Steven C. H. Hoi
Jiebo Luo
Susanne Boll
Dong Xu
Rong Jin
Irwin King *Editors*

Social Media Modeling and Computing

Springer

Social Media Modeling and Computing

Steven C.H. Hoi • Jiebo Luo • Susanne Boll •
Dong Xu • Rong Jin • Irwin King
Editors

Social Media Modeling and Computing

Springer

Editors
Asst. Prof. Steven C.H. Hoi
Nanyang Technological University
School of Computer Engineering
Singapore, Singapore
chhoi@ntu.edu.sg

Dr. Jiebo Luo
Kodak Research Laboratories
Lake Avenue 1999
14650 Rochester, NY
USA
Jiebo.luo@kodak.com

Prof. Susanne Boll
University of Oldenburg
Media Informatics and Multimedia Systems
Escherweg 2
26121 Oldenburg
Germany
susanne.boll@uni-oldenburg.de

Asst. Prof. Dong Xu
Nanyang Technological University
School of Computer Engineering
Singapore, Singapore
dongxu@ntu.edu.sg

Assoc. Prof. Rong Jin
Michigan State University
Dept. Computer Science and Engineering
Engineering Building 3115
48824 East Lansing, MI
USA
rongjin@cse.msu.edu

Prof. Irwin King
AT&T Labs Research
San Francisco
USA
and
The Chinese University of Hong Kong
Dept. Computer Science and Engineering
Shatin
Hong Kong SAR
king@cse.cuhk.edu.hk

ISBN 978-0-85729-435-7 e-ISBN 978-0-85729-436-4
DOI 10.1007/978-0-85729-436-4
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2011923787

Cover design: deblik

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Recent years have witnessed a growing number of user-centric multimedia applications, especially with the popularity of web 2.0. Examples include Flickr, YouTube, Facebook, Twitter, MySpace, etc. The emerging applications on social web and social networks have produced a new type of multimedia content, termed as "social media" here, as it is created by people using highly accessible and scalable publishing technologies for sharing via the web. With social media technology, images, videos and audios are generally accompanied by rich contextual information, such as tags, categories, title, metadata, comments, and ratings, etc. Massive emerging social media data offer new opportunities for solving some long-standing challenges in multimedia understanding and management, such as the semantic gap issue. These new media also introduce a number of new and challenging research problems and many exciting real-world applications.

This book presents recent advances on several aspects of emerging social media modeling and social media computing research. It is designed for practitioners and for researchers of all levels of expertise, from novice to expert. It targets various groups of people who need information on social media modeling and social media computing. They include:

- People who need a general understanding of social media. They are high-level managers and professional engineers who are interested in emerging social media modeling and computing technologies.
- Software developers who apply social media modeling and computing techniques. It also includes practitioners in related disciplines such as multimedia content management, information retrieval, web search, data mining, and machine learning.
- Researchers and students who are working on social media, multimedia, web search, data mining, and machine learning, and related disciplines, as well as anyone who wants a deep understanding of techniques for social media modeling and computing.

Regarding the contents and organization, this book consists of 12 chapters that present a variety of emerging technologies on social media modeling and comput-

ing. In particular, these book chapters can be summarized in the following three major aspects:

- *Social media content analysis*: The first part of the book is related to the application of multimedia content analysis techniques to the emerging social media data. It includes social image tag analysis (chapter "Quantifying Visual-Representativeness of Social Image Tags using Image Tag Clarity"), social image tag ranking (chapter "Tag-Based Social Image Search: Towards Relevant and Diverse Results"), and tag-based social image search (chapter "Social Image Tag Ranking by Two-View Learning"), social media content analysis by combining multimodal features (chapter "Combining Multimodal Features for Social Media Analysis"), and multi-label social image annotation by exploring group structures (chapter "Multi-label Image Annotation by Structural Grouping Sparsity").
- *Social media system design and analysis*: The second part of the book is devoted to social media system design and analysis. It includes the design of effective social media mechanism for incentivizing social media contributions (chapter "Mechanism Design for Incentivizing Social Media Contributions"), the design of efficient access control for privacy and security issues in multimedia social networks (chapter "Efficient Access Control in Multimedia Social Networks"), the analysis of users and their online behaviors in social video sharing portals (chapter "Call Me Guru: User Categories and Large-Scale Behavior in YouTube"), and visual analytic tools for social event analysis (chapter "Social Media Visual Analytics for Events").
- *Social media applications*: The last part of the book is related to the development of emerging social media applications by exploring emerging user-contributed social media data. It includes the application of social media information to music recommendation (chapter "Using Rich Social Media Information for Music Recommendation via Hypergraph Model"), the application of user-contributed Geotag information to automatic image annotation (chapter "Using Geotags to Derive Rich Tagclouds for Image Annotation"), and the application of social media techniques to analyze and improve real-world photobooks (chapter "Social Aspects of Photobooks: Improving Photobook Authoring from Large-scale Multimedia Analysis").

Each of the above book chapters can be considered as a compact, self-contained mini-book in its own right under its title. They are, however, organized and presented in relation to the basic principles and practice of social media modeling and computing. We also note that this book can be used as advanced materials by graduate students of information technology related subjects, such as computer science, computer engineering, and information systems, either in a classroom or for self-study.

Finally, this book was first initialized during the organization of the first international workshop on social media (WSM2009). It was later developed by soliciting contributions from a number of international experts on social media modeling and computing to present their best knowledge and practice on specific social media related topics. Some chapters of this book were originated from recent studies in in-

ternational conferences and workshops, including the SIGMM international Workshop on Social Media (WSM), and ACM International Conference on Multimedia (ACM Multimedia), and ACM International conference on Web Search and Data Mining (WSDM). As co-editors of this book, we would like to thank all the authors of the book chapters for their great efforts in providing the high quality contents to this book, and our colleagues who helped us during the organization of the WSM workshops and the book editing process.

