Mining and Representing User Interests: The Case of Tagging Practices

Hak-Lae Kim, John G. Breslin, Stefan Decker, and Hong-Gee Kim

Abstract—Social tagging in online communities has become an important method for reflecting classified thoughts of individual users. A number of social Web sites provide tagging functionalities and also offer folksonomies within or across the sites. However, it is practically not easy to find users' interests based on such folksonomies. In this paper, we provide a novel approach for clustering user-centric interests by analyzing tagging practices of individual users. To do this, we collect Really Simple Syndication data from blogosphere, find conceptual clusters using formal concept analysis, and then evaluate the significance of these clusters. The results of the empirical evaluation show that we can effectively recommend different collections of tags to an individual or a set of users.

Index Terms—Concept analysis, Semantic Web, social tagging, tag ontology.

I. INTRODUCTION

THERE is considerable interest in emergent virtual networks on social Web sites, owing particularly to user-created content [30]. People have voluntarily participated in content creation and sharing on social Web sites via diverse objects such as videos, audios, tags, bookmarks, and so on. Because specific interests from these activities can be formed as a network that consists of people and their objects of interest, as opposed to a people-to-people network, the attractiveness of an object-based network in which it is built is very important—this is an object-centered social network that refers to the phenomenon where a social tie is created not just between people but between people and objects or around objects from sociologist Karin Knorr Cetina [8], [9], [11].

In order to construct an object-centered network across different Web sites, however, it is necessary to take into consideration a guideline that collects interests from users' activities beyond objects. People are creating their activities on social Web sites with diverse objects. As their activities can be shared and aggregated across heterogeneous sites, people have turned their interest to find out relevant communities according to their activities. In addition, they will continuously seek the most rel-

evant people and groups for their interests, since these communities can be created in decentralized spaces regardless of where they carried out their online activities such as social networking, social tagging, social bookmarking, or social content sharing.

To put the solutions described in this paper into context, one must first understand what objects are common to create a virtual community. Tags are used to depict people's interest to online resources, and tags in most social Web sites have become an available feature. Since tags are used to organize individuals' ideas or thoughts as well as to encourage their social interaction [15], [28], [38], tagging activities on social Web sites can be considered a new way of collective authorship [35]. Once a set of tags is assigned to a resource, a network structure can be constructed given a number of users and the tags that they assign to a set of shared resources. An interest group based on tagging data is extracted from tagging behaviors; sets of tags can be used to build social networks and promote their use by other people in online communities.

The remainder of this paper is structured as follows. We start by providing an overview of social tagging and folksonomies and introducing the Semantic Web. Formal concept analysis (FCA) is introduced and discussed in Section III. Section IV describes an algorithm to compute a significance of tagging practices based on tag frequencies. Section V describes the characteristics of given sample data. The evaluation and the results of the experiments are discussed in Sections VI and VII, respectively. Section VIII discusses the representation of an "interest group" at a semantic level. Finally, in Section IX, we present some conclusions and explore future directions.

II. LITERATURE REVIEW

A. Social Tagging and Folksonomy

Tagging has become a popular method for annotating and organizing online resources, and it is also a widely used function in social software [37]. Social tagging or folksonomies as new ways of information categorization and indexing have received a great deal of attention from the Semantic Web and Web 2.0 communities. Many studies have already been carried out on social tagging ranging from social networks [2], [30], data mining and clustering [15], [17], [28], information retrieval [37], and visualization [10]. These research efforts have also deduced their results by using tagging data generated via user participation.

Most people agree that a tag is not simply a keyword anymore but that there is semantic information around a "tag" [34], [38]. When individual tags are shared among different users and used with other tags, they evolve into "social" tags. Thus,

folksonomies (which embed social relationships) have shared meaning and understanding, reflecting a conscious perception of individual users. Since any act of classification affects the definition of the entire tag, people who use a common set of tags in a certain community can be viewed as a *potential* interest group given their tagging practices.

