Personalized and Image based Search System to Improve Search Results

Sneha A. Taksande[1], A. V. Deorankar[2]

[1]PG Scholar, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India

[2]Associate Professor, Department of Information Technology, Government College of Engineering Amravati, Maharashtra, India

Abstract: ***In today's condition, information is the need of society as it has significant importance for contemporary success in every field. As there is vast growth of information available to the end users on the Web through internet, a search engine come to play a crucial role in information retrieval. Various search engines are there but sometimes they can be insufficient to specify the precise users need as the queries may be short and ambiguous. They also ignores the images in HTML pages. As users understand the pictures more easily than thousands of words, pictures may also help to the users who don't have good hand at searching. This paper present an approach to design a personalized search system based on user interest to improve the result delivery. Several parameters for personalization can be taken as users profile, his search history, preferences, ratings, etc. It also gives image based more relevant web pages re-ranked technique. The accuracy of web search results can be improve by considering the textual contents of the images in correlation with the personalization. Thus the integration of both the mechanism implies to enhance the search results.***

Keywords: Search Engine, information retrieval, Web Search, Personalization.

1. Introduction

As we know World Wide Web contains vast amount of interlinked HTML web documents. Though there are different search available using which it is easy to retrieve the information. But retrieving the most relevant information is still a challenging task. Since the users depends upon the search engine for accessing a great range of information from such a huge collection. So when a user submits his query to find some information, a search engine must be able to retrieve the documents that satisfies user's specific need. But search engine shows the list of ranked documents link based upon the words present in the query, they shows only the text documents. Pictures in HTML pages are ignored by them. Sometimes these results do not match with users interests either because of absence of important terms in the query or query words are ambiguous. It might be the case that two users uses the same words from different perspective. For example, two users searching for the term "Apple" having different intensions. In such cases we cannot determine whether this term is related to fruit or iPhone. For this problem personalization can be the solution that implies to the fact that, if user is from IT background then should get the result for iPhone in the top rank and other search results below it. In this paper we propose a novel document retrieval approach that incorporates the users profile and also uses the images in web pages to find the relevant documents. It uses the content of the pictures in the Web pages to boost the accuracy of pure text-based search engines. Web personalization means customizing the web environment according to user profile or users interests which can be inferred from user's action, browsed page history, users preferences and ratings, etc. The proposed method uses personalization. When user specifies the query, the search system fetches data from google using googles API. This data is stored in system database. By analyzing the query, system will formulate the users query. The searched data of the google is then re-ranked according to user's profile, users search history, preferences and ratings given by the other users. Also the images in web documents are used to get most relevant documents.

This paper is organized as follows: section 2 describes the related work in this area. Proposed approach is given in section 3. Finally conclusion is given in section 4.

2. Related Work

Andrew W. Fitzgibbon, *et al.* [1] uses the modern methods and representations for image understanding, to improve web document search. In this, the work is focused on re-ranking strategy. It has used the text features as relevance score and ranking position of document in the ranking list. The 'relevance score' feature is a numerical value indicating the relevancy of the document for query. The 'ranking position' is the position of document in the ranking list. By including these two features we can leverage the high-accuracy of modern text-based search.

Xiaoou Tang, *et al*, [2]this propose a novel Internet image search approach which only requires one-click user feedback. Expanded keywords are used to extend positive example images and also enlarge the image pool to include more relevant images. This framework makes it possible for industrial scale image search by both text and visual content.

Shipra Kataria and Pooja Sapra, [3] proposed a new methodology for rank improvement using search engine query logs. The most important part of this architecture is the use of Panda algorithm to find the relevancy of URLs based on the relevancy of content corresponding to them.
Shilpa Sethi and Ashutosh Dixit [4] gives personalization as a vital tool to meet the user's information need, as it gives

the results which are more relevant to the user. The advantage of proposed system is that it reduces the language gap between the user and the search engine. Further, the results which are more relevant to the users query but low in rank are up-ranked by proposed method by maintaining user browsing history. So, this made another advantage of relevancy increase.

This paper [6] will articulate the requirements of Web Document Clustering and reports on the clustering methods belonging in this domain. The work is focus on the methods that create their clusters based on the characters or individual terms rather than showcasing them as a single phrase with a meaning and sequence of words. For that, it has uses the different clustering algorithms such as Hierarchical Agglomerative Clustering (HAC), K-Mean

3. Proposed Approach

Various personalized methods are there which suggest that, if user's behavior is taken into consideration, the search results can be improved significantly. Sometimes the users are not able to formulate their query well, so this proposed methodology can work for them.

