INTERNATIONAL JOURNAL OF ADVANCED STUDIES
IN COMPUTER SCIENCE AND ENGINEERING
IJASCSE VOLUME 5, ISSUE 5, 2016
05/31/2016

Analysis and Design of an Efficient Search Technique Using Visual Classification

Nikhila T Bhuvan
Phd Scholar
Division of Computer Engg,
School of Engineering, Ernakulam, India

Dr. Sudheep Elayidom
Associate Professor
Division of Computer Engg,
School of Engineering, Ernakulum, India

Abstract **- Search engines and their ranking algorithms play a vital role in Information Retrieval from World Wide Web. Each search engine has got their own ranking algorithms with which they rank the pages and display it to the users. But the user's expectations are outpacing the experiences with Search engines. Traditional Web search engines mainly exploit the text and the in-link or out-link in the web pages to do the ranking of the web pages. They do not consider the multimedia in the HTML pages to find relevancy of the web pages against a given user query. The idea behind the new approach is, if a webpage is judged to be relevant to a query considering multiple factors other than text, there would be a better chance of a page to be really relevant. The traditional web search engines could be exploited for finding the initial set of web pages related with a query keyword which could be re-ranked using images provided by user's one click feedback. A visual classifier is trained using the set of images generated by user's click on an image. The trained classifier classifies the images in the web pages as relevant or irrelevant. The pages are re-ranked based on this classification scores providing a better user experience.**

Keywords: web page re-ranking, one click user feedback, visual classifier

I. INTRODUCTION

Images play a vital role in almost all areas of computing, like secure data transmission by means of steganography, ensuring the authenticity of the data by watermarking, disease analysis and traffic analysis by means of image processing, image modeling in the field of architecture, car modeling etc. The abundance of high quality capturing devices paved way for the fabrication of superior images. The online storage and growth of social media moved these innumerable private images from the local computing devices to be part of the World Wide Web. Why can't we leverage these data for a better information retrieval experience?

The problem in information retrieval is, searching the desired information from this very high volume and variety of data. One of the fundamental problem of text search is, the user may have to toggle between listed pages to find out the relevant information not only that, the top listed links may not have the required information. This means that there is space for further improvement of the performance of search engines. Each search engine has got their own ranking algorithms with which they rank the pages and display it to the users. But the user's expectations are outpacing the experiences with Search engines. The reasons for the underperformance of search engines may be SEO (Search Engine Optimization), the process of affecting the visibility of a web page or a website by search engine's results which is done by link farming, keyword stuffing etc.

Most of the modern search engines pay no attention to the multimedia on the web pages and retrieves documents purely based on the query keywords. The search engines normally uses the text annotations of the images for the search results [9]. In such cases there can be concerns like 1) the annotations provided may not be correct 2) The surrounding text of the HTML image may be too noisy, that is it may not contain accurate description which describe the image. Moreover, the query may also be confusing, as, it may be tricky for the users to correctly express the visual content of the target web pages only using keywords.

The problem is the semantic gap between the search query and the search results returned. The search results may or may not have the required information. The search results could be improved if any additional information can be provided along with the search query. The idea behind the approach is, if a web page is judged to be relevant based on multiple information modalities, then, there is a better

chance for it to be really relevant. The searching technology combining text along with inlinks and outlinks was studied in some works [1][3][4].Combining the textual and pictorial information has proved to be effective method for improving ranking accuracy[2]. However, different user preferences will be different, their interests and choices will be different, so a single global page ranking may not match their requirements. The current search engines do not refine the search results based on the user's interest; they do incorporate user interest for displaying the advertisements on the user's screen. So, the users are forced to distill down the information provided to them manually.

It would be better if the search results are re-ranked to the user based on some user feedback mechanism. Thus, the idea of re-ranking the web search results using the image similarity evolved. The user would provide his interest by clicking on an image from the set of image that the search engine retrieved for his text query. The URLs which are already retrieved by the search engine would be re-ranked according to the user feedback provided through user's click. The proposed algorithm combines global ranking of a search engine with content–based features of an image retrieved using one time user click feedback mechanism.