B. Tagging on the Semantic Web

Typical social tagging systems do not provide explicit links among the entities that realize tagging activities (i.e., users, tags, and resources), nor do they expose their data in a standard form. Semantic-Web-based approaches can support a standardized metadata schema to represent both structures and semantics of tagging data. The Semantic Web, a Web of data [4], enables the Web to understand the requests of people and machines to use the Web content semantically and automatically. It aims to provide a common framework that allows data to be shared and reused across application, enterprise, and community. Semantic Web technologies, in general, allow us to expose human knowledge to machines to perform automatic data linking and data integration [23].

According to Gruber [16], an ontology for tagging is not just a way to define meanings of certain tags, but it can also robustly represent the relationships among the entities that shape tagging activities, explicitly stating the knowledge structure of tagging data. In this perspective, he emphasizes the need for folksonomies and ontologies to work together, aiming to identify and formalize a conceptualization of tagging data at a semantic level.

Tag ontologies, in general, aim to provide a common conceptualization of the tagging process and the generated data, providing a standardized way to collect, interpret, or use shared tagging data. There is an agreement that the building blocks of a tagging model consist of the taggers, the tags, and the actual resources being tagged [23]. One of the advantages of tag ontologies is that isolated tagging data can be easily made mobile and integrated across applications. Tags, users, and their relations in a particular application can be represented in a representation language such as the Resource Description Framework (RDF) [1] or the Web Ontology Language [3]. Eventually, these data become accessible and mobile on the Web as Linked Data [5]. This can be considered as a starting point for the sharing and exchange of separate tagging activities on different platforms. Therefore, tag ontologies will become a very powerful means to make tagging data portable and to keep such data connected across systems. So far, there are several tag ontologies that have different objectives [23], and Kim *et al.* [22] proposed a novel approach to combine vocabularies among different tag ontologies. Social Semantic Cloud of Tags (SCOT)[1] aims to describe folksonomic characteristics and to offer social interoperability of semantic tag data across heterogeneous sources [20]. This model can express the structure of, features of, and relationships between tags and users, allows the exchange of semantic tag metadata for reuse in social applications, and enables interoperation among data sources, services,

[1]http://scot-project.org

TABLE I
EXAMPLE OF A FORMAL CONTEXT $\mathcal{C}$

	book	movie	travel	photo
Adam	X	X	X	
Bill	X		X	X
John	X		X	
Alice			X	X
Bob		X	X	

or agents in a tag space [19]. These features are a cornerstone to being able to identify, formalize, and interoperate a common conceptualization of tagging activity at a semantic level.

III. FCA IN SOCIAL TAGGING

FCA is a mathematical theory used for conceptual data analysis and unsupervised machine learning. The term was invented by Rudolf Wille in the early 1980s [44]. FCA has been applied to a wide range of disciplines, including software engineering, psychology, artificial intelligence, and linguistics [33]. In computer science, various FCA efforts have been performed in many fields such as information retrieval [18], [40], social network analysis [12], [21], [36], and knowledge representation and conceptual graphs [39].

A. Formal Context

FCA models the world of data through the use of objects and attributes. The relations between objects and attributes in a data set form the *formal context*. This is represented by the triple (G, M, I), where G is a set of *objects*, M refers to a set of attributes, and $I \subseteq G \times M$ specifies the binary relation between G and M. If an object $g \in G$ has an attribute $m \in M$, the relation is represented by $(g, m) \in I$ or gIm.

A formal context G is often described by a 2-D matrix. Table I illustrates the "context" of a number of people and the hobbies that they have. The objects (persons) correspond to the rows of the table, whereas the attributes (hobbies) are represented by the columns. The cross in the table indicates that an object g has an incidence relation with an attribute m. In this example, the objects are the persons $g =$ {'Adam', 'Bill', 'John', 'Alice', 'Bob'}, while the attributes are the hobbies $m =$ {'book', 'movies', 'travel', 'photo'}.

