Journal of Computer Science 4 (10): 842-846, 2008
ISSN 1549-3636
© 2008 Science Publications

Exploiting Surrounding Text for Retrieving Web Images

[1]S.A. Noah, [2]A. Azilawati, [1]T.M. Tengku Sembok and [1]T.W. Tengku Siti Meriam
[1]Department of Information Science, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43650 Selangor, Malaysia
[2]Faculty of Information Technology and Quantitative Science,
University Teknologi MARA, Perak, 32600 Sri Iskandar, Perak, Malaysia

Abstract: Web documents contain useful textual information that can be exploited for describing images. Research had been focused on representing images by means of its content (low level) description such as color, shape and texture, little research had been directed to exploiting such textual information. The aim of this research was to systematically exploit the textual content of HTML documents for automatically indexing and ranking of images embedded in web documents. A heuristic approach for locating and assigning weight surrounding web images and a modified tf.idf weighting scheme was proposed. Precision-recall measures of evaluation had been conducted for ten queries and promising results had been achieved. The proposed approach showed slightly better precision measure as compared to a popular search engine with an average of 0.63 and 0.55 relative precision measures respectively.

Key word: Information retrieval, image retrieval, precision recall

INTRODUCTION

The amount of multimedia data, particularly images have been increasing. On August 2005, Google alone had indexed almost 2.2 billion images on the web and that was an increased from 1.3 billion images reported a year before that. This vast amount is only considering those images available on the web, not including personal collections, pictures archives and photo CD-ROMs. As a result, there has been a great demand for solutions in the form of image retrieval systems or search engines that can organize and effectively search these images.

Solutions for better image retrieval have been conducted extensively within two main research interests[1]; the computer vision research and the data management research. The computer vision researcher believes that the best way of retrieving images is through a content-based approach (via visible information such as color, texture and shape); whereas the data management researcher insists on image annotations which can further supported conventional text-based approach. Although, content-based approach seems to be a natural way for describing and retrieving images, a successful example of such implementation is still yet to be seen. Furthermore, the effort to describe the information needs by means of sketches and other forms of input are still difficult and impractical. Users are more comfortable transforming their information needs by submitting keywords or phrases[6]. Therefore, the conventional textual approach of retrieval is still a popular choice among human searchers. Coincidently vast amount of images over the web are provided with short description and explanation somewhere in the documents. The main task is to find which part of the text can be used to represent the content of the images[7].

In this study, we proposed a method for representing and indexing web images by exploiting the HTML's document structure, the surrounding text and the description provided to images. We proposed a modified tf.idf weighting scheme by considering local weight of the indexed images.

Problem statement: Most of the early image retrieval systems such as[2,3] attempt to adapt human perceptual capabilities by using various features extraction models and spatial information to represent images. Techniques from pattern recognition fields were also favored by many early approaches in image retrieval systems. The main notion was to analyze the image and extract perceptually salient visual features such as color, shape and texture, which will be numerically represented for indexing purpose. These image representations will be

Corresponding Author: Shahrul Azman Noah, Department of Information Science, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43650 Selangor, Malaysia Tel: +60389216178 Fax: +60389256732

mapped to point in a multidimensional feature space where Spatial Access Methods (SAMs) will be used to locate other points close to it. The general idea was that closer points meant higher similarity value from the human perception. The distance between points is known as similarity distance. These systems, however, did not pay great attention on how human perceptions about images can be formulated in the form of query. As a result a number of sketch based approach in query formulation have been proposed[4,5].

As can be seen research on content-based image retrieval are still very actively conducted and seems to show promising result. However, a keyword-based approach is still a popular way for expressing information needs among web searchers. The common approach of text indexing for web images is based from conventional approach i.e., by considering all the text segments available in the document. Little effort however, has been focused to systematically locate, extract and index suitable text segments that potentially describe the images.

MATERIALS AND METHODS

Exploiting HTML document structure and surrounding text: The proposed approach exploits the textual content of web documents that can be associated with images i.e. text or terms located in HTML tags containing the image and terms that are located nearest to the image. The former feature allows us to assign specific weight to HTML's tag such as the <title> tag that gives the document's title and the <bold> tag that can be assumed as some stress given to the bold terms or phrases. The <alt> tag is given the highest weight as it is frequently used to provide explicit information or what we might wrongly called as "tool tip text". In this sense the values of the terms weight are assigned based on the heuristically importance of the HTML tags surrounding the images. The weights scheme used in this research is as shown in Table 1 which was initially proposed by[8].