II. PROBLEM DEFINITION

"To improve the text based search results using image similarity"

Even though the users are provided with a set of search results within a fraction of second, there are less number of satisfied users. The user's expectation is not being fulfilled by most of the search engines. The focus here is to provide the user with the preferred links with the help of one click user feedback mechanism. The two major information retrieval methods are Content based and Text based. In content based method the visual features like edges, color, texture etc are used to index and find a match, where as in text based the text annotations are used. Content based search would provide us with a better search result, but its computational complexity makes it less preferred to text based one.

The semantic gap between the query and the search results could be reduced if we could re-rank the global rank list provided by the search engine based on user's feedback. The user feedback mechanism used here is the one time user click on any one of the images retrieved by the same text query. Most search engines use this one time feedback mechanism for the confusing words or the words with multiple meanings, for example apple, sun etc. When we search for such words we can see there would be multiple categories on top of the search, if we click on any of these categories the search results will be re-ranked. This could even be incorporated for searching non confusing keywords. Here, we don't have categories to click for, the user will be made to click on an image and the related images will be displayed near to the image. Most of the search engines now implements this one click feedback mechanism to list out the images that are close to the clicked image [8].This related images could be used to re-rank the web URLs of the search.

The related image set will form the positive training image for the image classifier. This image classifier would re-rank the text based search results. Online image re-ranking [6] [7] [8], 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 is focuses on improving the users experience while surfing internet for useful information using one click user feedback mechanism.

III. LITERATURE SURVEY

A very few literature is available for combining multiple modalities like image and keyword based for re-ranking is available. There are lots of papers available on re-ranking of images in the image search but to use image to re-rank web pages are least explored. The main works are 1) Xue, Zhou and Zhang[4] use of image snippets to improve Web search 2)Yu, Shi, Wen, and Ma's[5] work on improving the raking score calculation for each pages considering the relevancy of images in the webpage. 3)Zhou and Dai's[1] who combined rank list generated using text query and the image from the ranked web pages to re-rank the web pages 4)Sergio, Lorenzo, Andrew's[2], where the input query was simultaneously inputted to an image and a text search engine to extract the top N images from the image search engine to train a visual classifier.

This trained visual classifier re-ranks the candidate web pages. The candidate web pages are those ranked by the text based search engine. The web-page will be having a rank based on the search result of the text based search engine, it would be combined with the new rank created based on the image similarity score returned by the classifier to form the new rank list.

Xue, Zhou and Zhang[4] uses the image snippets extracted from Web pages and text snippets to present the results to the user. The web pages are divided into many segments. For each block the relevancy score and importance score is calculated and aggregated. The image snippet would be the representation of the entire theme of the web page. That segment that is having the most relevant information in the web page is found out calculating its relevancy and importance. The relevancy can be calculated using query and content of each segment and the importance score is calculated using web block evaluator. The web block evaluator uses the spatial features and content features of each segment. Spatial features are the position and size of the segment where as the Content features are the number and size of images, links, etc in segment. The top ranked segment with atleast one image is selected to extract the image snippet from each web page.

Yu, Shi, Wen, and Ma's[5] used machine learning techniques to find the importance of an image in the web page and calculated the relevancy for each images. The relevancy calculation was based in the resemblance of the image annotations, surrounding text and <alt> tag of the image with the user query. The ranking score for the webpage is calculated giving weightage for the importance of the images and its relevancy score.

In the work by Zhou and Dai[1], a keyword based search provides the initial rank list. Related images are extracted from these candidate web pages. This method follows an unsupervised learning method to find the related images using the image annotations and surrounding text. They do the content search to identify the related images using a Bayesian-Network Based Text Retrieval method. Using these related images a visual prototype is created and the distance between the images in each webpage and the visual prototype is used to re-rank the web pages. They conducted the experiment on 15 different live queries and were able to prove that judging pages on multiple modalities will really improve their relevance.

