A Reactive Search Re-ranking model based on Images of User Interest

Nikhila T Bhuvan#1, Sudheep Elayidom#2
#1Rajagiri School of Engineering and Technology, Rajagiri Valley, Kakkanad, Kochi 682 039, Kerala, India
#2School of Engineering, CUSAT, Ernakulam 682 022, India

ABSTRACT

The volume and variety of data on web is increasing day by day, which is making the task of search engines difficult. The search engines typically use the text, inlinks and outlink to rank the web pages based on the user query. The techniques of Search Engine Optimisation like keyword stuffing, link farming etc made the least relevant pages to get a good position when ranked by search engines that puts the user in trouble. The user would have to waste their huge amount of time toggling between pages to find the relevant information. This problem can be surmounted by the use of multiple information modalities like images, videos etc along with the text to judge the relevancy of the web page for a user query. The paper explains how the images in the web pages can be used to re-rank them according to user’s interest. The user would have to provide a user click feedback for each query and the system would re-rank the web pages based on the user click. The method was tested on a set of real time sample queries and a better user satisfaction was achieved.

Keywords: Visual classifier; User click feedback; Search re-ranking

1. INTRODUCTION

This era showed an enormous intensification of digital data because of the growth of online storage and social media. Huge amount of high quality digital images and videos are available online now. In spite of all these available data, the information retrieval done by search engines still mainly depended on the text and on the inlinks and outlinks available on the web page. The multimedia content on the webpage play a very less role in ranking of the web pages even though they dominate the web. The content of the image or the video are not analyzed during the ranking procedure even though the image annotations and the video annotations are used. Not only that, the reasons for the underperformance of search engines may be SEO (Search Engine Optimization). All people are now trying to mark their presence on web. Everybody needs their website to be listed on top, making use of the SOE techniques to increase their visibility by link farming, keyword stuffing etc. The basic idea of this paper is to utilize the idle information in the form of multimedia to re-rank the web pages. All the web users require the information to be in front of them in a click of a button, the relevance of the information comes next. If the search engines are doing multimedia analysis to re rank, more accurate results could be provided. Even though this would take a little bit of time to populate the results users would not mind it as they get more precise results than before.

In order to enhance the web search, an innovative method to re-rank the web pages using images and user input overcoming the semantic gap between the text query and retrieved web pages is implemented. The idea behind the new approach is, if a webpage is judged to be relevant to a query in multiple information modalities and user feedback, then the page would have a better relevancy than others. The relevance of multiple information modalities other than
the text, inlinks and outlinks is already a very relevant area[12,1,10]. Combining the textual and pictorial information has proved to be an effective method for improved ranking accuracy[8]. Most of the search engines now implements this one click feedback mechanism to list out the images that are closer to the clicked image[11]. The rank of the web pages on the other side, that is, at the text based retrieval side is not getting affected. Online image re-ranking [11,3,4], which limits users’ effort to just one-click feedback, is an effective way to improve search results and its interaction is simple enough. This paper focuses on this issue and improves the text based search results by considering one click user feedback. Here, an improved retrieval mechanism is implemented which could provide better ranking using the multiple information modalities like text, image and user feedback. The image which is used to re-rank the web pages is selected by the user from the set of image of the image search of the query. The related images of the selected image is used re-rank the ranked web pages returned by the search engine.

2. PROBLEM DEFINITION

“To design, implement and test an effective searching technique taking into account the user click feedback reference from image search for a query”

In most of the image based search engines, user request is satisfied providing a pool of images related to the user query. Image retrieval could be done by two methods 1) Content – based[7] 2)Text –Based. In Content-based retrieval, images are given as the input for the search engines, the visual characteristics of the images are extracted for indexing and to find a match. The images are ranked based on the distance metric calculated. In text based image retrieval, query will be keywords. The images here would be stored based on the annotations provided along with the images. The images would be ranked based on the similarity score calculated using these descriptions in the form of annotation. The adoption of content based indexing and retrieval is computationally costly when it is compared with text based retrieval method. So, the text based method is preferred over the other. The best suited method for a search would be content based search, but to avoid the computational complexity we would prefer the text based one itself.

How to use these images from the search pool effectively to provide a better user experience with the search engines is what is being addressed in this paper. This pool of images would be fetched from the stored word-image index created by the crawlers using image annotations by image search engines using text based technique. If the user selects an image from the pool which would reflect user’s search objective from the collection, the images would be re-ranked using content based technique based on their visual resemblance with the user clicked image. The traditional web search engines could be exploited for finding the initial set of web pages related with a query keyword which is re-ranked using the extracted visual information. The visual information is retrieved using user intervention.

