Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Improving Web Image Search Re-Ranking Using Hybrid Approach

Kirti Yadav, Sudhir Singh
Department of Computer Science
Kanpur Institute of Technology,
Kanpur,India.

Abstract-The explosive growth and widespread accessibility of community contributed media content on the Internet have led to a surge of research activity in image search. Approaches that apply text search techniques for image search have achieved limited success as they entirely ignore visual content as a ranking signal.
We propose an adaptive visual similarity to re-rank the text based search results. A query image is first categorized as one of several predefined intention categories, and a specific similarity measure has been used inside each of category to combine the image features for re-ranking based on the query image.

***Keywords**: Content Based Image Retrieval, Image Ranking, Image Searching, Text based image retrieval, Hybrid approach Re-ranking, Visual Re-ranking, Image Ranking and Retrieval Techniques.*

I. INTRODUCTION

The ever-growing number of digital images on the *Internet* (such as in the online photo sharing Website, the online photo forum and so on), retrieving relevant images from a large collection of database images has become an important research topic. Over the past decades, many image retrieval systems have been developed, such as text-based image retrieval (TBIR), content-based image retrieval (CBIR) and hybrid approach [7].

1.1 Text-Based Approaches: The TBIR has been widely used in popular image search engines (e.g. Google, Bing and Yahoo!). Specifically, a user is required to input a keyword as a textual query to the retrieval system. Then the system returns the ranked relevant images whose surrounding texts contain the query keyword, and the ranking score is obtained according to some similarity measurements (such as cosine distance) between the query keyword and the textual features of relevant images.

Text-based search techniques have been verified to perform well in textual documents; they often result in mismatch when applied to the image search. The reason is that metadata cannot represent the semantic content of images. For example, a search by the keyword “Paris” nets a large number of images of a model Paris Hilton and the city Paris in the meantime.

1.2 Content-Based Approaches: Most search engine works on Text Based Approaches but there exist alternative approach, content based image retrieval that require a user to submit a query image, and return images that are similar in content .Google is one of the search engine that works on Content based Image re-ranking.

The extracted visual information is natural and objective, but completely ignores the role of human knowledge in the interpretation process. As the result, a red flower may be regarded as the same as a rising sun, and a fish the same as an airplane etc.

1.3 Hybrid approaches: Recent research combines both the visual content of images and the textual information obtained from the Web for the WWW image retrieval. Such methods exploit the usage of the visual information for refining the initial text-based search result. Especially, through user’s relevance feedback, i.e., the submission of desired images or visual content-based queries, the re-ranking for image search results can achieve significant performance improvement [8].

II. RELATED WORK

The methods for image search re-ranking can be classified into supervised and unsupervised ones, according to whether human labeled data has been used to derive the re-ranking model or not. The unsupervised re-ranking methods do not rely on human labeling of relevant images but require prior assumptions on how to employ the information contained in the underlying text- based result for re-ranking [9].

III. PROPOSED WORK

After the user log in, first user log displays the information about the previous user recently searched images. From that a particular query selected by the user or a new query given by the user which retrieves images from the database .Or user can directly search for the image query in database. In the existing system, classification of images can be displayed by means of semantic signature.

In our approach visual and textual features can be compared with the user selected image by means of texture, shape and color.

