*i*AVATAR: An Interactive Tool for Finding and Visualizing Visual-Representative Tags in Image Search

Aixin Sun Sourav S Bhowmick Yao Liu

School of Computer Engineering, Nanyang Technological University, Singapore

axsun|assourav@ntu.edu.sg

ABSTRACT

Tags associated with social images are valuable information source for superior image search and retrieval experiences. Due to the nature of tagging, many tags associated with images are not visually descriptive. Consequently, presence of these noisy tags may reduce the effectiveness of tags' role in image retrieval. To address this problem, we demonstrate *i*AVATAR (interActive VisuAl-representative TAgs Relationship) system that uses the notion of *Normalized Image Tag Clarity* (NITC) to find *visual-representative tags*. A visual-representative tag effectively describes the visual content of the images. Further, we visually demonstrate relationships between *popular* tags and visual-representative tags as well as *co-occurrence* likelihood of a pair of tags associated with a *search tag* or image using *tag relationship graph* (TRG). We demonstrate various innovative features of *i*AVATAR with a real-world dataset and show that it enriches users' understanding of various important tag features during image search.

1. INTRODUCTION

With the advances in digital photography and social media sharing web services, a huge number of multimedia content is now available online. Most of these services enable users to annotate images with free tags (e.g., `aircraft`, `lake`, `sky`). A key consequence of the availability of such tags as meta-data is that it has significantly facilitated web image search and organization as this rich collection of tags provides more information than we can possibly extract from content-based algorithms. However, it has been widely recognized that realizing a tag-based image retrieval system is technically challenging due to noisy and imprecise nature of tags [2]. Two similar images may be associated with significantly different sets of tags from different users. Further, tags associated with an image may describe the image from significantly different perspectives. For example, consider a photo uploaded by Sally which she took using her Canon camera at Sentosa when she traveled to Singapore in 2009. This image may be annotated by tags such as `Canon`, `2009`, `Singapore`, `beach`, `sentosa`, and many others. Notice that some of the tags (e.g., `2009` and `Canon`) do not effectively describe the visual content of the image. Consequently, presence of these noisy tags may reduce the effectiveness of tags' role in image retrieval. Needless to say that "de-noising" tags has been recently identified as one of the key research challenges in [2].

In this demonstration, we present a novel graphical *noisy tag-aware* social images retrieval system, called *i*AVATAR (**inter**Active **V**isu**A**l-representative **TA**gs **R**elationship), that takes a concrete step to address the above challenge. Given a search tag t as input, *i*AVATAR retrieves a ranked list of images, denoted by T, that is annotated with t in the image database. A key feature of this system is that for t (resp. for each image $d \in T$) it identifies a set of *visual-representative* tags [8] related to t (resp. d) and how these tags are *associated* with other related tags using a color-coded *tag relationship graph* (TRG). Each tag is a labeled colored node in the TRG where the font size of the label and the color intensity of the node are proportional to the *tag frequency* and *visual-representativeness*, respectively. A pair of nodes is connected by a labeled edge if the corresponding tag pair co-occur together among images in the dataset beyond certain threshold.

Intuitively, a tag is *visual-representative* if it effectively describes the *visual content* of the images. A visual-representative tag (e.g., `sky`, `sunset`) easily suggests the scene an image may describe even before the image is presented to a user. On the other hand, tags like `2009` and `Asia` often fail to suggest anything meaningful with respect to the visual content of the annotated image. Clearly, identification of visual-representative tags from all tags assigned to images enables end-users to eliminate noisy tags. Further, when a user selects a visual-representative tag t_v in the TRG, *i*AVATAR retrieves a fraction of images associated with t_v to validate that it indeed describes the visual content of the images.

Additionally, *i*AVATAR supports two interactive graphical features: the *filtering mechanism* and the *difference viewer*. The *filtering mechanism* enables users to filter or expand the TRG to view different sets of tags associated with t (resp. d) and their relationships based on different *threshold* values for *visual-representativeness*, *tag frequency*, and HOP *distance*. The *difference viewer* provides a graphical view of the effects of different types of tag co-occurrence measures (e.g., cosine, Jaccard coefficient, KL divergence) on the tag relationships in the TRG. Such interactive features of *i*AVATAR pave way to superior image retrieval experience as they not only enrich users with the knowledge of noisy or non-noisy tags but also provides an in-depth understanding of the relationships between tags associated with the retrieved images.

