Learning to Recommend Descriptive Tags for Questions in Social Forums

LIQIANG NIE, YI-LIANG ZHAO, and XIANGYU WANG, National University of Singapore
JIALIE SHEN, Singapore Management University
TAT-SENG CHUA, National University of Singapore

Around 40% of the questions in the emerging social-oriented question answering forums have at most one manually labeled tag, which is caused by incomprehensive question understanding or informal tagging behaviors. The incompleteness of question tags severely hinders all the tag-based manipulations, such as feeds for topic-followers, ontological knowledge organization, and other basic statistics. This article presents a novel scheme that is able to comprehensively learn descriptive tags for each question. Extensive evaluations on a representative real-world dataset demonstrate that our scheme yields significant gains for question annotation, and more importantly, the whole process of our approach is unsupervised and can be extended to handle large-scale data.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Question annotation, social QA, knowledge organization

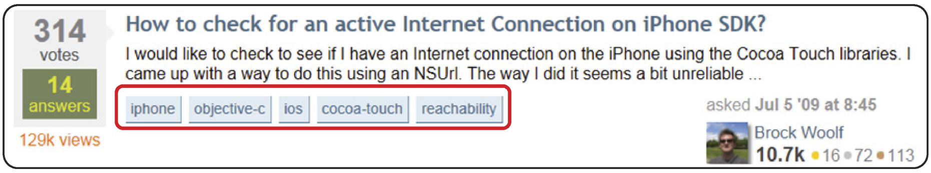
ACM Reference Format:
Nie, L., Zhao, Y.-L., Wang, X., Shen, J., and Chua, T.-S. 2014. Learning to recommend descriptive tags for questions in social forums. *ACM Trans. Inf. Syst.* 32, 1, Article 5 (January 2014), 23 pages.
DOI:http://dx.doi.org/10.1145/2559157

1. INTRODUCTION

Recent years have seen a flourishing of community-driven question answering (cQA) portals, which have emerged as an effective paradigm for disseminating diverse knowledge, seeking precise information, and locating outstanding expert. Yahoo! Answers and AnswerBag are typical examples of such cQA services with widespread adoption. With deep exploitation of human intelligence, they have largely relieved the challenges faced by traditional automatic QA, such as deep understanding of complex questions, sophisticated syntactic information extraction, and semantic answer summarization [Nie et al. 2011, 2013]. However, along with the proliferation of accumulated QA pairs and community participants, cQA has exposed several intrinsic problems, including conversational trends, varying content qualities, overwhelming numbers of fresh questions, data redundancies, and fixed taxonomies.

This work was supported by NUS-Tsinghua Extreme Search project under the grant number: R-252-300-001-490.
Author's address: L. Nie; email: nieliqiang@gmail.com.

DOI:http://dx.doi.org/10.1145/2559157



314
votes
14
answers
129k views

How to check for an active Internet Connection on iPhone SDK?

I would like to check to see if I have an Internet connection on the iPhone using the Cocoa Touch libraries. I came up with a way to do this using an NSUrl. The way I did it seems a bit unreliable ...

iphone objective-c ios cocoa-touch reachability

asked Jul 5 '09 at 8:45
Brock Woolf
10.7k ● 16 ● 72 ● 113

Fig. 1. Illustration of question-tag instance selected from stack overflow.

Table I. Average Number of Tags over Questions

Tag Number	Zero	One	Two	Above Two
Question Percentage	18.08%	19.34%	18.89%	43.69%

To thoroughly break the cQA dilemmas, more social-oriented QA sites turn up with the exploding growth of social networks, such as Quora[1] and Zhihu.[2] They bring in some fresh features which guide the development of communities toward healthy directions, ensuring canonical and reusable contents. First, they reconstruct the real-life connections by integrating with third-party social platforms, such as Facebook and Twitter. That provides a socially viral adoption and prevents casual or malicious users from valueless operations. Second, they track the reputation of each user from the start by maintaining histories and making them publicly viewable. This allows the community to police themselves for quality content through awarding expert users and disapproving trolls. Third, the spirit of crowdsourcing is boldly attempted via enabling any authenticated user to perform editing on any question, answer, and even topic, which gradually results in refined community content. Fourth, they dramatically improve the response likelihood for newly posted questions through propagating these questions among the tied followers or recommending specific experts with first-hand experiences in terms of social and historical data. Also, they constrain duplicate questions through surfacing textually similar questions in the search interface and encouraging users to refer to these existing questions before adding a new question.

Besides the aforementioned features, the most transformative initiative is question annotation due to the following facts. First, it enables each question to have multiple manually assigned topics without constraints on the vocabulary. These tags summarize question content in a coarse-grained but semantically meaningful level. One typical example of question annotation is demonstrated in Figure 1. Second, it leads to a faster and more accurate answer. Tags immediately put the question into the feeds of related topic-followers, which can aggregate more attention, including attention from experts. Third, question annotation naturally and greatly facilitates the first-order processing of user-generated content from various angles, such as statistics, sharing, indexing, and searching [Luhn 1958]. Last, but most important, question annotation benefits higher-order knowledge exploration, such as hierarchical structure organization.

However, the incomplete characteristic of question annotation is statistically observed as a noticeable phenomenon, as illustrated in Table I, which is gathered from approximately 200,000 questions crawled from Zhihu. It shows that more than 37% of questions contain zero or at most one tag. The incompleteness of tags is caused by incomprehensive question understanding or informal tagging behaviors. This severely hinders the performance of tag-based systems. For example, the performance of

[1] https://www.quora.com/
[2] http://www.zhihu.com/

question search may be degraded because of the absence of potentially relevant tags which are utilized to expand the questions. Also, it limits higher-level tasks of ontological question organization and effective question routing. Automatic question annotation with available tag vocabulary is the most straightforward approach to tackling this difficulty. However, some user-provided tags in the vocabulary are often biased towards personal perspectives or specific contextual information [Agarwal et al. 2006], which are usually subjective and inconsistent with the more frequently used terms. These types of tags are not very stable or reliable. Conversely, some tags with extremely high frequencies are too broad to describe individual question content. Therefore, new approaches towards automatically refilling or enlarging questions with objective and informative tags are highly desired.

It is worth mentioning that there already exist several efforts dedicated to annotation which roughly fall into two categories: media entity annotation and textual entity annotation. The former aims to make visual and audio entities more accessible to the general public. This is accomplished via either estimating the joint probabilities between a tag and the given extracted media features [Duygulu et al. 2002; Jeon et al. 2003; Monay and Gatica-Perez 2004; Mori et al. 1999; Sigurbjörnsson and van Zwol 2008; Yang et al. 2013] or treating tags as classes and employing the trained classifiers to annotate the media entities [Carneiro and Vasconcelos 2005; Fu et al. 2011; Kang et al. 2006; Qi et al. 2007; Xiang et al. 2009; Yang et al. 2006]. However, these technologies can hardly be applied to question annotation directly due to different modalities between tags and entities. Specifically, both tags and questions are in terms of text, and hence, the lexical properties can be utilized to partially infer the relevance scores between tags and questions. However, this is impossible in media annotation. Textual entity annotation is also promising in organizing, indexing, and searching textual resources. Textual entity may be individually annotated by extracting interesting terms from the document itself [Brooks and Montanez 2006; Luhn 1958; Narr et al. 2011; Subramanya and Liu 2008; Wu et al. 2010], or may be collaboratively annotated via globally selected tags from the whole collection by considering both coverage and popularity at the same time [Mishne 2006; Sood et al. 2007; Xu et al. 2006]. However, these approaches either view the textual entities independently or ignore the connections among the attributes of entities, let alone the relationships among users. Another reason that results in these methods are not applicable to question annotation is that these textual documents are much longer than questions.

Annotating questions with appropriate tags poses new challenges due to the following reasons. First, unlike normal documents, these questions are typically short. They thus do not provide sufficient word co-occurrences or shared contexts for effective similarity measure. The data sparseness hinders general machine learning methods in achieving desirable accuracy, and thus, in-depth mining of other cues to compensate for this limitation is neccessary, such as answer knowledge and tag sharing information. Second, differently from traditional community QA, social QA highlights the social relationship among users. These social connections hold considerable potential value for question annotation because they are able to transmit and propagate tags among a group of questions. These questions are asked by a group of users who are linked by a common characteristic, such as mutual interests. Therefore, the relationships should be incorporated into the model to boost the annotating performance. Third, even though various benchmarks on QA are available online, the representative social QA data for question annotation cannot be easily collected.

To solve this problem, we hypothesize the following.

(1) Similar questions from the same semantic space share the same tag space.
(2) Tags with excessive or rare occurrences are less informative or stable.

