# Table of Contents

MM 2012 Conference Organization .................................................................................................................. xxxi

MM 2012 Technical Program Committee ............................................................................................................ xxxiii

MM 2012 Technical Demo Program Committee .................................................................................................. xxxvii

MM 2012 Area Chairs .......................................................................................................................................... xxxviii

MM 2012 Additional Reviewers .......................................................................................................................... xxxix

MM2012 Sponsor & Supporters ............................................................................................................................. xl

## Plenary Talk 1

Session Chair: Noboru Babaguchi (*Osaka University*)

- **Future Direction of Digital Content**  ........................................................................................................ 1
  Masahiro Fujita (*Sony Corporation*)

## Panel 1: 20th Anniversary Panel

Session Chairs: Klara Nahrstedt (*University of Illinois at Urbana-Champaign*) and Malcolm Slaney (*Microsoft*)

- **Coulda, Woulda, Shoulda: 20 Years of Multimedia Opportunities (20th Anniversary Panel)** ................................................................................................................................. 3
  Klara Nahrstedt (*University of Illinois at Urbana-Champaign*), Malcolm Slaney (*Microsoft*)

## Plenary Talk 2

Session Chair: Kiyoharu Aizawa (*The University of Tokyo*)

- **Decoding Visual Experience from the Human Brain** .................................................................................. 5
  Yukiyasu Kamitani (*ATR Computational Neuroscience Laboratories*)

## Panel 2: Panel Discussion

Session Chairs: Lexing Xie, (*Australian National University*), David A. Shamma, (*Yahoo! Research*)
and Cees Snoek (*University of Amsterdam*)

- **Content is Dead; Long-Live Content!** ........................................................................................................ 7
  Lexing Xie (*Australian National University*), David Ayman Shamma (*Yahoo! Research*),
  Cees Snoek (*University of Amsterdam*)

## Best Paper Session

Session Chair: John Smith (*IBM*)

- **Finding Perfect Rendezvous On the Go: Accurate Mobile Visual Localization and Its Applications to Routing** ......................................................................................................................... 9
  Heng Liu (*University of Science and Technology of China*), Tao Mei (*Microsoft Research Asia*),
  Jiebo Luo (*University of Rochester*), Houqiang Li (*University of Science and Technology of China*),
  Shipeng Li (*Microsoft Research Asia*)

- **Right Buddy Makes the Difference: An Early Exploration of Social Relation Analysis in Multimedia Applications** .............................................................................................................................. 19
  Jitao Sang, Changsheng Xu (*Chinese Academy of Sciences & Singapore Institute of Digital Media*)

- **Propagation-Based Social-Aware Replication for Social Video Contents** ............................................... 29
  Zhi Wang, Lifeng Sun (*Tsinghua University*), Xiangwen Chen (*The Chinese University of Hong Kong*),
  Wenwu Zhu (*Tsinghua University*), Jiangchuan Liu (*Simon Fraser University*),
  Minghua Chen (*The Chinese University of Hong Kong*), Shiqiang Yang (*Tsinghua University*)

- **Action Recognition for Human-Marionette Interaction** ............................................................................... 39
  Shih-Yao Lin, Chuen-Kai Shie, Shen-Chi Chen, Yi-Ping Hung (*National Taiwan University*)
Right Buddy Makes the Difference: an Early Exploration of Social Relation Analysis in Multimedia Applications

Jitao Sang1,2, Changsheng Xu1,2
1National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
2China-Singapore Institute of Digital Media, Singapore, 119613, Singapore
{jtsang, csxu}@nlpr.ia.ac.cn

ABSTRACT

Social media is becoming popular these days, where user necessarily interacts with each other to form social networks. Influence network, as one special case of social network, has been recognized as significantly impacting social activities and user decisions. We emphasize in this paper that the inter-user influence is essentially topic-sensitive, as for different tasks users tend to trust different influencers and be influenced most by them. While existing research focuses on global influence modeling and applies to text-based networks, this work investigates the problem of topic-sensitive influence modeling in the multimedia domain.

We propose a multi-modal probabilistic model, considering both users’ textual annotation and uploaded visual image. This model is capable of simultaneously extracting user topic distributions and topic-sensitive influence strengths. By identifying the topic-sensitive influencer, we are able to conduct applications like collective search and collaborative recommendation. A risk minimization-based general framework for personalized image search is further presented, where the image search task is transferred to measure the distance of image and personalized query language models. The framework considers the noisy tag issue and enables easy incorporation of social influence. We have conducted experiments on a large-scale Flickr dataset. Qualitative as well as quantitative evaluation results have validated the effectiveness of the topic-sensitive influencer mining model, and demonstrated the advantage of incorporating topic-sensitive influence in personalized image search and topic-based image recommendation.

Categories and Subject Descriptors
H.5.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.4.m [Information Systems]: Miscellaneous

General Terms
Algorithms, Experimentation, Performance

Keywords
social relation analysis, multimedia application

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

1. INTRODUCTION

Social media has revolutionized the way people share and access information. In social media services, users necessarily interact with each other into communities, i.e., social networks. The social interactions include two-way links like “connect” in LinkedIn and “add friend” in Facebook, and one-way links like “follow” in Twitter, “contact” in Flickr and “subscribe” in Youtube. The social links are well recognized forces that govern the behaviors of involved users as well as the dynamics of social networks. For example, the colleagues on LinkedIn will largely impact one’s choices in work, while the friends from Facebook have strong influence on one’s preferences in daily life. Therefore, understanding social links can benefit various applications such as viral marketing [1], collaborative recommendation [2] and collective information retrieval [3]. In this work, we first focus on modeling the one-way social link, or influence [4], in photo sharing websites, and then exploit it for the application of personalized image search.

