Computing and Informatics, Vol. 30, 2011, 165–188

TAGS ARE RELATED: MEASUREMENT OF SEMANTIC RELATEDNESS BASED ON FOLKSONOMY NETWORK

Chao Wu, Bo Zhou

Department of Computer Science
Zhejiang University
Yugu Road 165
Hangzhou, Zhejiang, China
e-mail: {wuchao, bzhou}@zju.edu.cn

Manuscript received 22 June 2009; revised 16 September 2009
Communicated by Patrick Brézillon

Abstract. Folksonomy and tagging systems, which allow users to interactively annotate a pool of shared resources using descriptive tags, have enjoyed phenomenal success in recent years. The concepts are organized as a map in human mind, however, the tags in folksonomy, which reflect users' collaborative cognition on information, are isolated with current approach. What we do in this paper is to estimate the semantic relatedness among tags in folksonomy: whether tags are related from semantic view, rather than isolated? We introduce different algorithms to form networks of folksonomy, connecting tags by users collaborative tagging, or by resource context. Then we perform multiple measures of semantic relatedness on folksonomy networks to investigate semantic information within them. The result shows that the connections between tags have relatively strong semantic relatedness, and the relatedness decreases dramatically as the distance between tags increases. What we find in this paper could provide useful visions in designing future folksonomy-based systems, constructing semantic web in current state of the Internet, and developing natural language processing applications.

Keywords: Folksonomy, semantic relatedness, tag, tagging system, del.icio.us, measurement

Mathematics Subject Classification 2000: 94A15

1 INTRODUCTION

A tag is a non-hierarchical keyword or term assigned to a piece of information. Tagging systems, such as those used in social bookmarking sites like Del.icio.us[1], photograph sharing services like Flickr[2], and electronic marketplace like eBay[3], are exploding in popularity on the Internet. Tagging systems are often placed in opposition to taxonomic models, for tags are emerged from community of users, rather than being defined by a single person or an organization. The collaborative tagging, along with the user-generated taxonomy, is usually called 'folksonomy', originated from combining the words 'folk' and 'taxonomy'. Folksonomy is one of the most noticeable features in the current Web 2.0, and many research efforts have been paid on studying the structure of these tagging systems and users' behavior in them.

Many research efforts have been paid on studying tags. Recently, the main concern is tagging system's application on social networking, such as work in [16, 22, 35, 43, 72]. However, the application of tagging system currently outpaces our understanding of its inherent features. There is a lack of some foundational research on tagging and folksonomy, such as tags' semantic information discussed in this paper. Currently in collaborative tagging, tags are isolated without semantic relatedness. It is not coincident with the fact that concepts are connected as a map in human mind. As a result, we try to identify tags' connections and their semantic information in this paper, because it is important to identify a stable and meaningful folksonomy before fully adopting it to different applications.

Tag provides a simple and direct mechanism to create annotations. However, Mathes [40] believed the simplicity and ease of use of tagging would lead to problems with current folksonomy systems: Tags have little semantics and thus cause local variations. It is difficult to say the resulting set of tags could always correctly and consistently represent the tag authors' mental model. Besides, without the meanings and correlations between a set of tags, it is difficult to aggregate tagging data from different applications or services. All these limitations are due to the lack of a uniform structure and semantic representation in tagging systems.

As in the case of ontology, most of the methodologies for building ontologies rely on specialized people or organizations, rather than the daily users. However, it is usually acknowledged that ontologies are shared understandings that have to be constructed within social processes. In current state, social semantics could provide a valuable source of empirically-derived knowledge to enrich and validate controlled semantics [28]. For example, there are some efforts trying to create annotation from collaborative tagging [20]. However, one key challenge is that the user-generated tags show lack of semantic relations between them, which is needed in ontologies as relations between classes of concepts and instances. As a result, to leverage the social

[1] `http://delicious.com`
[2] `http://www.flickr.com`
[3] `http://www.ebay.com`

power of subject metadata description, we need to investigate semantic information in tag sources. Solving this problem is a critical first step to successfully deploying social semantic.

As a result, in this paper we try to study the semantic information, specifically semantic relatedness, within tagging system. We use different approaches to connect tags. Although tags are non-hierarchical, they could be related by co-occurrence or resource context, and thus form folksonomy networks. Upon these networks, we measure the semantic relatedness between tags with multiple methods. After the experiment, we find relatively high relatedness within folksonomy networks.

We hope what we have done will make a step forward to gain better insights into folksonomy and tagging. Specifically, we think our work can contribute in the following three aspects:

1. For better ontologies: Tag ontology can robustly represent entities and relationships. The network of folksonomy contains the semantic information of underlying resource space. The similarity distance between tags can also be viewed as distance of represented resources. This kind of machine-processable data could be applied in many applications, such as automatic classifying documents.
2. For learning users: The tag network can be used to get insights into the tagging behavior of users. It reflects tag authors' usage of words in describing resources, showing a map of concepts in human's mind.
3. For better tag-based system: We believe that users should not have to choose between pure tag-based models and pure taxonomic models with closed vocabularies. Many systems such as question answering, multi-document summarization, and information retrieval need robust numerical measures of lexical relatedness. The best practice will leverage statistical natural language processing techniques together with domain knowledge, and form a tagging system that preserves the flexibility of the tagging interface for annotation while also benefiting from the power and utility of a faceted ontology in the search and browse interface.

The rest of this paper is organized as follows: We provide additional background and related work in Section 2. The formation of folksonomy network with both methods of co-occurrence and resource context will be stated in Section 3, and different measures of semantic relatedness in Section 4. The experiment and its result will be shown in Section 5, and discussion on the result in Section 6. Finally, we conclude with plans for future work in Section 7.

2 RELATED WORK

Within online social networking sites, tagging systems allow users to interactively annotate a pool of shared resources using descriptive tags. Many works have been done on tags' usage and properties. Li [38] thought tagging was not only a method for organizing contents to facilitate the users who created it, but also a navigation

mechanism for users to discover interesting resources. In [21], the authors stated advantages of tagging systems as low barrier to entry, dynamic information, and decentralization. Besides, particular aspects of tagging system have been elaborated in more detail, such as ranking of contents [23] and discovering trends in the tagging behavior of users [12, 24].

For semantic web, research community has also realized the importance of tagging systems, especially in current state of the Internet [46]. Tagging and resulting folksonomy provide a kind of social semantic, made by a large number of normal Web users with implicit social interactions on the open Web without a pre-defined formal ontology [70]. Golder [17] thought collaborative tagging offers an alternative to current efforts at semantic web ontologies. Gruber [19] emphasized the need for folksonomy and ontologies to work together. Qin [28] stated that social semantics and controlled semantics can benefit from each other in a profound way in metadata description. Specia and Motta [63] proposed to integrate folksonomy and ontologies to enrich tag semantics; and in [9], the author believed that the successful application of the tagging paradigm could be seen as a lowercase semantic web. One can then build upon this lowercase semantic web as a basis for the introduction of more semantics, thus advancing further towards the Web 2.0 ideas.

