Combination of Keyword and Visual Feature based Image Retrieval System

Htwe Htwe Kyaing, Mi Mi Nge
University of Computer Studies, Yangon
htwehtwekyaing@gmail.com

Abstract

*Keyword-based image retrieval systems have become popular for many image database applications. To improve the performance of keyword-based web image queries, combination of keyword and visual feature based image retrieval system is presented in this paper. Firstly, DOM (Document Object Model) trees are constructed from collected web pages. And several text blocks are segmented based on text cohesion. Then, visual features are extracted from color images in RGB (Red, Green and Blue) color space by using color histogram. When user query is entered, text blocks which contain web images are taken as the associated texts of corresponding images and TF*IDF values are used to index web images. Finally, keyword and visual features are combined by using Gaussian Mixture Model to produce the relevance images.*

Keywords: DOM tree, text cohesion, keyword, visual feature, Gaussian mixture model

1. Introduction

Information retrieval is the processing of user requests, commonly referred to as queries, to obtain relevant information. Web images have been becoming one of the most important information types on the Web. Web images are used by Web page authors as visual presentations for their Web documents.

With millions of images on web and to retrieve relevant images is important in today's image retrieval process. Therefore, image retrieval process drew more and more attention in research area. It needs mature and agreeable techniques to support effective image queries. Text-based methods take the associated text as input and try to derive the content of a web image. Typically, image file names, anchor texts, surrounding paragraphs, even the whole texts of the Web documents are considered for the extractions.

In most systems, web images are represented into vectors of term-weight pairs. In order to correctly correlate terms to a web image, the associated text of the web image is partitioned into text blocks according to the structure of the text with respect to the web images. When a user enters keyword, result images are generated by image indexing and searching algorithms. This paper presents the image retrieval system by combination of text-based relevance model with visual feature relevance model.

The rest of the paper is organized as follows. Section 2 is related work. Section 3 is web page segmentation and visual feature extraction is presented in section 4. In section 5, image retrieval process is described. In section 6, proposed system design is described. System implementation is presented in Section 7. Section 8 is the conclusion of the system.

2. Related Work

Most of the web image search systems are based on only keyword based searches. Those systems use surrounding blocks of text to index the corresponding images. Web images, with close visual features, may have great differences in their semantics. R.Lempel and A.Soffer presented web image search system based on web links [5]. Two web images are supposed to be similar if they are co-cited by many web pages. It is extended from some well known link-based web page schemes to the context of web images retrieval. In V.Harmandas , M.Sanderson and M. D. Dunlop, links are used to track source pages of the image container page [4]. Then, container page and its source pages are used together to index the corresponding image. Results of above systems are a bunch of images which may include non-relevant images.

In general, traditional solutions often employ linear models to combine text and visual features for search of web images. In H.Feng and T-S.Chua, X-J Wang, W-Y Ma and G-R Xue and K.Yanai and K. Barnard, such methods cannot make use of the co-enforcement intrinsic between text similarity space and visual similarity space [1,7,8]. H.Feng and T-S.Chua described a bootstrapping method which uses a text classifier and a visual feature classifier to successively co-train the relationships between web images and text concepts [1]. But it needs to start its work with a small set of manually labeled sample images. In K.Yanai and K. Barnard, a Gaussian mixture model was employed as the visual feature classifier model [8]. In X-J Wang, W-Y Ma and G-R Xue, web image retrieval was improved with reinforcement between text similarity model and visual feature similarity model [7]. Image similarity is reinforced by propagation of similarities between text data set and image data set. Text data refer to text blocks which contain web images, and image data refer to web images themselves.

3. Web Page Segmentation

In the web page segmentation process, DOM tree is constructed from collected web pages. Web pages are represented as a DOM tree, with nodes as HTML tags. Then text blocks are segmented according to DOM tree by using text cohesion algorithm. Sibling elements under the same parents are merged if the text cohesion among them is over a pre-defined threshold. Thus a web page can be segmented into several text blocks tb_1, tb_2,..., tb_n.