A pure content based technique could not be used for searching because it may consume huge amount of distributed resources. So, better is a combination of a text based and a content based technique[2], the text based technique is used to provide an initial set of web pages that are to be re-ranked. These web pages are re-ranked using the output of the content based technique. The advantage of this system is that the content based technique would have to search only in a small domain, that is, the already provided web images related to a query. It would be more efficient than the content based scheme. When you compare it with the traditional search engines it would be time consuming because of the additional overhead in performing the image
similarity checking and re-ranking. Ultimately, the user will be provided with a better user satisfaction level which is the primary goal of the proposed system.

3. RELATED WORKS

Some search engines use queries as keywords for searching and rely on surrounding text of the HTML image for ranking pages[9]. There can be two major concerns: 1) the surrounding texts are often noisy or too few to accurately describe the image content, and 2) the human annotations given to the images may not be perfect and are resourcefully expensive. Furthermore, the query may be ambiguous, as, it may be difficult for users to precisely describe the visual content of target web pages only using keywords.

Although lots of literature is available for image search re-ranking, very few is available for combination of image and keyword based re-ranking. The idea behind is, if web pages are judged to be relevant on the basis of multiple information modalities, then there is an increased chance for it to be really relevant. Some of the related works includes the idea of Xue et al[10] who used image snippets to improve Web search and the idea of Yu et al.[6] who improved the ranking calculating score for each pages considering the relevancy of images in each webpage. Another work is by Zhou and Dai[12] who used the text query and the image from the ranked web pages to re rank the web pages. The latest contribution is by Sergio et al.[8] where the input query is simultaneously provided to an image and well as a text search engine and the top N images from the image search engine would be taken as positive training images for the query to create a visual classifier. This visual classifier is used to re-rank the candidate web pages generated by text search engines to create a new ranking taking into account several image features.

Xue et al.[10] used image snippets to improve Web search. The web pages are segmented into many blocks. The relevancy score and importance score for each segment is calculated and aggregated to rank these blocks. The relevancy score is calculated using query and contents of blocks in each segments and the importance score is calculated using web block evaluator. The relevance of the block is identified comparing query with the content of the block. The output of the evaluator and the relevancy score can be used to calculate the rank of the blocks. The top ranked block with atleast one image is selected to extract the image snippet from each web page. The image snippet score along with the original ranked score of the page is used to re-rank the web pages. A relevancy feedback mechanism is also included to label if the retrieved pages are relevant or not.

Yu et al.[6] calculated the relevancy of images in the retrieved web pages with respect to the query using the existing text retrieval techniques like image annotations, surrounding text, ALT tag and from the image URL. Importance of each image in a Webpage is found using a machine learning algorithm. The overall score of each image is calculated combining the evidence of all the above mentioned criterias and the web pages were re-ranked based on this score.

In the work by Zhou and Dai[12], a keyword based search is done to provide the initial rank list of the links. Related images are extracted from these candidate web pages, considering the annotation tags associated with the images. 40 words surrounding the image and metatags of the web pages are extracted using a Bayesian-Network Based Text Retrieval method. An image based ranking for the links containing the related images are calculated measuring the distance between the pictures in the page and the visual prototype of the query. This is combined with the keyword based rank provided by the search engine to produce new ranking for web pages. The experiment was conducted on live web search considering 15 different queries and proved that
judging pages on multiple modalities will really improve the ranking. The disadvantage with the method adopted was, an unsupervised learning method for ranking the images was used.

An even better ranking method using supervised ranking method is proposed by Sergio et al.\textsuperscript{6} where they used the top N images from the image search for the same query keyword to train the visual classifier. The resultant links of the text based search for the keyword is re-ranked using the score calculated by the trained visual classifier. As this algorithm uses a supervised visual model, there would be a better accuracy than the previous method. The system was evaluated against the benchmark of the TREC 2009 Million Query Track (MQ09)\textsuperscript{[2]}, which involving ad-hoc retrieval over a large set of queries. The test data was little outdated, as it was formed in the year 2009. No real time analysis was done in the work

Our method does a real time analysis of some sample query keywords. A re-ranked web page list is retrieved which would reflect the user interest based on user click. The algorithm trains a cloud based visual classifier using the related images based on user click.

4. PROPOSED SYSTEM

Here, a query Q would be simultaneously searched for in a text and image search, which would retrieve the initial set of URLs and images. Let the rank list generated by the search engine S be R. The top ‘N’ URLs are considered for re-ranking and generates a rank list R’ using one click user feedback, where R’ ⊂ R.

