Unsupervised Tag Sense Disambiguation in Folksonomies

Kaipeng Liu, Binxing Fang, Weizhe Zhang
Research Center of Computer Network and Information Security Technology,
Harbin Institute of Technology, Harbin 150001, China
Email: {liukaipeng, bxfang, zwz}@pact518.hit.edu.cn

***Abstract*— Disambiguating tag senses can benefit many applications leveraging folksonomies as knowledge sources. In this paper, we propose an unsupervised tag sense disambiguation approach. For a target tag, we model all the annotations involving it with a 3-order tensor to fully explore the multi-type interrelated data. We perform spectral clustering over the hypergraph induced from the 3-order tensor to discover the clusters representing the senses of the target tag. We conduct experiments on a dataset collected from a real-world system. Both the supervised and unsupervised evaluation results demonstrate the effectiveness of the proposed approach.**

***Index Terms*— social tagging, sense disambiguation, spectral clustering**

I. Introduction

The emergence of social tagging systems, which allow collaborative users to submit shared resources and to annotate them with descriptive tags, forming the so-called *folksonomies*. As shown in Fig. 1, a folksonomy can be seen as a structure $\mathbb{F} := (U, R, T, Y)$ consisting of i) a set U of *users*, ii) a set R of *resources*, iii) a set T of *tags*, and iv) the ternary relation between them, i.e. $Y \subseteq U \times R \times T$, called *annotations*.

The success of tagging mainly relies on the easy-to-use user interface that allows users to annotate the resources with free-chosen keywords, which does not require the users to be familiar with a predefined vocabulary. Though such a user interface allows the users to achieve immediate benefits from the system without too much overhead, it brings the problem of being lack of semantics in folksonomies. For example, the ambiguous usage of tags may have a great impact on the performance of tag-based information retrieval. To overcome this problem, in this paper, we study the problem of tag sense disambiguation in folksonomies. This study can benefit many applications that leverages folksonomies as data and knowledge sources, such as

Navigation interfaces The navigation interfaces of social tagging systems, such as tag clouds and most popular tag lists, usually leverages the tags for quick accessing to a specific category of resources. However, the different senses of tags are not considered in these interfaces, which leads to an unsatisfied user experience. Disambiguating tag senses can bring significant improvement for these navigation interfaces.

Information retrieval Retrieving resources based on social annotations [1]–[3] heavily relies on the ability of distinguishing different senses of tags. For instance, if we can index the different senses of tags, then queries having an explicit sense intent can be better served.

Ontology learning Learning ontologies from folksonomies becomes an active research topic in recent years [4]–[6]. One of the most import issue in ontology learning is that, the different senses of tags should be disambiguated to better distinguish the different concepts representing by the same tag.

The problem of tag sense disambiguation (TSD) is somewhat analogous to the well-known problem of word sense disambiguation (WSD). WSD has been considered as an AI-complete problem [7], which means that the difficulty of WSD is equivalent to solving central problems of artificial intelligence (AI), e.g. the Turing Test [8]. The acknowledged difficulty of WSD does not arise from a single cause, but rather from a variety of factors, such as the difficulty of choosing the representation of a word sense (ranging from the enumeration of a finite set of senses to rule-based generation of new senses) and the granularity of sense inventories (from subtle distinctions to homonyms). Moreover, WSD heavily relies on external knowledge. In fact, the skeletal procedure of any WSD algorithm can be summarized as: given a set of words,

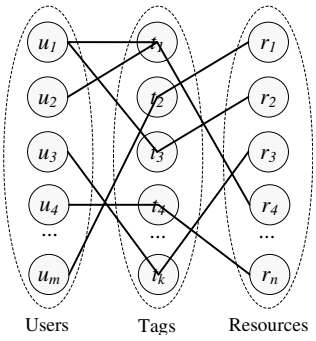

Figure 1. Illustration of the tripartite hypergraph structure of a folksonomy.

This work was supported in part by the National Natural Science Foundation of China under Grant No.60703014 and No.60933005, the National Grand Fundamental Research 973 Program of China under Grant No.G2007CB311100 and the National High-Tech Research and Development Plan of China under Grant No.2006AA010105-02, No.2007AA01Z416, No.2007AA01Z442 and No.2009AA01Z437.

doi:10.4304/jcp.5.11.1715-1722

a technique, which leverages one or more knowledge sources, is employed to assign the most appropriate senses to words in context. Knowledge sources can vary considerably from corpora of texts to more structured resources, such as machine-readable dictionaries. Without knowledge, it would be impossible for both humans and machines to accomplish the task of WSD. TSD share the same characteristics described above with WSD, except for the differences stated below:

Selection of Senses Unlike WSD, the targets of TSD is free form tags, which may not be included in a lexicon. Thus, the selection of senses cannot directly rely on such external knowledge.

