Image Based Model for Document Search and Re-ranking

Rutuja G. Mhatre[1], Rajesh H. Bhise[2]

[1]M.E.(Comp) Pursuing, Department of Computer Engineering, PHCET, Rasayni, Maharashtra, INDIA.
[2]Assistant Professor, Department of Computer Engineering, PHCET, Rasayni, Maharashtra, INDIA.

Abstract—*Traditional Web search engines do not use the images in the web pages to search relevant documents for a given query. Instead, they are typically operated by computing a measure of agreement between the keywords provided by the user and only the text portion of each web page. This project describes whether the image content appearing in a Web page can be used to enhance the semantic description of Web page and accordingly improve the performance of a keyword-based search engine. A Web-scalable system is presented in such a way that exploits a pure text-based search engine that finds an initial set of candidate documents as per given query. Then, by using visual information extracted from the images contained in the pages, the candidate set will be re-ranked. The computational efficiency of traditional text-based search engines will be maintained by the resulting system with only a small additional storage cost that will be needed to predetermine the visual information.*

Keywords—Web Pages, search engines, multimedia search, ,document ranking.

I. INTRODUCTION

"A picture is worth a 1000 words." regardless of this old saying, recent Web search engines overlook the images in web pages and retrieve documents only by comparing the query keywords with the text in the documents[1]. This text includes the words that are related to image captions and markup tags, but ignores the pixels themselves. This lack of consideration to the visual information contrasts with the current state of the Web, which over the last 20 years has evolved from a collection of mostly textual documents to the current fast-growing large-scale repository of multimedia where nearly every page contains several pictures or videos.

Most of the Authors frequently use images in documents to present important information. An image embedded in document is commonly referred to as a figure. Basically, images are created from a screenshot, a photographic picture, a graphics, a statistical plot, etc. Image search has become more popular and shows that end-users often seek to search for images and figures in documents.

Advanced indexing techniques, information extraction, and image processing that integrate image content with text can allow both keyword-based and image-based document queries. For example, if a user asks for documents with a specific description and certain illustrations, documents that are most relevant in terms of both textual and image relevance can be selected. Currently, many of the general-purpose search engines index textual information present in multimedia documents. Users don't extract, analyze or index image content in such documents. The non-textual information present in documents is increasing, so it becomes important for search engines to utilize both texts as well as image information so that they can better assisting end-users to find relevant documents.

A. RELATED WORK

In this project a novel web document retrieval approach is proposed that uses the content of the pictures (which are embedded in the Web pages) to boost the accuracy of pure text-based search engines. At high-level users expect that, for example, for the query "Ferrari Formula 1", users will go for documents containing images of Ferrari cars that would be more relevant than pages with unrelated images. Therefore our expectation is that a web search system that combines the textual information with the visual information extracted from the pictures will yield improved accuracy. As there are large literatures on combining text and image data for image search, only few prior attempts are known to improve Web document search using image content. An example which is represented by the model of Yu et al. [2] who expressed improved ranking by using simple image measures like size, aspect ratio, and high-level features such as blurriness. In contrast, a current image recognition system is used to provide rich data on the picture content. The system of Zhou and Dai [3] is another related approach. This prior system offers the benefit of being fully unsupervised, whereas a text-based image search is created in order to obtain training data to learn a visual model of the query,. However, this unsupervised learning of the visual model that is demonstrated for a given query is computationally much more expensive and results in lower accuracy compared to our system.

B. LITERATURE SURVEY:

Many Internet scale image search methods are text-based and are limited by the fact that query keywords cannot describe image content accurately. Content-based image retrieval uses visual features to evaluate image similarity. Many visual

features were developed for image search in recent years. Some were global features such as GIST and HOG. Some quantized local features, such as SIFT, into visual words, and represented images as bags-of-visual words (BoV) . Spatial information was encoded into the BoV model in multiple ways to preserve the geometry of visual words,.

Work that relates to it falls into a number of categories: retrieval of "multimedia" documents using image and text; automatic textual annotation of images; the combination of image and text features to improve image retrieval; and the use of image features to boost relevance in text document retrieval.