Singapore — Steven C.H. Hoi
USA — Jiebo Luo
Germany — Susanne Boll
Singapore — Dong Xu
USA — Rong Jin
USA — Irwin King

Contents

Contributors

Sourav S. Bhowmick School of Computer Engineering, Nanyang Technological University, Singapore, Singapore, assourav@ntu.edu.sg

Joan-Isaac Biel Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, jibiel@idiap.ch

Susanne Boll University of Oldenburg, Oldenburg, Germany, susanne.boll@uni-oldenburg.de

Jiajun Bu Zhejiang Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China, bjj@zju.edu.cn

Chun Chen Zhejiang Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China, chenc@zju.edu.cn

Nicholas Diakopoulos School of Communication and Information, Rutgers University, 4 Huntington St., New Brunswick, NJ 08901, USA, nicholas.diakopoulos@gmail.com

Sabu Emmanuel School of Computer Engineering, Nanyang Technological University, Singapore, Singapore, asemmanuel@ntu.edu.sg

Andrew Gallagher Corporate Research and Engineering, Eastman Kodak Company, Rochester, USA

Daniel Gatica-Perez Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, gatica@idiap.ch

Eirini Giannakidou Informatics & Telematics Institute, Thermi, Thessaloniki, Greece, igiannak@iti.gr; Department of Computer Science, Aristotle University of Thessaloniki, Thessaloniki, Greece

Yahong Han College of Computer Science, Zhejiang University, Hangzhou, China, yahong@zju.edu.cn

Xiaofei He State Key Laboratory of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310027, China, xiaofeihe@cad.zju.edu.cn

Steven C.H. Hoi School of Computer Engineering, Nanyang Technological University, Singapore, Singapore, chhoi@ntu.edu.sg

Xian-Sheng Hua Media Computing Group, Microsoft Research Asia, Beijing 100080, China, xshua@microsoft.com

Ramesh Jain University of California, Irvine, Irvine, USA, jain@ics.uci.edu

Dhiraj Joshi Corporate Research and Engineering, Eastman Kodak Company, Rochester, USA, dhiraj.joshi@kodak.com

Mohan Kankanhalli National University of Singapore, Singapore, Singapore, mohan@comp.nus.edu.sg

Funda Kivran-Swaine School of Communication and Information, Rutgers University, 4 Huntington St., New Brunswick, NJ 08901, USA

Ioannis Kompatsiaris Informatics & Telematics Institute, Thermi, Thessaloniki, Greece, ikom@iti.gr

Phoury Lei Corporate Research and Engineering, Eastman Kodak Company, Rochester, USA

Jiebo Luo Corporate Research and Engineering, Eastman Kodak Company, Rochester, USA

Mor Naaman School of Communication and Information, Rutgers University, 4 Huntington St., New Brunswick, NJ 08901, USA

Spiros Nikolopoulos Informatics & Telematics Institute, Thermi, Thessaloniki, Greece, nikolopo@iti.gr; School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK

Ioannis Patras School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK, i.patras@eecs.qmul.ac.uk

Amit Sachan School of Computer Engineering, Nanyang Technological University, Singapore, Singapore, amit0009@ntu.edu.sg

Philipp Sandhaus OFFIS – Institute for Information Science, Oldenburg, Germany, sandhaus@offis.de

Vivek K. Singh University of California, Irvine, Irvine, USA, singhv@uci.edu

Aixin Sun School of Computer Engineering, Nanyang Technological University, Singapore, Singapore, axsun@ntu.edu.sg

Shulong Tan Zhejiang Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China, shulongtan@zju.edu.cn

Athena Vakali Department of Computer Science, Aristotle University of Thessaloniki, Thessaloniki, Greece, avakali@csd.auth.gr

Meng Wang AKiiRA Media Systems Inc, Palo Alto, CA 94301, USA, eric.mengwang@gmail.com

Fei Wu College of Computer Science, Zhejiang University, Hangzhou, China, wufei@zju.edu.cn

Kuiyuan Yang Department of Automation, The University of Science and Technology of China, Hefei, Anhui 230027, China, yky@ustc.edu

Tayebeh Yazdani School of Communication and Information, Rutgers University, 4 Huntington St., New Brunswick, NJ 08901, USA

Jie Yu Corporate Research and Engineering, Eastman Kodak Company, Rochester, USA

Hong-Jiang Zhang Microsoft Advanced Technology Center, Beijing 100080, China, hjzhang@microsoft.com

Jinfeng Zhuang School of Computer Engineering, Nanyang Technological University, Singapore, Singapore, zhua0016@ntu.edu.sg

Yueting Zhuang College of Computer Science, Zhejiang University, Hangzhou, China, yzhuang@zju.edu.cn

Part I
Social Media Content Analysis

Quantifying Visual-Representativeness of Social Image Tags Using Image Tag Clarity

Aixin Sun and Sourav S. Bhowmick

Abstract Tags associated with images in various social media sharing web sites are valuable information source for superior image retrieval experiences. Due to the nature of tagging, many tags associated with images are not visually descriptive. In this chapter, we propose *Image Tag Clarity* to evaluate the effectiveness of a tag in describing the visual content of its annotated images, which is also known as the image tag visual-representativeness. It is measured by computing the zero-mean normalized distance between the *tag language model* estimated from the images annotated by the tag and the *collection language model*. The tag/collection language models are derived from the bag of visual-word local content features of the images. The visual-representative tags that are commonly used to annotate visually similar images are given high tag clarity scores. Evaluated on a large real-world dataset containing more than 269K images and their associated tags, we show that the image tag clarity score can effectively identify the visual-representative tags from all tags contributed by users. Based on the tag clarity scores, we have made a few interesting observations that could be used to support many tag-based applications.

1 Introduction

With the advances in digital photography (e.g., digital cameras and mobile phones) and social media sharing web sites, a huge number of multimedia content is now available online. Most of these sites enable users to annotate web objects including images with free tags (e.g., `aircraft`, `lake`, `sky`). For instance, most images accessible through Flickr[1] are annotated with tags from their uploaders as well as

[1] http://www.flickr.com.

This chapter is an extended version of the paper [11] presented at the first ACM SIGMM Workshop on Social Media (WSM), held in conjunction with ACM Multimedia, 2009.

A. Sun (✉) · S.S. Bhowmick
School of Computer Engineering, Nanyang Technological University, Singapore, Singapore
e-mail: axsun@ntu.edu.sg

S.S. Bhowmick
e-mail: assourav@ntu.edu.sg

S.C.H. Hoi et al. (eds.), *Social Media Modeling and Computing*,
DOI 10.1007/978-0-85729-436-4_1,

other users. 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.