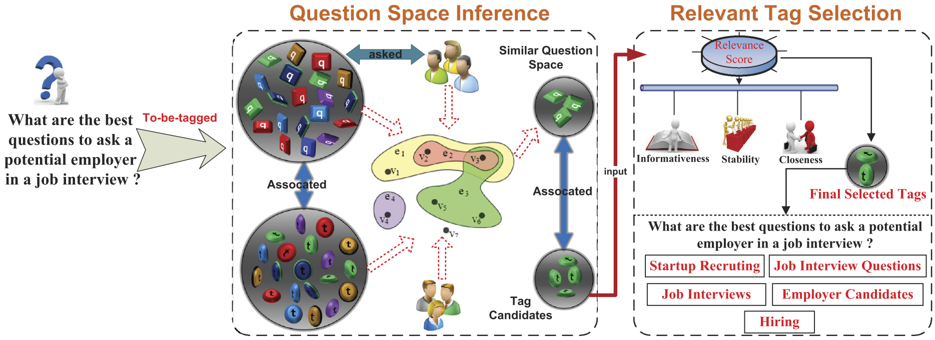


Fig. 2. The schematic illustration of the proposed automatic question annotation scheme for social QA services. Given a question, our scheme automatically recommends tags to this question, comprising two components. The first component attempts to construct a hypergraph by integrating multiple facets, including QA content analytics, tag-sharing information, as well as user connections. The outcome of the first component is a similar question space, and then the potential tag candidates can be collected. The second component comprehensively selects the most descriptive tags from the tag candidates by simultaneously considering informativeness, stability, and closeness.

These two assumptions have motivated us in proposing a novel scheme which is to enhance automatic question annotation by exploring the cues from both content analytics and social tagging behaviors. The scheme comprises two main components, as illustrated in Figure 2. The first component roughly identifies a collection of probably relevant tags via finding a similar question space, which aims to narrow down the suggested tag candidates. This task is accomplished through an adaptive probabilistic hypergraph learning, where the vertices denote the questions, and hyperedges are generated in terms of QA content analysis, tag sharing networks, as well as user social behaviors. The distinction among hyperedge influences is considered with a regularizer on hyperedge weights. The learning process iteratively and alternatively updates between the vertex relevances and hyperedge weights until convergence is reached. The second component deals with relevant tag selection by taking into consideration informativeness, stability, and question closeness at the same time. It intends to comprehensively evaluate each tag candidate and select the most appropriate tags for annotation. The whole process of our approach is unsupervised and can be extended to handle large-scale data.

By conducting experiments on the representative real-world dataset, we demonstrate that our proposed scheme achieves significant gains in question annotation.

The main contributions of this research are as follows.

- To the best of our knowledge, this is the first work that targets automatic question annotation for emerging social QA services. This work unravels the incomplete and biased problems of question tags.
- It proposes an adaptive probabilistic hypergraph learning approach to identify semantically similar question space. Intrinsically different from the conventional hypergraph learning with fixed hyperedge weights, our approach iteratively updates the weights to really reflect the different effects of users, questions, and tags.
- It proposes a heuristic approach to further filter the tag candidates by jointly integrating multifaceted cues simultaneously, that is, tag informativeness, tag stability, and question closeness. This effectively eliminates subjective, ambiguous, and generic tags.

The remainder is structured as follows. Sections 2 and 3, respectively, introduce the related work and the annotation scheme. In Section 4, we discuss our adaptive probabilistic hypergraph learning approach for question-space inference. Section 5 details the proposed heuristic approach for relevant tag selection. Experimental results and analysis are presented in Section 6, followed by our concluding remarks in Section 7.

2. RELATED WORK

2.1. Annotation of Media Entities

The prevalence of visual and audio capture devices and the growing popularity of media-sharing technologies have created massive multimedia content available online which are distributed in community-contributed sites, such as Flickr, Youtube, and MeeMix. Meanwhile, the user-generated tags play an essential role in making these media entities more accessible to the general public [Ames and Naaman 2007] via summarizing low-level features with semantic descriptors [Wu et al. 2011]. However, the grassroot-provided social media tags suffer from labor-intensive [Tang et al. 2010], incomplete [Liu et al. 2010], biased [Golder and Huberman 2006], and imprecise issues [Chua et al. 2009; Liu et al. 2009]. Several recent studies from multimedia, computer vision, as well as machine learning domains have been conducted to tackle these issues. These efforts can be broadly categorized into generative model approaches [Duygulu et al. 2002; Jeon et al. 2003; Monay and Gatica-Perez 2004; Mori et al. 1999; Sigurbjörnsson and van Zwol 2008] and discriminative model approaches [Carneiro and Vasconcelos 2005; Fu et al. 2011; Kang et al. 2006; Qi et al. 2007; Xiang et al. 2009; Yang et al. 2006].

The idea behind the generative approaches is to annotate visual or audio entities by estimating the correlations or joint probabilities between a tag and the given extracted features. The candidates with the highest probabilities could then be reserved as the final recommended tags. A variety of statistical machine learning models have been successfully applied to automatic media data annotation, such as the co-occurrence model [Mori et al. 1999], machine translation model [Duygulu et al. 2002], latent space analysis [Monay and Gatica-Perez 2004], as well as relevance language model [Jeon et al. 2003]. The proposed co-occurrence model [Mori et al. 1999] counted co-occurrences of words and image regions created using a regular grid. Three years later, it was improved by a classical machine translation model via translating vocabulary of visual blobs to textual tags [Duygulu et al. 2002]. Following that, Monay et al. utilized latent semantic analysis to capture co-occurrence information between low-level features and concepts [Monay and Gatica-Perez 2004]. In addition, a cross-media relevance model [Jeon et al. 2003] was introduced to annotate media data and observed improved effectiveness compared to translation models.

Alternatively, the discriminative approaches apply classification techniques by treating tags as classes and by employing the trained classifiers to annotate an input entity. Earlier studies were devoted to develop binary classifiers, while most recent literature viewed the tagging problem as a multi-class classification task [Xiang et al. 2009]. Yang et al. [2006] presented an asymmetrical support vector machine for region-based image annotation. Carneiro and Vasconcelos [2005] formulated the image annotation as a supervised multi-class problem and learned the distribution model for each class. These aforementioned methods, however, do not explicitly investigate the discriminative information between different classes. Kang et al. [2006] noticed this research issue and deeply exploited the correlations between class labels by extending the standard label propagation algorithms to propagate multiple labels. Furthermore, classifications for automated detection of the video and audio concepts were also comprehensively studied [Fu et al. 2011; Qi et al. 2007].

Though great success has been achieved for entity annotation in the media domain, these techniques cannot be directly applied to the social QA domain due to different modalities between tags and entities.

2.2. Annotation of Textual Entities

For textual entity annotation, tags are also promising in organizing, indexing, and searching textual resources, such as blogs and tweets. Typically, two general approaches exist. One is individually tagging by extracting interesting terms from the post itself. Brooks and Montanez [2006] developed a system to automatically extract three terms with the top TFIDF scores from each post and suggest them as tags. A more sophisticated work was proposed in Narr et al. [2011] by utilizing an advanced natural language processing approach to distill semantic annotations from Twitter and transform them into a reusable knowledge base. Wu et al. [2010] designed a novel system applying the TextRank algorithm to extract personalized tags to label Twitter users' interests and concerns. However, tags extracted from a single individual face the challenge of vocabulary variability [Subramanya and Liu 2008]. To overcome this problem, the second type a approach considers the tags collaboratively contributed by the crowds over a large collection of posts. Xu et al. [2006] developed a method to globally select tags from the whole collection by simultaneously considering the criteria of high coverage, least effort, and high popularity. A system called *AutoTag* automatically suggested tags to the given blog post via finding similar blog posts [Mishne 2006]. The principle of this work is concordant with our first assumption. An improved version of this system named *TagAssist* [Sood et al. 2007] annotates a post by generating search queries from the given post content, searching a collection of blog posts using those queries, and selecting suitable tags from the retrieved posts.

Overall, literature regarding text annotation is still relatively sparse, and the existing approaches either view the entities independently or overlook the social connections of entities' attributes. Most importantly, no reported work touches the annotation problem for one of the dominant thought-exchanging platforms, that is, social QA services.

3. QUESTION ANNOTATION SCHEME

Let $\mathcal{Q} = \{q_1, q_2, ..., q_N\}$ and $\mathcal{T} = \{t_1, t_2, ..., t_M\}$, respectively, represent a repository of questions and their associated tags. The target of this article is to automatically select appropriate tags from $\mathcal{T}$ to annotate a given question q. To accomplish this task, two components are involved, as illustrated in Figure 2.

The first component is question-space inference. It attempts to identify a subset of questions $\mathcal{Q}_s = \{q_1^s, q_2^s, ..., q_n^s\}$ from $\mathcal{Q}$, each of which is semantically close to q. This space is constructed by estimating the semantical similarity score f_i between each q_i ($q_i \in Q$) and q via our proposed adaptive probabilistic hypergraph learning approach. Then, all the associated tags of our inferred question space are straightforward to form a tag space $\mathcal{T}_s = \{t_1^s, t_2^s, ..., t_m^s\}$, $\mathcal{T}_s \subseteq \mathcal{T}$. As a byproduct, this component quantitatively outputs the semantical closeness between each question in $\mathcal{Q}_s$ and q, which is the question to be annotated.