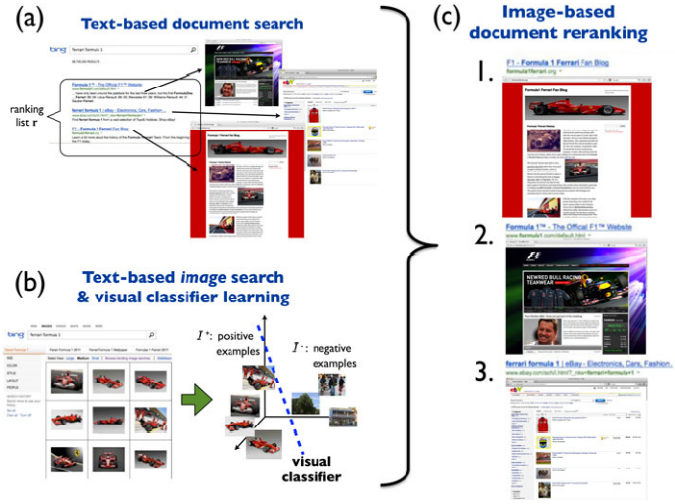
A representative work belonging to the last of these genres, is the approach of Yu et al. [2], he has collected a feature vector for each image in a document which includes metadata such as aspect ratio, height and width , as well as looking at the pixels to compute "colourfulness", "blurriness", a flag indicating presence/absence of faces, and a graphic/photo flag. Then, from training data where users rate images by "importance" within a document, they learn an "image importance" classifier, which is applied to each image in the document. They have shown that adding this feature improves judged relevance in a document search task. In contrast to their work, this project builds a query-dependent document representation which uses the image content at a semantic level. The system proposed by Yeh et al. [7] is another example of multimedia search. However, additional user input was required in their method, in the form of an image accompanying the text query. The approach that most closely relates to our own is the work of Zhou and Dai [3]. They were the first to show that content extracted from the pictures of Internet pages can be used to improve Web document search. Their system uses an unsupervised method to discover a visual representation of the query from the images of Web pages retrieved via text search. The visual model of the query is computed via an iterative technique for density estimation aimed at finding the region of the visual feature space that contains the highest concentration of image examples associated to the query. These image examples are then averaged to form a single prototypical representation of the query. Then, an image-based rank of candidate Web documents is computed by measuring the distance between the pictures in the page and the visual prototype of the query. This image-based rank is fused with a traditional keyword-based rank to form the final sorted list of documents. In our approach the costly and brittle unsupervised method of this prior system is replaced with the supervised earning of a visual classifier by exploiting as training data the photos retrieved by a text-based image search engine. This acquires much higher accuracy as compared to the system of Zhou and Dai. Furthermore, this approach has much lower runtime compared to the algorithm of this prior work. Finally, while the image-based system of [3] was tested only on 15 hand-selected visual queries, results are reported on a large-scale independent benchmark for Internet retrieval (the TREC 2009 Million Query Track (MQ09) [8]).As mentioned, there is a huge amount of work which attempts to retrieve images using textual query terms. To summarize the state of the art (in terms of methodology rather than benchmark results), the recent paper of Schroff et al. [9] serves as an adequate exemplar. This work combines text, metadata and visual features in order to achieve a completely automatic ranking of the images pertaining to a given query. Their approach begins with Web pages, recovered by text search for the query. Then images in the pages are reranked using text and metadata features, and finally a form of pseudo-relevance feedback (PRF) [1] is used: a classifier is trained to predict high rankings, and rerank the image list. This is useful addition to our system, perhaps improving the training set for our image model, but as with any PRF system, the results are an amplification of successes and failures of the base algorithm, so preference is given to test the baseline systems without PRF. Among the methods for content-based image search using text keywords, the work of Krapac et al. [10] is mentioned since, similarly to our approach, it also uses a query-relative representation. However, as already discussed above, using text features to improve image search is a qualitatively different problem as the one is introduced here. This work may be viewed as part of a more general endeavor: using images to help with problems in language. Barnard and Johnson [11] address the problem of word sense disambiguation in the context of words in image captions, and thus could hope to segment the results for ambiguous query terms. This can be useful in a PRF addendum to our class of system. Another sweep of related work is on automatic image annotation. Typically, classifiers are trained to label images with the object classes represented within. The key limitation of such methods from our point of view is that the number of classes is fixed in advance. Even the most ambitious current work looks at only thousands of classes [12]. However, in the context of search, there are millions of possible queries, and because of the "long tail" it is unsatisfactory to focus only on the most common ones. Furthermore, even if 10000 classes were pre-trained, this would add thousands of bytes to each document, while our method enables search of all possible classes with less per-document data.