Due to the popularity of tags, there have been increasing research efforts to better understand and exploit tag usage patterns for information retrieval and other related tasks. One such effort is to make better use of the tags associated with images for superior image retrieval experiences. However, this is still a challenging research problem, as it is well known that tags are noisy and imprecise [1]. As discussed in [4], tags are created by users more for their personal use than for others' benefit. Consequently, two similar images may be associated with significantly different sets of tags from different users, especially when images can only be annotated by users with *tagging permissions* (e.g., in Flickr, only the uploader and his/her contacts can tag an image). 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 40D camera at Sentosa when she traveled to Singapore in 2008. This image may be annotated by different tags such as `Canon`, `40D`, `2008`, `Singapore`, `travel`, `beach`, `Sentosa`, and many others. Notice that tags like `2008` and `Canon` do not effectively describe the visual content of the image, but more on providing contextual information about the image. Consequently, these tags maybe considered as noise in many applications (e.g., content-based tag recommendation). As the presence of such noise may reduce the usefulness of tags in image retrieval, "de-noising" tags has been recently identified as one of the key research challenges in [1]. Such de-noising of tags also enables us to build more effective tag ranking and recommendation services [8].

In this chapter, we take a step toward addressing the above challenge. We focus on identifying and quantifying *visual-representative* tags from all tags assigned to images so that less visually representative tags can be eliminated. Intuitively, a tag is *visual-representative* if it effectively describes the *visual content* of the images. A visual-representative tag (such as `sky`, `sunset`, and `tiger`) easily suggests the scene or object that an image may describe even before the image is presented to a user. On the other hand, tags like `2008` and `Asia` often fail to suggest anything meaningful with respect to the visual content of the annotated image as any image taken in 2008 or in Asia could be annotated by the two tags.

We propose the notion of *image tag clarity* to identify visual-representative tags. It is inspired by the *clarity score* proposed for query performance prediction in ad-hoc information retrieval for textual documents [2]. Note that clarity score cannot be directly applied to annotated images as keywords of a query literally appears in the retrieved text documents whereas the tags associated with an image do not explicitly appear in it. Informally, the *image tag clarity* is computed by the Kullback–Leibler (KL) divergence between the tag language model and collection language model and further normalized with zero-mean normalization. The tag/collection language models are derived from the local content features (i.e., bag of visual-words) of the images. Our experimental study with the NUS-WIDE dataset [1], containing 269,648 images from Flickr, demonstrates that the proposed clarity score measure can effectively identify the visually representative tags.

Based on the experimental results, we further investigated the relationships between tag visual-representativeness and tag frequency. Our study revealed that they are weakly correlated. That is, frequently used tags are more likely to be visually representative. We also observed that for images having three to 16 tags, the percentage of visually representative tags increases with the increase in number of tags. Furthermore, the visual-representativeness of a tag and its *position* with respect to other tags for a given image are correlated. That is, the first few tags assigned to an image are more likely to be visually representative compared to tags assigned later. This probably reflects the phenomenon that users tend to first tag an image based on its visual content and later add other tags to describe it from different perspectives. Lastly, the visually (resp. non-visually) representative tags have higher chance of *co-occurring* strongly with other visually (resp. non-visually) representative tags. These interesting observations could be very useful in supporting a wide range of tag-based applications such as tag recommendation and social image retrieval.

The rest of the chapter is organized as follows. In Sect. 2, we review the related work with emphasis on clarity score for query performance prediction as well as image tagging. Section 3 discusses the notion of image tag clarity. The details of the dataset and experimental results are reported in Sect. 4. The observations are presented in Sect. 5 and we conclude this chapter in Sect. 6.

2 Related Work

Recall that our proposed image tag clarity measure is inspired by the notion of clarity score proposed for query performance prediction in ad-hoc retrieval. Hence, we begin by reviewing the clarity score measure. Next, we discuss relevant research efforts in annotating web objects with tags.

2.1 Clarity Score

Query performance prediction is to predict the effectiveness of a keyword query in retrieving relevance documents from a document collection [2]. The prediction enables a search engine to answer poorly performing queries more effectively through alternative retrieval strategies (e.g., query expansion) [5, 15, 19, 20]. Depending on whether documents need to be retrieved for the query, the query performance prediction algorithms can be classified into two types: *pre-retrieval* and *post-retrieval* algorithms. Pre-retrieval algorithms rely on the statistics of the words in both the query and the collection. For instance, queries consisting of words with low document frequencies in the collection tend to perform better than queries with high document frequency words. Post-retrieval algorithms predict query performance based on the properties of the retrieved documents from the collection using the query. Among various post-retrieval algorithms, one significant contribution is the *clarity score* [2].

The *clarity score* of a query is computed as the *distance* between the *query language model* and the *collection language model*. If a query is effective in retrieving topically cohesive documents, then the query language model contains unusually large probabilities of words specific to the topic covered by the retrieved documents. Consequently, the distance between the query and the collection language models is large. If a query is ambiguous, then the documents covering various topics are likely to be retrieved. That is, the retrieved set of documents is similar to a set of documents through random sampling. As the word distribution in the retrieved documents is similar to that in the collection, the distance between them is small.

Formally, let Q be a query consisting of one or more query words $\{q|q \in Q\}$ and R be the set of top-K documents retrieved by Q from the collection $\mathscr{D}$. The value of K is predefined and set to 500 in [2]. Let w be an arbitrary word in the vocabulary. Then, the query language model $P(w|Q)$ is estimated by Eq. (1), where $P(d|Q)$ is estimated using Bayes' theorem as shown in Eq. (2).

$$P(w|Q) = \sum_{d \in R} P(w|d)P(d|Q), \tag{1}$$

$$P(Q|d) = \prod_{q \in Q} P(q|d). \tag{2}$$

Observe that in both Eqs. (1) and (2), $P(w|d)$ (resp. $P(q|d)$) is the relative frequency of word w (resp. q) in the document d linearly smoothed by w's relative frequency in the collection. The collection language model, $P(w|\mathscr{D})$, is estimated by the relative frequency of w in $\mathscr{D}$. Then, the clarity score of Q is the Kullback–Leibler (KL) divergence between $P(w|Q)$ and $P(w|\mathscr{D})$, and is given by the following equation.

$$KL(Q\|\mathscr{D}) = \sum_{w} P(w|Q) \log_2 \frac{P(w|Q)}{P(w|\mathscr{D})}. \tag{3}$$

Tagging is a popular technique for annotating objects on the web. In our previous work [13], we introduced the notion of *tag clarity* in the context of user behavior study in self-tagging systems, i.e., blogs. The clarity score of a tag is defined by the KL divergence between the tag language model (estimated from the blog posts associated with the tag) and the collection language model estimated from all blog posts. As blogs are self-tagging, i.e., only the blogger could annotate his/her blog posts, the tag clarity was proposed to study whether users implicitly develop consensus on the semantic of the tags. We observed that frequently used tags are topic discriminative. This finding is partially consistent with the findings in this proposed work although the object (text vs. image) of annotation and tagging rights (self-tagging vs. permission-based tagging) are different.