2. RELATED SYSTEMS AND NOVELTY

To the best of our knowledge, this is the first work that uses visual-representativeness as a new dimension to better understand tag properties. The proposed finding of visually-representative tags

Proceedings of the VLDB Endowment, Vol. 3, No. 2
Copyright 2010 VLDB Endowment 2150-8097/10/09... $ 10.00.

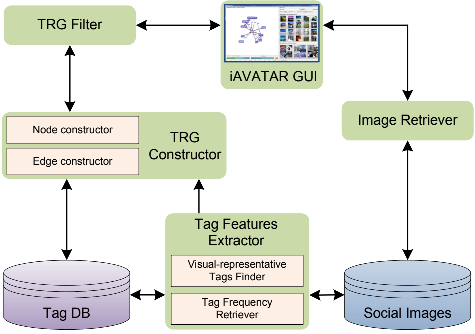


Figure 1: Architecture of *i*AVATAR.

exploits the techniques in the area of query performance prediction in Web search [3]. A query is unambiguous if all its matched documents are topically cohesive; analogously, a tag is visually-representative if its associated images are visually cohesive.

Tag cloud is the most widely adopted tag visualization technique. In a tag cloud, the tags are often ordered in alphabetical order and the font sizes are proportional to their frequencies. Other than frequency, tags have also been visualized by their spatial/temporal aspects. *Map-based tag cloud* visualizes tags on top of a map interface with geo-referenced social images [1]. The temporal evolution of tags within the Flickr is visualized in [4] where the font size of the tag is proportional to its interestingness derived from the frequency evolution of the tag along the timeline. To visualize the relationships between tags, [6] displays tag clouds on top of a topographical image so that related tags are closer to each other. However, none of the existing approaches explore relationships between tags by visual-representativeness, frequency, and co-occurrence, coherently.

3. SYSTEM OVERVIEW

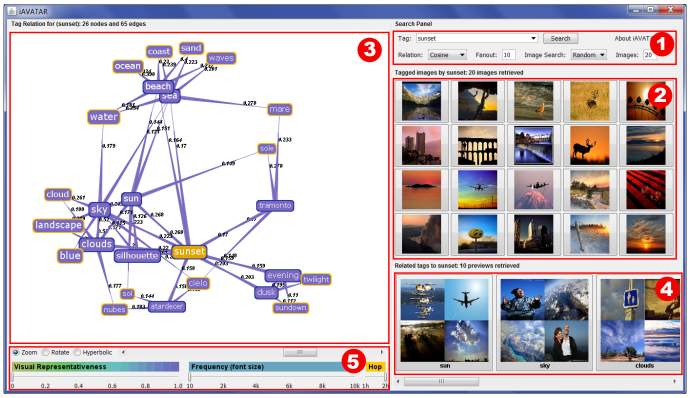
The *i*AVATAR system is implemented in Java using open-source libraries TouchGraph, Lucene and JGraphT. Figure 1 shows the system architecture of *i*AVATAR and mainly consists of the following modules.

The *i*AVATAR GUI Module: Figure 2 depicts the screenshot of the visual interface of *i*AVATAR[1]. It consists of five main panels. A user may formulate a search query in several ways. He may enter a *search tag* in Panel 1 and specify the number of images he may wish to view in the *Images* field. For example, in Figure 2 a user has entered "sunset" as the search tag and wishes to view 20 images containing this tag. It is also possible to initiate a search by double clicking on a node in the TRG in Panel 3. The search results are displayed in Panel 2.

The TRG *Viewer Panel* (Panel 3) depicts the area for visualizing the *tag relationship graph* (TRG). The TRG of the sunset tag is shown in Figure 2. The nodes in the graph are color-coded based on their visual-representativeness. The labels on the edges represent the tag co-occurrence values of pairs of tags. We shall elaborate on the structure of TRG later. Note that we can also view the TRG of a retrieved image by clicking on the image in Panel 2. Figure 5(a) depicts the TRG when the first image in Panel 2 is clicked.

The *Tag Preview Panel* (Panel 4) displays a preview of some of the highly visual-representative tags in Panel 3. Four images are randomly picked for each tag for preview. For example, Figure 2 shows preview of the tags sun, sky, and clouds among others. If any of the preview tag is clicked, then the clicked tag becomes the new search tag. Note that we can also view the preview of a specific visual-representative tag in the TRG by clicking on the corresponding node.

