Summarizing Social Image Search Results

Boon-Siew Seah, Sourav S Bhowmick, and Aixin Sun
School of Computer Engineering, Nanyang Technological University, Singapore
seah0097@ntu.edu.sg, assourav@ntu.edu.sg, axsun@ntu.edu.sg

ABSTRACT

Most existing social image search engines present search results as a ranked list of images, which cannot be consumed by users in a natural and intuitive manner. Here, we present a novel algorithm that exploits both visual features and tags of the search results to generate high quality image search result *summary*. The summary not only breaks the results into *visually* and *semantically coherent* clusters, but it also maximizes the *coverage* of the original search results. We demonstrate the effectiveness of our method against state-of-the-art image summarization and clustering algorithms.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*

Keywords

Image Search Summarization; Tag-based Image Search; Flickr

1. INTRODUCTION

Image search results are typically presented as a ranked list of images often in the form of thumbnails. Such thumbnail view of ranked images enables end users to quickly glance through a set of images. However, it suffers from two key limitations. First, it fails to provide a view of common visual objects or scenes *collectively*. For example, the result images of "fly" query can be clustered by visual objects (*e.g.,* aeroplane, insect) and activities (*e.g.,* jump). Organized image search results naturally enable a user to quickly identify and zoom into a subset of results that is most relevant to her query intent. Second, a thumbnail view fails to provide a bird eye view of different concepts in a query results. It will be beneficial to users if a suitable exemplar image from each concept can be selected to create a "summary" of the search results. Here, we take a systematic step towards addressing these limitations associated with social image search results. A social image refer to an image contributed to image sharing platform (*e.g.,* Flickr), which is often annotated with tag(s).

An appealing way to organize social image search results is to generate a set of *image clusters* from them, such that images in each cluster are *semantically and visually coherent*, and the clusters *maximally cover* the entire result set. Subsequently, at least one exemplar image from each cluster can be selected to generate an *exemplar summary* of the entire result set to give a bird eye view of different concepts in it. We advocate that such image clusters must satisfy the following desirable features: 1) *Concept-preserving* – each cluster should be annotated by a *minimal* set of tags generated from the images within to semantically describe *all* images in the cluster. Users therefore can easily associate the tag(s) with the images in a cluster at a glance; 2) *Visually coherent* – visually similar images must be clustered together and dissimilar images must be separated in different clusters; 3) *High coverage* – the image clusters should cover as much of the result set as possible in order to maximize incorporation of all possible query intent.

Recently, *early fusion* [6] and *late fusion* [4] approaches have attempted to summarize image search results. The former exploits the tags and visual content of the images jointly whereas the latter considers them independently. However, these techniques do not ensure that the generated summaries are concept-preserving and maximally cover the image results; instead they embody many-to-many association between a set of images and a "soup" of tags. As the tag-image associations are disrupted, the set of tags may not completely represent all images in a cluster. Even for image categorization techniques provided by Web image search engines (*e.g.,* Google and Bing),[1] where data associated with images are not as sparse as social images, there is little evidence whether they maximally cover the result set. Here, we propose a novel approach that models social image search result summarization as a weighted k-set cover problem that maximizes the above desirable features.

2. PROBLEM FORMULATION

Let Q be a search query with one or more tags and $\mathcal{D}$ be its search result images. Each $i \in \mathcal{D}$ has a d-dimensional visual feature vector and a set of tags T_i. The image-image visual similarities $\mathcal{D}$ is represented as a graph $G = (V, E, w)$, with $w : E \rightarrow \mathbb{R}$ indicating the degree of visual similarity between images.

The key intuition of social image search results summarization is to optimally decompose G into a set of *concept subgraphs* from which exemplar images are drawn to create the summary. A concept subgraph $C_T = (V_T, E_T, T)$ is a subgraph of G such that every image in the subgraph must share the set of tags T. Consequently, C_T can be represented by an *exemplar node* of 1-to-3 representative images labeled with T. More specifically, a summary decomposes G into a set of concept subgraphs $\mathcal{S} = \{C_{T^1}, C_{T^2}, \ldots C_{T^k}\}$ and a *remainder* subgraph R (containing images not in $\mathcal{S}$). We consider 3 properties of $\mathcal{S}$ that help identify a desirable decomposition:

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://http://dx.doi.org/10.1145/2567948.2577296.

