Learning Structured Knowledge from Social Tagging Data

A critical review of methods and techniques

Hang Dong, Wei Wang, Hai-Ning Liang
Department of Computer Science and Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, China
{Hang.Dong, Wei.Wang03, HaiNing.Liang}@xjtlu.edu.cn

***Abstract*—For more than a decade, researchers have been proposing various methods and techniques to mine social tagging data and to learn structured knowledge. It is essential to conduct a comprehensive survey on the related work, which would benefit the research community by providing better understanding of the state-of-the-art and insights into the future research directions. The paper first defines the spectrum of Knowledge Organization Systems, from unstructured with less semantics to highly structured with richer semantics. It then reviews the related work by classifying the methods and techniques into two main categories, namely, learning term lists and learning relations. The method and techniques originated from natural language processing, data mining, machine learning, social network analysis, and the Semantic Web are discussed in detail under the two categories. We summarize the prominent issues with the current research and highlight future directions on learning constantly evolving knowledge from social media data.**

Keywords—Knowledge Engineering, Knowledge Extraction, Social Media data, Social tagging data, Folksonomy, Knowledge Organization Systems, Ontology Learning

I. INTRODUCTION

Data on collaborative social websites contributed by millions of online users represent an essential source for the "collective intelligence", which many researchers have attempted to explore and exploit. The tags and the resulting folksonomies are originally utilized as efficient means for content annotation, organization and discovery. However, over the years they remain tacitly, gradually developing into a dormant collection of noisy, low quality and often ambiguous "keywords", which demonstrate little usefulness for content discovery and search on the information rich social Web of tremendous scale. To a great extent, the situation can be attributed to the lack of effective methods and techniques for deriving real semantics or knowledge from the user generated data.

In recent years, research on knowledge engineering from social media data has attracted great interests of the research communities, e.g., data mining, machine learning, natural language processing, information retrieval and the Semantic Web. One specific, challenging research topic in this line is concerned with learning structured knowledge by exploring and exploiting social tagging data or folksonomies. There have been numerous methods and techniques to induce semantics from the noisy and unstructured folksonomies, and to construct structured knowledge with rich semantics, which have been shown useful for many application areas, such as domain ontology enrichment [1], information retrieval and navigation [2-4], recommender systems [5-6], academic communication [7], e-learning [8], mood mining [9], geotagging [10], etc.

We believe that a comprehensive study on the related work is necessary and essential to researchers in this exciting research area. There have been several surveys about associating semantics to social tagging data, however, most of them are without a clear and concise taxonomy to categorize the key methods used in the studies, despite the fact that they cover many important elements, including preprocessing, spam detection, distance measures and evaluation [15], [53]. One exception is the widely cited work [18] published in 2012, which proposed a set of formal steps to learn structured knowledge from folksonomies and divided methods into clustering techniques, ontologies and hybrid categories. Nevertheless, there is a lack of a comprehensive taxonomy to organize the latest methods and techniques to learn structured knowledge from social tagging data. It is also important to know the issues of each type of methods and techniques.

Our main contribution in this paper is to provide a categorized view on the state-of-the-art, which enables readers to gain insights into the different methods and techniques and the specific problems that they can effectively solve. Moreover, we discuss the main limitations and pertinent issues with current research and identify future directions. The rest of the paper is organized as follows. In Section II, we review several types of Knowledge Organization Systems, varying from unstructured with less semantics to highly structured with richer semantics. In Section III, we present a categorization of current methods and technologies in learning knowledge structures from social media data. Under each sub-category, the methods and techniques and the problems that they intend to solve are discussed in detail. The limitations and issues with the existing work are presented in Section IV. We conclude the survey and discuss the future research issues in Section VI.

II. KNOWLEDGE ORGANIZATION SYSTEMS

The general term Knowledge Organization Systems (KOSs) is intended to encompass all kinds of schemes for organizing information and managing knowledge [19]. KOSs vary from unstructured to structured types and form a spectrum from weak semantics to strong semantics [12-14]. In the Semantic Web research, KOSs have received a lot of attention [11], for example, ontology representation using the Simple Knowledge Organization System[1] (SKOS). Many of the recent studies in the Semantic Web domain use ontologies of different formality (lightweight and heavyweight) to express different types of KOSs to facilitate knowledge representation and automated reasoning [12], [13], [52].

Pertinent to this paper are five kinds of KOSs, namely, folksonomies, term lists, concept hierarchies, taxonomies, and Ontologies. Folksonomies are highly unstructured and uncontrolled KOSs where users can freely add tags to annotate resources without constraints in most social tagging systems [15-16]. Therefore, Folksonomies inherit many of the problems in human natural language. Without special processing and treatment, systems are not able to discriminate different syntactic variations, e.g., polysemy, homonymy, synonymy and specificity (hypernym/hyponym) of tags and their senses [15], [17]. In addition, a large portion of tags is created for individual's use, e.g., "toread", which is hardly useful for others. Compared to ontologies, folksonomies lack a uniform representation to facilitate their sharing and reuse [18].

Term lists, such as for example, glossaries and gazetteers, are distinct from folksonomies, and they include widely accepted terms with clear definitions of their senses [19], rather than undefined terms with ambiguous meanings. Therefore, term lists are considered more structured than folksonomies, although less structured than concept hierarchies and taxonomies.

