Image Discovery and Insertion for Custom Publishing

Lei Liu
HP Labs
1501 Page Mill Rd
Palo Alto, CA 94304
lei.liu2@hp.com

Jerry Liu
HP Labs
1501 Page Mill Rd
Palo Alto, CA 94304
jerry.liu@hp.com

Shanchan Wu
HP Labs
1501 Page Mill Rd
Palo Alto, CA 94304
shanchan.wu@hp.com

ABSTRACT

Images in reading materials make the content come alive. Aside from providing additional information to the text, reading material containing illustration engages our spatial memory, increases memory retention of the material. However, despite the plethora of available multimedia, adding illustrations to text continues to be a difficult task for the amateur content publisher. To address this problem, we present a semantic-aware image discovery and insertion system for custom publishing. Compared to image search engines, our system has the advantage of being able to discern among different topics within a long text passage and recommend the most relevant images for each detected topic with semantic "visual words" based relevance.

1. INTRODUCTION

Eye-catching illustrations make reading experience more attractive, result in increasing reading engagement and memory retention rate. However, associating the image illustrations with the appropriate text content can be a painful task which takes much time and efforts, especially for the amateur content creator. These content creators may be a blogger sharing her subject matter expertise, a father creating a PTA newsletter, or even a teacher authoring her own class material. These subject matter expert can author text quite fluently but may often find locating meaningful illustrations to be a painful task requiring significant time and effort. Thus, custom publications from non-professionals often lack the richness of illustration found in their professional counterparts. To find illustrations relevant to given long text reading content, a usual practice is to submit the entire text string as a query to an image search engine, like Bing Image. However, as existing search engines are designed to accept a few words as the query, the output from the search engine from a query string will be an error indicating that the query is *too long to process*. The other way is to manually summarize the long query passage to create a query consisting of a few words to find the relevant images. However, this approach is inefficient and may not accurately represent the query passage. Another key disadvantage with current image recommendation systems is that although there may be more than one topic underlying the long query content, existing search engines fail to consider this factor and treat all of these concepts as a single topic with which to find the relevant images. In addition, as search engines usually transform the query and candidate resources into bags or vectors of words, the semantic topics underlying the content are totally overlooked. Topics are a better choice for truly understanding both the query and the image illustrations.

To address these challenges, we created a novel system that recommending illustration for custom publishing by enabling search with text passages of any length and by recommending a ranked list of images that match the different topics covered within the queried passage. In summary, our system has these contributions:(1) Our system recommends images for text queries of any length. (2)Our system detects underlying topics from multi-topic content, then recommends illustration for each topic.(3) Our system introduces a novel semantic "visual words" based image ranking system. Using text content from the web page where the image originated from, we determine "visual words" as the semantic topic features with probabilistic topic modeling technique.

2. METHODOLOGIES

We address this problem by developing a semantic-aware image discover and insertion system for custom publishing. In a nutshell, given query text of any length, our system first detects the underlying topics from the text. We then recommend a list of images that are semantically relevant for each detected topic.

Query Topics Discovery: Given a text content in any length as a query, we utilize topic models to discover the abstract topics underlying the query. Intuitively, provided that a selected text content is about one or more topics, one expects particular words to appear in each topic more or less frequently. After generating the topics, each topic is represented by a set of words that frequently occur together. In this paper, we use Latent Dirichlet Allocation (LDA). We represent each topic with a set of terms to indicate the concept for each single topic. **Topic Compression:** As the number of topics to be generated is given as input to LDA, this number associated with the queried passage is unknown, it is possible that multiple topics are generated but about similar concepts. In order to remove such redundancy, we propose the idea of topic compression by considering the word distribution of each topic, and then remove duplicate topics if they are discussing the similar concept. To identify if two topics are about similar concepts, we use Pearson correlation in this paper. Then for each remaining query topic, we fetch the top K (K=40 in this paper) relevant images from Bing Image with the topic represent terms as query.

Directly use the images with the original ranking from Bing Image is not appropriate for publishing purpose. To illustration this, we provide an example in Figure 1. Where a query passage from a chemistry book "... *explain why fireworks different colors with the knowledge of fireworks chemistry* ...", with Bing Image API,

Copyright is held by the authors. RecSys 2015 Poster Proceedings, September 16-20, 2015, Austria, Vienna.

Figure 1: Images Directly from Search Engine is inappropriate

both two images in Figure 1 are returned. However, the left one is more appropriate for publishing as it is semantic related to the book content and has the illustration ability.

Distinguishing the most semantic relevant images from other ones is critical for publishing purpose, since it is important to embed the image that has the illustration or semantic content explaining value to the surrounding book content. To perform this, we use these images from Bing as candidates and re-rank them based on the semantic "visual words", where we compare the relevance between each query topic and surrounding content in the original page where candidate images originated from. **Visual Words Generator:** To generate "visual words", we combine content from original page containing the candidate images for each query topic with the corresponding represented query topic content as a bucket. For each bucket, we generate a set of topics as the semantic "visual words" with LDA[2] (No. of topics can be selected via cross validation. In this paper, we generate 50 topics for each content bucket). **Relevance Discovery and Ranking:** With the "visual words" representation matrix, where the row is the extracted query topic or candidate images. The column are the "visual words". We apply cosine similarity to measure the relevance in this paper[3], other distance methods also can be applied here[5][4]. Finally, we select the top n (n < K, n = 20 in experiment section) images to show for each query topic discovered from the query passage. Besides this, our system allows user to customize image preferences, including privacy, sources, formats, resolutions, size, etc.