The other component is relevant tag selection. It ranks the tags in $\mathcal{T}_s$ by seamlessly integrating multiple analysis, including informativeness obtained from user tagging behaviours, stability defined based on the neighbour voting approach, and question closeness learned by component one. Based on the ordered tag list, we then select the top tags for the given question.

4. QUESTION-SPACE INFERENCE

In this section, we present our proposed adaptive probabilistic hypergraph learning approach which identifies the semantically similar question space by jointly considering the QA content analysis, tag-sharing networks, as well as users' social behaviors. We first introduce the hypergraph learning theory and then detail the alternating optimization process for our proposed model. Finally, we prove the learning consistency between simple graph and hypergraph.

4.1. Probabilistic Hypergraph Construction

A hypergraph $G(\mathcal{V}, \mathcal{E}, \mathbf{W})$ is composed of the vertex set $\mathcal{V}$, the hyperedge set $\mathcal{E}$, and the diagonal matrix of hyperedge weight $\mathbf{W}$. Here, $\mathcal{E}$ is a family of hyperedges e connecting arbitrary subsets of $\mathcal{V}$, and each hyperedge e is assigned weight $W(e)$. Unlike a simple graph, where each edge only conveys the pairwise relations that overlook the relations in higher order, a hypergraph contains the summarized local grouping information by allowing each hyperedge to link more than two vertices simultaneously. A probabilistic hypergraph G can be represented by a $|\mathcal{V}| \times |\mathcal{E}|$ incidence matrix $\mathbf{H}$ with the following entries:

$$h(v_i, e_j) = \begin{cases} P(v_i, e_j), & \text{if } v_i \text{ is linked by } e_j, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $P(v_i, e_j)$ describes the probability that vertex v_i falls into hyperedge e_j. Based on $\mathbf{H}$, the vertex degree of $v_i \in \mathcal{V}$ is estimated as

$$d(v_i) = \sum_{e_j \in \mathcal{E}} W(e_j) h(v_i, e_j). \tag{2}$$

Then the initial weight for each hyperedge is computed as

$$W(e_j) = \sum_{v_i \in e_j} h(v_i, e_j). \tag{3}$$

The magnitude of the hyperedge weight indicates to what extent the vertices in a hyperedge belong to the same group [Agarwal et al. 2006].

For a hyperedge $e_j \in \mathcal{E}$, its degree is defined as

$$\delta(e_j) = \sum_{v_i \in e_j} h(v_i, e_j), \tag{4}$$

where $v_i \in e_j$ means vertex v_i involves hyperedge e_j. We denote the vertex degrees and hyperedge degrees by $\mathbf{D}_v$ and $\mathbf{D}_e$, respectively.

In our work, the to-be-annotated question q and the N questions from $\mathcal{Q}$ are regarded as vertices, and thus the generated hypergraph has $N + 1$ vertices. Meanwhile, three types of hyperedges are constructed, and Figure 3 gives the illustration. The first type takes each vertex as a centroid and forms a hyperedge by circling around its k-nearest neighbors based on semantical QA content similarities. This procedure was firstly adopted in Huang et al. [2010]. The second type is tag-based by grouping all the questions sharing the same tag. The third one takes users' social behaviors into consideration via rounding up all the questions asked by the same user and its followees.[3] Therefore, up to $N + M + 1 + U$ hyperedges are generated in our hypergraph, where U and M, respectively, denote the number of involved users and tags. For each

[3] If A follows B, then A is B's follower and B is A's followee.

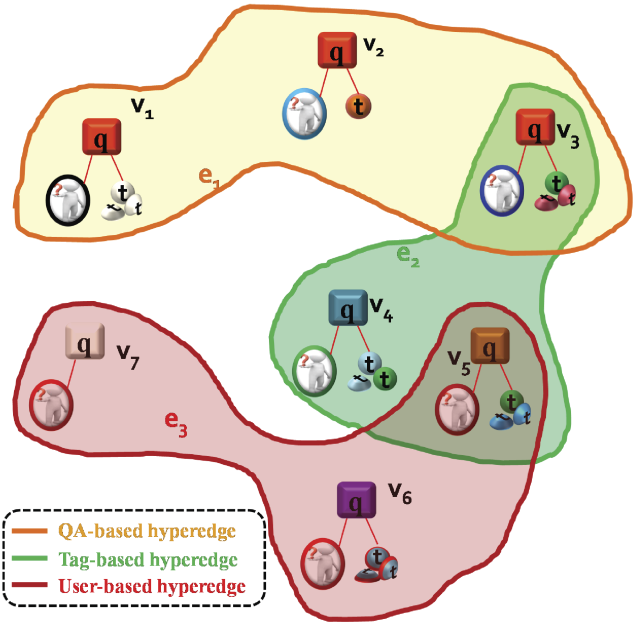


Fig. 3. Illustration of hyperedges construction. Three kinds of hyperedges are involved in our hypergraph which are, respectively, generated by grouping the local sharing information among users, questions, and tags.

hyperedge, the likelihood of each its constituent question belonging to its local group is defined according to its hyperedge type,

$$P(v_i, e_j) = \begin{cases} 1, & \text{Tag-based hyperedges;} \\ K(\mathbf{q}_i, \mathbf{q}_j), & \text{QA-based hyperedges;} \\ 1, & \text{User-based hyperedges.} \end{cases} \tag{5}$$

$K(\cdot, \cdot)$ is the Gaussian similarity function Nie et al. [2012a] defined as

$$K(\mathbf{q}_i, \mathbf{q}_j) = exp(-\frac{||\mathbf{q}_i - \mathbf{q}_j||^2}{\sigma^2}), \tag{6}$$

where the radius parameter, σ, is simply set as the median of the Euclidean distances among all QA pairs.

4.2. Adaptive Probabilistic Hypergraph Learning

We first rank all the questions in $\mathcal{Q}$ in descending order according to their relevance to q, which is estimated via our adaptive probabilistic hypergraph model. We then select the top-n questions to form the semantic space. In this article, the relevance estimation is viewed as a transductive inference problem [Yu et al. 2012; Zhou et al. 2006] formulated as a regularization framework,

$$\arg\min_{\mathbf{f}} \Phi(\mathbf{f}) = \arg\min_{\mathbf{f}} \left\{ \Omega(\mathbf{f}) + \lambda R(\mathbf{f}) \right\}, \tag{7}$$

where $\Omega(\mathbf{f})$ and $R(\mathbf{f})$ denote the regularizer on the hypergraph and empirical loss, respectively. The parameter λ is a regularization parameter for balancing the empirical loss and the regularizer.

Inspired by the normalized cost function of a simple graph [Nie et al. 2012b; Zhou et al. 2004], $\Omega(\mathbf{f})$ is defined as

$$\frac{1}{2}\sum_{e\in\mathcal{E}}\sum_{u,v\in e}\frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{f(u)}{\sqrt{d(u)}}-\frac{f(v)}{\sqrt{d(v)}}\right)^2, \tag{8}$$

where vector **f** contains the relevance probabilities that we want to learn. By defining $\Theta = \mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-\frac{1}{2}}$, we can further derive that

$$\begin{aligned}\Omega(\mathbf{f}) &= \sum_{e\in\mathcal{E}}\sum_{u,v\in e}\frac{w(e)h(u,e)h(v,e)}{\delta(e)}\left(\frac{f^2(u)}{d(u)}-\frac{f(u)f(v)}{\sqrt{d(u)d(v)}}\right)\\ &= \sum_{u\in\mathcal{V}}f^2(u)\sum_{e\in\mathcal{E}}\frac{w(e)h(u,e)}{d(u)}\sum_{v\in\mathcal{V}}\frac{h(v,e)}{\delta(e)}-\\ &\quad\sum_{e\in\mathcal{E}}\sum_{u,v\in e}\frac{f(u)h(u,e)w(e)h(v,e)f(v)}{\sqrt{d(u)d(v)\delta e}}\\ &= \mathbf{f}^T(\mathbf{I}-\Theta)\mathbf{f},\end{aligned} \tag{9}$$

where **I** is an identity matrix. Let $\Delta = \mathbf{I} - \Theta$, which is a positive semidefinite matrix, the so-called hypergraph Laplacian [Zhou et al. 2006], then $\Omega(\mathbf{f})$ can be rewritten as

$$\Omega(\mathbf{f}) = \mathbf{f}^T\triangle\mathbf{f}. \tag{10}$$

For the loss term, after introducing a new vector **y** containing all the initially estimated relevance probabilities, it is stated as a least squares function,

$$R(\mathbf{f}) = \|\mathbf{f}-\mathbf{y}\|^2 = \sum_{v\in\mathcal{V}}\big(f(v)-y(v)\big)^2. \tag{11}$$

By minimizing $\Phi(\mathbf{f})$, the first term in Eq. (7) guarantees that the relevance probability function is continuous and smooth in the semantic space. This means that the relevance probabilities of semantically similar questions should be close. The empirical loss function forces the relevance probabilities to approach the initial roughly estimated relevance scores. These two implicit constraints are widely adopted in reranking-oriented approaches [Nie et al. 2012b; Tian et al. 2008].