II. PROPOSED APPROACH

A novel Internet document search approach is proposed here. It requires the user to give only one click on a query image and documents from a pool retrieved by text-based search are re-ranked based on their visual and textual similarities to the query image. Believe is made on that users will tolerate one-click interaction which has been used by many popular text-based search engines. For example, Google search engine requires a user to select a suggested textual query expansion by one-click to get additional results. The key problem to be

solved in this project is how to capture user intention from this one-click query image.

A. IMPLEMENTATION:



The architecture of this system is schematically illustrated in Fig. 1. Consider, D be the database of Web pages. In order to generate the list of relevant documents for an input query q, a reranking strategy is used that combines traditional text-retrieval methods with the visual classifier learned for query q:

1) The query q is made input to a text-based search engine S operating on D to generate a ranking list r of K candidate pages (Fig. 1a).

2) At the same time, the query q is issued to a text-based image search engine; the top M image results I+ are used as positive examples for training a visual classifier recognizing the query in images(Fig. 1b).

3) The list of pages r is resorted (Fig. 1c) by considering several image features including the classification scores generated by evaluating the visual classifier on the pictures of the K candidate pages. The key intuition is that when the query represents a visual concept, i.e., a concept that can be recognized in images rather than text, the learned visual classifier can be applied to increase or decrease the relevancy of a candidate page in the ranking list depending on whether the document contains images exhibiting that visual concept.

B. THE QUERY-DOCUMENT FEATURES

Next, the choice of query-document features for image-based reranking is presented. For clarity only those features are presented that are found to be beneficial in terms of improving the ranking accuracy.

The vector for query-document pair (q,i) comprises the following features.

- **Text features ()** 'relevance score' and 'ranking position' of document i in the ranking list r produced by the text-based engine S for input query q. The 'relevance score' feature is a numerical value indicating the relevancy of the document i for query q, as estimated by S, purely based on textual information. The 'ranking position' is the position of i in the ranking

list r. By including these two features, the high-accuracy of modern text-based search can be leveraged. Because our reranking function uses the ranking scores and positions generated by S, it can be viewed as an extended version of S, where visual information is exploited in addition to the traditional text features.

- **Visual metadata features ()**

'# linked images' and '# valid images'. These attributes are used to describe whether the document contains many images. This information can be useful to the image-based reranker as it reveals whether the page contains a good amount of visual information. The feature '# linked images' is simply the number of images linked in the Web page. A potential problem is that Web pages often include a large number of small images corresponding to banners, clipart, icons and graphical separators. These images typically do not convey any information about the semantic content of the page. To remove such images from consideration, the classeme descriptor only from pictures having at least 100 pixels per side is extracted. The feature '# valid images' gives the total number of images in the page for which the classeme descriptor was computed. The '# linked images' and '# valid images' jointly inform the image-based reranker on whether the document is likely to contain advertisement or rather pictures potentially useful to check the semantic agreement between the query and the content of the page.