2.2 Tagging Images

Recent years have witnessed increasing research efforts to study images annotated with tags in social media sharing web sites like Flickr. Tag recommendation, tag

ranking, and tag-based classification are identified as key research tasks in this context [1]. Only few works exploit the relationship between a tag to the content of its annotated images. For a given image and its annotated tags, the *relevance* between the image and each tag is estimated through kernel density estimation in [8] and through k-nearest neighbor voting in [7]. In simple words, a tag is relevant to an image I if the tag has been used to annotate many images similar to I. The relevance score for a tag is therefore image-specific, whereas in our case the tag clarity score is *global*. For a given tag, the score reflects its effectiveness in visually describing all its annotated images. In this context, our work is also related to [10] where the main focus is on searching for high-level concepts (e.g., `sunset`) with little semantic gaps with respect to image representation in visual space. In [10], for a given image I, its confidence score is derived based on the coherence degree of its nearest neighbors in both visual and textual spaces, assuming that each image is surrounded by textual descriptions. The high-level concepts are then derived through clustering those images with high confidence scores. In contrast, our work differs in the following ways: (i) the computation of clarity score of a tag is purely based on its annotated images represented in visual space only; (ii) our task is to measure the visual-representativeness of a tag (i.e., a given concept) and not to mine concepts from textual descriptions; and (iii) our work does not rely on neighborhood relationships between images.

Very recently, *Flickr distance* was proposed to model two tags' similarity based on their annotated images [17]. For each tag, a visual language model is constructed from 1000 images annotated with the tag and the Flickr distance between the two tags is computed using the Jensen–Shannon divergence. Our work is significantly different from [17] in three aspects. First, our main research objective is to measure the visual-representativeness of a single tag and not the relationship between tag pairs. Second, the language models are estimated from different image representations. Our language models are estimated on top of the widely adopted bag of visual-words representation [9] while visual language model has its own definition in [17]. Third, we analyze the impact of tag frequency on its language modeling. In their work, a fixed number (i.e., 1000) of images for each tag were sampled for estimating its language model.

In [16], a probabilistic framework was proposed to resolve *tag ambiguity* in Flickr by suggesting semantic-orthogonal tags from those tags that co-occurred with the given set of tags. Although tag ambiguity is highly related to tag clarity, the approach in [16] was purely based on tag co-occurrence without considering the content of annotated images.

3 Image Tag Clarity

Intuitively, a tag is visually representative if all the images annotated with the tag are visually similar to each other. In this sense, we heuristically consider the assignment of a tag t to an image I as a sampling process. We assume that a user samples images from a large collection and decides whether t shall be assigned to

some images. Based on this setting, if all users have an implicit common understanding on the sampling process with respect to the visual content of the images, then the assignment of the visual-representative tags is a biased sampling process such that only those images that contain certain visual concepts (e.g., sunset scene) will be assigned as a visual-representative tag (e.g. `sunset`). On the other hand, the assignment of a non-visually representative tag is an unbiased sampling process to images regardless of their visual content. Most contextual tags describing the time and location in general (e.g., the country where the photo was taken) belong to the latter case. For instance, any image taken in Singapore during year 2008 can be tagged by `Singapore` and `2008` and either tag hardly describes the visual content of the tagged images. Based on this heuristic, for a given tag t, we can consider the set of images I_t annotated by t and compare this set to a randomly drawn set of images of similar size, denoted by I'_t ($|I_t| \simeq |I'_t|$), where t' denotes a dummy tag randomly assigned to images. If I_t is similar to any I'_t, randomly drawn from a large collection of images in terms of visual content, then t is unlikely to describe any specific visual concepts. Otherwise, if I_t is significantly different from I'_t, demonstrating common visual content features, then we consider t to be visually representative. In the following, we present image tag clarity measure to quantify tag visual-representativeness.

The image tag clarity score is based on the following framework. 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 retrieval model (returns an image as long as the image is annotated with the tag with equal relevance score). Then the clarity score proposed for query performance prediction can be adopted to measure tag clarity if the visual content of the images can be represented by "word" vectors similar to that for representing textual documents. That is, if all images associated with the tag are visually similar, then the language model estimated from the set of retrieved images (or the tag language model) shall contain some "words" with unusually high probabilities specific to the tag making the distance between the tag and the collection language models large. Among the various low-level features that are commonly used to represent images, the *bag of visual-words* feature represents images very much like textual documents [9]. In the sequel, we assume that a bag of visual-words has been extracted to represent each image.[2] We also use "image" and "document" interchangeably due to this representation.

3.1 Image Tag Clarity Score

Let I_t be the set of images annotated by a tag t and $\mathscr{I}$ be the image collection. Based on the clarity score definition in Eq. (3), the *image tag clarity* score of t, denoted by $\tau(t)$, is defined as the *KL* divergence between the *tag language model* ($P(w|I_t)$)

[2]Nevertheless, we believe that the image tag clarity score is generic and can be computed using other feature representations.

and the *collection language model* ($p(w|\mathscr{I})$), where w denotes a visual-word. It is expressed by the following equation.

$$\tau(t) = KL(I_t \| \mathscr{I}) = \sum_w P(w|I_t) \log_2 \frac{P(w|I_t)}{P(w|\mathscr{I})}. \tag{4}$$

As a collection language model is often estimated by the relative word frequency in the collection, our main focus in this section is to estimate the tag language model $P(w|I_t)$. This is a challenging issue for the following reason. In textual documents, keywords in a query Q literally appear in the retrieved documents. Hence, the degree of relevance between a document d and query Q (i.e., $P(d|Q)$) can be estimated using Eq. (2). However, in a bag of visual-words representation, the tag and the words are from two different feature spaces. As a tag does not literally appear in images, the degree of relevance of an image to a tag is unknown. That is, $P(d|Q)$ in Eq. (1) (or $P(I|I_t)$ in our setting) has to be estimated differently, as Eq. (2) cannot be directly applied.

Intuitively, there are at least two approaches to estimate the tag language model. First, we can simply treat all images equally representative of a tag t, so all the images annotated with t have uniform probability to be sampled. Second, we can estimate the representativeness of images based on their distances to I_t's centroid. Images that are more close to the centroid of I_t are considered more representative and shall contribute more to the estimation of the tag language model.