Concept hierarchies represent a set of concepts that are organized in a hierarchical fashion, typically with a subsumption relation. They are frequently used as the backbone of ontologies [13]. Taxonomies are well-structured, hierarchical and sometimes exclusive schemes to organize knowledge, such as the famous Dewey Decimal classification, and the personal or organizational file systems [20]. They provide good subject browsing facilities and interoperability with other services [21]. Building a concept hierarchy or taxonomy is resource demanding and often needs constant maintenance manually [22].

Ontology is defined as "a formal explicit specification of a shared conceptualization of a domain of interest" [13]. This definition captures the characteristics of "formality, explicitness, consensus, conceptuality and domain specificity" for knowledge specification [13]. Interrelation (relation between concepts), instantiation (assigning individual objects to classes), subsumption (is-a relationship), exclusion (is-different-from relationship) and axiomatization (complex statement about a domain) are the five essential elements in a formal ontology [13]. In the spectrum of formality, lightweight ontologies are less formal and process no or few axioms [13], but are more flexible and easier to maintain and use compare to heavyweight ontologies [23].

Much of the attention has been dedicated to learning structured KOSs with rich semantics (e.g., concept hierarchy, taxonomy and ontology) from less structured KOSs (e.g., folksonomies [15]) or unstructured KOSs such as scientific abstracts and text corpora [24-25]. More recently, social media data has become an important source for such learning tasks [26]. In this paper, we focus on studies using folksonomies, i.e. social tagging data, as a source to learn structured KOSs, e.g., term lists, concept hierarchies and ontologies.

III. OVERVIEW OF METHODS AND TECHNIQUES

Learning structured KOSs from folksonomies suffers from the problems associated with social tagging data, such as different syntactical variations, ambiguity of meanings (polysemy, homonymy and synonymy) and the noise in the tagging data. Existing methods and techniques approximately fall into two main categories: associating semantics to folksonomies and deriving structured KOSs from folksonomies. The former is to learn term lists (or concept lists) from folksonomies by expressing sense of terms using groups of tags and/or entries in external lexical resources such as Wikipedia[2] and WordNet[3]; the latter is to learn relations of tags and to organize them into more structured form (e.g., lightweight ontologies).

A. Learning Term Lists from Folksonomies

Term lists in this context refer to terms which represent clear senses derived from the shared conceptualizations in one or more social tagging systems. In literature, researchers also use the terms "discovering shared conceptualizations" [27], "[topic] sense induction" [15], [17] to describe the same task.

By extending the categorization in [17], we divide methods and techniques for learning term lists to Word Sense Disambiguation (WSD) and Word Sense Induction (WSI). A tree structure of research is illustrated below in Figure 1.

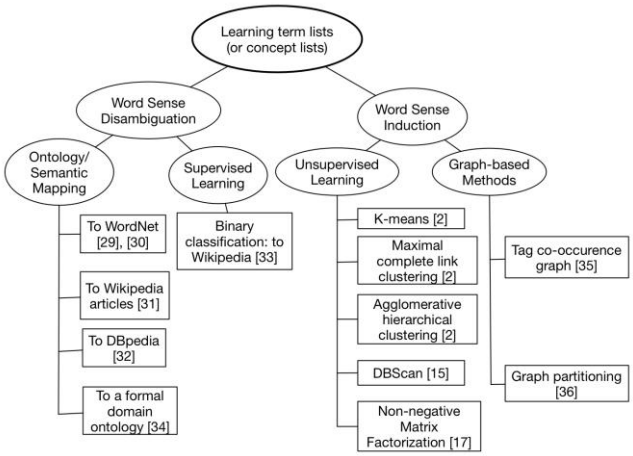

Fig. 1. An overview of methods and techniques to learn term lists from social tagging data

[1] http://www.w3.org/2004/02/skos/

[2] http://www.wikipedia.org/

[3] http://wordnet.princeton.edu/

1) Word Sense Disambiguation

In natural language processing, Word Sense Disambiguation (WSD) is defined as the task of selecting the correct sense of a word for a particular context [28], normally from a fixed inventory of potential senses. Choosing inventories for sense mapping and context definition of social tags are pertinent to methods in this area.

There have been a number of studies using WSD based methods to deal with ambiguous tag senses. The work in [29] has attempted to define tag senses using a set of elements or instances. They formalize a controlled tag as a social tag connected to an ordered list of linguistic concepts. To convert the del.icio.us[4] tags to controlled tags, manual validation was carried out after the automatic tokenization and lemmatization based on WordNet. Then, disambiguation was conducted based an algorithm making use of the tag collocation features: tag collocated as tokens in a split or in the annotation of same resources. Part-Of-Speech tagger was used to boost the score of the similarity of two words if they have the same part-of-speech. It is discovered that WordNet can only represent less than a half (48.7%) of the tags due to its static nature. Specifically, some tags (e.g., "apple", "mac") present in WordNet, but do not have the intended senses; while many other tags do not exist in WordNet [15].

WordNet is also used by [30] to associate semantics to folksonomies. The study created FLOR, an automatic framework to perform tag preprocessing, semantic definition and enrichment processes. Lexical representations of tags, along with tag synonyms derived through tag co-occurrence analysis in Flickr[5], are mapped against WordNet. For the disambiguation purpose, similarity between tags is calculated based on the number of common ancestors of the tags in WordNet and the length of their connecting path. Finally, each tag is assigned to at least one entity.