Representation of context The context in WSD is often constructed with word around the target word in the text, usually in a context window or in the same paragraph. In TSD, there is no such context. Thus we have to develop new methods to model the context of the target tag.

Knowledge sources There are no public available knowledge sources focusing on the vocabulary of tags. Thus, training and evaluating TSD algorithms are hard to perform.

Due to i) the applications mentioned above usually do not require us to map the induced senses of the target tag to a predefined set of senses and ii) the problem of lack of knowledge source mentioned above, we adopt an unsupervised method to *discriminate* tag senses in this study. Specifically, we make the following contributions in this paper,

- We model the contexts of a target tag with a 3-order tensor to to fully explore the multi-type interrelated data associated with the target tag.
- We perform spectral clustering over the tripartite hypergraph induced from the 3-order tensor to discover the clusters representing the senses of the target tag.
- We conduct experiment on a dataset collected from a real-world social tagging system and the experimental result demonstrate the effectiveness of the proposed method.

The remainder of this paper is organized as follows: Section II reviews the related work; Section III introduces our algorithm for tag sense disambiguation; Section IV presents the experimental results; Section V concludes our work.

II. Related Work

In this section, we review some research efforts closely related to our study, including unsupervised WSD, TSD in folksonomies and the spectral clustering technique.

A. Unsupervised WSD

Unlike supervised WSD, in which senses for a target word are selected from a closed list based on a dictionary or lexicon, unsupervised WSD tries to induce word senses directly from the training corpus. The main approaches for unsupervised WSD includes methods based on context clustering, word clustering and cooccurrence graphs. The methods based on context clustering employ the vector space model of a target word. The vectors representing the occurrences of a target word are clustered into groups, each identifying a sense of the target word. Schütze [9] proposed an context-group discrimination algorithm, which grouped the occurrences of an ambiguous word into clusters of senses based on the contextual similarity between word occurrences. For their algorithm, the contextual similarity was calculated with the cosine between the corresponding vectors; the clustering was performed with the Expectation Maximization algorithm, which was an iterative maximum likelihood estimation procedure of a probabilistic model [10]. Pedersen and Bruce [11] proposed a different clustering approach employing the agglomerative clustering technique. Initially, each word occurrence formed a singleton cluster. Then, the agglomerative clustering method merged the most similar pair of clusters. The procedure continued with successively less similar pairs until a stopping threshold was reached.

On the other hand, the methods based on word clustering identifying words that are similar to the target word and use the clusters of words to convey a specific sense. Lin [12] proposed a word clustering approach to identify the words $\mathbf{w} = (w1, \ldots, wk)$ similar to a target word w_0. The similarity between w_0 and w_i was determined based on the information content of their single features, which were given by the syntactic dependencies, such as subject-verb, verb-object, adjective-noun, etc., that occurred in a given corpus. The more dependencies the two words shared, the higher the information content. As for context vectors, however, the words in $\mathbf{w}$ would cover all the senses of w_0. Thus, they applied a word clustering algorithm to discriminate between the senses. Let $\mathbf{w}$ be the list of similar words ordered by similarity to w_0. They first created a similarity tree T which contains a single node w_0. Then, for each $i \in \{1, \ldots, k\}$, they added $w_i \in \mathbf{w}$ as a child of w_j in the tree T such that w_j was the most similar word to w_i in $\{w_0, \ldots, w_{i1}$. After a pruning step, each subtree rooted at w_0 was considered as a distinct sense of w_0. Lin and Pantel also proposed in [13] a subsequent approach, called the clustering by committee (CBC) algorithm, which employed a different word clustering method.

The methods based on cooccurrence graphs provide a different view of WSD. These approaches are based on the notion of a cooccurrence graph $G = (V, E)$ whose vertices V correspond to words and edges E connect the words having certain syntactic relations such as in the same paragraph or in a larger context. Véronis proposed an approach called HyperLex [14]. First, a cooccurrence graph was built such that nodes were words occurring in the paragraphs of a text corpus in which a target word occurred, and an edge between a pair of words was added to the graph if they co-occur in the same paragraph. Each edge was assigned a weight according to the relative cooccurrence frequency of the two words connected by

the edge, by which words with high frequency of cooccurrence were assigned a weight close to zero, whereas words which rarely occurred together received weights close to 1. Edges with a weight above a certain threshold were discarded. In the second step, an iterative algorithm was applied to the cooccurrence graph. At each iteration, the node with the highest relative degree in the graph was selected as a hub. All its neighbors were no longer considered as hub candidates. The algorithm stopped when the relative frequency of the word corresponding to the selected hub was below a fixed threshold. The entire set of hubs selected in the above procedure was said to represent the senses of the target word. A similar approach based on PageRank was proposed by Agirre et al in [15].