However, in the constructed hypergraph, the effects of hyperedges cannot be treated on an equal footing, since they are generated from different angles, spanning from semantical similarities between QA pairs and tag-sharing networks to users' social behaviors. Even though all the hyperedges are initialized with reasonable weights based on local information, further globally adaptive refinement and modulation are still necessary. Inspired by Yu et al. [2012] and Gao et al. [2012], we extend the conventional hypergraph to an adaptive one by integrating a two-norm regularizer on **W**. Therefore, Eq. (7) is restated as

$$\arg\min_{\mathbf{f},\mathbf{W}}\left\{\mathbf{f}^T\triangle\mathbf{f}+\lambda\|\mathbf{f}-\mathbf{y}\|^2+\mu\|diag(\mathbf{W})\|^2\right\}, \tag{12}$$

where μ is a positive parameter. For model simplicity, all the entries in **W** are confined to be nonnegative and sum to one. We alternatively optimize **f** and **W**.

First, **W** is fixed, and derivatives with respect to **f** are taken on the objective function. We have

$$\mathbf{f} = (1-\eta)(I-\eta\Theta)^{-1}\mathbf{y}, \tag{13}$$

where $\eta = \frac{1}{1+\lambda}$. Next, we fix $\mathbf{f}$ and optimize $\mathbf{W}$ with the help of Lagrangian, which is frequently utilized in optimization problems [Gao et al. 2012]. The objective function is transformed into

$$\arg\min_{\mathbf{W},\xi} \left\{ \mathbf{f}^T \triangle \mathbf{f} + \mu \|diag(\mathbf{W})\|^2 + \xi(\sum_i W_{ii} - 1) \right\}. \tag{14}$$

Differentiating the trace of the preceding formulation with respect to $\mathbf{W}$, it can be derived that

$$\mathbf{W} = \frac{\Gamma^T \Gamma \mathbf{D}_e^{-1} - \xi \mathbf{I}}{2\mu}, \tag{15}$$

where Γ denotes $\mathbf{f}^T \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H}$. Replacing $\mathbf{W}$ in Eq. (14) with Eq. (15), and taking derivatives with ξ, we obtain

$$\xi = \Gamma \mathbf{D}_e^{-1} \Gamma^T - 2\mu. \tag{16}$$

In the whole iterative process, we alternatively update $\mathbf{f}$ and $\mathbf{W}$. Each step decreases the objective function $\Phi(\mathbf{f})$ whose lower bound is zero. Therefore, convergence of our scheme is guaranteed [Gao et al. 2012; Yu et al. 2012]. Another noteworthy issue is that the initial relevance probability of each question in $\mathcal{Q}$ to the given question q are estimated based on Eq. (6).

4.3. Discussions

It is intuitive that the conventional simple graph is a special case of hypergraph, where all hyperedges have degree two and represent only pairwise relationships. To further investigate the learning process relationships between these two kinds of graphs, we develop a regularization framework $\Phi_{\mathbf{s}}(\mathbf{f})$ for simple graph.

$$\frac{1}{2} \sum_{i,j} W_{ij} \left(\frac{f_i}{\sqrt{D_{ii}}} - \frac{f_j}{\sqrt{D_{jj}}} \right)^2 + \lambda \sum_i (f_i - y_i)^2. \tag{17}$$

The first term is the normalized cost function controlling the smoothness, where $\mathbf{D}$ is a diagonal matrix with its (i,i)-element equal to the sum of the ith row of the affinity matrix $\mathbf{W}$. Let $\Theta_{\mathbf{s}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}$; the simple graph Laplacian can be denoted as $\Delta_{\mathbf{s}} = \mathbf{I} - \Theta_{\mathbf{s}}$. It can be shown that the first term is equivalent to $\mathbf{f}^T \Delta_{\mathbf{s}} \mathbf{f}$, which is similar to the regularizer on the hypergraph in Eq. (10). Analogous to the empirical loss function of the hypergraph, the second term is utilized to constrain the fitting, which means a good classifying function should not change too much from the initial label assignment [Zhou et al. 2004]. The parameter λ is a regularization parameter to balance the empirical loss and the regularizer. Differentiating $\Phi_{\mathbf{s}}(\mathbf{f})$ with respect to $\mathbf{f}$, we have

$$\mathbf{f} - \Theta_{\mathbf{s}} \mathbf{f} + \lambda(\mathbf{f} - \mathbf{y}) = \mathbf{0}. \tag{18}$$

With $\eta = \frac{1}{1+\lambda}$, we get

$$\mathbf{f} = (1 - \eta)(I - \eta \Theta_{\mathbf{s}})^{-1} \mathbf{y}, \tag{19}$$

where vector $\mathbf{f}$ contains the relevance probabilities f_i $(i \in [1,N])$ that stand for the semantical similarity score between q_i $(q_i \in Q)$ and q. It is observed that the regularization framework of the simple graph and its derived result completely coincide

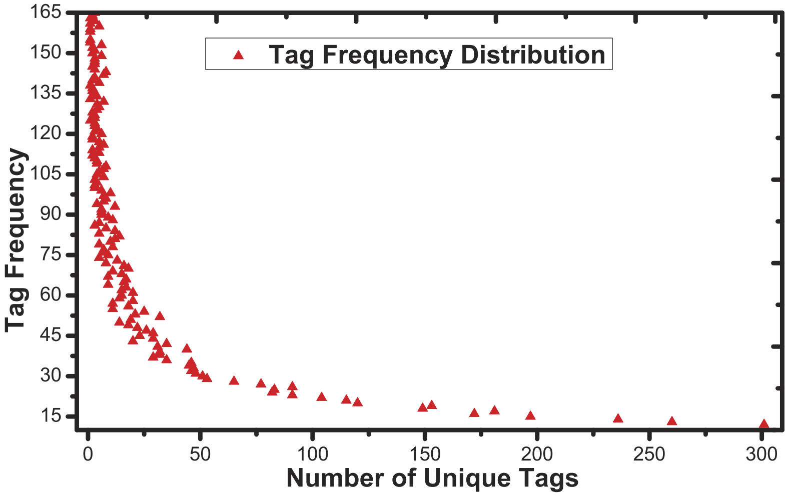


Fig. 4. The tag frequency distribution with respect to the number of distinct tags over our representative dataset.

with those of the hypergraph. This further proves that hypergraphs are generalizations of simple graphs in terms of both intrinsic attributes and corresponding learning approaches.

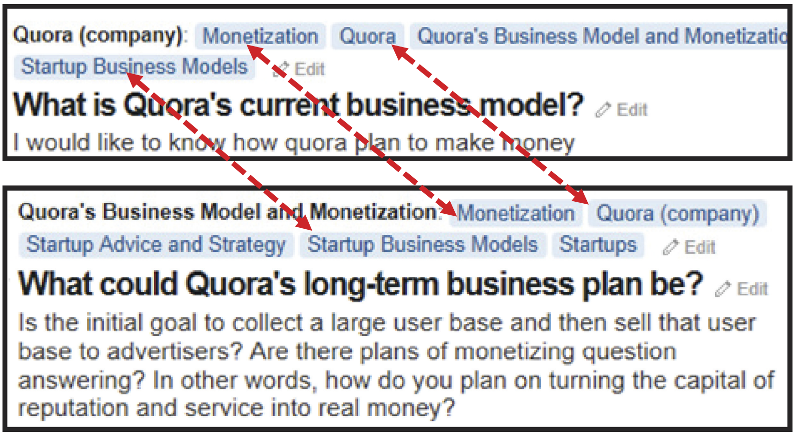
It is worth clarifying the loop training data problem. For each to-be-annotated question q, its tag candidates are automatically selected via its semantically similar question-space inference. The similar question-space is actually a small subset of our question dataset Q, where each question is manually labeled with multiple tags by the website's grassroot users. After automatic annotation, question q and its associated tags will not be added to Q for further prediction. This aims to avoid performance reduction after several rounds of iterations by utilizing the "predicted result" as training data.

5. RELEVANT TAG SELECTION

Based on the first component, a tag space shared by the inferred question space can be generated effortlessly. However, not all the roughly selected tag candidates are able to well summarize the given question content. A heuristic tag relevance estimation approach is proposed in this section to further filter the tag candidates by integrating multifaceted cues. Following that, the complexity of our scheme is analyzed.