Wikipedia is used as an important source for tag sense disambiguation. In [31], articles in Wikipedia are mapped to social tags in del.icio.us. Similar to [30], tag groups are created based on tag co-occurrence in annotating resources. The relevance of a tag to a Wikipedia topic is computed by the sum of multiplication of the frequency of all grouped tags in the Wikipedia article with the weight of tag co-occurrences. Qualitative analysis is conducted and shows satisfied precision and recall.

The work in [32] uses DBpedia[6], a semantic representation of Wikipedia information, as the inventory to disambiguate social tags. More comprehensive situations of tag co-occurrence were proposed to analyze tags collocation in annotating resources, user profiles, user social networks and the whole folksonomy. However, similar to other studies, only the first co-occurrence pattern in annotating resources is implemented. The disambiguation algorithm is based on measuring the distance between tag contexts (grouped tags) and senses in the repository.

A supervised machine learning based method to map social tags in a Q&A website, StackOverflow[7], to Wikipedia, is introduced in [33]. An open-source toolkit, Wikipedia-Miner[8] [54], is used to detect concepts appearing in the tags' Wikipedia pages. The task is then cast as a binary classification problem by determining those candidates as equivalent or non-equivalent concepts. Features of the learning are those about Wikipedia concepts such as frequency, occurrence, max link probability, disambiguation confidence, etc. Precision, recall and F-measure are performed using classifiers including Bayes Network, KNN, SVM, decision tree, random forest and so on. It reported that the best F1 score achieved is 99.6% based on 10-fold cross-validation.

Domain expert ontologies are also used in tasks of tag sense disambiguation, for example, a formal ontology in the music domain has been used to disambiguate Last.fm[9] concepts [34]. A common feature for all the methods under this category is that they all heavily rely on the use of external lexical resources, such as WordNet, Wikipedia or domain ontologies. However, in many situations, authoritative lexical resources can only cover certain portion of all tags (the work in [29] reported the coverage is about 50%).

2) Word Sense Induction

Compared to WSD, Word Sense Induction (WSI) does not need external lexical resources, since each set of senses, e.g. a cluster of tags, is created automatically from the instances of words in the training set [28]. Most methods based on WSI employ unsupervised learning techniques, more specifically, clustering [28]. Computing similarity between tags is fundamental for these methods. Nearly all current methods have made use of co-occurrence features of tags in annotating same resources (tag-resource matrix) or created by same users (tag-user matrix), as summarized in [18].

The work in [2] evaluates three unsupervised learning techniques, namely, Hierarchical Agglomerative Clustering (HAC), Maximal Complete-Link Clustering and *K*-means clustering for generating tag clusters to enhance personalization and navigation in del.icio.us. The similarity between two tags is computed by treating each tag as a vector over the set of resources and using the adapted TF-IDF (term frequency-inverse document frequency) as the weighting scheme. The HAC algorithm begins by placing each tag in a singleton cluster, and continuously joins these clusters until all tags have been aggregated into one cluster. Similarity of two clusters is calculated based on the distance between their centroids. During iterations, it is suggested to tune several parameters, including step, generalization level and division coefficient to control the levels of hierarchies and to prevent over-clustering. The work reported that HAC has superior performance over the other two clustering techniques for personalized search tasks [2]. The second method evaluated in [2] is the Maximal Complete Link Clustering, a graph theoretical clustering technique. By setting a minimum similarity threshold, a graph can be generated by turning each tag a node and connecting nodes based on their similarity. The algorithm identifies all

[4] https://delicious.com/
[5] http://flickr.com/
[6] http://dbpedia.org/
[7] http://stackoverflow.com/
[8] http://wikipedia-miner.cms.waikato.ac.nz/
[9] http://www.last.fm/

maximal cliques in the graph to discover maximal complete clusters. Overlapping clusters are permitted in this method, which may well reflect the overlapping senses of tags. The downside is that the algorithm is computationally expensive for large datasets. The third method is the *K*-means clustering, a classical flat clustering technique. The number of clusters *k* is required be predetermined. Each tag is randomly assigned to one of *k* sets, and updated iteratively based on similarity measure between itself and all the cluster centroids. The main limitation of the *K*-means clustering is that it is not able to isolate irrelevant tags and to discriminate subtle senses of tags.

DBScan is a density-based clustering algorithm and has also been employed for tag sense induction [15], [29]. It does not require specifying the number of clusters *k* and is relatively resistant to noise and can handle clusters that have arbitrary shapes and sizes. The method used in [29] follows a two-step procedure. In the first step, it clusters the user-resource bipartite graph for each tag to discover the senses of the tag. The common user-resource-tag triplet is then substituted by user-resource-sense triplet. In the second step, it clusters the tag-senses in this user-resource-sense tripartite to find synonyms. Different collocation-based distance measures are employed and the results are evaluated using precision, recall and F-measure. Due to the sparsity of the dataset, the precision of the second step is extremely low.

A recent study proposes the use of non-negative matrix factorization (NMF), a dimensionality reduction technique, for tag sense induction [17]. NMF can factor the original tag-resource matrix into two matrices A and B, where the number of columns in A is equal to the number of rows in B (the number is denoted as *K*). The value of *K* can be interpreted as the number of tag clusters, which represent the topic senses. The work also performs an experimental study and shows that how the performance of recommendation for tags can be improved by controlling the value of *K*. The evaluation results based on del.icio.us dataset are impressive: very subtle senses can be discriminated in a tag such as "job" (having polysemous senses) and "ps" (having homonymous senses).