Towards utilizing semantic information in folksonomy, there have been a number of efforts to add more structure and semantics to conventional tagging systems. Cattuto et al. [7] observed small world effects by analyzing a network structure of folksonomy from Bibsonomy[4] and del.icio.us. Golder [18] analyzed the structure of collaborative tagging systems, as well as their dynamic aspects, and tried to discover regularities in users' activity, tag frequencies, kinds of tags used, and bursts of popularity in tagging.

In this paper, we construct networks of folksonomy and apply semantic relatedness measures upon them. The construction of folksonomy network needs to identify tag relations, which can be done by applying statistical methods with tag co-occurrences and resource similarity (tag context). There are some studies [4, 42, 61] demonstrating the associative and hierarchical relationships of similarity between tags inferred from tag concurrence analysis, while some other works [14, 33] utilized the context of tags (text in documents, information from Wikipedia, etc.) to identify their relationships. We will use both approaches to form folksonomy networks in this paper.

We will conduct measurement of semantic relatedness on folksonomy networks. Semantic relatedness refers to human judgments of the degree to which a given pair of concepts is related. Measures of semantic similarity between concepts are widely used in natural language processing, such as word sense disambiguation [48, 64], information retrieval [13], interpretation of noun compounds [30], and textual inference [53]. Lots of practices have been done on this area over the years. For example, Rada et al. [52] traversed MeSH, a term hierarchy for indexing articles in

[4] `http://www.bibsonomy.org`

Medline, and devised a "semantic distance" measure based on semantic networks. Taking a similar approach, Caviedes and Cimino [8] developed the CDist measure for finding path lengths in the UMLS hierarchy.

Many techniques have been proposed to automatically calculate the semantic relatedness of concepts. Most methods could fall in following three categories:

1. **Path finding measures** combine the words with a lexical resource, which have been shown to have a high correlation with those of human subjects, to find the shortest path between two words. Early in the 1960's, Quillian [51] had used the content of a machine readable dictionary to make inferences about word meanings and proposed that the contents of a dictionary be represented in a semantic network. Following his idea, many works [31, 32, 71] used the semantic network from lexical resources (English dictionary, Roget's Thesaurus, etc.) to define word similarity. Currently, most methods of this kind utilize WordNet for measurement. Budanitsky and Hirst [5] provided a survey of these WordNet-based measures of lexical similarity.
2. **Information content measures** calculate concept relatedness with derived statistical information from text corpora. The limitation of purely path based measures is that the degree of semantic similarity implied by a single link is not consistent. Resnik [54] attempted to address this problem by augmenting concepts with a corpus-based statistics known as information content, which is essentially a measure of the specificity of a concept. The information content of a concept is estimated by counting the frequency of that concept in a large corpus of text. With similar idea, Sahami and Heilman [57] proposed to use the Web as a source of additional knowledge for measuring similarity. More recently, measures incorporating information from Wikipedia are widely used. Strube and Ponzetto [65] were the first to compute measures of semantic relatedness using Wikipedia. Their approach, known as WikiRelate!, took familiar techniques that had previously been applied to WordNet and modified them to take advantage of the data found within Wikipedia; and Milne [45] proposed the "Wikipedia Link Vector Model", which used only the hyperlink structure of Wikipedia rather than its full textual content in measuring relatedness between concepts.
3. **Context vector measure** compares context vectors derived from statics of large corpora with vectors. It was firstly developed by Wilks et al. [69]. Their method expanded the dictionary glosses with related words. Then a lot of other methods were proposed. Niwa and Nitta [47] compared context vectors with co-occurrence statistics derived from the path lengths in a network that represented their co-occurrence in dictionary definitions. Patwardhan and Pedersen [50] used the co-occurrence information along with the WordNet definitions to build gloss vectors corresponding to each concept in WordNet. Numeric scores of relatedness are assigned to a pair of concepts by measuring the cosine of the angle between their respective gloss vectors. Gabrilovich and Markovitch [15] proposed Explicit Semantic Analysis (ESA), a machine learning method that

represents the meaning of texts from Wikipedia as a weighted vector, to compute semantic relatedness.

Other than above kinds of method, there are some other ways of relatedness measurement. For example, research done by Spasic and Ananiadou [62] defined a similarity metric based on a variation of edit distance [68] applied at a word level. In their approach, the semantic similarity of two terms was the cost associated with converting one term to another, using insert, delete and replace operations on words (instead of letters). Hughes and Ramage [26] presented the application of random walk Markov chain theory to measuring lexical semantic relatedness.

3 FOLKSONOMY NETWORK

Tags in folksonomy could be related. What we did in this paper is to identify their relationships and check whether they have semantic information. Here in this part, we will try to form the network of folksonomy, connecting tags with different methods.

The idea of forming folksonomy network is natural. In human mind, concepts are connected to form a map; and within collaborative tagging, tags are not isolated but connected by users' tagging behavior (actually, tagging systems such as del.icio.us have already used some mechanism to bundle tags and form a hierarchical structure). Although the connection among tags is implicit, we could infer it from the explicit connection between tags and resources. Here we use both methods of tag co-occurrence and resource context to infer these connections. Before we present our method of forming folksonomy network, let us first introduce the concept of it.

3.1 Concept of Folksonomy Network

We could consider folksonomy's underlying data structure as a hyper-graph. Although folksonomy is non-hierarchical, tags are related. Words trigger reasoning at a much deeper level that manipulates concepts – the basic units of meaning that serve humans to organize and share their knowledge; and humans usually agree with the relative relatedness of concepts [6, 41, 44, 56]. So it is necessary to organize concepts in a network, rather than isolated words.

In our opinion, tag acts as a bridge in three occasions:

- Tag connects users and resources. It is users' subjective description on objective resources. Tag provides a mechanism for users to find resources and for resources to associate with users.
- Tag connects resources. Different resources can share a common tag and are thus clustered. For example, in social bookmarking system like Del.icio.us, bookmarks are clustered by tags, and users can find other bookmarks that may interest them from the current bookmark's tags.

- Tag connects users. Users are linked by collaboration tagging and are able to share resources. By this way, they can find users having similar interest via shared tags.

We combine these three connections and extract the relationships between tags from them, as shown in Figure 1.