[1] For clarity, we recommend viewing all diagrams presented in this section directly from the color PDF, or from a color print copy.


iAVATAR

Tag Relation for (sunset): 26 nodes and 65 edges

Search Panel

Tag: sunset

Relation: Cosine

Fanout: 10

Image Search: Random

Images: 20

Tagged images by sunset: 20 images retrieved

Related tags to sunset: 10 previews retrieved

Figure 2: Visual interface of *i*AVATAR.

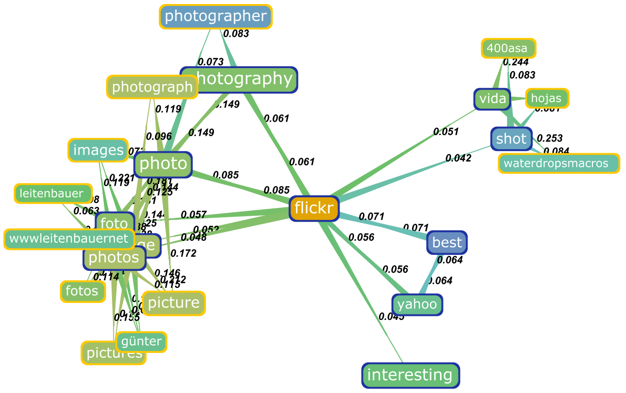


Figure 3: TRG for "Flickr".

Lastly, the *Filtering Panel* (Panel 5) enables the user to filter the nodes in the TRG in real time by modifying the thresholds of visual-representativeness and tag frequency. It also allows a user to view related tags that are one or two hops away from the search tag.

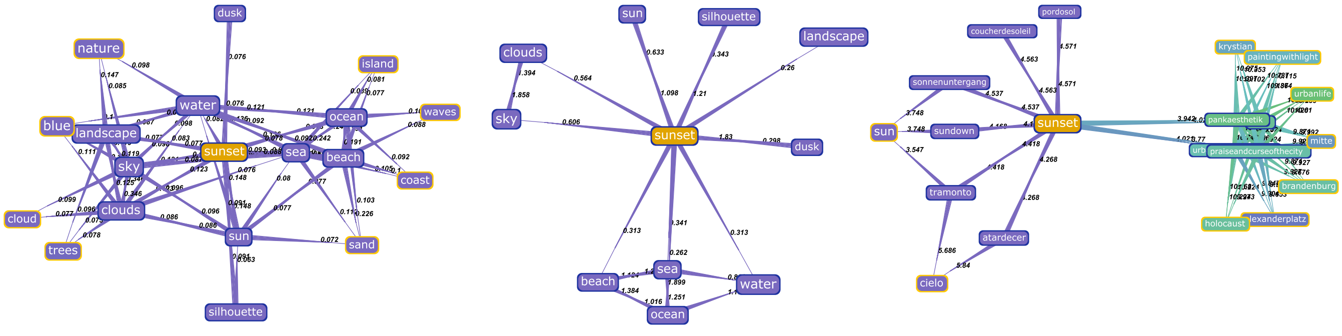
The Image Retriever Module: Given a search tag, this module retrieves a ranked list of images from the image repository that are annotated (or related) with the given tag. A user may choose a ranking method by selecting from the drop down menu associated with the *Image Search* field in Panel 1, such as TFIDF-based and tag expansion-based ranking methods. The default approach is *random ranking* where as long as an image is annotated by the searched tag, it has equal probability of being displayed in Panel 2.

The Tag Features Extractor Module: This module is the core engine of *i*AVATAR and consists of the following submodules.

The Tag Frequency Retriever Module. The purpose of this module is to compute the *frequencies* of tags in the image repository and store them in the tag features database (*Tag* DB). *Tag frequency* is the number of images a tag t is associated with in the given dataset.

The Visual-Representative Tags Finder Module. This module finds visual-representative tags from the image dataset and store them in *Tag* DB. Intuitively, a tag is visually representative if all the images annotated with the tag are visually similar to each other. We use the notion of *normalized tag clarity score* (NITC) to measure the visual-representativeness of a tag. We briefly describe NITC here. The reader may refer to [8] for details.