A graph-based co-occurrence model for sense induction is proposed in [35]. The work studies two graphs of connected tags: one constructed from tag collocation in annotating resources and another from tag collocation in user profiles. In both graphs, tags are nodes and collocation relations between two tags are edges. The strength of associations between two tags is determined by co-occurrence analysis. A similar graph representation of tag-resource collocation is in [36]. The authors show that the simple method can help learn lightweight ontologies from the tagging data.

B. Learning Relationships from Folksonomies

Research in this category aims to learn relations among tags, which can help generates a hierarchy of concepts or tags, [35], [37]. These hierarchies can further extended to lightweight or even formal ontologies. A variety of techniques from the areas of social network analysis, machine learning, data mining, natural language processing, the Semantic Web, etc., have already been proposed in literature. Figure 2 depicts the overview of methods and techniques in this category.

1) Social Network Analysis

Generally speaking, Social Network Analysis (SNA) can be seen as the study of relations mainly by means of graph theory [38]. In SNA, "centrality" refers to a set of metrics to quantify how influential or powerful a particular node (or group) is within a network [38-39]. The idea of using graph centrality to measure generality of tags is firstly proposed in [37], which aims to extract taxonomies from tagging data. A graph based on similarity of tags is created where tags are nodes and similarity measure meeting the threshold is established as edges. Nodes are ranked in an order by generality measured using the betweenness centrality. More general nodes are considered as being close to the root. The algorithm inserts leaf nodes to the current, more general node based on the betweenness value and iterates through all the tags.

The research in [26] extends the work in [37] by using WSI techniques in the preprocessing step to disambiguate tag senses, and using degree centrality instead of betweenness centrality to measure tag generality. The evaluation shows that the preprocessing step is crucial to generate a concept hierarchy with high quality.

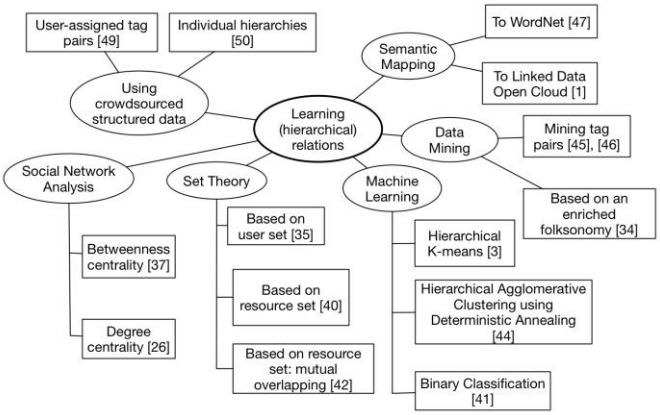

Fig. 2. An overview of methods and techniques to learn relations from social tagging data

2) Set Theory

In [35], when detecting the broader/narrower relations among tags, the authors make an assumption that a tag A is broader than a tag B if the set of entities (users or resources) classified under B is a subset of entities under A. They successfully use the co-occurrence analysis results and the associated user sets to generate tag hierarchies.

The study in [40] measures the inclusion and generalization degree between two tags based on the resources set. Assuming that there are two tags t_i and t_j; R_i, R_j are the set of resources annotated with the two tags, respectively. The generalization degree of t_i vs. t_j is computed as the ratio of $|(R_i \cap R_j)|$ to $|R_j|$. If two tags t_i and t_j co-occur, then the relationship "t_i subsumes t_j" can be defined if two conditions hold: first, t_i labels more resources than t_j; second, the generalization degree of t_i vs. t_j is greater than a predefined threshold. More recently, a similar algorithm based on the resource set, mutual overlapping, is also proposed in [42]. It is demonstrated that mutual overlapping performs well on the dataset collected from the e-business

website Taobao[10]. The two metrics, inclusion-generalization metric and mutual overlapping, are later used as two of the features for the supervised learning approach in [41].

3) Machine Learning

In terms of unsupervised learning, divisive (or top-down) hierarchical clustering techniques have been used to generate hierarchical structures from social tags. It is reported that in many circumstances, divisive algorithms produce more accurate hierarchies than agglomerative algorithms [43]. The work in [44] recursively splits the whole cluster into smaller groups which contain semantically coherent and precise elements (tags). By using the Deterministic Annealing (DA) algorithm, the number and the sizes of clusters are automatically determined. Although their results have a decent hierarchical appearance, the final tag structure cannot discriminate all sub, related and parallel relations from the ancestor and child nodes, as discussed in the study.

In the evaluation study [3], divisive hierarchical *K*-means clustering is performed on five social tagging datasets, Bibsonomy, CiteULike, del.icio.us, Flickr and Last.fm. The research evaluates the outcome of algorithm using semantic (based on reference) and pragmatic methods (based on greedy search to measure network navigability), and compared the results to two other methods using graph centrality measures [26], [37]. Results show that tag hierarchies generated from hierarchical clustering methods are less similar to gold-standard references, such as WordNet or Wikipedia, compare to the centrality-based algorithms [26], [37]. In addition, parameter values on the number of clusters are experimented, and no significant changes are found regarding the quality of the generated tag hierarchies. The most significant advantage of the hierarchical *K*-means clustering is that it is easy to implement and computationally efficient.