Social influence analysis has attracted extensive research interests, such as investigating the existence of influence [5, 6], maximizing influence propagation [7, 8] and identifying the evoked influencer [9]. For social influence-based retrieval, the basic premise behind is that the preferences of other users, who are the influencers to the searcher, provide a good indication for the searcher’s preference. The influence link between influencers and searcher has two intrinsic characteristics: (1) The influence is continuous. This is easy to understand, as binary influence link (i.e., influencer or not) is too coarse to model the relationship strength. Recently, continuous influence modeling has been addressed in [10, 11]. (2) The influence is topic-sensitive. Given a one-way influence network, the actual influencer is task-dependent. We use a toy example to elaborate this. Fig. 1 shows the contact network of user Bob in Flickr which includes three influencers: Tom, Emily and Jason. On the right we illustrate the expertise of each influencer regarding topics on travel, fashion and technology. Assuming Bob
is searching photos of “Tahtí” for his honeymoon trip, obviously Tom’s preference will influence him most. While, when Bob searches photos of “D&G fashion show”, the advice from Emily will be most appreciated. This demonstrates that some influencers in certain topics might be more trustworthy than others and the influence value is topic-sensitive. However, little work has investigated the problem of topic-level influence modeling, especially in multimedia domain.

In this paper, we investigate the problem of topic-sensitive influencer mining in social media communities. In particular, we take Flickr, one of the most popular photo sharing websites, as the social media platform in our study. In Flickr, users are allowed to add others as their contacts, which is analogous to the influencers. Besides the explicit one-way influence link between users, the uploaded images and associated tags can also be leveraged. The multi-modal information available on photo sharing websites provide challenges as well as opportunities to the problem of topic-sensitive influence mining. As shown in Fig.2(a), we formulate the multi-modal Topic-sensitive Influence Modeling (mmTIM) problem as follows: The input is users’ contact networks, annotation sets and uploaded image sets. We propose a probabilistic generative model to infer the output, which inverses the generation process of image content as well as the associated tags. With users as nodes, user-influencer social links as edges, the output includes (1) the learned topic space, (2) the topic distribution of each node and (3) the topic-sensitive edge strength 1. The edge strengths correspond to the topic-sensitive influence values, while the node topic distribution indicates the topic expertise of corresponding user. After mining the topic-sensitive influencer in the contact network, we apply it to the application of personalized search. Following the risk minimization-based information retrieval scheme, we present a general framework for personalized image search. The new framework benefits from the consideration of influencer preference and annotation confidence in the modeling process. Topic-sensitive influence and user expertise can be readily integrated in the expansion of the language models (LM) for queries and image documents. Therefore, the main contributions of this paper can be summarized as follows:

- We address the topic-sensitive influencer mining problem for applications of collaborative recommendation and collective search in multimedia domain.
- A multi-modal topic-sensitive influence model is presented, which enables simultaneous extraction of node topic distribution, topic-sensitive edge strength and the topic space.
- We propose a risk minimization-based theoretical framework for personalized image search. Topic-sensitive influence and annotation confidence can be seamlessly incorporated.

2. RELATED WORK

Our work is closely related to two research topics: (1) social influence modeling and (2) personalized information retrieval.

2.1 Social Influence Modeling

One research line in social influence modeling is to investigate the influence spread patterns, such as influence propagation maximization [7, 8], influence diffusion over social networks [12, 13] and influence network evolution [9]. Several works focus on another research line by analyzing influence for social activity understanding. For example, in [14] the authors demonstrated from results of questionnaire that influencer behavior subjects to influence and impacts influence network evolution. [5] studied the correlation between influence and social activities, and concluded that social influence plays an important role in dictating user behaviors. Ronald et al. [15] proposed an agent-based model to describe social activities between users in influence network. They found influence network has a significant effect on the number and type of the social activities.

The influence networks used in above works are either binary or continuous. For binary influence modeling, a number of efforts have been made to identify the existence of social influence in online social networks. Anagnostopoulos et al. [5] provided a theoretical justification to identify influence as a source of social correlation. In the same year, Crandall et al. [6] analyzed the binary influence link by investigating the correlation between social similarity and influence. Compared with binary influence, continuous influence leads to more precise description of the social relation. The study in [10] developed a link-based latent variable model to estimate influence strength from user similarity and interaction activities (e.g., communication, collaborative tagging). Zhuang et al. [11] further addressed this problem by exploring the heterogeneous data in photo sharing websites and proposing a multiple kernel-based learning-to-rank method to combine them.

As discussed in the previous section, in many cases, the influence strength is actually topic-sensitive. Topic-sensitive influence modeling, which provides more delicate and informative descriptions of social relations between people, is necessary to applications...
like collaborative recommendation and collective search. So far as we know, only two papers have addressed this problem, which are focused on text-based citation alike networks. Tang et al. [16] introduced topic-sensitive social influence modeling and proposed a Topical Affinity Propagation (TAP) approach to tackle the problem. In their formulation, the topic space and topic distribution for each user node are assumed available and employed as inputs. In [17], the authors moved one step further and presented a graphical probabilistic model to learn the node topic distributions and topic-sensitive influence jointly. The information they utilized include the heterogeneous link and the textual content associated with each node. Our work is different from them in that we are investigating the topic-sensitive influence in the multimedia context. Besides the explicit influence link and the textual annotation, we provide a multi-modal framework simultaneously modeling visual content.

2.2 Personalized Information Retrieval

Extensive efforts have been focusing on personalized information retrieval these days. The resources being leveraged include explicit user profile, relevance feedback, click-through data, context information, social annotation and social network. Chirita et al. [18] conducted an early work by re-ranking search results according to the cosine distance between each URL and explicit user interest profiles. Kraft et al. [19] utilized the search context information collected from users’ feedback to refine the raw query for second search. In [20], the authors utilized user’s click-through history to construct user preference vector and extended Topic-Sensitive PageRank for personalized search. [21] introduces a local search system by considering context information (e.g., time, location, weather) as well as user query log for search result personalization. [22] demonstrates that user-generated metadata indicates user’s interest and can be used to personalize information. They extracted latent topics from tagging data and used the latent user interest profile for personalized image search. In similar spirits, Lu et al. [23] utilized user activities of tagging and joining interest group to extract latent interest profile and re-rank the returned images by combining latent interest-based user preference and query relevance. Recently, an interesting work is performed in [24], where the authors proposed to extract user-specific topic spaces by expanding the raw annotation set for each user. Queries are mapped to the derived user-specific topics spaces and the personalized relevance score over images are calculated.