[1] Google: `http://images.google.com` Bing: `http://www.bing.com/images`

- $visual\_coherence(\mathcal{S}) = \frac{1}{|\mathcal{S}|}\sum_{C_T\in\mathcal{S}}\frac{\sum_{e\in E_T} w(e)}{|E_T|}$ reflects the average visual similarity of images in each $C_T \in \mathcal{S}$.
- $distinctiveness(\mathcal{S}) = \frac{\left|\bigcup_{C_T\in\mathcal{S}} V_T\right|}{\sum_{C_T\in\mathcal{S}}|V_T|}$ captures the degree of summary redundancy. A decomposition that creates clean separation of concept subgraphs is desirable.
- $coverage(\mathcal{S}) = \frac{\left|\bigcup_{C_T\in\mathcal{S}} V_T\right|}{|V|}$ quantifies the fraction of images in V that also appears in $\mathcal{S}$.

Given $\mathcal{Q}, \mathcal{D}$ and G, the goal of ***social image search results summarization*** is to find an optimal set of concept subgraphs $\mathcal{S}$ such that $visual\_coherence(\mathcal{S})$, $coverage(\mathcal{S})$ and $distinctiveness(\mathcal{S})$ are maximized. The exemplar summary $\mathcal{M}$ is constructed by mapping each $C_T \in \mathcal{S}$ into an exemplar node.

To solve the above problem, we propose to decompose G into $\mathcal{S}\cup\mathcal{R}$ using a weighted minimum k-set cover optimization model [2]. It incurs a weight (*i.e.,* cost) for each C_T or R in $\mathcal{S}\cup\mathcal{R}$. For C_T, it incurs a *visual incoherence cost* (maximize $visual\_coherence(\mathcal{S})$). For R, it incurs a *remainder penalty* cost (maximize $coverage(\mathcal{S})$). We find the minimum weight of $\mathcal{S}\cup\mathcal{R}$ needed to cover V (controlling $distinctiveness(\mathcal{S})$).

Given G, let $\mathcal{E}$ be the family of all concept subgraphs of G and $\mathcal{F}$ be the family of all subgraphs of G. Let k be the cardinality constraint. The optimal $\mathcal{S}\cup\mathcal{R}$, where $\mathcal{S}\subset\mathcal{E}$ (set of concept subgraphs) and $\mathcal{R}\subset\mathcal{F}$ (set of remainder subgraphs), is the minimum cost set that covers V:

$$\arg\min_{\mathcal{S}\cup\mathcal{R}} f(\mathcal{S}\cup\mathcal{R}) = \arg\min_{\mathcal{S}\cup\mathcal{R}} \sum_{C_T\in\mathcal{S}} c(C_T) + \sum_{R\in\mathcal{R}} r(R)$$

subject to $V = \left(\bigcup_{C_T\in\mathcal{S}} V_T\right) \cup \left(\bigcup_{V_R\in\mathcal{R}} V_R\right)$ and $|\mathcal{S}| + |\mathcal{R}| \le k$, where the *visual incoherence* cost function $c : \mathcal{E} \to \mathbb{R}$ and the *remainder penalty* cost function $r : \mathcal{F} \to \mathbb{R}$ are defined as follows:

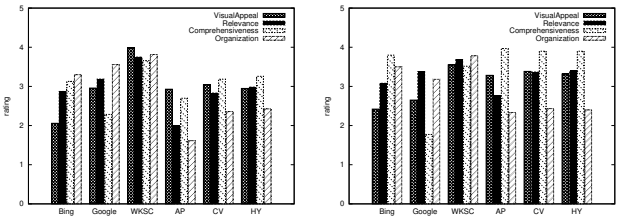
$$c(C_T) = \frac{|E_T|}{\sum_{e\in E_T} w(e)} \qquad r(R) = (|V_R| + 1)\max_{C_T\in\mathcal{E}} c(C_T)$$

It can be proven that an optimal solution of the problem is a set of concept subgraphs $\mathcal{S}$ and at most a single remainder subgraph R, and the remainder subgraph does not overlap with $\mathcal{S}$.