Sergio's, Lorenzo's and Andrew's[2] is the most latest paper in this area which utilized a supervised learning method overcoming the disadvantages of the method proposed by Zhou and Dai. A visual classifier is created and trained using the top N images from the image search. Using this classifier the scores of images in the webpage is calculated. The calculated score is used to re-rank the web pages. 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), which is a little outdated, no real time analysis was shown.

The proposed method represented in figure 1 does a real time analysis of some predefined query keywords. A re-ranked web page list is retrieved which would reflect the user interest based on user click.

IV. PROPOSED SYSTEM

A text based query keyword would be simultaneously searched for in the image and text search, which provides a list of ranked images and ranked URLs respectively. Let the rank list generated by the search engine S be R. The top 'N' URLs from the text based search engine is considered for re-ranking. The algorithm re-ranks the first N URLs based on the classification score provided by the classifier. The Classifier is trained using the related images from the one time user click. Let the re-ranked rank list be R', where $R' \subseteq R$.

Proposed algorithm is represented in figure1 and the algorithm for the the score calculation of the top N URLs is explained below. The algorithm for Score Cal() is given below. Score Cal() involves the creation of a new classifier using CreateClassifier() where the classifier is trained using related images of the user click.

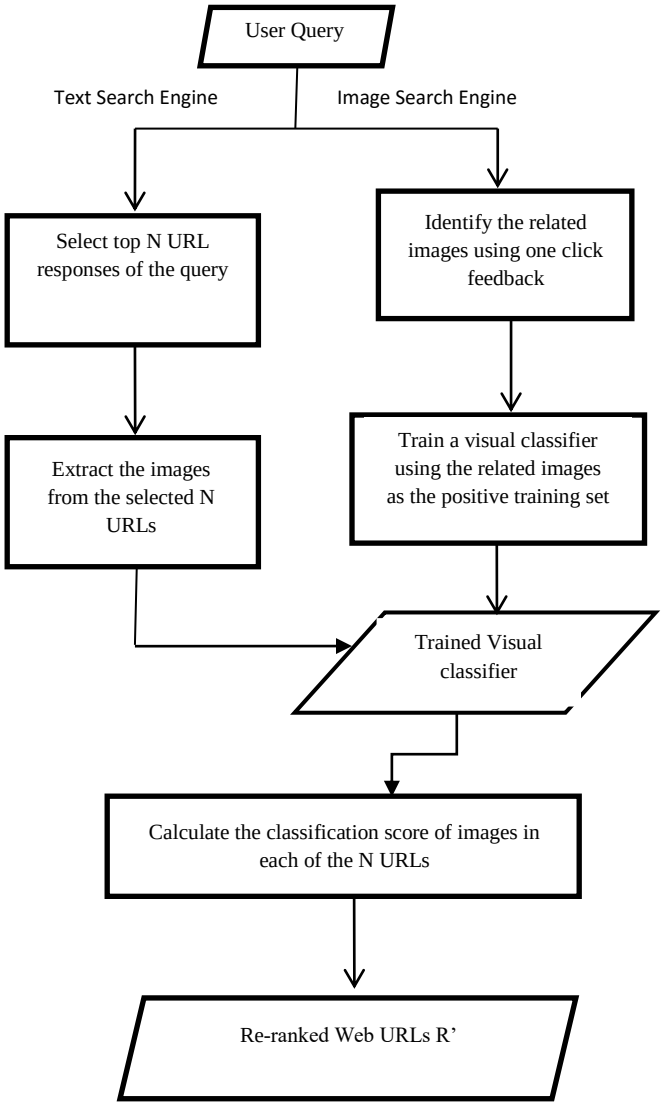

Figure1: Flowchart of the proposed methodology for re-ranking

```
Algorithm ScoreCal(R,I')
 {
   CreateClassifier(VisualRecognition Service);
   for i:=1 to N //N top ranked urls from R
   {
        Document doc=Jsoup.connect(url[i]).get();
        Elements images = doc.select("img");
        for each image in the url
        Repeat
          If (width(images)<=100 &&
                    height(images)<=100)
            Ignore the image as it is a banner image
          Else
            getScore("Hibiscus cross section",img);
   }

 }
```

