Learning Image-Text Associations

Tao Jiang and Ah-Hwee Tan, *Senior Member, IEEE*

Abstract—Web information fusion can be defined as the problem of collating and tracking information related to specific topics on the World Wide Web. Whereas most existing work on web information fusion has focused on text-based multi-document summarization, this paper concerns the topic of image and text association, a cornerstone of cross-media web information fusion. Specifically, we present two learning methods for discovering the underlying associations between images and texts based on small training data sets. The first method based on vague transformation measures the information similarity between the visual features and the textual features through a set of predefined domain-specific information categories. Another method uses a neural network to learn direct mapping between the visual and textual features by automatically and incrementally summarizing the associated features into a set of information templates. Despite their distinct approaches, our experimental results on a terrorist domain document set show that both methods are capable of learning associations between images and texts from a small training data set.

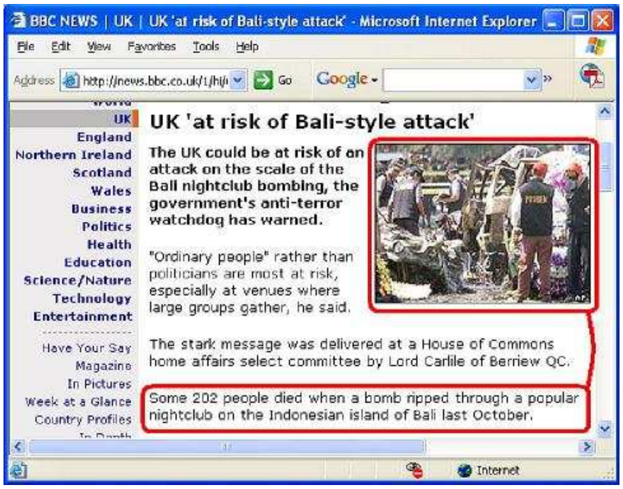
Index Terms—Data Mining, Multimedia Data Mining, Image-Text Association Mining.

1 INTRODUCTION

The diverse and distributed nature of the information published on the World Wide Web has made it difficult to collate and track information related to specific topics. Although web search engines have reduced information overloading to a certain extent, the information in the retrieved documents still contains a lot of redundancy. Techniques are needed in web information fusion, involving filtering of irrelevant and redundant information, collating of information according to themes, and generation of coherent presentation. As a commonly used technique for information fusion, document summarization has been discussed in a large body of literatures. Most document summarization methods however focus on summarizing *text* documents [1] [2] [3]. As an increasing amount of non-text content, namely images, video, and sound, is becoming available on the web, collating and summarizing multimedia information has posed a key challenge in the web information fusion.

Existing literatures of hypermedia authoring and cross-document text summarization [4], [1] have suggested that understanding the interrelation between information blocks is essential in information fusion and summarization. In this paper, we focus on the problem of learning to identify relations between multimedia components, in particular, *image and text associations*, for supporting cross-media web information fusion. An image-text association refers to a pair of image and text segment that are semantically related to each other in a web page. A sample image-text association is shown in Figure 1. Identifying such associations enables one to provide a coherent multimedia summarization of the web documents.

Note that learning image-text association is similar to the task of automatic annotation [5] but has important differences. Whereas image annotation concerns annotating images using a set of predefined keywords, image-text association links images to free text segments in natural language. The methods for image annotation are thus not directly applicable to the problem of identifying image-text associations.



UK 'at risk of Bali-style attack'

The UK could be at risk of an attack on the scale of the Bali nightclub bombing, the government's anti-terror watchdog has warned.

"Ordinary people" rather than politicians are most at risk, especially at venues where large groups gather, he said.

The stark message was delivered at a House of Commons home affairs select committee by Lord Carlile of Berriew QC.

Some 202 people died when a bomb ripped through a popular nightclub on the Indonesian island of Bali last October.

Fig. 1. An associated text and image pair.

A key issue of using a machine learning approach to image-text associations is the lack of large training data sets. However, learning from small training data sets poses the new challenge of handling implicit associations. Referring to Figure 2, two associated image-text pairs (I-T pairs) share partial visual (smokes and fires)

- *T. Jiang is now with the ecPresence Technology Pte Ltd, Singapore 609966. He was previously with the School of Computer Engineering, Nanyang Technological University.
E-mail: jian0006@ntu.edu.sg*
- *A.-H. Tan is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798.
E-mail: asahtan@ntu.edu.sg*

Manuscript received May 24, 2007; revised January 17, 2008; accepted June 25, 2008.

Digital Object Indentifier 10.1109/TKDE.2008.150

and textual features ("attack") but also have different visual and textual contents. As the two I-T pairs are actually on similar topics (describing scenes of terror attacks), the distinct parts, such as the visual content ("black smoke" which can be represented using low-level color and texture features) of the image in I-T pair 2 and the term "Blazing" (underlined) in I-T pair 1, could be potentially associated. We call such useful associations, which convey the information patterns in the domain but are not represented by the training data set, *implicit associations*. We can imagine that the smaller the data set is, the more useful association patterns cannot be covered by the data samples and the more implicit associations exist. Therefore, we need an algorithm which is capable of generalizing the data samples in small data set to induce the missing implicit associations.

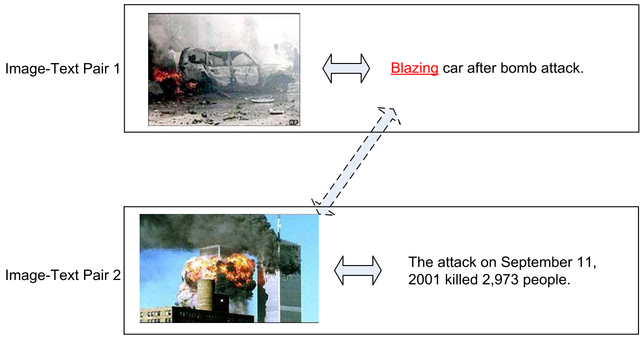


Image-Text Pair 1
Blazing car after bomb attack.
Image-Text Pair 2
The attack on September 11, 2001 killed 2,973 people.

Fig. 2. An illustration of implicit associations between visual and textual features.

In this paper, we present two methods, following the multilingual retrieval paradigm [6] [7] for learning image-text associations. The first method is a textual-visual similarity model [8] with the use of a statistical vague transformation technique for extracting associations between images and texts. As vague transformation typically requires large training data sets and tends to be computationally intensive, we employ a set of domain-specific information categories for *indirectly* matching the textual and visual information at the semantic level. With a small number of domain information categories, the training data sets for vague transformation need not be large and the computation cost can be minimized. In addition, as each information category summarizes a set of data samples, implicit image-text associations can be captured (see Section 3.3 for more details). As information categories may be inconsistently embedded in the visual and textual information spaces, we further employ a *visual space projection* method to transform the visual feature space into a new space, in which the similarities between the information categories are comparable to those in the textual information space. Our experiments show that employing visual space projection can further improve the performance of the vague transformation.

Considering that indirectly matching the visual and textual information using an intermediate tier of information categories may result in a loss of information, we develop another method based on an associative neural network model called fusion ART [9], a direct generalization of Adaptive Resonance Theory (ART) model [10] from one feature field to multiple pattern channels. Even with relatively small data sets where *implicit associations* tend to appear, fusion ART is able to automatically learn a set of prototypical image-text pairs and therefore can achieve a good performance. This is consistent with the prior findings that ART models can efficiently learn useful patterns from small training data sets for text categorization [11]. In addition, fusion ART can directly learn the associations between the features in the visual and textual channels without using a fixed set of information categories. Therefore, the information loss might be minimal.

The two proposed models have been evaluated on a multimedia document set in the terrorist domain collected from the BBC and CNN news web sites. The experimental results show that both vague transformation and fusion ART outperform a baseline method based on an existing state-of-the-art image annotation method known as Cross-Media Relevance Model (CMRM) [12] in learning image-text associations from a small training data set. We have also combined the proposed methods with a pure text-matching based method matching image captions with text segments. We find that though the text based method is fairly reliable, the proposed cross-media methods consistently enhance the overall performance in image-text associations.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 describes our approaches of using vague transformation and fusion ART for learning image-text associations. Section 4 reports our experiments based on the terrorist domain data set. Concluding remarks are given in the final section.

2 Related Work

2.1 Automatic Hypermedia Authoring for Multimedia Content Summarization

Automatic hypermedia authoring aims to automatically organize related multimedia components and generate coherent hypermedia representations [13] [14]. However, most of automatic hypermedia authoring systems assume that multimedia objects used for authoring are well-annotated. Usually, the annotation tasks are done manually with the assistance of annotation tools. Manual annotation can be very tedious and time consuming. Identifying links (relations) within information is an important subtask for automatic hypermedia authoring. In [13], authors defined a set of rules for determining the relations between multimedia objects. For example, a rule can be like "if there is an image A whose subject is equivalent to the title of a text document B, the relation between the image and the text document is *depict(A, B)*". However, this method requires that the multimedia objects (the image A and the text document B in the example) have well-annotated metadata based on which the rules can be defined and applied. As most of web content are raw multimedia data without any well-structured annotation, existing techniques for automatic hypermedia authoring cannot be directly applied to fusion or summarization of the multimedia contents on the web.

For annotating images with semantic concepts or keywords, various techniques have been proposed. These techniques can mainly be classified into three categories, including image classification, image association rules, and statistical based image annotations, reviewed in the following sections.

2.2 Image Classification for Image Annotation

Classification is a data mining technique used to predict group membership for data instances [15]. For example, classification can be used to predict whether a person is infected by dengue disease. In the multimedia domain, classification has been used for annotation purposes, i.e. predicting whether certain semantic concepts appear in media objects.

An early work in this area is to classify the indoor-outdoor scenarios of the video frames [16]. In this work, a video frame or image is modelled as sequences of image segments, each of which is represented by a set of color histograms. A group of one-dimension Hidden Markov Models (HMM) are first trained to capture the patterns of image segment sequences and then used to predict the indoor-outdoor categories of new images.