Several approaches have directly or indirectly employed user’s social relations for personalized search. Bender et al. [25] proposed a PageRank-like algorithm to exploit the utilization of social relations. They assumed that a document receives an extra “friendship” score when tagged by the searcher’s friends. Carmel et al. [3] explicitly defined similar and similar scores to model relation-ship between users. With familiar score estimated from explicit social link and similar score calculated from collaborative activities, personalization is conducted according to the searcher’s relationship strength with users in the social network. Similarly, in [26], a multilevel actor similarity method is proposed to first measure the social affinity between users. The personalized ranking score is then calculated by combining the query-video relevance and the affinity score between the searcher and the video’s owner. Lin et al. [27] leveraged the output of community discovery and proposed a community-oriented re-ranking scheme to aggregate query-video relevance and user-community-video preference.

Our work differentiates from existing social network-based approaches in that, we exploit the topic-sensitive social relations between users and integrate query-dependent social strength into personalized ranking score calculation. Moreover, for the first time, we present a theoretical personalized image search framework by jointly modeling social influence preference and annotation confidence.

3. TOPIC-SENSITIVE INFLUENCER MINING

This section introduces the topic-sensitive influencer mining problem. Table 1 lists the key notations. First, we formally define the problem:

**Definition 1 (Topic-sensitive Influencer Mining).** Given a collection of Flickr users \(U\), each user \(u \in U\) corresponds to a three-dimensional tuple \([C_u, D_u, T_u]\). The goal of topic-sensitive influencer mining is to learn (1) Topic space \(\Phi^u \subseteq \mathbb{R}^{T \times |V|}\) and \(\Phi^v \subseteq \mathbb{R}^{T \times |V|}\); (2) Node topic distribution \(\Omega_u \subseteq \mathbb{R}^{T \times 1}\) for each user \(u\); and (3) Topic-sensitive edge strength \(\Psi(t) \subseteq \mathbb{R}^{[|T| \times |D|]}\), \(t = \{1, \ldots, T\}\).

We propose a probabilistic generative model for topic-sensitive influencer mining. Since the model specifies the generation process of the visual image content as well as the textual tag words for each user, we call it multi-modal Topic-sensitive Influence Model (mmTIM).

3.1 Multi-Modal Topic-sensitive Influence Model

It is well recognized that users’ online activity is impacted a lot by their influencers [5, 15]. The uploaded images and annotated tags can be viewed as observations of users’ activities. Inspired by this, for each user, we assume that his/her tag words and uploaded images are generated in two ways, either innovative—mainly depending on his/her own interest or influenced—being more influenced by the influencers. For example, when Bob uploads an image, the theme of the image may match his own original interest, or follow one of his influencers. Another example, the tag Bob choos-

---

2 We use topic distributions over tag words and visual descriptors to represent the topic space: \(T\) is the number of topics, \(|V|\) is the size of tag vocabulary, \(|V|\) is the size of visual-descriptor vocabulary.

3 \(\Psi_{u_1, u_2}(t)\), \(u_2 \in C_{u_1}\) measures the influence strength from user \(u_2\) to \(u_1\) in \(t\)-th topic.
The generative process of visual descriptors is similar other than collaborative recommendation.

Fig. 3 illustrates our understanding of the generation process as graphical structure. For image visual content, we first construct a visual vocabulary and represent each user by the visual descriptor responses $v$ of his/her uploaded images. In the model, we use binary variables $s^w_u$ and $s^s_u$ to control whether the tag words and visual descriptors are generated by the users themselves or according to one of the influencers. When $s^w_u = 1$, the tag word is generated based on user’s own interest $\Omega_u$. When $s^w_u = 0$, the tag word is influenced by one of the users in the contact list $C_u$, and it is generated based on the interest $\Omega_{w, u}$ of the sampled influencer $e^w_u$. The details of the generative process for tag words of user $u$ are as follows:

For each tag word $w_{u, i} \in w_u$:

1. Draw a switch variable from a bernoulli distribution: $s^w_{u, i} \sim Bernoulli(\lambda)$;
2. If $s^w_{u, i} = 0$, then
   i. Draw an influencer from $u$’s contact list, which follows multinomial distribution parameterized by $\gamma$: $c^w_{u, i} \sim Multi(\gamma)$;
   ii. Draw a topic from influencer $c^w_{u, i}$’s topic distribution: $z^w_{u, i} \sim Multi(\Omega_{w, u})$;
3. If $s^w_{u, i} = 1$, then
   i. Draw a topic from $u$’s own topic distribution: $z^w_{u, i} \sim Multi(\Omega_u)$;
4. Draw a word $w_{u, i}$ from the topic-word distribution: $w_{u, i} \sim Multi(\Phi^w_{z_{u, i}})$.

The generative process of visual descriptors is similar other than that the visual descriptor $v_{u, i}$ is sampled from topic-visual descriptor distribution $\Phi^v_{z_{u, i}}$. Actually, during the model learning process, we assume the prior distributions follow symmetric Dirichlet, which are conjugate priors for multinomial (or Beta for Bernoulli). For simplification, we do not draw the hyperparameters in the graphical model.

Note that the proposed mmTIM is much inspired by the model presented in [17], but with two important differences: (1) [17] focuses on social network based on textual information, while we present its multimodal version to explore the multimedia social network, by considering both textual annotation and visual image information; (2) Instead of building influencer set at document-level as in [17], we build the influencer set at user-level, which is more intuitive for user-oriented applications like personalized search and collaborative recommendation.

### 3.2 Learning mmTIM by Gibbs Sampling

The mmTIM model includes three sets of latent variables, the binary switch labels $s^w_u, s^s_u$, the influencer $c^w_u, c^s_u$, and the topic assignments $z^w_u, z^s_u$. We use Gibbs sampling [28] for model learning. When using Gibbs sampling to train a generative model, a Markov chain is formed. The joint distribution is approximated by drawing a sequence of samples and each latent variable is iteratively updated by fixing other variables. The derivation of update rules is detailed in the appendix. We list the update rules for latent variables concerning tag words as follows.