3.1. DOM Tree

For the web page segmentation, HTML tags of web pages are represented as nodes of DOM tree. The HTML DOM views a HTML document as a node-tree. All the nodes in the tree have relationships to each other. All nodes can be accessed through the tree. Example DOM tree is shown in Figure 1.

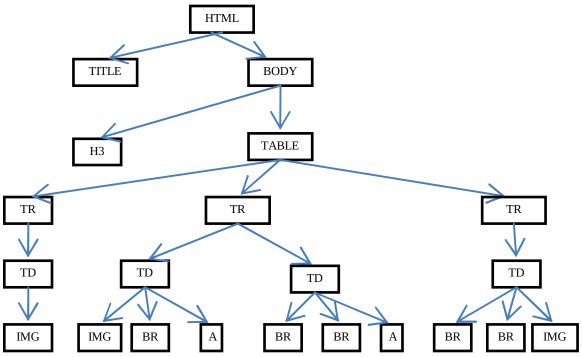

Figure 1: Example DOM Tree for HTML Web Page

The tree starts at the root node and branches out to the text nodes at the lowest level of the tree. Some of the nodes such as <script>, <font> and comments <!-- --> are not processed. Tags such as <IMG> and <A> cannot be parent node, they can only be leaf node in building DOM tree. Each element has exactly one parent node.

3.2. Text Cohesion

Text contained in the same HTML tag element with image i is more relevant to that image. Relevant text block of web image i is expanded to include all the siblings of tb if their text cohesion is higher than a given threshold. In order to find the text cohesion between text blocks, Chameleon algorithm is used. Cohesion of text block (tb) is measured by the semantic relevance between blocks b_1, b_2. With Chameleon algorithm, cohesion is computed based on relative closeness and relative inter-connectivity between blocks b_1, b_2 as follows.

$$coh\ (tb) = RC\ (b_1, b_2)\ RI\ (b_1, b_2) \qquad \text{Eqn.(1)}$$

Where,

RC = Relative closeness between b_1 and b_2

RI = Relative inter-connectivity between b_1 and b_2

RC and RI can be computed as follows:

$$RC(b_1, b_2) = \frac{S_{EC\ (b1,\ b2)}}{\frac{|b_1|}{|b_1|+|b_2|} S_{EC\ (b1)} + \frac{|b_2|}{|b_1|+|b_2|} S_{EC\ (b2)}} \qquad \text{Eqn.(2)}$$

$$RI(b_1, b_2) = \frac{EC_{\ (b1,\ b2)}}{\frac{1}{2}\left(EC_{\ (b1)} + EC_{\ (b2)}\right)} \qquad \text{Eqn.(3)}$$

Where, $|b_1|$ = size of blocks b_1

$|b_2|$ = size of blocks b_2

$S_{EC(b1,\ b2)}$ = average value of edges across b_1, b_2,

$S_{EC(bi)}$ = average value of edges within block b_i,

EC (b_1, b_2) = sum of edges across b_1 and b_2,

EC(b_i) = the sum of edges within block b_i.

RC = relative closeness between these two blocks

RI = relative interconnectivity between these two blocks.

4. Visual Feature Extraction

In image processing, feature extraction is a special form of dimensionality reduction. RGB values of input image are transformed into a reduced set of features. Transforming into the set of features is called feature extraction. Features set extract the relevant information from the input data using this reduced representation instead of the full size input. In the Visual Feature Extraction phase, color images of RGB color space and visual features are extracted through color histogram. A color model is an abstract mathematical model describing the way colors can be represented as tuples of numbers, typically as three or four values or color components.

