Information Fusion 12 (2011) 223–230

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Using Multi-Modal Semantic Association Rules to fuse keywords and visual features automatically for Web image retrieval

Ruhan He [a,*], Naixue Xiong [b], Laurence T. Yang [c], Jong Hyuk Park [d]

[a] College of Computer Science, Wuhan University of Science and Engineering, Wuhan 430073, China
[b] Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
[c] Department of Computer Science, St. Francis Xavier University, Antigonish, Canada
[d] Department of Computer Science and Engineering, Seoul National University of Technology, 172 Gongreung 2-dong, Nowon-gu, Seoul 139-742, Republic of Korea

ARTICLE INFO

Article history:
Received 15 April 2008
Received in revised form 26 December 2009
Accepted 18 February 2010
Available online 1 March 2010

Keywords:
Web image retrieval
Multi-Modal Semantic Association Rule (MMSAR)
Association rule mining
Inverted file
Relevance Feedback (RF)

ABSTRACT

A recent trend for image search is to fuse the two basic modalities of Web images, i.e., textual features (usually represented by keywords) and visual features for retrieval. The key issue is how to associate the two modalities for fusion. In this paper, a new approach based on Multi-Modal Semantic Association Rule (MMSAR) is proposed to fuse keywords and visual features automatically for Web image retrieval. A MMSAR contains a single keyword and several visual feature clusters, which crosses and associates the two modalities of Web images. A customized frequent itemsets mining algorithm is designed for the particular MMSARs based on the existing inverted file, and a new support–confidence framework is defined for the mining algorithm. Based on the mined MMSARs, the keywords and the visual features are fused automatically in the retrieval process. The proposed approach not only remarkably improves the retrieval precision, but also has fast response time. The experiments are carried out in a Web image retrieval system, *VAST* (VisuAl & SemanTic image search), and the results show the superiority and effectiveness of the proposed approach.

1. Introduction

Multi-source information fusion is rapidly emerging as a discipline to reckon with and is finding ever-increasing applications in biomedical, industrial automation, aerospace systems and environmental engineering, etc. The latest advances in image processing technology also seek to combine image data from several diverse types of sensors in order to obtain a more accurate view of the scene [1,2]. The fusion is expected to give better spatial coverage, redundancy, robustness and accuracy.

In the field of Web image retrieval, the current commercial search engines, including Google image search, Lycos and AltaVista photo finder, use text (i.e., surrounding words) to look for images, without considering image content [3]. However, when the surrounding words are ambiguous or even irrelevant to the image, the search based on text only will result in many unwanted result images. In contrast, Content-Based Image Retrieval (CBIR) systems, such as QBIC [4], Photobook [5], VisualSEEk [6], SIMPLicity [7] and Blobworld [8], are proposed to utilize only low-level features, such as color, texture and shape, to retrieve similar images. They do not incorporate textual information to improve the semantic quality of image search, which leads to many of the retrieved images are visually similar but not semantic related to the query.

In fact, the images on the Web are generally exposed as part of Web contents [9]. It has two-source information or two-modal characteristics (i.e., textual and visual features), which bring information that are complementary and that can disambiguate each other [10]. Therefore, a Web image should be specified not only by the image itself, but also with respect to the Web contents surrounding the image [3]. Hence, it is necessary and valuable to fuse textual features (usually represented by keywords) and visual features for image retrieval in the Web context.

Based on information fusion idea, a recent trend for image search is to fuse the two basic modalities of Web images (i.e., textual features and visual features) for retrieval. The key issue is how to associate the two modalities for fusion. The existing methods for the fusion can be categorized into the automatic method and the non-automatic method. The non-automatic method is mainly the interactive Relevance Feedback (RF) [11–16], which needs the user to interact with the system for feedback after the initial retrieval. Meanwhile, the automatic method mainly includes the Pseudo-Relevance Feedback (Pseudo-RF) method [17], the online clustering method (i.e., re-ranking by online clustering) [18–20] and the long-term RF learning (LT-RF) [21]. In these works, little attention has been paid to associating and fusing the two basic modalities of

[*] Corresponding author. Tel.: +86 027 87181191; fax: +86 027 87403983.
E-mail addresses: heruhan@gmail.com (R. He), nxiong@cs.gsu.edu (N. Xiong), ltyang@stfx.ca (L.T. Yang), parkjonghyuk1@hotmail.com (J.H. Park).

1566-2535/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2010.02.001

Web images by association rule mining technology despite it is extensively used in CBIRs and text-based search engines.

In addition, according to recent statistical analysis [22,23], there are three typical behaviors (which can be called "lazy user" problem [21]) for Web users: (1) usually input very short queries; (2) rarely make use of feedback to revise the search; (3) almost always pay attention to only the top 10–20 retrieved results. According to the second behavior, we know that the interactive Relevance Feedback is rarely employed by Web users despite it is very effective. Therefore, the practicability of the interactive RF method is limited in real-life Web image retrieval systems because Web users seldom use it. Consequently, the long-term RF learning (i.e., the offline RF learning) is also limited for the insufficient data of the log of RF sessions that is difficult, expensive or time consuming to obtain. Hence, we think, it is better to be an auxiliary or optional means for the RF methods (including interactive RF and long-term RF learning) in the Web system due to the "lazy user" problem.

Therefore, in this paper, we mainly focus on the non-RF methods that fuse text and visual features automatically (i.e., without user's additional assistance) in Web image retrieval. From data mining perspective, we propose a new approach based on the particular Multi-Modal Semantic Association Rule (MMSAR) to fuse keywords and visual features automatically for Web image retrieval. A MMSAR contains a single keyword and several image visual feature clusters. It includes two modalities (i.e., textual and visual feature space) of Web image, which is different from the traditional single modality association rule. In our approach, based on the existing inverted file, which relates keywords and their associated images, a customized mining process is also developed for the particular semantic association rules. The MMSARs effectively associate keywords and visual feature clusters and are used to fuse the two modalities of Web images automatically for Web image retrieval. The proposed approach is implemented in our Web image search prototype system (i.e., *VAST* system [24]) and reveals excellent performance.