5.1. Tag Relevance Estimation

According to our statistics, the tag frequency distribution in our dataset with respect to the number of distinct tags is displayed in Figure 4. We observe that the tags distributed in the head part of the curve tend to be the phrases in high-level semantics, such as "technology", "life", "entertainment", and so on. They are too generic to be informative tags. On the other hand, the tail of the curve contains the tags with very low collection frequencies that are usually extremely specific. They are either unpopular abbreviations, personalized terms, or informal spellings [Li et al. 2009], such as "iSteve", "WEBLOC", etc. Actually, these two phenomenons fit our second assumption. It is also found that the semantically closer two questions are, the higher the probabilities that the tags can be shared between them. This is coherent with our first assumption. The typical example is illustrated in Figure 5.



Quora (company): Monetization Quora Quora's Business Model and Monetizatio
Startup Business Models Edit

What is Quora's current business model? Edit

I would like to know how quora plan to make money

Quora's Business Model and Monetization: Monetization Quora (company)
Startup Advice and Strategy Startup Business Models Startups Edit

What could Quora's long-term business plan be? Edit

Is the initial goal to collect a large user base and then sell that user base to advertisers? Are there plans of monetizing question answering? In other words, how do you plan on turning the capital of reputation and service into real money?

Fig. 5. The illustrative instance of semantically similar questions sharing the same tags.

The aforementioned analysis strongly suggests that the tag relevance estimation should simultaneously damp generic tags, penalize specific tags, as well as reward tags from semantically closer questions. It is mathmatically stated as

$$Score(q, t^s) = D(t^s) \times S(\mathcal{Q}_s, t^s) \times C(q, t^s), \tag{20}$$

where t^s denotes a tag candidate collected from the inferred question space $\mathcal{Q}_s$, and q is the to-be-annotated question.

The first term is the informativeness and descriptiveness measurement which ensures that tags with high frequencies will have lower relevance scores. It is defined as

$$D(t^s) = \frac{1}{\log(o(t^s) + 1)}, \tag{21}$$

where $o(t^s)$ refers to the occurrence frequency of tag t^s in the entire data collection.

The second term measures the stability of tags, written as

$$S(\mathcal{Q}_s, t^s) = \frac{|\mathcal{Q}_t|}{|\mathcal{Q}_s|}, \tag{22}$$

where $\mathcal{Q}_t \subseteq \mathcal{Q}_s$, is defined as $\{q_t^s | q_t^s \in \mathcal{Q}_s \ \& \ t^s \in TagSet(q_t^s)\}$. The set $TagSet(q_t^s)$ means the associated tags of question q_t^s. Here, specific tags with lower collection frequencies are treated as less stable. This equation can be intuitively interpreted as follows: question space $\mathcal{Q}_s$ and its questions can be, respectively, viewed as a family and family members. Then the popularity of tag t^s in the family is estimated by averaging the voting from all family members. Practically, if different community participants annotate more distinct questions from the same semantically similar space using the same tags, these tags are more likely to reflect the objective aspects of the semantic content, and they are more reliable than tags with much lower collection frequencies. Through the algorithm, unambiguous and objective tags that receive the most neighbor voting will stand out.

The last term in Eq. (20) analyzes the tag relevance from the perspective of average neighbour distances, stated as

$$C(q, t^s) = \frac{\sum_{q_t^s \in \mathcal{Q}_t} f(q_t^s)}{|\mathcal{Q}_t|}, \tag{23}$$

Table II. Meta Information of Our Data Collection

User Num	Question Num	Answer Num	Tag Num	Distinct Tag Num
105.57 K	218.35 K	900.40 K	541.51 K	32.05 K

where $f(q_t^s)$ is obtained based on the proposed adaptive probabilistic hypergraph learning approach. Compared to the hard voting depicted by the second term, it is a kind of soft voting.

5.2. Complexity Analysis

The computational complexity of our scheme mainly comes from three parts: (1) feature extraction (including both of questions and answers); (2) adaptive probabilistic hypergraph learning; and (3) the heuristic approach for tag selection. Undoubtedly, feature extraction is the most computationally expensive step but can be handled offline. Actually, the complexity of the relevant tag selection can be ignored due to the smaller size of tag candidates inferred by our first component. For the proposed hypergraph learning, the computational cost magnitude is analyzed as

$$O\left(t(E^3 + 2VE^2 + 2EV^2 + V^3) + dV^2\right), \tag{24}$$

where t is the time of iterations and is usually below 10 in our work. d stands for the feature dimension. The sizes of considered vertices and hyperedges are, respectively, denoted as V and E—both in the order of thousands if we only select the top one thousand questions based on the initial relevance probabilities. Thus the computational cost is very low.

6. EXPERIMENTS

6.1. First-Order Analytics on Our Dataset

To evaluate our methods, a large real-world dataset was crawled from Zhihu, which officially announced that it had approximately 300,000 users as of March 2012.[4] Our dataset was collected in July 2012, comprising of more than 105,000 connected users and all their associated data, including asked questions, replied answers, the social connections among users, and the historical edit log. It accounts for a large fraction of the whole website and hence is comparatively representative for statistical analytics.

Table II displays the meta information of our data collection. Based on this table, we can easily calculate that the number of tags per question and the number of unique tag occurrences on average is 2.48 and 16.9, respectively. These two values explicitly reveal the tag incompleteness problem and the high rate of tag-repeating utilization, correspondingly. The reuse of tags demonstrates the rationality of selecting appropriate tags from the inferred tag vocabulary. Meanwhile, the repeating times determine the size of tag-based hyperedges, that is, around 16.9 questions on average are grouped by one tag-based hyperedge. Also, this finding suggests that we only need consider the 17-nearest neighbours on average when constructing the question-content-based hyperedges.

Figure 6 shows the distribution of the number of users with respect to followees and followers. Both them comply with power law distributions except the two bottom-left points. These two points refer to hundreds of users either having no followers or having no followees. Besides, it is derived that the average followers per user, 44, is relatively larger than the average followees per user, 28. This is why we chose information from

[4] `http://tech.sina.com.cn/i/2012-03-16/16476844824.shtml`

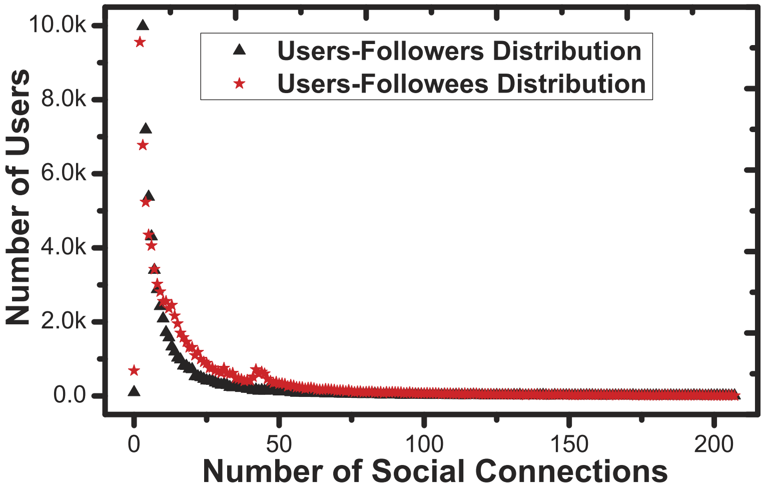


Fig. 6. The distribution of the number of users with respect to the number of social connections. “k” refers to thousand.

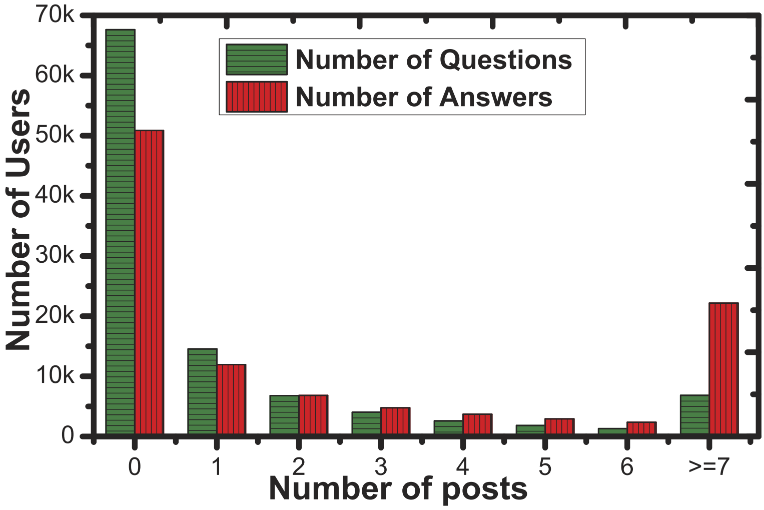


Fig. 7. The distribution of the number of users with respect to the number of posts. The leftmost bars mean more than 65,000 users and around 50,000 users never asked a question and never answered a question, respectively. These bars clearly reflect that most of the QA users are browsing the contents rather than actively contributing something.

followees to construct the user-based hyperedges, that is, keeping the simplicity of our hypergraph. Also, Figure 7 shows the distribution of the number of users over the categorical posts, including questions and answers. This figure provides conclusive evidence that more than half of the users are not active, never asking or answering questions. Instead, they may browse, tag, or refine others' post. From the angle of statistics, community participants seem to prefer answering (8.53 answers per user and 4.12 answers per question) to asking (2.07 questions per user). Jointly analyzing these basic statistical data, another important inference is the average size of user-based hyperedges: 60 questions on average are gathered together by each user-based hyperedge.