4.1. Histogram Computation

In image processing, a color histogram is a representation of the distribution of colors in an image. A digital image has a set of pixels and each pixel stores color values. Image features are generated via image histogram. They are useful in a variety of applications, especially image classification and image clustering. In this system, image features are used to compute the probability distribution model. Histogram values are computed for each pixel in the image file into a set of values (double array). Red, Green and Blue values of each pixels are grouped into 4 regions (0-63, 64-127, 128-191, 192-255), and computed by following equations.

binCount = number of regions (default value = 4)

Index for each color value is computed as follows:

*idx = colorValue * binCount / 255* Eqn.(4)

After computing the idx for each color value (R, G, B), then histogram indexes are computed as follows:

*idx = i1 + i2 * b1 + i3 * b1 * b2* Eqn.(5)

Where, i1 = red index, i2 = green index, i3 = blue index

b1 = bincount for red, b2 = bincount for green, b3 = bincount for blue

Then all histogram values are normalized to get value between 0 and 1.

5. Image Retrieval Process

For each query q, the results for q can be a long list. There can be large amount of noise or irrelevant images in the results. To reduce the influence of those noise images, this paper combines TF*IDF based relevance with visual feature distributions. Important images for the same concept are similar in visual features. Gaussian Mixture Model is used to describe the visual feature distribution for retrieving relevant images. Image's relevance to the concept in visual feature space is defined as the ratio of positive web image distribution density over negative web image distribution density. Thus, a combined relevance model is defined by using both text-based relevance and feature-based relevance.

For any web image i_j, let T_1, T_2, T_3 and T_4 represent the block types of image's owner page title and related text block, image ALT, image name and image caption respectively. For any term w_k, the weight of term w_k associated to image i_j is defined as

$$w(w_k, i_j) = \sum_{l=1}^{4} \omega_l * tf\ (T_l(i_j), w_k) * idf\ (T_l(i_j), w_k)$$

Eqn.(6)

Where, $tf(T_l(i_j),w_k)$ = the term frequency of w_k in text block $T_l(i_j)$,

$idf(T_l(i_j),w_k)$ = the inverse document frequency of w_k in all the text blocks of type T_l,

ω_l = the weight of block type T_l and its value is determined with using cosine similarity algorithm.

IDF (Inverse Document Frequency) is an important measure that represents the scaling factor, or the importance, of a term t. If a term t occurs in many documents, its importance will be scaled down.

$$\text{IDF}\ (t) = \log\ \frac{1 + |d|}{|d_t|}$$

Eqn.(7)

Where |d| = number of all documents in document collection

$|d_t|$ = number of documents containing term *t*.

For measuring document similarity based on relative term occurrences of document vectors, cosine similarity measure is widely used and it can be defined as follows:

$$\text{Sim}\ (S_a, S_b) = \frac{\sum_{i=1}^{n}(S_{ai} * S_{bi})}{\sqrt{\sum_{i=1}^{n} S^2_{ai} * \sum_{i=1}^{n} S^2_{bi}}}$$

Eqn.(8)

Then conditional probability of term w_k, given a web image i_j is defined as:

$$P_T(w_k \mid i_j) = \frac{w(w_k, i_j)}{\sum w_{k \in T(ij)} w(w_k \mid i_j)}$$

Eqn.(9)

Similarly, conditional probability of a web image i_j, given a term w_k is defined as:

$$P_T(i_j \mid w_k) = \frac{w(w_k, i_j)}{\sum_{Ij \in I} w(w_k \mid i_j)}$$

Eqn.(10)

Because the training set $I(w_k)$ for P_I may contain irrelevant web images, this system uses Gaussian Mixture Model to alleviate the influences of those irrelevant training images. w_k web images are used to model the distribution of w_k images and non-w_k web images to model the distribution of non-w_k images. Relevance is defined as the ratio of positive distribution over negative distribution.

$$R_F(i_j \mid w_k) = \frac{P_1(i_j \mid w_k)}{P_2(i_j \mid \text{non}–w_k)}$$

Eqn.(11)

5.1. Gaussian Mixture Model

A Gaussian mixture (GM) is defined as a convex combination of Gaussian densities. A Gaussian density in a dimensional space is characterized by its mean and covariance matrix. For the calculation of relevance model, Gaussian Mixture Models are constructed as follows:

$$\text{EXP}_l = \exp^{-\frac{1}{2}(i - \mu_{l,j})^T \sum_{l,j}^{-1} (i - \mu_{l,j})}$$