The remainder of this paper is organized as follows. In Section 2, we introduce some related work. In Section 3, the mining process of the MMSAR is described. In Section 4, we depict how the MMSARs are exploited in our system to fuse keywords and visual features automatically for Web image retrieval. In Section 5, experiments are carried out to evaluate the proposed method. Finally, we conclude our work and give some future work.

2. Related work

Based on information fusion idea, the methods that fuse text and visual features in Web image retrieval can be categorized into the automatic method and the non-automatic method. The non-automatic method is mainly the interactive Relevance Feedback (RF), which needs the user to interact with the system for feedback after the initial retrieval. Most of the past Web image retrieval systems, such as Pictoseek [11], WebMars [12], Webseer [13], ImageRover [14], Cortina [15] and Atlas WISE [16], fuse text and visual features together by the interactive RF. Meanwhile, the automatic method mainly includes the Pseudo-Relevance Feedback (Pseudo-RF) method [17], the online clustering method (i.e., re-ranking by online clustering) [18–20] and the long-term RF learning (LT-RF) [21].

The Pseudo-RF method [17,25,26] in information retrieval always assumes the top-*N* documents are the relevant ones for feedback. In Web image retrieval, the top-*N* images are assumed to be the relevant images for the query. In [17], SVM-based learning is used to learn the top-*N* images. However, the performance of the Pseudo-RF method in Web image retrieval strongly relies on the correctness of the assumption (i.e., the good quality of the top-*N* images that are used for feedback). When there are only a few relevant images in the top-*N* images, this method might be ineffective.

Some work in image retrieval uses visual feature-based online clustering method to improve the retrieval performance [18,20,27]. In [27], the system clusters the keyword-based retrieval results by the visual features of image and thus automatically reranks the results set. However, these methods cannot improve the retrieval effectiveness when only a few correct images are contained within the top-*N* results [27]. Moreover, these online clustering methods increase the response time of the query remarkably, which reduce the usability of the Web system.

In traditional CBIR, LT-RF learning is commonly combined with interactive RF [28–32]. In Web image retrieval, the log of the user's interactive RF is used to discover the relationship between low-level and high-level features [21]. The text description is combined with the low-level image features in the image similarity assessment and thus improves the image retrieval performance. However, as mentioned in Section 1, the LT-RF method is not appropriate to be a major means, but an auxiliary or optional means in Web systems due to the "lazy user" problem.

In [33,34], Latent Semantic Indexing (LSI) has been applied to find relationships between keywords and images. But they are jointed with the interactive RF. In fact, they use the interactive RF to fuse keywords and image features into one LSI space. They belong to the category of the interactive RF.

Cortina [15] uses MPEG-7 visual features [41] to cluster the images, and mining the association between the keywords and the MPEG-7 visual features. It adopts the standard support–confidence framework for the association rule mining, which results in extremely low support of the association rule (most of rules are lower than 0.1% with support). Therefore, the low support maybe reduces the usability of the rules. In addition, the associations are also utilized in the interactive RF process.

Motivated by the success of data mining in CBIRs and text-based search engines, our approach mines the semantic rules containing one keyword and several visual feature clusters offline based on the existing inverted file, and applies the semantic rule online to realize the fused retrieval automatically.

3. Multi-Modal Semantic Association Rule (MMSAR) mining

To discover the Multi-Modal Association between the high-level keywords and low-level visual features of Web image, we need to quantify the visual features by clustering, because the keyword space is discrete while the visual feature space is continuous in general. Therefore, we aim to associate the keywords and the visual feature clusters.

The mining process of MMSARs is as follows:

(1) Construct the transaction database *D* and the basic candidate 2-itemsets based on the existing inverted file.

We do not start from 1-itemset because the visual features are very high dimensional and the associations between keywords (which are single modality association rules) are much stronger than the associations between keywords and low-level features or low-level visual clusters. If starting from 1-itemset, the keywords and visual feature cluster are equally treated, and then most of the created 2-itemsets based on 1-itemset are keyword and keyword, but few of keyword and visual feature cluster. Our goal is not the association between keyword and keyword. We are interested in the association between keywords and visual feature clusters. Therefore, only the itemsets containing at least one keyword and one visual feature cluster are considered.

To find the rules that consist of one keyword and several visual feature clusters, we must construct the basic 2-itemsets. The existing inverted file relates the keywords to their associated images. A record of the inverted index table *I*, which is responding to the inverted file, is defined as follows:

$$(Q_j, \{I_{j,i} | i = 1, \ldots, N_j\}) \tag{1}$$

where Q_j is a keyword, $I_{j,i}$ is an image that contains the keyword Q_j and N_j is the number of the images containing the keyword Q_j.

Assume each image has *h* different visual features, i.e., $f = \{f_1, f_2, \ldots, f_h\}$. For each image $I_{j,i}$, we build an Image-Cluster mapping defined as follows:

$$(I_{j,i}, \{C^f_{j,i,k} | 1 \leqslant f \leqslant h, 1 \leqslant k \leqslant 2\}) \tag{2}$$

where $C^f_{j,i,1}$ is the cluster over the feature *f* that $I_{j,i}$ belongs to and $C^f_{j,i,2}$ is the cluster over the feature *f* next closest to $C^f_{j,i,1}$, i.e.,

$$C^f_{j,i,1} = C^f_m, \quad C^f_{j,i,2} = C^f_n \tag{3}$$

where

$$m = \arg_u MIN\left(D_f\left(f(I_{j,i}), \overline{C^f_u}\right)\right) \tag{4}$$

$$n = \arg_{v \neq m} MIN\left(D_f\left(f(I_{j,i}), \overline{C^f_v}\right)\right) \tag{5}$$

The $\overline{C^f_u}$ and $\overline{C^f_v}$ represent the centroid of the *u*th and the *v*th clusters over the visual feature *f*. $f(I_{j,i})$ represents the visual feature of image $I_{j,i}$ over the visual feature *f*. $D_f(f_1, f_2)$ is the distance between the two features f_1 and f_2 over the visual feature *f*. We select two clusters for each image over each visual feature type for increasing the frequency of the visual feature clusters and the number of co-occurrences of keywords and visual feature clusters.