- The first approach estimates the tag language model as the average relative visual-word frequency in the images with equal importance $\frac{1}{|I_t|}$. Hence, the tag language model, denoted by $P_s(w|I_t)$, is given by the following equation.

$$P_s(w|I_t) = \sum_{I \in I_t} \frac{1}{|I_t|} P_{ml}(w|I). \tag{5}$$

  Observe that it is consistent with the *small document model* used in [3] for blog feed search. Similar approach has also been used in modeling blog tag clarity in our earlier work [13].
- In the second approach, also known as the *centrality document model*, the tag language model $P_c(w|I_t)$ is estimated using Eq. (6), where $P(I|I_t)$ reflects the relative closeness of the image I to I_t's centroid defined in Eq. (7).

$$P_c(w|I_t) = \sum_{I \in I_t} P_{ml}(w|I) P(I|I_t), \tag{6}$$

$$P(I|I_t) = \frac{\varphi(I, I_t)}{\sum_{I \in I_t} \varphi(I, I_t)}, \tag{7}$$

$$\varphi(I, I_t) = \prod_{w \in I} P_s(w|I_t)^{P_{ml}(w|I)}. \tag{8}$$

  In Eq. (7), $\varphi(I, I_t)$ is a *centrality function* which defines the similarity between an image I to the tagged collection I_t. Let $P_s(w|I_t)$ be the tag language model estimated with small document model in Eq. (5) and $P_{ml}(w|I)$ be the relative

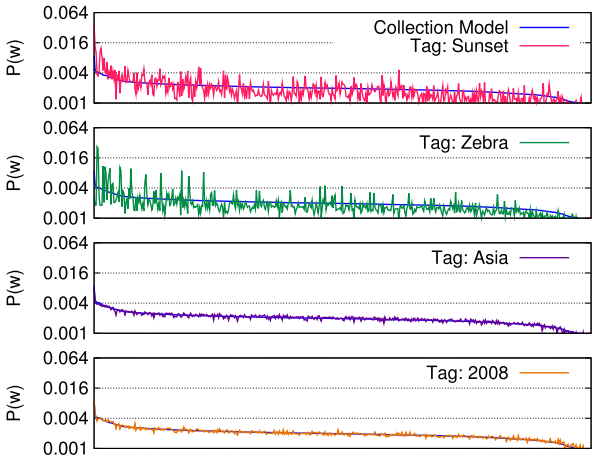


Fig. 1 Tag language models against collection language model. The x-axis is the visual-words ordered according to $P(w|\mathscr{I})$ in descending order, and the y-axis shows the $P(w|I_t)$ and $P(w|\mathscr{I})$ for each tag

visual-word frequency of w in image I. Then based on [3], $\varphi(I, I_t)$ is defined to be the weighted geometric mean of word generation probabilities in I_t as shown in Eq. (8). The weight of each visual-word is its likelihood in image I.

The estimated tag language model is further smoothed using the Jelinek–Mercer smoothing with $\lambda = 0.99$.

$$P_{\text{smoothed}}(w|I_t) = \lambda P_c(w|I_t) + (1-\lambda)P(w|\mathscr{I}). \tag{9}$$

Intuitively the centrality document model better simulates the clarity score compared to the small document model. However, the distance between an image I to the tagged collection I_t is an estimation which may not necessarily reflect the relevance between the image I and the tag t. Our experimental study revealed that the two models deliver nearly identical results. Hence in this chapter, we report the results based on the small document model due to its simplicity.

Figure 1 illustrates four example tag language models against the collection language model derived from the NUS-WIDE dataset (see Sect. 4). The x-axis is the visual-words ordered according to $P(w|\mathscr{I})$ in descending order; and the y-axis shows the $P(w|I_t)$ and $P(w|\mathscr{I})$, respectively. Clearly, the tag language models for `Sunset` and `Zebra` are significantly different from the collection language model, while the models for `Asia` and `2008` are very similar to that of the collection model. That is, the images tagged by either `Asia` or `2008` are similar to a randomly sampled set of images from the collection. For either `Sunset` or `Zebra`, one may expect the annotated image contains or describes the scene or object expressed by the semantic of the tag, making these images similar to each other and distinctive from the entire collection. Table 1 reports the tag clarity scores of these four tags. Observe that $\tau(\textit{Sunset})$ and $\tau(\textit{Zebra})$ are much larger than $\tau(\textit{Asia})$ and $\tau(2008)$.

Table 1 Clarity scores, normalized clarity scores and tag frequencies for the four example tags

Tag	Tag clarity $\tau(t)$	Normalized tag clarity $\tau_z(t)$	Tag frequency $\lvert I_t \rvert$
`Sunset`	0.294	285.3	10962
`Zebra`	0.412	97.9	627
`Asia`	0.004	−2.5	3878
`2008`	0.005	−1.9	4388

3.2 Normalized Image Tag Clarity Score

The aforementioned example demonstrates that `Sunset` and `Zebra` are more visually representative than `Asia` and `2008`. However, it is hard to determine a *threshold* for tag clarity $\tau(t)$ such that if a tag t has its clarity score above the threshold value then it is considered visually representative. Recall from Sect. 2.1, the query language model is estimated from a fix number of top-K documents (e.g., $K = 500$ in [2]). The clarity scores for all queries are therefore computed based on the same number of documents. However in tagging, the tag distribution follows power-law distribution where a small set of tags are much more frequently used than other tags (see Sect. 4). The sizes of the I_t for different tags can therefore be significantly different. We address this issue by *normalizing* the image tag clarity score.

Reconsider the task of assigning a tag to an image as a sampling process of picking up images from a large collection (i.e., $\mathscr{I}$). If the sampling is unbiased (i.e., uniform sampling), then the language model of the sampled images $P(w|I_t)$ naturally gets closer to $P(w|\mathscr{I})$ as I_t gets larger. Hence, the distance $KL(I_t\|\mathscr{I})$ becomes smaller. Therefore, $KL(I_t\|\mathscr{I})$ may not accurately reflect the clarity of a tag as it is expected that $KL(I_{t1}\|\mathscr{I}) < KL(I_{t2}\|\mathscr{I})$ if $|I_{t1}| > |I_{t2}|$ when both t_1 and t_2 are uniformly sampled, i.e., not visually representative.