Eqn.(12)

$$P_1(i \mid w_k) = \sum_{j=1}^{m_1} \omega_{1,i} \frac{1}{\sqrt{(2\pi)^N \mid \Sigma_{1,j} \mid}} EXP_1$$

Eqn.(13)

$$P_2(i \mid non - w_k) = \sum_{j=1}^{m_2} \omega_{2,i} \frac{1}{\sqrt{(2\pi)^N \mid \Sigma_{2,j} \mid}} EXP_2$$

Eqn.(14)

where P_1 and P_2 are density functions for conditional probability given w_k and non-w_k respectively, i represents any image, N is the dimension of the image feature, m_1 is the number of positive components $\omega_{1,j}$, $\mu_{1,j}$, $\Sigma_{1,j}$ represents the weight, mean vector and the covariant matrix of the j-th positive component.

6. Proposed System

This paper presents image retrieval by combining keyword and visual feature extraction. Overview process of the proposed system is shown in Figure 2, where there is a bunch of web documents in web corpus. For each web document in this collection, DOM tree is built and text cohesion is computed to segment the web page. Then terms and indexes of each image in web page stored in image database as indexing step to fasten the image search process. When a user enters keyword to search, term vector is prepared based on input query (keyword). Then cosine similarity is used to compute the resulted images of input keyword, known as keyword search results. It may contain the irrelevant images, and Gaussian Mixture Model is used to filter out irrelevant images. In this step, visual features of result images are applied to compute the probabilities of distribution model.

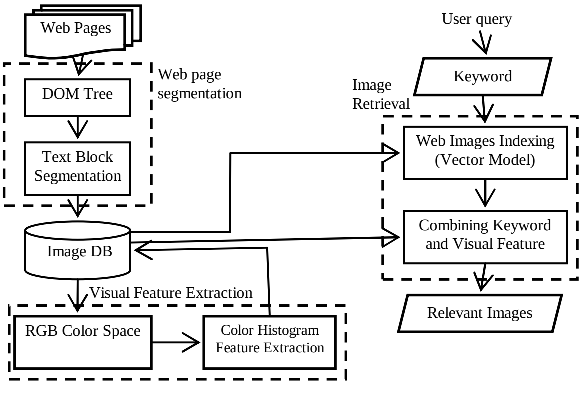

Figure 2: System overview

Proposed system mainly includes three components:

- **Web page segmentation**: web document is built into DOM tree structure and text cohesion is computed to get the relevance between text blocks.
- **Visual feature extraction**: histogram values of images are used as visual features in this system. Features are applied in Gaussian Mixture model to compute the relevance value in the later process.
- **Image retrieval**: this phase includes web image indexing and combining of keyword and visual features to get more relevant images. In order to compute the relevance probability, extracted features of result images are applied to Gaussian Mixture Model. Probabilities for positive samples and negative samples are computed by Gaussian Mixture Model and ratio of probability of positive samples to probability of negative samples is used to find the relevance of input image features.

7. System Implementation

This system is implemented as web-based image retrieval system by using C# language. Step by step process for a single web page is presented as a case study. Web pages containing images of computers and accessories are collected for the search process. Web page segmentation, visual feature extractions and image indexing are pre-processed and terms and image indexes are stored in the database. When the user enters keywords, image results are generated by keyword based search and relevant image results are filtered by Gaussian Mixture Model.

7.1. Case Study

In this paper, four hundred web pages are downloaded and used as input for storing terms, images indexes and images in the database. A web page including thirty lines is prepared with two blocks in the same table. Firstly, DOM Tree is built for the input web page. According to DOM Tree, there are two blocks under the same parent and text cohesion of those two blocks is necessary to compute. In text cohesion computing, the total numbers of edges in first and second blocks have to be counted.