Supervised learning algorithms have also been used in detecting hierarchies from social tagging data based on learning of subsumption relations. A binary classification approach has been proposed in [41]. Features are extracted based on association rule mining, similarity measures, inclusion and generalization measures [40], mutual overlapping [41] and taxonomy search [34]. The positive and negative classes are labeled using WordNet and ConceptNet[11]. An under-sampling method, Tomek Link, is used to mitigate the class imbalance problem (e.g., there are significantly more negative examples) and the class overlapping problem (e.g., some examples share very similar characteristics). The classifiers used include C4.5, Random Forest, SVM, Naïve Bayes, Logistic Regression and AdaBoost. The result of F-measure achieves nearly 100% based on a ten times stratified 10-fold cross-validation. The results show that supervised learning based methods perform superiorly in extending existing knowledge hierarchies with the learned tag/concept relations.

4) Data Mining

Association analysis is a common technique in data mining. The mined association rules from tag pairs can be viewed as candidates to generate tag hierarchies. However, they are not strong enough to determine tag relations. Therefore existing studies often adopt some additional techniques to enhance the relation detection process. Another issue of employing association rule mining is to aggregate the User-Tag-Resource tripartite structure (*U*, *T*, *R*) into two-dimensional contexts.

The work in both [45] and [46] successfully derives tag pairs from del.icio.us by using association rule mining. It projects the three-dimensional folksonomy to different combinations two-dimensional contexts, and calculates the supports. To construct hierarchies from tag pairs, the two studies consider the relations between the resource set annotated by each tag, similar to the idea in [40]. The study in [46] takes into account the textual information of annotated documents to generate tag representation, which is then used to compute similarity as a supplement of association rule mining.

A more sophisticated approach is presented in [34]. The study uses association rule mining techniques based on an enriched folksonomy from a domain expert ontology. The original user-tag-resource triplet is enriched with user-concept-resource. Then the enriched folksonomy is projected onto a widely used transactional dataset. To learn a concept hierarchy ("ontology"), an important assumption is proposed, which is, the more popular a tag is, the more general it is and thus the higher the level it occupies in the hierarchical structure (known as the "generality-popularity" assumption [51]).

5) Semantic mapping

Semantic mapping, or semantic grounding, in this context refers to a series of methods to match social tags to external resources in order to obtain the relations between tags. Methods in this sub-category are in fact extensions of WSD methods but going further to seek relations between entities.

In [47], the subsumption relations among tags are defined by using their mapped concepts in WordNet. For a mapped concept *c*, the algorithm retrieves the available relations from *c* to the top of the hierarchical structure in WordNet. By combining the extracted structures from all mapped concepts, a tag ontology can be built to support the tag recommendation task.

A slightly different mapping approach is proposed in [1]. The study uses the Linked Open Data Cloud[12] as external resources. DBpedia in the Linked Open Data Cloud not only represents the knowledge in Wikipedia, but also links to a large number datasets in different domains. The dataset used in this research is del.icio.us folksonomy in the financial domain. For two preprocessed tags (local classes) lc_i and lc_j, the algorithm finds the path between mapped ontology classes Oc_i and Oc_j. Then the relation between lc_i and lc_j are established the same as each relation found between Oc_i and Oc_j. Based on this method, a lightweight ontology in financial domain is generated, which consists of 187 classes linking to each other and to 212 classes in three other ontologies by the *owl:sameAs* relation. The ontology also defines other relations that include *rdfs:subclassOf*, *owl:disjointWith* and so on. Qualitative evaluation by domain experts demonstrates a high precision at around 80.67%.

[10] http://www.taobao.com/

[11] http://conceptnet5.media.mit.edu/

[12] http://lod-cloud.net/

6) Crowdsourcing

Recently, crowdsourcing methods have been proposed to derive hierarchical structures from social tags, under the motivation that even the most sophisticated computational techniques cannot substitute the knowledge of human [48]. It is possible to assign the task of building and maintaining ontologies to users as everyday work processes. In [49], a new system, TagTree is implemented which allows users to annotate with tag pairs and single tags. By giving users an option to annotate tag pairs to describe resources, the method shows notable performance based on graph centrality measures [26], [37], however, the learned hierarchy is not as rich and expressive as the structure generated with single tags [49]. In addition, the new annotation process can have the risk of losing the simple and flexible style of folksonomies [49].

Some systems already have implicit function for users to assign hierarchical descriptions to resources. For example, Flickr let users group related photos into *sets* and related *sets* into *collections*, as discovered by [50]. Personal hierarchies are created in this way by users and are shown to be a rich resource of evidence for learning concept hierarchies. The study then adopts a relational clustering algorithm to aggregate these hierarchies into a "bushier tree". However, it is challenging to handle this special form of data.

IV. Issues of Current Methods and Techniques

In this section, we discussed some of the prominent issues related to the current research on learning knowledge from social tagging data. Based on these issues, we elicit some of the important future research directions in Section V.

A. Issues in Learning Term Lists

Learning term lists using the WSD based methods heavily relies on the external lexical resources or domain ontologies. The main issue of these methods is concerned with the coverage and quality of external resources. WordNet is the most widely used resource for tag sense disambiguation, but it does not cover new tags and their senses. The study [17] discovered that WordNet is not directly suited to senses of folksonomies, since the word sense definition in the folksonomies under study may be too fine (e.g., 16 meanings in "job"), or too coarse (e.g., all 3 different tag clusters representing 3 senses derived from folksonomies can only fit into 1 meaning out of 16 meanings in "job"). In addition, WordNet has only English words, thus unable to be employed for datasets compiled in other languages. Besides the issue of coverage, there seems no discussion in the literature about the Quality Assurance of external lexical resources, which has been widely used for auditing ontologies in the biomedical domain [55].