\[
p(s^w_{i} = 0|s^w_{i-1}, u^w_{i-1}, c^w_{i}, z^w_{i-1}, \cdot) \propto \frac{N_{U,S}(u^w_{i-1}, 0) + \alpha_\lambda - 1}{N_{U}(u^w_{i-1}) + 2\alpha_\lambda - 1} \cdot \frac{N_{U,Z}(c^w_{i}, z^w_{i}) + \alpha_\Omega - 1}{N_{U}(c^w_{i}) + \alpha_\Omega - 1} \cdot \frac{1}{N_{U,S}(u^w_{i-1}, 1) + T\alpha_1 - 1}
\]

\[
p(s^w_{i} = 1|s^w_{i-1}, u^w_{i-1}, c^w_{i}, z^w_{i-1}, \cdot) \propto \frac{N_{U,S}(u^w_{i-1}, 1) + \alpha_\lambda - 1}{N_{U}(u^w_{i-1}) + 2\alpha_\lambda - 1} \cdot \frac{N_{U,Z}(c^w_{i}, z^w_{i}) + \alpha_\Omega - 1}{N_{U}(c^w_{i}) + \alpha_\Omega - 1} \cdot \frac{1}{N_{U,S}(u^w_{i-1}, 1) + T\alpha_1 - 1}
\]

\[
p(c^w_{i}|s^w_{i-1}, w^w_{i}, \cdot) \propto \frac{N_{U,C,S,Z}(u^w_{i}, c^w_{i}, 0, z^w_{i}) + \gamma}{N_{U,S,Z}(u^w_{i}, 1, z^w_{i}) + \gamma} \cdot \frac{N_{U,S,Z}(w^w_{i}, 1, z^w_{i}) + \alpha_\phi - 1}{N_{U,S,Z}(w^w_{i}, 1) + \gamma} \cdot \frac{N_{U,Z}(z^w_{i}) + |W|\alpha_\phi - 1}{N_{U,Z}(z^w_{i}) + |W|\alpha_\phi - 1}
\]

\[
p(z^w_{i}|s^w_{i-1}, c^w_{i}, \cdot) \propto \frac{N_{U,C,S,Z}(u^w_{i}, c^w_{i}, 0, z^w_{i}) + \gamma}{N_{U,S,Z}(u^w_{i}, 1, z^w_{i}) + \gamma} \cdot \frac{N_{U,S,Z}(w^w_{i}, 1, z^w_{i}) + \alpha_\phi - 1}{N_{U,S,Z}(w^w_{i}, 1) + \gamma} \cdot \frac{N_{U,Z}(z^w_{i}) + |W|\alpha_\phi - 1}{N_{U,Z}(z^w_{i}) + |W|\alpha_\phi - 1}
\]

where $u^w_i$ denotes the user to which the $i$-th word belongs, $z^w_i$ denotes the topic assignment of the $i$-th word, $z^w_i \sim Multi(\Omega_u)$, $\alpha_\lambda, \alpha_\phi, \alpha_\Omega$, and $\gamma$ are symmetric hyperparameters controlling the corresponding Dirichlet prior distributions. $N(\cdot)$ stores the number of samples satisfying certain requirements during the iterative sampling process. For example, $N_{U,C,S,Z}(u^w_i, c^w_i, 0, z^w_i)$ represents the number of tag words for user $u^w_i$ which are supposed to be influenced by contact user $c^w_i$ and generated from topic $z^w_i$. The update rules for variables concerning visual descriptors are similar and omitted here for the space reason.

### 3.3 Parameter Estimation

After Gibbs sampling, we can obtain the sampled latent variables of $s^w_u, s^s_u, c^w_u, c^s_u, z^w_u, z^s_u$. Topic-word and topic-visual descriptor distributions $\Phi^w, \Phi^v$, which represent the learned topic space, can be easily computed from sampled topic assignments $z^w_u, z^s_u$. Since $\Phi^w_{z_{u, i}}$ actually measures the probability of the $j$-th tag word in the $t$-th topic, it can be estimated by normalizing the counter $N_{Z,W}(\cdot)$. It is similar to $\Phi^v$. Therefore, we have:

\[
\Phi^w_{z_{u, i}} = \frac{N_{Z,W}(Z_t, w_{i}) + \alpha_\phi}{N_{Z,W}(Z_t) + |W|\alpha_\phi} \quad (6)
\]

\[
\Phi^v_{z_{u, i}} = \frac{N_{Z,V}(Z_t, v_{i}) + \alpha_\phi}{N_{Z,V}(Z_t) + |V|\alpha_\phi} \quad (7)
\]

where $Z_t$ denotes the $t$-th topic, which is different from the topic assignment variables $z^w_u$ and $z^s_u$. The node topic distribution for $m$-th user $U_m$ can be computed by:

\[
\Omega_{m,t} = \frac{N_{U,S}(U_m, 1, Z_t) + N_{U,S,Z}(U_m, 1, Z_t) + \alpha_\Omega}{N_{U,S}(U_m, 1) + N_{U,S,Z}(U_m, 1) + T\alpha_1} \quad (8)
\]
The t-th topical influence strength $\Psi_{m1,m2}(t)$ from user $U_{m1}$ to user $U_{m2}$, can be estimated by the number of words/visual descriptors for user $U_{m2}$ which are influenced by $U_{m1}$ in t-th topic, i.e., $N_{U,C,S,Z}(U_{m2}, U_{m1}, 0, Z_t)$:

$$\Psi_{m1,m2}(t) = \frac{N_{U,C,S,Z}(U_{m2}, U_{m1}, 0, Z_t) + \alpha_t}{N_{U,S,Z}(U_{m2}, 0, Z_t) + N_{U,C,S,Z}(U_{m2}, U_{m1}, 0, Z_t) + C_{U_{m2}}|\alpha_t}$$

This equation is quite intuitive in that if one user $U_{m2}$ has more tag words or visual images from topic $Z_t$ likely to be influenced from the contact user $U_{m1}$, then $U_{m1}$ is supposed to influence $U_{m2}$ strongly in t-th topic.

4. PERSONALIZED IMAGE SEARCH UNDER RISK MINIMIZATION

In this section, we propose a risk minimization-based framework to incorporate the derived topic-sensitive influences for multimedia application of personalized image search. Risk minimization is a popular information retrieval framework with solid theoretical foundation [29]. It formulates query and document by language models (LM) [30], where queries and documents are modeled respectively from a generative process. Risk minimization views the retrieval of relevant documents from the perspective of Bayesian decision theory, and the goal is equivalent to minimizing the expected loss.

4.1 Risk Minimization Framework

In the context of personalized image search, we view the query as generated from a probabilistic process associated with searcher $u$, and each image as generated from a probabilistic process associated with the candidate image set. Specifically, query is the result of first choosing a language model, and then generating the query from that model. The generative process of each image is similar. Note that the query and image language models provide entrance to the incorporation of rich information. For example, the query model can encode detailed user information and searching context when issuing the query (time, location, searcher’s mood, etc.); the image model can encode information of uploader, annotator and comments of the candidate images. In this paper, we incorporate user interest into the query model and annotator information into the image model.