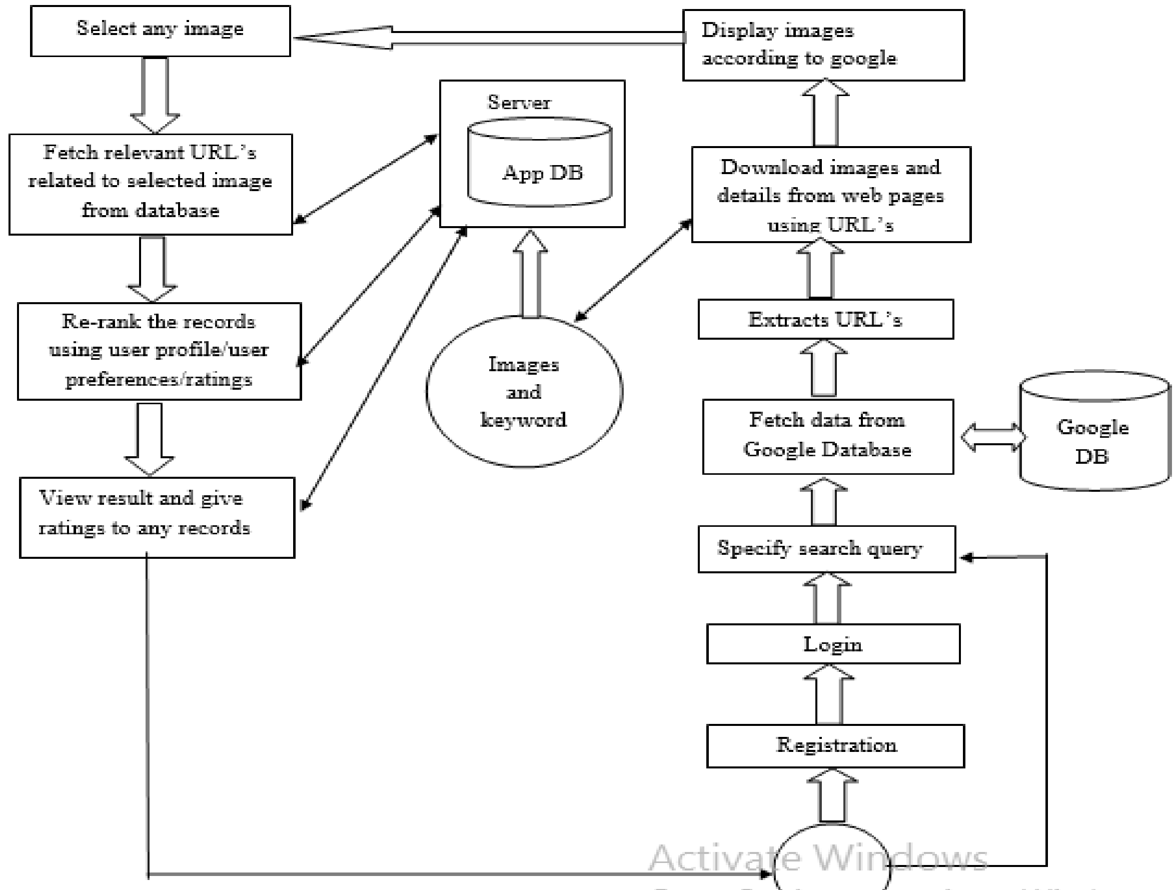
Select any image
Display images according to google
Server
App DB
Fetch relevant URL's related to selected image from database
Download images and details from web pages using URL's
Re-rank the records using user profile/user preferences/ratings
Images and keyword
Extracts URL's
Fetch data from Google Database
Google DB
View result and give ratings to any records
Specify search query
Login
Registration

Figure 1: Architecture for Proposed System

The proposed method has the following modules.
1) User management
2) Relevant Query Suggestions
3) Google Data Extraction
4) Web page images downloads
5) Keywords storage
6) Image based search system

The system architecture is as given in figure 1.The steps for above workflow can be given as below:
1) Specify query in search box.
2) If user is logged in user then users profile wise query reformation is done.
3) Send this query to google database and get google data.
4) Result set will be re-rank using previous search history/ratings if available.
5) Otherwise user can use image based approach by selecting a query related specific image, which is extracted from google database.
6) Further the most relevant documents according to image selection will be re-ranked by comparing the keywords of query with those present in HTML web pages containing the images.