According to Eqs. (1) and (2), there are $2h * N_j$ 2-itemsets for the keyword Q_j, i.e.,

$$\left(Q_j, C^f_{j,i,k}\right) \quad (1 \leqslant f \leqslant h, 1 \leqslant i \leqslant N_j, 1 \leqslant k \leqslant 2) \tag{6}$$

We filter out the duplicate itemsets, and obtain M_j non-duplicated 2-itemsets for the keyword Q_j. Similarly, many non-duplicated 2-itemsets for other keywords can be deduced from other records of *I*. Therefore, the basic candidate 2-itemsets will be successfully constructed.

Considering the images with close size in the same Web page always have semantic homogeneity and generally have some common keywords, we assume they belong to the same category. One webpage contains one or multiple images in the inverted file, i.e.,

$$(Q_j, P_{j,s} | 1 \leqslant s \leqslant W_j) \tag{7}$$

$$P_{j,s} = \{I_{j,l} | l = 1, \ldots, W_{j,s}\} \tag{8}$$

where $P_{j,s}$ represents the images within one given webpage *s* for the keyword Q_j in the inverted file, W_j is the number of pages that contain the keyword $Q_j, W_{j,s}$ represents the number of images in $P_{j,s}$, and $\sum_{s=1}^{W_j} W_{j,s} = N_j$. Therefore, we get

$$(Q_j, \{I_{j,l} | l = 1, \ldots, W_{j,s}\}) \tag{9}$$

According to the formulas (2) and (9), we can get one itemset of transaction in database *D*, i.e.,

$$(Q_j, \{C^f_{j,l,k} | 1 \leqslant f \leqslant h, 1 \leqslant k \leqslant 2, 1 \leqslant l \leqslant W_{j,s}\}) \tag{10}$$

Therefore, there are W_j transactions for the keyword Q_j in *D*. Similarly, many transactions for other keywords can be added into *D*. Thus, the transaction database *D* is constructed.

(2) Define the support–confidence

If we directly utilize the standard support–confidence for the semantic rules, their support is extremely low, which will affect the rules mining. For different keyword Q_j, the size of the corresponding set is extremely imbalance. Only the keywords that have big result sets can obtain high support and confidence value under the standard support–confidence framework. To overcome this drawback, we define the support and the confidence of the semantic rule as follows:

$$supp(Q_j) = \frac{count(Q_j)}{|D|} \tag{11}$$

$$supp(C_i) = \frac{count(C_i)}{|D|} \tag{12}$$

$$supp(Q_j => C_i) = \frac{count(Q_j, C_i)}{count(Q_j)} \tag{13}$$

$$conf(Q_j => C_i) = \frac{count(Q_j, C_i)}{\max_t(count(Q_j, C_t))} \tag{14}$$

where *count*(*A*) is the number of itemsets that contain *A* in *D*. Obviously, $count(Q_j) = W_j$ and $|D|$ is the number of all itemsets in *D*. Similarly,

$$supp(Q_j => \{C_i | i = 1, \ldots, m\}) = \frac{count(Q_j, \{C_i | i = 1, \ldots, m\})}{count(Q_j)} \tag{15}$$

$$conf(Q_j => \{C_i | i = 1, \ldots, m\}) = \frac{count(Q_j, \{C_i | i = 1, \ldots, m\})}{\max_t(count(Q_j, C_t))} \tag{16}$$

The new definition of the support–confidence eliminates the imbalance for different keywords, and the calculation of support and confidence is restricted within the result set of the keyword respectively. However, the new definition also has a bad effect that a few of keywords with very small result set maybe get a high support and confidence. This drawback will be alleviated by defining a threshold *min_count* to filter out these keywords in the later rules mining process.

(3) Frequent itemset mining

To discover all frequent patterns of the association between keywords and visual feature clusters, a customized frequent itemsets mining algorithm is given in Table 1 based on A-priori algorithm [35,36] and the above formulas. It is used to identify the frequent itemsets and prepare for generating strong association rules. To perform a search, the user has to specify *min_supp* and *min_count* for frequent itemsets. Every subset of a frequent itemset is also frequent. Each superset of frequent itemsets with cardinality *K* belongs to a set of candidate itemsets of size *K* (i.e., S_K). It must be noted that the itemset is particular, which includes and crosses two modalities (i.e., including one keyword and several visual feature clusters). The mining process starts from 2-itemsets in fact. Moreover, the frequent itemset must contain one and only one keyword. The antecedent of the itemset is a single keyword and the consequent is several clusters.

(4) Identify the strong semantic rules

After the frequent itemsets mining, we get the frequent itemsets. By defining two thresholds (i.e., *min_supp* and *min_conf*), we can get the strong association rules like $(Q_j, \{C_i | i = 1, \ldots, m\})$. Based on the statistics, we set *min_count*, *min_supp* and *min_conf* to be 300, 10% and 30% respectively in practice. The *min_count* is used to filter out the keywords with only a few images in the inverted file.