Recently, many efforts aim to classify and annotate images with more concrete concepts. In [17], a decision tree is used to learn the classification rules that associate color features, including global color histograms and local dominant colors, with semantic concepts such as sunset, marine, arid images, and nocturne. In [18], a learning vector quantization (LVQ) based neural network is used to classify images into outdoor-domain concepts, such as sky, road, and forest. Image features are extracted via Haar wavelet transformation. Another approach using vector quantization for image classification was presented in [19]. In this method, images are divided into a number of image blocks. Visual information of the image blocks is represented using HSV colors. For each image category, a concept-specific codebook is extracted based on training images. Each codebook contains a set of codewords, i.e. representative color feature vectors for the concept. New image classification is performed based on finding most similar codewords for its image blocks. The new image will be assigned to the category whose codebook provides the most number of the similar codewords.

At current stage, image classification mainly works for discriminating images into a relevant small set of categories that are visually separable. It is not suitable for linking images with free texts, in which ten-thousands of the different terms exists. On one hand, the concepts represented by those terms, such as "sea" and "sky", may not be easily separable be the visual features. On the other hand, training classifier for each of these terms would need a large amount of training data which is usually unavailable and the training process would also be extremely time-consuming.

2.3 Learning Association Rules between Image Content and Semantic Concepts

Association rule mining (ARM) is originally used for discovering association patterns in transaction databases. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq \mathcal{I}$ (called itemsets or patterns) and $X \cap Y = \emptyset$. In the domain of market-basket analysis, such an association rule indicates that the customers who buy the set of items X are also likely to buy the set of items Y. Mining association rule from multimedia data is usually a straightforward extensions of ARM in transaction databases.

In this area, many efforts are conducted to extract the association between low-level visual features and high-level semantic concepts for image annotation. Ding et al. [20] presented a pixel based approach to deduce associations between pixels' spectral features and semantic concepts. In [20], each pixel is treated as a transaction, whilst the set ranges of the pixel's spectral bands and auxiliary concept labels (e.g. crop yields) are considered as items. Then pixel-level association rules of the form "Band 1 in the range [a, b] and Band 2 in the range [c, d] are likely to imply crop yield E". However, Tesic et al. [21] pointed out that using individual pixel as transaction may cause the lose of the context information of surrounding locations which are usually very useful for determine the image semantics. This motivated them to use image and rectangular image regions as transactions and items. Image regions are first represented using Gabor texture features and then clustered using self-organizing map (SOM) [22] and learning vector quantization (LVQ) to form a visual thesaurus. The thesaurus is used to provide the perceptual labelling of the image regions. Then, the first- and second-order spatial predicate associations among regions are tabulated in spatial event cubes (SECs), based on which higher-order association rules are determined using Apriori algorithm [23]. For example, a third-order itemset is in the form of "If a region with label u_j is a right neighbor of a u_i region, it is likely that there is a u_k region on the right side of u_j". More recently, Teredesai et al. [24] proposed a framework to learn multi-relational visual-textual associations for image annotation. Within this framework, keywords and image visual features (including color saliency maps, orientation and intensity contrast maps) are extracted and stored separately in relational tables in a database. Then a FP-Growth algorithm [25] is used for extracting multi-relational associations between the visual features and keywords from the database tables. The extracted rules, such as "4 Yellow → EARTH, GROUND", can be subsequently used for annotating new images.

In [26], the author proposed a method to use associations of visual features to discriminate high-level semantic concepts. To avoid combinatory explosion during the association extraction, a clustering method is used to organize the large number of color and texture features into a visual dictionary where similar visual features are grouped together. Then each image can be represented using a relevant small set of representative visual feature groups. Then for each specific image category (i.e. semantic concept), a set of associations is extracted as a visual knowledge base featuring the image category. When a new image comes in, it considered related to a image category if it globally verifies the the associations associated with that image category. In this method, associations were only learnt among visual feature groups, not between visual features and semantic concepts or keywords.

Due to the pattern combinatory explosion problem, the performance of learning association rule is highly depended on the number of items (e.g. image features and the number of lexical terms). Although existing methods that learning association rules between image features and high-level semantic concepts are applicable for small set of concepts/keywords, they may encounter problems when mining association rules on images and free texts where a large amount of different terms exist. This may not only cause significant increasing in the learning times but also result in a great number of association rules which may also lower the performance during the process of annotating images as more rules need to be considered and consolidated.

2.4 Image Annotation Using Statistical Model

There have been many prior work on image annotation using statistical modelling approaches [5]. Barnard and Forsyth [27] proposed a method to encode the correlations of words and visual features using a co-occurrence model. The learned model can be used to predict words for images based on the observed visual features. Another related work [12] by Jeon et al presented a cross-media relevance model for annotating images by estimating the conditional probability of observing a term w given the observed visual content of an image. In [28], Duygulu et al. showed that image annotation could be considered as a machine translation problem to find the correspondence between the keywords and the image regions. Experiments conducted using IBM translation models illustrated promising results. In [29], Li and Wang presented an approach that trained hundreds of two-dimensional multiresolution hidden Markov models (2D MHMMs), each of which encoded the statistics of visual features in the images related to a specific concept category. For annotating an image, the concept of which the 2D MHMMs generates that image with the highest probability will be chosen. In [30], Xing et al propose a dual-wing harmonium model for learning the statistical relationships between the semantic categories and images and texts. The dual-wing harmonium is an extension of the original basic harmonium which models a random field represented by the joint probabilities of a set of input variables (image or text features) and a set of hidden variables (semantic categories). As the harmonium model is undirected, the inferencing can be in two directions. Therefore, dual-wing harmonium model can be used for annotating an image by first inferencing the semantic categories based on its visual features and then predicting the text keywords based on the inferred semantic categories.

A major deficiency of the existing machine learning and statistic based automatic multimedia annotation methods is that they usually assign a fixed number of keywords or concepts to a media object. Therefore, there will inevitably be some media objects assigned with unnecessary annotations and some others assigned with insufficient annotations. In addition, image annotation typically uses a relatively small set of domain-specific key terms (class labels, concepts, or categories) for labelling the images. Our task of discovering semantic image-text associations from Web documents however do not assume such a set of selected domain specific key terms. In fact, although our research focuses on image and text contents from the terrorist domain, both domain-specific and general-domain information (e.g. general-domain terms "people" or "bus") are incorporated in our learning paradigm. With the above considerations, it is clear that the existing image annotation methods are not directly applicable to the task of image-text association. Furthermore, model evolution is not well supported by the above methods. Specifically, after the machine learning and statistic models are trained, they are difficult to update. Moreover, the above methods usually treat the semantic concepts in media objects as separate individuals without considering relationships between the concepts for multimedia content representation and annotation.

3 Identifying Image-Text Associations

3.1 A Similarity-Based Model for Discovering Image-Text Associations

The task of identifying image-text associations can be cast into an *information retrieval* (*IR*) problem. Within a web document d containing images and text segments, we treat each image I in d as a query to find a text segment TS that is most *semantically related* to I, i.e. TS is most similar to I according to a predefined similarity measure function among all text segments in d. In many cases, an image caption can be obtained along with an image in a web document. Therefore, we suppose that each image I can be represented as a visual feature vector, denoted as $\mathbf{v^I} = (v_1^I, v_2^I, ..., v_m^I)$, together with a textual feature vector representing the image caption, denoted as $\mathbf{t^I} = (t_1^I, t_2^I, ..., t_n^I)$. For calculating the similarity between an image I and a text segment TS represented by a textual feature vector $\mathbf{t^{TS}} = (t_1^{TS}, t_2^{TS}, ..., t_n^{TS})$, we need to define a similarity measure $sim_d(<\mathbf{v^I}, \mathbf{t^I}>, \mathbf{t^{TS}})$.

To simplify the problem, we assume that, given an image I and a text segment TS, the similarity between $\mathbf{v^I}$ and $\mathbf{t^{TS}}$ and the similarity between $\mathbf{t^I}$ and $\mathbf{t^{TS}}$ are independent. Therefore, we can calculate $sim_d(<\mathbf{v^I}, \mathbf{t^I}>, \mathbf{t^{TS}})$ with the use of a linear mixture model as follows:

$$sim_d(<\mathbf{v^I}, \mathbf{t^I}>, \mathbf{t^{TS}}) = \lambda \cdot sim_d^{tt}(\mathbf{t^I}, \mathbf{t^{TS}}) + (1-\lambda) \cdot sim_d^{vt}(\mathbf{v^I}, \mathbf{t^{TS}}). \quad (1)$$

In the subsequent sections, we first introduce our method used for measuring the textual similarities between image captions and text segments (Section 3.2). Then, two cross-media similarity measures based on vague transformation (Section 3.3) and fusion ART (Section 3.4) are presented.

3.2 Text-Based Similarity Measure

Matching between text-based features is relatively straightforward. We use the cosine distance

$$sim_d^{tt}(\mathbf{t^I}, \mathbf{t^{TS}}) = \frac{\sum_{k=1}^{n} t_k^I \cdot t_k^{TS}}{\parallel t^I \parallel \parallel t^{TS} \parallel} \quad (2)$$

to measure the similarity between the textual features of an image caption and a text segment. The cosine measure is used as it has been proven to be insensitive to the length of text documents.

3.3 Vague Transformation Based Cross-Media Similarity Measure

Measuring the similarity between visual and textual features is similar to the task of measuring relevance of documents in the field of multilingual retrieval for selecting documents in one language based on queries expressed in another [6]. For multilingual retrieval, transformations are usually needed for bridging the gap between different representation schemes based on different terminologies.

An open problem is that there is usually a basic distinction between the vocabularies of different languages, i.e. word senses may not be organized with words in the same way in different languages. Therefore, an exact mapping from one language to another language may not exist. This is known as the *vague problem* [7], which is even more challenging in visual-textual transformation. For individual image regions, they can hardly convey any meaningful semantics without considering their contexts. For example, a yellow region can be a petal of a flower or can be a part of a flame. On the contrary, words in natural languages usually have a more precise meaning. In most cases, image regions can hardly be directly and precisely mapped to words because of the ambiguity. As vague transformations [31] [7] have been proven useful in the field of multilingual retrieval, in this paper, we borrow the idea from statistical vague transformation methods for cross-media similarity measure.

3.3.1 Statistical Vague Transformation in Multilingual Retrieval

A statistical vague transformation method is first presented in In [31] for measuring the similarity of the terms belonging to two languages. In this method, each term t is represented by a feature vector in the training document space, in which each training document represents a feature dimension and the feature value, known as an association factor, is the conditional probability that given the term t, t belongs to this training document, i.e.

$$z(t,d) = \frac{h(t,d)}{f(t)}, \quad (3)$$

where $h(t,d)$ is the number of times that the term t appears in document d; and $f(t)$ is the total number of times that the term t appears.