To determine whether a tag t is visually representative, its tag clarity score $\tau(t)$ is compared with the clarity score of a dummy tag t' which is randomly assigned to images in $\mathscr{I}$ such that t and t' have the same tag frequency. That is, if a tag t is visually representative, then its image tag clarity score $\tau(t)$ is expected to be significantly larger than $\tau(t')$ where t' is a dummy tag randomly assigned to the same number of images as of t (or $|I_{t'}| = |I_t|$).[3] In our experiments, we observed that $\tau(t')$ follows a normal distribution for all dummy tags having the same tag frequency $|I_{t'}|$. Hence we apply zero-mean normalization and the normalized image tag clarity score $\tau_z(t)$ is given in Eq. (10), where $\mu(t')$ and $\sigma(t')$ are the *expected tag clarity score* and its *standard deviation* derived from multiple dummy tags, respectively.

$$\tau_z(t) = \frac{\tau(t) - \mu(t')}{\sigma(t')}. \tag{10}$$

[3]Recall that both $\tau(t)$ and $\tau(t')$ are computed purely from visual content features of their tagged images.

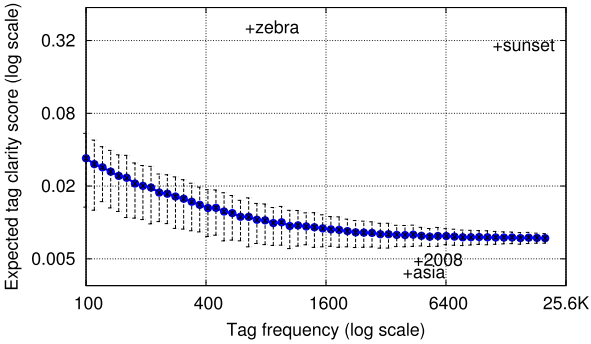


Fig. 2 The expected tag clarity scores and their standard deviations derived from 500 dummy tags with respect to the tag frequency on the x-axis. The four example tags are also plotted with their frequencies and clarity scores

The normalized image tag clarity score is the number of standard deviations a tag is observed with respect to a randomly assigned dummy tag with the same tag frequency. Note that tag frequency is an important measure here as both $\mu(t')$ and $\sigma(t')$ are heavily affected by it. As discussed earlier, the larger is $|I_{t'}|$ the closer its language model to the collection model. This is illustrated by Fig. 2, which reports the expected tag clarity scores and their standard deviations derived from 500 dummy tags with respect to the tag frequencies on the x-axis. The four example tags are also plotted with their frequencies and clarity scores. Observe that although $\tau(Sunset) < \tau(Zebra)$, $\tau_z(Sunset) > \tau_z(Zebra)$ after the normalization process (Table 1).

3.3 Time Complexity

The proposed tag language model can be estimated in $O(N)$ time for a tag associated with N images. Note that the expected tag clarity scores and standard deviation need to be computed only once for all tags with tag frequency N in a given dataset. Moreover, the computation of expected tag clarity scores and standard deviation can be further reduced by binning the tag frequencies and computing the expected tag clarity scores and standard deviations for each frequency bin.

In our experiments, we are interested in the tags that have been used to tag at least 100 images. We set our first frequency bin to cover tag frequency from $b_0 = 100$ to $b_1 = 110$. Subsequently, we set $b_{n+1} = (1 + 10\%) \times b_n$ $(n \geq 0)$ until the last bin covers the tag with highest tag frequency in our dataset. For each bin starting with b_n, 500 dummy tags with tag frequency randomly generated within $[b_n, b_{n+1})$ are used to derive the expected tag clarity and standard deviation (shown in Fig. 2). A given tag clarity score is then normalized by $\mu(b_n)$ and $\sigma(b_n)$ where b_n is the bin $|I_t|$ belongs to. Observe that in this setting every tag is normalized using dummy tags generated with frequencies within 10% of its frequency.[4]

[4]Note that the way we bin the tag frequencies and the number of dummy tags used for estimation of the expected tag clarity scores and their standard deviations are different from that in [11]. This leads to differences in the normalized tag clarity scores reported in Sects. 4 and 5.

4 Performance Evaluation

Evaluation of the techniques for quantifying tag visual-representativeness is a challenging research issue for two reasons. Firstly, there is a lack of widely adopted metric for the performance evaluation. Secondly, there is a lack of benchmark data for the evaluation. In the following, we present a summary of the evaluations conducted in our earlier work [12] and then report the observations from the experimental results. The technical details of the evaluation can be found in [12].

We used the NUS-WIDE dataset[5] containing 269,648 images from Flickr [1]. The images are assigned with zero, one or more categories (or concepts) from a pre-defined list of 81 categories. The dataset provides six types of low-level features including both global features (e.g., color histogram, edge direction histogram, wavelet texture features and others) and local features (500-D bag of visual-words). The normalized image tag clarity score discussed in this chapter is also known as *SClarL* method in [12] and the method is compared against another six methods for quantifying image tag visual-representativeness using either global features or local features. In our evaluation, we formulate the task of quantifying tag visual-representativeness as a classification task to distinguish visual-representative tags (i.e., positive tags) from non-visual-representative tags (i.e., negative tags).

Two sets of labeled tags were used in the experiments. In the first set of labeled tags, the 81 categories (which also appear as tags) in the NUS-WIDE dataset were used as positive tags and another 78 frequently used tags in the dataset were identified as negative tags. These 78 tags include 17 tags related to time (e.g., 2004–2008, January–December) and 61 location tags related to continent and country names (e.g., Europe, Japan). In the second set of labeled tags, 1576 frequently used tags were manually labeled including 814 positive tags (for object, scene, activity, color, and picture type) and 762 negative tags (for location, self-reference, opinion, camera model, and time).

The experimental evaluation adopted three performance metrics, namely, Average Precision, Precision@N, and Coverage@N. Among the seven methods evaluated, image tag clarity performed very well, with very good precision and fairly good coverage. In particular, the average precisions for the first and second sets of labeled tags were 0.89 and 0.74, respectively. The detailed results are reported in [12].

5 Observations Related to Image Tag Clarity Scores

In the NUS-WIDE dataset, there are more than 420K distinct tags that appear at least once. The tag distribution is reported in Fig. 3. Similar to statistics related to many studies on user-generated content, the tag frequency distribution follows a power-law distribution. Among the 420K distinct tags, 5981 tags have been used

[5] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm Accessed June 2009.