Table 1
The customized frequent itemsets mining algorithm.

```
Input: Transaction Database $D$;
    Minimal Support Threshold $min\_supp$;
    Minimal Count Threshold $min\_count$;
Output: Frequent Itemset $L$
$L_1 = \{Q_j | count(Q_j) > min\_count\}$;
$L_1 = L_1 \cup \{C_i | count(C_j) > min\_count\}$;
$L_2 = \{(Q_j, C_i) | Q_j \in L_1, C_i \in L_1, supp(Q_j, C_i) > min\_supp\}$;
for ($K = 3; L_{K-1} \neq \phi; K++$) do
    $S_K = \{s_k | s_k$ is $K$-itemset with only one keyword in it and a combination
    of frequent sets from $L_{K-1}\}$;
    for ($\forall T \in D$) and ($\forall s_k \in S_K$) do
        if ($s_k \subseteq T$) do
            assume the only keyword that $s_k$ contains is $Q_j$;
            $supp(s_k) = supp(s_k) + 1/count(Q_j)$;
        end if
    end for
    $L_K = \{s_k | supp(s_k) > min\_supp\}$;
end for
$L = \cup\{L_K | K \geqslant 2\}$;
return $L$;
```

Table 2
Some examples of the semantic rules.

Semantic rule	Support (%)	Confidence (%)
(Great Wall, C11, C34, C65)	10.8	32.1
(Car, C22, C56)	14.4	56.7
(Great Wall, C65)	21.8	72.8

Table 2 shows some examples of the semantic rule. The keywords in the semantic rules are the English interpretation for Chinese keywords in Table 4. The C11 represents the cluster with ID 11, and the rest C34, C65, C22, can be deduced by analogy.

4. Fusing keywords and visual features by MMSARs

Based on the mined MMSARs, we re-rank the results of initial keyword-based search automatically. The semantic rules are used to identify the interesting low-level visual feature clusters. The images in these interesting clusters will rank higher. The clusters in these semantic rules can be sorted by their weighted confidence in descending order. The weighted confidence of C_i for the keyword Q_j is defined as follows:

$$w\_conf_{Q_j}(C_i) = \sum_{(Q_j, C_i) \subseteq S} length(S) * conf(S) \tag{17}$$

where S represents the rule that contains (Q_j, C_i) and $length(S)$ represents the length of the rule S.

The automatic re-ranking process can be defined as fellows:

(1) A user enters a query keyword Q_j.
(2) Run the plain textual ranked boolean query based on the inverted index and get the initial retrieved images result $R(Q_j)$.
(3) Look up the association rules table, if exists the association rules containing the keyword Q_j, like $(Q_j, \{C_i | i = 1, \ldots, m\})$, then go on; else go to step 9.
(4) According to formula (17), we order the clusters in these semantic rules containing keyword Q_j by their weighted confidence.
(5) For the images in $R(Q_j)$, the images in the cluster with the largest weighted confidence ranks first, then subsequently the images in the cluster with the second largest weighted confidence, and the same is valid for the rest.
(6) The images within one cluster are ranked by their visual distance (e.g. EMD [37]) to the cluster centroid in ascending order.
(7) The images not in these clusters are put in the last by their original order.
(8) Get the re-ranked results based on (5)–(7).
(9) Return the result to user.

It can be seen from the above, if there exists association rules for the query keyword Q_j, the images that both are contained in the interesting clusters of the rules and include the keyword Q_j have higher ranks. Otherwise, the re-ranking process reduces to pure keyword-based search. In addition, the selection of visual features, which is another difficult problem in image retrieval area, is automatic and dynamic in our re-ranking process.

5. Experiments and evaluations

5.1. Experimental setup

We have evaluated the proposed approach in a prototype of Web image retrieval system, *VAST* (VisuAl & SemanTic image search) [24], with a database of about 1,000,000 Web images collected from the Websites listed in Table 3. We choose these Websites because they are abundant of images, have faster download speed, and give a comparatively accurate textual description for the images. These images are quite heterogeneous, including peoples, natural scenes, animals, plants, buildings, indoor scenes and sports, etc.

We keep only JPEG, PNG, GIF and BMP images, and filter out those images whose aspect ratios are greater than 3 or less than 1/3, or whose widths or heights are less than 80 pixels. The filtered images are most probably logos or advertisements or with low quality.

The textual features of Web images are extracted from multiple sources on the Web page that contains the image, including image filename and URL, ALT (alternate) text, surrounding text, page title, image hyperlinks, anchor text, etc. The $TF * IDF$ method [38] is used to weight each keyword in the textual feature vector and the *consine* metric is used to calculate the textual similarity between the query and one image.

As for the image visual features and their clustering, we adopt the region features and k-means clustering algorithm [42]. The selected color feature space is $L * U * V$ color space. We use a fast image segmentation algorithm, which is mainly based on watershed algorithm and c-means algorithm [39]. Considering the speed and the perceptual character of human being, we set the number of the segmented regions to be six at present. The Earth Mover's Distance (EMD) [37] is used to compute the visual similarity of two images. The distance metric of two regions is Euclidean distance in the EMD algorithm. For each region, we extract their average $L * U * V$ color and 3-level wavelet texture currently. The average color in $L * U * V$ color space are used to represent the color distribution. In 3-level Db4 wavelet decomposition , the mean and standard deviation of the coefficients in each subband uniquely characterize a texture [43].

Table 3
Website list of our image database.

www.china-image.cn
www.photolib.chinanews.com.cn
www.fotosearch.cn
www.phototime.cn
www.chinatuku.com
www.picture.com.cn
www.southcn.com/travel/photo/
www.tour-magazine.com

The EMD distance distribution of the region color feature can be matched with a known normal distribution ($\mu = 209,\ \sigma = 59$) as Fig. 1, which is based on the statistics of about 250,000 real Web images in dataset. Based on the distribution, the similarity is normalized. Similarly, texture feature is also normalized.

The *k*-means algorithm [42] is chosen to cluster each visual feature due to its high effectiveness. Based on the EMD distribution of the images in Fig. 1, we take a sample to calculate the cluster centroids initially in the 250,000 real Web images. Initially, 100 such clusters are created over each feature type (the max radius of the cluster is set to be $\sigma/2$ initially), which offer a good trade-off between speed and accuracy. Because the distribution of the images over the clusters is very uneven, we partition the big cluster that is larger than a threshold *T* (currently be set to 20,000) into smaller clusters of about 10,000 images per cluster. Each cluster is labeled with a unique id.