There are also several issues with the WSI based methods using unsupervised learning techniques, for example, clustering methods, such as *K*-means, need to specify the number of clusters first, thus may be unsuitable for tag sense induction [15]; clustering techniques may also generate unreasonable ("odd") sense groups [2] that are difficult to interpret.

The problem of biases in using co-occurrence features for learning concepts from social tags have also been raised in literature. In [15], the authors report that collocation-based distances are not precise enough in many situations. In [35], the authors argue that the roles of users are often ignored when considering tag collocation in annotating resources. Therefore, the clustered tags cannot reflect the accumulated knowledge contributed by the community. On the contrary, when considering tag collocation with user profiles, resources are ignored and the clustered tags cannot reflect the true use of tags.

B. Issues in Learning Relations

In the task of learning relations, methods using social network analysis generally apply graph centrality metrics to measure the popularity of tags, based on the assumption that central or popular tags are more general (called "popularity-generality" assumption in [51]). The evaluation results show that 77% accuracy based on the "popularity-generality" can be achieved. This approach is also proved to be more efficient than hierarchical clustering approaches in [6] according to different evaluation methods. Similarly, co-occurrence based on user-tag-resource triplet provides another common way to learn the broader/narrower relations between tags. However, these methods lack of theoretical foundation. A common limitation of the methods in this category is that they are not able to differentiate different types of relations.

Semantic mapping based methods also rely on authoritative external resources to learn relations. They also suffer from the limited coverage because of the external resources used. Another major limitation is that these relatively static (or slow-evolving) lexical resources or domain expert ontologies may not help efficiently capture the rapid changes or evolution in social media data.

Crowdsourcing based methods incorporate human users' efforts to alleviate the problems in other relation learning methods. Users, as well as domain experts, can annotate resources using structured forms of language in either explicit or implicit way. However, the extracted hierarchies completely relying on human participation might not be expressive enough to be employed many practical applications. More importantly, this approach requires special system design and can have the risk of losing the simplicity and flexibility of folksonomies [49].

V. Conclusion and Future Work

In this survey paper, we have reviewed the representative works in learning structured knowledge from social tagging data. We categorized the state-of-the-art methods into two groups, those of learning term lists and those learning relations and focused our study on the detailed methods and techniques. Moreover, we discussed the specific problems that the existing methods and techniques can solve and identify their advantages and limitations. We believe that the survey will give the researchers from the broad communities of data engineering and knowledge discovery a highly structured view of the vast number of studies, and benefit them by providing a more comprehensive understanding of the relevant research towards harvesting the "collective intelligence". Different from many of the existing studies, our paper overviewed the core methods and techniques deriving concepts and their relations from social tagging data. For related review papers focusing on data

preprocessing, evaluations, and formal steps to associate semantics to folksonomies, the readers are advised to refer to [15], [3], [18] respectively.

Based on our survey, we believe that there are a number of important future research directions: first, we need to perform larger scale studies on the existing social media data and to build highly usable knowledge structures. Second, we need to develop efficient methods and techniques that can model the evolution of the knowledge structure along with the constantly changing social media data in order to capture the up-to-date knowledge and changes in the communities. This important aspect, i.e. evolution of knowledge in folksonomies, has not been studied sufficiently. Third, we need to develop learning techniques that can work with the trends of "big social media data". Social media data is being produced in an unprecedented pace and new techniques need to be able to process this tremendous amount of data both effectively and efficiently. This also provides an essential opportunity for researchers to integrate data in different sources and to derive more comprehensive knowledge representing the "wisdom of the crowd".

REFERENCES

[1] A. García-Silva, L. J. García-Castro, A. García, and O. Corcho, "Social tags and Linked Data for ontology development: A Case Study in the Financial Domain," in *the Proc. 4th Int. Conf. Web Intelligence, Mining and Semantics (WIMS14)*, Thessaloniki, Greece, 2014.

[2] J. Gemmell, A. Shepitsen, M. Mobasher, and R. Burke, "Personalization in folksonomies based on tag clustering," in *the Proc. 6th Workshop Intelligent Techniques for Web Personalization and Recommender Systems*, Chicago, Illinois, USA, 2008.

[3] M. Strohmaier, D. Helic, D. Benz, C. Körner, and R. Kern, "Evaluation of folksonomy induction algorithms," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1-22, Sep. 2012.

[4] E. Bouillet, Z. Liu, A. Ranganathan, and A. Riabov, "Method for enhancing search and browsing in collaborative tagging systems through learned tag hierarchies," U.S. Patent 8799294 B2, Aug. 2014.

[5] F. Gedikli and D. Jannach, "Recommender systems, semantic-based," in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj & J. Rokne, Eds. New York: Springer, 2014, pp. 1501-1510.

[6] I. Cantador, I. Konstas, and J. M. Jose, "Categorising social tags to improve folksonomy-based recommendations," *Web semantics: science, services and agents on the World Wide Web*, vol. 9, no. 1, pp. 1-15, Mar. 2011.

[7] D. Parra and P. Brusilovsky, "Collaborative filtering for social tagging systems: an experiment with CiteULike," in *the Proc. 3rd ACM Conf. Recommender systems (RecSys'09)*, New York, USA, 2009.