B. TSD in Folksonomies

The problem of TSD have been studied by many researchers in recent years. Yeung et al. proposed an unsupervised method for tag sense disambiguation through the analysis of the tripartite structure of folksonomies [16]. Based on the GN algorithm [17], which was proposed for discovering community structures with networks, they divided the graph consisting of resources associated with the target tag into clusters, each represented one sense of the target tag. First, the tagging data that associated the target tag t was collected and a one-mode graph of resources was constructed with the tagging data. Then, the edge with the highest betweenness value within this graph was removed, which was followed by an update of the best division of the graph based on the calculation of the modularity of the current division. This procedure was repeated until no more edges left in the graph. Finally, the division with the highest value of modularity was obtained. The clusters identified by this division were selected as the senses of the target tag. The most frequently used tags in each cluster were chosen as the signature of the corresponding tag sense.

Knowledge-based approaches [18], [19] were also proposed for TSD. In [18], Lee et al. used Wikipedia as a reference to the tag vocabulary. They developed a method to map each occurrence of a target tag to a topic in Wikipedia. First, the local neighbor tags and global neighbor tags of the target tag were identified with the co-occurrence relations. These tags were used as the context of the occurrence of the target tag. Then, the topic relevance values between this context and all the Wikipedia topics were calculated to find the best mapping from the occurrence to the Wikipedia topic. Analogously, in [19], Garcia-Silva et al. used DBpedia, which was a compiled version of Wikipedia, to label tag senses. They used the similarities between the context of a tag occurrence and a tag sense represented by the bag-of-words model of a topic entry in DBpedia to select the best mapping from tag occurrences to tag senses.

C. Spectral Clustering

Spectral clustering has many fundamental advantages compared to the "traditional algorithms" such as k-means and single-linkage clustering and hence has attracted many research efforts. The most relevant works to our study are those focusing on multi-type relational data [20] and k-partite graphs [21], [22]. Long et al. proposed in [20] a method to cluster multi-type interrelated objects simultaneously. They proposed a general model, called the collective factorization on related matrices, for multi-type relational data clustering, which was applicable to relational data with various structures. Under this model, they developed the spectral relational clustering algorithm to cluster multi-type interrelated data objects simultaneously. The algorithm iteratively embedded each type of data objects into low dimensional spaces and benefited from the interactions among the hidden structures of different types of data objects. Zhou et al. applied spectral clustering to hypergraph for unsupervised learning [21], in which relationships among data objects were used to improve the cluster quality of interrelated data objects through an iterative reinforcement clustering process. The link structure derived from relationships of the interrelated data objects was used to differentiate the importance of objects and the learned importance was also used in the clustering process to further improve the clustering results. Chen and Saad developed a co-clustering algorithm for high order relational data using spectral hypergraph partitioning [22]. They generalized the methodology of spectral clustering which originally operated on undirected graphs to hypergraphs, and further developed algorithms for hypergraph embedding and transductive classification on the basis of the spectral hypergraph clustering approach. Refer to [23] for a comprehensive survey of spectral clustering methods.

III. Tag Sense Disambiguation

In this section, we develop a TSD algorithm based on the basic idea that although tags may be used for different meanings, one can still figure out what the particular sense is used for each occurrence based on its context. We first discuss the modeling of contexts in the following subsection, and then describe how to leverage the contexts to perform TSD in the next subsection.

A. Modeling Contexts

As mentioned in Section I, unlike WSD, modeling context for a target tag in TSD is not straight-forward. In WSD, words in the same paragraph or in a larger context window are usually considered in the same context. In TSD, however, there is no such syntactic relation. However, we can resort to leverage the multi-type interrelated data model inherent in folksonomies. In this study, we simultaneously model all the contexts of the target tag as a 3-order tensor (see Fig. 2).

A tensor is a multidimensional array of data whose elements are referred by using multiple indices, each of which represents a mode of the tensor. The number of indices required is called the order of the tensor. For a target tag t_0, all the annotations involving t_0 is

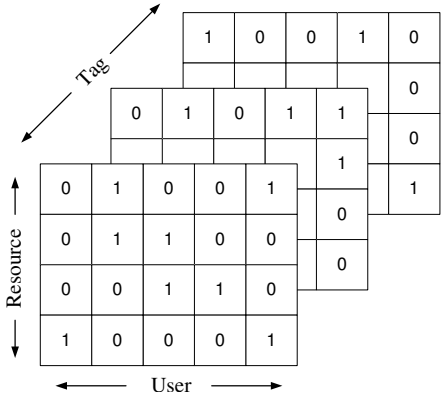

Figure 2. Illustration of the 3-order tensor representing all the contexts of a target tag.