We consider a tag to be a keyword query and the set of images annotated with the tag are the retrieved documents based on a boolean



(a) Tag Relation: Jaccard

(b) Tag Relation: Pointwise KL

(c) Tag Relation: Pointwise MI

Figure 4: Tag Co-occurrence using different measures.

retrieval model (which returns an image as long as the image is annotated with the tag with equal relevance score). Among the various low-level features that are commonly used to represent image content, *bag of visual words* feature represents images very much like textual documents [7]. If all images associated with the tag are visually similar, then the *tag language model* estimated from the set of retrieved images shall contain some "visual words" with unusually high probabilities specific to the tag making the distance between the tag and the collection language models large.

Based on the above model, we assume a bag of visual words is extracted to represent each image. Because of this representation, we use "image" and "document" interchangeably and use d to denote an image. We now define the notion of *image tag clarity*. Let $\mathcal{D}$ be the set of images and $T \subseteq \mathcal{D}$ be the set of images annotated by a tag t. Let w be an arbitrary visual word in the vocabulary. The *image tag clarity* score of t, denoted by $\textsc{itc}(t)$, is defined as the Kullback–Leibler(KL)-divergence between the *tag language model* ($P(w|T)$) and the *collection language model* ($P(w|\mathcal{D})$). It is expressed by the following equation.

$$\textsc{itc}(t) = KL(T||\mathcal{D}) = \sum_{w} P(w|T) \log_2 \frac{P(w|T)}{P(w|\mathcal{D})} \qquad (1)$$

The collection language model $P(w|\mathcal{D})$ is estimated from the relative visual word frequency in $\mathcal{D}$. The tag language model $P(w|T)$ is estimated using Equation 2, where $P(d|T)$ reflects the relative closeness of the image d to T's centroid defined in Equation 3.

$$\begin{aligned} P(w|T) &= \sum_{d \in T} P_{ml}(w|d) P(d|T) & (2) \\ P(d|T) &= \frac{\varphi(d,T)}{\sum_{d \in T} \varphi(d,T)} & (3) \\ \varphi(d,T) &= \prod_{w \in d} P_s(w|T)^{P_{ml}(w|d)} & (4) \end{aligned}$$

In Equation 3, $\varphi(d,T)$ is a *centrality function* which defines the similarity between an image d to T, adopted from [5]. Let $P_{ml}(w|d)$ be the relative word frequency of w in image d. Let $P_s(w|T)$ be the tag language model estimated from the expected word frequency in the tagged images with equal importance $\frac{1}{|T|}$, i.e., $P_s(w|T) = \sum_{d \in T} \frac{1}{|T|} P_{ml}(w|d)$. Then $\varphi(d,T)$ is defined to be the weighted geometric mean of word generation probabilities in T shown in Equation 4. The estimated tag language model is further smoothed using the Jelinek-Mercer smoothing with $\lambda = 0.99$.

$$P_{smoothed}(w|T) = \lambda P(w|T) + (1-\lambda) P(w|\mathcal{D}) \qquad (5)$$

In tagging, the tag distribution follows a power-law distribution with a small set of tags much more frequently used than other tags [8]. To overcome the impact of tag frequency, we applied zero-mean normalization to the image tag clarity scores. The *expected* image tag clarity score with respect to t is computed by randomly assigned dummy tags with the same frequency to images in the dataset. Let $f(t)$ be the frequency of a tag t in the image dataset. Let $\mu(f(t))$ and $\sigma(f(t))$ be the *expected tag clarity* and *standard deviation* obtained by assigning multiple dummy tags having the same frequency $f(t)$. Then, the *normalized image tag clarity* score, denoted by $\textsc{nitc}(t)$, is given by Equation 6. A tag t is considered visual-representative if $\textsc{nitc}(t) \geq 3$ (i.e., the tag clarity is 3 standard deviations away from the expected tag clarity of randomly assigned tags of the same frequency).

Table 1: Tag co-occurrence measures.