- **Query visualness features()..**

'visual classifier accuracy' and 'visual concept frequency'. These entries are features dependent only on the query (i.e., they are constant for all documents) and describe the general ability of the visual classifier learned for query q to recognize that concept in images. In particular, 'visual classifier accuracy' gives the cross-validation accuracy of the classifier trained on the examples retrieved by Bing Images for query q. A 5-fold cross validation is used to determine the SVM hyperparameter and then store the best cross-validation accuracy over all hyperparameter values in the feature 'visual classifier accuracy'. While this feature provides us with an estimate of how reliably the classifier recognizes visual concept q in images, it does not convey how frequently this visual concept is present in pictures of Web pages. This information is captured by feature 'visual concept frequency' which is computed as the fraction of times the visual classifier for query q returns a positive score on the images of the database D. The intuition is that the joint analysis of the two query visualness features may provide the reranker with an indication of the usefulness of employing the visual classifier for query q to find relevant pages.

- **Visual content features ()..** the visual content features consist of the 'histogram of visual scores' and the

'document relevancy probability'. The 'histogram of visual

scores' is a five-bin histogram () representing the quantized distribution of the classification scores (i.e., the SVM outputs) produced by the visual classifier of

query q on the images of document i. The histogram is unnormalized and thus the sum of histogram values is equal to '# valid images'. The bin bounds are set to correspond to the following percentiles of classification scores, estimated from a large number of

queries: 30, 45, 60 and 80 percent. Thus, the histogram gives us the number of images in the document that yield classification score exceeding these thresholds. The histogram captures a measure of the semantic compatibility between the images in i and the query q. The 'document relevancy probability' () is an estimate of the posterior probability that the document i is relevant for query q given the observed classification scores of the images contained in the page, i.e., $p(i \text{ is relevant} \mid s_1, \ldots ; sn_i)$, where $s_1 \ldots s_n$ are the binarized scores that the SVM for query q produces on the ni (valid) images of document i. This probability is computed via standard application of Bayes's rule under the assumption of conditional independence (also known as the Naive Bayes assumption) [17]. In our case, conditional independence means that the classification scores are independent given the relevancy status of the document. In other words assume that

$p(_{su} \mid i \text{ is relevant}, s_u) = p(_{su} \mid i \text{ is relevant})$

and that

$p(_{su} \mid i \text{ is not relevant}, s_u) = p(_{su} \mid i \text{ is not relevant})$ for $u \neq v$.

In conclusion, the 'document relevancy probability' feature provides us directly with an estimate of the relevancy of the document purely based on the visual content of the images in the page. Note that, while it may appear that the 'document relevancy probability' and the

'histogram of visual scores' capture similar information, they actually represent the outputs of different classification models and the inclusion of both these features is found to be

beneficial to improve the reranking accuracy. Finally, if a document does not contain any valid image, features and are set to zero.

III. RERANKING MODEL

In Gradient Boosted Regression Trees (GBRT). Gradient Boosted Regression Trees (GBRT) were firstly introduced in [21] and have been proven to be among the best known models for document ranking (e.g., the best performing systems in the recent Yahoo Learning to Rank Challenge [22] use some form of GBRT). This model also uses averaging the outputs of P regression trees for prediction. However, contrasting in case of the random forest where the trees are high-variance classifiers independently learned, the GBRT trees are trained in series and are constrained to have small depth so that each individual tree has a high bias. Each tree is optimized to correct the prediction of the training documents that are responsible for the current regression error (more details on the learning procedure see [23]). Here P is chosen via brute force search on the cross-validation error. The GBRT and the random forest are trained with the code from [24].

IV. CONCLUSION AND FUTURE SCOPE

In this project, the largely unexplored topic of how to use images to improve Web document search is investigated. This is demonstrated 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 pictures appearing in it. It shows that this yields a 33 percent relative improvement in accuracy over a state-of-the-art text-based retrieval baseline. All this is achieved at the small cost of a few additional hundred bytes of storage for each page. While in this work focus is made 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.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, United Kingdom: Cambridge University Press, 2008.

[2] Q. Yu, S. Shi, Z. Li, J.-R. Wen, and W.-Y. Ma, "Improve ranking by using image information," in Proc. Adv. Inf. Retrieval, 29th Eur. Conf. IR Res., Rome, Italy, 2007, pp. 645–652.