collected as the corpus $\mathbb{F}_0 = (U, R, T, Y)$. A 3-order tensor $A \in \{0,1\}^{|U|\times|R|\times|T|}$ is constructed based on the corpus. The element in A is defined as $a_{u,r,t} = 1$ if and only if $(u, r, t) \in Y$. The 3-order tensor can be modeled as a hypergraph $G = (V, E, w)$ where $V = U \cup R \cup T$, $E \subseteq U \times R \times T$, and each hyperedge e contains exactly 3 vertices and is assigned with a weight of 1, i.e. $w(e) = w(u, r, t) = 1$. This hypergraph model of the data tensor shares similarities with the tripartite graph model [24], which also uses V as the vertex set, with U, R, and T as a partite set. However, the hypergraph contains hyperedges connecting exactly one vertex from each partite set, with each hyperedge having a single weight, while the tripartite graph contains edges connecting vertices from only two partite sets. In other words, the hypergraph models interrelationships among all the partite sets, while the tripartite graph models only pairwise relationships between the partite sets. Thus, one can view hypergraphs as natural and convenient models for data tensors. For the TSD setting, modeling the contexts as a 3-order tensor and inducing it to a hypergraph can provide us the ability to fully explore the multi-type interrelated data associated with the target tag.

B. Sense Discrimination via Spectral Clustering

Now we can completely represent the complex relationships among multi-type objects in folksonomies by using a tensor model, which can be induced to a hypergraph. However, a new problem arises. How to partition this hypergraph to discover the senses of the target tag? By extending the idea of co-clustering the rows and columns of a data matrix which is modeled as a bipartite graph [25], [26], we can simultaneously cluster all the modes of the 3-order tensor which is modeled as a hypergraph to discriminate the senses of a target tag. One of the best choices for this task is spectral clustering. Spectral clustering is a powerful technique for discovering cluster structures in multi-type interrelated data. In this study, we use spectral clustering to discover the clusters in the hypergraph modeling the contexts of the target tag.

Spectral clustering methods roots in the theory of spectral graph. The basic idea of spectral clustering is to construct a weighted graph with the application-specific data set, in which each node corresponds to an object and each weighted edge models the interrelation between two objects. In this framework, the clustering problem can be seen as a graph partitioning problem, which can be tackled with the aid of spectral graph theory. The core of spectral graph theory is the eigenvalue decomposition on the Laplacian matrix of the graph constructed from the relational data. In fact, a close relationship between the second smallest eigenvalue of the Laplacian and the graph cut can be identified [27], [28].

To perform spectral clustering, the Laplacian of the hypergraph $G = (V, E, w)$ induced from the 3-order tensor, which is used to model all the contexts involving a target tag, is defined as

$$L = D - \Psi, \tag{1}$$

where D is a $|V|\times|V|$ diagonal matrix and Φ is a $|V|\times|V|$ matrix. Due to the special structure of the hypergraph which is induced from a 3-order tensor, its Laplacian is also structured. By splitting D and Ψ into blocks with respect to the vertex subsets U, R and T, the Laplacian reads

$$\begin{aligned} L &= D - \Psi \\ &= \begin{bmatrix} D^U - \Psi^{UU} & -\Psi^{UR} & -\Psi^{UT} \\ -\Psi^{RU} & D^R - \Psi^{RR} & -\Psi^{RT} \\ -\Psi^{TU} & -\Psi^{TR} & D^T - \Psi^{TT} \end{bmatrix}, \end{aligned}$$

where the elements of D blocks are

$$\begin{aligned} D^U_u &= \sum_{r\in R, t\in T} w(u, r, t) = |Y_u|, \\ D^R_r &= \sum_{u\in U, t\in T} w(u, r, t) = |Y_r|, \\ D^T_t &= \sum_{u\in U, t\in T} w(u, r, t) = |Y_t|, \end{aligned}$$

and the elements of Ψ blocks are

$$\begin{aligned} \psi^{UU}_{u,u} &= \frac{1}{3}\sum_{r\in R, t\in T} w(u, r, t) = \frac{1}{3}|Y_u|, \\ \psi^{RR}_{r,r} &= \frac{1}{3}\sum_{u\in U, t\in T} w(u, r, t) = \frac{1}{3}|Y_r|, \\ \psi^{TT}_{t,t} &= \frac{1}{3}\sum_{u\in U, r\in R} w(u, r, t) = \frac{1}{3}|Y_t|, \\ \psi^{UR}_{u,r} = \Psi^{RU}_{r,u} &= \frac{1}{3}\sum_{t\in T} w(u, r, t) = \frac{1}{3}|Y_{u,r}|, \\ \psi^{UT}_{u,t} = \Psi^{TU}_{t,u} &= \frac{1}{3}\sum_{r\in R} w(u, r, t) = \frac{1}{3}|Y_{u,t}|, \\ \psi^{RT}_{r,t} = \Psi^{TR}_{t,r} &= \frac{1}{3}\sum_{u\in U} w(u, r, t) = \frac{1}{3}|Y_{r,t}|, \end{aligned}$$

where Y_i is the set of annotations associated with the object i and $Y_{i,j}$ is the set of annotations associated with both the object i and j.