3. ALGORITHM

Because the weighted k-set cover problem is NP-hard, we present a greedy heuristic solution [2]. It consists of five key phases:

1. Given $\mathcal{D}$, we use image-image visual cosine similarity as the edge weights of G.
2. This phase enumerates $\mathcal{E}$ from G (approximately). We construct a directed acyclic graph to allow structured enumeration of concept subgraphs (forming $\mathcal{E}^*$). Every non-root node represents a concept subgraph. Let $C_T^0 = \left(V, E, T^0 = \emptyset\right)$ be the root node at depth $i = 0$. Given C_T^i, we construct $C_T^{i+1} = \left(V_T^{i+1}, E_T^{i+1}, T^{i+1}\right)$ satisfying the following: 1) $T^{i+1} = T^i \cup \{t'\}$, where t' is one additional tag. 2) V_T^{i+1} is the set of all images in V_T^i sharing T^{i+1} and $V_T^{i+1} \neq V_T^i$, 3) C_T^{i+1} induced by V_T^{i+1} has at least one edge.
3. We now find a subset $\mathcal{S}\subset\mathcal{E}^*$ and $\mathcal{R}\subset\mathcal{F}$ that optimally decomposes G, by adopting a H_k-approximation greedy algorithm following [2]. The basic idea is to select, at each iteration, C_T with least cost greedily and let the last iteration be R.
4. Starting with $\mathcal{S}$, this phase aggregates concept subgraphs iteratively to form summaries at reduced level of detail. The successor $\mathcal{S}^{i+1}$ of $\mathcal{S}^i$ is formed by *contracting* pair of concept subgraphs. The contraction of pairs C_{T^1} and C_{T^2} removes both subgraphs and replaces them with $C_{T^1\cup T^2} = \left(V_{T^1}\cup V_{T^2}, E_{T^1}\cup E_{T^2}\right)$. At each iteration, we contract the pair with most concept similarity.
5. Given the most compressed $\mathcal{S}^i$, the final phase involves selection of one to three exemplar images from each concept subgraph to form an exemplar summary.



(a) Single-tag queries (b) Multi-tag queries

Figure 1: Comparative Evaluation.

4. EXPERIMENTS

Experiments were conducted on the NUS-WIDE dataset containing 269,648 *Flickr* images [1]. We selected 30 representative queries of a variety of abstract (*e.g.,* cute, fly) and concrete (*e.g.,* asia, animal) subjects for our study. For each query, the 1000 top-ranked images result form $\mathcal{D}$. We compared our method (WKSC) with the following baseline clustering/summarization techniques: Canonical View Summarization (CV) [5], Affinity Propagation (AP) [3] and H^2MP (HY) [6]. We also compared visual summaries constructed by *Google Images* (Categories) and *Bing Images* (Related Topics) for the same query. Total 12 volunteers were engaged to rate quality of the summaries.[2] Summaries generated by the algorithms are presented as a set of exemplars but without the names of the specific algorithms producing the summaries. A human assessor rates (from 1 for most unsatisfactory to 5 for most satisfactory) the summaries based on four aspects: *visual appeal*, *relevance*, *comprehensiveness* and *organization*. Figure 1 shows the results of the user study for single- and multi-tag queries. The rating for each question-algorithm pair is the average rating from multiple queries chosen by the assessors. The results clearly demonstrate the superiority of our method as assessors consider its summaries to be easiest to interpret, comprehensive, most conceptually relevant, and visually appealing. This underlines the importance of having concept preservation to obtain precise clusters.

In summary, we present an algorithm that meets three desirable features of a good social image search results summary: concept-preservation, visual coherence and coverage. Our empirical study demonstrated its superiority over existing techniques.

5. REFERENCES

[1] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from National University of Singapore. In *ACM CIVR*, pages 48:1–48:9, 2009.
[2] V. Chvatal. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.
[3] B. J Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.
[4] P.-A. Moëllic, J.-E. Haugeard, and G. Pitel. Image clustering based on a shared nearest neighbors approach for tagged collections. In *ACM CIVR*, pages 269–278, 2008.
[5] I. Simon, N. Snavely, and S. M. Seitz. Scene Summarization for Online Image Collections. In *IEEE ICCV*, pages 1–8, 2007.
[6] H. Xu, J. Wang, X.-S. Hua, and S. Li. Hybrid image summarization. In *ACM Multimedia*, pages 1217–1220, 2011.

[2] None of the volunteers are authors of this paper.