Diversifying Landmark Image Search Results by Learning Interested Views from Community Photos

Yuheng Ren[†*], Mo Yu[†*], Xin-Jing Wang[‡], Lei Zhang[‡], Wei-Ying Ma[‡]
[‡]Microsoft Research Asia, 49 Zhichun Road, Beijing, China
[†]Harbin Technology Institute, Harbin, 150001, P.R.China
[†]{yuhengren,yumo}@vilab.hit.edu.cn; [‡]{xjwang, leizhang, wyma}@microsoft.com

ABSTRACT
In this paper, we demonstrate a novel landmark photo search and browsing system, Agate, which ranks landmark image search results considering their relevance, diversity and quality. Agate learns from community photos the most interested aspects and related activities of a landmark, and generates adaptively a Table of Content (TOC) as a summary of the attractions to facilitate user browsing. Image search results are thus re-ranked with the TOC so as to ensure a quick overview of the attractions of the landmarks. A novel non-parametric TOC generation and re-ranking algorithm, MoM-DPM Sets, is proposed as the key technology of Agate. Experimental results based on human evaluation show the effectiveness of our model and user preference for Agate.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, clustering.* G.3 [**Mathematics of Computing**]: Probability and Statistics – *nonparametric statistics.* H.5.3 [**Information Interface and Presentation**]: Group and Organization Interfaces – o*rganizational design, Web-based interaction.*

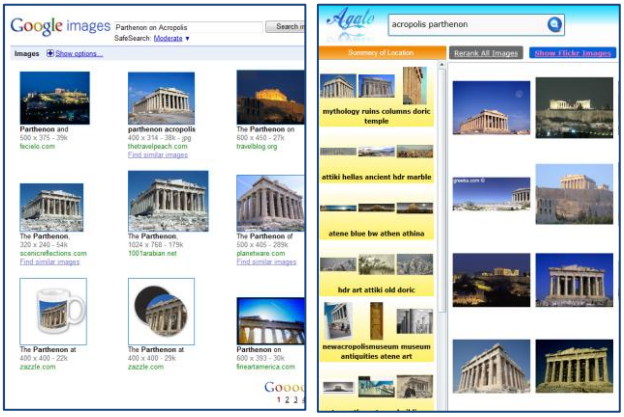
General Terms
Algorithms, Performance, Human Factors

Keywords
User interest modeling, set-based ranking, landmark image search.

1. INTRODUCTION
Online travel services have become increasingly important in user experience. The pre-trip behavior of a typical user is first to seek for inspiration or destination guidance, collect information, do research and comparison, and then plan and book their trip [4]. In this loop, existing search engines are generally serving as a hub to help redirect users to travel agent sites. We believe that search engines should be able to contribute more. Imagine that the online photo search of a tourist resort is so advanced that it presents the user a comprehensive overview of the attractions of this place, e.g., when the user searches "Bora-Bora island", beach, underwater activities, and revelry, etc. are outlined. Obviously image search engines will become competitive travel guidance agents in the online travel market.

* The work was performed at Microsoft Research Asia.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.



(a) (b)

Figure 1. Google (left) and Agate (right) results of "Parthenon on Acropolis"

Unfortunately, existing image search engines are far from satisfying such user needs: they return unorganized list of images based on keyword matching, which would decelerate user's convergence to the attractive aspects of a landmark. Meanwhile, text-based image retrieval suffers from word ambiguity and visual redundancy. For example, Figure 1(a) shows the Google image search results of "Parthenon on Acropolis"; redundant images which differ only in size and irrelevant images like cups and cup pads were ranked high.

In this work, we demonstrate a new landmark search and browsing system, Agate, as an attempt to sketch the attractions of a place and deliver relevant, diverse and high-quality photo search results. As shown in Figure 1(b), Agate automatically generates a table of content (TOC) as a navigation panel on the left, which summarizes the attractive aspects of Acropolis, and rendered search results with improved relevance and diversity.

Similar ideas of generating TOC were addressed by Wang et al. [8] and the IGroup system [3]. Both these work group web images to improve the search result. Wang et al. [8] made use of the visual context of an image while IGroup identifies salient key phrases from image surrounding texts. They attempted to structuralize image search results to facilitate user browsing, but neither of them identified the most interested aspects, nor did they re-rank image search results to punish irrelevance and encourage diversity.

