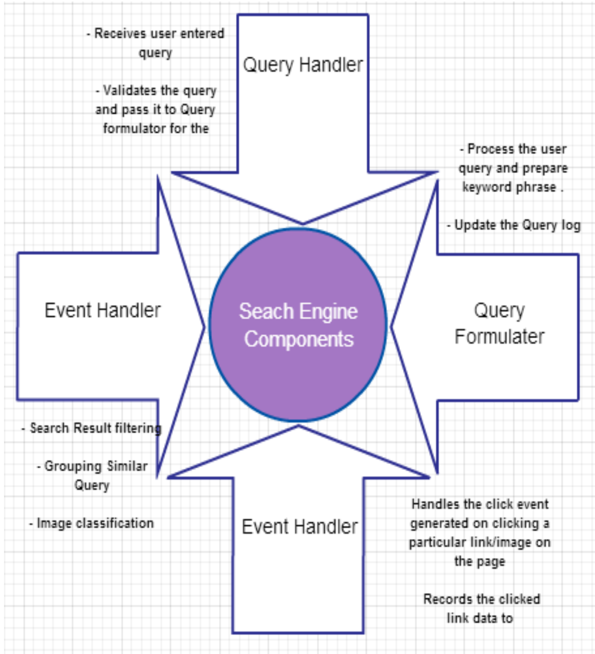

Fig.4.Components of System Architecture

A. *Query Handler*

Query submitted by a user it received by the request handler. This component maintain request load by queuing the request and provide inputs for query processing. It initially validates the query before proceeding for query formulation.

B. *Query Formulation*

Query Formulation is a key component of the framework which process the user query and prepare keyword phrase which pose to search engine for searching. The submitted query might not be in appropriate structure for submitting to a search engine. Query formulator processes the user query to prepare the keywords and phrases, and at the same time it updates the Query log database.

C. *Event Handler*

Event handler handles the click event generated on clicking a particular link/image on the page, and records the clicked link data to clickthrough database. The result provides by the result handler appends event methods to the links to detect the event and helps to records the clickthrough by the user. The recorded clickthrough log helps in machine learning to makes search more precise on new searching by the user.

D. *Result Handler*

Result Handler is the key component of the framework which handles the search results return by the search engine. It does the following activities for improvising the search results.

1. Search Result filtering by removing duplicate results and organize as per original ranking.
2. Grouping Similar Query using Clickthrough database based on query semantic similarity approach.

3. Semantically Image classification through machine learning using Density-based method on clickthrough database related to user query.

Re-Ranking the search result based on the image classification and query similarity.

VII. CONCLUSION

In this paper, we concentrated 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, we focused to examine the effective association between User Queries and Click through data log and updates search results according to each individual preferences/interests. We also showed a ranking method to score the images that are retrieved using the proposed approach.

REFERENCES

[1]. J. Cui, F. Wen, and X. Tang. Intentsearch: interactive on-line image search re-ranking. In MM ’08, pages 997–998, 2008.

[2]. Y. Luo, W. Liu, J. Liu, and X. Tang. Mqsearch: image search by multi-class query. In CHI ’08, pages 49–52, 2008.

[3]. G. P. Nguyen and M. Worring. Optimization of interactive visual-similarity-based search. TOMCCAP, 4(1), 2008.

[4]. R. Yan, A. Natsev, and M. Campbell. Multi-query interactive image and video retrieval -: theory and practice. In CIVR ’08, pages 475–484, 2008.

[5]. Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In MM ’09, pages 15–24, 2009.

[6]. F. Mahmoudi, J. Shanbehzadeh A. Eftekhari-Moghadam and H. Soltanian- Zadeh, “Image retrieval based on shape similarity by edge orientation autocorrelogram,” IEEE Trans. on Pattern Recognition, vol. 36, pp. 1725-1736, 2003.

[7]. U. Lee, Z. Liu and J. Cho, “Automatic identification of user goals in web search,” WWW, pp. 391-400, 2005.

[8]. X. Li, Y-Y. Wang and A. Acero, “Learning query intent from regularized click graphs,” SIGIR, pp. 339-346, 2008

[9]. X. Wang and C-X. Zhai, “Learn from web search logs to organize search results,” SIGIR, pp. 87-94, 2007.

[10]. H-J. Zeng, Q-C. He, Z.Chen, W-Y. Ma and J. Ma, “Learning to cluster Web search results,” SIGIR, pp. 210-217, 2004.

[11]. Z-J. Zha, L-J. Yang, Z-F. Wang, T-S. Chua and X-S. Hua, “Visual query suggestion: towards capturing user intent in internet image search,” ACM Trans. On Multimedia Comput. Commu. Appl. 6, 3, Article, August 2010.

[12]. Xiao gang Wang, Ken Liu and Xiao Tang „Query-Specific Visual Semantic Spaces for Web Image Re-ranking.. In Proceeding of the 14th ACM International Conference on Multimedia,(2011) .

[13]. Y. Rui, T. S. Huang, and S. Mehrotra, “Content-based image retrieval with relevance feedback in MARS,” in Proc. IEEE Int. Conf. Image Process., 1997, pp. 815–818.

[14]. [1]. L. Chen, D. Xu, I. W. Tsang, and J. Luo, “Tag-based web photo retrieval improved by batch mode re-tagging,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn., 2010, pp. 3440–3446.

[15]. [2]. W. H. Hsu, L. S. Kennedy, and S.-F. Chang, “Video search reranking via information bottleneck principle,” in Proc. 14th ACM Int. Conf. Multimedia, 2006, pp. 35–44.

[16]. [3]. Y. Liu, D. Xu, I. W. Tsang, and J. Luo, “Textual query of personal photos facilitated by large-scale web data,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 1022–1036, May 2011.

[17]. [4]. C. Zhang, J. Y. Chai, and R. Jin, “User term feedback in interactive text-based image retrieval,” in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 51–58.

[18]. [5]. Z.-H. Zhou and H.-B. Dai, “Exploiting image contents in web search,” in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 2928–2933.

[19]. [6]. Google Image Search. http://images.google.com/.

[20]. [7]. Microsoft Bing Image Search. http://images.bing.com/