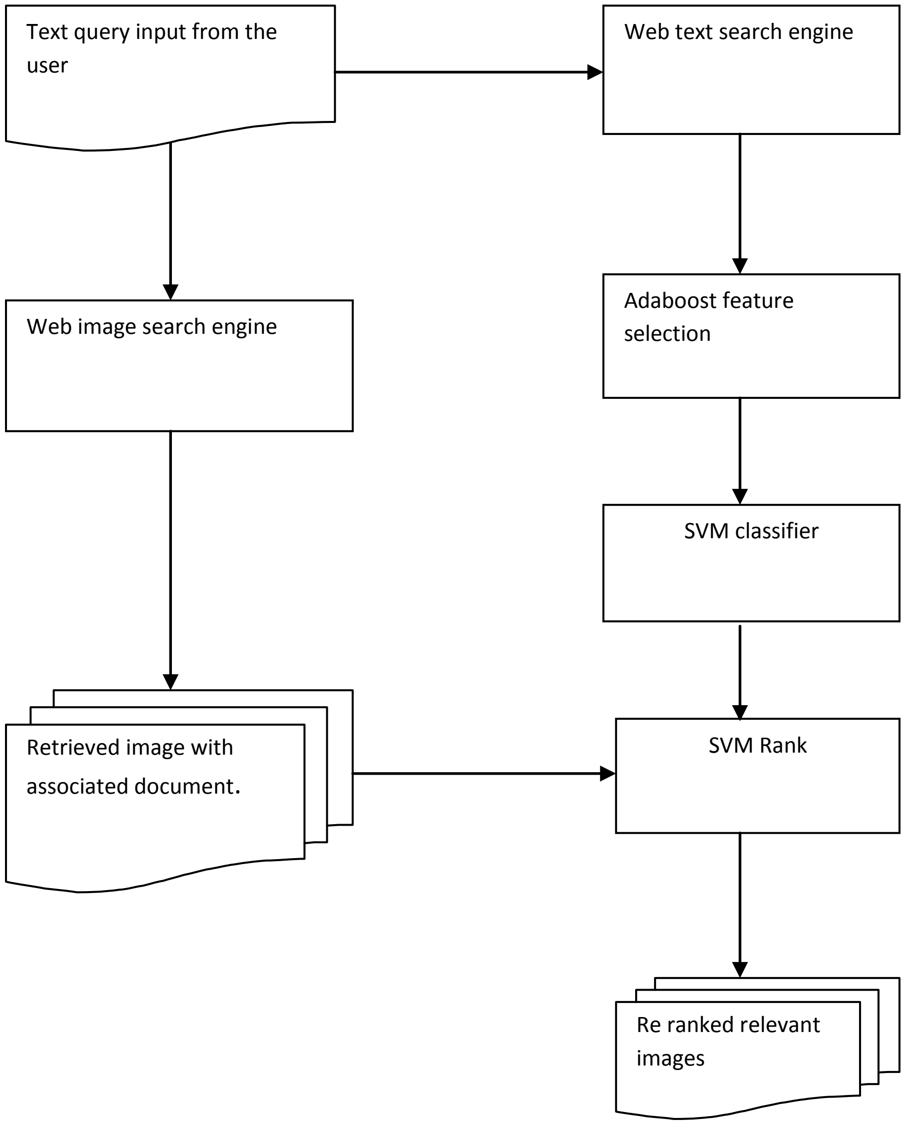

Fig 1 System architecture

Generally it performs the following five operations.

1. Stores the image URL from Google
2. Extracts the image sources and saves it
3. Adaboost selects the important features
4. SVM is used for Classification and Ranking
5. Performance comparison

3.1 ADABOOST

AdaBoost is one of the most promising, fast convergences, and easy to be implemented machine learning algorithm. It requires no prior knowledge about the weak learner and can be easily combined with other method to find weak hypothesis, such as support vector machine.

Feature selection is an optimization process to reduce a large set of original rough features to a relatively smaller feature subset which containing only significant to improve the classification accuracy fast and effectively.

Algorithm :AdaBoost for Feature Selection

Given example images: $(x_1, y_1), \cdots, (x_n, y_n)$ where $x_i \in X, y_i \in Y = \{0,1\}$ for negative and positive examples respectively

Initialize weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0,1$ respectively, where m and l are the number of negatives and positives examples respectively.

For t = 1,⋯, T:

Normalize the weights,

$$w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}.$$

For each feature, j, train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to w_t, $\varepsilon_t = \sum_i w_i \left| h_j(x_i) - y_i \right|$

Choose the classifier h_t, with the lowest error ε_t

Update the weights:

$$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$$

where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and

$$\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}.$$

The final strong classifier is:

$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases}$$

where $\alpha_t = \log\frac{1}{\beta_t}$.