![Fig. 1. An overview of the new system implemented](image-url)
Fig.1 gives an idea about the proposed system. Different modules of the proposed method are:

Step1. Retrieval of images from the ranked web pages
Step2. Retrieval of related images using one click user feedback
Step3. Training the classifier using the related images, images from user feedback is taken as the positive training set and a set of handpicked is considered as negative images.
Step4. Eliminating the banner images from the initially ranked pages and calculate the classification score for the remaining images in the webpage.
Step5. Re-ranking the pages based on the calculated score

4.1 Retrieval of images from the ranked web pages

The images from the first 100 URL are extracted using Jsoup[13]. Jsoup can scrap and parse text, images and links from an URL. Jsoup can deal with various components of a webpage given its URL. The basic syntax of Jsoup is:  

```java
Document doc1 = Jsoup.connect(http://hibiscus.org/*).get();
```

4.2 Retrieval of related images using user intervention

Let I be the set of images ranked by the image search engine. User can select any selects one of the image based on his interest from the list. The search engine using content based image retrieval technique retrieves the images which are similar to the selected image, let it be I’, where \( I' \subset I \). These images are even extracted using Jsoup[13]. This I’ is considered as the positive training set for the classifier. For the time being no mechanism for automatic generation of the negative training images has been devised. Negative image sets are manually generated for the 15 sample queries that are being tested using this algorithm. Each web pages listed initially in the rank list R is taken and the images are extracted one by one.

4.3 Creating and Training the classifier

The classifier used is the visual classifier service provided by IBM Watson[14]. It is a cloud service provided by IBM Bluemix. Provided an image the service returns the score form relevant classifier representing such objects. A customised classifier can even be created, after creating a classifier, providing it with the positive and negative data set, it could be trained to classify the images that we need to classify.

4.4 Eliminating the banner images

Those images whose size is less than 100x100 is considered as banner images and are eliminated from the score value calculation and those pages which are having many number of such images are considered as advertisement pages and are eliminated from the re-ranked list. The source of the images is also checked for to see if they are advertisement related pages which is also a criteria to avoid the pages from the re-ranked list.

4.4 Re-ranking the pages

There may be many images in a web page, let the images in the web page, \( W \) be \( \{I_1,I_2,I_3,...I_n\} \). If the size of images are below 100x100 the classifier does not calculate any score for it under the assumption that it is a banner image. All the other images will be provided as the input for the visual classifier, let the classification score provided by the classifier for each image be \{S(I_1),S(I_2),S(I_3)...S(I_m)\} where \( m \leq n \).

The score of the web page is \( \text{MAX}\{S(I_1),S(I_2),S(I_3)...S(I_m)\} \)

This procedure will be done for all the first N web pages. The result of the proposed system would be the re-ranked web pages.
5. EXPERIMENTAL SETUP

The heart of the proposed system is the visual classifier which is a cloud based service provided by IBM Bluemix. The IBM Watson provides a cloud based API to analyze the contents of the images as well as videos. IBM Watson is used here, as all the other approaches require us to write the image analysis and classification codes which are very time consuming. This cloud service uses semantic classifiers built with machine-learning technology. It recognizes the visual properties like color, texture, shapes, and edges and gives a score based on their evaluation about the image that is uploaded to the classifier. It is very easy to work with. Similar services are provided by Amazon web services but this one is the simplest of all.

The Java Image I/O API\[13\] from the javax.imageio package helps to work with the images stored in web pages. These are pluggable APIs that provides flexibility while working with images in the web pages, for their loading and saving. This is used to identify the source of the images in the web pages and to analyze if the page is an advertisement or not. The \texttt{<img>} tag is the one defining the image in an HTML page. This tag has two attributes: src and alt. The src defines the source of the image in the webpage. Here, it is checked if the src URL contains the keyword “googleads”. If it is so, this is considered as an advertisement and it is also not provided to train the classifier to find the score. The webpage under consideration is also checked for to make out if it contains more pictures which are advertisements, if so, these webpages are not considered during re-ranking. The width and height attributes of the images are also analyzed to see if they are banner images. The banner images are also not considered as a valid image for classification score calculation. The images below 100x100 width x height proportion is taken as banner images.

The services of IBM Watson can be availed after registering using the ID provided. Using this ID, a visual classifiers can be created for different queries. The positive and negative images are used to create the required classifier using createClassifiers(). After creating the classifier score of each page can be calculated using getScore() function. The attributes for getScore() function are search string and link URL. The function takes the link/URL that appeared in the search results for the search query to calculate the classification score using the Visual Classifier already created using 'createClassifiers()'.Finally, the function getModifiedSearchResults() reorders the links based on the user click. The attributes for ‘getModifiedSearchResults()’ are search string and the original search result.

