Applied Mathematics & Information Sciences
*An International Journal*



# Automatic Clustering of Social Tag using Community Detection

*Weisen Pan, Shizhan Chen* and Zhiyong Feng*

School of Computer Science and Technology, Tianjin University, Tianjin 300072 P.R. China

Received: 17 Oct. 2012; Revised 21 Nov. 2012; Accepted 1 Dec. 2012
Published online: 1 Mar. 2013

**Abstract:** Automatically clustering social tags into semantic communities would greatly boost the ability of Web services search engines to retrieve the most relevant ones at the same time improve the accuracy of tag-based service recommendation. In this paper, we first investigate the different collaborative intention between co-occurring tags in Seekda as well as their dynamical aspects. Inspired by the relationships between co-occurring tags, we designed the social tag network. By analyzing the networks constructed, we show that the social tag network have scale free properties. In order to identify densely connected semantic communities, we then introduce a novel graph-based clustering algorithm for weighted networks based on the concept of edge betweenness with high enough intensity. Finally, experimental results on real world datasets show that our algorithm can effectively discovers the semantic communities and the resulting tag communities correspond to meaningful topic domains.

**Keywords:** Social tag, web service, semantic communities, scale free, community detection.

## 1. Introduction

In recent years Service Oriented Architectures (SOA) [1] has become an emerging and promising approach for supporting the rapid development of low-cost, interoperable, and evolvable distributed applications. Research activities have focused on different challenges [2,3] about SOA. One of essential challenges is how to find the desired Web services for user [4]. However, it is becoming more difficult and time consuming task with dramatically growing of Web services on the internet.

Collaborative tagging provides a convenient way to annotate shared content by allowing users to use any tag or keyword. Social tags provide meaningful descriptions of items, and allow users to organize and index their contents. User social tags not only have proven to be a useful when browsing large collections of documents [5], but also useful to have a quick understanding of a particular service and service classification. Recently, some web service search engines, such as Seekda [6], allow users to manually annotate web services using tags. Fig.1.1 shows two examples of collaboratively tagged Web services in Seekda. WSRoomSearchService in Fig. 1.1 (a) is a Web service which provides the function of room search. It has three tags, tourism, room_search_service, and internal. The tourism tag shows which topic domain of Web service belongs to. The room_search_service tag describes the function of this Web service. The internal tag may tell users that service requester cannot invoke this Web service for free, it is an internal state. Fig. 1.1 (b) shows another Web service providing predator information, which is very important for bioinformatics. If we utilize the tag bioinformatics in the search engine, this service will be included in the search result about bioinformatics. From these two examples, we can find that the tagging data can help to retrieve more relevant web services, and these tags between co-occurring represent quite a few different aspects of the Web services.


Figure 1.1: Example of collaboratively tagged Web services in Seekda

* Corresponding author: e-mail: shizhan@tju.edu.cn

One of main advantage of social tag is that they are very easy to create, where users do not need any constraint or experience. However, this issue implies a number of limitations on the Web service discovery and recommendation mechanisms. Users have different intentions when tagging. Tags not only describe the functions of Web services, but also express additional contextual and semantical information, for example by providing the organization information about who created the Web services or the topic domain (e.g., travel, biology and business) in which a Web service belongs to. Furthermore, tags may depict the qualities of Web services such as free, real time and internal etc. Current tag-based discovery and recommendation engines do not take into account the above distinction of tags, and run their content retrieval algorithms in the entire tag space. The problem is that some tags type may be useful for some particular users when searching, but they may not bring any benefit to others. On the other hand, although useful qualities tags are for the purposes of an individual, still they may fail to be of benefit when recommending Web services to other users. These findings are also supported by previous research Ref. [7–9]. Therefore, automatically clustering tags into semantic communities, like travel-related, business-related and biology-related etc., will greatly boost the ability of Web services search engines to retrieve the most relevant ones and at the same time improve the accuracy of tag-based service recommendation. Unfortunately, the importance of tag clustering according to different user intention is largely ignored by existing tagging systems.

The goal of our work is to automatically clustering social tags through community detection. The contributions of our paper are as follows:

1) Exploiting the semantic relationships between co-occurring tags, we have designed the social tag network. Network analysis results show that social tag network have the properties of scale free.
2) We have proposed a novel graph-based clustering algorithm for weighted networks based on the concept of edge betweenness with high enough intensity.
3) We have conducted an empirical study to evaluate the effect of our algorithm. The experiments have been performed with a real world datasets obtained from Seekda. The experimental results show that our algorithm can effectively discovers the semantic communities and the resulting tag communities correspond to meaningful topic domains, which can be beneficial to improve the accuracy of service discovery and recommendation.