3.2 SUPPORT VECTOR MACHINE

Support vector machine was developed by Vapnik from the theory of Structural Risk Minimization. However, the classification performance of the practically implemented is often far from the theoretically expected. In order to improve the classification performance of the real SVM, some researchers attempt to employ ensemble methods, such as conventional Bagging and AdaBoost. SVM is essentially a stable and strong classifier [9]

Considering the problem of classifying a set of training vectors belonging to two separate classes,

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \subseteq X \times Y$$

where

$$x_i \in X \subset R^n, y_i \in Y = \{-1, +1\}, i = 1, 2, \cdots, l.$$

SVM can be trained by solving the following optimization problem:

$$\min_{w} \Phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i, i = 1, 2, \cdots l$$

where $\xi_i > 0$ is the i-th slack variable and C is the regularization parameter.

The above optimization problem can be solved in its dual form:

$$\alpha^* = \arg\max_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K\langle x_i, x_j\rangle - \sum_{k=1}^{l} \alpha_k$$

where $K\langle x_i, x_j\rangle$ is the kernel function performing the nonlinear mapping into feature space.

most frequently used kernel are Radius Basis Functions (RBF):

$$k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

After Re-ranking Before Re-ranking

FIG-2

Ranking SVM is an application of Support vector machine, which is used to solve certain ranking problems. The algorithm of ranking SVM was published by Torsten Joachims in 2003. The original purpose of Ranking SVM is to improve the performance of the internet search engine [1].

Ranking SVM, one of the pair-wise ranking methods, which is used to adaptively sort the web-pages by their relationships (how relevant) to a specific query. A mapping function is required to define such relationship. The mapping function projects each data pair (inquire and clicked web-page) onto a feature space. These features combined with user's click-through data (which implies page ranks for a specific query) can be considered as the training data for machine learning algorithms.

Generally, Ranking SVM includes three steps in the training period:

a. It maps the similarities between queries and the clicked pages onto certain feature space.

b. It calculates the distances between any two of the vectors obtained in step (a).

c. It forms optimization problem which is similar to SVM classification and solve such problem with the regular SVM solver.

IV. RESULT and ANALSIS

The comparison of performance before and after re-ranking is shown in (Fig 2).The average precision at the top 50 documents, i.e. in the first two to three result pages of Google. Here it remove the irrelevant image and place the relevant image..

We tested the idea of re-ranking on three text queries to a large-scale web image search engine, Google Image Search, which has been on-line since July 2001. As of March 2003, there are 425 million images indexed by Google Image Search. With the huge amount of indexed images, there should be large varieties of images, and testing on the search engine of this scale will be more realistic than on an in-house, small-scale web image search system [4].