The knowledge of the most attractive views of a landmark, however, is embedded in user-generated content (UGC), e.g. the votes that a Flickr image obtains suggest to certain extent the corresponding view's attractiveness. In this study, we propose a novel algorithm called Multimodal Dirichlet Process Mixture Sets (MoM-DPM Sets) to learn such knowledge from UGC data, and

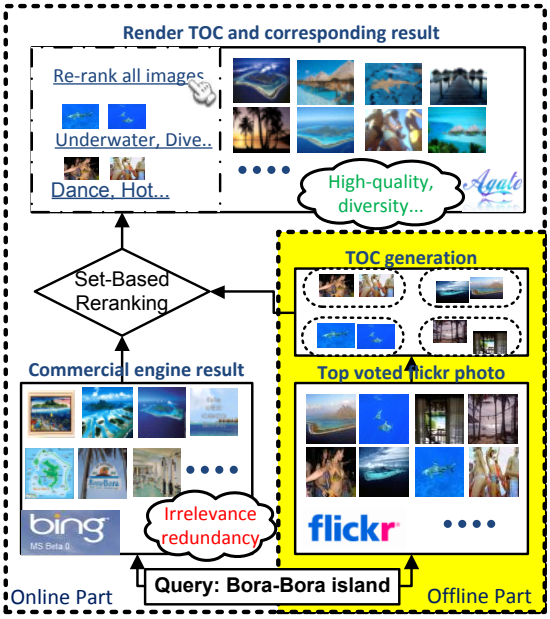


Figure 2. Sketch of Agate's framework: 1) collect top-voted Flickr photos, 2) image auto-group to generate TOC, 3) re-rank the image search results of commercial engine, and 4) deliver the TOC and ranking results in the UI.

then apply the knowledge to re-rank image search results accordingly.

2. THE AGATE SYSTEM

2.1 The Framework

As shown in Figure 2, Agate consists of three parts: 1) collecting top-voted Flickr images, 2) generating TOC, and 3) re-ranking image search results.

Given a location name, we collect the most interested images from Flickr.com using its "search by interestingness" service. After filtering out stop words (including general stop words and the query location name), extracting visual features (color sift descriptor) [6] and textual features (word occurrence), we group these images into clusters and then assign a name for each cluster. We use the clusters to re-rank Bing image search results and present the output to the user.

The UI of Agate is shown in Figure 1(b). On the left is a TOC navigation panel and on the right shows the image search result. Given a location query, the TOC panel shows the thumbnails along with cluster names about the corresponding attractions, ranked in descending order of their importance, and the search result panel shows the image search results re-ranked against the attractions, which have improved relevance, diversity and quality.

The user can also select to browse images of a certain view by clicking on the corresponding thumbnails. Image search results will then be ranked against this certain category so that the less interested aspects of the query location will be ranked lower.

2.2 Top-voted Flickr Image Collection

Mining from UGC data has enabled many interesting research on computer vision nowadays. However, most of them use only images and their tags; there are still many useful metadata that have not been fully taken advantages of.

In this work, we propose to leverage the metadata of user votes associated with Flickr images to discover the most interested aspects of a landmark. A user votes for an image due to various reasons, e.g. it is a professional shot; it catches the most famous view, or some noisy reasons such as friendship. However, it is reasonable to assume that the majority of user votes will favor the high-quality and attractive contents [5]. This is the intuition behind our TOC generation technique.

There are many measurements of interestingness of an image. In our current approach, we simply used the "interestingness" property provided by Flickr since the Flickr ratings should have considered user preference and is thus more trustable. Given a location name, we query the Flickr search service and sort the image results in their descending order of "interestingness". We crawl at most top 500 images as well as their user-submitted tags for TOC generation.

2.3 TOC Generation and Image Re-Ranking

A natural way to generate TOC from a photo collection is clustering [8][3][5]. Then we re-rank the original image search results against these clusters.

There are three major challenges: 1) different landmarks will have different number of attractive aspects, and how to determine the number adaptively is challenging; 2) both visual and textual features are valuable to ensure the clustering effectiveness, while how to fuse these heterogeneous features is still an open research topic; and 3) how to provide a unified measurement for both clustering and re-ranking.

Recent research achievements on topic modeling suggest a promising solution for feature fusion [7][2] [1], while the Dirichlet Process (DP) technique [7] gives a solution to automatically determine the number of clusters. Combining these two, the Multi-Modal Dirichlet Process Mixture model (MoM-DPM) [1] is able to address the former two challenges mentioned above. However, as to our knowledge, there are no such previous works which are able to simultaneously solve all the three challenges with a single model.