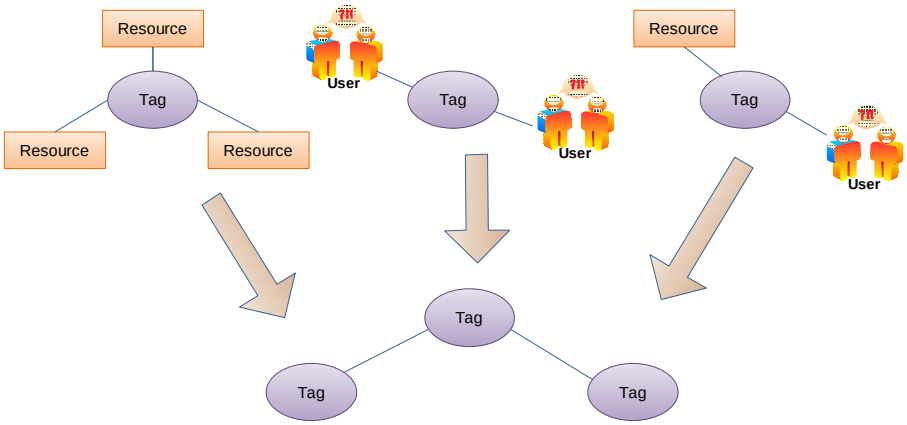


Fig. 1. Combine the tags' connections with resources and users to form a tags' network

3.2 Methods of Forming Folksonomy Network

Before beginning, let us clarify some of the generic terminology we use. We adopt Gruber's formalization of tagging as a three-place relation [19]:

Tagging(resource, tag, tagger)

Resources are digital information objects such as web pages, photos or video clips. A *tag* is a user-defined string associated to resources in the act of tagging. A *tagger* is the user who gives the tag to the target resource. *Tagging* is defined as the process of attaching tags to resources by taggers.

Upon this definition, we adopt the formal notation for folksonomy introduced in [23]:

A folksonomy is a tuple

$$F := (U, T, R, Y)$$

where:

- U, T, and R are finite sets of taggers, tags and resources
- Y is a ternary relation between them, i.e. $Y \subseteq U \times T \times R$ called tag assignments.

Next, let us introduce our methods of tag's relationship extraction. The method of relationship extraction could be tags' co-occurrence and resource context.

3.2.1 Tag Co-Occurrence

Since the process of tagging is inclusive [17], a large overlap often exists among resources marked with different tags. We could introduce a network of tag co-occurrence, with an assumption: If two tags frequently co-occurred, there ought to be some type of relation between them or else they would not be frequently tagged together by users. The idea is similar to "association rules" described in [1, 2], that if many transactions contain both two items, these two have an association. Here, if there are many resources with both tag A and tag B, we will connect the two in network graph.

Different tags are concurrently used to describe a resource for many reasons, some are even very arbitrary. However, although each tagging behavior seems somehow irregular, a high frequency of co-occurrences is not coincidental; rather, it exposes the semantic aspects underlining collaborative tagging, such as homonymy, synonymy, hierarchical relations among tags and so on. With this idea, these connected tags form a network upon which we can measure their semantic similarity.

The concrete methods of forming network are variable. Here we used 3 approaches. The reason we use multiple methods to form the network is to eliminate the impact of network formation while measuring semantic relatedness between tags.

1. Link weight

The first approach is straightforward and simple. Two tags t_1 and t_2 are linked if they have been associated with the same resource. To avoid noise, a link weight can be introduced by defining the weight of the link between t_1 and t_2 as the number of posts where they appear together. We calculated the linked weight for any two tags, which resulted in an $n \times n$ matrix, where n is the count of all tags. Tag pairs with low linked weight (less than threshold ω , which is set to 10 as an empirical value) were removed from the matrix and the remaining high linked weight pairs were retained.

So the folksonomy network will be formed by connecting tags t_1 and t_2, if:

$$Y' = \{U' \times \{t_1 t_2 \cup T'\} \times R'\}$$

and:

$$|Y'| > \omega$$

where $U' \subseteq U, T' \subseteq T, R' \subseteq R$.

With this method, we connect tags by co-occurrence and thus form a network of folksonomy.

2. Subsumption

Sanderson and Croft [58] described a simple statistical model for subsumption in which X subsumes Y if:

$$P(x|y >= 0.8) \,\&\&\, P(y|x < 1).$$

Here we will adopt this method to identify the relationships between tags. Specifically, tags t_1 and t_2 are connected if there exists Y' that:

$$Y' = \{U' \times \{t_1 t_2 \cup T'\} \times R'\}$$

and for t_1 and t_2:

$$P(t_1|t_2 >= 0.8) \,\&\&\, P(t_2|t_1 < 1)$$

or:

$$P(t_2|t_1 >= 0.8) \,\&\&\, P(t_1|t_2 < 1).$$

With this method, we connect tags with co-occurrence by statistical model for subsumption.

3. MI scores

We use an index of mutual information (MI) to find pairs of tags that frequently co-occurred. The mutual information between any two tags is calculated based on the normalized co-occurrence between two tags. The well-established formula of computing MI score by Shannon [60] is shown below:

$$MI(t_1, t_2) = P(t_1, t_2) \log \frac{P(t_1, t_2)}{P(t_1)P(t_2)}.$$

We calculate the MI scores for any two tags, which resulted in an $n \times n$ matrix, where n is the count of all tags. Tag pairs with low MI scores are removed from the matrix and the remaining high MI score pairs are retained.

3.2.2 Resource Context

Other than by co-occurrence, we may also connect tags according to their context of resources. The resources could be tagged documents or other sources like Wikipedia texts. Then the network of folksonomy is automatically generated as an undirected graph $G = (V, E)$ where each vertex v in V is a tag and $D(v)$ is a set of documents tagged as v or other resources describing v and each edge e_{ij} in E is the similarity between $D(v_i)$ and $D(v_j)$. Here we adopt two approaches to form such networks.

1. Context from Wikipedia

Wikipedia provides a very large domain-independent encyclopedic repository. Its texts contain information of specific tags. Here we use the method utilized in [33], a statistical model for deriving subsumption relationships between tags by computing the frequency of each tag on Wikipedia texts. The model adopted the basic idea from [58] and made a slight modification to the original model. It is defined as follows, for two tags, x and y, x subsumes y if:

$$TF(y|Wiki(x)) < TF(x|Wiki(y))$$

and:

$$\mu < TF(x|Wiki(y))$$

where $Wiki(a)$ is the Wikipedia text that tag a is mapped to, $TF(b|Wiki(a))$ is the term frequency of tag b on the $Wiki(a)$, and μ is the threshold value that is determined empirically. In other words, tag x subsumes tag y if 1) x is more frequent on the Wikipedia text of y than y is on the Wikipedia text of x, and 2) x occurs on the Wikipedia text of y to some degree. Empirically, the quality was best when μ was 0.01.