We assume each candidate image $d$ is associated with visual content $v_d$, and annotation $w_d$. The generative process of query $q$ and image $d$ is illustrated in Fig. 4. $\theta_Q$ denotes the parameter of the query model by user $u$, $\theta_D$, $\theta_D^v$ denote the parameters of the image model for generating the visual content and textual annotation. For each image $d$, there are two hidden relevance variable $r(q, u, v_d)$ and $r(q, u, w_d)$, which depends on $[\theta_Q^v, \theta_Q^v]$ and $[\theta_D, \theta_D^v]$, respectively. User $u$ generates the query $q$ by first selecting $\theta_Q^v$ and then choose a query from the selected model. Similar hierarchical process happens to image visual content $v_d$ and textual annotation $w_d$. Therefore, the generative process in Fig. 4 can be summarized by:

$$u \rightarrow \theta_Q^v \rightarrow q, V \rightarrow \theta_D^v \rightarrow v_d \text{ and } W \rightarrow \theta_D^v \rightarrow w_d.$$ 

We consider the task of personalized image search as returning a list of images to the issued query $q$ according to preference of searcher $u$. In the context of Bayesian decision theory, to each action, there is an associated loss $L$, which, in our case, is the loss for returning an individual image to the searcher. Under this framework, the expected risk of returning individual image $d$ is decomposed into two components:

$$L(q, u, d) = \mu R(q, u, v_d) + (1 - \mu) R(q, u, w_d)$$

Figure 4: Generative process of query $q$ and image $d$ in personalized image search.

$$R(q, u, d) = \mu R(q, u, v_d) + (1 - \mu) R(q, u, w_d)$$

$$= \mu \int \int \int p(q, u, d|\theta_Q^v, \theta_D^v) p(\theta_Q^v|\theta_Q^v) p(\theta_D^v|\theta_D^v) \times$$

$$L(\theta_Q^v, \theta_D^v, r(q, u, v_d)) d\theta_Q^v d\theta_D^v + (1 - \mu) \int \int \int p(q, u, d|\theta_D^v, \theta_D^v) p(\theta_D^v|\theta_D^v) \times$$

$$L(\theta_D^v, r(q, u, w_d)) d\theta_D^v \theta_D^v$$

where $\mu$ is weight parameter controlling the strength of visual content and tag words, $p(r|\theta_Q^v, \theta_D^v)$ is the probability of generating relevance $r$ given the query and image model parameters $\theta_Q^v$ and $\theta_D^v$. Following the derivation operation from [29], we consider the case that the loss function $L$ depends only on model parameters $\theta_Q^v$ and $\theta_D^v$. Formally, let $L$ be proportional to a distance measure $\Delta(\cdot)$ between model parameters, i.e.,

$$L(\theta_Q^v, \theta_D^v, r) \propto \Delta(\theta_Q^v, \theta_D^v)$$

Intuitively, if the models $\theta_Q^v, \theta_D^v$ are similar, $\Delta(\theta_Q^v, \theta_D^v)$ should be small. With this assumption, we can have

$$R(q, u, d) \propto \mu \int \int \int p(q, u, d|\theta_Q^v, \theta_D^v) p(\theta_Q^v|\theta_Q^v) p(\theta_D^v|\theta_D^v) \times$$

$$\Delta(\theta_Q^v, \theta_D^v) \times (1 - \mu) \int \int \int p(q, u, d|\theta_D^v, \theta_D^v) p(\theta_D^v|\theta_D^v) \times$$

$$\Delta(\theta_D^v, \theta_D^v) \propto \mu p(q, u, d|\theta_Q^v, \theta_D^v) \Delta(\theta_Q^v, \theta_D^v) + (1 - \mu) p(q, u) \Delta(\theta_Q^v, \theta_D^v)$$

$$\theta_Q^v, \theta_D^v$$ are posterior point estimate of the model parameters:

$$\hat{\theta}_Q^v = \arg\max \theta_Q^v p(q, u)$$

$$\hat{\theta}_D^v = \arg\max \theta_D^v p(q, u)$$

Since $p(\theta_Q^v|q, u)$ does not depend on $d$ and we further assume $p(\theta_D^v|v_d)$ and $p(\theta_D^v|w_d)$ are the same for all $d$ for ranking. The risk minimization framework finally reduces to measurement of the similarity between LMs. We employ the Kullback-Leibler divergence to measure $\Delta(\cdot)$:

$$R(q, u, d) \propto \mu \Delta(\theta_Q^v, \theta_D^v) + (1 - \mu) \Delta(\theta_Q^v, \theta_D^v)$$

$$\propto \mu \sum_{t} p(Z_t|\theta_Q^v) \log \frac{p(Z_t|\theta_Q^v)}{p(Z_t|\theta_D^v)} + (1 - \mu) \times$$

$$\sum_{t} p(Z_t|\theta_D^v) \log \frac{p(Z_t|\theta_Q^v)}{p(Z_t|\theta_D^v)}$$

(11)
With the expected risk for returning individual image \( R(q, u, d) \), we make simplification by further assuming the risk of returning each image is independent of returning others. It can be easily derived that, the final rank of each image \( \text{rank}(q, u, d) \) which minimizes the overall risk is inversely proportional to its individual risk:

\[
\text{rank}(q, u, d) \propto \frac{1}{R(q, u, d)} \quad (12)
\]

In the following subsection, we will instantiate the query and image LMs by incorporating annotation confidence and topic-sensitive influences.

### 4.2 Query and Image Language Models

In personalized search, the query model should consider both the query and user information. Following the topic-sensitive influence modeling discussed in previous sections, we define query model \( \theta_Q^u \) as the distribution over the learned topics given query \( q \) and user \( u \):

\[
\theta_Q^u \triangleq p(z|q, u) = \frac{p(z|u)p(q|z, u)}{p(q|u)} \quad (13)
\]

Since the denominator \( p(q|u) \) does not affect ranking, and \( q \) has no direct relation with \( u \), we have

\[
\theta_Q^u \propto p(z|u)p(q|z) \propto p(z|u) \prod_{w_i \in q} p(w_i|z) \quad (14)
\]

where query \( q \) = \( w_1, \ldots, w_{n_q} \), \( p(z|u) \) and \( p(w_i|z) \) can be directly obtained from user topic distribution \( \Omega_u \) and topic-word distribution \( \Phi^w \). Note that we actually utilize the unigram language model in this paper.