2.3.1 The MoM-DPM Sets model

We propose the Multi-Modal Dirichlet Process Mixture Sets algorithm (MoM-DPM Sets) to fill in the vacancy. Firstly, it is a DP mixture model which generates adaptively a number of latent topics, each indexes a cluster. Secondly, it is multimodal which conditionally independently generates visual and textual representations of an image given a topic. Thirdly, it formulates the re-ranking step as a set-based Bayesian inference problem. Rather than learning a single (optimal) parameter set from training data and measuring a new image against the model as previous works did [1][2], MoM-DPM Sets identifies which images should be in one cluster, and measures a new image against the images in a cluster given all possible distributions of model parameters. Note that since all parameter distributions are taken into consideration in MoM-DPM Sets, it gives a truly Bayesian inference, which is a fundamental theoretic difference to the previous models. This formulation provides a great capability to summarize community images and re-rank web images (in a different domain) in a fundamental way.

We represent an image both by (a) a bag-of-visual-words v representing the visual features and (b) a bag-of-terms t generated from its surrounding tags. Let θ^v be the multinomial distribution over visual words with a Dirichlet prior H^v, and θ^t be the Ber-

noulli distribution over terms with the Beta prior $H^t = \{\boldsymbol{\gamma}^0, \boldsymbol{\gamma}^1\}$. We use Bernoulli distribution for terms because generally unique Flickr tag appears only once for an image.

The generative process of MoM-DPM Sets is shown in Table 1. Let v_i, t_i, z_i represent the visual features, textual features, and cluster label of the i-th image respectively. Let $\boldsymbol{z}_{\setminus i}$ be the cluster labels of all the observed images with the i-th image removed, and V_z, T_z be the visual and textual features of images in a certain cluster z. Let ϕ^v and ϕ^t be the parameter set corresponding to visual and textual features respectively, the model is to learn the probability $p(z_i = z|v_i, t_i, V_z, T_z, \phi^v, \phi^t)$. Note that the key difference of this model from the previous work [1] lies in the existence of V_z, T_z. This is the key of set-based Bayesian inference.

We solve this model with Gibbs sampling, as below:

For an existing (active) topic $z_i = z \in \{1, .. K\}$

$$p(z_i = z|v_i, t_i, V_z, T_z, \phi^v, \phi^t) \propto \frac{n_{\setminus i}^z}{n - 1 + \alpha} p(v_i|z_i = z, V_z, \phi^v) p(t_i|z_i = z, T_z, \phi^t) \quad (1)$$

And for a new (inactive) topic, $z_i = z, for\ \forall z_j \in \{1, .., K\}, z \neq z_j$

$$p(z_i = z|v_i, t_i, \alpha, \phi^v, \phi^t) \propto \frac{\alpha}{n - 1 + \alpha} p(v_i = v|H^v) p(t_i = t|H^t) \quad (2)$$

Where K is number of existing (active) clusters in current iteration, $n_{\setminus i}^z$ is the number of images (except the i-th) labeled by topic z. This was in the same form of algorithm 3 presented by Neal et al. [7], while our approach can be regarded as its multi-modal extension.

Taking the specific multinomial and Bernoulli distribution into the form, we have:

$$\begin{aligned} p\left(v_i \middle| z_i = z, V_{z_{\setminus i}}, \phi^v\right) &= \sum_{\theta^v_{z,\setminus i}} Mul\left(v_i = v|\theta^v_{z,\setminus i}\right) Dir\left(\theta^v_{z,\setminus i} \middle| V_{z_{\setminus i}}, H^v\right) \\ &= \frac{\Gamma\left(\sum_k (H_k^v + n_{k,\setminus i}^{v,z})\right)}{\Pi_k \Gamma\left((H_k^v + n_{k,\setminus i}^{v,z})\right)} \frac{\Pi_k \Gamma\left(H_k^v + n_{k,\setminus i}^{v,z} + v_k\right)}{\Gamma\left(\sum_k (H_k^v + n_{k,\setminus i}^{v,z} + v_k)\right)} \end{aligned} \quad (3)$$

$$\begin{aligned} p\left(t_i = t \middle| z_i = z, T_{z_{\setminus i}}, \phi^t\right) &= \sum_{\theta^t_{z,\setminus i}} Ber\left(t_i = t|\theta^t_{z,\setminus i}\right) Beta\left(\theta^t_{z,\setminus i} \middle| T_{z_{\setminus i}}, H^t\right) \\ &= \prod\nolimits_k \frac{\left(\gamma_k^1 + n_{k,\setminus i}^{t,z}\right)^{t_k} (\gamma_k^0 + \sum_j n_{j,\setminus i}^{t,z} - n_{k,\setminus i}^{t,z})^{(1-t_k)}}{(\gamma_k^1 + \gamma_k^0 + \sum_j n_{j,\setminus i}^{t,z})} \end{aligned} \quad (4)$$

where v_k denotes the count of the k-th visual word in the query image, and $t_k = 1$ means that term k appears in the query image's tag list and $t_k = 0$ otherwise. $n_{k,\setminus i}^{v,z}$ and $n_{j,\setminus i}^{t,z}$ indicate the number of images with topic z in their visual and textual appearances respectively.