In this way, we can connect tags t_1 and t_2 if t_1 subsumes t_2 or t_2 subsumes t_1, and thus form a folksonomy network.

2. Context from document

We could also compute the tag context from document. Here we adopted residual document frequency (RDF) presented in [14]. We use RDF to measure the difference between document frequency of tag t_1 within the documents tagged t_2 (df_{t_1}, t_2) and document frequency predicted by assuming a Poisson distribution:

$$rdf_{t_1,t_2} = df_{t_1,t_2} - df_{t_2}(1 - e^{-\frac{f_{t_1}}{n}})$$

where df_{t_1} is document frequency tagged t_2, f_{t_1} is the total frequency of term t_1, and n is the total number of documents. In this way, we can form network of folksonomy, generating edges between tags with higher rdf than threshold.

Till now, we have setup networks of folksonomy with different methods. In the next part, we will present our measures of semantic relatedness, which will be applied on these folksonomy networks.

4 SEMANTIC RELATEDNESS MEASUREMENT

We give a set of methods to measure semantic relatedness between tags. We view the semantic relatedness between tags as the semantic relatedness between words, which has been studied for a long history. Specifically, we pick the semantic similarity as our target properties and measure the semantic distance between words. Semantic similarity has some differences from semantic relatedness because semantic relatedness includes concepts as antonymy and meronymy. However, in essence, semantic similarity and semantic relatedness all mean "How much does term A have to do with term B?".

There have been many proposals estimating semantic distance. The result distance is usually a number, between 0 and 1, where 1 signifies extremely high similarity/relatedness, and 0 signifies little-to-none. These methods could be intrinsic or extrinsic. Intrinsic metrics employ no external evidence, i.e. no knowledge sources except for the conceptual network itself [32, 37, 59]. Extrinsic metrics require additional knowledge, e.g. information content values of concepts computed from corpora [29, 39, 54]. The type of knowledge source employed also varies: metrics

can either employ a machine readable dictionary, i.e. textual definitions of words therein as an underlying knowledge base [37, 49], or operate on the structure of a conceptual network [29, 59]. Different way of measures will influence the result. So here we adopt multiple methods.

4.1 Measure Based on Roget's Thesaurus

Our first method belongs to path finding measures, using Roget's thesaurus as lexical resource. Roget's thesaurus [55], containing around 250,000 words, is based on a well-constructed concept classification. Concepts are organized in a hierarchy, where more general concepts are near the root of the hierarchy, and more specific ones near at the leaves.

It is convenient to measure similarity in Roget's thesaurus according to the path lengths between concepts. The measure is the straightforward edge counting method of Rada et al. [52], which defines semantic distance as the number of nodes in the taxonomy along the shortest path between two conceptual nodes. So the distance of two terms in Roget's thesaurus equals to the number of edges in the shortest path: the words from the same semicolon group have the shortest distance of 0; and the longest distance is 16, as shown in Table 1.

Distance	Relationship	Example
0	same semicolon group	journey's end – terminus
2	same paragraph	devotion – abnormal affection
4	same part of speech	popular misconception – glaring error
6	same head	individual – lonely
8	same head group	finance – apply for a loan
10	same sub-section	life expectancy – herbalize
12	same section	love – inspired
14	same class	book – blind eye
16	in the Thesaurus	nag – lightning

Table 1. Word distance in Roget's thesaurus

Then the similarity of two concepts t_1 and t_2 is defined as:

$$sim_{path}(t_1, t_2) = \frac{1}{Dis(t_1, t_2)}$$

where $Dis(t_1, t_2)$ is the distance of two tags in Roget's thesaurus.

We use the system of Open Roget[5], which is based on Roget's thesaurus, to evaluate the distance between words. The performance of Roget Thesaurus based approach is better than WordNet-based similarity measures according to [27].

[5] `http://rogets.site.uottawa.ca/`

4.2 Measure Based on WordNet

WordNet is a lexical database of concepts and relations, where each unique meaning of a word is represented by a synset. Synsets are connected to each other through explicit semantic relations that are defined in WordNet (synonymy, antonymy, is-a, part-of, etc.). This creates a network where related concepts can be identified by their relative distance from each other. Due to its increasing scope and free availability, WordNet has become a popular resource for identifying taxonomic and networked relationships among concepts.

For nouns, the most common and useful relation is the "is-a" relation. These comprise over 70 % of the total relations for nouns. Leacock and Chodorow [32] proposed a measure based on the shortest paths between noun concepts in an is-a hierarchy in WordNet. According to this method, the similarity of tags t_1 and t_2 is computed as:

$$\mathrm{sim}_{lch}(t_1, t_2) = -\log \frac{\mathrm{length}(t_1, t_2)}{2\mathrm{D}}$$

where $length(t_1, t_2)$ is the number of nodes along the shortest path between the two nodes in WordNet. D is the maximum depth of the taxonomy, which is a constant of 16 for all noun concepts.

4.3 Measure Based on Wikipedia

Wikipedia provides a semantic network for computing semantic relatedness with more coverage than WordNet. It contains entries on a vast number of named entities and very specialized concepts. Articles can be assigned to one or more categories, which are further categorized to provide a category tree.

Here we use the method of WikiRelate!, developed by Strube and Ponzetto [65] for computing semantic relatedness. This method has been proven to correlate well with human judgments. Given a pair of words, WikiRelate! searches for Wikipedia articles, that respectively contain the pair of words in their titles. Semantic relatedness is then computed using distance measures of the articles.

Specifically, the method measures relatedness between two words with a function of text (i.e. gloss) overlap. Given two texts t_1 and t_2 taken as definitions for the words w_1 and w_2, the overlap score overlap(t_1, t_2) is computed as: $\sum_n m^2$ for n phrasal m-word overlaps.

The relatedness score is given by applying a double normalization step to the overlap score. We first normalize by the sum of text lengths and then take the output as the value of the hyperbolic tangent function in order to minimize the role of outliers skewing the score distribution:

$$sim(t_1, t_2) = tanh(\frac{overlap(t_1, t_2)}{length(t_1) + length(t_2)}).$$

4.4 MSRs

We also conduct the measurement with a semantic relatedness toolset called MSRs[6]. MSRs provide a set of computational means for calculating the association strength between terms. MSRs have been used in many applications such as augmented search engine technology, semantic relevancy maps, and essay-grading algorithms for ETS, etc.