The remainder of the paper is organized as follows. Section 2 describes the social tag network. Section 3 introduces our proposed graph-based clustering approach in more detail. Section 4 presents the conducted experiments, and Section 5 provides a discussion of obtained results. Section 6 is involved in related works. Finally, in Section 7 we conclude and present ideas for future work.

## 2. Social Tag Network

In social tagging systems, users usually have different intentions when tagging. Therefore, social tags may describe quite a few different aspects of the item. From the Fig.1.1, we can see that different type tags that annotate same Web service may have internal semantic relations, for example the taxonomic relation that bioinformatics tag describes the taxonomy domain of predator tag. The design of social tag network is inspired by above-mentioned internal semantic relations.

Social tag network is a cross-linked social graph, or a bipartite graphs (a network with two classes of vertices). These cross-linked social graphs model the associations between co-occurring tags, tags and Web services. In order to model network of social tag at an abstract level, we will represent such system as bipartite graphs with edges.

Definition 1. A social tag network is defined as bipartite graph $BG=< T, WS, E >$, where $T$ and $WS$ represent different nodes set. In this paper, node can be classified as tags shaped like an ball and Web services like a ellipse in Fig. 2.1. $T$ is a set of n tag nodes $T=\{tag_1, tag_2, ..., tag_n\}$. $WS$ is a set containing $m$ Web service nodes $WS=\{ws_1, ws_2, ..., ws_m\}$. These two sets are linked by an involvement relation $E \in T \times WS$, it contains two type relations: one between co-occurring social tags for annotate the same Web service like solid lines, the other between tags and Web services like dashed lines in Fig. 2.1.

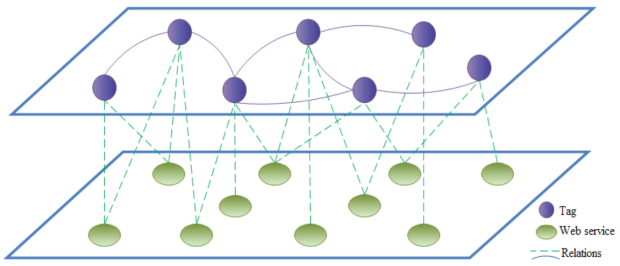


Figure 2.1: Schematic diagram of social tag network

Bipartite graphs are rather cumbersome to work with. In the paper, we are more concerned with the associations between co-occurring tags, which can help us achieve the goal automatically clustering social tags. Thus, we can reduce such a graph into one-mode graphs with regular edges. These one-mode graphs model the associations between co-occurring tags. Tag co-occurrence is measure of tag relatedness that can be measured when two tags are used to annotate the same Web service regardless of the annotator.

Definition 2. Social tag network is reducing as one-mode weighted graph $Graph=< V, E, W >$, where $V$ is a set of tags. $E$ represents a set of edges. An edge exists between two tags if the users annotate the same resource use two different tags. $W$ is a weight matrix. The weight $w_{ij}$ is equal to the number of times $tag_i$ occurred together with $tag_j$ within the same resource.

The resulting social tag network has power law distribution in its degree, as shown in Section.4.2, so their clus-

tering structure is scale free and there is no typical community size. Therefore, in order to discovery communities, we partition the graph using a nonlocal process exploiting the concept of edge betweenness.

## 3. Community-based Social Tag Clustering

After the social tag network generation process, clustering is performed to identify the community structure. There is no formal definition for a community of vertices within a graph. A graph can be said to have community structure if it consists of subsets of vertices, with many edges connecting vertices of the same subset, but few edges lying between subsets [10]. Finding communities within a graph is an efficient way to identify semantic communities of related tags.

In order to identify densely connected semantic communities, we partition social tag network using a nonlocal process exploiting the concept of edge betweenness. The edge betweenness is defined as the number of shortest paths connecting pairs of nodes that go through that edge [11]. Newman and Girvan [12] have been proposed an algorithm to detect community structure for unweighted networks based on the concept of edge betweenness. One of the important questions of the Newman and Girvan algorithm is only applicable to unweighted networks with high degree of accuracy, but is not suitable for weighted networks. In the Newman and Girvan method, one simply chooses the edge of highest betweenness and removes it. However, this choice is somewhat arbitrary, because edges links weighted can critically affect the structure of category community. Another question is when is good community modularity achieved for weighted networks. In order to solve above-mentioned problem, we introduce a novel graph-based clustering algorithm for weighted networks by applying concept of subgraph intensity and coherence to Newman and Girvan method.