In our evaluation, the Gibbs sampling generally converges in about 30 iterations. And then we save the learnt clusters for the re-ranking step.

2.3.2 Cluster name generation

In order to display the TOC categories in plain sight, we randomly select three images in each cluster and show them in the navigation panel. Meanwhile, we assign a name to each cluster to make the semantics clearer.

Generative Process of MoM-DPM Sets

- draw $\boldsymbol{\pi} \sim GEM(\alpha)$ using a stick-breaking process
- for each image $i = 1, 2, \ldots, N$
  - draw a topic $z_i \sim Discrete(\boldsymbol{\pi})$
  - if z_i not exist in $\mathbf{z}$
    - draw a multinomial distribution over visual words, $\theta^v_{z_i} \sim Dir(H^v)$
    - draw a Bernoulli distribution over terms, $\theta^t_{z_i} \sim Beta(H^t)$
  - draw $v \sim Mult(\theta^v_{z_i})$ using Eq.(3)
  - draw $t \sim Ber(\theta^t_{z_i})$ using Eq.(4)

Table1: The generative process of MoM-DPM sets

The clustering effectiveness is ensured in two aspects: 1) it is applied onto top-voted search results, so that the image collection used to learn the clusters is comparatively clear; 2) the MoM-DPM Sets is effective in clustering. Therefore, a simple key phrase extractor is able to produce representative cluster names. We observed that after filtering stop words, meaningful tags which describe the attractions have high frequency within certain clusters but have relatively low frequency over all clusters. For example, in the cluster of night views of the Golden Gate Bridge, we observed that those representative words like "night", "light" have relatively high frequency, while those noisy keywords such as "awesome" and "Nikon" are fairly common among all clusters. So we adopt the TF-IDF measure to score the associated words of images in a cluster. Concretely, we collect all the words in a cluster as one single document, and compute the TF-IDF score for each word among all clusters. Then for each cluster, we sort their words in the descending order of their TF-IDF score and the top-ranked four words are selected as the cluster name. We observed from our demo that this simple approach is fairly effective.

2.3.3 Re-ranking image search results

Our demo supports two types of re-ranking: 1) re-ranking the image search results against all learnt clusters, and 2) re-ranking against a certain cluster. This is helpful when the user is only interested in a certain attractive view.

We used the Bing image search engine for our evaluation. In the case of re-ranking against all clusters, a new image is scored by Eq. (5):

$$score(v^q, t^q) = \max_{z \in \boldsymbol{z}} p(z|v^q, t^q, \boldsymbol{z}, V_z, T_z, \phi^v, \phi^t) \propto \max_{z \in \boldsymbol{z}} \{\frac{n_z}{n + \alpha} p(v^q|V_z, \phi^v) p(t^q|T_z, \phi^t)\} \quad (5)$$

Where v^q, t^q represent the features of the query image. n is the number of images in all the learnt clusters and n_z is cluster size of a topic z, and $\boldsymbol{z}$ represents all the available topics.

In the case of re-ranking against one single cluster, we adopt Eq. (6):

$$score(v^q, t^q) = p(z|v^q, t^q, \boldsymbol{z}, V_z, T_z, \phi^v, \phi^t) \quad (6)$$

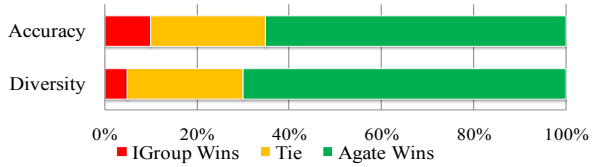


Figure 3. Percentage of queries wined by Agate and IGroup on TOC quality

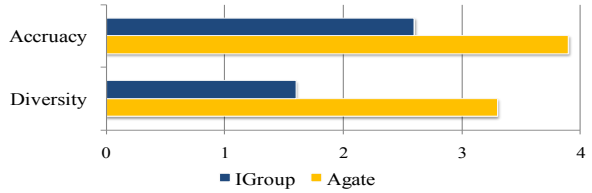


Figure 4. Average user scoring of TOC quality

3. EVALUATIONS

Since there is no benchmark dataset or ground truth data available, we asked ten volunteers to manually evaluate the system. The participants were asked to act as travelers searching for information about their destination landmarks.