As each document also exists in the corpus in the second language, a document feature vector representing a term in the first language can be used for calculate the corresponding term vector for the second language. As a result, given a query term in one language for multilingual retrieval, it is not translated into a single word using a dictionary, but rather transformed into a weighted vector of terms in the second language representing the query term in the document collection.

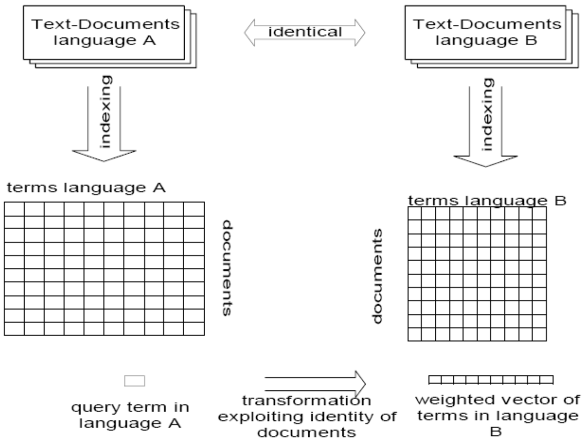

Text-Documents language A
identical
Text-Documents language B
indexing
indexing
terms language A
documents
terms language B
documents
query term in language A
transformation exploiting identity of documents
weighted vector of terms in language B

Fig. 3. An illustration of bilingual vague transformation adopted from [7].

Suppose that there is a set of identified associated image-text pairs. We can treat such an image-text pair as a document represented in both visual and textual languages. Then, statistical vague transformation can be applied for identifying image-text associations. However, this requires a large set of training data which is usually unavailable. Therefore, we consider incorporate high-level information categories to summarize and generalize useful information.

3.3.2 Single-Direction Vague Transformation for cross-media information retrieval

In our cross-media information retrieval model described in Section 3.1, the images are considered as the queries for retrieving the relevant text segments. Referring to the field of multilingual retrieval, transformation of the queries into the representation schema of the target document collection seems to be more efficient [7]. Therefore, we first investigate a *single-direction vague transformation* of the visual information into the textual information space.

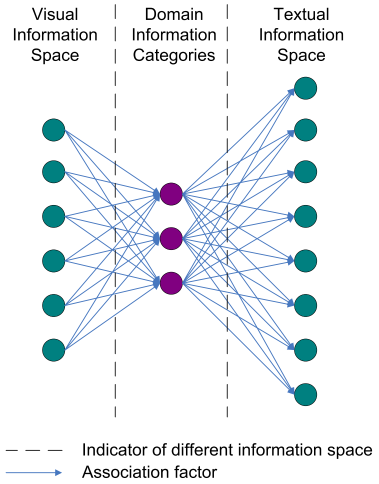

Visual Information Space
Domain Information Categories
Textual Information Space
Indicator of different information space
Association factor

Fig. 4. An illustration of cross-media transformation with information bottleneck.

A drawback of the existing methods for vague transformation, such as those presented in [32] and [31], is that they require a large training set to build multilingual thesauruses. In addition, as the construction of the multilingual thesauruses requires calculating an association factor for each pair of words picked from two languages, it may be computationally formidable. To overcome these limitations, we introduce an intermediate layer for the transformation. This intermediate layer is a set of domain information categories which can be seen as another vocabulary of a smaller size for describing domain information. For example, in the terror attack domain, information categories may include *Attack Details*, *Impacts*, and *Victims* etc. Therefore, our cross-media transformation is in fact a concatenation of two sub-transformations, i.e. from visual feature space to domain information categories and then to textual feature space (see Figure 4). This is actually known as the *information bottleneck* method [33], which has been used for information generalization, clustering and retrieval [34]. For each sub-transformation, as the number of domain information categories is small, the size of the training data set for thesaurus construction needs not be large and the construction cost can be affordable. As discussed in subsection 2.4, for an associated pair of image and text, their contents may not be an exact match. However, we believe that they can always be mapped in terms of general domain information categories.

Based on the above observation, we build two thesauruses in the form of transformation matrices, each of which corresponds to a sub-transformation. Suppose the visterm space $\mathcal{V}$ is of m dimensions, the textual feature space $\mathcal{T}$ is of n dimensions, and the cardinality of the set of high-level domain information categories $\mathcal{C}$ is l. Based on $\mathcal{V}$, $\mathcal{T}$, and $\mathcal{C}$, we define the two following transformation matrices:

$$M^{\mathcal{VC}} = \begin{pmatrix} m_{11}^{\mathcal{VC}} & m_{12}^{\mathcal{VC}} & . & . & m_{1l}^{\mathcal{VC}} \\ m_{21}^{\mathcal{VC}} & m_{22}^{\mathcal{VC}} & . & . & m_{2l}^{\mathcal{VC}} \\ . & . & & & . \\ . & . & & & . \\ m_{m1}^{\mathcal{VC}} & m_{m2}^{\mathcal{VC}} & . & . & m_{ml}^{\mathcal{VC}} \end{pmatrix} \tag{4}$$

and

$$M^{\mathcal{CT}} = \begin{pmatrix} m_{11}^{\mathcal{CT}} & m_{12}^{\mathcal{CT}} & . & . & . & m_{1n}^{\mathcal{CT}} \\ m_{21}^{\mathcal{CT}} & m_{22}^{\mathcal{CT}} & . & . & . & m_{2n}^{\mathcal{CT}} \\ . & . & & & & . \\ m_{l1}^{\mathcal{CT}} & m_{l2}^{\mathcal{CT}} & . & . & . & m_{ln}^{\mathcal{CT}} \end{pmatrix}, \tag{5}$$

where $m_{ij}^{\mathcal{VC}}$ represents the association factor between the visual feature v_i and the information category c_j; and $m_{jk}^{\mathcal{CT}}$ represents the association factor between the information category c_j and the textual feature t_k. In our current system, $m_{ij}^{\mathcal{VC}}$ and $m_{jk}^{\mathcal{CT}}$ are calculated by

$$m_{ij}^{\mathcal{VC}} = P(c_j|v_i) \approx \frac{N(v_i, c_j)}{N(v_i)} \tag{6}$$

and

$$m_{jk}^{\mathcal{CT}} = P(tm_k|c_j) \approx \frac{N(c_j, tm_k)}{N(c_j)}, \tag{7}$$

where $N(v_i)$ is the number of images containing the visual feature v_i; $N(v_i, c_j)$ is the number of images containing v_i and belonging to the information category c_j; $N(c_j)$ is the number of text segments belonging to the category c_j; and $N(c_j, tm_k)$ is the number of text segments belonging to c_j and containing the textual feature (term) tm_k.

For calculating $m_{ij}^{\mathcal{VC}}$ and $m_{jk}^{\mathcal{CT}}$ in Eq. 6 and Eq. 7, we build a training data set of texts and images that have been manually classified into domain information categories (see Section 4 for details).

Based on Eq. 4 and Eq. 5, we can define the similarity between the visual part of an image $\mathbf{v^I}$ and a text segment represented by $\mathbf{t^{TS}}$ as $(\mathbf{v^I})^T M^{\mathcal{VC}} M^{\mathcal{CT}} \mathbf{t^{TS}}$. For embedding into Eq. 1, we use its normalized form

$$sim^{\mathcal{VT}}(\mathbf{v^I}, \mathbf{t^{TS}}) = \frac{(\mathbf{v^I})^T M^{\mathcal{VC}} M^{\mathcal{CT}} \mathbf{t^{TS}}}{\| ((\mathbf{v^I})^T M^{\mathcal{VC}} M^{\mathcal{CT}})^T \| \| \mathbf{t^{TS}} \|}. \tag{8}$$

3.3.3 Dual-Direction Vague Transformation

Eq. 8 calculates the cross-media similarity using a single-direction transformation from visual feature space to textual feature space. However, it may still have the vague problem. For example, suppose there is a picture I , represented by the visual feature vector $\mathbf{v^I}$, belonging to a domain information category *Attack Details*, and two text segments TS_1 and TS_2 , represented by the textual feature vectors $\mathbf{t^{TS_1}}$ and $\mathbf{t^{TS_2}}$, belonging to the categories of *Attack Details* and *Victims* respectively. If the two categories *Attack Details* and *Victims* share many common words (such as *kill*, *die*, and *injure*), the vague transformation result of $\mathbf{v^I}$ might be similar to both $\mathbf{t^{TS_1}}$ and $\mathbf{t^{TS_2}}$. To reduce the influence of common terms on different categories and utilize the strength of the distinct words, we consider another transformation from the word space to the visterm space. Similarly, we define a pair of transformation matrices $M^{\mathcal{TC}} = \{m_{kj}^{\mathcal{TC}}\}^{n \times l}$ and $M^{\mathcal{CV}} = \{m_{ji}^{\mathcal{CV}}\}^{l \times m}$, where $m_{kj}^{\mathcal{TC}} = P(c_j|tm_k) \approx \frac{N(c_j, tm_k)}{N(tm_k)}$ and $m_{ji}^{\mathcal{CV}} = P(v_i|c_j) \approx \frac{N(v_i, c_j)}{N(c_j)}$ $(i = 1, 2, ..., m,\ j = 1, 2, ..., l$, and $k = 1, 2, ..., n)$. Here, $N(tm_k)$ is the number of text segments containing the term tm_k; $N(c_j, tm_k)$, $N(v_i, c_j)$, and $N(c_j)$ are same as those in Eq. 6 and 7. Then, the similarity between a text segment represented by the textual feature vector $\mathbf{t^{TS}}$ and the visual content of an image $\mathbf{v^I}$ can be defined as

$$sim^{\mathcal{TV}}(\mathbf{t^{TS}}, \mathbf{v^I}) = \frac{(\mathbf{t^{TS}})^T M^{\mathcal{TC}} M^{\mathcal{CV}} \mathbf{v^I}}{\| ((\mathbf{t^{TS}})^T M^{\mathcal{TC}} M^{\mathcal{CV}})^T \| \| \mathbf{v^I} \|}. \quad (9)$$