[8] L. Zhuhadar, S. R. Kruk, and J. Daday, "Semantically enriched Massive Open Online Courses (MOOCs) platform," *Computers in Human Behavior*, vol. 51, part B, pp. 578-593, Oct. 2015.

[9] P. Saari and T. Eerola, "Semantic computing of moods based on tags in social media of music," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2548-2560, Oct. 2014.

[10] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *the Proc. 30th Annu. Int. ACM SIGIR Conf. Research and development in information retrieval*, Amsterdam, The Netherlands, 2007.

[11] R. R. Souza, D. Tudhope, and M. B. Almeida, "Towards a taxonomy of KOSs: Dimensions for classifying Knowledge Organization Systems," in *the 11th ISKO Int. Conf.*, Rome, Italy, 2010.

[12] M. Bergman. (2007, May 16). *An intrepid guide to ontologies* [Online] AI3:::Adaptive Information. Available: http://www.mkbergman.com/374/an-intrepid-guide-to-ontologies/

[13] S. Grimm, A. Abecker, J. Völker, and R. Studer, "Ontologies and the Semantic Web," in *Handbook of Semantic Web Technologies*, J. Domingue, D. Fensel, and J. Hendler, Eds. Berlin Heidelberg: Springer, 2011, pp. 507-579.

[14] L. M. Zeng, "Knowledge Organization Systems (KOS)," *Knowledge Organization*, vol. 35, n. 2-3, pp. 160-182, 2008.

[15] P. Andrews and J. Pane, "Sense induction in folksonomies: a review," *Artificial Intelligence Review*, vol. 40, no. 2, pp. 147-174, Aug. 2013.

[16] T. V. Wal. (2007, Feb. 2). *Folksonomy* [Online]. Available: http://vanderwal.net/folksonomy.html

[17] J. Chen, S. Feng, and J. Liu, "Topic sense induction from social tags based on non-negative matrix factorization," *Information Sciences*, vol. 280, pp. 16-25, Oct. 2014.

[18] A. García-Silva, O. Corcho, H. Alani, and A. Gómez-Pérez, "Review of the state of the art: Discovering and associating semantics to tags in folksonomies," *The Knowledge Engineering Review*, vol. 27, no. 01, pp. 57-85, Mar. 2012.

[19] G. Hodge, *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: The Digital Library Federation Council on Library and Information Resources, 2000.

[20] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol, 32, no. 2, pp. 198-208, Apr. 2006.

[21] B. M. Villazón-Terrazas, *A Method for Reusing and Re-engineering Non-Ontological Resources for Building Ontologies*. Amsterdam: IOS Press, 2012.

[22] E. Tsui, W. M. Wang, C. F. Cheung, and A. S. M. Lau, "A concept–relationship acquisition and inference approach for hierarchical taxonomy construction from tags," *Information Processing & Management*, vol. 46, no. 1, pp. 44-57, Jan. 2010.

[23] C. V. Damme, M. Hepp, and K. Siorpaes, "Folksontology: An integrated approach for turning folksonomies into ontologies," in *Int. Workshop Bridging the Gap between Semantic Web and Web located at the 4th European Semantic Web Conf.* (ESWC 2007), Innsbruck, Austra, 2007, pp. 57-70.

[24] S. Li, Y. Sun, and D. Soergel, "A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis," *Scientometrics*, vol. 103, no. 3, pp. 1023-1042, June 2015.

[25] P. Cimiano, A. Hotho, and S. Staab, "Learning concept hierarchies from text corpora using Formal Concept Analysis," *J. Artif. Intell. Res.(JAIR)*, vol. 24, pp. 305-339, Aug. 2005.

[26] D. Benz, A. Hotho, S. Stützer, and G. Stumme, "Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge," in *Proc. WebSci10: Extending the Frontiers of Society On-Line*, Raleigh, NC: US, 2010.

[27] R. Jäschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme, "Discovering shared conceptualizations in folksonomies," *Web semantics: science, services and agents on the World Wide Web*, vol. 6, no. 1, pp. 38-53, Feb. 2008.

[28] D. Jurafsky and J. H. Martin, "Chapter 18: Computing with word senses," in *Speech and language processing* [online]. Draft of July 2015. Available: http://web.stanford.edu/~jurafsky/slp3/18.pdf

[29] P. Andrews, J. Pane, and I. Zaihrayeu, 'Semantic disambiguation in folksonomy: A Case Study,' in *Advanced Language Technologies for Digital Libraries*, Bernardi, R., Chambers, S., Gottfried, B., Segond, F., and Zaihrayeu, I., Eds. Springer Berlin Heidelberg, 2011, pp. 114-134.

[30] S. Angeletou, M. Sabou, and E. Motta, 'Semantically enriching folksonomies with FLOR,' in *Proc. 1st Int. Workshop Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) at The 5th Annu. European Semantic Web Conf. (ESWC 2008)*, Tenerife, Spain, 2008.

[31] L. Kangpyo, K. Hyunwoo, S. Hyopil, and K. Hyoung-Joo, "Tag sense disambiguation for clarifying the vocabulary of social tags," in *Computational Science and Engineering, 2009. CSE '09. International Conference on* , vol. 4, pp. 729-734, 2009.

[32] A. García-Silva, M. Szomszor, H. Alani, and O. Corcho, "Preliminary results in tag disambiguation using dbpedia," in *Knowledge Capture (K-*

Cap'09)–1st Int. Workshop Collective Knowledge Capturing and Representation (CKCaR'09), Redondo Beach, California, USA, 2009 ©ACM.