Algorithm 1: The TSD algorithm

Input: The target tag t_0.
Input: The given number k of tag senses.
Construct a tensor A for t according to the ways described in Section III-A.
Compute the unnormalized Laplacian L for the hypergraph $G = (V, E)$ induced from A with Eq. (1).
Compute the first k eigenvectors $u_1, \ldots, u_k$ of L.
Let $U \in \mathbb{R}^{|V| \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ as columns.
For $i = 1, \ldots, |V|$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
Cluster the points y_i, $i = 1, \ldots, |V|$ with the k-means [29] algorithm into clusters $C_1, \ldots, C_k$.
Output clusters $S_1, \ldots, S_k$ with $S_i = \{j | y_j \in C_i\}$.

Given the definition of the Laplacian of the hypergraph, the TSD algorithm based on spectral clustering is shown in Algorithm 1. This algorithm use spectral clustering to discover the clusters of users, resources and tags simultaneously. Each cluster identified by the clustering algorithm corresponds to a particular induced sense for the target tag.

For a new occurrence of the target tag t_0, i.e. a user u post a resource r with tags $t_0, t_1, \ldots, t_n$, a score vector $\mathbf{s}$ is used to assign an induced sense to it. The element s_i in $\mathbf{s}$ corresponds to the induced sense S_i. Each object x in $c = \{u, r, t_1, \ldots, t_n\}$ contributes 1 to the score of the cluster that contains it, i.e. $s_i = |\{x \in S_i | x \in c\}|$. The induced sense with the largest score is chosen.

IV. Experiments

In this section, we conduct experiments on a dataset collected from a real-world social tagging system and report the experimental results. We first introduce the dataset used in the experiments and then perform both the qualitative and quantitative evaluation.

A. Dataset

We conduct experiments on a dataset collected from a real-world system, namely Delicious (http://delicious.com/), for online sharing bookmarks. The dataset is a partial dump of Delicious representing annotating activities during a certain period of time. Starting at Dec 2007, we crawled thousands of web pages from Delicious and extracted post information such as user, resource, post date and corresponding tags. To reduce the impact of idiosyncratic tags, we preprocess the dataset by computing the p-core [30], [31] at level 10 of the tripartite hypergraph representation of the dataset. The p-core at level t has the property that each user, resource and tag has/occurs in at least t annotations. The algorithm for p-core computation can be found in [30]. There were 282,016 users, 90,790 resources, 32,615 tags and 30,902,845 annotations in the preprocessed dataset.

Tag	#Ann.	#User	#Res.	#Tag
apple	70406	31375	2390	6036
bank	4305	3395	598	1097
bass	1273	840	182	659
cambridge	765	663	142	664
opera	50908	30211	1507	5743

TABLE I.
THE DATASET USED IN THE EXPERIMENTS. FIVE TAGS ARE MANUALLY SELECTED. FOR EACH TARGET TAG, THE NUMBER OF ANNOTATIONS ASSOCIATED WITH IT AND THE THE NUMBER OF USERS, RESOURCES, TAGS COOCURRED WITH IT ARE LISTED.

We manually selected five tags for performance evaluation. For each tag, we use its word senses in WordNet 3.0 [32] as the real senses, except for the tag "apple", for which a sense of "things related to Apple Inc." is added. For each tag, we randomly selected 200 posts associated it and labeled each occurrence of the tag with the appropriate sense, forming the labeled set D_L. All the other annotations forms the unlabeled set D_U. We use the unlabeled set to discover the senses for the target tag. The statistics of the dataset are shown in Table I.

B. Qualitative Insights

To obtain qualitative insights into the TSD method, we present some clustering results in this subsection. Table II list the 5 most frequently used tags within each cluster discovered by the proposed algorithm for the 5 tags. From these tables, we can see that different clusters corresponding to different senses of the target tags are discovered. For the tag "apple", two clusters are found. The first cluster contains tags such as "mac" and "osx" identifying the sense of "things related to Apple Inc.". The second one contains tags such as "recipe" and "dessert identifying the sense of "fruit". For the tag "bank", three clusters are found. The first cluster contains tags such as "finance" and "money" identifying the sense of "depository financial institution". The second one contains tags such as "stock" and "government", identifying the sense of "a supply or stock held in reserve for future use". The third one contains tags such as "photo" and "green", identifying the sense of "sloping land". For the tag "bass", two clusters are found. The first cluster contains tags such as "music" and "guitar" identifying the sense of "the member with the lowest range of a family of musical instruments". The second one contains tags such as "fishing" and "lake", identifying the sense of "any of various North American freshwater fish with lean flesh". For the tag "cambridge", three clusters are found. The first cluster contains tags such as "english" and "dictionary". This cluster seems to corresponds to the sense of "english dictionary published by cambridge", which is not in WordNet. The second one contains tags such as "research" and "education" identifying the sense of "a university in England". The third one contains tags such as "boston" and "mit" identifying the sense of "a city in Massachusetts just to the north of Boston".

