Discovering Multilingual Concepts from Unaligned Web Documents by Exploring Associated Images

Xiaochen Zhang† Xiaoming Jin† Lianghao Li‡ Dou Shen§
†School of Software, Tsinghua University, Beijing 100084, P.R. China
‡Hong Kong University of Science and Technology, Hong Kong, P.R. China
§Baidu, Beijing, P.R. China
jasiazhang@gmail.com, xmjin@tsinghua.edu.cn, llias@ust.hk, doushen@live.com

ABSTRACT

The Internet is experiencing an explosion of information presented in different languages. Though written in different languages, some articles implicitly share common concepts. In this paper, we propose a novel framework to mine cross-language common concepts from unaligned web documents. Specifically, visual words of images are used to bridge articles in different languages and then common concepts of multiple languages are learned by using an existing topic modeling algorithm. We conduct cross-lingual text classification in a real-world data set using the mined multilingual concepts from our method. The experiment results show that our approach is effective to mine cross-lingual common concepts.

Categories and Subject Descriptors

L.1.4 [**Knowledge and Media**]: Semantic Web; I.2.7 [**Artificial Intelligence**]: Language Models

Keywords

Multilingual, Image, Common Concepts

1. INTRODUCTION

Discovering multilingual articles sharing common topics is significant for many real-world applications. Additionally, articles on the Internet always appear together with one or more associated images, which provide a convenient way to bridge multilingual articles. Intuitively, similar images convey the same topics. Thus even written in different languages, articles with similar or the same images are likely to illustrate the same topics. Articles shown in Figure 1 serve as an good example. The two articles displayed in this figure both talk about 2012 Olympic Games. Since the images in these two articles are very similar (almost the same except the color styles are different), for an Internet user whose native language is English but knows nothing about Chinese, the user can still guess that the article in Chinese talks about 2012 Olympic Games as well.

Motivated by this observation, we propose a novel method to bridge multilingual articles by building co-occurrence between different languages with articles' associated images. Specifically, we represent each article using both the textual features extracted from its textual document and the visual features learned from the associated images. The textual features, such as bag-of-words, can be very different under various language domains. However, the visual features of images are shared by all language domains. By using the shared visual features as a bridge, we can connect multiple language domains and jointly learn common concepts from all articles. Unlike many existing cross-langauge/domain topic mining methods [3, 4], we only utilize the shared image features instead of the article itself to bridge different languages. So our method does not require the articles to be the same or aligned beforehand.

(a) Wikipedia (b) Baidu Baike

Figure 1: Two articles about 2012 Olympic Games in different languages

2. PROBLEM AND OUR SOLUTION

Our method takes multi-lingual articles as input, each of which consists of a textual document and one or more images. By jointly extracting topics from the articles, our method outputs a set of cross-lingual common concepts, which then can be used to enrich the feature representation of each article. In the following, we first introduce some definitions for our problem. Then, our proposed multi-lingual concept mining method is discussed in details.

Definition 1. (**Multi-Lingual Corpus**) Given a multi-language corpus C, it consists of articles written in one of N languages. For each language L_i, we denote its vocabulary as $W^i = \{w_1^i, w_2^i, ..., w_{n_i}^i\}$, and the number of articles written in L_i as M_i. Each article A_j^i consists of a textual document T_j^i and a set of corresponding images I_j^i.

In this paper, textual document T_j^i is represented as a bag-of-words feature vector $\{x_1^i, x_2^i, ..., x_{n_i}^i\}$, where x_t^i is the amount of times that term w_t^i has been observed in document T_j^i. Additionally, we represent all images in the same visual feature space. Specifically, we adopt the SURF algorithm [1] to extract features from images and use k-means to quantize image features into k visual words. Specifically, I_j^i is represented as a bag-of-visual-words vector of $\{x_{n_i+1}^i, x_{n_i+2}^i, ..., x_{n_i+k}^i\}$, where $x_{n_i+l}^i$ is the learned feature weight for visual feature v_l^i.

By combining both the textual features and image fea-

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

tures, we can re-represent each article in the i^{th} language as $\{x^i_1, ..., x^i_{n_i}, x^i_{n_i+1}, ..., x^i_{n_i+k}\}$, where image features are shared by articles in all languages. Thus we build the co-occurrence between words in different languages by using the shared image features as a bridge. Then all articles represented by combined text and image features are used as input corpus to run the multi-lingual concept mining model.

In this paper, we choose Latent Dirichlet Allocation (LDA) [2] as our multi-language concept mining model. As a generative graphical model, LDA can be represented as a three level hierarchical Bayesian model. Given a corpus (the multi-lingual corpus in our case) consisting of M documents, and suppose there are K topics $\beta = \beta_{1:K}$. The generation process for a document W, is described as follows:

1. Choose $\theta \sim Dirichlet(\alpha)$
2. For each of the N words w_n:
   (a) Choose topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta_{1:K})$, a multinomial probability conditioned on the topic z_n.