[3] Z.-H. Zhou and H.-B. Dai, "Exploiting image contents in web search." in Proc. Int. J. Conf. Artif. Intell., 2007, pp. 2922–2927.

[4] L. Torresani, M. Szummer, and A. W. Fitzgibbon, "Efficient object category recognition using classemes," in Proc. Eur. Conf. Comput. Vis., 2010, pp. 776–789.

[5] Google images. [Online]. Available: http://images.google.com

[6] Bing images. [Online]. Available: http://bing.com/images

[7] T. Yeh, J. J. Lee, and T. Darrell, "Photo-based question answering," in Proc. 16th ACM Int. Conf. Multimedia, ser. MM '08, New York, NY, USA: ACM, 2008, pp. 389–398.

[8] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas, "Million query track 2009 overview," TREC, 2009.

[9] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 4, pp. 754–766, Apr. 2011.

[10] J. Krapac, M. Allan, J. J. Verbeek, and F. Jurie, "Improving web image search results using query-

relative classifiers," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, pp. 1094–1101.

[11] K. Barnard and M. Johnson, "Word sense disambiguation with pictures," Artif. Intell., vol. 167, no. 1, pp. 13–30, 2005.

[12] K. L. Jia Deng, Alexander C. Berg and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in Proc. Eur. Conf. Comput. Vis., vol. LNCS 6315, 2010, pp. 71–84.

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 365–372.

[14] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth, "Describing objects by their attributes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1778–1785.

[15] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 951–958.

[16] Lscom: Cyc ontology dated (2006-06-30). [Online]. Available: http://lastlaugh.inf.cs.cmu.edu/lscom/ontology/LSCOM-20060630.txt, http://lscom.org/ontology/index.html

[17] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, August 2006.

[18] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in Proc. Adv. Large Margin Classifiers, MIT Press, 2000, Ch. 7, pp. 115–132.

[19] T. Joachims, "Training linear SVMs in linear time," in Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining,ser. KDD '06, New York, NY, USA: ACM, 2006, pp. 217–226.

[20] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5– 32, 2001.

[21] [21] J. H. Friedman, "Greedy Function Approximation: A gradient boosting machine," Ann. Statist., vol. 29, pp. 1189–1232, 2001.

[22] Yahoo learning to rank challenge. [Online]. Available: Website, http://learningtorankchallenge.yahoo.com/

[23] Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, and G. Sun, "A general boosting method and its application to learning ranking functions for web search," in Proc. Neural Inf. Process. Syst., 2007, pp. 1697–1704.

[24] A. Mohan, Z. Chen, and K. Q. Weinberger, "Web-search ranking with initialized gradient boosted regression trees," J. Mach. Learn. Res.—Proc. Track, vol. 14, pp. 77–89, 2011.

[25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," J. Mach. Learn. Res., vol. 9, pp. 1871–1874, 2008

[26] (2009). Carnegie Mellon University, Language Technologies Institute. The ClueWeb09 Dataset. [Online]. Available: http://lemurproject.org/clueweb09.php/

[27] TREC.TREC 2009 Million Query Track—Prels relevance judgements. [Online]. Available: http://trec.nist.gov/data/million. query/09/prels.20001-60000.gz

[28] W. Zheng and H. Fang, "Axiomatic approaches to information retrieval- university of delaware at TREC 2009 million query and web tracks," in Proc. Text Retrieval Conf., 2009.

[29] TREC.UDMQAxQEWeb ranking lists at TREC 2009 million query track. [Online]. Available: http://trec.nist.gov/results.html

[30] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language-model based search engine for complex queries," in Proc. Int. Conf. Intell. Anal., vol. 2, no. 6, 2005, pp. 2–7.

[31] Amazon mechanical turk. [Online]. Available: https://mturk.com [32] O. Alonso and S. Mizzaro, "Using crowdsourcing for TREC relevance assessment," Inf. Process. Manag., vol. 48, no. 6, pp. 1053–1066, Nov. 2012.