B. Lattice of Formal Concepts

A *formal concept* refers to the relationship between a set of formal objects and a set of attributes. A concept of the context (G, M, I) models a duality relationship between a subset of objects denoted by A and a subset of attributes denoted by B

$$A^I := \{m \in M | \forall g \in A\} \tag{1}$$

$$B^I := \{g \in G | \forall m \in B\} \tag{2}$$

where A^I is the set of attributes common to the objects in A and B^I is the set of objects common to the attributes in B. A (formal) concept of the context (G, M, I) is a pair (A, B), where $A \subseteq G, B \subseteq M, A = B^I$, and $B = A^I$. The set A is called the extent and B is called the intent of the concept (A, B).
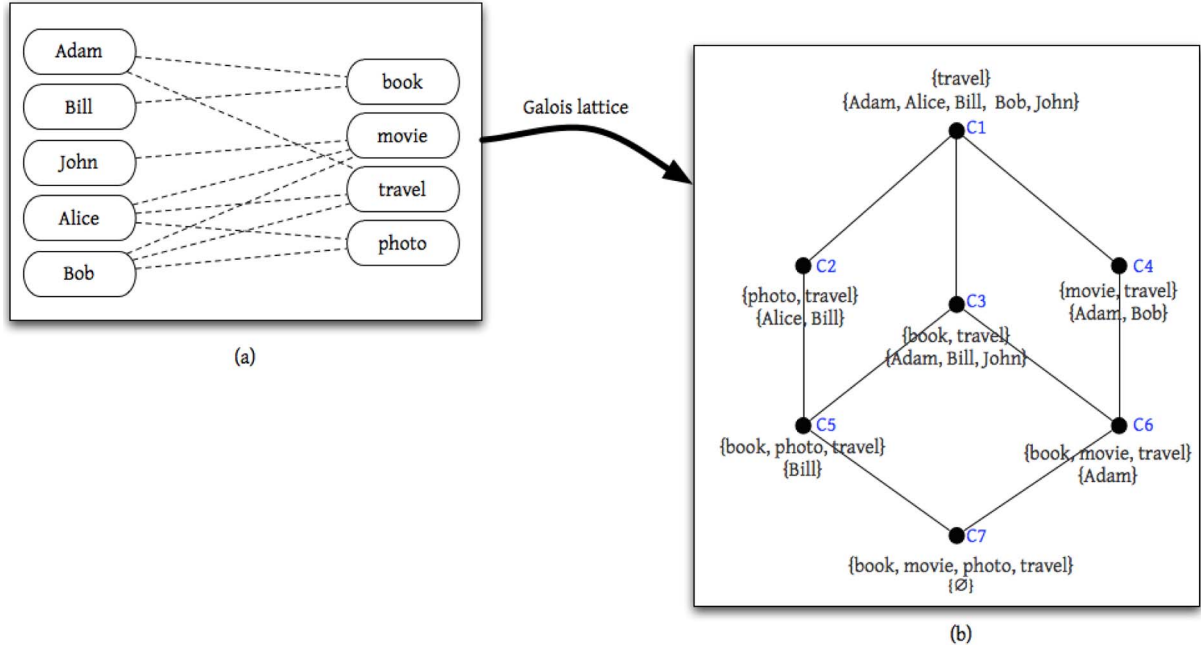
Fig. 1. Concept lattice for the formal context in Table I. (a) to (b) illustrates the way how the formal context is transformed into Galois connections. The nodes in (b) represent formal concepts.

The concepts of FCA are often arranged in a *hierarchical order* based on a $\leq$ relation between concepts. A concept (A_1, B_1) is a *subconcept* of a concept (A_2, B_2) if $A_1 \subseteq A_2$ or $B_2 \subseteq B_1$. Correspondingly, (A_2, B_2) is a *superconcept* of (A_1, B_1); hence, $(A_1, B_1) \leq (A_2, B_2)$. This relation is referred to as a lattice. That is, a lattice is a partial set in which two any elements have both a least upper bound and a greatest lower bound [43]. The lattice maps an ordering from the most general to the most specific concept, from top to bottom. The topmost concept (the largest subconcept) is called the *supremum*, and the concept at the very bottom (the smallest subconcept) is called the *infimum*.