Based on an investigation of image's textual annotations, we selected 30 Chinese query keywords shown in Table 4 for experiments, which have both the largest frequencies and the explicit meaning for objective evaluation. This is because we need to ensure that sufficient images can be retrieved for evaluation. For convenience of expression, the Chinese keywords are replaced by its corresponding English interpretation in the following.

We gather the top-*K* images from results for experiments, and evaluate the effectiveness in comparing the Precision and Recall value. Precision and Recall are calculated as follows: Precision = the number of relevant retrieved images/the total number of retrieved images, Recall = the number of relevant retrieved images/the number of all relevant images. For the evaluation, we marked relevant images and irrelevant images manually about top-*K* images. We set $K = 100$ in our experiments.

Four methods have been used for comparison, namely the keyword only method, the Pseudo-RF method, the online clustering method and our approach (i.e., the semantic rules method). In the Pseudo-RF method, the top-*N* images are used for feedback. We set $N = 10$ based on the investigation in [40]. In the online clustering method, we adopt the centroid-of-largest-cluster mode with the max radius of the cluster equal to $\sigma/2$, which is the best reported in [27].

The interactive RF and the LT-RF method are not considered in the experiment because the former is not an automatic mode and the latter lacks the learning data. In fact, they can be combined with the Pseudo-RF method, the online clustering method, and the semantic rules method as an auxiliary or optional method in the Web system for further improving the retrieval performance, but this is out of the scope of this paper.

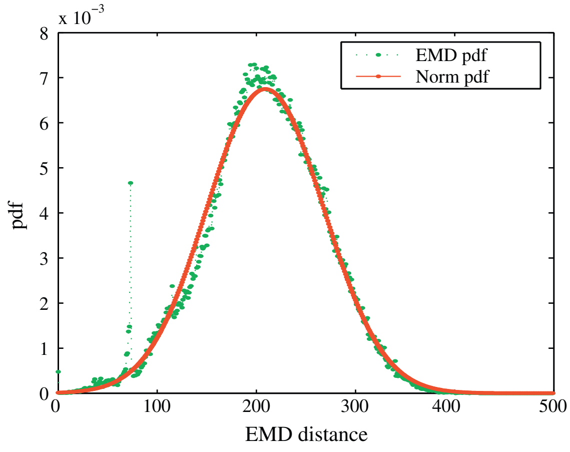

Fig. 1. The EMD distance distribution of region color feature.

Table 4
Chinese query keywords for experiment.

Chinese query keywords	汽车，长城，狮子，老虎，海滩，蝴蝶， 小鸟，雪，天空，日落，紫罗兰，大象， 湖泊，婴儿，熊， 小船，猫，云彩， 狗，鱼，青蛙，狐狸，马，房屋，山脉， 河流，树林，玫瑰花，故宫，黄河
Interpretations in english	Car, Great Wall, lion, tiger, seabeach, butterfly, bird, snow, sky, sunset, tulip, elephant, lake, baby, bear, boat, cat, cloud, dog, fish, frog, fox, horse, house, mountain, river, tree, rose, the Imperial Palace, Yellow River

5.2. Top-K retrieval precision

Fig. 2 shows the average Precision/Recall graphs for the 30 query keywords in Table 4. It indicates that the average precision of the semantic rules method is better than the other three methods, which suggests the effectiveness and superiority of our method. Moreover, the precision of the keyword only method is the worst in these four methods, which confirms that the single-modal retrieval is inferior to the multi-modal retrieval. Furthermore, the online clustering method is better than the Pseudo-RF method, which is consistent with the conclusion in [27].

In fact, our method, the Pseudo-RF method and the online clustering method, are all based on the initial text-based retrieval results. Therefore, the quality of the initial text-based retrieval results significantly affects the effectiveness of them. Considering the quality of the top-*K* images in the initial text-based retrieval results, we further divide the queries into three typical cases for comparison in detail. We select three query keywords to represent the three typical cases, respectively. The query of "car" represents the case of many relevant images in the initial top-*K* retrieved images (i.e., the good case) while the query of "lion" represents the case of only a few relevant images (i.e., the bad case). The query of "Great Wall" represents the case in the middle (i.e., the middle case), which is the most often cases in the queries.

Fig. 3 shows the Precision/Recall graph for the query keyword "car". In this case, the Pseudo-RF, the online clustering method and our method all obtain a good retrieval precision, and the semantic rules method is the best. The semantic rules method obtains strong MMSARs in this case, which makes it perform excellently. The high precision of the Pseudo-RF method and the online clustering method in Fig. 3 lies in that their assumptions are satisfied in this case. However, the Pseudo-RF method and the online clustering method are still inferior to the semantic rules method.

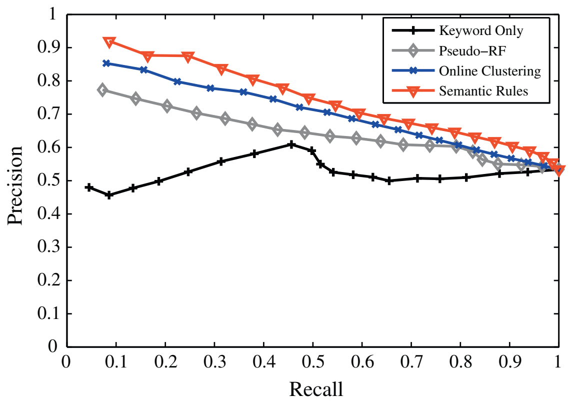

Fig. 2. Average Precision/Recall graph for the 30 query keywords.