The experiment was conducted for some 25 sample queries like hibiscus cross section, Maruthi Baleno interior, microprocessor 8086, etc. The system re-ranked the web pages based on the user click feedback providing the user with better ranked URLs avoiding the advertisement pages. The user need not toggle between the pages in search of the required information providing a better user satisfaction level.

6. RESULTS

The search results of Google search engine for the query keyword “hibiscus cross section” is shown in below in Fig.2.

The ranked links for the query by Google search engine is given in the left hand side and the corresponding web pages of the first 6 results are shown to the right hand side of the Fig. 2. You can see that none of the top 6 results gives proper information for the query.
Fig 2: Search result of a search engine for the query “hibiscus cross section”

Fig 3: Reranked web pages for the query “hibiscus cross section” using the new algorithm
In Fig 3, we can see a considerable change in the search results. Here, in the results for the proposed method we have taken into account only the image in the reranked web page. From the results it is very clear that the ranking could be made better considering the images along with the text.

5. CONCLUSION

As the proposed method showed a better user satisfaction on the sample test queries, this could be implemented on a larger collection of queries. We can evaluate our system on the benchmark of the TREC 2009 Million Query Track (MQ09) [12]. The dataset which can be used for our experiment may be “Category B” ClueWeb09 dataset[15] which includes roughly 50 million English pages crawled from the Web, but the images in the webpage are not stored in dataset. The associated images should be manually downloaded and populated. There may be one more problem that arises. The dataset was formed in early 2009; many of the web pages in the set may be now unavailable. So, we would have to restrict our experimental analysis only to the pages for which successful downloading of all images linked to the pages are available.

One of the disadvantages of the new algorithm is that its time consuming as a number of images from the web pages need to be downloaded to do the calculation. This could be improved by the implementation of this algorithm in a distributed manner. A better re-ranking would be provided if we consider the surrounding text of the image along with the image classification.

REFERENCE

[1] A Woodruff, A Faulring, R Rosenholtz, J Morrison and P. Pirolli. Using thumbnails to search the Web. In Proceeding of the SIGCHI Conference on Human Factors in Computing Systems, pages 198–205, Seatle, WA, 2001.

[2] B. Carterette, V. Pavlu, H. Fang and E. Kanoulas,. TRECMillion Query Track 2009 Overview. in TREC, 2009.

[3] J Cui and F Wen, X Tang. Intent Search: Interactive on-Line Image Search Re-Ranking. Proc. 16th ACM Int’l Conf. Multimedia, 2008.

[4] J Cui and F Wen X Tang. Real Time Google and Live Image Search Re-Ranking. Proc. 16th ACM Int’l Conf. Multimedia, 2008.

[5] Juan Manuel Barrios, Diego Diaz-Espinoza and Benjamin Bustos, Text-Based and Content-Based Image Retrieval on Flickr: DEMO”, Proceedings of the Second International Workshop on Similarity Search and Applications, 2009, SISAP ’09,Pages 156157, ISBN: 978-0- 7695-3765-8

[6] Q Yu, S Shi, Z Li, J R Wen and W Y Ma. Improve ranking by using image information. in Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings, 2007, pp. 645–652.

[7] Ritendra Datta, Jia Li, and James Z. Wang, Content-Based Image Retrieval - Approaches and Trends of the New Age, MIR’05, ACM, November 11-12, Singapore, 2005.
[8] Sergio Rodriguez-Vaamonde, Lorenzo Torresani, Andrew W Fitzgibbon. What Can Pictures Tell Us About Web Pages? Improving Document Search using Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI 10.1109/TPAMI.2014.2366761

[9] T A S Coelho, P P Calado, L V Souza, B Ribeiro-Neto and R Muntz, 2004 Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417, 2004

[10] X B Xue, Z H Zhou and Z Zhan, Improve Web search using image snippets. *In Proceeding of the 21st National Conference on Artificial Intelligence*, pages 1431–1436, Boston, WA, 2006.

[11] X Tang, K Liu, J Cui, F Wen, X Wang. Intent Search: Capturing User Intention for One-Click Internet Image Search. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1342-1353, July 2012.

[12] Z H Zhou and H B Dai, 2007. Exploiting image contents in web search. *IJCAI*, 2007, pp. 2922–2927.

[13] Use DOM Methods to Navigate a document. Available: https://jsoup.org/cookbook/extracting-data/dom-navigation [Accessed : Feb 2016]

[14] Getting started with the Visual Recognition service. Available: http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/doc/visual-recognition [Accessed: Feb 2016]

[15] Carnegie Mellon University, Language Technologies Institute. The ClueWeb09 Dataset. Website, 2009, Available: http://lemurproject.org/clueweb09.php/. [Accessed: Jan 2016]