Fig. 1(a) shows the sample group, with relationships between persons and interests shown using a dashed line, and Fig. 1(b) shows a concept lattice corresponding to the formal context represented in Table I. The move from (a) to (b) illustrates the way that the formal context is transformed into Galois connections. The nodes in (b) represent formal concepts. FCA provides the advantage that Galois connections and the set of formal concepts can be visualized. The lines in Fig. 1(b) indicate hierarchic relationships from top (most general) to bottom (most specific). The attributes "movies" and "travel" are the most popular interests that people have (i.e., {movies; John, Alice, Bob} and {travel; Alice, Bob, Adam}), while a single person is interested in all of the objects.

Based on FCA, we present a formal framework that describes an interest group based on a user's tagging practices.

Definition 1) *Interest group.* Given a set of users $\mathcal{U}$ and considering the tags that they have in common, the interest group of $\mathcal{U}$ is the set of users who are using these tags.

Definition 2) *Intent.* The intent of a set of users $\mathcal{U}$ is the set of tags which are used by every user in $\mathcal{U}$.

Definition 3) *Extent.* The extent of a set of tags $\mathcal{T}$ is the set of users using every tag in $\mathcal{T}$.

Thus, an interest group would be a set of users that use a significantly similar collection of tags to identify their resources.

C. Lattice Construction

Because of the inherent complexity of the lattice-building algorithms, a concept lattice may be feasible for small- to medium-size collections. The size of concept lattices can grow exponentially with respect to the number of contexts [27]. However, this occurs rarely in practical applications, as witnessed by a number of theoretical and experimental findings. Several algorithms have been developed for constructing a concept lattice [24], [29]. These algorithms aim to solve the problem of generating the set of all concepts and the concept lattice for a formal context. From the perspective of the worst case complexity, an algorithm constructing all concepts and/or a concept lattice can be considered *optimal* if it generates the lattice with polynomial time delay and is space linear in terms of the number of all concepts [26]. The algorithm by Ganter and Kuznetsov [13] computes the set of all concepts in time $O(|G|^2|M||L|)$, where $|L|$ is the size of the concept lattice, and has polynomial delay $O(|G|^2|M|)$. The algorithm by Berry *et al.* [6] uses a tree for fast storing and retrieval of concepts. The complexity of this algorithm is $O(|G||M|^2|L|)$ and has a polynomial delay $O(|G||M|^2)$. The close-by-one algorithm [25], which is based on a computation of closures, uses a similar notion of canonicity. This is a similar method for selecting subsets and employs an intermediate structure that helps to compute closures more efficiently using the generated concepts. Its time complexity is $O(|G|^2|M||L|)$, and its polynomial delay is $O(|G|^3|M|)$. The Norris algorithm [31] is basically an incremental version of the close-by-one algorithm. The time complexity of the algorithm is $O(|G|^2|M||L|)$. The best theoretical worst time complexity is $O(|L||M|(|G| + |M|))$, exhibited by the algorithm presented by Nourine and Raynaud [32].

A comparative study of several algorithms for constructing concept lattices is extensively studied in the literatures [26], [27]. According to Kuznetsov and Obiedkov [26], the choice of an algorithm for the construction of the concept lattice should be based on the characteristics of input data; in practice, the behavior may significantly vary depending on a number of factors including the relative sizes of G and M, the size of I, and the density of the context, i.e., the size of I relative to the product $|G||M|$. They recommended the close-by-one and Norris algorithms for small and sparse contexts, while Berry *et al.* perform well on contexts of average density. In particular, we use the algorithm by Ganter and Kuznetsov [13] for constructing a concept lattice since this algorithm is probably the best known algorithm for generating a concept lattice [27].

D. Limitations

The main advantage of a Galois lattice is the focus on subset-oriented visualization that is particularly appropriate for representing the relationships between users and tags in a given data set. This approach, however, has shortcomings when it comes to analyzing interest groups, including the following.

1) From an analytic point of view, the challenge for research in FCA is the size of the concept lattices. We are able to reduce significantly the size of the concept lattice using a conceptual scaling technique [14], [42], which reduces the complexity of concept lattices. However, the resulting lattice may grow exponentially to the size of the data set [42].