Finally, we can define a cross-media similarity measure based on the *dual-direction transformation* which is the arithmetic mean of $sim^{\mathcal{VT}}(\mathbf{v^I}, \mathbf{t^{TS}})$ and $sim^{\mathcal{TV}}(\mathbf{t^{TS}}, \mathbf{v^I})$ given by

$$sim_d^{vt}(\mathbf{v^I}, \mathbf{t^{TS}}) = \frac{sim^{\mathcal{VT}}(\mathbf{v^I}, \mathbf{t^{TS}}) + sim^{\mathcal{TV}}(\mathbf{t^{TS}}, \mathbf{v^I})}{2}. \quad (10)$$

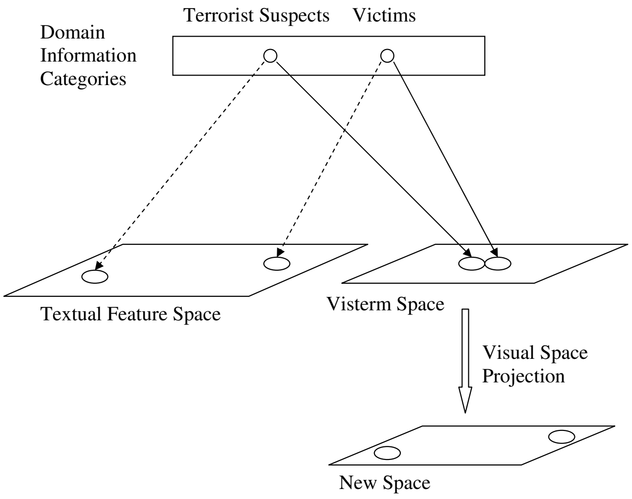


Fig. 5. Visual space projection.

3.3.4 Vague Transformation with Visual Space Projection

A problem in the reversed cross-media (text-to-visual) transformation in dual-direction transformation is that the intermediate layer, i.e. information categories, may be *embedded* differently in the textual feature space and the visterm space. For example, in Figure 5, two information categories, "Terrorist Suspects" and "Victims", may contain quite different text descriptions but somewhat similar images, e.g. human faces. Suppose we translate a term vector of a text segment into the visual feature space using a cross-media transformation. Transforming a term vector in "Victims" category or a term vector in "Terrorist Suspects" category may result in a similar visual feature vector as these two information categories have similar representation in the visual feature space. In such a case when there are text segments belonging to the two categories in the same web page, we may not be able to select a proper text segment for an image about "Terrorist Suspects" or "Victims" based on the text-to-visual vague transformation.

For solving this problem, we need to consolidate the differences in the similarities between the information categories in the textual feature space and the visual feature space. We assume that text can more precisely represent the semantic similarities of the information categories. Therefore, we project the visual feature space into a new space in which the information category similarities defined in the textual feature space can be preserved.

We treat each row in the transformation matrix $M^{\mathcal{CV}}$ in Eq. 9 as a visual feature representation of an information category. We use a $m \times m$ matrix X to transform the visterm space into a new space, wherein the similarity matrix of information categories can be represented as $(M^{\mathcal{CV}} X^T)(X M^{\mathcal{CV}^T}) = \{s_v(c_i, c_j)\}^{l \times l}$, where $s_v(c_i, c_j)$ represents the similarity between the information categories c_i and c_j. In addition, we suppose $D = \{s_t(c_i, c_j)\}^{l \times l}$ is the similarity matrix of the information categories in the textual feature space, where $s_t(c_i, c_j)$ is the similarity between the information categories c_i and c_j. Our objective is to minimize the differences between the information category similarities in the new space and the textual feature space. This can be formulated as an optimization problem of

$$min_X \| D - M^{\mathcal{CV}} X^T X M^{\mathcal{CV}^T} \|^2 . \quad (11)$$

The similarity matrix D in the textual feature space is calculated based on our training data set, in which texts and images are manually classified into the domain specific information categories. Currently, two different methods have been explored for this purpose as follows.

- **Using Bipartite Graph of the Classified Text Segments** For constructing the similarity matrix of the information categories in the textual feature space, we utilize the bipartite graph of the classified text segments and the information categories as shown in Figure 6.

  The underlying idea is that the more text segments that two information categories share, the more similar they are. We borrow the similarity measure used in [31] for calculating the similarity between information categories which is originally used for calculating term similarity based on bipartite graphs of terms and text documents. Therefore, any $s_t(c_i, c_j)$ in D can be calculated as

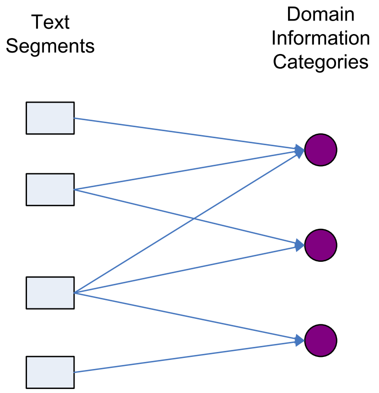


Fig. 6. Bipartite graph of classified text segments and information categories.

$$s_t(c_i,c_j)=\sum_{TS_k\in c_i\cap c_j} wt(c_i,TS_k)\cdot wt(c_j,TS_k), \qquad (12)$$

where

$$wt(c_i,TS_k)=\frac{1/|TS_k|}{\sqrt{\sum_{TS_l\in c_i}(1/|TS_l|)^2}}, \qquad (13)$$

where $|TS_k|$ and $|TS_l|$ represent the sizes of the text segments TS_k and TS_l respectively.

- **Using Category-To-Text Transformation Matrix** We also attempt another approach that utilizes the category-to-text vague transformation matrix $M^{\mathcal{CT}}$ in Eq. 5 for calculating the similarity matrix of the information categories. We treat each row in $M^{\mathcal{CT}}$ as a textual feature space representation of an information category. Then, we calculate the similarity matrix D in the textual feature space by

$$D=M^{\mathcal{CT}}\cdot {M^{\mathcal{CT}}}^T. \qquad (14)$$

With the similarity matrix D of the information categories calculated above, the visual space projection matrix X can be solved based on Eq. 11. By incorporating X, Eq. 9 can be redefined as

$$sim^{\mathcal{TV}}(\mathbf{t^{TS}},\mathbf{v^I})= \qquad (15)$$
$$\frac{(\mathbf{t^{TS}})^T M^{\mathcal{TC}}M^{\mathcal{CV}}X^TX\mathbf{v^I}}{\|\,((\mathbf{t^{TS}})^T M^{\mathcal{TC}}M^{\mathcal{CV}}X^TX)^T\,\|\|\,\mathbf{v^I}\,\|}.$$

Using this refined equation in the dual-direction transformation, we expect that the performance of discovering the image-text associations can be improved. However, solving Eq. 11 is a non-linear optimization problem of a very large scale because X is a $m\times m$ matrix, i.e. there are m^2 variables to tune. Fortunately, from Eq. 15 we can see that we do not need to get the exact matrix X. Instead, we only need to solve a simple linear equation $D=M^{\mathcal{CV}}X^TX{M^{\mathcal{CV}}}^T$ to obtain a matrix

$$A=X^TX={M^{\mathcal{CV}}}^{-1}D{{M^{\mathcal{CV}}}^{-1}}^T, \qquad (16)$$

where ${M^{\mathcal{CV}}}^{-1}$ is the pseudo-inverse of the transformation matrix $M^{\mathcal{CV}}$.

Then, we can substitute X^TX in Eq. 15 with A for calculating $sim^{\mathcal{TV}}(ts,i_v)$, i.e. the similarity between a text segment represented by $\mathbf{t^{TS}}$ and the visual content of an image represented by $\mathbf{v^I}$.

3.4 Fusion ART Based Cross-Media Similarity Measure

In the previous subsection, we present the method for measuring the similarity between visual and textual information using a vague transformation technique. For the vague transformation technique to work on small data sets, we employ an intermediate layer of information categories to map the visual and textual information in an *indirect* manner. A deficiency of this method is that for training the transformation matrix, additional work on manual categorization of images and text segments is required. In addition, matching visual and textual information based on a small number of information categories may cause a loss of detailed information. Therefore, it would be appealing if we can find a method that learns *direct* associations between the visual and textual information. In this subsection, we present a method based on the fusion ART network, a generalization of Adaptive Resonance Theory (ART) model [10], for discovering direct mappings between visual and textual features.

3.4.1 A similarity measure based on Adaptive Resonance Theory

As discussed, small data set does not have enough data samples, and thus many useful association patterns may appear in the data set implicitly. Those implicit associations may not be reflected in individual data samples but can be extracted by summarizing a group of data samples.

For learning implicit associations, we employ a method based on the fusion ART network. Fusion ART can be seen as multiple overlapping Adaptive Resonance Theory (ART) models [35] each of which corresponds to an individual information channels. Figure 7 shows a two-channel fusion ART (also known as Adaptive Resonance Associative Map [36]) for learning associations between images and texts. The model consists of a F_2^c field, and two input pattern fields, namely F_1^{c1} for representing visual information channel of the images and F_1^{c2} for representing textual information channel of text segments. Such a fusion ART network can be seen as a simulation of a *physical resonance phenomenon* where each associated image-text pair can be seen as an information "*object*" that has a "natural frequency" in either visual or textual information channel represented by the visual feature vector $\mathbf{v}=(v_1,v_2,...,v_m)$ or the textual feature vector $\mathbf{t}=(t_1,t_2,...,t_n)$. If two information objects have similar "natural frequencies", strong resonance occurs.

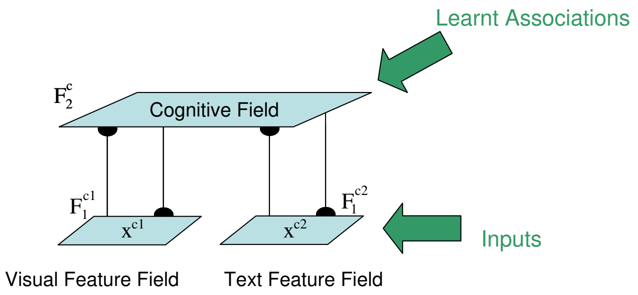


Fig. 7. Fusion ART for learning image-text associations.

The strength of the resonance can be computed by a resonance function.