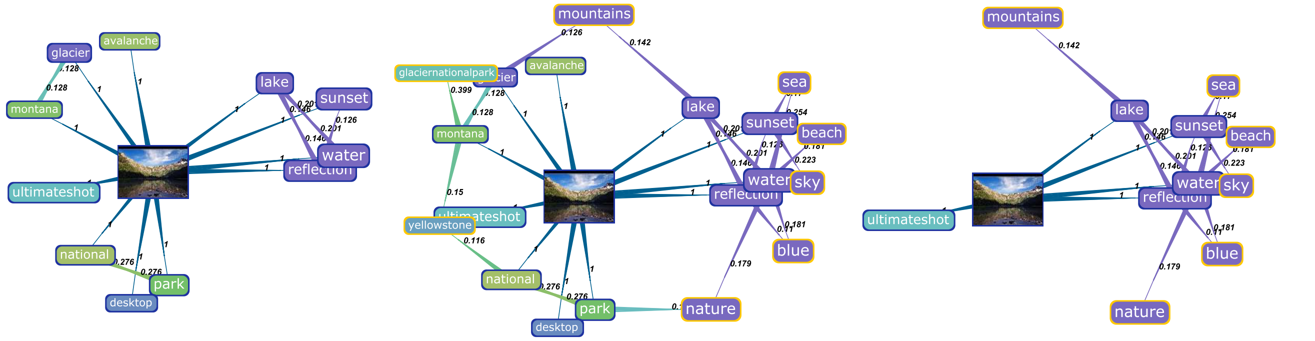
Co-occur probability	$\frac{f(t_a \wedge t_b)}{f(t_a)}$
Cosine	$\frac{f(t_a \wedge t_b)}{\sqrt{f(t_a) \times f(t_b)}}$
Jaccard Coefficient	$\frac{f(t_a \wedge t_b)}{f(t_a)+f(t_b)-f(t_a \wedge t_b)}$
Pointwise KL divergence	$\frac{f(t_a \wedge t_b)}{f(t_a)} \times \log_2(\frac{f(t_a \wedge t_b) \times N}{f(t_a) \times f(t_b)})$
Pointwise Mutual Information	$\log_2 \frac{f(t_a \wedge t_b) \times N}{f(t_a) \times f(t_b)}$

$$\textsc{nitc}(t) = \frac{\textsc{itc}(t) - \mu(f(t))}{\sigma(f(t))} \qquad (6)$$

The proposed tag language model can be estimated in $O(N)$ time for a tag associated with N images and requires at most three scans of the images (for computing Equations 2, 3, and 4). Note that the expected tag clarity scores need to be computed only once for a given dataset. As $\textsc{nitc}(t)$ values are in the range of $(-\infty, +\infty)$, the values are further normalized into [0,1] using a sigmoid function in iAVATAR. The evaluation of tag visual representativeness will be reported in a separate study.

The TRG Constructor Module: Given the visual-representative tags and tag frequencies, the objective of this module is to construct the tag relationship graph (TRG) of the search tag or image. It consists of two submodules as follows.

The Node Constructor Module. The nodes of a TRG of a given search tag or image is constructed by this module. Each node is labeled with a tag t and the font size of the label is proportional to the frequency of t. That is, the larger the font the more frequent is the tag, similar to most tag clouds. A node is color-coded based on its visual-representativeness (except for the search tag whose node is highlighted in orange color). The more visually representative (higher NITC value) a tag is the darker is the color (the spectrum of color codes used in iAVATAR is shown in Panel 5) of its node. For example, Figures 2 and 3 depict two examples of TRGs for the search tags `sunset` and `flickr`, respectively. Notice that



(a) Visual Rep: 0, Frequency: 0, Hop 1

(b) Visual Rep: 0, Frequency: 0, Hop 2

(c) Visual Rep: 0.6, Frequency: 1k, Hop 2

Figure 5: Filtering mechanism in iAVATAR.

the TRG of `flickr` has many lighter colored nodes as most the associated tags are not visually representative. In contrast, the TRG of `sunset` has many violet colored nodes as the associated tags have high NITC values.