Cluster	most frequently used tags
1	mac, osx, software, news, blog
2	recipes, recipe, cake, dessert, cookies

(a) Clusters for the tag "apple"

Cluster	most frequently used tags
1	finance, money, banking, financial, business
2	stock, government, economy, social, advice
3	photo, home, cool, gallery, green

(b) Clusters for the tag "bank"

Cluster	most frequently used tags
1	music, guitar, lessons, tabs, tutorial
2	fishing, lake, recipes, water, boat

(c) Clusters for the tag "bass"

Cluster	most frequently used tags
1	english, dictionary, language, reference, dictionaries
2	research, education, university, uk, science
3	boston, mit, museum, travel, food

(d) Clusters for the tag "cambridge"

Cluster	most frequently used tags
1	browser, software, tools, web, tips
2	music, classical, concert, mp3, art

(e) Clusters for the tag "opera"

TABLE II.
THE CLUSTERS DISCOVERED BY THE PROPOSED METHOD. FOR EACH TAG, THE 5 MOST FREQUENTLY USED TAGS WITHIN EACH CLUSTER ARE LISTED.

For the tag "opera", two clusters are found. The first cluster contains tags such as "browser" and "software" identifying the sense of "a commercial browser". The second one contains tags such as "music" and "classical" identifying the sense of "a drama set to music". These observations indicate that the proposed algorithm has the ability to discover useful clusters representing different senses of the target tag.

C. Quantitative Evaluation

Unlike the supervised alternative, evaluating unsupervised sense disambiguation algorithm is not straightforward. There are some alternatives to evaluate an unsupervised sense disambiguation algorithm. One is to manually examine the correctness of the sense assigned to each occurrence of the target tag. This approach has the drawbacks such as expensive to perform and subjective bias. Another alternative is to evaluate the algorithm according to certain performance metrics within an real-world application, such as an information retrieval system. The drawbacks of this method are that i) it is time consuming to build such an evaluation application and ii) it is difficult to separate the reasons for a good or bad performance. Suppose there are k induced senses and n real senses for the target tag, we adopt the following two alternatives for evaluation,

Supervised evaluation We adopt the evaluation method proposed in [14]. Besides the unlabeled set, the labeled set D_L is divided into two portions, namely the *mapping set* D_M, which contains 80% of the instances of D_L, and the *test set* D_T, which contains 20% of the instances of D_L. We use the mapping set to compute a $m \times n$ matrix M that maps the induced senses to the real senses. Each element $m_{i,j}$ in M is the probability of a tag having real sense j given that it has been assigned to the induced sense i. For each occurrence in the test set, the real sense corresponding to the largest element in $\mathbf{s}$ is chosen, where $\mathbf{s}$ is the score vector mentioned in Section III-B. The performance metric used here is accuracy

$$\text{Accuracy} = \frac{\text{\#correct answer provided}}{|D_T|}.$$

Unsupervised evaluation In [15], Agirre et al. proposed an unsupervised evaluation method. Since the induced senses are in fact some clusters of tags, we can trait the gold standard senses as classes and use performance measures from the clustering literature, such as entropy and purity, to evaluate the performance of the sense disambiguation algorithm. For the labeled set D_L, let $\Omega_1, \ldots, \Omega_k$ be the clusters generated by the TSD algorithm while $C_1, \ldots, C_n$ be the classes for the labeled senses. The performance metrics used here are entropy

$$\text{Entropy} = -\sum_{i=1}^{k} \frac{|\Omega_i|}{|D_L|} \sum_{j=1}^{n} p_{i,j} \log_2 p_{i,j},$$

where

$$p_{i,j} = \frac{|\Omega_i \cap C_j|}{|\Omega_i|},$$

and purity

$$\text{Purity} = \frac{1}{|D_L|} \sum_{i=1}^{k} \max_{j=1}^{n} |\Omega_i \cap C_j|.$$

We use the *first sense* method as the baseline. This method select the first sense within a ranking of all the senses of the target tag based as the result sense, independent of its context. In this study, we rank the senses of a target tag based on the frequency of occurrence of each sense in the labeled dataset. In Table I, the senses are listed according this ranking. We also compare the proposed method with the method proposed in by Yeung et al. in [16], which is also an unsupervised TSD method.