Our modifications were necessary to make the method applicable to weighted social tag network. We introduce an extension of Newman and Girvan method that takes into account the link weights in a more delicate way by incorporating the subgraph intensity defined in Ref. [13] into the search algorithm. We use the concept of subgraph intensity to characterise how compact or tight the subgraph is. The subgraph intensity allows us to characterise the interaction patterns within communities. By denoting $V_{sg}$ the set of nodes and $E_{sg}$ the set of links in the subgraph with weights $W_{ij}$, we can express subgraph intensity as the geometric mean of its weights as

$$I(sg) = \left( \prod_{(ij)\in E_{sg}} W_{ij} \right)^{1/E_{sg}} \quad (1)$$

Where $|E_{sg}|$ is the number of links in $E_{sg}$. The subgraph intensity $I(sg)$ may turn out to be low because one of the weights is very low, or it may result from all of the weights being low. In order to distinguish between these extremes, we use the concept of subgraph coherence $Q(sg)$ and defined as the ratio of the geometric mean to the arithmetic mean of the weights

$$Q(sg) = I(sg) / \sum_{(ij)\in E_{sg}} W_{ij} \quad (2)$$

Where $Q \in [0,1]$. We choose subgraph that subgraph coherence higher than the subgraph coherence threshold $\lambda$ as a category community.

The main step of our algorithm is the computation of the edge betweenness of all the edges and then removal of those with the highest value in the network. This process is repeated until the parent network splits, producing two separate subgraph networks. For each subgraph, we calculate subgraph coherence value. If the subgraph coherence values higher than the subgraph coherence threshold $\lambda$, output it as a communities. The subgraph can be split further in the same way until they contain only one node. The clustering algorithm which we propose here is a partition algorithm. It starts with an edge with a high betweenness. Algorithm 1 provides a pseudo code version of our algorithm.

**Algorithm 1:** Tag clustering.

**Input:** $Graph =< V, E, W >$ a graph /* V set of vertices; E set of edges; W is weight matrix. */

**Output:** $C = C_1, C_2, ...C_n$ /* C is subsets of vertices (clusters) partitioning V. */

```
1:  begin:
2:  SG ← Graph ;
3:  for each subgraph SG_i ∈ SG do
4:      L ← SG_i ;  /* List L with all edges betweenness value in decreasing order. */
5:      while L ≠ ∅ do
6:          L := remove(L_h)  /* Remove edge with highest betweenness From List L. */
7:          if SG_i split into two or more subgraph g_j then
8:              SG ← remove(SG_i)
9:              G ← g_j;
10:             for each subgraph g_j ∈ G do
11:                 s = Q(g_j);
12:                 if s ≥ λ then
13:                     C ← g_j;  /*If subgraph intensity values higher than intensity threshold λ. */
14:                 else
15:                     SG ← g_j;
16:                     goto line 3;
17:                 end if
18:             end for
19:         else
20:             gote line 6;
21:         end if
22:     end while
23: end for
24: return C;
25: end
```

## 4. Experiments

In this section, we present the results of our study. We first introduce the data collection. And then we present topological landscape of social tag networks with respect to the properties of scale free, finally show the results of automatic social tag clustering.

### *4.1. Data Collection*

For the purposes of construct social tag network and automatic social tag clustering based on community detection. We collected live data from Seekda. Seekda is an online search engine for Web services, which crawls and indexes Web Services Description Language (WSDL) files from the Web. It allows users to manually annotate web services using keywords or tags, which describe the function of the web service or provide additional contextual and semantical information.

We crawl 21,273 real Web services from Seekda. For each web service, we get the data of service name, WSDL document, tags, country, and the name of service provider. We select 5,030 web services which contain 1,038 tags as the dataset for experiments. We first remove the noise tags, including stop word, abbreviations, and special symbol. The resulting networks consist of 1,023 nodes and 9,469 edges.