[33] A. Joorabchi, M. English, and A. E. Mahdi, "Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow," *Journal of Information Science*, published online before print, May 2015.

[34] L. B. Marinho, K. Buza, and L. Schmidt-Thieme, "Folksonomy-based collabulary learning," in *The Semantic Web - ISWC 2008*. Springer Berlin Heidelberg, 2008, pp. 261-276.

[35] P. Mika, "Ontologies are us: A unified model of social networks and semantics," in *The Semantic Web–ISWC 2005*. Springer, 2005, pp. 522-536.

[36] G. Begelman, P. Keller, and F. Smadja, "Automated tag clustering: Improving search and exploration in the tag space," in *the Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland, 2006.

[37] P. Heymann and H. Garcia-Molina. (2006, Apr. 26). *Collaborative creation of communal hierarchical taxonomies in social tagging systems* [online]. Technical Report, Stanford InfoLab. Available: http://ilpubs.stanford.edu:8090/775/

[38] M. Tsvetovat and A. Kouznetsov, *Social Network Analysis for Startups: Finding Connections on the Social Web*. Sebastopol, CA: O'Reilly Media, 2011.

[39] J. Scott and P. J. Carrington. *The SAGE Handbook of Social Network Analysis*: SAGE Publications, 2011.

[40] P. De Meo, G. Quattrone, and D. Ursino, "Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies," *Information Systems*, vol. 34, no. 6, pp. 511-535, Sep. 2009.

[41] A. S. C. Rêgo, L. B. Marinho, and C. E. S. Pires, "A supervised learning approach to detect subsumption relations between tags in folksonomies," in *the Proc. 30th Annu. ACM Symp. Applied Computing (SAC '15)*, Salamanca, Spain, 2015, pp. 409-415.

[42] S. Cai, H. Sun, S. Gu, and Z. Ming, "Learning concept hierarchy from folksonomy," in *the Web Information Systems and Applications Conference (WISA), 2011 Eighth*, Chongqing, China, 2011, pp. 47-51.

[43] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. New York: Cambridge University Press, 2009.

[44] M. Zhou, S. Bao, X. Wu, and Y. Yu, "An unsupervised model for exploring hierarchical semantics from social annotations," in *The Semantic Web*, Springer, 2007, pp. 680-693.

[45] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme, "Mining association rules in folksonomies," in *Data Science and Classification*, V. Batagelj, H.-H. Bock, A. Ferligoj, & A. Žiberna, Eds. Springer Berlin Heidelberg, 2006, pp. 261-270.

[46] G. Solskinnsbakk and J. Gulla, "A hybrid approach to constructing tag hierarchies," in *On the Move to Meaningful Internet Systems, OTM 2010*, Vol. 6427, R. Meersman, T. Dillon, & P. Herrero, Eds. Springer Berlin Heidelberg, 2010, pp. 975-982.

[47] E. Djuana, Y. Xu, and Y. Li, "Learning personalized tag ontology from user tagging information," in *the 10th Australasian Data Mining Conference (AusDM 2012)*, Sydney, Australia, December 2012, Conf. in Research and Practice in Information Technology (CRPIT), Vol. 134, 2012.

[48] H. Lin, and J. Davis, "Computational and crowdsourcing methods for extracting ontological structure from folksonomy," in *The Semantic Web: Research and Applications*, Vol. 6089, L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, & T. Tudorache, Eds. Springer Berlin Heidelberg, 2010, pp. 472-477.

[49] F. Almoqhim, D. Millard, and N. Shadbolt, "An approach to building high-quality tag hierarchies from crowdsourced taxonomic tag pairs," in *Social Informatics*, Vol. 8238, A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin, & B. Dai, Eds. Springer International Publishing, 2013, pp. 129-138.

[50] A. Plangprasopchok, K. Lerman, and L. Getoor, "Growing a tree in the forest: constructing folksonomies by integrating structured metadata," in *the Proc. of the 16th ACM SIGKDD Int. Conf. Knowledge discovery and data mining*, Washington, DC, USA, 2010.

[51] F. Almoqhim, D. Millard, and N. Shadbolt, "Improving on popularity as a proxy for generality when building tag hierarchies from folksonomies," in *Social Informatics*, Vol. 8851, L. Aiello and D. McFarland, Eds. Springer International Publishing, 2014, pp. 95-111.

[52] B. Smith, C. Welty, 'FOIS introduction: Ontology---towards a new synthesis,' in *Proc. Int. Conf. Formal Ontology in Information Systems, (FOIS* '01*)*, Ogunquit, Maine, USA, 2001, pp. iii-ix.

[53] F. Jabeen, S. Khusro, A. Majid, and A. Rauf, "Semantics discovery in social tagging systems: A review", *Multimedia Tools and Applications*, published online, pp. 1-33, Oct. 2014.

[54] D. Milne and I. H. Witten, "An open-source toolkit for mining Wikipedia", *Artificial Intelligence*, vol. 194, pp. 222-239, Jan. 2013.

[55] X. Zhu, J.-W. Fan, D. M. Baorto, C. Weng, and J. J. Cimino, "A Review of Auditing Methods Applied to the Content of Controlled Biomedical Terminologies", *Journal of biomedical informatics*, vol. 42, no. 3, pp. 413-425, Mar. 2009.