From the standard language model-based information retrieval in [29], image model \( \theta_D^v \) can be directly represented as the topic distribution of the image visual content \( v_d = \langle v_{d1}, \ldots, v_{dn_v} \rangle \):

\[
\theta_D^v \triangleq p(z|v_d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(z|v_{di}) = \frac{1}{n_d} \sum_{i=1}^{n_d} \frac{p(v_{di}|z)p(v_{di})}{p(v_{di})} \quad (15)
\]

where \( n_d \) indicates the number of visual descriptors in image \( d \), \( p(v_{di}|z) \) is the topic-visual descriptor distribution, \( p(v_{di}) \) is the topic prior distribution and \( p(v_{di}) \) is visual descriptor prior distribution \( 4 \).

We consider noisy issue of user-generated annotations when representing image model \( \theta_D^v \). Besides aggregating each tag’s topic-word distribution, we incorporate the annotator authority as annotation confidence as weight for the corresponding tag word. Formally, we have:

\[
\theta_D^v \triangleq p(z|w_d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(z|u_{wdi}) \frac{p(w_{di}|z)p(w_{di})}{p(w_{di})} \quad (16)
\]

where \( n_d \) indicates the number of tag word in annotation of image \( d \), \( u_{wdi} \) is the annotator for tag \( w_{di} \). Since user’s dominant topic distribution can be viewed as his/her expertise, we employ the annotator’s topic distribution \( p(z|u_{wdi}) \) to represent the authority on topic \( z \), which determines the trust we place on each tag. Now we have introduced the basic personalized image search framework which is based on risk minimization. As discussed at the beginning of this section, one important advantage of language model is its expansibility. We have successfully encoded annotation confidence to tackle with noisy user-generated social tags by Eq. (16). In the

\[4\] We assume that visual descriptor prior and tag word prior all follow Uniform distribution, i.e., \( p(v_{di}) = \frac{1}{|\mathcal{V}|} \) and \( p(w_{di}) = \frac{1}{|\mathcal{T}|} \)

following, we will discuss how to incorporate the topic-sensitive influence analysis between users into the proposed framework.

Based on the above formulation, we can compute a personalized risk value for each individual image when searching by \( u \) as \( R(q, u, d) \). In the context of influence network, searcher’s influence will impact the personalized rank by modifying the risk as follows:

\[
R(q, u, d) = \rho R(q, u, d) + \frac{1 - \rho}{|C_u|} \sum_{c \in C_u} \left( \sum_{t \in T} p(Z_t|u, c)p(Z_t|q) \right) R(q, c, d) \quad (17)
\]

where \( R(q, c, d) \) is the risk value when searching by user \( c \), \( \rho \) is weight term balancing contributions from the searcher and his/her influencers, \( p(Z_t|u, c) \) is the topic-sensitive influence strength between user \( u \) and \( c \) on topic \( Z_t \). The intuition here is that, if one influencer has high influence to the searcher on the query-concerning topic \( \{ \sum_{t \in T} p(Z_t|u, c)p(Z_t|q) \} \), his/her preference to each candidate images \( R(q, c, d) \) should be emphasized and attached larger weight. Recall the toy example in Fig. 1, since word “Tahiti” closely relates to the travel topic, when Bob searches “Tahiti”, Tom’s suggestion will be highly valued and contribute most to the final decision.

### 5. EXPERIMENTS

#### 5.1 Dataset

The data are collected from a large-scale Flickr dataset NUS-WIDE [31], which consists of 269,648 images. We crawled the image’s uploader information and obtained 5,120 unique users. Since the focus in this paper is social influence analysis from user perspective, we remove the users uploading less than 15 images and conduct experiments on the remaining users and their images. For each user, we collected his/her annotation set and contact list. If there exists contact relation between two users, we refer to it as one contact edge. The statistic of resulted dataset is shown in Table 2.

In this work, we choose to represent the image visual content by region-level Maximally Stable Extremal Region (MSER) feature. Compared with key-point based descriptor, MSER regions indicate local homogeneous parts in objects and shows higher distinctness [32]. In our implementation, each detected elliptical region from MSER is normalized into a circular region, from which a
SIFT descriptor is computed. About 6M MSER descriptors are extracted from sampled 50,000 images, which are further quantized to constitute a dictionary of 5,000 visual words. In the following, we will show the experimental results for topic-sensitive influencer mining and personalized image search, respectively.

5.2 Topic-sensitive Influencer Mining

5.2.1 Topic Number $T$ Selection

In topic modeling, the selection of topic number $T$ is not trivial. We resort to the perplexity measure in this paper, which is a standard measure for estimating how well one generative model fits the data. The lower the perplexity score is, the better the performance. The perplexity of a set of tag-visual descriptor test pairs $(w_d, v_d)$, for all $d \in D_{test}$ images, is defined as the exponential of the negative normalized predictive log-likelihood using the trained model. Formally,

$$\text{Perplexity}(D_{test}) = \exp \left\{ -\frac{\sum_{d \in D_{test}} \ln p(w_d|v_d)}{\sum_{d \in D_{test}} n_d^w + n_d^v n + 1} \right\}$$

(18)

where

$$p(w_d|v_d) = \prod_{(w_d, v_d) \in (w_d, v_d)} \sum_{t=1}^T p(w_d|Z_t) p(Z_t|v_d)$$

We first test the perplexity of model on a held-out set of 2,000 images for different settings of topic number $T$. The perplexity scores of the proposed mmTIM over iterations are shown in Fig. 5. We can see that the perplexity scores decreases dramatically during the first several iterations and converges to a stable level within 20 iterations. Since larger topic number requires more computation cost and captures weak semantics, we prefer the smallest $T$ that yields small perplexity and fast convergence. It is clear that the perplexity decreases much slower when $T > 20$ and we choose the desired topic number $T$ to be 20 for mmTIM in the following experiments.