Alg.1: Algorithm to calculate the rank for a webpage

```
Algorithm CreateClassifier()

{
      File positiveImages0 = new
                    File("positive.zip");
      File negativeImages0 = new
                    File("negative.zip");
      VisualClassifier Hibiscus=
                    service.createClassifier
                    ("Hibiscus",
                    positiveImages,
                    negativeImages);

   }
```

Alg.2: Algorithm to create a Classifier

The images are extracted from the top N URLs and processed by the classifier one by one. The score for the images in the URLs are calculated using the getScore() function. For the time being no mechanism for automatic generation of the negative training images has been devised. Those images whose size is less than 100x100[2] is considered as banner images and are eliminated from the score value calculation. The score of a webpage is the maximum of the scores of images in the webpage. If there are large numbers of banner images in a webpage then those pages are omitted from re-ranking under the consideration that those are advertisement pages. The source of the images can be even checked for to see if the image in the webpage is an advertisement or not.

V. EXPERIMENT SETUP

The main component of the experiment is the cloud based service of IBM Bluemix, IBM Watson, an image classifier. The positive and negative image set generated would be fed to the IBM Watson[10], a cloud based Visual Recognition service which provides an API to analyze the contents of the images as well as videos. Unless all the other approaches, the Watson service does not requires us to write the time consuming codes for image analysis and classification. This cloud service uses the visual

properties like edges, color texture, shape etc to built a semantic classifiers using machine-learning technology. The service helps us to create a custom classifier and returns an ID of the classifier. The classifier during creation can be trained providing the negative and positive images. The trained classifier presents us a prediction based on their assessment about the image that is uploaded to the classifier.

The Java Image I/O API[11] from the javax.imageio package are the pluggable APIs which helps to work with the images in the web pages. These APIs provide flexibility in handling images in the web pages, for their loading and saving. The images in the web pages could be extracted using Jsoup[14]. Jsoup is a java based HTML parser. It provides an API to parse HTML or URLs to extract and manipulate data in them. The source of an image in the webpage could be fetched for and can be checked to analyze if the images are advertisements or not. The src and alt are two attributes if an <img> tag in an HTML page. The src defines the source of the image in the webpage. Here, it is checked if the src URL contains the keyword "googleads", if so it is considered as an advertisement.

The experiment was conducted for some sample query keywords like hibiscus cross section, Maruthi Baleno interior, microprocessor 8086, etc. The proposed system re-ranked the web pages based on the user click feedback. The user need not toggle between the pages in search of the required information. The user satisfaction level is also found to be better.

VI. CONCLUSION AND FUTURE WORK

The proposed method showed a better user satisfaction on some sample test queries. The method can be evaluated on a large scale using the benchmark of TREC 2009 Million Query Track (MQ09) [12]. Using TREC 2009 Million Query Track ad-hoc retrieval over a large set of queries can be made. The computational complexity and delay in providing the result is a bottle neck for our system as it does the real time image analysis of the images in N number of pages retrieved for a text search. But the user satisfaction will outpace the complexity. This complexity could be improved by setting up a distributed map reduce implementation of our algorithm.

REFERNCES

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

[2] Sergio Rodriguez-Vaamonde, Lorenzo Torresani, and 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

[3] 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.

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

[5] 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.

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

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

[8] X Tang, K Liu, J Cui, F Wen, and 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.

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

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

[11] " The Java Image I/O API", Available: https://docs.oracle.com/javase/8/docs/technotes/guides/imageio/spec/imageio_guideTOC.fm.html

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

[13] "Carnegie Mellon University, Language Technologies Institute.The ClueWeb09 Dataset," Website, 2009, Available: http://lemurproject.org /clueweb09.php/.

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