In size of block b_1 includes 12 edge- DELL LATITUDE D810 2130MHz 1024MB 60GB (WIN XPP+) 15.4 DVDRW 10/100 NIC.

In size of block b_2 includes 12 edge- DELL LATITUDE D610 2130MHz 512MB 40GB (WIN XPP+) 14 DVDRW 10/100 NIC

To compute the text cohesion, coh (tb) of blocks b_1 and b_2 by Equation (1), relative closeness RC(b_1, b_2) by

Equation (2), and relative inter-connectivity $RI(b_1, b_2)$ by Equation (3), should be calculated first.
$S_{EC(b1, b2)}$ = Average value of Edges across b_1, b_2 = (8 + 4) / 16 = 0.75
$S_{EC(b1)}$ = (8 + 2) / 12 = 0.833
$S_{EC(b2)}$ = (8 + 2) / 12 = 0.833
$EC(b_1, b_2) = 12$
$EC(b_1) = 10$
$EC(b_2) = 10$
$|b_1| = 12$
$|b_2| = 12$

$RC(b_1, b_2)$ = 0.75 / ((0.5 * 0.833) + (0.5 * 0.833)) = 0.9
$RI(b_1, b_2)$ = 12 / 10 = 1.2
$coh(b_1, b_2)$ = 1.08

Therefore, those two blocks are merged since they have strong cohesion. After computing text cohesion, color images of visual features can be extracted through color histogram computed according to Equation (4) and (5). Image features are then stored in the database as indexes.

Vector space model is constructed based on user keyword. User keyword is the keyword entered by user, also known as user query. User query is tokenized into words and stop words (non-interesting words such as ‘of’, ‘at’) are removed. In this paper, images for **computer related terms** are collected. Therefore keywords relating to computer terms can be searched in this case study.

User Keyword = Dell D810
First Block
T_1 = “Dell – Auction” /Title/
T_2 = “Dell Latitude D810” /Alt/
T_3 = “D810.jpg” /Name/
T_4 = “DELL LATITUDE D810 2130MHz 1024MB 60GB (WIN XPP+) 15.4 DVDRW 10/100 NIC” & “DELL LATITUDE D610 2130MHz 512MB 40GB (WIN XPP+) 14 DVDRW 10/100 NIC” /Caption/

In this case study, we assume these facts. There are 10 documents. DELL contains in 5 documents and D810 contains in 1 document. Thus, the IDF values of terms “DELL” and “D810” can be computed by using Equation (7) as follow:
IDF (DELL) = log (1 + 10) / 5 = 0.3424
IDF (D810) = log (1 + 10) / 1 = 1.041

Table: Vector Model of First Image Block and Second Image Block

	Dell (tf * idf)	**D810 (tf * idf)**
User Query	1	1
Block 1, T1	1 * 0.3424 = 0.3424	0 * 1.041 = 0
Block1, T2	1 * 0.3424 = 0.3424	1 * 1.041 = 1.041
Block1, T3	0 * 0.3424 = 0	1 * 1.041 = 1.041
Block1, T4	2 * 0.3424 = 0.6848	1 * 1.041 = 1.041
Block1	4 * 0.3424 = 1.3696	3 * 1.041 = 3.123

Similarity of user query and b_1 can be computed by using Equation (8) which is described as follows:

$$W_1 = \frac{(1 * 1.3696) + (1 * 3.123)}{\sqrt{(1 + 1) * (1.3696 + 3.123)}}$$

W_1 = 4.4926 / Sqrt (2 * 4.4926)

W_1 = 4.4926 / 2.9975 = 1.4987

Weight for First Image Block is W_1 = 1.4987.
$W(w_k, img_1)$ for term Dell = (1.4987 * 1) + (1.4987 * 1) + (1.4987 * 0) + (1.4987 * 2) = 5.9948
$W(w_k, img_1)$ for term D810 = (1.4987 * 0) + (1.4987 * 1) + (1.4987 * 1) + (1.4987 * 1) = 4.4961