3. EXPERIMENT

We have implemented the system, which is currently being piloted with multiple local universities and high schools[1]. While the pilot is ongoing, early feedbacks show that our system is seen favorably by the users. To show the effectiveness of our method, we randomly select 100 query passages from each of 6 testing books in varies grades and subjects. We consider 4 different ways to extract the key terms from the selected query (summarized in Table 1) and 2 ways to discover the relevant resources (summarized in Table 2). Consequently, we have implemented 8 scheme combinations from Tables 1 and 2.

Table 1: Key Terms Extraction Scheme S1 ∼ S4

Scheme	Words	Weighting
S1	words	Frequency-based weighting
S2	nouns	Frequency-based weighting
S3	noun phrases	Phrase Weighting
S4	topic words	topic word distribution

The words are extracted with the largest $tf * idf$ value from the selected passage in S1, nouns and noun phrases are identified using off the shelf POS tagger. Both words and nouns are weighted by Frequency-based weighting ($tf * idf$). The noun phrases are weighted by phrase weighting. The details of the weighting methods are provided as follow:

- *Frequency-based weighting*: For any term t_i, the frequency based weighting $f_t(t_i)$ is computed by using the $tf*idf$ weighting scheme widely used.
- *Phrase Weighting*: Let s be a phrase and $t_i \in s$ be a term contained in s. Then: $f_s(s) = f_t(s)\frac{\sum_{t_i \in s} f_i}{|s|}$. The phrase weighting $f_s(s)$ considers two factors: $f_t(s)$, the phrase TF*IDF score by treating each phrase as a single term, and the average frequency of all terms contained in the phrase $\sum_{t_i \in s} \frac{f_i}{|s|}$, where $|s|$ is the length of the phrase, i.e., the number of terms. The first factor considers the importance of the phrase as a whole unit $f_w(s)$, the second factor is the relevance to the document reflected by the average frequency of its contained terms.

To evaluate the performance of different query extraction schemes, we conducted a user study for the results . We select top 20 results to show with different relevance methods (T1∼T2) and key term schemes (S1∼S4) for each query passage. The results are manually judged to be appropriate or not. **Precision** is used as the evaluation metric. Stanford POS Tagger [1] is used to extract nouns and noun phrases. The noun phrases are extracted by regular expression (Adjective|Noun)*(Noun Preposition)?(Adjective|Noun)*Noun. We repeat the experiment with randomly selected passage queries from 6 testing books (B1∼B6). For each book, the average precisions of all passage queries with all 8 combinations (S1∼S4 with T1∼ T2) are calculated. Figure 2 plots the overall results.

Table 2: Relevance Discovery T1 ∼ T2

	Relevance
T1	Topic Features with Cosine Similarity
T2	Directly from Bing Image

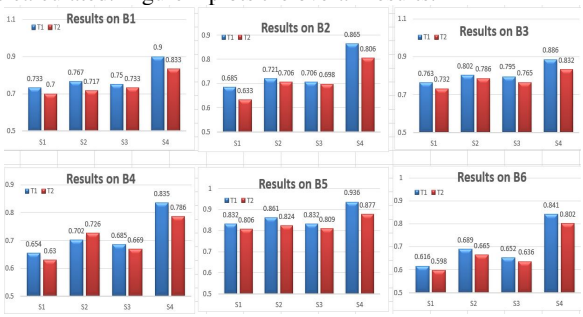

Figure 2: Results of the 8 combinations on B1∼B6

From the results, we have the following observations: (1) T1 is better than T2 with all key term extraction Schemes (S1∼S4) across all 6 books(B1∼B6), which justify our argument that topics underlying the content offers a better way to truly understand both passage query and candidate resources. (2) Nouns and Noun phrases achieve similar performance, and they are better than words selected without POS tagger. (3)Our method (S4 with T1) achieves the best performance for all the queries across 6 testing books.

4. REFERENCES

[1] J. M. Hailpern, R. Vernica, M. Bullock, U. Chatow, J. Fan, G. Koutrika, J. Liu, L. Liu, S. J. Simske, and S. Wu. To print or not to print: hybrid learning with METIS learning platform. In *ACM EICS 2015*, pages 206–215, 2015.

[2] G. Koutrika, L. Liu, and S. Simske. Generating reading orders over document collections. In *31st IEEE International Conference on Data Engineering, ICDE'2015*, pages 507–518, 2015.

[3] L. Liu, G. Koutrika, and S. Wu. Learningassistant: A novel learning resource recommendation system. In *International Conference on Data Engineering (ICDE'2015)*, 2015.

[4] L. Liu and P.-N. Tan. A framework for co-classification of articles and users in wikipedia. In *Web Intelligence'10*, pages 212–215, 2010.

[5] P. Mandayam Comar, L. Liu, S. Saha, A. Nucci, and P.-N. Tan. Weighted linear kernel with tree transformed features for malware detection. In *CIKM'12*, pages 2287–2290, 2012.

[1] http://nlp.stanford.edu/software/tagger.shtml