### *4.2. Scale Free*

The network is scale free with its distribution of nodes following a power law. A power law distribution often occurs in complex systems where a majority of nodes have very few connections, while a few nodes have a high degree of connections. Typical power law function has the form of $y = Cx^{-\Upsilon}$, and is captured as a straight line in log-log plots. Here $\Upsilon$ is power law exponent and $C$ is constants. Recent study [14] show that most of real-world power law distributions exhibited an exponent of $2 < \Upsilon < 3$, and that random graphs with this exponent consist of one giant connected component and other small components [15].

The degree distribution of the social tag network follows a power law distribution to some extent, as we show in Fig. 4.1. X-axis is the degree k of a node; it is a measure of the number of tags interaction with other tags in social tag network. Y-axis is the probability $P(k)$; it is a measure to find a tag with this degree and is an indicator of the popularity of an available tags. $\Upsilon$ (=1.3286) is slightly smaller than 2. However, this does not affect entire power law tendency of social tag network. The whole distribution also shows power-law-like properties. This result implies that node with high degree has a capacity to facilitate interactions between the nodes that it links. It is evident to observe the existence of ”hub” tags with huge number of degrees while majority has only a few links, such as bioinformatics, tourism, business, and free etc.

For the community detection perspective, this result suggests this network has obvious community structure. Furthermore, we can also observe some densely connected semantic communities from the social tag network, like biology-related and travel-related etc. In what follows, we focus exclusively on identifying the densely connected semantic communities within the giant component.

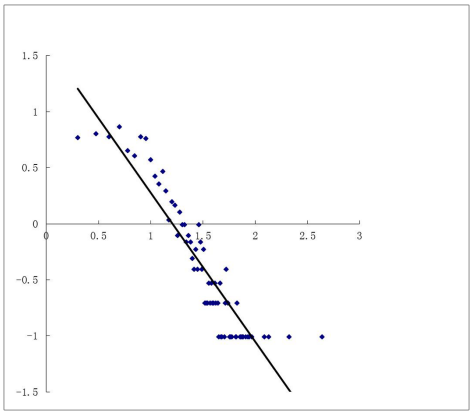


Figure 4.1: Degree distribution of nodes in dual logarithmic coordinates. X-axis is the degree k of nodes, and Y-axis is the probability P(k).

Node betweenness centrality is defined as the number of shortest paths between pairs of nodes that pass through a given node [16]; it can be regarded as a measure of the extent to which a node has control over information flowing between others. The correlation between the node degree and the node betweenness centrality value, which corresponds to the fact that the central nodes are most likely high degree hubs of the network, is shown in Fig. 4.2. We can find that few ”hub” tags with huge number of degrees while majority has highly node betweenness, which is a valuable hint to community detection algorithm.

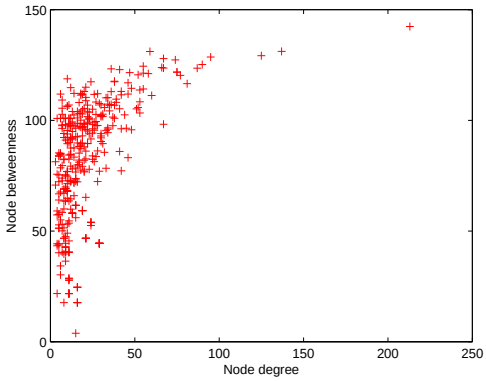


Figure 4.2: The correlation between node degree and node betweenness centrality

### *4.3. Clustering Results*

In this section, we present and discuss the results obtained by the clustering of social tags. We first carried out com-

munity detection process of Section 3 using different intensity coherence threshold $\lambda$ values. Table 1 summarizes some of the community detection results we obtained.

**Table 1** The results of community detection. The columns correspond to the following: $\lambda$ algorithm threshold, number of communities (NC), average community size (ACS), number of tags assigned to communities (NT), percentage of tags assigned to communities (PC).

| $\lambda$ | NC | ACS | NT | PC |
|---|---|---|---|---|
| 0.9 | 65 | 11.3 | 732 | 84.6% |
| 0.8 | 37 | 20.7 | 764 | 88.3% |
| 0.7 | 14 | 57.6 | 806 | 93.2% |
| 0.6 | 11 | 72.1 | 793 | 91.7% |

From Table 1, we can see the community numbers are growing along with the threshold $\lambda$, and most of tags can be assigned to particular community. Through further analyze of each communities in different threshold, we derive the proper result of tag categorization. At the level of $\lambda = 0.7$, we found six cohesive semantic communities that we identified as communities related to biology, travel, business, location, organization and quality (see Table 2).