2) The users and the tags in a lattice structure have a high degree of overlap among concepts. Therefore, it is difficult to recommend the concepts when some queries are performed. Although the scaling techniques using nominal, ordinal, or interval scales are useful, they are limited when handling numerical data such as tag frequencies.

The strengths and weaknesses of FCA discussed earlier hint at their complementarity. In particular, using some characteristics of tagging data may help refine the results over large sets of users with a rich collection of attribute types. In the following section, we will discuss how tag frequencies provide an insight into users' tagging practices and propose a way to compute the significance of each concept based on the tag frequency.

IV. TAG FREQUENCIES

A tag has a weighted occurrence that is associated with or assigned to certain resources. The more a tag appears, the stronger its influence might be on the eventual classification and retrieval of a resource. Thus, tag occurrence plays an important role to infer topics of interest for the user given the tagging data. In order to represent the proper occurrences of a tag in a formal concept, we consider different types of tag frequencies and discuss the fundamental issue of "tag occurrence." We need a measure to assign weights to the concepts obtained through FCA by taking into account the popularity of tags within the same concepts.

A. Absolute Versus Relative Frequency

When we say "*a tag is popular*," the observation can be based on either the real frequency of a certain tag or its frequency that is proportional to the frequencies of other tags. That is, a frequency of a tag can be expressed as the absolute or relative (normalized) frequency. The former is based on raw observations by means of a sequential process of observing and counting a tag that have not normalized with respect to the base rates of the event. On the other hand, the latter means a frequency that is expressed in relation to a sample size or rate. In general, the absolute format is useful since this simply gives the number of occurrences of each tag. However, this does not indicate the magnitude of a tag in terms of a proportion of the total number of tags in a given data. It is a problem when two data sets with different sizes are being compared. Thus, using the absolute format on its own cannot be used to correctly compare the occurrence of tags in two or more groups with different sizes.

The most basic way to calculate the relative frequency of a tag for a concept is

$$R(t) = \frac{O_{(t)}}{\sum O_{(i)}} \tag{3}$$

where $R(t)$ is the proportion of tag t, $O_{(t)}$ is the number of occurrences of tag t, and $\sum O_i$ is the total occurrence of the tags in the sample. This format is normally used to display a tag cloud where the size of each tag represents its popularity as a ratio to the popularity of all tags.

B. Refinement of the Formal Concepts

In our case, each formal concept has different users and tags, and tagging behaviors between the users also differ (i.e., there are different numbers of resources, tags, and occurrences of each tag). We need to evaluate the significance of each concept by taking into consideration the absolute and relative frequencies of tags within that concept. To achieve this, we compute the significance score of a concept $S(c)$ by using both the absolute and relative frequencies of the individual tags within the concept tag cluster

$$S(c) = \frac{1}{N} \left(\sum_{i=1} \sum_{k=1} F(u_i t_k) \times R(u_i t_k) \right) \tag{4}$$

where $F(u_i t_k)$ is the absolute frequency of tag t_k of user u_i in tag cluster c and $R(u_i t_k)$ [computed by (3)] is the normalized frequency of the tag of user u_i in the concept tag cluster. Thus, to obtain $S(c)$ in each cluster, we multiply the absolute and relative occurrences of each tag and sum the score up for each individual user. The user's scores are then summed up and averaged.

If we computed $S(c)$ via the usual mean method, each tag occurrence would have an equal impact on the result regardless of its popularity. Using this method, the effect of less popular tags on the concept weight is reduced.