Conditional Probability of term “Dell” given a web image img_1 and term “D810” given a web image img_1 can be calculated by using Equation (9).
$P_T(w_k|img_1)$ for term Dell = 5.9948 / ((0.3424 * 5.9948) + (0.3424 * 5.9948) + (0 * 5.9948) + (0.6848 * 5.9948)) = 0.7301
$P_T(w_k|img_1)$ for term D810 = 4.4961 / ((0 * 4.4961) + (1.041 * 4.4961) + (1.041 * 4.4961) + (1.041 * 4.4961)) = 0.3202
Therefore, Average conditional probability for both terms is obtained as P_T = 0.52515.

Conditional Probability of web image img_1 given a term “Dell” and web image img_1 given a term “D810” can be calculated by using Equation (10).
$P_T(img_1|w_k)$ for term Dell = 5.9948 / (3 * 5.9948) = 0.33
$P_T(img_1|w_k)$ for term D810 = 4.4961 / (3 * 4.4961) = 0.33
Therefore, Average conditional probability for web image img_1 is obtained as P_T = 0.33.

If we set threshold = 0.3, for the indexing of web images, we got image from block 1 as the results for user query “Dell D810”. In the practical results, there may be a large amount of images produced by keyword search algorithm and Gaussian Mixture Model is used to filter out the irrelevant results. It is necessary to compute probability distribution over positive samples (P_1) by Equation (13) and probability distribution over negative samples (P_2) by Equation (14). Then ratio of P_1 over P_2 is calculated to measure the relevance of the input result image by Equation (11).

8. Conclusion

This system presents the processes of combining keyword and visual feature based web image retrieval. Firstly, web pages into several text blocks based on their semantic cohesions are segmented, blocks which contain web images as associated texts for the corresponding web images. A probability model is created to define the terms’ relevant degrees to the images. Non relevant images are then removed using Gaussian Mixture Distribution Model to enhance the relevant degrees from the aspect of visual feature distributions. Therefore, this system uses not only keyword based search algorithm, in order to enhance the relevant degrees, visual feature enhancement is also used, improving the results of the web image retrieval process.

9. References

[1] Feng H. and Chua T-S., "A bootstrapping approach to annotating large image collection" In: Sebe N, Lew M S, and Djeraba C (eds) Proc. ACM Int'l. Conf. Multimedia Information Retrieval, pp. 55-62, 2003.

[2] Gong Z., Hou L.U. and Cheang C.W., "Web Image Indexing by Using Associated Texts", Faculty of Science and Technology, University of Macau, Macao, PRC, Jun 2005.

[3] Gong Z., Hou L.U. and Cheang C.W., "Web Image Semantic Clustering", In proceeding of Springer-Verlag Berlin Heidelberg, CoopIS/DOA/ODBASE 2005, LNCS 3761, pp. 1416 – 1431, 2005.

[4] Harmandas V., Sanderson M., and Dunlop M. D., "Image retrieval by hypertext links", In: Proc. SIGIR-97, 20th ACM Int'l. Conf. Research and Development in Information Retrieval. Philadelphia PA, USA, pp. 296-303, 1997.

[5] Lempel R. and Soffer A. "PicASHOW: Pictorial authority search by hyperlinks on the Web. ACM Transactions on Information Systems, vol. 20, no. 1, pp. 1-24, 2002.

[6] Verbeek J.J., Vlassis N. and Kr¨ose B., "Efficient Greedy Learning of Gaussian Mixture Models", Published in Neural Computation 15(2), pages 469-485, 2003.

[7] Wang X-J, Ma W-Y and Xue G-R, "Multi-model similarity propagation and its application for web image retrieval", In: Proc. ACM int'l Conf. MM, pp. 944-951, 2004.

[8] Yanai K. and Barnard K., "Probabilistic web image gathering" In: Proc. Seventh ACM SIGMM Int'l. Workshop Multimedia Information Retrieval. ACM Press, New York, NY, USA, pp. 57-64, 2005.