There are many different API provided by MSRs, we used the following ones: PMI-G, PMI-Gwikipedia, NSS-G, and NSS-Gwikipedia, where:

- PMI (Pointwise mutual information [65]) quantifies the discrepancy between the probability of their coincidence given their joint distribution versus the probability of their coincidence given only their individual distributions and assuming independence.
- NSS (Normalized Search Similarity) is proposed by Cilibrasi [10] and adapted from Normalized Google Distance.
- Measures that end with "-G" use the Google search engine, and "-Gwikipedia" searches only *.wikipedia.org.

We will apply all these measures on networks of folksonomy to estimate semantic relatedness among tags. The last thing to mention is that the performances of these measures are different, according to [15], as shown in Table 2:

Measure	Correlation with manual judgment
WordNet	0.33–0.35
Roget's Thesaurus	0.55
WikiRelate!	0.19–0.48

Table 2. Performance of semantic relatedness measures

5 EXPERIMENT AND RESULT

In this part, we check the semantic relationships between connected tags, by applying different measures on folksonomy network. We first introduce our data source and method to collect data. Then we conduct the experiment and show its results.

5.1 Dataset

We pick del.icio.us as our data source and setup the dataset. Del.icio.us provides user with a viable and effective mechanism to organize Web resources via bookmark; and it is said to be the true implementation of collaborative tagging. Tags are used to describe the bookmarks posted by users.

[6] `http://cwl-projects.cogsci.rpi.edu/msr/`

Del.icio.us is an ideal choice for us because it has accumulated plenty data of tags and resources. We collect publicly accessible data from del.icio.us to setup the dataset. Our crawling algorithm uses breadth-first search. From Oct 13 to 28, 2008, we used crawler to download bookmarks pages in a recursive way, and obtained the result dataset of folksonomy with $|U| = 21\,620$ users, $|T| = 312\,573$ tags and $|R| = 1\,873\,035$ resources, related by in total $|Y| = 9\,476\,904$ tag assignments. Tags appearing less than 5 times in the tag set were deleted in order to reduce computation cost.

The breadth-first method is efficient but has the risk to leave those isolated nodes out [34]. So we verified the result to make sure what we have obtained covers most part of tags' graph. We randomly chose a set of users, collected all their tags (340 users, 3548 tags), and found that we had already crawled 97.8 % of tags, showing our method covering the vast majority of the tags population.

5.2 Result

With the dataset, we formed 5 folksonomy networks described in Section 3, and then measured the semantic relatedness of directly connected tag pairs, with methods described in Section 4. Table 3 presents the result. All results are normalized between 0 and 1 for comparison, where 1 signifies extremely high similarity/relatedness, and 0 signifies little-to-none. "Avg" represents the average value of semantic relatedness of connected tags, "s.d." represents the standard deviation of semantic relatedness of connected tags, and "Random" represents the average value of semantic relatedness of random pairs of tags.

	Roget's thesaurus			*WordNet*			*WikiRelate!*		
	Avg	s.d.	Random	Avg	s.d.	Random	Avg	s.d.	Random
Folksonomy network by co-occurrence									
Link Weight	0.37	0.11	0.18	0.26	0.09	0.12	0.3	0.08	0.2
Subsumption	0.43	0.09	0.13	0.32	0.06	0.17	0.36	0.09	0.13
MI Score	0.41	0.13	0.14	0.35	0.09	0.13	0.31	0.16	0.15
Folksonomy network by context									
Wikipedia Context	0.48	0.08	0.21	0.52	0.11	0.18	0.67	0.09	0.23
Document Context	0.51	0.11	0.18	0.47	0.08	0.13	0.49	0.11	0.16

Table 3. Semantic relatedness in folksonomy network

We also conducted the measurement with MSRs. The result is shown in Table 4.

	PMI-G		*PMI-Gwikipedia*		*NSS-G*		*NSS-Gwikipedia*	
	Avg	Random	Avg	Random	Avg	Random	Avg	Random
Folksonomy network by co-occurrence								
Link Weight	0.563	0.320	0.511	0.093	0.636	0.383	0.490	0.273
Subsumption	0.604	0.389	0.529	0.089	0.673	0.468	0.461	0.302
MI Score	0.625	0.275	0.561	0.106	0.738	0.453	0.472	0.263
Folksonomy network by context								
Wikipedia Context	0.694	0.384	0.796	0.182	0.820	0.395	0.687	0.319
Document Context	0.722	0.258	0.641	0.147	0.754	0.387	0.614	0.329

Table 4. Semantic relatedness with MSRs

We could find the semantic similarity of connected tags is much stronger than random picked tag pairs, regardless of what folksonomy we use and what measures we adopt. As a result, we can get the conclusion that connection of tags contains semantic information. Another finding from the experiment result is that the tag relatedness in folksonomy networks formed by resource context is higher than that formed by co-occurrence. That is reasonable, for the connection formed with resource context contains more information about resources and thus better reflects tags meaning. However, the difference is not so notable, which shows that users' collaborative tagging can well reflect tag's usage in describing the resource.

We also measure the relatedness between tags that are not directly connected, to investigate the change of semantic relatedness. The result is shown in Table 5. D represents the shortest distance between two tags. For example, D = 2 means the shortest distance of two tags is 2, and they need to be connected by another tag as intermediate.

	Roget's thesaurus			*WordNet*			*WikiRelate!*		
	D = 1	D = 2	D = 3	D = 1	D = 2	D = 3	D = 1	D = 2	D = 3
Folksonomy network by co-occurrence									
Link Weight	0.37	0.19	0.12	0.26	0.12	0.06	0.30	0.19	0.17
Subsumption	0.43	0.21	0.09	0.32	0.13	0.09	0.36	0.13	0.11
MI Score	0.41	0.17	0.11	0.35	0.18	0.13	0.31	0.16	0.14
Folksonomy network by context									
Wikipedia Context	0.48	0.24	0.06	0.52	0.13	0.04	0.67	0.21	0.07
Document Context	0.51	0.18	0.13	0.47	0.11	0.07	0.49	0.18	0.09

Table 5. Average semantic relatedness between tags with different distance

We could find that semantic relatedness decreases dramatically as the distance between tags increases. The phenomenon is coincidental with small world property of folksonomy [7]. Because the diameter of folksonomy is small, an increase of distance will include a lot of unrelated tags into vision. From the other side, it could be one of the reasons of small world property of folksonomy. Plenty of connections exist among tags and semantic related tags are clustered; so short average path length and high clustering coefficient could be produced.

6 DISCUSSION

From the experiment result, we could see that there exists semantic relatedness among tags; they are even linked by co-occurrence without using semantic implication. The result confirms our idea of users' usage of folksonomy as a semantic map of concept.