To represent the content of each question q, we first merged all its answers together with the question itself. This aims to address the information insufficiency problem of short questions. Based on our dataset, 218,350 QA pairs were formed. We then filtered out the QA pairs without any tag, resulting in 178,870 QA pairs left. Following that, the openNLP tool was utilized to segment all 178,870 QA pairs [Chang et al. 2009], which output more than 200,000 chunks. After removing stop words and filtering the chunks with frequencies smaller than 5 in the 178,870 QA pairs, we obtained a 29,802-dimensional bag-of-chunks histogram. Therefore, for each QA pair, it is represented by a 29,802-dimensional feature vector indicating its embedded chunk distribution. The feature vector is employed to calculate the pairwise question similarities via Euclidian distances:

$$d(q_i, q_j) = \sqrt{\sum_{k=1}^{dim} (d_{ik} - d_{jk})^2}. \tag{25}$$

The median of the Euclidean distances among all QA pairs was feed into the hypergraph model in Eq. (6). Meanwhile, we randomly selected 50 questions as testing data.

6.2. Learning Performance Comparison

To evaluate the ranking-based question-space inference, we adopted NDCG@n as our metric,

$$NDCG@n = \frac{rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}}{IDCG}, \tag{26}$$

where rel_i is the relevance score of the ith question in the ranked list, and $IDCG$ is the normalizing factor that makes NDCG@n being 1 for a perfect ranking.

To demonstrate the effectiveness of our proposed approach, we comparatively evaluated the following methods.

- —*PRF: Pseudo-Relevance Feedback* [Yan et al. 2003]. A support vector machine (SVM) classifier was trained to perform the reranking based on the assumption that the top-ranked questions are more relevant than the low-ranked results in general. The initial question ranking list was generated based on Eq. (6). (Baseline 1).
- —*RW: Random Walk-Based Reranking* [Hsu et al. 2007]. This is a typical simple graph-based reranking method jointly exploiting both initial relevance probabilities and semantic similarity between questions. The stationary probability of random walk was used to compute the final relevance scores. The initial relevance probabilities of each question was estimated based on Eq. (6). (Baseline 2).
- —*CHL: Conventional Hypergraph Learning* [Huang et al. 2010]. The weights of different hyperedges were not dynamically learned, which are fixed according to initial estimation, as described in Section 4.1. (Baseline 3).
- —*APHL: Adaptive Probabilistic Hypergraph Learning*. This is our proposed approach with alternative optimization between **W** and **f**.

For each method mentioned, the involved parameters were carefully tuned with a hold-out set, and the parameters with the best performance were used to report the final comparison results. Meanwhile, the ground truth was created by a manual labeling procedure through a pooling method. Specifically, each testing question has a pool that was constructed by merging the top-50 semantically similar questions recommended based on each strategy. Then five human annotators were invited to label all the questions pool by pool, including two Masters students and one Ph.D. student in computer science, one Masters student in information system, as well as one software engineer.

Table III. Illustrative Explanation of “Very Relevant”, “Relevant”, and “Irrelevant” Questions to the Given Question

Questions	Related Topics	Relevant Level
Which method of Perl provides the same function as Java substring() ?	Perl method, Java substring(), Equivalent function	Very relevant
Does someone have examples of using the Java substring() method ?	Examples, Java substring()	Relevant
Where can I download the perl debugger ?	Download location, Perl debugger	Irrelevant

Note: The given question is “What Perl method is equivalent to the Java substring() method”, which is talking about the topics “Perl method”, “Java substring()”, and “equivalent function”.

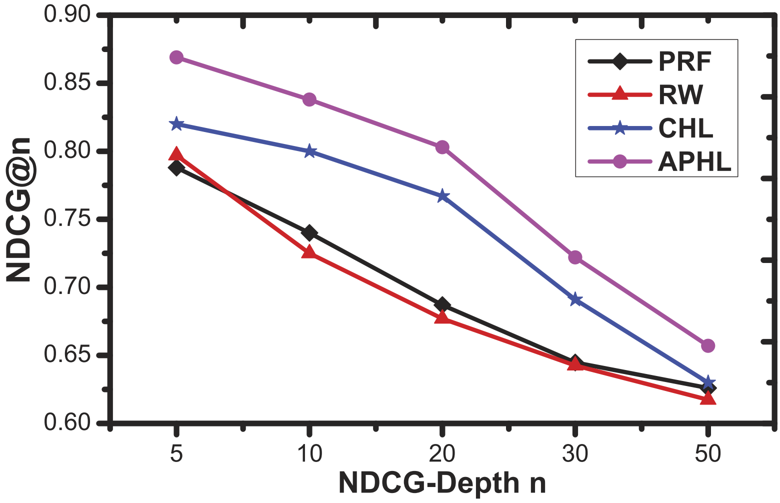


Fig. 8. Performance comparison among different reranking-based question-space inference approaches in terms of NDCG at different depths.

The labelers were trained with a short tutorial and a set of typical examples, as shown in Table III. Each question was assigned to be very relevant (score 2), relevant (score 1), or irrelevant (score 0) with respect to the given question. One question usually relates to multiple topics. If two questions are identical or semantically similar (i.e., they are related to the same topics), then they are regarded as very relevant. If two questions share at least one topic but not all, they are treated as relevant. And if two questions do not share any topic, they are viewed as irrelevant. We performed a voting method to establish the final relevance level of each question. For the cases in which there were two classes having the same number of ballots, a discussion was carried out among the labelers to decide the final ground truths.

We need to admit that the ground truth labeling is subjective, but a majority voting among the five labelers can partially alleviate the problem. We have also analyzed the interrater reliability of the labeling tasks with the Kappa method [Warrens 2010], and the Kappa value is much greater than 0.7, which indicates an adequate interrater agreement.

Figure 8 illustrates the experimental results. From this figure, our observation confirms that the proposed approach consistently and substantially outperforms other current publicly disclosed state-of-the-art reranking algorithms across various depths of NDCG. Among these four methods, the two hypergraph-based learning approaches

Table IV. Illustration of Representative Questions with Recommended Tags and Their Labeling Results

Question Samples	Recommended Tags Judged as Positive by Assessors	Recommended Tags Judged as NOT Positive by Assessors
What was the economic impact of the 2009 H1N1 Swine Flu epidemic ?	H1N1, Economics, Epidemiology, Impact, H1N1 outbreak,	Airport Security, Vaccination, Doctors, International Travel
How many active users does Flickr have ?	Flickr, Flickr Users, Active Users, Statistics	Picasa, Instagram, Startup and Companies, Mobile Photo Opportunity
Why is it normally hard to monetize Web applications ?	Web Apps, Monetization, Web-based Startup, Business Models, Revenues	Digital Advertising, E-Commerce, Retail, Social Media, Web 2.0

Note: Here we do not display the verbose answers due to the limited space.

show superiority over the other two approaches. One possible reason is the unreliable initial ranking list resulting from rough estimation. The other main reason is that hypergraph-based learning is able to capture the high-order relationships among questions, that is, the summarized local grouping information, in contrast to simple pairwise relationships characterized by the other two approaches. From this figure, we can also observe that our proposed method performs stably better than the conventional hypergraph learning approach. This verifies that it is better to simultaneously learn the question relevance score and hyperedge weights.

6.3. Relevant Tag Selection

It is well known that for the annotation task, precision is usually more important than recall. Therefore, we adopted two metrics that are able to capture precision from different aspects. The first is average *S@K* over all testing questions, which measures the probability of finding a relevant tag among the top-*K* recommended tags. To be specific, for each testing question, *S@K* is assigned one if a relevant tag was ranked in the top-*K* positions and zero, otherwise. The second metric is average *P@K* that stands for the proportion of recommended tags that is relevant. *P@K* is defined as

$$P@K = \frac{|C \cap R|}{|C|}, \tag{27}$$

where C is a set of the top-K tags and R is the manually labeled positive ones. For the ground truth construction, analogous to Section 6.2, the pooling method was employed by simply asking five volunteers to judge a suggested tag as positive or not. The five labelers were trained with the following instructions: if one tag is able to capture one aspect or topic of the given QA pair, it is annotated as “positive”; otherwise, it is annotated as “not positive”. A suggested tag was ultimately assumed as positive only if three out of five labelers marked it as positive. Some tagging examples are listed[5] in Table IV. The interrater agreement was also measured by the Kappa method, which achieved sufficient interrater agreement with Kappa values bigger than 0.7. We did not observe any testing question for which all the tags were marked as “not positive” by the assessors.