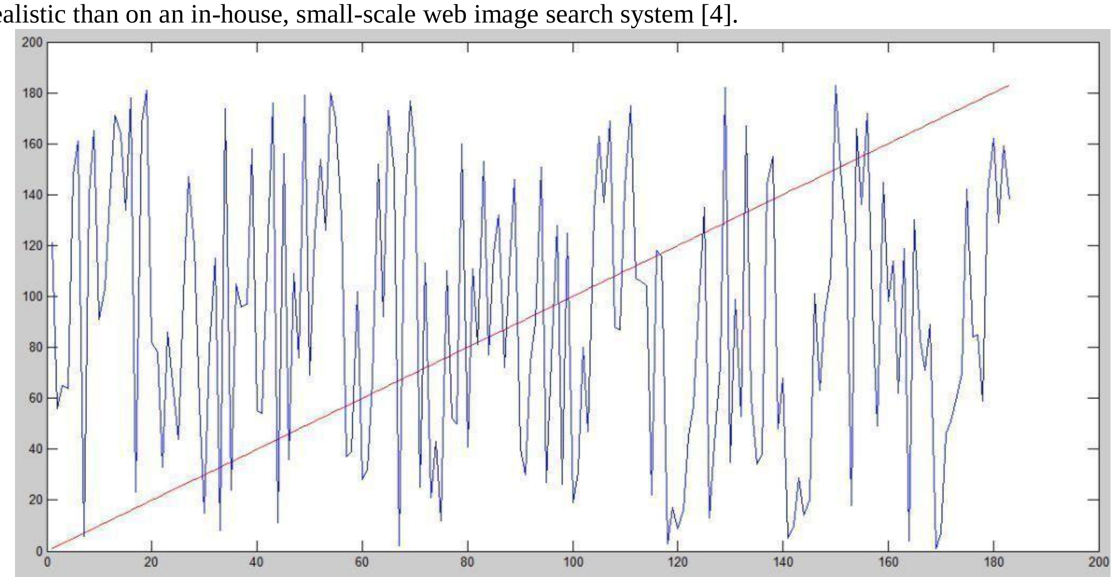

FIG: 3-Graph between the rank and index

Google Web Search, based on Page Rank algorithm, is a large-scale and heavily-used web text search engine. As of March 2003, there are more than three billions of web pages indexed by Google Web Search. There are 150 millions queries to Google Web. Here the fig-3 shows the comparison between the Google image search based on page rank method and hybrid approach based on relevance model. The straight line indicates google ranking and zigzag lines indicate the ranking of images after applying hybrid approach.
Search every day with the huge amounts of indexed web pages, we expect top-ranked documents will be more representative, and relevance model estimation will be more accurate and reliable for each query, we send the same keywords to Google Web Search and obtain a list of relevant documents via Google Web APIs. Before calculating the statistics from these top-ranked HTML documents, we remove all HTML tags, filter out words appearing in the query stop word list, count the word how many times it occurs in the page, which are all common pre processing in the Information Retrieval systems and usually improve retrieval performance. From the (Table 1) show the result of three search queries and the number of relevant images in top 200.

Table 1.Three search queries

Query No.	**Text Query**	**Number of Relevant Images In Top 200**
1	Lion	100
2	birds	60
3	fish	80

V. CONCLUSION and FUTURE WORK

Re-ranking web image retrieval can improve the performance of web image retrieval, which is supported by the experiment results. The re-ranking process based on relevance model utilizes global information from the image's HTML document to evaluate the relevance of the image. The relevance model can be learned automatically from a web text search engine without preparing any training data. The reasonable next step is to evaluate the idea of re-ranking on more and different types queries. At the same time, it will be infeasible to manually label thousands of images retrieved from a web image search engine. An alternative is task-oriented evaluation, like image similarity search. Given a query, we re-rank images returned from a web image search engine and use top-rank images to find similar images in the database. We then can evaluate the performance of the re-ranking process on similarity search task as a proxy to true objective function.
Although we apply the idea of re-ranking on web image retrieval in this paper, there are no constraints that re-ranking process cannot be applied to other web media search. Re-ranking process will be applicable if the media files are associated with web pages, such as video, music files, MIDI files, speech wave files, etc. Re-ranking process may provide additional information to judge the relevance of the media file.

References

[1] www.wikipedia.com.

[2] V. Lavrenko and W. B. Croft .Relevance-based language models .In Proceedings of the International ACM SIGIR Conference 2001.

[3] R. Lempel and A. Soffer. Pic A SHOW: Pictorial authority search by hyperlinks on the web. In Proceedings of the 10th International World-Wide Web Conference, 2001

[4] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," Multimedia Syst., vol. 8, no. 6(2003)536–544.

[5] T.-Y. Liu, "Learning to rank for information retrieval," Found. Trends Inf. Retr., vol. 3 (Mar. 2009) 225–331.

[6] X. Tian D.Tao, X.-S. Hua, and X. Wu, "Active re-ranking for web image search," IEEE Trans. Image Process., vol. 19, no. 3, (Mar.2010) 805–820.

[7] Chaitali Jambhulkar, Ankita Borana, Chaitali Deshmukh, Priyanka Nanaware & Chandrama G. Thorat "Personalized Image Search using Hybrid Re-ranking Based on Ontology" 2319 – 2526, Volume-1, Issue-2, 2012

[8] A Novel Approach for Image Searching using Visual Re-ranking Technique (Dec. 20115] L.Ramadevi1, Ch. Jayachandra2, D.Srivalli3"Improving Web Image Search Re-ranking" *2278-8727Volume 13, Issue 1 (Jul. - Aug. 2013),*

[9] Sabitha M G1, R Hariharan2 *"Hybrid* Approach for Image Search re-ranking" Volume 2 Issue 5, May 2013**.**