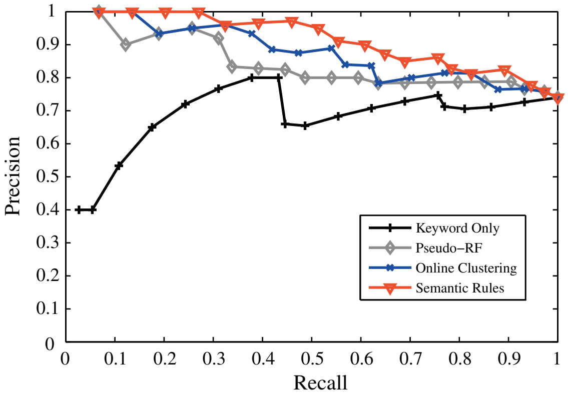

Fig. 3. Precision/Recall graphs for query “car”.

Fig. 4 shows the Precision/Recall graph for the query keyword “Great Wall”. Same as in Fig. 3, the online clustering method, the Pseudo-RF method and our method have better retrieval Precision than the keyword only method. Furthermore, our method is remarkably better than the Pseudo-RF method and the online clustering method, and the online clustering method is a little better than the Pseudo-RF method. Obviously, the semantic rules method indicates significant superiority to the other methods in this case.

Fig. 5 shows the Precision/Recall graph for the query keyword “lion”. Compared with the keyword only method and the Pseudo-RF method, the semantic rules method and the online clustering method obtain better effectiveness remarkably in this case, which demonstrates the effectiveness of the clustering for the initial retrieved results. Moreover, the Pseudo-RF method does not show the superiority to the keyword only method because only a few of the initial retrieved top-*N* images are relevant to the query. In addition, our method is a little better than the online clustering method only in the low-recall stage, which also demonstrates the effectiveness of our method because it puts the relevant images to the front.

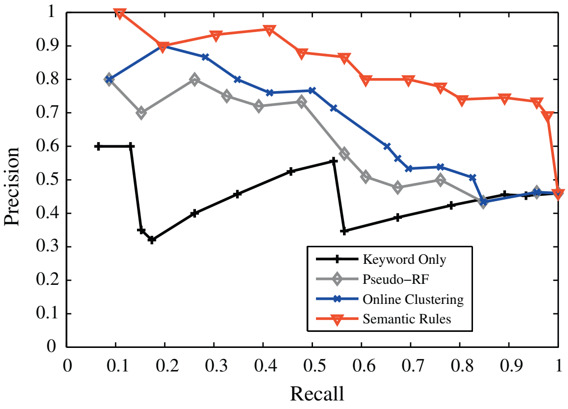

Fig. 4. Precision/Recall graphs for query “Great Wall”.

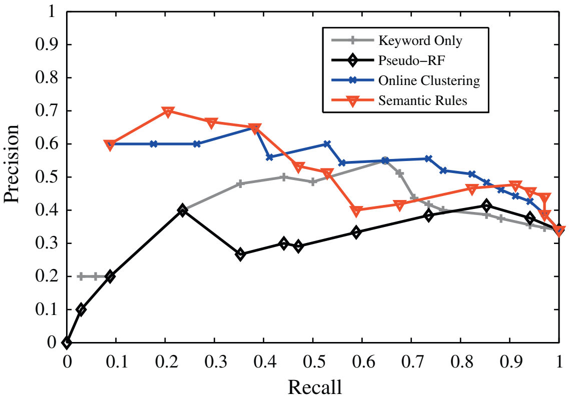

Fig. 5. Precision/Recall graphs for query “lion”.

It is seen from the above figures that the semantic rules method consistently performs better than the other three methods. It improves the retrieval effectiveness either significantly in the good and the middle cases or a little in the bad case, compared with the other three methods. The semantic rule method obviously outperforms the keyword only method due to the fusion of the keywords and the visual features. The online clustering method is the closest to the semantic rules method because they are all based on the visual feature clustering. However, the semantic rules method clusters on different visual features separately based on the whole inverted index for keyword, and automatically select the most appropriate visual feature clusters to represent the Web images in visual feature space.

The query “Great Wall” is an example for the middle case, which represents the most cases in the real-life queries. For the query “Great Wall”, we compare the top-40 retrieval images for the four method, which is shown in Table 5. There are 21 images relevant to “Great Wall” in the retrieved top-40 images for the keyword only method, 22 images for the Pseudo-RF method, 25 images for the online clustering method and 32 images for the semantic rules method. Obviously, the semantic rules method achieves the best results.

To sum up, our method (i.e., the semantic rules method), is generally better than the other three methods, which proves the effectiveness of our method. Especially in the good case and the middle case, the superiority of our method is remarkable. In contrast, the keyword only method is usually the worst, which further confirms the necessity of the combination of the two modalities of Web images.

The reason that our method outperforms the others on retrieval precision maybe lies in: (1) our method clusters the images over each feature type separately; (2) the clustering of our method is based on the whole inverted index for keyword, whereas the online clustering method only considers the top-*K* images; (3) based on the semantic rules, our method automatically selects the clusters over the most appropriate visual feature types to associate with each keyword (i.e., different keywords are associated with different feature types), whereas the online clustering method and the Pseudo-RF method treat with all query keywords equally (i.e., performing on the same feature types for each keyword); (4) our method inherits the merits of data mining technology, which keeps the interesting association rules and filters out some rules with low confidence.

5.3. Response time

All the methods used to fuse the keywords and the visual features unavoidably increase the response time of the query due to its additional processing for visual features. To evaluate the re-

Table 5
The number of the relevant images for “Great Wall” in the top-40 results.

The method	The number of relevant images
The keyword only method	21
The Pseudo-RF method	22
The online clustering method	25
The semantic rules method	32

Table 6
The comparison of the increased response time.

Method vs. scope	Semantic rules (ms)	Pseudo-RF (ms)	Online clustering (ms)
Top-100	86.2	20.2	189.3
Top-200	112.3	34.6	316.2
Top-300	134.1	47.5	456.8
Top-400	173.6	72.8	671.1
Top-500	278.7	143.2	1293.7

sponse time for different methods, we performed experiments on Red Hat Advanced Server 2.6.9-34.EL, Itanium 2 with 1000 MHz and 2G RAM.

