9-2006

Discovering image-text associations for cross-media web information fusion

Tao JIANG
Nanyang Technological University

Ah-Hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Citation

JIANG, Tao and TAN, Ah-Hwee. Discovering image-text associations for cross-media web information fusion. (2006). *Knowledge Discovery in Databases: 17th European Conference, PKDD 2006, Berlin, Germany, September 18-22: Proceedings*. 4213, 561-568.

Available at: https://ink.library.smu.edu.sg/sis_research/6770

Discovering Image-Text Associations for Cross-Media Web Information Fusion

Tao Jiang and Ah-Hwee Tan

School of Computer Engineering
Nanyang Technological University, Nanyang Avenue, Singapore 639798
{jian0006, asahtan}@ntu.edu.sg

Abstract. 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. Whereas most existing work on web information fusion has focused on multiple document summarization, this paper presents a novel approach for discovering associations between images and text segments, which subsequently can be used to support cross-media web content summarization. Specifically, we employ a similarity-based multilingual retrieval model and adopt a vague transformation technique for measuring the information similarity between visual features and textual features. The experimental results on a terrorist domain document set suggest that combining visual and textual features provides a promising approach to image and text fusion.

1 Introduction

The diverse and distributed nature of the information on the World Wide Web has made it difficult to collate and track information related to specific topics. Techniques for web information fusion, involving filtering of redundant information, collating of information according to themes, and generation of coherent presentation, are needed for information users. As a useful 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. As an increasing amount of non-text content, namely images, video, and sound, is becoming available on the web, summarizing multimedia information has posed a key challenge in web information fusion.

In this paper, we focus on one of the important problems in multimedia fusion, namely the extraction of association between multimedia components, in particular, images and texts. Our approach is consistent with those found in the literatures of hypermedia authoring and cross-document text summarization, that understanding the interrelation between information blocks are essential for collating information and generating final presentations.

By extending a process for multi-document summarization [1], we present a procedure for web document fusion (Figure 1) consisting of five stages as follows.

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 561–568, 2006.

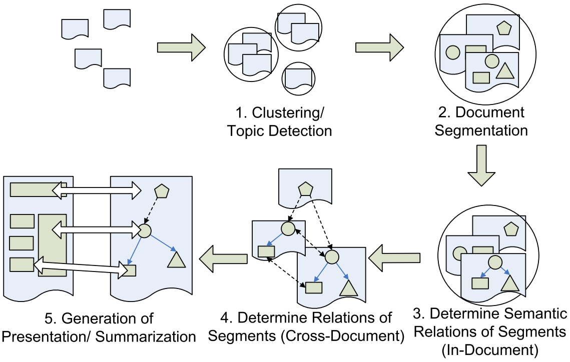

Fig. 1. An overview of the web information fusion process

1. The raw web documents are first clustered according to their topics.
2. Each document is divided into several atomic segments (atomic description unit) according to sub-topics and media types.
3. For each document, the relations among document segments are determined. In our work, we focus on the relations across media types.
4. Within a document cluster, the cross-document relations among document segments are determined. Duplicated contents are detected.
5. The document segments are reorganized and presented according to a summarization template and the in-document and cross-document relations.

We see that techniques developed for text documents can be used in the first two stages of the cross-media summarization process, i.e. document clustering and segmentation (where each multimedia object itself can be seen as a segment). For detecting the associations and relationships between multimedia components (e.g. text segments and images) within or across documents, we present a textual-visual vague transformation technique, borrowed from the field of multilingual retrieval [2], for extracting associations between images and texts from news web documents. The extracted image-text associations can be subsequently used for the third, fourth, and fifth stages of the summarization.

Note that our method is different from the existing efforts on image indexing using statistical modelling approaches originally proposed in the field of natural language processing [3]. Image indexing tends to establish the correspondence between keywords (concepts) and particular image regions. Our task, however, does not require such a correspondence between the contents of the text segments and the associated images. In the next two sections, we present our methods for data preprocessing and image-text association learning. Section 4 reports our experiments. Concluding remarks are given in Section 5.

2 Harvesting and Preprocessing of Texts and Images

We develop an image crawler, named “ICrawl”, based on Yahoo Search API. Upon receiving a query, ICrawl searches images through Yahoo search engine, downloads the images retrieved, and extracts the textual contents from the web pages wherein the images appear. The extracted textual contents include the tips and captions of the images, the keywords extracted from the URLs/tips/captions, and long text paragraphs (more than 15 words).

2.1 Textual Feature Extraction

Currently, we treat each text paragraph extracted from web pages as a text segment. We tokenize the text segments, add part-of-speech tags, remove stop words, replace tokens with their stems, filter out terms with unwanted POS tags (only nouns, verbs and adjectives are left), and finally generate term vectors. For images downloaded from the web, their surrounding texts, including captions, tips, and keywords in URL, are extracted. Like text segments, the extracted surrounding texts are processed to form term vectors.