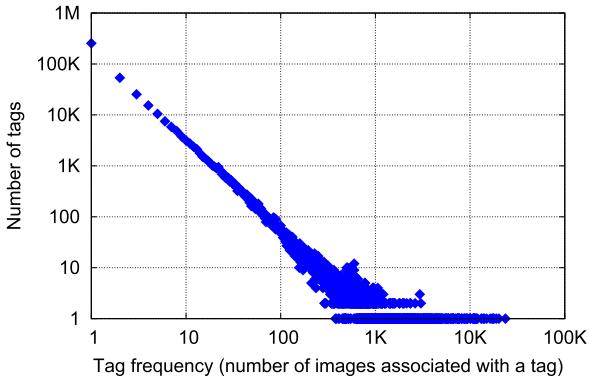


Fig. 3 Tag frequency distribution

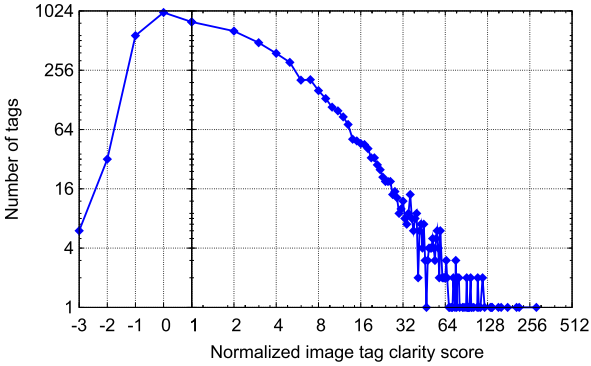


Fig. 4 Normalized image tag score distribution. Tags are binned by $floor(\tau_z(t))$ shown on the x-axis and the y-axis plots the corresponding number of tags in each bin

to annotate at least 100 images each.[6] We consider these tags as *popular* tags and report the observations made on their image tag clarity scores. In the sequel, all tag clarity scores refer to the normalized scores.

5.1 Image Tag Clarity Score Distribution

We first report the image tag clarity score distribution as well as the top-25 most and least visual-representative tags identified through our experiments.

The relationship between the number of tags (tags are binned by $floor(\tau_z(t))$) with their image tag clarity scores is shown in Fig. 4. Observe that among 5981 popular tags, 2950 tags (or about 49.3%) have tag clarity scores greater than 3. Recall that the normalized image tag clarity score is the number of standard deviations a tag is observed with respect to a randomly assigned dummy tag with the same tag frequency. If $\tau(t) \geq \mu(t') + 3\sigma(t')$, then the chance of t being randomly assigned

[6]The number reported here is slightly different from that reported in [1] probably due to different pre-processing. Nevertheless, the tag distribution remains similar.

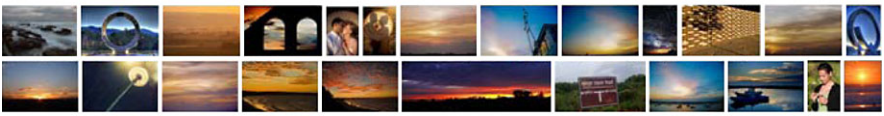
Table 2 The top-25 most and least visual-representative tags with their image tag clarity scores $\tau_z(t)$ and frequency percentile $P_f(t)$. The tags that match the category labels in the NUS-WIDE dataset are shown in bold

	Tag	$\tau_z(t)$	$P_f(t)$	Tag	$\tau_z(t)$	$P_f(t)$
1	**sunset**	285.3	100	people	−2.8	100
2	fog	215.4	97	brown	−2.7	97
3	**sky**	206.4	100	asia	−2.5	98
4	silhouette	178.5	98	japan	−2.4	98
5	sunrise	160.3	98	france	−2.1	98
6	charts	153.4	78	washington	−2.1	97
7	**sun**	138.3	99	2008	−1.9	99
8	mist	137.4	95	china	−1.8	97
9	sea	134.2	100	photograph	−1.6	89
10	**clouds**	122.3	100	july	−1.6	86
11	lightning	118.4	74	picture	−1.5	92
12	**beach**	118.3	99	virginia	−1.5	87
13	landscape	114.7	100	religion	−1.3	95
14	minimalism	111.4	78	india	−1.3	97
15	dunes	110.0	83	ohio	−1.3	87
16	blue	109.9	100	august	−1.2	80
17	dawn	108.8	91	photographers	−1.2	86
18	horizon	102.0	92	royal	−1.2	73
19	moon	99.1	95	finepix	−1.2	65
20	**ocean**	98.7	99	pic	−1.2	59
21	**zebra**	97.9	82	smorgasbord	−1.2	61
22	storm	97.5	96	world	−1.2	95
23	sketches	95.7	82	may	−1.2	84
24	**lake**	94.4	99	global	−1.1	66
25	windmills	93.7	76	2005	−1.1	96

to the images (independent of their visual content) is less than 0.1%, and we consider t to be visually representative. Here, we use three standard deviations as the threshold as it is often used as threshold to determine outliers statistically for normal distributions [6]. Nevertheless, this threshold value is only used in this chapter for analysis and can be adjusted according to a specific application. For brevity, we refer to tags that are visually representative (e.g., $\tau_z(t) \geq 3$) as *visual tags* and others as *non-visual tags*. There are 2950 visual tags and 3031 non-visual tags, respectively.

The top-25 most and least visual-representative tags are listed in Table 2 together with their normalized tag clarity score $\tau_z(t)$ and frequency percentiles (denoted by $P_f(t)$). Observe that many of the top-25 most visual-representative tags describe common scenes (e.g., `sunset`, `lightning`, `sea`, and `sky`) or objects (e.g., `zebra` and `moon`). As these are commonly used words, most users could easily use them to describe images containing the scenes or objects. Consequently, it creates strong connection between the user-specified tags and the images demonstrating the

(a) Images for tag search: `people` (least visual-representative tag)



(b) Images for tag search: `sunset` (most visual-representative tag)

Fig. 5 Images returned by Flickr for search tags `people` and `sunset`

aforementioned scenes and objects, making these tags highly visual-representative. Further, the frequency percentile values associated with the tags suggest that a large user group indeed develops consensus implicitly to use a relatively small set of common tags to describe a large number of images. Specifically, among the top 25 most visually representative tags, 18 tags have frequency percentile above 90, indicating that these are extremely popular tags.

Observe that most of the least visual-representative tags are locations (e.g., `asia`, `washington`, `japan`, `france`, `china`), or temporal such as `2008`, `july`, `august`, `may`, or high-level descriptions including `pic`, `photograph`, `picture`. All these tags do not convey much information related to the visual content of the images. For instance, images accompanied with the `asia` tag are very diverse and can range from the busy street scenes in Bangkok to images of Gobi desert in Mongolia. Such results show that the proposed image clarity score seems to be a good measure reflecting the semantic relationship of an assigned tag to the visual content of the image.