5.2.2 Illustration of Discovered Topics

In order to interpret the derived latent topic space, we visualize two of the discovered 20 topics by providing five top-ranked tag words and the most five related images. As represented in Fig. 6, the tag words are sorted by their probability of being generated from the corresponding topic $p(w_i|Z_t)$, while images are sorted by counting the topic indicator variables of visual descriptors and tag words $p(Z_t|w_d, v_d)$:

$$p(Z_t|w_d, v_d) = \frac{\sum_{i=1}^{n_d^w} I(z^w_{d,i} = t) + \sum_{i=1}^{n_d^v} I(z^v_{d,i} = t)}{n_d^w + n_d^v}$$

where $z^w_{d,i}$ and $z^v_{d,i}$ are the topic assignments for the $i$-th tag word and visual descriptor for image $d$, and $I(\cdot)$ is indicator function returning 1 is it is true and 0 otherwise.

By providing a combination of representative words and image, it becomes very easy to interpret the domain knowledge associated with each topic. We can see that, by considering both textual tag words and visual image content, the discovered latent topics show high consistency between semantic concepts and visual themes.

5.2.3 Qualitative Case Study

We demonstrate the effectiveness of mmTIM on topic-sensitive influencer identification of sampled users. Fig. 7 shows two test users and influencers who impact them most on discovered Topics #2 and #13. We show the node topic distribution by lengthy color bars, which indicate users’ interest proportion on the corresponding topics. Since users explicitly express their preferences by adding Favorite marks to images, the exemplary favorite images are displayed below the two test users for user preference illustration. For each influencer, we provide the number of their followers, five exemplary images from their uploaded image set and the tag clouds. While #follower indicates the general influence strength, the uploaded images and annotations offer knowledge about the influencers’ topic-related expertise.

We can see that, mmTIM shows its capability in identifying the most influential contact users on topic level. The identified influencers have high #follower and show strong expertise on the corresponding topics. For example, user “95386698@N00” has major interest in topic #2 and #13. Combining the large #follower and the decent design and fashion images uploaded by user “26324110@N00”, we may roughly conclude that he/she is an expert on fashion topics, i.e., topic #13.
5.2.4 Quantitative Evaluation

Now we conduct quantitative evaluation of topic-sensitive influencer mining. Recall that the goal is to identify contact users with the most influence to the target user on specific topic. Observing that many of the images to which user add Favorite marks are uploaded by his/her contact users, we leverage user’s Favorite images to generate ground-truth information for the evaluation. Specifically, for each Favorite image \( d \) of test user \( U_i \), we calculate its topic proportion and assume the image belongs to the dominant topic, which is denoted as \( z_d \). On each topic, the contact user \( U_j \) owning the top \( k \) images \( \text{Favorite} \) by \( U_i \) is considered most influential to \( U_i \) on this topic. Formally, the most influential contact user to user \( U_i \) on the \( t \)-th topic is defined as:

\[
I_{U_i}^t = \arg \max_{U_j \in C_{U_j}, d \in F_{U_i}} \sum_{u \in U_j} I(z_d = t, u_d = U_j)
\]

where \( u_d \) is uploader of image \( d \), \( F_{U_i} \) is Favorite image set of \( U_i \).

We consider the following two topic-sensitive influence modeling methods for comparison:

- **Topical Affinity Propagation (TAP [16]):** a method learning topic space and topic-sensitive influences separately by inputting the nodes’ topic mixtures;
- **Mining Topic-level Influence on Heterogeneous networks (mTH [17]):** a probabilistic model exploiting the heterogeneous link and textual content information, which targets at text-based citation networks.

Note that the topic space of TAP is pre-extracted by running a standard LDA on the user-annotation corpus with each user’s annotation set as one document and specific tag as word, and the citation link between users in [16, 17] is replaced by the contact relation in our implementation.

We utilize top-k accuracy as the evaluation metric. For each test user \( U_i \), we rank the contact users by their influence values on the \( t \)-th topic. Denoting the rank of ground-truth influencer \( I_{U_i}^t \), in contact list \( C_{U_i} \) as \( \pi(I_{U_i}^t) \), the top-k accuracy is calculated as:

\[
\text{Accuracy}(k) = \frac{\sum_{t,U_i} I(\pi(I_{U_i}^t) \leq k)}{T \cdot |U|} \tag{19}
\]

The results are shown in Fig.8. We can see that the proposed mmTIM consistently outperforms the two baselines. Note that the average #contact user is around 15. The top-1 accuracy of mmTIM is about 25%, which indicates that one out of four trials, mmTIM succeeds to identify the real topic-sensitive influencer at the first rank. The problem of TAP can be summarized in two-fold: First, TAP assumes that topic space and node topic distribution are available before social influence modeling. However, the pre-defined topic space extracted from user-annotation may be insufficient to capture the semantics in social influence links. Second, when modeling topic-sensitive influence, TAP only utilizes the link information of user-contact user, which loses information of relations between users and images. mTH only utilizes the link information of user-contact user, which loses information of relations between users and images. mTH obtains better performance than TAP by addressing the above two problems. However, mTH is proposed for text-based citation networks and does not address the multi-modality issue, which results in the inferior performance to mmTIM. Moreover, compared with mmTIM, mTH is more focused on the document-level (i.e., image or paper) instead of user-level, by explicitly building the influence set from user-document relations. While, under social media settings, the influence set of specific user is actually available by follow or friend list.

5.3 Personalized Image Search Evaluation

The derived topic-sensitive influence values can be applied to applications like social search, friend recommendation, group suggestion, etc. Based on the proposed risk minimization-based theoretical framework, we evaluate the effectiveness of mmTIM on personalized image search and topic-based image recommendation.

In the query and image language models, according to Eq. (17), there is a weight parameter \( \rho \) controlling the strength of searcher and influencer. We choose \( \rho \) based on the assumption that the searcher himself/herself should be more trustworthy if he/she has much annotation activities, otherwise the influencer should be more trusted. Formally, for searcher \( U_i \), \( \rho \) is set as:

\[
\rho = \frac{|T_{U_i}|}{|T_{U_i}| + \frac{1}{|C_{U_i}|} \sum_{U_j \in C_{U_i}} |T_{U_j}|}
\]

By this setting, if the searcher owns more #tag than the average #tag of his/her influencers, \( \rho > 0.5 \), and \( R(q, u, d) \) will contribute more the final rank list. Otherwise, the influencer’s preference \( R(q, c, d) \) will be emphasized more.