The choice of folksonomy networks influences the semantic relatedness among tags within it. The resource context approaches with additional information provide more accurate semantic information than co-occurrence approaches. However, the difference is not so obvious. In application, the resource context approaches need more computing efforts because they need to retrieve and process related resources, such as article in Wikipedia or annotated documents. As a result, the methods

of co-occurrence might be more suitable for tagging systems with huge dynamic contents.

Besides, in co-occurrence methods, MI score and statistical model for subsumption have more accurate semantic information than linked weight, while the performance of linked weight is most efficient. In resource context methods, the RDF method needs more computing than the Wikipedia way, but provides similar accurateness of semantic information.

The correlation with manual judgment of different measures has been shown in Table 2. Unlike this, these measures have their own properties. Roget's thesaurus is simple and the algorithm is efficient. However, the thesaurus has been created in 1805, so much of the vocabulary of tags could not be found in it. The WordNet method contains much more vocabulary but it also cannot cover all folksonomy. The strength of Wikipedia lies in its size, which could be used to overcome current knowledge bases' limited coverage and scalability issues. Besides, the content is Wikipedia is dynamic with users' edit work, so it is very suitable to accompany the evolution of folksonomy. However, the algorithm of Wikipedia measure needs more computing cost than the others.

The research described in this paper is motivated by the larger goal of improving the way in which we locate information. Our finding, along with the folksonomy network, has a wide range of applications in knowledge organization and information retrieval.

For tagging systems, the semantic information discovered here could be used as a bridge between the query terms and the topics/terminology of the documents available; and the semantic relatedness could be used as an indicator to predict related tags in tag suggestion and other occasions. Besides, the semantic information within tag connection can be utilized in search engines, recommendation systems, etc.

For semantic web, the relatedness of folksonomy could provide valuable information for automatic generation of ontologies. The links of tags can give some insight about the semantic relatedness between words in human's cognition of resources, which is useful for both researchers and system designers. For example, the Open Directory project uses human effect to maintain the list of Web resources and their categories. With the map of tag graph, we can automatically achieve the same function with bookmarks with associated tags. Besides, it is slow for new words (such as a new product name, like "iphone") to be included in dictionary. However, they can be immediately used by users in tagging. In this way, we overcome the well known knowledge acquisition bottleneck by deriving a knowledge resource from a very large, collaboratively created data.

For NLP, computers need access to vast amounts of common-sense and domain-specific world knowledge [36]; this is essential for high-level linguistic tasks which require language understanding capabilities such as question answering [25] and recognizing textual entailment [3, 66]. However, there are not many knowledge bases available which provide a large amount of information on named entities and contain continuously updated knowledge for processing current information. The

folksonomy network could shed some light on this problem, providing large amount of semantic knowledge with dynamic evolution.

Certain limitation of this work must be mentioned. Our measures mainly focus on semantic similarity. We need to point out that actually semantic relatedness and semantic similarity are two separate notions. Semantic relatedness is a more general notion of the relatedness of concepts, while similarity is a special case of relatedness that is tied to the likeness (in the shape or form) of the concepts. Two concepts can be related without being similar. For example, two concepts may be related because they are antonyms. The measures of semantic similarity described here are based on is-a relations that link concepts (directly or indirectly), trying to return a numeric score that quantifies how much two concepts are alike. Measures of semantic relatedness are more general, and can include information about other relations, or may be based on co-occurrence statistics from corpora. However, there have been relatively few attempts to develop measures that rely on relations beyond is-a.

7 AN EXAMPLE OF APPLICATION

In this part, we will give an example of application, utilizing the folksonomy network and semantic relatedness in it. We will develop a keyword extraction method by modeling a text document as a graph of terms.

At present, keyword extraction is mainly done by statistical learning method, which needs training and does not perform well with noise. We find for a document text that the keywords are always related to the main topics and tend to interconnect with each other in semantic relationship, while the irrelevant terms for main topics tend to be isolated. In this way, we can first get all the terms in a text, and use them to construct a term network. The weight of connection between terms is computed with the folksonomy network as stated before. Then a core community detecting method would be used upon term network to find the keyword. Below are the main steps:

1. **Text extraction.** We first extract text document from a web page, and then query each term with WordNet and get all the nouns. It is reasonable to do so, because most keywords of text are nouns.
2. **Term network construction.** On the set of terms, we calculate each pair's semantic relatedness in folksonomy network we got before. Here we choose the folksonomy network formed by co-occurrence for its good performance, and use relatedness measurement based on Wikipedia for its good accurateness and wide coverage. Thus, we construct a network of terms, in which the weight of links is semantic relatedness.
3. **Core community detecting.** In this step, we try to find the densely connected terms. Many algorithms could be used to detect the core community in a network. Here we adopt the method proposed by Clauset et al. [11]. It is highly

effective at discovering core community structure of complex network. Then we could choose the terms in core community as document keyword.

As experiment, we adapted the method on a set of blogs from blogspot[7], and found it works very well and is resistant to noise. Below is a comparison of manual keywords and auto-generated keywords by our method of an example blog post. We found the accurateness of keywords extraction is relatively high.

Blog title	"Developing flash-alike gallery with jQuery"
Manual keywords	jQuery, gallery, javascript, programming, RIA, flash, css
Generated Keywords	jQuery, gallery, javascript, web, flash, photo, css

Table 6. Word distance in Roget's thesaurus

This application is just an example of the folksonomy network with semantic relatedness; there could be many other usages.

8 CONCLUSION AND FUTURE WORK

Our goal in this paper is to evaluate semantic relatedness of tags in folksonomy. Due to its core position in online social networks, tags can reflect properties of both resource space and user space; and the semantic relatedness among tags can reveal the relationship between resources, and can reflect users' tagging behavior. To achieve this, we proposed different ways (co-occurrence and resource context) of forming networks of folksonomy, with the data gathered from social bookmarking site del.icio.us. Upon this network, multiple relatedness measures utilizing Roget's thesaurus, WordNet, Wikipedia and MSRs are applied to pairs of tags with different distances. The results all show that the semantic relatedness is relatively strong within tag connections, which means the networks of folksonomy containing semantic information. We believe this information can give some inspiration in designing future tagging-based systems, constructing semantic web in current state of the Internet, and developing natural language processing applications. For example, the folksonomy network could be used in extracting keywords from text with network analysis methods.