For calculating the term weights of the term vectors, we use a model, named TF-ITSF, similar to a traditional TF-IDF model. For a text segment or an image text description (the surrounding text of an image) ts in a web document d, we use the following equation to weight a term w:

$$w^d(ts) = tf(ts, w) \cdot \log \frac{N^d}{tsf^d(w)} \tag{1}$$

where $tf(ts, w)$ denotes the frequency of w in the text segment ts, N^d is the total number of text segments and text descriptions of images in web document d, and $tsf^d(w)$ is the text segment frequency of term w in web document d.

2.2 Visual Feature Extraction from Images

For an image downloaded, we first segment it into 10×10 rectangle regions. For each region we extract a visual feature vector, consisting of 6 color features and 60 gabor texture features which have been proven to be useful in many applications. Color features are the means and variances of the RGB color spaces. Texture features are extracted by calculating the means and variations of the Gabor filtered image regions on 6 orientations at 5 scales (frequencies). After the visual feature vectors of the image regions are extracted, all image regions are clustered using the k-means algorithm with k=500. The generated clusters, called *visterms*, are treated as a vocabulary for the images. For enriching this vocabulary of visterms with the high-level semantic features, a face detection model is used for detecting the faces in the images, which we found useful for understanding the contents of images in the domain of terror attack. Finally, a vector of visterm frequencies (501 dimensions) is extracted for each image.

3 Identifying Associations Between Texts and Images

3.1 Similarity-Based Retrieval Model

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. Suppose each image i is represented by a visterm vector, denoted as i_v, together with a term vector of its surrounding text, denoted as i_t. For calculating the similarity between the images and text segments, we need to define a similarity measure $sim_d(i, ts) = sim_d(< i_v, i_t >, ts)$.

For simplifying the problem, we assume that, once an image i and a text segment ts are given, the similarity between i_v and ts and the similarity between i_t and ts are independent. Therefore, we can calculate $sim_d(i, ts)$ with the use of a linear mixture model as follows:

$$sim_d(i, ts) = sim_d(< i_v, i_t >, ts) = \lambda \cdot sim_d^{tt}(i_t, ts) + (1 - \lambda) \cdot sim_d^{vt}(i_v, ts). \quad (2)$$

3.2 Text-Based Similarity Measure

We use cosine distance as in Eq. 3 to measure the similarity between the textual features of an image and a text segment. The cosine distance measure is used as it has been proven to be insensitive to the length of text documents.

$$sim_d^{tt}(i_t, ts) = \frac{\sum_{k=1}^{n} w_k^d(i_t) w_k^d(ts)}{\| i_t \| \| ts \|} \quad (3)$$

3.3 Cross-Media Similarity Measure

Measuring 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. 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 problem can be more serious in visual-textual transformation. Individual visterms can hardly convey any meaningful semantics without considering the contexts where they are placed. However, words in natural languages usually have relatively complete meanings. We can imagine that in most cases visterms can hardly be directly and precisely mapped to words because of the ambiguity. Vague transformations [2] [4] which have been proved useful in IR seem suitable for solving the vague problem of mapping visual representations of images to textual representations of text segments. In this paper, we borrow the idea from statistical vague transformation methods in multilingual retrieval for our cross-media similarity measure.

A drawback of the existing methods [2] [4] is that they require a large training set to build multilingual thesauruses. Such a training set is usually unavailable. 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 obstacles, we introduce

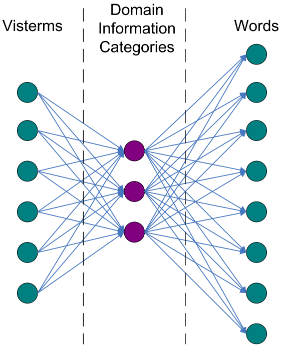

Fig. 2. An illustration of cross-media transformation with information bottleneck

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 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 visterm space to domain information categories and then to word space (see Figure 2). This is actually known as the *information bottleneck* method. 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 Section 1, for an associated pair of image and text, their contents may not be an exact match or mapping. However, we believe that they can always be matched on a general domain information category.