Table 1: Word weights based on HTML tags

HTML tags	Weight
<alt> tag on image	6.00
<title>	5.00
<H1>	4.00
<H2>	3.60
<H3>	3.35
<H4>	2.40
<H5>	2.30
<H6>	2.20
<B>	3.00
<Em>	2.70
<I>	2.70
<strong>	2.50
(without tags)	1.00

The latter features on the other hand assumed a simple heuristic whereby texts that are located near to the indexed image are claimed to be describing or explaining the particular image[9]. Such a heuristic also assumed that these texts provide more meaningful information to the indexed image. Figure 1, for instance illustrates an example of a web document containing an image and the textual content associated to it. In this study, we consider the ten terms located before and after the images as index.

Each term is assigned a special weight based upon their proximity to the image by using the following equation:

$$w_t = \rho.e^{-2.0\times {pos_t}/{N}} \tag{1}$$

Where:

w_t = The weight of term t
pos_t = The position of term t from the image
N = The maximum number of term near the image
ρ = The damping factor

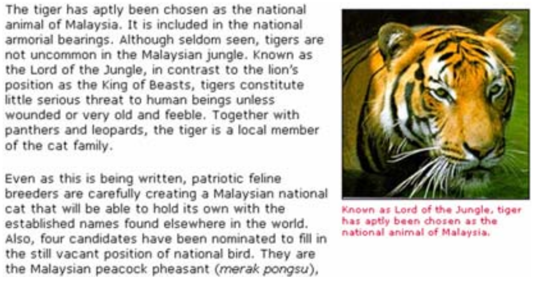
In this study we consider ρ = 5, such that term nearest to the image will have a value which is slightly less than the <alt> tag but almost equals to the tag <title>.

Based upon the specific weight of each term considered, we can construct the local weight of each image based upon the nearest surrounding text as follows:

$$L(c) = \log\left(\frac{\sum_{s\in Q} v_s + \sum_{t\in R} w_t}{|Q| + |R|}\right) \tag{2}$$

Where:

Q = The set of terms located in HTML tags
R = The set of terms located before and after image c
|Q| and |R| = The number of terms located in HTML tags and terms located before and after image w respectively



The tiger has aptly been chosen as the national animal of Malaysia. It is included in the national armorial bearings. Although seldom seen, tigers are not uncommon in the Malaysian jungle. Known as the Lord of the Jungle, in contrast to the lion's position as the King of Beasts, tigers constitute little serious threat to human beings unless wounded or very old and feeble. Together with panthers and leopards, the tiger is a local member of the cat family.

Even as this is being written, patriotic feline breeders are carefully creating a Malaysian national cat that will be able to hold its own with the established names found elsewhere in the world. Also, four candidates have been nominated to fill in the still vacant position of national bird. They are the Malaysian peacock pheasant (*merak pongsu*),

Known as Lord of the Jungle, tiger has aptly been chosen as the national animal of Malaysia.

Fig. 1: An example of image with description

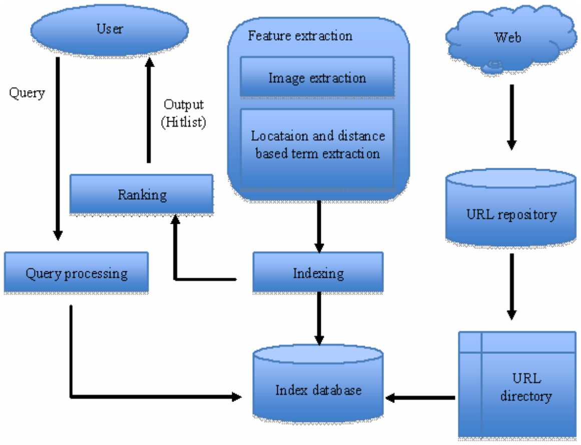


Fig. 2: Architecture of the prototype

The common fid equation as follows:

$$w_{ik} = tf_{ik} \times \log\frac{N}{n_k} \tag{3}$$

Where:

w_{ik} = The weight of term k for image i
tf_{ik} = The term frequency of term k related to image i
N = The total number of terms in the database
n_k = The total number of images containing the term k is modified by considering the local weight L(c) as follows:

$$w_{ik} = tf_{ik} \times \log\frac{N}{n_k} + L(C) \tag{4}$$

The prototype search engine: In order to evaluate the effectiveness of the aforementioned approach, a prototype indexing and retrieval engine has been developed to automatically index and processed the HTML documents according the aforementioned heuristic. The architecture of the prototype is shown in Fig. 2. We therefore briefly described the processes involved.

As shown in Fig. 2, the first module is the feature extraction involving image extraction and term extraction. All the features available in the HTML documents and which are related to the image will be extracted and submitted to the indexing engine. All the documents and images were collected from the crawling process of our own spider and stored in the repository. At the same time, the URL of the images will be recorded and stored in the index database. Each HTML document will be preprocessed for finding the image and their relative location in the HTML document. Only the images with the size of 99 pixels will be considered. Images that do not fulfill this requirement are considered to be non-content bearing and should not be indexed. The name of the images will then be extracted as well as terms existed in the <alt> tag.