Much work still remains. We expect to enhance our work presented here in a number of respects. First of all, our measurement only considered the semantic similarity, which was a special case of semantic relatedness: for example, cars and gasoline would seem to be more closely related than cars and bicycles, but the latter pair is certainly more similar. Although the similarity is the most used relatedness between tags, some computational linguistics applications often require measures of relatedness rather than the more narrowly defined measures of similarity [5]. So we believe that developing measures that includes relations beyond similarity is an important content of future work. Second, in our work, although we identify

[7] `http://www.blogspot.com`

that semantic relatedness are high in folksonomy network and connection among tags contains semantic information, further investigation into the meaning of this relatedness and their properties should be done. Thirdly, there are ongoing efforts by us to find better algorithm for forming folksonomy network, with high performance and high semantic accuracy.

REFERENCES

[1] Agrawal, R. et al.: Mining Association Rules Between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Washington, D.C., United States 1993.

[2] Agrawal, R.—Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20^{th} International Conference on Very Large Data Bases, Morgan Kaufmann Publishers 1994.

[3] Bos, J.—Markert, K.: Recognising Textual Entailment With Logical Inference. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada 2005.

[4] Brooks, C. H.—Montanez, N.: Improved Annotation of the Blogosphere Via Autotagging and Hierarchical Clustering. In Proceedings of the 15^{th} International Conference on World Wide Web, ACM: Edinburgh, Scotland 2006.

[5] Budanitsky, A.—Hirst, G.: Evaluating WordNet-Based Measures of Lexical Semantic Relatedness. Comput. Linguist., Vol. 32, 2006, No. 1, pp. 13–47.

[6] Carnine, D.—Kameenui, E.—Coyle, G.: Utilization of Contextual Information in Determining the Meaning of Unfamiliar Words. Reading Research Quarterly, Vol. 19, 1984, pp. 188–204.

[7] Cattuto, C. et al.: Network Properties of Folksonomies. AI Comm., Vol. 20, 2007, No. 4, pp. 245–262.

[8] Caviedes, J. E.—Cimino, J. J.: Towards the Development of a Conceptual Distance Metric for the UMLS. J. of Biomedical Informatics, Vol. 37, 2004, No. 2, pp. 77–85.

[9] Chirita, P.—A. et al.: P-TAG: Large Scale Automatic Generation of Personalized Annotation Tags for the Web. In Proceedings of the 16^{th} International Conference on World Wide Web, ACM: Banff, Alberta, Canada 2007.

[10] Cilibrasi, R.—Vitanyi, P.: Similarity of Objects and the Meaning of Words. In Theory and Applications of Models of Computation. 2006, pp. 21–45.

[11] Clauset, A.—Newman, M. E. J.—Moore, C.: Finding Community Structure in Very Large Networks. Physical Review E, 70:066111, 2004.

[12] Dubinko, M. et al.: Visualizing Tags over Time. In Proceedings of the 15^{th} International Conference on World Wide Web, ACM: Edinburgh, Scotland 2006.

[13] Finkelstein, L. et al.: Placing Search in Context: The Concept Revisited. ACM Trans. Inf. Syst., Vol. 20, 2002, No. 1, pp. 116–131.

[14] Fujimura, S.—Fujimura, K.—Okuda, H.: Blogosonomy: Autotagging Any Text Using Bloggers' Knowledge. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society 2007.

[15] Gabrilovich, E.—Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India 2007, pp. 1606–1611.

[16] Ganesan, K. A.—Sundaresan, N.—Deo, H.: Mining Tag Clouds and Emoticons Behind Community Feedback. In Proceeding of the 17th International Conference on World Wide Web, ACM, Beijing, China 2008.

[17] Golder, S. A.–Huberman, B. A.: The Structure of Collaborative Tagging Systems. HP Labs technical report 2005.

[18] Golder, S. A.—Huberman, B. A.: Usage Patterns of Collaborative Tagging Systems. J. Inf. Sci., Vol. 32, 2006, No. 2, pp. 198–208.

[19] Gruber, T.: Ontology of Folksonomy: A Mash-Up of Apples and Oranges. International Journal on Semantic Web and Information Systems, Vol. 3, 2007, No. 1, pp. 1–11.

[20] Jiafeng, G. et al.: Exploring Collaboratively Annotated Data for Automatic Annotation. In Proceeding of CAW2.0 2009, ACM, Madrid, Spain.

[21] Hammond, T. et al.: Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 2005, Vol. 11, 2005, No. 4.

[22] Hassan-Montero, Y.—Herrero-Solana, V.: Improving Tag-Clouds as Visual Information Retrieval Interfaces. In I International Conference on Multidisciplinary Information Sciences and Technologies, InSciT 2006, Mrida, Spain.

[23] Hotho, A. et al.: Information Retrieval in Folksonomies: Search and Ranking. In The Semantic Web: Research and Applications, 2006, pp. 411–426.

[24] Hotho, A. et al.: Trend Detection in Folksonomies. In Semantic Multimedia 2006, pp. 56–70.

[25] Hovy, E. et al.: Question Answering in Webclopedia. In Proceedings of the Thirteenth Text Retrieval Conference 2001, Gaithersburg, Md.

[26] Hughes, T.—Ramage, D.: Lexical Semantic Relatedness with Random Graph Walks. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007), Association for Computational Linguistics: Prague, Czech Republic.

[27] Jarmasz, M.—Szpakowicz, S.: Roget's Thesaurus and Semantic Similarity. In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003), Borovets, Bulgaria, pp. 212–219.

[28] Jian, Q.: Controlled Semantics vs. Social Semantics: An Epistemological Analysis. In Proceedings of the 10th International ISKO Conference: Culture and Identity in Knowledge Organization, Montreal 2008.

[29] Jiang, J.—Conrath, D.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan 1997.

[30] Kim, S. N.—Baldwin, T.: Automatic Interpretation of Noun Compounds Using Wordnet Similarity. In Natural Language Processing – IJCNLP 2005, pp. 945–956.

[31] Kozima, H.—Furugori, T.: Similarity Between Words Computed by Spreading Activation on an English Dictionary. In Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics 1993, Association for Computational Linguistics: Utrecht, The Netherlands.

[32] Leacock, C.—Chodorow, M.: Combining Local Context and Wordnet Similarity for Word Sense Identification. WordNet: An electronic lexical database, C. Fellbaum (Ed.), MIT Press 1998.

[33] Lee, K. et al.: Folksoviz: a Subsumption-Based Folksonomy Visualization Using Wikipedia Texts. In Proceeding of the 17th International Conference on World Wide Web 2008, ACM, Beijing, China.

[34] Lee, S. H.–Kim, P. J.—Jeong, H.: Statistical Properties of Sampled Networks. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), Vol. 73, 2006, No. 1, pp. 016102-7.

[35] Lee, S. S.—Won, D.—McLeod, D.: Tag-Geotag Correlation in Social Networks. In Proceeding of the 2008 ACM Workshop on Search in Social Media, 2008, ACM, Napa Valley, California, USA.