Given a set of multimedia information objects (associated image-text pairs) for training, the fusion ART learns a set of *multimedia information object templates*, or *object templates* in short. Each object template, recorded by a category node in the F_2^c field, represents a group of information objects that have similar "natural frequencies" and can strongly resonate with each other. Initially, no object template (category node) exists in the F_2^c field. When information objects (associated image-text pairs) are presented one at a time to the F_1^{c1} and F_1^{c2} fields, the object templates are incrementally captured and encoded in the F_2^c field. The process of learning object templates using fusion ART can be summarized in the following stages:

1) Code Activation: A bottom-up propagation process first takes place when an information object is presented to the F_1^{c1} and F_1^{c2} fields. For each category node (multimedia information object template) in the F_2^c field, a *resonance score* is computed using an *ART choice function*. The ART choice function varies with respect to different ART models, including ART 1 [10], ART 2 [37], ART 2-A [38] and fuzzy ART [39]. We adopt the ART 2 choice function based on the cosine similarity which has been proven to be effective for measuring vector similarities and insensitive to the vector lengths. Given a pair of visual and textual information feature vectors $\mathbf{v}$ and $\mathbf{t}$, for each F_2^c category node j with a visual information template $\mathbf{v^{c_j}}$ and a textual information template $\mathbf{t^{c_j}}$, the resonance score T_j is calculated by:

$$T_j = \gamma \frac{\mathbf{v} \cdot \mathbf{v^{c_j}}}{\parallel \mathbf{v} \parallel \parallel \mathbf{v^{c_j}} \parallel} + (1-\gamma) \frac{\mathbf{t} \cdot \mathbf{t^{c_j}}}{\parallel \mathbf{t} \parallel \parallel \mathbf{t^{c_j}} \parallel}, \qquad (17)$$

where γ is the factor for weighing the visual and textual information channels. For giving the equal weight to the visual and textual information channels, we set the γ value to 0.5. $\frac{\mathbf{v} \cdot \mathbf{v^{c_j}}}{\|\mathbf{v}\|\|\mathbf{v^{c_j}}\|}$ and $\frac{\mathbf{t} \cdot \mathbf{t^{c_j}}}{\|\mathbf{t}\|\|\mathbf{t^{c_j}}\|}$ are actually the ART 2 choice function for visual and textual information channels.

2) Code Competition: A code competition process follows under which the F_2^c node j with the highest resonance score is identified.

3) Template Matching: Before the node j can be used for learning, a template matching process checks that for each (visual and text) information channel, the object template of node j is sufficiently similar to the input object with respect to the norm of the input object. The similarity value is computed by using the ART 2 match function [37]. The ART 2 match function is defined as

$$T_j = \gamma \frac{\mathbf{v} \cdot \mathbf{v^{c_j}}}{\parallel \mathbf{v} \parallel \parallel \mathbf{v} \parallel} + (1-\gamma) \frac{\mathbf{t} \cdot \mathbf{t^{c_j}}}{\parallel \mathbf{t} \parallel \parallel \mathbf{t} \parallel}. \qquad (18)$$

Resonance can occur if for each (visual and textual) channel, the match function value meets a vigilance criterion. At current stage, the vigilance criterion is an experience value manually tested and defined by us.

4) Template Learning: Once a node j is selected, its object template will be updated by linearly combining the object template with the input information object according to a predefined learning rate [37]. The equation for updating the object template is defined as follows:

$$v^{c_j} = (1-\beta_v) \cdot v^{c_j} + \beta_v \cdot v \qquad (19)$$

and

$$t^{c_j} = (1-\beta_t) \cdot t^{c_j} + \beta_t \cdot t, \qquad (20)$$

where β_v and β_t are learning rates for visual and textual information channels.

5) New Code Creation: When no category node is sufficiently similar to the new input information object, a new category node is added to the F_2^c field. Fusion ART thus expands its network architecture dynamically in response to incoming information objects.

The advantage of using fusion ART is that its object templates are learnt by incrementally combining and merging new information objects with previously learnt object templates. Therefore, a learnt object template is a "summarization" of characteristics of the training objects. For example, the three training I-T pairs as shown in Figure 8 can be summarized by fusion ART into one object template and thereby the implicit associations across the I-T pairs can be captured. In particular, the implicit associations between frequently occurred visual and textual contents (such as the visual content "black smoke" which can be represented using low-level visual features and the term "blazing" in Figure 8) can be learnt for predicting new image-text associations.

Suppose a trained fusion ART can capture all the typical multimedia information object templates. Given a new pair of image and text segment that are semantically relevant, their information object should be able to strongly resonate with an object template in the trained fusion ART. In other words, we can measure the similarity or relevance between an image and a text segment according to their resonance score (see Eq. 18) in a trained fusion ART. Such resonance based similarity

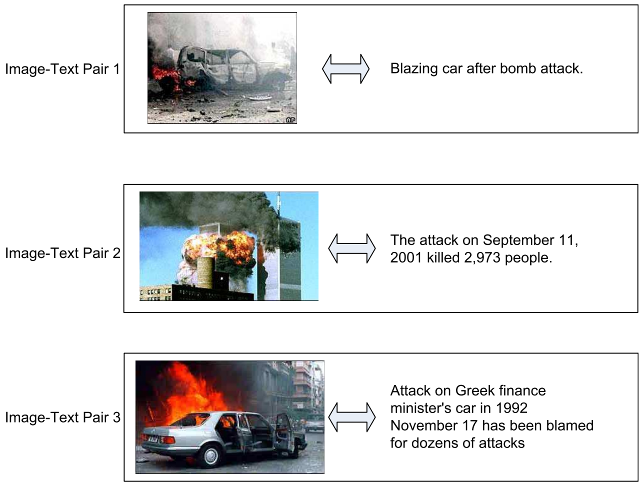


Image-Text Pair 1
Blazing car after bomb attack.
Image-Text Pair 2
The attack on September 11, 2001 killed 2,973 people.
Image-Text Pair 3
Attack on Greek finance minister's car in 1992 November 17 has been blamed for dozens of attacks

Fig. 8. Training samples for fusion ART.

measure has been used in existing work for data terrain analysis [40].

3.4.2 Image Annotation Using fusion ART

In the above discussion, we define a cross-media similarity measure based on the resonance function of the fusion ART. Based on this similarity measure, we can identify an associated text segment represented by textual feature vector $\mathbf{t}$ that is considered most "similar" to an image represented by the visual feature vector $\mathbf{v}$. At the same time, we will identify a F_2^c object template j with a textual information template $\mathbf{t^{c_j}}$ having the strongest resonance with the image-text pair represented by <$\mathbf{v}$, $\mathbf{t}$>. Based on the $\mathbf{t}$ and $\mathbf{t^{c_j}}$, we can extract a set of keywords for annotating the image. The process is described as follows:

1) For the kth dimension of the textual information vectors $\mathbf{t}$ and $\mathbf{t^{c_j}}$, if $min\{t_k, t_k^{c_j}\} > 0$, extract the term tm_k corresponding to the kth dimension of the textual information vectors for annotating the image.
2) When tm_k is extracted, a confidence value of $min\{t_k, t_k^{c_j}\}$ is assigned to tm_k based on which all extracted keywords can be ranked for annotating the image.

We can see that the image annotation task is possible because the fusion ART can capture the direct associations between the visual and textual information. The performance evaluation of using fusion ART for image annotation is beyond the scope of this paper. However, the fusion ART based image annotation method has a unique advantage over the existing image annotation methods, namely it does not require a predefined set of keywords. A set of examples of using the fusion ART based method for image annotation will be presented in the next section.

3.4.3 Handling Noisy Text

Web text data may contain various noises such as wrongly spelled words. Existing research shows that the web text noise may lower the performance of the data mining tasks on web text data [41], [42]. However, the ART based systems have shown advantages on handling noisy web text and have been successfully used in many applications for web/text mining tasks [43], [44], [41], [45], [46]. In normal practices, the ART based systems may handle noisy data through two ways:

- Firstly, noisy reduction can be performed in the data preprocessing stage. For example, in the web/text classification systems, feature reduction is usually used for eliminating the features that do not have strong association with the class labels [11]. However, for image-text associations, the information object templates are automatically learnt by the fusion ART and thus no class labels exist. Therefore, in our data preprocessing stage, we use another commonly used method that employs the notion of term frequency to prune the terms with very high (above 80%) or very low frequencies (below 1%). Such a feature pruning method has been effective in removing irrelevant features (including noises) in web/text mining tasks.
- In addition, fusion ART can control the influence of the noises through varying the learning rates. Refering to Eq. 19 and 20, when using small learning rates, noisy data with low frequencies will not cause large changes to information object templates. With sufficient patterns, the influence of the noises will be kept low.

4 Experimental Results

4.1 Data Set

The experiments are conducted on a web page collection, containing 300 images related to terrorist attacks, downloaded from the CNN and BBC news web sites. For each image, we extract its caption (if available) and long text paragraphs (with more than 15 words) from the web page containing the image.

For applying vague transformation based methods that utilize an intermediate layer of information categories, we manually categorize about 1500 text segments and 300 images into 15 predefined domain information categories, i.e. *Anti-Terror, Attack Detail, After Attack, Ceremony, Government Emergency Preparation, Government Response, Impact, Investigation, Rescue, Simulation, Attack Target, Terrorist Claim, Terrorist Suspect, Victim*, and *Others*. We note that the semantic contents of images are usually more concentrating and easy to be classified into a information category. However, for texts, a text segment may be a mixture of information belonging to different information categories. In this case, we assign the text segment into multiple information categories. In addition, a PhD student in School of Computer Engineering, Nanyang Technological University manually inspect

TABLE 1
Statistics of the data set.

N_i	N_w	N_c	Avg. N_t	Distribution of N_t	N_{tm}
300	287	297	5	$>$ 5: 39%; 3−5: 53%; 2: 8%	8747

N_i denotes the number of images, N_w denotes the number of web pages where the images and texts are extracted, N_c denotes the number of images with captions, *Avg.* N_t denotes the average number of texts segments appearing along with an image in a web page, N_{tm} denotes the number of text features (terms) extracted for representing web texts.

the web pages where the 300 images are extracted to identify the associated text segments for the images. This forms our ground truth for evaluating the correctness of the image-text pairs extracted by the proposed image-text association learning methods. As there are the text captions associated with the images which can provide a lot of clues, the text segments associated with the images usually can be identified from the web pages without difficulties. Table 1 lists the statistics of the data set and the detailed data preprocessing methods are described as follows.