As an example, we calculate the significance measure of concepts C_2 [i.e., ({Alice, Bill}, {photo, travel})] and C_3 [i.e., ({Adam, Bill, John}, {book, travel})] in the concept lattice

TABLE II
OCCURRENCES OF EACH KEYWORD

	book	movies	travel	photo	...	$\sum$
Adam	55	9	35	0	...	120
Bill	25	0	20	16	...	75
John	40	0	60	0	...	150
Alice	0	0	15	2	...	50
Bob	0	10	15	0	...	40

(see Fig. 1). Using the occurrences of each keyword in Table II, we can compute S_c as follows:

$$S_{(C_2)} = \frac{1}{2} \times \left(\left(2 \times \frac{2}{50} + 16 \times \frac{16}{75} \right) \right.$$
$$\left. + \left(15 \times \frac{15}{50} + 1 \times \frac{1}{75} \right) \right)$$
$$= 4.003$$
$$S_{(C_3)} = \frac{1}{3} \times \left(\left(55 \times \frac{55}{120} + 25 \times \frac{25}{75} + 40 \times \frac{40}{150} \right) \right.$$
$$\left. + \left(35 \times \frac{35}{120} + 20 \times \frac{20}{75} + 60 \times \frac{60}{150} \right) \right)$$
$$= 29.597.$$

From the aforementioned example, the significance of two concepts can be directly calculated from the frequencies of attributes (keywords) per user. The significance depends on the frequencies of objects and attributes rather than on the total number of them (i.e., how many there are). In general, FCA assumes that, the more objects and attributes a concept has, the more similar the objects in the concept are. The proposed model gives an effective method to calculate a significance measure by using numerical features of objects and attributes in the process of concept analysis. Compared with other similarity measures, using this method, a significance of individual concepts can be measured and compared. Moreover, this offers more convenient means to calculate concept similarity in a concept lattice with a large context. Obviously, calculating the significance of concepts is simpler than constructing a concept lattice.

V. DATA COLLECTION AND METHODS

Information and communication technology (ICT) bloggers tend to annotate their ICT-related topics. By analyzing their tagging practices, we may find out potential interest groups for a specific set of bloggers.

To do this, we gathered a collection of Really Simple Syndication (RSS) feeds from the blogosphere which focus on information technologies between January 2004 and May 2009. The list of RSS feeds was collected from the following links (aggregators and blog indexes):

1) http://radio.xmlstoragesystem.com/rcsPublic/rssHotlist;
2) http://c18-gdl-xw-lb.cnet.com/html/ne/blogs/ CNETNewsBlog100.opml;
3) http://www.currybet.net/download/opml/ top100ukblogs.xml;
4) http://planetrdf.com/index.rdf;
5) http://planet.apache.org/committers/opml.xml;
6) http://planet.xmlhack.com/bloggers.rdf;
7) http://nooperation.typepad.com/files/ top100devblogs.xml.
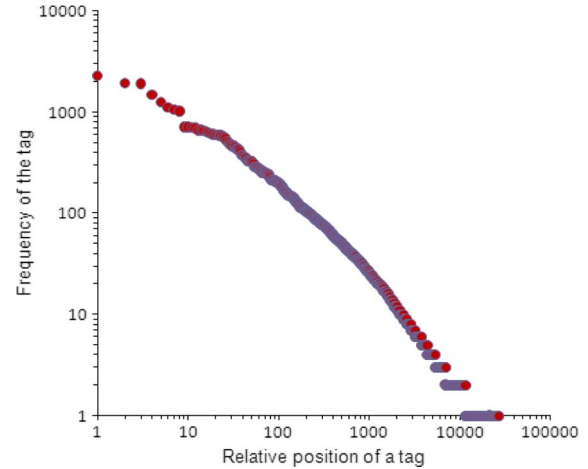

Fig. 2. Distribution of the frequency of tags in the given data set.

Initially, we aggregated RSS feeds for each blogger using the Google AJAX Feed API[2] that provides easy access to public data feeds through JavaScript. Some feeds from these sites are, however, filtered to just have posts from specific blog categories. In these cases, we found the URLs of the full feeds and aggregated them instead.

