Image Based Information Retrieval

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

PG Scholar, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India.[1]
Associate Professor, Department of Information Technology, Government College of Engineering Amravati, Maharashtra, India.[2]

Abstract-*In today's time retrieving the relevant information from huge collection of data has attracted a lot of attention. Various search system are available for that purpose but they should be able to find the most relevant search results according to users query that satisfies the users need. Different techniques are also there to retrieve these information. In traditional search engines usually text documents are considered and the images in these documents are ignored. Generally pictures in the HTML web pages are used to retrieve the other relevant pictures by comparing its textual as well as visual contents. Also in traditional text-based search engine relevant images can be retrieve using visual features by giving a textual query. Various systems and search engines are available for easy access and retrieval of relevant multimedia content. Most of them rely on textual data associated with the visual contents. So this paper present an approach that will produce the search results by considering the details of the images in web pages.*

Keywords: Search System, Web pages, Relevant Search Results, Images.

1. Introduction

As there is the tremendous growth of internet over the past few years, a large repository of data covering almost every area has been formed over the web and as a result of which search engine users are facing a lot of problems in retrieving the most appropriate information out of it, which is known as information overkill problem. So the search system designed should be able to retrieve the good and necessary information that will fulfills the user's needs. Due to the success of information retrieval, most commercial search engines employ text-based search techniques for image search by using associated textual information, such as file name, surrounding text, URL, etc. Even though text-based search techniques have achieved great success in document retrieval, text information is often noisy and even unavailable. In order to improve search performance, image search re-ranking, which adjusts the initial ranking orders by mining visual content or leveraging some auxiliary knowledge, is proposed, and has been the focus of attention in both academia and industry in recent years. Besides the well-known semantic gap, intent gap, which is the gap between the representation of users' query/demand and the real intent of the users, is becoming a major problem restricting the development of image retrieval in image search re-ranking. Most of the existing re-ranking methods utilize the visual information in an unsupervised and passive manner to overcome the "semantic gap" i.e. the gap between the low-level features and high-level semantics. However, when measuring image similarity and typicality, conventional re-ranking approaches only consider visual information and initial ranks of images, while overlooking the influence of click-through data. Though multiple visual modalities can be used to further mine useful visual information that can only achieve limited performance improvements. This is because these re-ranking approaches neglect the "intent gap". Users' real search intent is hard to measure and capture without users' participation and feedback. Therefore some researchers attempt to integrate users' interaction with the search process. There are a widely accepted assumption and a generally applied strategy for most image search re-ranking approaches respectively, i.e., visually similar images should be close in a ranking list, and images with higher relevance should be ranked higher than others.

In this paper we propose a novel document retrieval approach that uses the content of the pictures in the Web pages to boost the accuracy of search engines. This paper gives an approach for information retrieval based on the images. The text details of these images will be taken into account in order to compare them with the query keywords and provide the search results. Consequently our hope is that a search system will considers the textual information extracted from the pictures will yield improved accuracy of search system.

2. Related work

In this paper [1] it addressed the issue of leveraging click-through data to reduce the intent gap of image search. They propose a novel image search re-ranking approach, named spectral clustering re-ranking with click-based similarity and typicality (SCCST). In this proposed re-ranking scheme, click information is fully adopted to guide the image similarity learning and image typicality learning. Based on the learnt similarity measure, SCCST performs spectral clustering to group visually and semantically similar images into same clusters. The final re-rank list is obtained by calculating

pictures appearing in it. While this work also have focused on a reranking strategy.

This paper [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. This paper [4] present a novel sample-based online active spectral clusters typicality and within-clusters image typicality in descending order.

In this paper [2] they have explored the topic of how to use images to improve web document search. It is demonstrated that, by using modern methods and representations for image understanding, it is possible to enrich the semantic description of a Web page with the content extracted from the

clustering framework that actively selects pairwise constraint queries with the goal of minimizing the uncertainty of the clustering problem. The results validate our decomposition formulation and show that our method is consistently superior to existing state-of-the-art techniques, as well as being robust to noise and to unknown numbers of clusters.

3. Proposed Approach

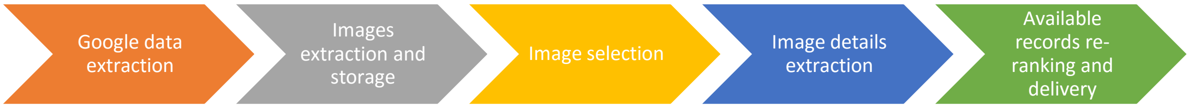
Google data extraction
Images extraction and storage
Image selection
Image details extraction
Available records re-ranking and delivery

Fig. 1. Workflow of Image Based Search System for Information Retrieval

The workflow for the proposed approach is as shown in figure. This approach is based on the reranking of relevant information considering the images in web pages. The top ranked images returned from the search engines are not always relevant because the textual terms associated with the images may not conceptually and semantically describe the content in the images. Therefore the images should be refined such that the result set contains only the images that are relevant to the user query.