4.1.1 Textual Feature Extraction

We treat each text paragraph as a text segment. Preprocessing the text segments includes tokenizing the text segments, part-of-speech tagging, stop word filtering, stemming, removing unwanted terms (retaining only nouns, verbs and adjectives), and generating the textual feature vectors where each dimension corresponds to a remaining term after the preprocessing.

For calculating the term weights of the textual feature vectors, we use a model, named TF-ITSF (term frequency and inverted text segment frequency), similar to the traditional TF-IDF model. For a text segment TS in a web document d, we use the following equation to weight the kth term tm_k in TS's textual feature vector $\mathbf{t^{TS}} = (t_1^{TS}, t_2^{TS}, ..., t_n^{TS})$:

$$t_k^{TS} = tf(TS, tm_k) \cdot \log \frac{N(d)}{tsf(d, tm_k)}, \tag{21}$$

where $tf(TS, tm_k)$ denotes the frequency of tm_k in the text segment TS, $N(d)$ is the total number of text segments in the web document d, and $tsf(d, tm_k)$ is the text segment frequency of term tm_k in d. Here, we use the text segment frequency for measuring the importance of a term for a web document.

After a textual feature vector $\mathbf{t^{TS}} = (t_1^{TS}, t_2^{TS}, ..., t_n^{TS})$ is extracted, L1-normalization is applied for normalizing the term weights into a range of $[0, 1]$:

$$\mathbf{t^{TS}} = \frac{(t_1^{TS}, t_2^{TS}, ..., t_n^{TS})}{max\{t_{i=1...n}^{TS}\}}, \tag{22}$$

where n is the number of textual features (i.e. terms).

4.1.2 Visual Feature Extraction from Images

Our visual feature extraction method is inspired by those used in the existing image annotation approaches [5] [27] [12] [28] [47]. During image preprocessing, each image is first segmented into 10×10 rectangular regions arranged in a grid. There are two reasons why we use rectangular regions instead of employing more complex image segmentation techniques to obtain image regions. The first reason is that existing work on the image annotation shows that using rectangular regions can provide performance gain compared with using regions obtained by automatic image segmentation techniques [47]. The second reason is that using rectangular regions is much less time-consuming and therefore suitable for the online multimedia web information fusion task.

For each image region, we extract a visual feature vector, consisting of six color features and 60 gabor texture features. The color features are the means and variances of the RGB color spaces. The texture features are extracted by calculating the means and variations of the Gabor filtered image regions in six orientations at five different scales. Such a set of color and textual features have been proven to be useful for image classification and annotation tasks [48] [49].

After the feature vectors are extracted, all image regions are clustered using the k-means algorithm. The purpose of the clustering is to discretize the continuous color and texture features [27] [28] [47]. The generated clusters, called *visterms* represented by $\{vt_1, vt_2, ..., vt_k\}$, are treated as a vocabulary for describing the visual content of the images. An image is described by a visterm vt_j if it contains a region belonging to the jth cluster. For the terrorist domain data set, the visterm vocabulary is enriched with a high-level semantic feature, extracted by a face detection model provided by the OpenCV. In total, a visterm vector of k+1 features is extracted for each image. The weight of each feature is the corresponding visterm frequency normalized with the use of L1-normalization.

A problem of using visterms is how we can determine a proper number of visterms k (i.e. the number of clusters for k-means clustering). Note that the images have been manually categorized into the fifteen information categories which reflect the images' semantic meanings. Therefore, images belonging to different information categories should have different patterns in their visterm vectors. If we cluster images into different clusters based on their visterm vectors, in the most ideal case, images belonging to different categories should be assigned into different clusters. Based on the above consideration, we can measure the meaningfulness of the visterm sets with different k values by calculating the *information gains* of the image clustering results. The definition of the information gain given below is similar to the one used in the Decision Tree for measuring partitioning results with respect to different data attributes.

Given a set S of images belonging to m information categories, the *information need* for classifying of the images in S is measured by

$$I(S) = -\sum_{i=1}^{m} \frac{s_i}{\| S \|} log(\frac{s_i}{\| S \|}), \tag{23}$$

where $\| S \|$ is the total number of images and s_i is the number of images belonging to the ith information category.

Suppose we cluster an image collection S into n clusters, i.e. S_1, S_2, ..., S_n. The information gain can be calculated as follows:

$$Gain = I(S) - \sum_{j=1}^{n} \frac{\| S_j \|}{\| D \|} I(S_j), \tag{24}$$

where $\| S_j \|$ is the number of images in the jth cluster.

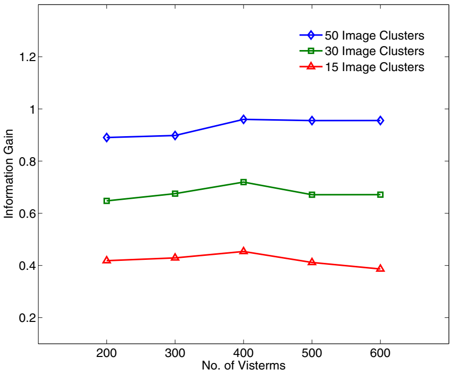


Fig. 9. Information gains of clustering the images based on a varying number of visterms.

Figure 9 shows the information gains obtained by clustering our image collection based on visterm sets with a varying number of visterms. We can see no matter how many clusters of images we generate, the largest information gain is always achieved when k is around 400. Based on this observation, we generate 400 visterms for the image visterm vectors.

Note that we employ information gain depending on information category for determine the number of visterms to use for representing image contents. This is an optimization for the data preprocessing stage, the benefit of which will be employed by all the learning models in our experiments. Therefore, we should see that it is not conflict with our statement that the fusion ART does not depend on the predefined information categories for learning the image-text associations.

4.2 Performance Evaluation Method

We adopt a five-fold cross-validation to test the performance of our methods. In each experiment, we use four folds of the data (240 images) for training and one fold (60 images) for testing. The performance is measured in terms of precision defined by

$$precision = \frac{N_c}{N}, \tag{25}$$

where N_c is the number of correctly identified image-text associations and N is the total number of images. We experimented with different λ values for our cross-media retrieval models (see Eq. 1) to find the best *balance point* of weighting the impact of textual and visual features. However, we should note that in principle, the best λ could be obtained by using an algorithm such as expectation-maximization (EM) [50].

Note that whereas most information retrieval tasks use both precision and recall to evaluate the performance, we only use precision in our experiment. This is simply due to the fact that for each image, there is only one text segment considered to be semantically relevant. In addition, we also extract only one associated text segment for each image using the various models. Therefore, in our experiments, the precision and recall values are actually the same.

4.3 Evaluation of Cross-Media Similarity Measure Based on Visual Features Only

We first evaluate the performance of the cross-media similarity measures, defined in subsection 3.3 and subsection 3.4, by setting $\lambda = 0$ in our linear mixture similarity model (see Eq. 1), i.e. using only visual contents of images (without image captions) for measuring image-text associations. As there has been no prior work on image-text association learning, we implement two baseline methods for evaluation and comparison. The first method is based on the cross-media relevance model (CMRM) proposed by Jeon et al [12]. The CMRM model is designed for image annotation by estimating the conditional probability of observing a term w given the observed visual content of an image. Another baseline method is based on the dual-wing harmonium (DWH) model proposed by Xing et al [30]. As described in Section 2, a trained dual-wing harmonium can also be used to estimate the conditional probability of seeing a term w given the observed image visual features. As our objective is to associate an entire text segment to an image, we extend CMRM and DWH models to calculate the average conditional probability of observing terms in a text segment given the visual content of an image. The reason of using the average conditional probability, instead of the joint conditional probability, is that we need to minimize the influence of the length of the text segments. Note that the longer a text segment is, the smaller the joint conditional probability tends to be. Table 2 summarizes the seven methods that we experimented for discovering image-text associations based on pure visual contents of images. The first four methods are vague transformation based cross-media similarity measures that we define in subsection 3.3. The fifth method is the the fusion ART (object resonance) based similarity measure. The last two methods are baseline

TABLE 2
The seven cross-media models in comparison.

Model	Descriptions
SDT	Single-direction vague transformation
DDT	Dual-direction vague transformation
DDT_VP_BG	DDT with visual space projection using bipartite graph based similarity matrix
DDT_VP_CT	DDT with visual space projection using category-to-text transformation based similarity matrix
fusion ART	The fusion ART based cross-media resonance model
CMRM	Cross-media relevance model
DWH	Dual-wing harmonium model

methods based on the CMRM model and the DWH model respectively. Figure 10 shows the performance of using the various models for extracting image-text associations based on a five-fold cross-validation.

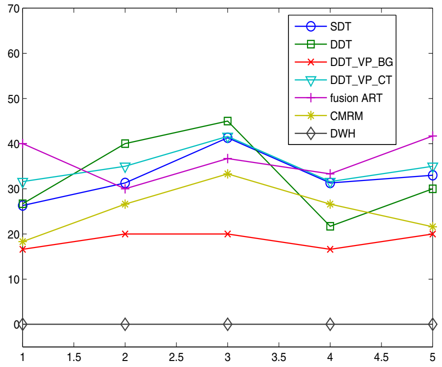


Fig. 10. Comparison of cross-media models for discovering image-text associations.

We see that among the seven methods, DDT_VP_CT and fusion ART provide the best performance. They outperform SDT and DDT which have a similar performance. All of these four models perform much better than DWH model, CMRM model and DDT_VP_BG. We can see that DWH model always obtains a precision of 0% and therefore cannot predict the correct image-text association for this particular experiment. The reason could be that the training data set is too small and on the contrary the data dimensions are quite large (501 for visual features and 8747 for textual features) to train a effective DWH model using Gibbs sampling [30]. It is surprising that DDT_VP_BG is the worst method other than the DWH model, hinting that the similarity matrix calculated based on bipartite graphs cannot really reflect the semantic similarity between the domain information categories. We shall revisit this issue in the next subsection. Note that although DDT outperforms SDT in most of the folds, there is a significant performance reduction in the fold 4. The reason could be that the reverse vague transformation results of certain text segments in the fold 4 are difficult to be discriminated due to the reason described in Section 3.3.4. Therefore, the reverse vague transformation based on text data may even lower the overall performance of the DDT. On the other hand, DDT_VP_CT performs much stable than DDT by incorporating the visual space projection.