The description for each module of system architecture is as given below.

A. User management

This module generates the users profile and provides the following facilities.
- Registration
- Login/logout
- Use search engine

- View search history
- Change password
- Password recovery
- Edit profile

Users profile may include user id, his profession, area of interests, etc. A user can view his search history, change the password for his account, can edit his profile.

B. Relevant Query Suggestions

System will analyze the specified query and suggest user more relevant query on the basis of his profile. A query is then normalized and system will also suggest the related queries submitted by the other users having similar profile as that of the user.

C. Google Data Extraction

When user specify the search query system will fetch the data from Google database with the help of API and the Google links will be collected in system database.

D. Web page images downloads

The links wise images of web pages will be downloaded. The web links fetched from google database will be filter out on the basis of historical search. System will automatically download the images of new links and stores on the application server.

E. Keywords storage

The web pages will be analyzed by the system and stop words will be stored in application database for future use. The keywords will be used to link images with web pages.

F. Image based search system

When user specifies query, system will display images of matching web pages. User will select any one image and system will find out the details of selected image. With the help of image details, system will filter unwanted data links and display the relevant links only. The links will be then re-ranked with the help of ratings given by the other users and Re-ranking of links will be done on the basis of user's profile, user's preferences and other user's ratings.

4. Conclusion

In this paper the integration of two methods i.e. personalization and image based search has been illustrated, that leads to improve the search result delivery. It also gives image based more relevant web pages re-ranked technique. The contents of images in HTML pages are used to improve accuracy of search system. It has used the text features as 'relevance score' and 'ranking position' that indicates the relevancy of the document for query and ranking position of document in the ranking list respectively. For the new users who don't have good hand at searching, it become difficult for them to reach to their need. There is wastage of time and labor. But, the proposed method reduces the time and labor for them. So this architecture will result into the most relevant information at the topmost position of the result list of search engine.

References

[1] Andrew W. Fiyzgibbon, Sergio Rodriguez-Vaamonde, Lorenzo Torresani, "What Can Pictures Tell Us About Web Pages? Improving Document Search Using Images,"IEEE, transactions on pattern analysis and machine intelligence, vol. 37, no. 6, June 2015.

[2] Xiaopeng Yang, Tao Mei, Yongdong Zhang, Jie Liu, and Shin'ichi Satoh, "Web Image Search Re-Ranking With Click-Based Similarity and Typicality,"IEEE transactions on image processing, vol. 25, no. 10, October 2016.

[3] Shipra Kataria and Pooja Sapra, "A Novel Approach for Rank Optimization using Search Engine Transaction Logs," IEEE, 2016.

[4] Shilpa Sethi, Ashutosh Dixit, "Design of Personalised Search System Based On User Interest and Query Structuring," 2ndInternational Conference on Computing for Sustainable Global Development, 2015.

[5] Lixin Duan, Wen Li, Ivor Wai-Hung Tsang, and Dong Xu, "Improving Web Image Search By Bag-Based Reranking," IEEE transactions on image processing, vol. 20, NO. 11, November 2011.

[6] Sanjib Kumar Sahu and Shalini Srivastava. "Review of Web Document Clustering Algorithms," IEEE, 2016.

[7] Ehsan Nowroozi, "Introduction to New Methodologies and Applications in Information Retrieval Indexing," 2nd International Conference on Mechanical and Electronics Engineering, IEEE, 2010.

[8] Caiming Xiong, David M. Johnson, and Jason J. Corso, "Active Clustering with Model-Based Uncertainty Reduction," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[9] Vimina E R, Ramakrishnan K, Navya Nandakumar and Poulose Jacob K "An Efficient Multi Query System for Content Based Image Retrieval Using Query Replacement,"IEEE, 2015.

[10] Lingxi Xie, Jingdong Wang, Bo Zhang, and Qi Tian,"Fine-Grained Image Search," IEEE Transactions On Multimedia, Vol. 17, No. 5, May 2015.

[11] Fabrizio Lamberti, Andrea Sanna, and Claudio Demartini, "A Relation-Based Page Rank Algorithm for Semantic Web Search Engines,"IEEE Transactions on Knowledge and Data Engineering, Vol. 21, NO. 1, January 2009.