We obtained a folksonomy with $|U| = 528$ users, $|T| = 27\,025$ tags, $|R| = 82\,047$ resources, and $|A| = 166\,639$ tag assignments in $|R|$. Fig. 2 shows the distribution of the tag frequency on a log-log scale. The horizontal scale gives the logarithm of the relative position (where the most used tag is in position 1, the second most used tag is in position 2, and so on). In other words, the x-axis denotes the frequency of the tags ordered by descending tag frequency. The y-axis refers to the tag frequency. We observe a power law in both distributions. This means that most tags appear only once or rarely, while a few tags are used frequently. In our data set, the tag *general* is the most popular with frequency $|F| = 2274$. The tags *uncategorized*, *semantic web*, *java*, *technology*, and *news* have frequencies of 1919, 1915, 1472, 1257, and 1119, respectively. The popular tags from this power law seem to demonstrate general and intuitive aspects of using tags in the data set. We will now describe a new method for finding out interest groups thorough an analysis of tagging practices.

VI. EXTRACTION AND EVALUATION OF CONCEPTS

As noted earlier, a formal context can be transformed into concepts in a lattice structure. The concepts can be considered as potential candidates of interest groups for individual users, since each concept consists of a set of users and a set of common tags. There are some applications supporting the production and visualization of a concept lattice [41]. However, the visualization of a Galois lattice can become quite complex as the number of users and/or the number of tags become large. It is difficult for users to find meaningful groups based on such visualization. Therefore, we extract a list of concepts from a formal context using the algorithm by Ganter and Kuznetsov,
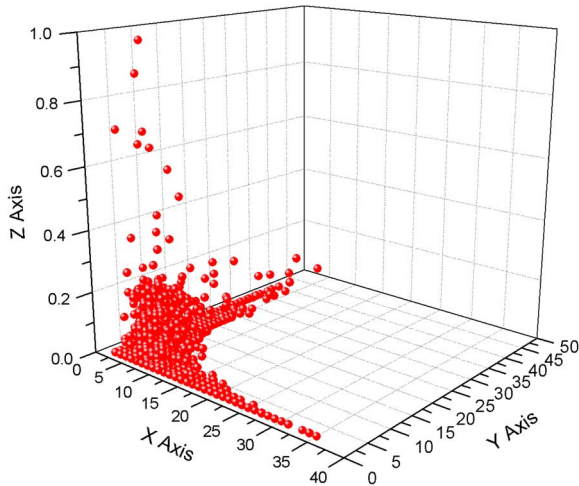
[2]http://code.google.com/apis/ajaxfeeds/

Fig. 3. Execution time for producing concepts. The x-axis refers to the number of users, the y-axis means the number of tags, and the z-axis refers to the execution time of a concept.

Fig. 5. Number of users and tags in the concepts. The x-axis refers to the number of users, and the y-axis means the number of tags.
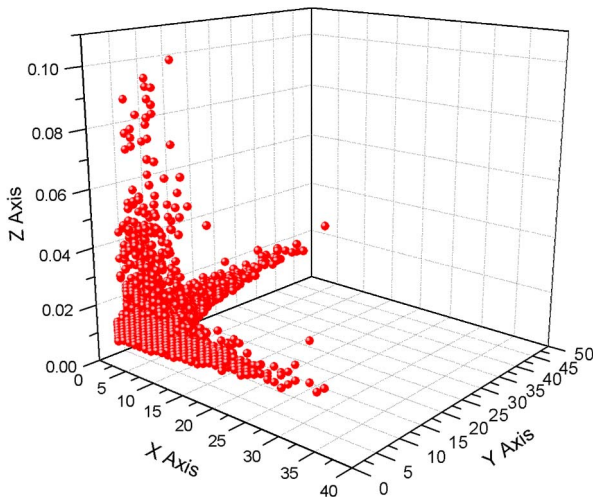

Fig. 4. Execution time for the significance measure of concepts. The z-axis refers to the overall time for calculating the significance of concepts.

Fig. 6. Significance values of concepts. The x-axis refers to the number of users, the y-axis means the number of tags, and the z-axis refers to a significance measure of concepts, which consist of a number of users and tags.