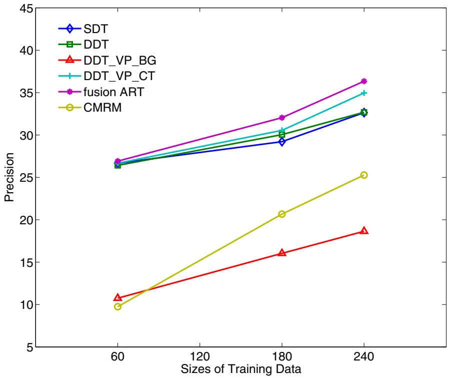


Fig. 11. Performance comparison of cross-media models with respect to different training data sizes.

For evaluating the impact of the size of training data on the learning performance, we also experiment with different data sizes for training and testing. As DWH model has been shown cannot be trained properly for this dataset, we leave it out in the rest of the experiments. Figure 11 shows the performance of the six cross-media similarity models with respect to training data of various sizes. We can see that when the size of the training data decreases, the precision of the CMRM model drops dramatically. In contrast, the performance of vague transformation and fusion ART drop less than 10% in terms of average precision. It shows that our methods also provide better performance stability on small data sets comparing with the statistical based cross-media relevance model.

4.4 Evaluation of Linear Mixture Similarity Model

In this section, we study the effect of using both textual and visual features in the linear mixture similarity model for discovering image-text associations. Referring to the experimental results in Table 3, we see that textual information is fairly reliable in identifying image-text associations. In fact, the pure text similarity measure ($\lambda = 1.0$) outperforms the pure cross-media similarity measure ($\lambda = 0.0$) by 20.7% to 24.4% in terms of average precision.

TABLE 3
The average precision scores (%) for image-text association extraction.

Methods	λ										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SDT	32.6	41.6	48.6	56.0	59.3	60.6	61.3	**61.6**	60.6	57.6	57.0
DDT	32.7	51.7	59.0	59.7	**61.7**	61.3	60.7	60.0	59.3	58.3	57.0
DDT_VP_BG	18.6	22.3	28.0	35	40.0	45.6	51.6	54.3	56.6	**57.3**	57.0
DDT_VP_CT	35.0	40.3	45.3	50.6	54.6	59.3	61.3	**62.6**	62.3	60.3	57.0
fusion ART	36.3	42.3	49.0	55.3	57.7	60.3	**62.0**	61.7	58.3	58.0	57.0

Caption In Web Pages	Police photograph the body of the gunman.	Wreckage of the base of the World Trade Center. The CIA searched the wreckage.	Injured man being helped away.
Text-Based Measure (λ = 1.0)	**At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.089)**	**The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.268)**	Others were not even able to do that. One witness said he saw several people lying on the floor of the bus, including one man whose legs had been blown off. (SC=0.110)
Cross-Media Measure (DDT_VP_CT, λ = 0.0)	**At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.157)**	A secret CIA office was destroyed in the 11 September attack on the World Trade Center, the New York Times reports. (SC=0.157)	**It was here on Thursday that a Palestinian suicide bomber blew himself up on board a crowded bus, killing five people and injuring about 50 others. (SC=0.087)**
Cross-Media Measure (Fusion ART, λ = 0.0)	**At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.430)**	**The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.97)**	**It was here on Thursday that a Palestinian suicide bomber blew himself up on board a crowded bus, killing five people and injuring about 50 others. (SC=0.265)**
Mixture Measure (DDT_VP_CT, λ = 0.7)	**At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.109)**	**The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.181)**	Others were not even able to do that. One witness said he saw several people lying on the floor of the bus, including one man whose legs had been blown off. (SC=0.093)
Mixture Measure (Fusion ART, λ = 0.6)	**At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.268)**	**The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.69)**	**It was here on Thursday that a Palestinian suicide bomber blew himself up on board a crowded bus, killing five people and injuring about 50 others. (SC=0.203)**

Fig. 12. A sample set of image-text associations extracted with similarity scores (SC). The correctly identified associated texts are bolded.

However, the best result is achieved by the linear mixture model using both the text-based and the cross-media similarity measures. DDT_VP_CT with $\lambda = 0.7$ can achieve an average precision of 62.6%, whilst the fusion ART with $\lambda = 0.6$ can achieve an average precision of 62.0%. On the average, the mixture similarity models can outperform the pure text similarity measure by about 5%. This shows that visual features are also useful in the identification of image-text associations. In addition, we observe that combining cross-media and text-based similarity measures improves the performance of pure text similarity measure on each fold of the experiment. Therefore, such improvement is stable. In fact, the keywords extracted from the captions of the images sometimes may be inconsistent with the contents of the images. For example, an image on the 911 attack scene may have a caption on the ceremony of 911 attack, such as “Victims’ families will tread Ground Zero for the first time”. In such a case, the visual features can compensate the imprecision in the textual features.

Among the vague transformation methods, dual-direction transformation achieves almost the same performance as single-direction transformation. However, visual space projection with dual-direction transformation can slightly improve the average precision. We can also see that the bipartite graph based similarity matrix D for visual space projection does not improve the image-text association results. By examining the classified text segments, we notice that only a small number of text segments belong to more than one categories and contribute to category similarities. This may have resulted in an inaccurate similarity matrix and a biased visual space projection.

On the other hand, the performance of fusion ART is comparable with that of vague transformation with visual space projection. Nevertheless, when using the pure cross-media model ($\lambda = 0.0$), fusion ART can actually outperform vague transformation based methods by about 1% to 3%. Looking into each fold of the experiment, we see that the fusion ART based method is much more stable than the vague transformation based methods in the sense that the best results are almost always achieved with $\lambda = 0.6$ or 0.7. For the vague transformation based methods, the best result of each experiment fold is obtained with rather different λ values. This suggests that the vague transformation based methods are more sensitive to the training data.

A sample set of the extracted image-text associations is shown in Figure 12. We find that cross-media models are usually good in associating general domain keywords with images. Referring to the second image in Figure 12, cross-media models can associate an image depicting the attack scene of 911 attack with a text segment containing the word “attack” which is commonly used for describing terrorist attack scenes. However, for the word “wreckage” that is a more specific word, cross-media models usually cannot identify it correctly. For such cases, using image captions may be helpful. On the other hand, as discussed before, image captions may not always reflect the image content accurately. For example, the caption of the third image contains the word “man”, which is a very general term, not quite relevant to the terrorist event. For such cases, cross-media models can be useful to find the proper domain-specific textual information based on the visual features of the images.

Figure 13 shows a sample set of the results by using fusion ART for image annotation. We can see that such annotations can reflect the direct associations between the visual and textual features in the images and texts. For example, the visual cue of “debris” in the images may be associated with words, such as “bomb” and “terror” in the text segments. Discovering such direct associations is an advantage of the fusion ART based method.

4.5 Discussions and Comparisons

In table 4, we provide a summary of the key characteristics of the two proposed methods. First of all, we note that the underlying ideas of the two approaches are quite different. Given a pair of image and text segment, the vague transformation based method translates features from one information space into another information space so that features of different spaces can be compared. The fusion ART based method, on the other hand, learns a set of prototypical image-text associations and then predicts the degree of association between an incoming pair of image and text segment by comparing it with the learned associations. During the prediction process, the visual and textual information are first compared in their respective spaces and the results consolidated based on a multimedia object resonance function (ART choice function).

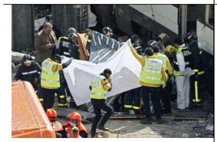
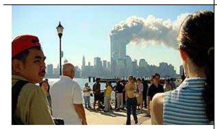
Images	Keyword Annotation Using fusion ART
	complex (0.5), house (0.33), attack (0.25), suicide (0.2), people (0.17), bomb (0.16), children (0.14)
	train (0.2), bomb (0.17), thursday (0.1)
	yorker (0.5), trade (0.5), terror (0.33), america (0.33), world (0.25), center (0.2), plane (0.2),

Fig. 13. Samples of image annotations using fusion ART.

Vague transformation is a statistic based method which calculates the conditional probabilities in one information space given some observations in the other information space. To calculate such conditional probabilities, we need to perform batch learning on a fixed set of training data. Once the transformation matrices are trained, they cannot be updated without building from scratch. In contrast, the fusion ART based method adopts an incremental competitive learning paradigm. The trained fusion ART can always be updated when new training data are available.

The vague transformation based method encodes the learnt conditional probabilities in transformation matrices. A fixed number of domain-specific information categories is used to reduce the information complexity. Instead of using predefined information categories, the fusion ART based method can automatically organize multimedia information objects into typical categories. The characteristics of an object category are encoded by a multimedia information object template. There are usually more category nodes learnt by the fusion ART. Therefore, the information in the fusion ART is less compact than that in the transformation matrices. In our experiments, around 70 to 80 categories are learnt by the fusion ART on a data set containing 300 images (i.e. 240 images are used for training in our five fold cross-validation).

In terms of efficiency, the vague transformation based

TABLE 4
Comparison of the vague transformation and the fusion ART based methods.

	Vague Transformation	fusion ART
Approach	Information are translated from one space to another space, so that information from different spaces can be compared	Visual and textual information are compared in their respective spaces and the results consolidated based on a multimedia object resonance function
Learning methodology	Statistic based batch learning	Incremental competitive learning
Information encoding	Using transformation matrices to summarize the information based on predefined information categories	Using self-organizing networks to learn typical categories of multimedia information objects
Network size	Number of predefined information categories is small. The information summarized in the transformation matrices is relatively more compact	More category nodes are created
Speed	Around 20 seconds for training; 20 seconds for testing; the training and testing time increases linearly with the number of data samples	Around 120 seconds for training; 30 seconds for testing; the training and testing time increases exponentially with the number of learnt object templates
Performance Stability	Unstable	Stable

method runs much faster than the fusion ART based method during both training and testing. However, the fusion ART based method produces a more stable performance than that of the vague transformation based method (see discussions in Section 4.4).

5 Conclusion

We have presented two distinct methods for learning and extracting associations between images and texts from multimedia web documents. The vague transformation based method utilizes an intermediate layer of information categories for capturing indirect image-text associations. The fusion ART based method learns direct associations between image and text features by employing a resonance environment of the multimedia objects. The experimental results suggest that both methods are able to efficiently learn image-text associations from a small training data set. Most notably, they both perform significantly better than the baseline performance provided by a typical image annotation model. In addition, while the text-matching based method is still more reliable than the cross-media similarity measures, combining visual and textual features provides the best overall performance in discovering cross-media relationships between components of multimedia documents.