Table 6 shows the comparison of the increased response time against top-*K* retrieved scope for the three methods, namely the semantic rules method, the Pseudo-RF method and the online clustering method. The increase response time is equal to the response time of the corresponding method minus the response time of keyword only method. The keyword only method is not considered because it does not use the visual feature.

It can be seen from Table 6 that the semantic rules method is faster than the online clustering method while slower than the Pseudo-RF method. Furthermore, for the different scope of the top-*N* initial retrieved images, the semantic rules method and the Pseudo-RF rise relatively slow while the online clustering method increases fast. In fact, the online clustering method consumes much time on the clustering process, which leads to the longer response time. The semantic rules method makes the clustering process and the association process done offline, which does not add much response time to the system.

In addition, the comparison in Table 6 is only for the case of one user. When a large number of Web users access simultaneously, the online clustering method will dramatically reduce the system performance because it needs additional memory cost and CPU cost. In contrast, the semantic rules method does not have this problem because the clustering is done offline beforehand.

In all, our method is superior to the online clustering method on both retrieval precision and response time, and significantly outperforms the Pseudo-RF method on retrieval precision.

6. Conclusions and future work

From information fusion prospective, we use association rule mining technology in our Web image retrieval system (i.e., *VAST* [24]) to discover the Multi-Modal Semantic Rule between keywords and image visual feature clusters, which includes and crosses the two modalities of Web image. Based on the Multi-Modal Association Rules, we fuse keywords and visual features automatically for Web image retrieval, which is the trend of image search engine. The characteristics and the advantages of the proposed approach can be highlighted as follows:

(1) The mining target is the particular Multi-Modal Association Rule, i.e., the relationship between keywords and image visual clusters (which belongs to two different modalities), not the relationship between image regions or relationship between different textual information in Web page (which belongs to a single modality). The particularity of the mining target make the mining process is particular and customized.
(2) The semantic association rules can be easily merged into the current text-based image search engines because they are all based on the existing inverted file.
(3) In retrieval process, based on the semantic rules, the selection of visual features is automatic and dynamic.
(4) The proposed approach based on the semantic rules fuses the two modalities of Web images automatically, and improves the retrieval precision remarkably, which demonstrates the benefit of multi-source information fusion technology.
(5) Association rule mining is done offline, which does not affect the response time for online retrieval. Therefore, the proposed approach has fast response time online.

Furthermore, we would try other more effective clustering and visual feature extraction algorithms for the preparation of data mining, and further investigate the association rules between text semantic clusters and image visual feature clusters and use them at Query-By-Example mode, which will make the Query-By-Example also utilize both the visual and the textual information for better retrieval performance. Based on these fusion ideas, we expect to explore the future generation image search engine, which will give people more reliable and comfortable services.

Acknowledgements

This work is supported by Hubei Provincial Department of Education (No. Q20091704) and Natural Science Foundation of Hubei Province (No. 2009CDB338).

References

[1] B.V. Dasarathy, A special issue on image fusion: advances in the state of the art, Information Fusion 8 (2) (2007) 114–118.
[2] N. Afzel, S. Richa, V. Mayank, Robust memory-efficient data level information fusion of multi-modal biometric images, Information Fusion 8 (4) (2007) 337–346.
[3] M.L. Kherfi, D. Ziou, A. Bernardi, Image retrieval from the world wide web: issues, techniques, and systems, ACM Computing Surveys 36 (1) (2004) 35–67.
[4] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, The QBIC project: querying images by content using color, texture and shape, in: Proceedings of SPIE Storage and Retrieval for Image and Video Databases, San Jose, CA, 1994, pp. 203–207.
[5] A. Pentland, R.W. Picard, S. Sclaroff, Photobook: content-based manipulation of image databases, International Journal of Computer Vision 18 (3) (1996) 233–254.
[6] J.R. Smith, S.F. Chang, VisualSEEk: a fully automated content-based image query system, in: Proceedings of the Forth ACM International Conference on Multimedia (ACM MM'96), Boston, MA, 1996, pp. 87–98.
[7] J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (9) (2001) 947–963.
[8] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation–maximization and its application to image querying, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (8) (2002) 1026–1038.
[9] K. Zettsu, Y. Kidawara, K. Tanaka, Retrieving web images based on their usage context for augmenting ubiquitous contents, in: Proceedings of 2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM 2003), Victoria, BC, Canada, 2003, pp. 923–926.
[10] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.
[11] T. Gevers, A.W.M. Smeulders, Pictoseek: combining color and shape invariant features for image retrieval, IEEE Transactions on Image Processing 9 (1) (2000) 102–119.
[12] M.O. Binderberger, S. Mehirotra, K. Chakrabarti, K. Porkaew, WebMARS: a multimedia search engine, in: Proceedings of SPIE Electronic Imaging 2000, Internet Imaging, San Jose, CA, 2000, pp. 23–28.
[13] C. Frankel, M.J. Swain, V. Athitsos, WebSeer: An Image Search Engine for the World Wide Web, Technical Report TR-96-14, Computer Science Department, University of Chicago, 1996.
[14] M.L. Cascia, S. Sethi, S. Sclaroff, Combining textual and visual cues for content-based image retrieval on the world wide web, in: Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries, Santa Barbara, California, 1998, pp. 24–28.
[15] D.G. Elisa, D.G. Joriz, G. Steffen, G. Pratim, X.J. Xu, M.R. Amir, M. Emily, Z.Q. Bi, B.S. Manjunath, CORTINA: searching a 10 million + images database, in: Proceedings of 33rd International Conference on Very Large Data Bases (VLDB'07), Vienna, Austria, 2007.
[16] M.L. Kherfi, D. Ziou, A. Andbernardi, Atlas WISE: a web-based image retrieval engine, in: Proceedings of the International Conference on Image and Signal Processing (ICISP 2003), Agadir, Morocco, 2003, pp. 69–77.