**Table 2** The six main cohesive semantic communities and some related tags

| Topic | Some related tags |
|---|---|
| biology | bioinformatics, gene, protein, homology, pathogen, predator, peptide, eukaryotic, clustalw, genomics, yeast, microarray |
| travel | tourism, car rental, weather, hotel, airline, trip, city, book flight, accomodation, destination, room search, distance |
| business | onsale, finance, prize, stock, salary, market, forward rate, fund, billing, invoice, swap, earning, pricing |
| location | seattle, australia, USA, england, german, french, chinese, japanese, korean |
| organization | amazon, microsoft, company, msn, individual, university, institute, xignite |
| quality | security, free, internal, test, open, cost, real time, validation,verification |

We count the percentage of different tags category in overall tags. The resulting distributions are shown in Fig. 4.3. The most obvious general conclusion is that tags mainly divides into six categories in Seekda. Specifically, the three most important categories for Seekda are travel, biology and business, these type tags describes what a Web services achievement functions. Location is an additional retrieval cue, the requester usually choose the nearest Web service. Organization is also common in Seekda, as it specifies who created the Web service. Generally speaking, the success services composition was occurred more easily from the same service provider. Quality is a little more frequent, because the fact that the quality of Web service is very hard to describe. Due to the tags fuzziness and randomness, some tags cannot fall into those six classes.

For these unrecognized tags, we tentatively propose some principles in annotation behavior to tackle such problems in future work.

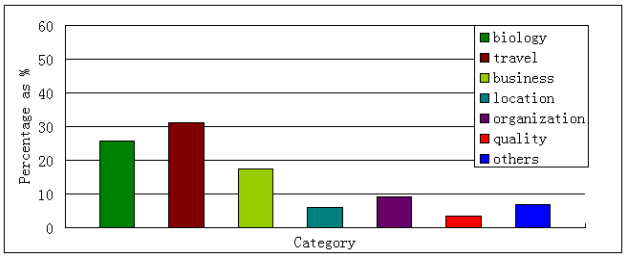


Figure 4.3: Tag categories distributions in Seekda

## 5. Discussion

Collaborative tagging is an act of organizing the resource through keywords or metadata. By automatic clustering these social tags, we can examine what kinds of distinctions are important to taggers. Through analyze the above community detection results, social tags can be divided into the following categories at abstract level.

Function Tags: Tags that describe what a Web services achievement functions. It is probably the most obvious way to describe the Web services, while such functionality information can partially be extracted from the content of WSDL. For some nonstandard Web services, it is not easily accessible. The function tags are further categorized to specific subcategory domains like biology, travel, and business etc.

Location Tags: Tags that provide the physical location of Web services in which the Web services was saved. For example, there is a Web service-the get book name system, it is locate in Seattle.

Organization tags: Some Web services are tagged according to who owns or created the Web services. Given the apparent popularity of Web services among Seekda users, identifying Web services ownership can be particularly important.

Quality tags: Tags can also comment subjectively on the quality of a Web service, expressing opinions based on social motivations typical for tagging systems, or are simply used as rating-like annotations for easing personal retrieval. Some examples of these tags are free, security and internal.

The clustering results show that social tags not only describe the function of Web services, but also express location, organization and quality aspects of the Web services. Obviously the function tag is the most important tags in Seekda, the proportion of tags assigned to those categories is 76.8%. Through the function tag, the requester can easily discovery the desired Web services for user. Meanwhile, the system can also recommend the related Web service to users in tag community. The others classes

tags also providing the complement for Web services discovery and composition. Another advantage of collaborative tag is their ability to rapidly adapt to new changes in terminologies and domains. Thus, it is also useful to complement existing knowledge bases, such as WordNet, YAGO [17] and other several ontologies used in semantic web applications.

Even though the proposed tag community detection algorithm can efficient detect the categorization communities. However, there are several limitations that one should consider before applying it to a new setting. On the one hand, a important issue troubling the application of the proposed community detection method arises from existing some synonym tags. User can use a tag to annotate a Web service while another user can use a synonym of that tag to annotate same Web service. In that case, sometimes two or more communities to be detected as belonging to the same community due to synonym tag that is connected to all of them. On the other hand, tag ambiguity is also one of the main problems. Some tags can be polysemous, where the same word has more than one meaning such apple as a fruit opposed to apple as computer brand. The same tags may be put into the different communities at the same time. In the future we need to investigate efficient ways to avoid the tag fuzziness and ambiguity by context identification.