The only parameter to the proposed TSD algorithm is the number of clusters k. There are many studies focusing on the automatic selection of k [33], [34]. In our study, we employ an empirical approach by systematically varying the value of $k \in \{2, \ldots, 2n\}$ to select the value yielding the best performance of purity.

The evaluation results are reported in Table III. Table III(a) shows the number of senses (clusters) identified

Tag	Baseline	Yeung	Proposed
apple	1	3	2
bank	1	4	3
bass	1	4	2
cambridge	1	5	3
opera	1	4	2

(a) Number of senses (clusters) identified by each algorithm

Tag	Baseline	Yeung	Proposed
apple	61.0	82.0	**92.0**
bank	36.0	76.0	**83.0**
bass	39.0	82.0	**89.0**
cambridge	51.0	84.0	**92.0**
opera	59.0	89.0	**95.0**

(b) Supervised (Accuracy)

Tag	Baseline	Yeung	Proposed
apple	20.3	12.5	**10.1**
bank	35.4	19.4	**16.4**
bass	30.3	18.7	**15.3**
cambridge	22.9	13.6	**11.0**
opera	21.4	14.1	**12.3**

(c) Unsupervised (Entropy)

Tag	Baseline	Yeung	Proposed
apple	63.1	84.1	**87.9**
bank	55.4	79.3	**81.7**
bass	57.2	83.1	**85.3**
cambridge	61.2	87.9	**89.1**
opera	62.3	86.7	**90.1**

(d) Unsupervised (Purity)

TABLE III.
SUPERVISED AND UNSUPERVISED EVALUATION RESULTS. THE BASELINE AND THE PROPOSED METHOD ARE COMPARED BY PERFORMANCE MEASURED BY ACCURACY (FOR SUPERVISED EVALUATION), ENTROPY AND PURITY (FOR UNSUPERVISED EVALUATION).

by each algorithm. Only one sense is identified by the baseline algorithm since the it assigns the first sense to each tag occurrence. Generally, Yeung's algorithm discovers more clusters than the proposed algorithm.

It is interesting to notice that all the performance metrics in this table are significantly higher than the typical values for the task of WSD. This is because that, for TSD, the context information in a relational form is much more informative for sense disambiguation than that of WSD, for which the context information is semi-relational given that it is collected from natural language snippets.

Comparing the algorithms, we can find that both the Yeung's and the proposed algorithm outperform the baseline for all the performance metrics in both supervised and unsupervised evaluation. This indicates that the task of TSD is nontrivial, hence it is necessary to apply TSD algorithms to the social annotation data to clarify the vocabulary of folksonomies.

In supervised evaluation, the proposed method outperforms Yeung's algorithm for all the target tags and achieves a 7.6 point improvement of accuracy in average. Yeung's method folds the tripartite hypergraph structure into a matrix encoding only the connections between resources. Relatedness data such as the user-resource, user-tag and resource-tag relations is lost in the folding process. Thus, the performance of Yeung's method is clearly affected by this loss of information. On the contrary, the proposed approach models all the social annotations associated with the target tag by a 3-order tensor, which keeps all the information available. By applying the spectral clustering algorithm over the hypergraph induced from the tensor, clusters representing the senses of the target tag can be effectively discovered. Moreover, the spectral approach applied is robust to the noise annotations. All these factors contribute to the good performance of the proposed method. The results of unsupervised evaluation are analogous to those of the supervised evaluation. The proposed method outperforms Yeung's method in all the cases. In summary, the proposed method can obtain a convincing performance in both supervised and unsupervised evaluation.

V. CONCLUSION

We have demonstrated our method to disambiguate tag senses in folksonomies. The TSD method proposed is an unsupervised algorithm based on spectral clustering. For a target tag, all the annotations associated with it are modeled as a 3-order tensor. The hypergraph induced from this tensor is then partitioned with a spectral clustering algorithm to find the clusters representing senses of the target tag. We conduct experiment on the Delicious dataset. We perform both supervised and unsupervised evaluation to access the performance of the proposed method. We find that, the proposed method is superior to another method ignoring the multi-type interrelated data. This indicates that, by completely representing the context information with a 3-order tensor and employing the spectral clustering technique to discovering the senses of the target tag, the performance of TSD system can be significantly improved.

REFERENCES

[1] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?" in *Proceedings of the international conference on Web search and web data mining (WSDM'08)*, 2008, pp. 195–206.

[2] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, 2007, pp. 501–510.

[3] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles, "Exploring social annotations for information retrieval," in *Proceeding of the 17th international conference on World Wide Web (WWW'08)*, 2008, pp. 715–724.

[4] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 1, pp. 5–15, 2007.