Based on the above discussion, we aim to build two thesauri in the form of transformation matrices, each of which corresponds to a sub-transformation. Suppose the visterm space $\mathcal{V}$ has m dimensions and the textual feature space $\mathcal{T}$ has n dimensions. In addition, we suppose 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:

$$M^{\mathcal{VC}}_{m\times l} = \begin{pmatrix} m^{\mathcal{VC}}_{11} & m^{\mathcal{VC}}_{12} & .. & m^{\mathcal{VC}}_{1l} \\ m^{\mathcal{VC}}_{21} & m^{\mathcal{VC}}_{22} & .. & m^{\mathcal{VC}}_{2l} \\ . & . & & . \\ m^{\mathcal{VC}}_{m1} & m^{\mathcal{VC}}_{m2} & .. & m^{\mathcal{VC}}_{ml} \end{pmatrix}, \quad M^{\mathcal{CT}}_{l\times n} = \begin{pmatrix} m^{\mathcal{CT}}_{11} & m^{\mathcal{CT}}_{12} & .... & m^{\mathcal{CT}}_{1n} \\ m^{\mathcal{CT}}_{21} & m^{\mathcal{CT}}_{22} & .... & m^{\mathcal{CT}}_{2n} \\ . & . & & . \\ m^{\mathcal{CT}}_{l1} & m^{\mathcal{CT}}_{l2} & .... & m^{\mathcal{CT}}_{ln} \end{pmatrix}; \qquad (4)$$

where $m^{\mathcal{VC}}_{ij}$ represents the association factor between the visterm v_i and the information category c_j; and $m^{\mathcal{CT}}_{jk}$ represents the association factor between the information category c_j and the textual feature t_k. Currently, $m^{\mathcal{VC}}_{ij}$ and $m^{\mathcal{CT}}_{jk}$ are calculated by

$$m^{\mathcal{VC}}_{ij} = P(c_j|v_i) \approx \frac{\#(v_i, c_j)}{\#(v_i)}, \quad m^{\mathcal{CT}}_{jk} = P(t_k|c_j) \approx \frac{\#(c_j, t_k)}{\#(c_j)}; \qquad (5)$$

where $\#(v_i)$ is the number of images containing the visterm v_i; $\#(v_i, c_j)$ is the number of images containing v_i and belonging to the information category c_j; $\#(c_j)$ is the number of text segments belonging to the category c_j; and $\#(c_j, t_k)$ is the number of text segments belonging to c_j and containing the term t_k.

Based on Eq. 4, we can define the similarity between the visual part of an image i_v and a text segment ts as $i_v^T M_{m\times l}^{\mathcal{VC}} M_{l\times n}^{\mathcal{CT}} ts$. For embedding into Eq. 2, we use its normalized form

$$sim^{\mathcal{VT}}(i_v, ts) = \frac{i_v^T M_{m\times l}^{\mathcal{VC}} M_{l\times n}^{\mathcal{CT}} ts}{\| i_v^T M_{m\times l}^{\mathcal{VC}} M_{l\times n}^{\mathcal{CT}} \| \| ts \|}. \tag{6}$$

Eq. 6 calculates the cross-media similarity using a *single-direction transformation* from visterm space to word space. However, it may still cause vague problems. For example, suppose there is a picture i belonging to a domain information category, *Attack Details*, and two text segments $ts1$ and $ts2$ 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 transformation result of i_v might be similar to both $ts1$ and $ts2$. To reduce the influence of common terms in different categories and employ the strength of the distinct words, we consider another transformation from word space to visterm space. We can similarly define another pair of transformation matrices $M_{n\times l}^{\mathcal{TC}} = \{m_{kj}^{\mathcal{TC}}\}^{n\times l}$ and $M_{l\times m}^{\mathcal{CV}} = \{m_{ji}^{\mathcal{CV}}\}^{l\times m}$, where $i = 1, 2, ..., m$, $j = 1, 2, ..., l$, and $k = 1, 2, ..., n$. Then, the similarity from a text segment ts to the visual part of an image i_v can be defined as

$$sim^{\mathcal{TV}}(ts, i_v) = \frac{ts^T M_{n\times l}^{\mathcal{TC}} M_{l\times m}^{\mathcal{CV}} i_v}{\| ts^T M_{n\times l}^{\mathcal{TC}} M_{l\times m}^{\mathcal{CV}} \| \| i_v \|}. \tag{7}$$

Finally, we can define a cross-media similarity measure using the *dual-direction transformation* which is the geometric mean of $sim^{\mathcal{VT}}(i_v, ts)$ and $sim^{\mathcal{TV}}(ts, i_v)$:

$$sim_d^{vt}(i_v, ts) = \sqrt{sim^{\mathcal{VT}}(i_v, ts) \cdot sim^{\mathcal{TV}}(ts, i_v)}. \tag{8}$$

4 Experiments

The experiments are conducted on an image collection, containing 285 images related to terrorist attacks, downloaded from the CNN and BBC news web sites. We manually categorize about 1500 text segments and 285 images into twelve domain information categories, i.e. *Anti-Terror, Attack Details, After Attack, Government Responses, Rescure, Impact, Investigation, Terrorist Claims, Terrorist Suspects, Victims, Ceremony*, and *Others*. We use a 5-fold cross-validation to test the performance of our method in terms of precision defined by $precision = \frac{\#(Correctly\ Identified\ Associations)}{\#(Total\ Images)}$. The correctness of the extracted image-text associations are judged by human by inspecting the web pages wherein the images appear.