[36] Lenat, D.—Guha, R.: Building Large Knowledge Based Systems. Addison Wesley 1990.

[37] Lesk, M.: Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proceedings of the 5th Annual International Conference on SystemsDdocumentation, 1986, ACM, Toronto, Ontario, Canada.

[38] Li, R. et al.: Towards Effective Browsing of Large Scale Social Annotations. In Proceedings of the 16th International Conference on World Wide We, 2007, ACM, Banff, Alberta, Canada.

[39] Lin, D.: An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. 1998.

[40] Mathes, A.: Folksonomies – Cooperative Classification and Communication Through Shared Metadata. In Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign 2004.

[41] McDonald, S.—Ramscar, M.: Testing the Distributional Hypothesis: the Influence of Context on Judgements of Semantic Similarity. In Proceedings of the 23rd Annual Conference of the Cognitive Science Society, Edinburgh, Scotland 2001, pp. 611–616.

[42] Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Proceedings of the 4th International Semantic Web Conference (ISWC 2005), LNCS 3729, Springer-Verlag 2005.

[43] Millen, D. R.—Feinberg, J.: Using Social Tagging to Improve Social Navigation. In Workshop on the Social Navigation and Community-Based Adaptation Technologies at AH 2006, Dublin, Ireland, pp. 532–541.

[44] Miller, G.—Charles, W.: Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, Vol. 6, 1991, No. 1, pp. 1–28.

[45] Milne, D.: Computing Semantic Relatedness Using Wikipedia Link Structure. In Proceedings of the New Zealand Computer Science Research Student Conference (NZC-SRSC '07), Hamilton, New Zealand 2007.

[46] Nickles, M.—Froehner, T.—Weiss, G.: Social Annotation of Semantically Heterogeneous Knowledge. In Proceedings of the 4th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot) 2004.

[47] Niwa, Y.—Nitta, Y.: Co-Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries. In Proceedings of the 15th Conference on Computational Linguistics, Volume 1, Association for Computational Linguistics, Kyoto, Japan 1994.

[48] Patwardhan, S.–Banerjee, S.—Pedersen, T.: Senserelate::Targetword: A Generalized Framework for Word Sense Disambiguation. In Proceedings of the ACL 2005 on Interactive poster and demonstration sessions, Association for Computational Linguistics, Ann Arbor, Michigan 2005.

[49] Patwardhan, S.–Banerjee, S.—Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In Computational Linguistics and Intelligent Text Processing 2008, pp. 241–257.

[50] Patwardhan, S.—Pedersen, T.: Using WordNet-Based Context Vectors to Estimate the Semantic Relatedness of Concepts. In Proceedings of the EACL 2006 workshop, Making sense of sense: Bringing computational linguistics and psycholinguistics together, Trento, Italy 2006, pp. 1–8.

[51] Quillian, M.: Semantic Memory. Semantic Information Processing, M. Minsky (Ed.), The MIT Press, Cambridge, MA 1968.

[52] Rada, R. et al.: Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics, Vol. 19, 1989, No. 1, pp. 17–30.

[53] Raina, R.—Ng, A.—Manning, C.: Robust Textual Inference Via Learning and Abductive Reasoning. In Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, PA 2005.

[54] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal 1995.

[55] Roget, P.: Roget's Thesaurus of English Words and Phrases. Longman Group Ltd., 1982.

[56] Rubenstein, H.—Goodenough, J. B.: Contextual Correlates of Synonymy. Commun. ACM, Vol. 8, 1965, No. 10, pp. 627–633.

[57] Sahami, M.—Heilman, T. D.: A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets. In Proceedings of the 15th International Conference on World Wide Web, ACM, Edinburgh, Scotland 2006.

[58] Sanderson, M.—Croft, B.: Deriving Concept Hierarchies from Text. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Berkeley, California, United States 1999.

[59] Seco, N.–Veale, T.—Hayes, J.: An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain 2004, pp. 1089–1090.

[60] Shannon—Claude, E.: The Mathematical Theory of Communication. Bell System Technology Journal, Vol. 27, 1948, pp. 379–423.

[61] Shaw, B.: Semidefinite Embedding Applied to Visualizing Folksonomies. 2005.

[62] Spasic, I.—Ananiadou, S.: A Flexible Measure of Contextual Similarity for Biomedical Terms. In Pacific Biocomputing Symposium 2005, pp. 197–208.

[63] Specia, L.—Motta, E.: Integrating Folksonomies with the Semantic Web. In The Semantic Web: Research and Applications, 2007, pp. 624–639.

[64] Stevenson, M.—Greenwood, M. A.: A Semantic Approach to IE Pattern Induction. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Ann Arbor, Michigan 2005.

[65] Strube, M.—Ponzetto, S. P.: Wikirelate! Computing Semantic Relatedness Using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence 2006, pp. 1419–1424.

[66] Tatu, M. et al.: COGEX at the Second Recognizing Textual Entailment Challenge. In Proceedings of the Second PASCAL Recognising Textual Entailment Challenge Workshop, Venice, Italy 2006, pp. 104–109.

[67] Turney, P. D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the 12th European Conference on Machine Learning., Springer-Verlag 2001.

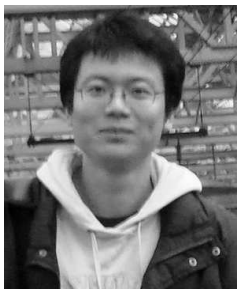
[68] V. L.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, Vol. 10, 1966, p. 707.

[69] Wilks, Y. et al.: Providing Machine Tractable Dictionary Tools. Machine Translation, Vol. 5, 1990, No. 2, pp. 99–154.

[70] Wu, X.—Zhang, L.—Yu, Y.: Exploring Social Annotations for the Semantic Web. In Proceedings of the 15th International Conference on World Wide Web, ACM, Edinburgh, Scotland 2006.

[71] Wu, Z.—Palmer, M.: Verbs Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, Las Cruces, New Mexico 1994.

[72] Xu, Z.et al.: Towards the Semantic Web: Collaborative Tag Suggestions. In Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference 2006.



Chao Wu is a research associate at Department of Computer Science, Zhejiang University. Currently he is visiting Imperial College London as postdoc researcher. He received his Doctor degree from Zhejiang University. His research interests include social networking, cloud computing, folksonomy system, data visualization, etc.

Bo Zhou is a Professor in Department of Computer Science, Zhejiang University. His research is in the areas of database management, CIMS, data mining, data warehouse, social networking, etc. He graduated in Computer Science from Zhejiang University and has a PhD in Computer Science at Zhejiang University.