An interesting observation is that `people` is rated as a least visually representative tag. A tag search of `people` on Flickr showed that most of the returned images indeed contained people in their visual content (see Fig. 5(a)). However, the images demonstrated a great variety especially with respect to the background settings. Hence the proposed technique may wrongly identify the tags that are indeed related to some visual content as non-visual tags when the visual pattern is not clear from the feature representation. That is, such visual pattern may require a certain high-level recognition. This calls for further study on how to detect visual tags related to complicated visual patterns like `people`. On the other hand, images returned in response to the tag `sunset` indeed show similar visual scenes (see Fig. 5(b)).

5.2 Tag Usage Pattern

5.2.1 Tag Visual-Representativeness vs. Tag Frequency

It is often assumed that extremely popular tags, like stop-words in textual documents, contain little information in image tagging [18]. However, as demonstrated

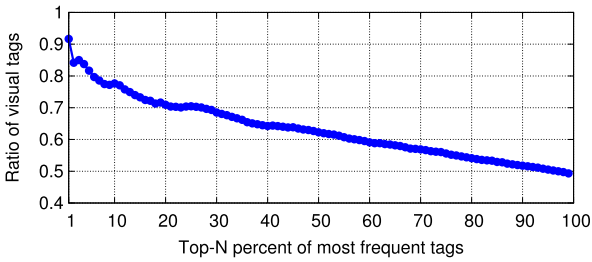


Fig. 6 Ratio of visual tags against the tag frequency at top N percent indicated along the x-axis

in our empirical study, many of the highly representative tags (e.g., the top-25 most representative tags) have frequency percentile above 90. One example is `sky` which is the third most popular tag in the dataset. It is also the third most visually representative tag and been used as a category label in the NUS-WIDE dataset. Using the notion of image tag clarity, we aim to have a deeper understanding on the relationship between tag clarity and its frequency.

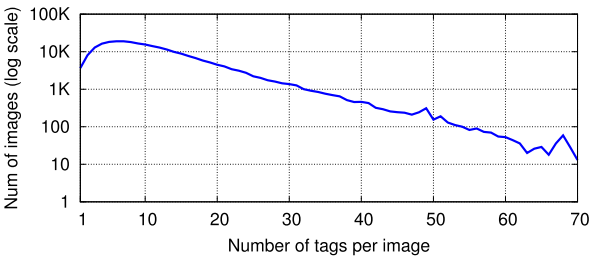
Our study showed that the 2950 visual tags are used 2,225,239 times to annotate images in the dataset[7]; while the 3031 non-visual tags are used 997,014 times only. That is, the visual tags are 2.23 times more frequently used than the non-visual tags. In other words, users are more likely to annotate images using tags related to the visual content of the images.

To further study the relationship between tag visual-representativeness and tag frequency, we sorted 5981 tags of interests according to their tag frequency in descending order. Figure 6 plots the ratio of visual tags among the top $N\%$ most frequent tags ($1 \leq N \leq 100$). The figure clearly indicates that the highly frequently used tags are more likely to be visual tags. For instance, more than 90% of the 60-most frequently used tags (or 1% of the 5981 tags) are visual tags. This is consistent with that listed in Table 2, where many of the most visually representative tags have high frequency percentiles. The Pearson's correlation coefficient between tag frequency and tag clarity score is 0.35. That is, they are weakly correlated and more frequent tags are in general more likely to be visually representative. This is not surprising as tags are in general considered resource annotations and the resource in this setting is images. The aforementioned observation also supports tag-based approach for social image retrieval as most frequently used tags are indeed visual tags.

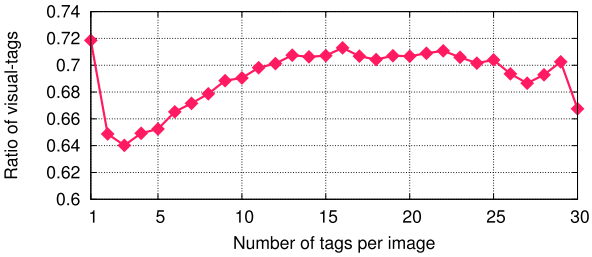
5.2.2 Visual Tags vs. Non-visual Tags

In this section, we study the distribution of visual and non-visual tags with respect to the number of tags associated with images as well as their positions. We first plot the tag distribution among images in the dataset in Fig. 7(a). In this plot, the x-axis is the number of tags per image and y-axis plots the number of images having that

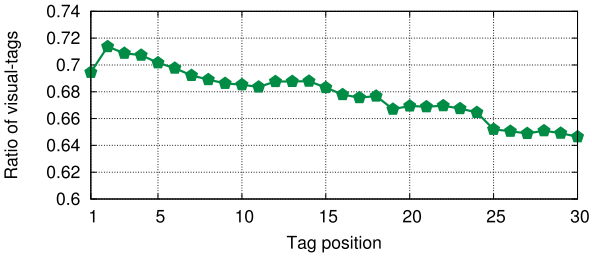
[7] One image may be annotated by multiple visual or non-visual tags, respectively.



(a) The number of images against the number of tags per image.



(b) Average ratio of visual tags among all images having the number of tags specified on x-axis.



(c) Ratio of visual-tags at different tag positions.

Fig. 7 Tag usage patterns between tag visual-representativeness, tag frequency and tag position

number of tags. Among the 269K images in the dataset, nearly 74% of images are associated with three to 16 tags (from the domain of 5981 tags of interest). Fewer than 5% of images have more than 30 tags each. Hence, we only focus on those images with no more than 30 tags for the study of tag position distribution.

Figure 7(b) plots the average ratio of visual tags among each image having K tags ($1 \leq K \leq 30$). The ratio of visual tags gradually increases from 0.64 to 0.72 with the increase of the number of tags from three to 16. Subsequently, the ratio remains relatively stable for images having 17 to 25 tags each. Figure 7(b) shows that the chance of an image being annotated by visual tags increases with the number of tags received. As many tags are received from the contacts of the image uploader in Flickr, these users may not know much about the image other than its visual content. The tags contributed by these users are more likely to be visual tags. Overall, the results also show that in general more visual tags are associated with images than non-visual tags with the ratio of visual tags well above 0.6. This is consistent with