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.
Cross-Media Measure (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=0.129)**	A secret CIA office was destroyed in the 11 September attack on the World Trade Center, the New York Times reports. (SC=0.142)	**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=0.085)**
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)
Mixture Measure (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=0.104)**	**The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.216)**	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=0.084)

Fig. 3. A sample set of image-text associations extracted with similarity scores (SC). The correctly identified associated texts are bolded.

As indicated by the experimental results shown in Table 1, we see that textual information is essential for identifying image-text associations. In fact, pure text similarity measure ($\lambda = 1.0$) outperforms pure cross-media similarity measure ($\lambda = 0.0$) by 23.0%-25.7% in terms of average precision. However, the best result (average precision of 67.2%) is achieved by the linear mixture model using both text-based and cross-media similarity measures (with $\lambda = 0.6$). This shows that visual features are useful for improving the performance of the identification task. In fact, we observe that keywords extracted from surrounding texts of images sometimes may be inconsistent with the contents of the images. Visual features can provide more information for the disambiguation of the image semantics and reducing the influence of the imprecision caused by textual features. Another important observation is that pure dual-direction transformation is better than pure single-direction transformation for measuring the cross-media similarity ($\lambda = 0.0$). In general, the overall precision of using dual-direction transformation is higher than that of using single-direction transformation.

A sample set of the extracted image-text associations is shown in Figure 3. We notice that when text contents in web pages are quite different, e.g. belonging

Table 1. The precision scores (%) for image-text association extraction

		λ										
Fold	Transformation Used	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1	Single-Direction	41.5	56.6	62.3	69.8	**71.7**	69.8	66.0	66.0	66.0	66.0	62.3
	Dual-Direction	43.4	56.6	64.2	67.9	**73.6**	69.8	69.8	69.8	67.9	67.9	62.3
2	Single-Direction	32.1	45.3	47.2	58.5	60.4	60.4	**64.2**	**64.2**	**64.2**	62.3	62.3
	Dual-Direction	37.7	45.3	47.2	60.4	64.2	62.3	**66.0**	64.2	64.2	62.3	62.3
3	Single-Direction	26.4	34.0	45.3	54.7	62.3	64.2	**69.8**	**69.8**	**69.8**	67.9	66.0
	Dual-Direction	26.4	34.0	43.4	54.7	62.3	66.0	**69.8**	**69.8**	**69.8**	67.9	66.0
4	Single-Direction	37.7	49.1	56.6	62.3	64.2	64.2	66.0	**67.9**	66.0	64.2	64.2
	Dual-Direction	41.5	47.2	56.6	66.0	64.2	66.0	**67.9**	**67.9**	**67.9**	64.2	64.2
5	Single-Direction	43.4	58.5	66.0	64.2	64.2	**66.0**	64.2	62.3	58.5	54.7	54.7
	Dual-Direction	45.3	54.7	60.4	64.2	**66.0**	64.2	62.3	62.3	60.4	54.7	54.7
Average	Single-Direction	36.2	48.7	55.5	61.9	64.5	64.9	**66.0**	**66.0**	64.9	63.0	61.9
	Dual-Direction	38.9	47.5	54.3	62.6	66.0	65.7	**67.2**	66.8	66.0	63.4	61.9

to different domain information categories, the cross-media similarity measure may be efficient enough to identify the associated text for an image (see the first and third column in Figure 3). For the case that contents in different text segments are related to each other, using only the cross-media similarity measure may not identify the most suitable text segment, but the extracted one can be semantically relevant (the second column in Figure 3).

5 Conclusion

In this paper, we present an approach for extracting associations between images and texts from web pages for cross-media information fusion. We use a similarity-based multilingual retrieval model and adopt a vague transformation technique for measuring the similarity between visual features and textual features. The experimental results suggest that combination of visual and textual features can produce better results than using visual or textual features alone.

References

1. Radev, D.R.: A common theory of information fusion from multiple text sources step one: cross-document structure. In: Proceedings of the 1st SIGdial workshop on Discourse and dialogue, Morristown, NJ, USA, Association for Computational Linguistics (2000) 74–83
2. Mandl, T.: Vague transformations in information retrieval. In: ISI. (1998) 312–328
3. Chang, S.F., Manmatha, R., Chua, T.S.: 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) 1005–1008
4. Sheridan, P., Ballerini, J.P.: Experiments in multilingual information retrieval using the spider system. In: SIGIR '96, New York, ACM Press (1996) 58–65