The flow of this work can be explained as follows: when the user specifies query by keywords for an image search, the other related terms for a given query term is obtained using the Tag Frequency-Inverse Document Frequency (TF-IDF) algorithm and also by the Knowledge Graph API an Application Programming Interface provided by Google. It returns all the related words for the given word from the knowledge graph technology which used by Google to store the relationship between the different words. All the related terms are used to construct the hierarchical feature based on the relationship between the words. For example, if the user gives query as "apple", the two major categories representing the query terms is "apple fruit" and "apple products" these categories can be further subcategorized as "red apple, green apple, etc." and "iPad, iPhone, etc." Similarly the topics can be sub categorized to many levels as shown below.

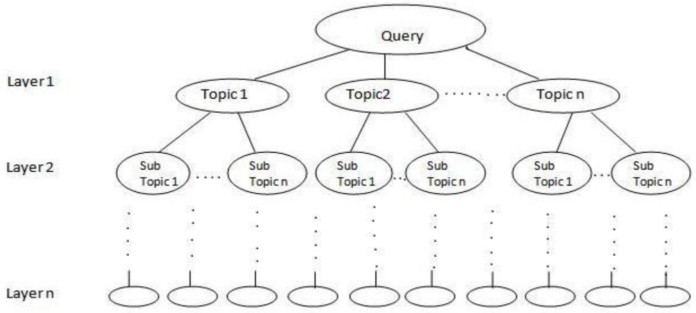

Fig. 2. Hierarchical Feature

The images are obtained for both the user query and the related terms for the user query using the Google image search API. Then the images obtained for those related terms and the initial search results are then refilled to obtain only the relevant images by calculating the relevance score of the images to the particular query. This relevance vector which contains the relevance score for all the images obtained by the relevance prediction method can be further used for topic aware re-ranking. The relevance based re-ranking boosts the relevant images to the top of the search result list while suppressing the irrelevant ones to the bottom, such that the user can access satisfying images in top positions. It tends to provide a list of images ranked with respect to the relevance score of each image to the query. Therefore initial results are reorders by their relevance score.

Modules for the above workflow diagram are given as below:

3.1 Google data Extraction

When user specifies the search query to the search system, system will fetch the data from google using googles API (Application Programming Interface) to the systems database. The links will be collected and stored in database for future use.

3.2 Image Extraction and Storage

From the collected link in above phase, images will be extracted. The extracted images are then shown to the users and also stored in database for future use.

3.3 Image Selection

User will select any image from the extracted images which best suits to his/her query. After that the details of the selected image will be extracted and the information relevant to this image and his query will be up ranked by the system.

3.4 Image details Extraction

HTML pages contains the images which may have some text in it. Text includes the words in image captions and markup tags. These textual details of the images is extracted that can be used to compare with the keywords in the query to obtain the results.

3.5 Records re-ranking and delivery

The searched records is then re-ranked according to their relevance with search query keywords. The most relevant data is ranked at the top of all the search results and the less relevant is ranked below it. For this the textual contents of the images in the web pages is taken into account that mapped with the query keywords and produces the results for particular search.

4. Conclusion

Sometimes a typical query returns thousands or millions of documents, but searchers hardly ever look beyond the first result page. This paper presents how to use images to improve Web document search. This work have also focused on a reranking strategy. This framework is sufficiently efficient to support in the near future the application of a single joint search model over text and images in the Web collection. This approach can be most suited for queries which cover wide range of topics. Ranking algorithms based on web content mining totally ignores importance of webpage. So this approach overcome this limitation as it is giving importance to the image contents of these web pages.

References

[1] 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.

[2] 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.

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

[4] Caiming Xiong, David M. Johnson, and Jason J. Corso "Active Clustering with Model-Based Uncertainty Reduction," IEEE Transactions On Pattern Analysis And Machine Intelligence, 2016.

[5] Geetha C and Geetha p, "Diversifying Image Search Results," International Conference on Computation of Power, Energy Information and Communication (ICCPEIC), IEEE, 2016.

[6] Dr. Daya Gupta and Devika Singh, "User Preference Based Page Ranking Algorithm," International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2016.

[7] Prakasha S, Shashidhar HR and Dr. G T Raju, "Structured Intelligent Search Engine for Effective Information Retrieval using Query Clustering Technique and Semantic Web," IEEE, 2014.