We use Gibbs Sampling for parameters estimation in LDA. After finishing Gibbs Sampling, the multinomial parameter sets $z_{k,t}$ and $\theta_{m,k}$ are computed as follows.

$$z_{k,t} = \frac{n^{(t)}_k + \beta_t}{\sum_{t=1}^{V} n^{(t)}_k + \beta_t}, \qquad \theta_{m,k} = \frac{n^{(k)}_m + \alpha_k}{\sum_{k=1}^{K} n^{(k)}_m + \alpha_k}$$

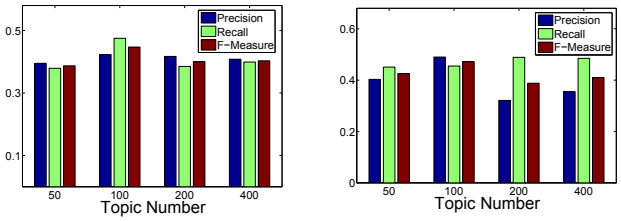
where $n^{(t)}_k$ denotes the number of times that word t is assigned to topic k; $\sum_{t=1}^{V} n^{(t)}_k$ is the total number of words assigned to topic k.

For a new document d, the topic distribution of d can be represented as $\theta^d = \{\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,k}\}$. Thus after topic inference, all documents in different languages can be represented in the same topic space and have the form of $d = \{\theta_{d,1}, \theta_{d,2}, ..., \theta_{d,k}\}$.

3. EXPERIMENT

In this section, we conduct experiments on real-world data sets to evaluate the effectiveness of our proposed method. We first adopt our method on *ImageCLEF* Wikipedia Image Retrieval Dataset to extract cross-language common concepts. The *ImageCLEF* dataset consists of 53,715 English articles and 39,744 German articles, each of which contains a textual document and one or more corresponding images. Then, we use the mined concepts as features and test its effectiveness in cross-lingual text classification (CLTC). We conduct our experiment on Amazon review dataset, which contains thousands of Book, DVD and Magazine reviews written in different languages. Each review is labeled as Book, DVD or Magazine according to the product class it talks about. In this paper, we adopt the reviews written in English and German in our experiment. We adopt the mined multilingual concepts to infer hidden concepts for Amazon product reviews. The Support Vector Machines (SVMs) are used as the basic classifier in all experiments. We test our method in two settings: ($En \rightarrow De$) and ($De \rightarrow En$). In each setting, we adopt product reviews written in one language as training data, and reviews written in the other language as test data.

We use Precision, Recall and F-measure to evaluate the classification performance. Figure 2(a) demonstrates the classification results on $En \rightarrow De$ CLTC task under various numbers of topics. We can see that the optimal results is



(a) Results on $En \rightarrow De$ (b) Results on $De \rightarrow En$

Figure 2: CLTC Results with Various Topics

Table 1: CLTC F-measure of Our Method and MCs

Topic Number	50	100	200	400
En → De CLTC Results				
Our Method	0.3701	0.4226	0.3899	0.3998
MC	0.6315	0.6216	0.6749	0.5385
De → En CLTC Results				
Our Method	0.3678	0.4078	0.3450	0.3598
MC	0.7274	0.7128	0.7059	0.6714

obtained with 100 common concepts on $En \rightarrow De$. Figure 2(b) shows similar trends as in $En \rightarrow De$ task. In addition, we can observe that our CLTC classification method performs better than random classification result (which is about 0.33 in precision, recall and F-measure).

To illustrate the effectiveness of our method, we build a set of monolingual classifiers (MCs), each of which is trained and tested on the product reviews written in the same language. We denote it as the upper bound baseline. Table 1 shows the classification performance of our method and the upper bound baseline MCs. We can observe that although our method outperforms random classification result, when compared with the MCs approach, there is still a large room to improve. In addition, we can find that our method and MCs share similar classification trends as the number of common concepts varies.

4. CONCLUSION

In this paper, we propose a novel method to mine cross-lingual common concepts from unaligned corpus using associated images. Our method postulates that articles even from different languages share some common concept as long as their associated images are similar. We conduct CLTC adopting the cross-lingual common concepts mined by our method and the results indicate that our method works well.

5. ACKNOWLEDGMENT

The work was supported by National Natural Science Foundation of China (60973103, 90924003).

6. REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proceedings of ECCV*, pages 404–417. Springer, 2006.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*,, 3:993–1022, 2003.

[3] L. Li, X. Jin, and M. Long. Topic correlation analysis for cross-domain text classification. In *Proceedings of AAAI*, pages 998–1004. AAAI Press, 2012.

[4] X. Ni, J.-T. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In *Proceedings of WSDM*, pages 375–384. ACM, 2011.