## 6. Related Works

Research in tag clustering has recently gained much attention due to clustering tags can aid in the personalization of search and navigation.

Some approaches are proposed to automatically classify functional similar tags in whole tag space. Specia and Motta [18] propose a semi-automatic approach using clustering techniques in order to group functional similar tags according to the resources they annotate or to the users who authored the tags. However, the authors do not evaluate how well the clusters of highly co-occurring tags in the similar tag sets help in the disambiguation of tag senses. Zhou and colleagues [19] describes a method to compute the similarity between tag sets and use it as the distance measure to cluster web documents into groups. Although its performance improvements over the traditional similarity measurement not only in the reliable derivation of clustering results, but also in clustering accuracies and efficiencies, the approach does not propose any method to eliminate ambiguous among tags. Giannakidou et al. [20] presents a statistical approach for discovering the semantic functional similar tags. This approach is based on a similarity measure that mixes tag co-occurrence with semantic similarity. Nevertheless, this approach clusters tags into disjoints groups. This means that a tag can belong to just one group and therefore if a tag has several meanings the approach will only identify the most frequent meaning for that tag according to the tag co-occurrence pattern.

Other clustering algorithms have been proposed for automatically identifying tag types. Wartena [21] investigate different types of tags and define a number of features that are typical for some of these classes. In order to classify tags automatically into the proposed categories they have used the logistic linear classifier. Nevertheless, these tags lack a uniform representation to facilitate their sharing and reuse. To solve the above-mentioned problem, TagExplorer [22] detects the tag type of Flickr tags using WordNet. In the TagExplorer system, Flickr tags are categorized into location, subject, names, activity and time. Each tag type is mapped to specific WordNet categories in order to give a uniform representation for tags. However, using WordNet alone may be insufficient coverage of the Flickr vocabulary. Bischoff and colleagues [8] manually classify a number of tag collections obtained from different social tagging systems in several tag types, and study the distributions of tags assigned to each type, analyzing their usage implications on search tasks. The obtained results provide insight into the use of different kinds of tags for improving search. On the basis of Bischoff et al. work, Cantador et al. [23] mapping different type tags to semantic concepts existing in the multidomain YAGO ontology, which is a Semantic Web knowledge base with structured information extracted from WordNet and Wikipedia, and do not perform any disambiguation technique.

Our work is inspired by tags co-occurrence with semantically related e.g., Ref. [18,20]. However, unlike those works, we find co-occurrence tags express the different aspects of resource not only functional similar. In fact, some tags may describe the function and other tags depict the domain, location, organization and qualities for the same resource.

## 7. Conclusion and Future work

Effective Web service discovery is an important issue, especially for non-semantic Web services. Automatically clustering tagged Web service into semantic communities can improve the accuracy of service discovery and recommendation. In this paper, we first design the social tag network based on tags social dimension. Network analysis results show that social tag network with respect to the scale free properties. In order to identify densely connected semantic communities, we then introduce a novel graph-based clustering algorithm for weighted network. Experimental results on real world data sets show that our algorithm can effectively discovers the semantic communities and the resulting tag communities correspond to meaningful topic domains. Our clustering approach can be integrated into search engines to improve the quality of Web service discovery and recommendation by helping to identify semantic related tag groups.

As future work, we plan to extend our graph-based clustering algorithm to the larger dataset for further evaluate our work. We also hope to investigate ways to deal with the limitations of the proposed algorithm more detailed in Section 5, namely tag fuzziness and ambiguity. In addition, we would like to propose a method for Web services recommendation that make use of the results of community detection.

## Acknowledgements

We would like to thank all members of service computing group from Institute of Knowledge Science and Engineering, Tianjin University. This work was supported by the National Natural Science Foundation of China under Grant No.61173155, the 985 Project of Tianjin University under Grant No.06050110000, the National High-Tech Research and Development Program of China under Grant No.2007AA01Z130, and the Innovation Foundation of Tianjin University under Grant No. 2010XG-0009.