[5] P. Heymann and H. Garcia-Molina, "Collaborative creation of communal hierarchical taxonomies in social tagging systems," Stanford University, Tech. Rep. 2006–10, April 2006.
[6] P. Schmitz, "Inducing ontology from flickr tags," in *Proceedings of the Collaborative Web Tagging Workshop at WWW'06*, 2006.
[7] J. C. Mallery, "Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers," Master's thesis. MIT Political Science Department, Cambridge, MA, USA, 1988.
[8] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 54, no. 236, pp. 433–460, 1950.
[9] H. Schütze, "Automatic word sense discrimination," *Computational Linguistics*, vol. 24, no. 1, pp. 97–123, 1998.
[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.
[11] A. Purandare and T. Pedersen, "Word sense discrimination by clustering contexts in vector and similarity spaces," in *Proceedings of the Conference on Computational Natural Language Learning (CoNLL'04)*, 2004, pp. 41–48.
[12] D. Lin, "Automatic retrieval and clustering of similar words," in *Annual Meeting of Association For Computational Linguistics (ACL'98)*, vol. 36, 1998, pp. 768–774.
[13] P. Pantel and D. Lin, "Discovering word senses from text," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)*, 2002, pp. 613–619.
[14] J. Véronis, "HyperLex: lexical cartography for information retrieval," *Computer Speech & Language*, vol. 18, no. 3, pp. 223–252, 2004.
[15] E. Agirre, D. Martınez, O. de Lacalle, and A. Soroa, "Two graph-based algorithms for state-of-the-art WSD," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 585–593.
[16] C. Yeung, N. Gibbins, and N. Shadbolt, "Tag meaning disambiguation through analysis of tripartite structure of folksonomies," in *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops*, 2007, pp. 3–6.
[17] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.
[18] K. Lee, H. Kim, H. Shin, and H. Kim, "Tag Sense Disambiguation for Clarifying the Vocabulary of Social Tags," in *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE'09)*, 2009, pp. 729–734.
[19] A. García-Silva, M. Szomszor, H. Alani, and O. Corcho, "Preliminary Results in Tag Disambiguation using DBpedia," in *Proceedings of the First International Workshop on Collective Knowledge Capturing and Representation (CKCaR'09)*, 2009.
[20] B. Long, Z. M. Zhang, X. Wú, and P. S. Yu, "Spectral clustering for multi-type relational data," in *Proceedings of the 23rd international conference on Machine learning (ICML'06)*, 2006, pp. 585–592.
[21] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Ma, "ReCoM: reinforcement clustering of multi-type interrelated data objects," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR'03)*, 2003, pp. 274–281.
[22] D. Zhou, J. Huang, and B. Schlkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Advances in Neural Information Processing Systems (NIPS) 19*, 2006, pp. 1601–1608.
[23] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recogn.*, vol. 41, no. 1, pp. 176–190, 2008.
[24] B. Long, X. Wu, Z. M. Zhang, and P. S. Yu, "Unsupervised learning on k-partite graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'06)*, 2006, pp. 317–326.
[25] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01)*, 2001, pp. 269–274.
[26] H. Zha, X. He, C. Ding, H. Simon, and M. Gu, "Bipartite graph partitioning and data clustering," in *Proceedings of the tenth international conference on Information and knowledge management (CIKM'01)*, 2001, pp. 25–32.
[27] F. Chung, *Spectral graph theory: CBMS Regional Conference Series in Mathematics*. American Mathematical Society Providence, RI, 1997.
[28] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973.
[29] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
[30] V. Batagelj and M. Zaversnik, "Generalized cores," *CoRR*, vol. cs.DS/0202039, 2002.
[31] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme, "Tag recommendations in social bookmarking systems," *AI Commun.*, vol. 21, no. 4, pp. 231–247, 2008.
[32] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. The MIT Press, May 1998.
[33] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. The MIT Press, 2004, pp. 1601–1608.
[34] D. Chakrabarti, "Autopart: parameter-free graph partitioning and outlier detection," in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04)*, 2004, pp. 112–124.

Kaipeng Liu is currently a Ph.D. candidate at Harbin Institute of Technology, China. He received his MS and BS degrees in computer science and technology from the Harbin Institute of Technology in 2006 and 2004, respectively. His research interests include information retrieval, data mining and machine learning.

Binxing Fang received his Ph.D. degree in computer architecture from Harbin Institute of Technology in 1989, his MS degree in computer science and technology from the Tsinghua University in 1984, and his BS degree in computer science and technology from Harbin Institute of Technology in 1981. He is currently a member of Chinese Academy of Engineering. His current research interests include information security, information retrieval, and distributed systems.

Weizhe Zhang doctor, associate professor. His research interests include network computing, cluster computing, parallel and distributed system. His undertaking projects are supported by the National Natural Science Foundation of China, 863 High-Tech Research Plan of China, Province and Ministry Science Foundation projects. He has published over 30 papers in journals and international conferences.