The term extraction stage involved extracting all the related terms located in the HTML tags as described previously. For example the sentence or terms under the <title> tag is used to extract the title of the documents or even the name of indexed images. Terms that are extracted from these tags are given special weight as previously shown in Table 1. Term extraction will also involved extracting terms located near the image and special weight is assigned based on the Eq. 4. Stopword that does not have any content bearing meaning is not included in the index. We used the conventional Vector Space Model (VSM) in constructing the inverted index and the ranking is done by means of cosine similarity measures as follows:

$$\mathrm{Cos}(D_i, Q) = \frac{\sum_{k=1}^{t}(d_{ik} \bullet q_k)}{\sqrt{\sum_{k=1}^{t} d_{ik}^{2} \bullet \sum_{k=1}^{t} q_k^{2}}} \tag{5}$$

Where:

d_{ik} = Represents the weight of term k in image i
q_k = The weight of term k in query q

In our case the weight of term k in query q is equal to 1.

RESULTS

In order to evaluate the effectiveness of the proposed approach, we have conducted a series of evaluative experiments. The first stage of the experiment is the population of the database with documents or images. Prior to this, we have constructed 10 queries for evaluation purposes. Apart from testing; these queries were used for database population whereby they were submitted to a search engine (other than Google) and the first 100 hitlists (containing images) were used to populate the database. For each query, relevant images were manually labeled, therefore producing a gold standard of evaluation benchmark[10].

We employed the standard precision recall measures as have been used many similar evaluative experiments such as those of[11]. Precision and recall are respectively formulated as:

$$\text{Precision} = \frac{\text{No. of relevant documents retrieved}}{\text{No. of documents retrieved}} \tag{6}$$

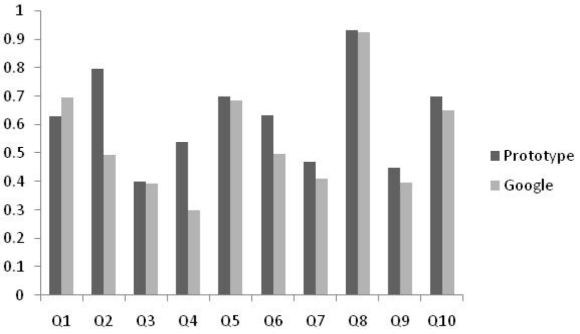


Fig. 3: Precision for the ten submitted queries

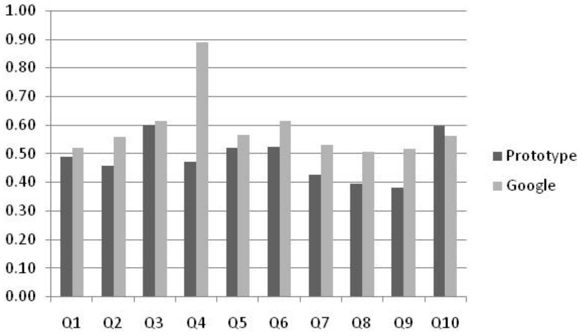


Fig. 4: Relative recall for the ten submitted queries

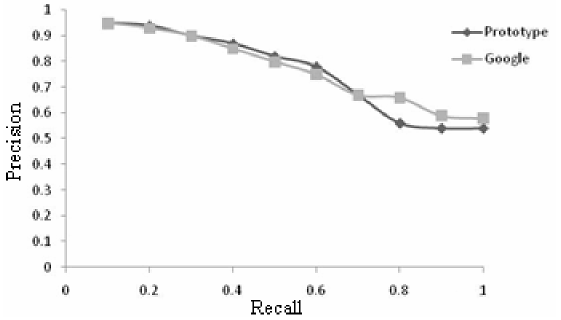


Fig. 5: The average precision graph comparing between the developed prototype and the Google search engine

$$\text{Recall} = \frac{\text{No. of relevant documents retrieved}}{\text{No. of relevantdocuments in the database}} \quad (7)$$

The results presented below are based on the average of the ten queries submitted to the prototype search engine and Google. Figure 3 shows that our proposed approach provide slightly higher precision measure in the majority of the queries. On average the proposed approach achieved 0.63 precision measures as compared to 0.55 for the Google search engine. Figure 4 however shows that the recall measures of our proposed approach are lower than of Google search engine. The average recall of our proposed approach is 0.49 as compared to 0.59 of the Google search engine. In this experiment we used the cutoff point of 20 documents.

Figure 5 shows the interpolated average precision-recall measures graph of the two results based from the ten submitted queries. As can be seen, the proposed approach in this study shown to have a slightly better result for the recall value <0.6. The performance however deteriorate for the recall value >0.6.