[17] J. He, M. Li, Z. Li, H.J. Zhang, H. Tong, C. Zhang, Pseudo relevance feedback based on iterative probabilistic one-class SVMs in web image retrieval, in: Proceedings of the 5th Pacific Rim Conference on Multimedia (PCM 2004), LNCS 3332, Springer, Berlin, 2004.

[18] G. Park, Y. Baek, H.K. Lee, Majority based ranking approach in web image retrieval, in: Proceedings of the International Conference on Image and Video Retrieval (CIVR 2003), Urbana-Champaign, IL, USA, 2003, pp. 111–120.

[19] A. Tombros, R. Villa, V. Rijsbergen, The effectiveness of query specific hierarchic clustering in information retrieval, Information Processing & Management 38 (4) (2002) 559–582.

[20] G. Park, Y. Baek, H.K. Lee, A ranking algorithm using dynamic clustering for content-based image retrieval, in: Proceedings of the International Conference on Image and Video Retrieval: The Challenge of Image and Video Retrieval (CIVR 2002), London, UK, 2002, pp. 316–324.

[21] Z. Chen, W.Y. Liu, F. Zhang, M.J. Li, H.J. Zhang, Web mining for web image retrieval, Journal of the American Society for Information Science and Technology 52 (10) (2001) 831–839.

[22] C. Silverstein, M. Henzinger, H. Marais, M. Moricz, Analysis of a very large web search engine query log, SIGIR Forum 33 (1) (1999) 6–12.

[23] M.W. Berry, P. Wang, Y. Yang, Mining longitudinal web queries: trends and patterns, Journal of the American Society for Information Science and Technology 54 (8) (2003) 743–758.

[24] H. Jin, R.H. He, Z.S. Liao, W.B. Tao, Q. Zhang, A flexible and extensible framework for web image retrieval system, in: Proceedings of the International Conference on Internet and Web Applications and Services (ICIW 2006), Guadeloupe, French Caribbean, 2006, pp. 193–198.

[25] S. Yu, D. Cai, J.R. Wen, W.Y. Ma, Improving pseudo-relevance feedback in web information retrieval using web page segmentation, in: Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest, Hungary, 2003, pp. 11–18.

[26] R. Yan, A. Hauptmann, R. Jin, Multimedia search with pseudo-relevance feedback, in: Proceedings of the International Conference on Image and Video Retrieval (CIVR'03), Urbana, IL, 2003, pp. 238–247.

[27] G. Park, Y. Baek, H.K. Lee, Web image retrieval using majority-based ranking approach, Multimed Tools Application 31 (2) (2006) 195–219.

[28] X. He, O. King, W.Y. Ma, M. Li, H.J. Zhang, Learning a semantic space from user's relevance feedback for image retrieval, IEEE Transactions on Circuits and Systems or Video Technology 13 (1) (2003) 39–48.

[29] M. Koskela, J. Laaksonen, Using long-term learning to improve efficiency of content-based image retrieval, in: Proceedings of the Third International Workshop on Pattern Recognition in Information Systems (PRIS 2003), Angers, France, 2003, pp. 72–79.

[30] J. Feng, M.J. Li, H.J. Zhang, B. Zhang, A unified framework for image retrieval using keyword and visual features, IEEE Transactions on Image Processing 14 (7) (2005) 979–989.

[31] S.C.H. Hoi, M.R. Lyu, R. Jin, Integrating user feedback log into relevance feedback by coupled SVM for content-based image retrieval, in: Proceedings of the 21st International Conference on Data Engineering (ICDE '05), Tokyo, Japan, 2005, pp. 1177–1177.

[32] S.C.H. Hoi, M.R. Lyu, R. Jin, A unified log-based relevance feedback scheme for image retrieval, IEEE Transaction on Knowledge and Data Engineering 18 (4) (2006) 509–524.

[33] R. Zhao, W.I. Grosky, Narrowing the semantic gap – improved text-based web document retrieval using visual features, IEEE Transactions on Multimedia 4 (1) (2002) 189–200.

[34] T. Westerveld, Image retrieval: content versus context, in: Proceedings of the International Conference on Content-based Multimedia Information Access (RIAO 2000), 2000, pp. 276–284.

[35] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Morgan Kaufmann, 1994, pp. 487–499.

[36] F. Bodon, A fast apriori implementation, in: Proceedings of IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), Melbourne, Florida, USA, 2003.

[37] Y. Rubner, C. Tomasi, L. Guibas, A metric for distributions with applications to image databases, in: Proceedings of IEEE International Conference on Computer Vision, Bombay, India, 1998, pp. 59–66.

[38] G. Salton, Automatic Text Processing, Addison-Wesley, 1989.

[39] K. Haris, S. Efstratiadis S, N. Maglaveras, A. Katsaggelos, Hybrid image segmentation using watersheds and fast region merging, IEEE Transactions on Image Processing 7 (12) (1998) 1684–1699.

[40] J. Montgomery, L. Si, J. Callan, D.A. Evans, Effect of varying number of documents in blind feedback, in: Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR'04), Sheffield, South Yorkshire, UK, 2004, pp. 476–477.

[41] ISO/IEC 15938-5 FDIS Information Technology. Mpeg-7: Multimedia content description. %3chttp://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm%3e.

[42] S.J. Phillips, Acceleration of *k*-means and related clustering algorithms, in: Proceedings of ALENEX-02 on the 4th International Workshop on Algorithm Engineering and Experiments, San Francisco, US, 2002, pp. 166–177.

[43] T. Chang, C.-C.J. Kuo, Texture analysis and classification with tree-structured wavelet transform, IEEE Transactions on Image Processing 2 (4) (1993) 429–441.