5.3.1 Personalized Image Search

Since users’ tagging activities indicate their personal relevance judgement, we employ social annotations for personalized search evaluation. The main assumption is that the images tagged by user \( U_i \) with tag \( w_j \) will be considered relevant if \( U_i \) issues \( w_j \) as a query. The random selected 100 users who tagged 50–100 images constitute the test user set \( U_{test} \). The overlapping 21 tags the 100 users used constitute test query set \( T_{test} \). In order to reduce the dependency between user annotation and evaluation, for the training process, we remove the tagging data related to the test queries.

Based on the proposed personalized image search framework, we consider the following four settings:

- **Basic**: basic personalized image search based on Eq. (12), which computes \( \theta_{U_i}^q \) without considering annotation confidence;
- **Basic + annotation confidence (Basic_AC)**: basic personalized image search leveraging annotator authority as annotation confidence in Eq. (16);
- **Social network with global influence (Social_global)**: social-based personalized image search by considering global influence from contact users, which modified Eq. (17) as:

\[
\hat{R}(q, u, d) = \rho R(q, u, d) + \frac{1 - \rho}{|C_u|} \sum_{c \in C_u} \left( \sum_{t} p(Z_t | u, c) \right) R(q, c, d)
\]

We estimate the global influence by simply aggregating influences over topics which is irrelevant to queries;

- **Social network with topic-sensitive influence (Social_topic)**: social-based personalized image search by considering topic-sensitive influences computed from mmTIM.
We use mMAP as the evaluation metric, which is the mean of the Mean Average Precisions (MAP) for all test users. mMAP is defined as:

$$mMAP = \frac{\sum_{t=1}^{|[U_{test}]|} \sum_{j=1}^{|[T_{test}]|} AP(i, j)}{|U_{test}|}$$

(20)

where $AP(i, j)$ denotes the Average Precision value of the $j$-th test query for the $i$-th user. The results are shown in Fig. 9(a). It is shown that Basic_AC outperforms Basic by 20%, which demonstrates the effectiveness of addressing the noisy annotation issue by incorporating annotator authority. By further considering the influence from contact users, it is weird that the performance of Social_global is inferior to that of Basic_AC. The reason may be that Social_global ignores the relation to the query and treats the social influence as constant for all topics. This result coincides with the motivation illustrated by the toy example we introduce at the beginning of the paper. By considering the topic-sensitive influence values for different queries, Social_topic obtain better performance than both Basic_AC and Social_global, which validates the advantage of the proposed framework to incorporate topic-sensitive influence modeling for personalized search.

5.3.2 Topic-based Image Recommendation

In social media websites, users will explicitly fill in interest profile or express their interests by following certain themes, e.g., sport, music, travel, etc. In this case, users will appreciate if resources based on the expressed user preferences are recommended. In this paper, we simulate the process of expressing interests by assuming user selecting one extracted topic and evaluate the performance of topic-based image recommendation.

The Favorite images are again employed as ground-truth for evaluation. The intuition is that, if image $d_j$ is marked Favorite by user $U_i$ and its dominant topic is $Z_t$, then $d_j$ is treated as relevant when $U_i$ selects topic $Z_t$. In our experiments, 100 randomly selected users who marked Favorite to 50 – 100 images constitute the test set and a total of 7, 25 Favorite images constitute the relevant image set for the topic-based test queries. Following the proposed risk minimization-based framework in Eq. (11) (12) (17), query $q$ is replaced by the selected topic $Z_t$ and we can compare between the three topic-sensitive influence modeling methods: TAP, mTIH and mmTIM. mMAP is utilized for evaluation and the results are demonstrated in Fig. 9(b). It is shown that the proposed mmTIM obtains better performance than the other two methods, which is consistent with the results shown in Fig. 8. This further validates the effectiveness of incorporating multi-modal information for image-related tasks. The fact that mmTIM and mTIH outperform TAP coincides with our motivation that more accurate topic-sensitive influence modeling contributes to better recommendation and search results.

6. CONCLUSIONS

In this paper, a multi-modal probabilistic model is proposed to deal with the problem of topic-sensitive influencer mining in social sharing websites. The proposed model is delicate and informative, which provides better representation of the influence network and social annotation activity. The derived topic-sensitive influence has demonstrated its effectiveness through our introduced LM-based general framework in applications of personalized image search and topic-based image recommendation. We hope this paper could serve as a good chance to further the agenda of social relation analysis in this community.

This work can be extended in the following four directions: (1) We exploit social annotation for the influence modeling in this paper. How the social influence correlates with other social activities, like Favorite, share, comment, rate, is not clear and remains an interesting question. (2) Since the semantic tag and low-level visual feature account for different-level descriptions of the image, it is arbitrary to mix them in the same topic. To develop models allowing for more flexible sampling of word topics and visual topics will be another future direction. (3) We have demonstrated the potential of topic-sensitive influence modeling by two simple applications. Based on the proposed generative model, more real applications can be designed, such as friend suggestion, user behavior prediction, and so on. (4) The proposed personalized image search framework provides one of the first theoretical attempts to address this challenging problem. It is important to find the correlations with other personalized search methods in the future.

7. ACKNOWLEDGEMENT

This work was supported in part by National Program on Key Basic Research Project (973 Program, Project No. 2012CB316304) and the National Natural Science Foundation of China (Grant No. 909020303, 61003161).

8. REFERENCES


[18] We derive the first component of Eq. (22). For the numerator, since $p(s^w_\alpha) \propto \int p(s^w_\alpha | \lambda) \cdot \lambda \cdot \prod_{i=1}^{d} \lambda_i^{w_i} \cdot \Phi_{\lambda}^{-1}(\lambda)$, we have

$$\int p(s^w_\alpha | \lambda) \cdot \lambda \cdot \prod_{i=1}^{d} \lambda_i^{w_i} \cdot \Phi_{\lambda}^{-1}(\lambda) \, d\lambda = \int p(s^w_\alpha | \lambda) \cdot \lambda \cdot \prod_{i=1}^{d} \lambda_i^{w_i} \cdot \Phi_{\lambda}^{-1}(\lambda) \, d\lambda$$

The second component of Eq. (22) can be derived analogously.