Our proposed methods so far have been tested on a terrorist domain data set. It is necessary to extend our experiments to other domain data sets to obtain a more accurate assessment of the systems' performance. In addition, as both methods are based on a similarity based multilingual retrieval paradigm, using advanced similarity calculation methods with visterm and term taxonomies may result in a better performance. These will form part of our future work.

Acknowledgment

The reported work is supported in part by the I^{2}R-SCE Joint Lab on Intelligent Media. We would like to thank the anonymous reviewers for providing many invaluable comments to the previous versions of the paper. Special thanks to Dr. Eric P. Xing and Dr. Jun Yang at the Carnegie Mellon University for providing matlab source code of the dual-wing harmonium model.

References

[1] D. Radev, "A common theory of information theory from multiple text sources, step one: Cross-document structure," in *Proceedings 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, 2000.

[2] R. Barzilay, "Information fusion for multidocument summarization: paraphrasing and generation," Ph.D. dissertation, 2003, adviser-Kathleen R. Mckeown.

[3] H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt, "Automatic ontology-based knowledge extraction from web documents," *IEEE Intelligent Systems*, vol. 18, no. 1, pp. 14–21, 2003.

[4] A. Ginige, D. Lowe, and J. Robertson, "Hypermedia authoring," *Multimedia, IEEE*, vol. 2, no. 4, pp. 24–35, 1995.

[5] S.-F. Chang, R. Manmatha, and T.-S. Chua, "Combining text and audio-visual features in video indexing," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05).*, 2005, pp. 1005–1008.

[6] D. W. Oard and B. J. Dorr, "A survey of multilingual text retrieval," College Park, MD, USA, Tech. Rep., 1996.

[7] T. Mandl, "Vague transformations in information retrieval." in *ISI 98.*, 1998, pp. 312–325.

[8] T. Jiang and A.-H. Tan, "Discovering image-text associations for cross-media web information fusion." in *PKDD/ECML 2006*, 2006, pp. 561–568.

[9] A.-H. Tan, G. A. Carpenter, and S. Grossberg, "Intelligence through interaction: Towards a unified theory for learning," in *D. Liu et al. (Eds.): International Symposium on Neural Networks (ISNN) 2007*, ser. Lecture Notes in Computer Science, vol. 4491, 2007, pp. 1098–1107.

[10] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics and Image Processing*, vol. 37, pp. 54–115, 1987.

[11] J. He, A.-H. Tan, and C. L. Tan, "On machine learning methods for chinese document categorization." *Appl. Intell.*, vol. 18, no. 3, pp. 311–322, 2003.

[12] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03*. New York, NY, USA: ACM Press, 2003, pp. 119–126.

[13] S. Little, J. Geurts, and J. Hunter, "Dynamic generation of intelligent multimedia presentations through semantic inferencing," in *ECDL '02*, London, 2002, pp. 158–175.

[14] J. Geurts, S. Bocconi, J. van Ossenbruggen, and L. Hardman, "Towards ontology-driven discourse: From semantic graphs to multimedia presentations." in *International Semantic Web Conference*, 2003, pp. 597–612.

[15] J. Han, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[16] H. H. Yu and W. H. Wolf, "Scenic classification methods for image and video databases," C.-C. J. Kuo, Ed., vol. 2606, no. 1. SPIE, 1995, pp. 363–371. [Online]. Available: http://link.aip.org/link/?PSI/2606/363/1

[17] I. K. Sethi, I. L. Coman, and D. Stan, "Mining association rules between low-level image features and high-level concepts," B. V. Dasarathy, Ed., vol. 4384, no. 1. SPIE, 2001, pp. 279–290. [Online]. Available: http://link.aip.org/link/?PSI/4384/279/1

[18] M. Blume and D. R. Ballard, "Image annotation based on learning vector quantization and localized haar wavelet transform features," S. K. Rogers, Ed., vol. 3077, no. 1. SPIE, 1997, pp. 181–190. [Online]. Available: http://link.aip.org/link/?PSI/3077/181/1

[19] A. Mustafa and I. K. Sethi, "Creating agents for locating images of specific categories," S. Santini and R. Schettini, Eds., vol. 5304, no. 1. SPIE, 2003, pp. 170–178. [Online]. Available: http://link.aip.org/link/?PSI/5304/170/1

[20] Q. Ding, Q. Ding, and W. Perrizo, "Association rule mining on remotely sensed images using p-trees," in *PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. London, UK: Springer-Verlag, 2002, pp. 66–79.

[21] J. Tesic, S. Newsam, and B. S. Manjunath, "Mining image datasets using perceptual association rules," in *SIAM Sixth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Third SIAM International Conference (SDM)*, May 2003. [Online]. Available: http://vision.ece.ucsb.edu/publications/03SDMJelena.pdf

[22] T. Kohonen, *Self-Organizing Maps*, T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[23] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of VLDB'94, Santiago de Chile, Chile*, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 1994, pp. 487–499.

[24] A. M. Teredesai, M. A. Ahmad, J. Kanodia, and R. S. Gaborski, "Comma: a framework for integrated multimedia mining using multi-relational associations," *Knowl. Inf. Syst.*, vol. 10, no. 2, pp. 135–162, 2006.

[25] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach." *Data Min. Knowl. Discov.*, vol. 8, no. 1, pp. 53–87, 2004.

[26] C. Djeraba, "Association and content-based retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 1, pp. 118–135, 2003.

[27] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *ICCV 2001.*, vol. 2, 2001, pp. 408–415.

[28] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV '02*, London, 2002, pp. 97–112.

[29] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, 2003.

[30] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*. Arlington, Virginia: AUAI Press, 2005, p. 633.

[31] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the spider system," in *SIGIR '96*. New York: ACM Press, 1996, pp. 58–65.

[32] P. Biebricher, N. Fuhr, G. Lustig, M. Schwantner, and G. Knorz, "The automatic indexing system air/phys - from research to applications," in *SIGIR '88*. New York: ACM Press, 1988, pp. 333–342.

[33] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377. [Online]. Available: citeseer.ist.psu.edu/tishby99information.html

[34] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*. New York, NY, USA: ACM, 2006, pp. 35–44.

[35] G. Carpenter and S. Grossberg, *Pattern Recognition by Self-Organizing Neural Networks*. Cambridge, MA: MIT Press, 1991.

[36] A.-H. Tan, "Adaptive resonance associative map," *Neural Netw.*, vol. 8, no. 3, pp. 437–446, 1995.

[37] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol. 26, pp. 4919–4930, 1987.

[38] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition," *Neural Netw.*, vol. 4, no. 4, pp. 493–504, 1991.

[39] ——, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system." *Neural Networks*, vol. 4, no. 6, pp. 759–771, 1991.

[40] W. Li, K.-L. Ong, and W. K. Ng, "Visual terrain analysis of high-dimensional datasets." in *PKDD*, ser. Lecture Notes in Computer Science, A. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, Eds., vol. 3721. Springer, 2005, pp. 593–600.

[41] F. Chu, Y. Wang, and C. Zaniolo, "An adaptive learning approach for noisy data streams," *icdm*, vol. 00, pp. 351–354, 2004.

[42] D. Shen, Q. Yang, and Z. Chen, "Noise reduction through summarization for web-page classification," *Inf. Process. Manage.*, vol. 43, no. 6, pp. 1735–1747, 2007.

[43] A. Tan, H. Ong, H. Pan, J. Ng, and Q. Li, "Foci: A personalized web intelligence system," in *IJCAI workshop on Intelligent Techniques for Web Personalization*, Seattle, Aug 2001, pp. 14–19.

[44] A.-H. Tan, H.-L. Ong, H. Pan, J. Ng, and Q.-X. Li, "Towards personalised web intelligence." *Knowledge and Information Systems*, vol. 6, no. 5, pp. 595–616, 2004.

[45] E. W. M. Lee, Y. Y. Lee, C. P. Lim, and C. Y. Tang, "Application of a noisy data classification technique to determine the occurrence of flashover in compartment fires," *Advanced Engineering Informatics*, vol. 20, no. 2, pp. 213–222, 2006.

[46] A. M. Fard, H. Akbari, R. Mohammad, and T. Akbarzadeh, "Fuzzy adaptive resonance theory for content-based data retrieval," *Innovations in Information Technology, 2006*, pp. 1–5, Nov. 2006.

[47] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation." in *CVPR (2)*, 2004, pp. 1002–1009.

[48] M. Sharma, "Performance evaluation of image segmentation and texture extraction methods in scene analysis," Master's thesis, 1998.

[49] P. Duygulu, O. C. Ozcanli, and N. Papernick, *Comparison of Feature Sets Using Multimedia Translation*, 2869th ed., ser. Lecture Notes in Computer Science, 2003.

[50] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

Tao Jiang received his Ph.D. degree from the Nanyang Technological University. Prior to that, he obtained his Bachelor of Science in Computer Science and Technology from Peking University in 2000. His research interests include data mining, machine learning and multimedia information fusion. Since Oct 2007, he has been with ecPresence Technology Pte Ltd where he is currently a project manager. From July 2000 to May 2003, he worked in Found Group, one of the biggest IT companies in China, earlier as a software engineer and later as a technical manager. Now, he also serves as a coordinator of "vWorld Online Community" project supported and sponsored by Multimedia Development Authority (MDA), Singapore.

Ah-Hwee Tan received the B.S. degree (first class honors) and the M.S. degree in Computer and information science from the National University of Singapore, in 1989 and 1991, respectively, and subsequently received the Ph.D. degree in cognitive and neural systems from Boston University, Boston, MA, in 1994. He is currently an Associate Professor and Head, Division of Information Systems at the School of Computer Engineering, Nanyang Technological University. He was the founding Director of the Emerging Research Laboratory, a research centre for incubating interdisciplinary research initiatives. Prior to joining NTU, he was a Research Manager at the A*STAR Institute for Infocomm Research (I2R), responsible for the Text Mining and Intelligent Agents research groups. His current research interests include cognitive and neural systems, information mining, machine learning, knowledge discovery, document analysis, and intelligent agents. Dr. Tan holds five patents and has published over eighty technical papers in books, international journals, and conferences. He is an editorial board member of Applied Intelligence and a member of the ACM.