We compared our question annotation approach with other state-of-the-art methods in terms of various metrics.

—*FTRCK: Flickr Tag Recommendation Based on Collective Knowledge* [Sigurbjörnsson and van Zwol 2008]. Basically, this approach is a statistical

[5] To improve the readability, we translate the Chinese QA pairs with tags into English.

Table V. Comparative Evaluation Results of Relevant Tag-Selection Approaches in Terms of *S@K*

Metrics Apporaches	S@1	S@2	S@3	S@4	S@5
Ours	**70.0%**	**88.0%**	**94.0%**	**96.0%**	**98.0%**
FTRCK	64.0%	82.0%	86.0%	90.0%	92.0%
TagAssist	66.0%	80.0%	88.0%	92.0%	94.0%

Table VI. Comparative Evaluation Results of Relevant Tag-Selection Approaches in Terms of *P@K*

Metrics Apporaches	P@1	P@2	P@3	P@4	P@5
Ours	**70.0%**	**65.0%**	**62.7%**	**59.0%**	**58.0%**
FTRCK	64.0%	60.0%	56.7%	54.0%	52.0%
TagAssist	66.0%	63.0%	59.3%	56.0%	53.2%

and data-driven method. To be specific, given a question with user-defined tags, an ordered list of m candidate tags is derived for each of the user-defined tags based on the tag co-occurrence. The lists of candidate tags are then used as input for tag aggregation and ranking, which ultimately produces the ranked list of n recommended tags. For the aggregation, we employed the vote-based strategy as introduced in Sigurbjörnsson and van Zwol [2008].

—*TagAssist*. This method was first introduced by Sood et al. [2007], which provides tag suggestions for new blog posts by utilizing existing tagged posts. It annotates a post by generating search queries from the given post content, searching a collection of blog posts using those queries, and selecting suitable tags from the retrieved posts. For question match, we employed the method in Wang et al. [2009].

Tables V and VI, respectively, present the precision of relevant tag selection based on *P@K* and *S@K*. It is observed that the performance in terms of *S*@1 is as high as 70%, which means that for up to 70% questions, our proposed annotation scheme can suggest a relevant tag at rank one. Moreover, the value of *S*@5 almost ensures that at least one tag is relevant among the top-five recommended tags. Besides, *P*@5 achieves 58% accuracy, which reflects that about three out of the top five tags on average are able to characterize the question topics well. From the view of efficacy, the performances of *S@K* and *P@K* unquestioningly confirm the high applicability of our proposed method in tag suggestion.

As displayed in Tables V and VI, our proposed approach consistently outperforms other state-of-the-art methods. FTRCK only statistically considers the co-occurrences among tags and completely ignores the pairwise similarities between QA pairs estimated based on lexical properties and also overlooks the social relationships among users. Even though TagAssist takes pairwise similarities into consideration, it does not consider the tag co-occurrence information and the social connections among users. These oversights cause their poor performances compared to our proposed method. Meanwhile, we can see that the performance of TagAssist is a bit better than FTRCK, which reflects that the lexical-properties-based similarity is more reliable than tag co-occurrence cues. However, it is worth highlighting that FTRCK is applicable to online services because of its easy implementation and faster speed.

We also performed T-test analysis. The results are displayed in Table VII. From this table, we can see that all the p-values are smaller than 0.05, which verifies that our proposed method significantly outperforms other two state-of-the-art algorithms.

Table VII. The p-Value of T-Test Between Different Methods

Metrics / Apporaches	T-test over S@K	T-test over P@K
Our Method VS. FTRCK	4.461E-05	1.084E-05
Our Method VS. TagAssist	0.00145	0.000932

Note: We can see that all the p-values$<$ 0.05, and they indicate that the difference is significant.

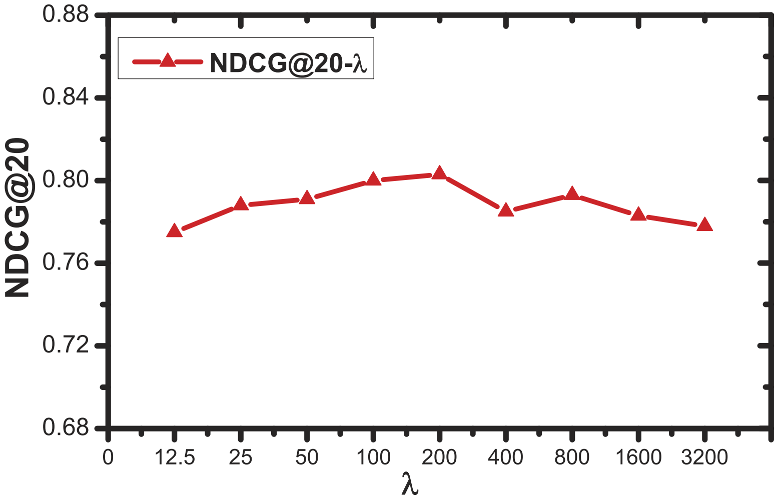


Fig. 9. The performance of question-space inference with various λ when μ is fixed as 0.001.

6.4. Sensitivity of Parameters

As shown in Eq. (12) and discussed in Section 4.2, the two positive parameters λ and μ play important roles in modulating the effects of empirical loss and weighting regularizer, respectively. The former is widely tuned in hypergraph learning algorithms. While for the latter, with variation from zero to infinity, the hyperedge weights will accordingly vary from an extremely balanced case to an extremely imbalanced case [Yu et al. 2012]. Specifically, when $\mu = \infty$, the proposed adaptive hypergraph will be reduced to conventional hypergraph, since the optimal solution will assign identical weights for all hyperedges. On the contrary, if μ tends to zero, then the optimal results will be that only one weight is one and all others are zero.

In this section, we conducted a series of experiments to investigate the sensitivity of these two parameters. We first performed grid search with flexible step size to seek λ and μ with optimal reranking performance in terms of NDCG@20. 200 and 0.001 were, respectively, located for λ and μ, which are also utilized to report the comparison results in Figure 8. The NDCG@20-λ curve is presented in Figure 9 with μ fixed as 0.001. As illustrated, the performance gradually increases with λ growing and arriving at a peak at a certain value, then the performance goes downward and finally becomes relatively stable. Similarly, Figure 10 shows the NDCG@20-μ curve with λ fixed as 200, where the performance varies according to different μ. With the increase of μ, more informative hyperedges are taken into consideration via updating the weights from zero to nonzero. It is also observed that when μ reaches a certain value, the performance starts to decrease. This is because more "incorrect" hyperedges are potentially introduced. However, based on the observations, we conclude that the performance of

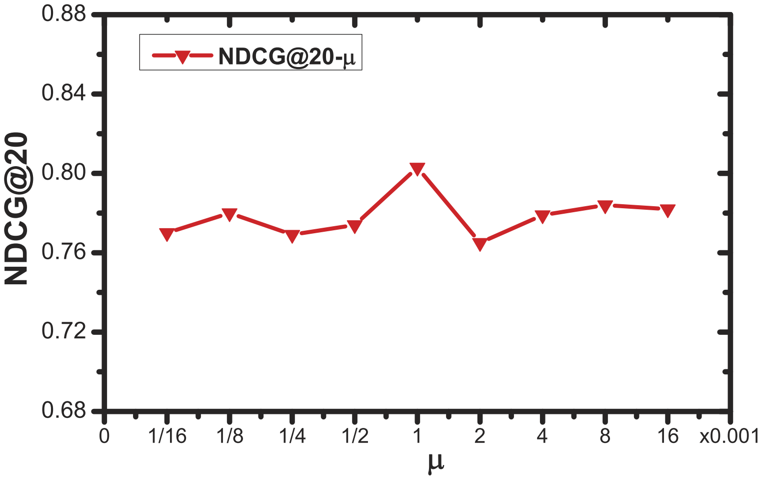


Fig. 10. The performance of question-space inference with various μ when λ is fixed as 200.

our proposed method changes between $(0.765, 0.803)$ when the parameters vary in a wide range, which is not very sensitive.