## References

[1] M. P. Papazoglou and W. J. Heuvel, Service oriented architectures: approaches, technologies and research issues, The VLDB Journal. **16**, 389-415 (2007).
[2] A. Becker, T. Widjaja and P. Buxmann, Value potentials and challenges of service-oriented architectures, Business Information Systems Engineering. **3**, 199-210 (2009).
[3] A. Maurizio, J. Sager, G. Corbitt and L. Girolami, Service oriented architecture: challenges for business and academia, Proc. The 41st Annual Hawaii International Conference on System Sciences, 315-323 (2008).
[4] D. Mukhopadhyay and A. Chougule, Survey on web service discovery approaches, Advances in Intelligent and Soft Computing. **166**, 1001-1012 (2012).
[5] J. R. Falleri, Z. Azmeh, M. Huchard and C. Tibermacine, Automatic tag identification in web service descriptions, Proc. The International Conference on Web Information Systems and Technology, (2010).
[6] M. Gawinecki, Analysis of seekda tags for web service matchmaking, Technical report, University of Modena and Reggio Emilia, (2009).
[7] G. Begelman, P. Keller and F. Smadja, Automated tag clustering: Improving search and exploration in the tag space, Proc. 15th World Wide Web Conference, ACM Press, 22-26 (2006).
[8] K. Bischoff, C.S. Firan, W. Nejdl and R. Paiu, Can all tags be used for search?, Proc. The 17th ACM Conference on Information and Knowledge Management, 203-212 (2008).
[9] M. Gawinecki, G. Cabri, M. Paprzycki and M. Ganzha, Structured collaborative tagging: is it practical for web service discovery?, Web Information Systems and Technologies, Lecture Notes in Business Information Processing. **75**, 69-84 (2011).
[10] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences. **99**, 7821-7826 (2002).
[11] T. Narayanan, M. Gersten, S. Subramaniam and A. Grama, Modularity detection in protein-protein interaction networks, BMC Research Notes. **4**, (2011).
[12] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Physical Review E. **69**, 026113 (2004).
[13] J. P. Onnela, J. Saramki, J. Kertsz and K. Kaski, Intensity and coherence of motifs in weighted complex networks, Physical Review E. **71**, 065103 (2005).
[14] M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, Contemporary Physics. **46**, 323-351 (2005).
[15] R. Albert and A. L. Barabasi, Statistical mechanics of complex networks, Reviews of Modern Physics. **74**, 47-97 (2002).
[16] B. Mirzasoleiman, M. Babaei and M. Jalili, Cascaded failures in weighted networks, Physical Review E. **84**, 046114 (2011).
[17] F.M. Suchanek, G. Kasneci and G. Weikum, YAGO: a large ontology from Wikipedia and WordNet, Journal of Web Semantics. **6**, 203-217 (2008).
[18] L. Specia and E. Motta, Integrating folksonomies with the semantic web, Proc. The 4th European Conference on the Semantic Web: Research and Applications, Innsbruck, Austria, 624-639 (2007).
[19] J. L. Zhou, X. J. Nie, L. J. Qin and J. F. Zhu, Web clustering based on tag set similarity, Journal of Computers. **6**, 59-66 (2011).
[20] E. Giannakidou, V. Koutsonikola, A. Vakali and Y. Kompatsiaris, Co-clustering tags and social data sources, Proc. The Ninth International Conference on Web-Age Information Management, 317-324 (2008).
[21] C. Wartena. Automatic classification of social tags, Proc. The 14th European Conference on Research and Advanced Technology for Digital Libraries, Berlin, Heidelberg, 176-183 (2010).
[22] B. Sigurbjornsson and R. Zwol, Tagexplorer: faceted browsing of flickr photos, Technical Report, Yahoo! Research, (2010).
[23] I. Cantador, I. Konstas and J. M. Jose. Categorising social tags to improve folksonomy-based recommendations, Journal of Web Semantics. **9**, 1-15 (2011).

**Weisen Pan** is a PhD Candidate in School of Computer Science and Technology, Tianjin University, Tianjin, China. His research interests are in the areas of service computing, and social computing.

**Shizhan Chen** received the PhD degree in Computer Application Technology from the School of Computer Science and Technology, Tianjin University in 2010. He is currently a university lecturer in Tianjin University. His research interests are in the areas of service computing.

**Zhiyong Feng** is a Professor at the School of Computer Science and Technology, Tianjin University. He received the PhD degree from Tianjin University in 1996. His research interests include knowledge engineering, services computing, social computing and security soft-ware engineering.