3.1 Quality of Generated TOC

The first session of user study is to compare the outputs of Agate and IGroup[3], a state of the art image grouping system, on the effectiveness of TOC generation and categorized search. In this session, twenty landmarks randomly selected from the Wikipedia list of landmarks[1] formed the testing query set.

The TOC outputs of both IGroup[3] and Agate are presented to participants simultaneously, but which results came from which system was kept blind to the labeler. They need to assign a score between one to five to measure 1) diversity, which evaluates whether the learnt categories contain diverse aspects of a landmark, and 2) accuracy, which measures whether the member images are relevant to the TOC concept. Figure shows the evaluation result. Most of participants agreed d that on 70% of the queries, Agate outperformed IGroup [3] in diversity, and on 65% of the queries, Agate won in accuracy. IGroup [3] was superior just on 5% of queries in diversity and 10% of queries in accuracy, while for the rest of queries, they tied. Figure shows the average scores of the two systems. We can see that Agate greatly outperformed IGroup [3] both in diversity and in satisfaction.

3.2 Effectiveness of Image Re-Ranking

The second session of user study is to measure whether Agate can improve the quality of image search results. We measure the quality with the following factors: 1) relevance, whether the number of irrelevant images is reduced; 2) diversity, whether the search results cover diverse topics about the query landmark; 3) image quality, whether low-resolution images will be ranked lower.

We used Bing as our baseline and Bing image search results as the resource of new images. Again all participants were asked to score one to five to each criterion above; the larger score the better. Meanwhile, we asked them to give an "overall" score to indicate their overall impression of the search results. The performance was tested on forty randomly selected landmark queries.

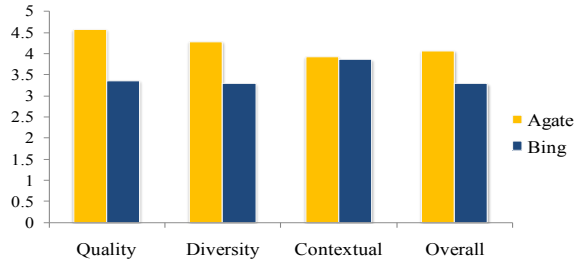


Figure 5. Image search performance of Agate vs. Bing.

Figure illustrates the result. It can be seen that the re-ranked sets outperformed original Bing image search results on all three criteria. And the overall impression of Agate greatly surpassed Bing. According to our close observation, if the score of a criterion is larger than four, then the search result volunteers would describe the search engine as "effective". And as it could be seen in the figure, Agate was scored higher than four in all criteria.

4. CONCLUSION

In this paper, we propose Agate, a novel landmark image search and browsing system. Agate attempts to enrich an image search engine with the service of travel guidance, via which the user obtains a comprehensive understanding of the attractions of a location. Agate seeks such knowledge from community photos, and then applies it to re-rank commercial image search results. A MoM-DPM Sets model was proposed as the key technology underlying Agate, which determines the number of attractive aspects adaptively, fuses visual and textual features, and unifies the clustering and ranking steps. A friendlier user interface was designed to facilitate user browsing and help her quickly discover the interested photos. Comprehensive user studies showed the effectiveness and the superiority of Agate to existing systems.

5. REFERENCES

[1] A. Velivelli and T.S. Huang. Automatic Video Annotation Using Multimodal Dirichlet Process Mixture Model. *ICNSC 2008.*

[2] B. David and M. Jordan. Modeling annotated data. *ACM SIGIR conference*, pp. 127–134, 2003.

[3] F. Jing, C. Wang, Y. Yao, K. Dong, L. Zhang and W. Ma. IGroup: Web Image Search Results Clustering. *ACM Multimedia,* Oct., 2006.

[4] J. Butcher. Travel 'Experiences': The Ancillary Product Context. *WTM Seminar of Isango.com*. Nov. 2008.

[5] I. Simon, N. Snavely et al. Scene Summarization for Online Image Collections. *ICCV*, 2007.

[6] K. Sande, T. Gevers and C. Snoek. Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Trans. On PAMI* (in press), 2010

[7] R.M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational And Graphical Statistics*, vol.9(2):249-265,2000

[8] X. Wang, W. Ma, Q. He and X. Li. Grouping Web Image Search Result. In *Proceeding of the 12th International ACM Conference on Multimedia,* Oct. 2004.

[1] http://en.wikipedia.org/wiki/List_of_landmarks