6.5. Discussions

Social QA originates from the traditional community QA (cQA) but is beyond cQA. To be specific, social QA brings in lots of novel social-oriented features, such as question tags, social relationships, and crowdsourcing. Our work studies the question annotation for social QA services, and our proposed model seamlessly integrates multifacet heterogeneous information cues, including both tag-sharing network and social connections among users. The current dominant cQA forums, however, do not support these features well. Even though we can view the categories of some cQA forums as tags, for example, Yahoo! Answer, it still suffers from limited tag vocabulary and problems of one tag per question.[6] Consequently, our model cannot be validated on cQA benchmarks. Currently, only two well-known social QA websites are launched: Quora and Zhihu. The former is well presented in English; it, however, strictly prohibits crawlers. On the other hand, Zhihu data is easy to collect and sufficient to support our evaluation.

Another issue which deserves further discussion here, is the biased tag problem. The results in Table IV reveal that approximately two tags on average in the top-five recommended tags may skew the original question topics. This problem is hard to tackle because our data distributes in general and broader domains, so we thus cannot ensure there exists sufficient real “neighbors” of each testing sample. The basic principle of our approach is to globally select tags from similar question space, so it is unable to obtain appropriate tags for the isolated questions. However, our approach can achieve greater performance on vertical and specific domains, such as the healthcare domain. Irrelevant tags may result in poor performance of tag-based operations, such as query expansion, statistics, as well as content organization, etc. This, however, can be completely avoided if we incorporate the interactive recommendation. To be specific, for each new coming question, our system automatically recommends tags to users,

[6] It is well known that Yahoo! Answers has only 1,263 leaf-level nodes distributed in 26 top-level categories. This category vocabulary is extremely limited.

including appropriate and inappropriate ones, and then the users can manually select the tags that are relevant to the intents/topics of the given question.

7. CONCLUSIONS AND FUTURE WORK

This article studies user tagging behaviors with a representative real-world dataset and presents a novel scheme to automatically annotate social questions which unravels the incomplete and biased problems of question tags. For a given question, the scheme first constructs an adaptive probabilistic hypergraph to infer the semantically similar question space. Based on this question space, a collection of probably relevant tags are roughly identified. Comprehensive information cues from users, questions, and tags are seamlessly integrated into this hypergraph. This step narrows down the suggested tag candidates. Our scheme then performs a heuristic approach to further filter the tag candidates by simultaneously damping generic tags, penalizing specific tags, as well as rewarding tags from closer neighbours. It greatly strengthens annotation by keeping off subjective, ambiguous, and generic tags. The experimental results have demonstrated that our scheme achieves promising performance for question annotation.

Our proposed scheme will benefit several tag-based operations, such as the knowledge organizer. The knowledge structures of conventional cQA forums are predefined, which suffer from issues associated with fixed taxonomies, such as being centralized, conservative, and ambiguous [Huang et al. 2010]. By leveraging question tags, the ontology of QA pairs can be flexibly and effortlessly reorganized via mapping associated tags into user needs-aware categories. The new generated knowledge hierarchy is user-navigable and reconfigurable, which greatly empowers users' Web surfing experiences.

However, the current approach overlooks informative terms extracted from QA pairs. Thus in-depth research remains to be investigated.

REFERENCES

Sameer Agarwal, Kristin Branson, and Serge Belongie. 2006. Higher order learning with graphs. In *Proceedings of the International Conference on Machine Learning*.

Morgan Ames and Mor Naaman. 2007. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Christopher H. Brooks and Nancy Montanez. 2006. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the International Conference on World Wide Web*.

Gustavo Carneiro and Nuno Vasconcelos. 2005. Formulating semantic image annotation as a supervised learning problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A real-world Web image database from National University of Singapore. In *Proceeding of the ACM International Conference on Image and Video Retrieval*.

Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*.

Zhouyu Fu, Guojun Lu, Kai ming Ting, and Dengsheng Zhang. 2011. A survey of audio-based music classification and annotation. *IEEE Trans. Multimedia 13*, 2, 303–319.

Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2012. Visual-textual joint relevance learning for tag-based social image search. *IEEE Trans. Image Process. 22*, 1, 363–376.

Scott A. Golder and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *J. Inf. Sci. 32*, 2, 198–208.

Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. 2007. Video search reranking through random walk over document-level context graph. In *Proceedings of the ACM International Conference on Multimedia*.

Yuchi Huang, Qingshan Liu, Shaoting Zhang, and Dimitris N. Metaxas. 2010. Image retrieval via probabilistic hypergraph ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jiwoon Jeon, Victor Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the International ACM SIGIR Conference*.

Feng Kang, Rong Jin, and Rahul Sukthankar. 2006. Correlated label propagation with application to multi-label learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xirong Li, Cees G. M. Snoek, and Marcel Worring. 2009. Annotating images by harnessing worldwide user-tagged photos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. In *Proceedings of the International Conference on World Wide Web*.

Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. 2010. Image retagging. In *Proceedings of the ACM International Conference on Multimedia*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Develop. 2*, 2, 159–165.

Gilad Mishne. 2006. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the International Conference on World Wide Web*.

Florent Monay and Daniel Gatica-Perez. 2004. Plsa-based image auto-annotation: Constraining the latent space. In *Proceedings of the ACM International Conference on Multimedia*.

Yasuhide Hironobu Mori, Hironobu Takahashi, and Ryuichi Oka. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*.

Sascha Narr, Ernesto William De Luca, and Sahin Albayrak. 2011. Extracting semantic annotations from Twitter. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval*.

Liqiang Nie, Meng Wang, Zheng-Jun Zha, Guangda Li, and Tat-Seng Chua. 2011. Multimedia answering: Enriching text QA with media information. In *Proceedings of the International ACM SIGIR Conference*.

Liqiang Nie, Meng Wang, Zheng-Jun Zha, and Tat-Seng Chua. 2012a. Oracle in image search: A content-based approach to performance prediction. *ACM Trans. Inf. Syst. 30*, 2, Article 3.

Liqiang Nie, Shuicheng Yan, Meng Wang, Richang Hong, and Tat-Seng Chua. 2012b. Harvesting visual concepts for image search with complex queries. In *Proceedings of the ACM International Conference on Multimedia*.

Liqiang Nie, Meng Wang, Yue Gao, Zheng-Jun Zha, and Tat-Seng Chua. 2013. Beyond text QA: Multimedia answer generation by harvesting Web information. *IEEE Trans. Multimedia 15*, 2, 426–441.

Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. 2007. Correlative multi-label video annotation. In *Proceedings of the ACM International Conference on Multimedia*.

Börkur Sigurbjörnsson and Roelof van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the International Conference on World Wide Web*.

Sanjay Sood, Sara Owsley, Kristian Hammond, and Larry Birnbaum. 2007. Tagassist: Automatic tag suggestion for blog posts. In *Proceedings of the International Conference on Weblogs and Social Media*.

Shankara B. Subramanya and Huan Liu. 2008. Socialtagger - Collaborative tagging for blogs in the long tail. In *Proceedings of the ACM Workshop on Search in Social Media*.

Jinhui Tang, Haojie Li, Guo-Jun Qi, and Tat-Seng Chua. 2010. Image annotation by graph-based inference with integrated multiple/single instance representations. *IEEE Trans. Multimedia 12*, 2, 131–141.

Xinmei Tian, Linjun Yang, Jingdong Wang, Yichen Yang, Xiuqing Wu, and Xian-Sheng Hua. 2008. Bayesian video search reranking. In *Proceedings of the ACM International Conference on Multimedia*.

Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based QA services. In *Proceedings of the International ACM SIGIR Conference*.

Matthijs J. Warrens. 2010. Inequalities between multi-rater kappas. *Adv. Data Anal. Classification 4,* 4, 271–286.

Pengcheng Wu, Steven Chu-Hong Hoi, Peilin Zhao, and Ying He. 2011. Mining social images with distance metric learning for automated image tagging. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

Wei Wu, Bin Zhang, and Mari Ostendorf. 2010. Automatic generation of personalized annotation tags for twitter users. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Yu Xiang, Xiangdong Zhou, Tat-Seng Chua, and Chong-Wah Ngo. 2009. A revisit of generative model for automatic image annotation using markov random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. 2006. Towards the Semantic Web: Collaborative tag suggestions. In *Proceedings of the International Conference on World Wide Web*.

Rong Yan, Alexander Hauptmann, and Rong Jin. 2003. Multimedia search with pseudo-relevance feedback. In *Proceedings of the International Conference on Image and Video*.

Changbo Yang, Ming Dong, and Jing Hua. 2006. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yang Yang, Yi Yang, and Heng Tao Shen. 2013. Effective transfer tagging from image to video. *ACM Trans. Multimedia Comput. Commun. Appl. 9*, 2, Article 14.

Jun Yu, Dacheng Tao, and Meng Wang. 2012. Adaptive hypergraph learning and its application in image classification. *IEEE Trans. Image Process. 21*, 7, 3262–3272.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *Proceedings of the Advances in Neural Information Processing Systems Conference*.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. 2006. Learning with hypergraphs: Clustering, classification, and embedding. In *Proceedings of the Advances in Neural Information Processing Systems Conference*.

Received February 2013; revised July 2013, September 2013; accepted November 2013