DISCUSSION

The results shown in the previous section indicated that the exploitation of surrounding text do have the potential to improve the precision of text-based image retrieval. As mentioned before, our approach shown to have a slightly better precision measure for the recall value <0.6. The deterioration of precision measure for the recall value>0.6 is due to the limited number of stored images as compared to the commercial search engine used during the testing. The result only provides comparison with the Google search engine. As Google has been the most widely used search engine among internet users, the result indicate promising solutions for further research and development. Populating the original index with other similar and related terms such as those extracted from the WordNet is one of the areas for further research in text based image retrieval. The result has also affirmed that text-based retrieval for images is sufficient enough for searching. However it cannot be denied that incorporating visual features will provide value added feature for indexed images.

CONCLUSION

This study provides a simple heuristic for exploiting web document structure and text regions surrounding images for supporting image retrieval. Our approach exploits the text located in the HTML tags and assigned specific weight based on the heuristic of importance. The approach also consider the text located near the image by assigning the weight based on their distance to the image. For these extracted text, a local weight is defined as the summation of all the term weights. A modified tf.idf weighting scheme is then proposed by incorporating such local weight. Thus, the overall weight of a term is defined as the sum of local term weight plus its tf.idf weight. Our experiments indicate slightly better result in terms of precision measures but at the same time exhibits limitation in the recall measures. Such limitation might due to the quality of our corpus and to the semantic relatedness of term which has not been considered. Ignoring the semantic aspect of terms causes documents containing the terms ‘kittens’, ‘pets’ and ‘felis catus’ to be ignored for the query ‘cat’.

Our future research is therefore directed towards incorporating semantic meaning to terms by using different levels of Natural Language Processing (NLP) techniques and forms of lexical ontology[12]. It concerns with semantic extraction, named entity recognition and the use of WordNet and ConceptNet lexical ontology for creating semantic index. We therefore anticipate the capacity of supporting semantic search for web images from the surrounding text segments.

ACKNOWLEDGEMENT

The researchers wish to thank the Ministry of Science, Technology and Innovation (MOSTI) for the fund provided for this project and also the anonymous referees for their helpful and constructive comments of this research.

REFERENCES

1. Lew, M.S., N. Sebe, C. Djeraba and R. Jain, 2006. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput., 2: 1-19. http://doi.acm.org/10.1145/1126004.1126005
2. Pentland, A., R.W. Picard and S. Sclaroff, 1996. Photobook: Content-based manipulation of image database. Int. J. Comput. Vision, 18: 233-254. DOI: 10.1007/BF00123143
3. Carson, C., S. Belongie, H. Greenspan and J. Malik, 2002. Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Patt. Anal. Mach. Intel., 24: 1026-1038. ieeexplore.ieee.org/iel5/34/22016/01023800.pdf
4. Bradshaw, B., 2000. Semantic based image retrieval: A probabilistic approach. Proceedings of the 8th ACM International Conference on Multimedia, 2000, Marina del Rey, California, United States, pp: 167-176. http://doi.acm.org/10.1145/354384.354456
5. Jaimes, A. and S.F. Chang, 2002. Concepts and Techniques for Indexing Visual Semantics. In: Image Databases Search and Retrieval of Digital Imagery, Castelli, V. and L.D. Berman (Eds.). John Wiley, New York, USA., ISBN: 9780471224631, pp: 497-565.
6. Zhang, C., J.Y. Chai and R. Jin, 2005. User term feedback in interactive text-based image retrieval. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM Press, New York, USA., pp: 51-55. http://doi.acm.org/10.1145/1076034.1076046
7. Gong, Z., H.U. Leong and C.W. Cheang, 2006. Web image indexing by using associated text. Knowl. Inform. Syst., 10: 243-264. DOI: 10.1007/s10115-005-0231-8
8. Rowe, N.C., 2002. Finding and labeling the subject of a captioned depictive natural photograph. IEEE Trans. Knowl. Data Eng., 14: 202-207. DOI: 10.1109/69.979983
9. Sclaroff, S., M.L. Cascia, S. Sethi and L. Taycher, 1999. Unifying textual and visual cues for content-based image retrieval on the World Wide Web. Comput. Vision Image Understand., 75: 86-98. http://portal.acm.org/citation.cfm?id=329175.329194.
10. Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Retrieval Evaluation. In: Modern Information Retrieval, Baeza-Yates, R. and B. Ribeiro-Neto (Eds.). Addison Wesley, New York, USA., ISBN: 0-201-39829-X, pp: 73-97.
11. Shafi, S.M. and R.A. Rather, 2005. Precision and Recall of five search engines for retrieval of scholarly information in the field of biotechnology. Webology, 2: 1-12. http://www.webology.ir/2005/v2n2/a12.html
12. Noah, S.A., A.C. Alhadi and L. Zakaria, 2005. A semantic retrieval of web documents using domain ontology. Int. J. Web Grid Servic., 1: 151-164. DOI: 10.1504/IJWGS.2005.008359