The Edge Constructor Module. This module constructs edges between the nodes in TRG. As the label of an edge represents the *tag co-occurrence* score of the connected tag pair, it first computes the tag co-occurrence of a pair of tags (t_a, t_b) related to the search tag (or image). Let $f(t_a \wedge t_b)$ be the number of images tagged by both t_a and t_b. Let N be the number of images in the given dataset. Then, the tag co-occurrence values are computed using one of the measures listed in Table 1, where $f(t_a)$ denotes t_a's frequency. A user can choose one of these measures using the drop down list associated with the *Relation* field in Panel 1. Observe that it is possible for t_a to co-occur with a large number of t_bs. Hence, iAVATAR let users control the number of pairs of (t_a, t_b) to view by specifying the *fanout* in Panel 1. For instance, in Figure 2 the fanout is specified as 10. Given a fanout k, at most top-k (t_a, t_b) pairs are selected based on the tag co-occurrence scores and labeled edges are added between these pairs in the TRG. The color of an edge is determined by the average NITC value of the tag pair.

Figure 4 depicts the TRGs of the `sunset` search tag for three tag co-occurrence measures. Observe that the TRG's structure can significantly change with the choice of tag co-occurrence metric. It is easy to see that this feature of iAVATAR enables users to visualize the effect of a specific tag co-occurrence measure on the TRG.

The Tag Filter Module. Finally, the objective of this module is to provide users flexibility to filter or expand the TRG based on different features of the tags. Currently, it supports tag frequency, visual-representativeness, and HOP distance-based filters (Panel 5). A user can modify the threshold of visual-representativeness (resp. tag frequency) by dragging the *Visual Representativeness* (resp. *Tag Frequency*) slider in Panel 5. Only tags whose visual-representativeness (resp. tag frequency) are greater than the threshold are displayed in the TRG. The HOP filter controls expansion of the TRG by determining whether tags that are related to the search tag or image indirectly by two hops shall be displayed. If they are displayed, then the nodes on the second hop are highlighted with orange-colored border. A node is shown as a second-hop node if it is related to at least two first hop nodes (to avoid the situation of too many nodes in Panel 3). Figure 5 depicts an example of this module for an image.

4. DEMONSTRATION

Our demonstration will be loaded with the NUS-WIDE dataset[2] containing 269,648 images from Flickr [2]. Provided by the dataset, the 500-D bag of visual words feature is used to compute tag visual-representativeness. The original tags (without cleaning) of images are used in this demonstration. Using this dataset, we aim to showcase the functionality and effectiveness of the iAVATAR system in identifying and visualizing visual-representative tags and their relationships with related tags during social image search and retrieval. Specifically, we will showcase the followings.

Visual-representative tags and their relationships: Through the iAVATAR GUI, we will demonstrate visual-representative tags associated to a search tag (or image) and how these tags are related based on the chosen tag co-occurrence measure. Example search tags illustrating relationships between related tags for different user input features (Panel 1) will be presented. Users can also write their own ad-hoc search tag through our GUI and drill into any image in the search results to visualize details related to associated tags. Such visualization enables us to distinguish between non-noisy and noisy tags during image search.

Interactive filtering of the TRG: By setting the sliders in Panel 5 at different threshold values, a user can view different features associated with tags (e.g., frequency, visual-representativeness) that depend on these threshold values. For instance, for a given search tag or image, a user can view details related to most *popular* tags (tags with high tag frequency), highly visual-representative tags (high NITC score), and their relationships at different HOP distance. An immediate benefit of this information is that it enables us to determine the relationship between popular tags and highly visual-representative tags. For example, shown in Figure 5(c), `lake` and `water` tags are highly visually representative and popular and they often co-occur together.

Differential views of tags and images: We shall demonstrate and compare with specific examples how different image ranking and tag co-occurrence metrics influence the ranking of images and relationships between tags, respectively. A user can also specify their own examples and interactively choose various metrics in Panel 1 to view the effect of his/her choices.

5. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, J. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In JCDL, 2007

[2] T.-S. Chua, J. Tang, R. Hong, et al. Nus-wide: A real-world web image database from national university of singapore. In *ACM CIVR*, 2009.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR*, 2002.

[4] M. Dubinko, R. Kumar, et al. Visualizing tags over time. In *ACM Trans. Web*, 1(2), 2007.

[5] J. L. Elsas, J. Arguello, et al. Retrieval and feedback models for blog feed search. *In SIGIR*, 2008.

[6] K. Fujimura, S. Fujimura, et al. Topigraphy: visualization for large-scale tag clouds. In *ACM WWW*, 2008.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[8] A. Sun, S. S. Bhowmick. Image Tag Clarity: In Search of Visual-Representative Tags for Social Images. *In WSM (in conj. with ACM MM)*, 2009.

[2] `http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.`