We also evaluated the algorithm for computing a significance measure of concepts. A list of concepts derived from the observed lattice is the basis for calculating the significance of a set of users and tags and for analyzing hidden structures in the concepts. The mean of the generation time is 0.0015 s. Comparing Figs. 3 and 4, the process of measuring significance does not need more time than the task for a concept's production. Although the two algorithms implemented here do not lead to improved complexity results, the overall process for constructing and measuring significance of a concept yields practical performance for a large number of formal contexts.

and the algorithm in (4) is used to compute a significance measure of each concept. These algorithms were implemented in Java in the Eclipse environment. All tests were run on an Intel Core 2 Duo processor and 2-GB RAM. Here, we present two charts that show how the execution time of the algorithms depends on various parameters (see Figs. 3 and 4).

A. Finding and Measuring Concepts

The resulting lattice has 41 450 concepts [i.e., pairs of a set of users (objects) and a set of tags (attributes)] with an average of 5.5 users and 5.1 tags per concept. Fig. 3 shows the execution time versus a concept consisting of a number of users and tags. Although the overall structure and the size of the lattice have some impacts on the running time, most concepts were built within a reasonable time. Except for concept C_{35146}, which has the maximum time (i.e., 7.502 s), the average generation time is 0.017 s.
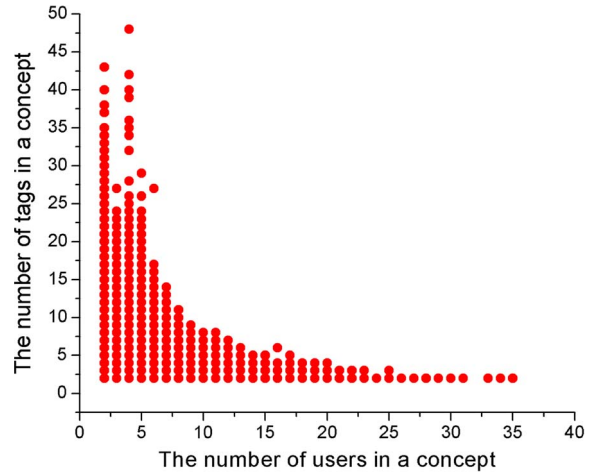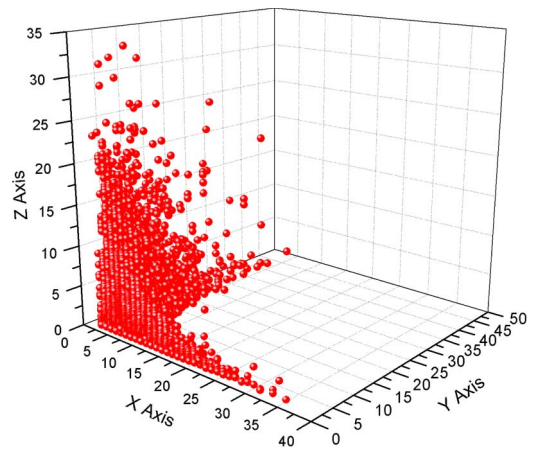
VII. RESULTS

A. Refining Concepts' Significance

We now turn to refining the concepts produced via FCA. Fig. 5 shows the relationships between the number of users and the number of tags within the concepts. Roughly, the numbers of tags are in inverse proportion to the number of users in the concepts. That is, the higher the co-occurrence of tags, the

TABLE III
TOP TEN CONCEPTS BASED ON THE SIGNIFICANCE MEASURE

Concept	$S(c)$	Users (extents)	Tags (intents)
4198	33.71	Mashable!, Google Blogoscoped, Minding the Planet, O'Reilly Radar - Insight, analysis, and research about emerging technologies., AI3:::Adaptive Information	google, internet, search, technology
649	31.31	Dave Naylor a UK SEO and Search Marketing, SEOptimise	google, marketing, yahoo, digg, socialnetworking, search, wordpress, links, webdesign, socialmedia, money, seo
18545	26.08	Mashable!, ReadWriteWeb	music, google, twitter, news, amazon, events, yahoo, identity, apple, facebook, socialnetworks, conferences, trends
22841	21.37	Dave Naylor a UK SEO and Search Marketing, Mashable!, SEOptimise, UMBC ebiquity	google, yahoo, socialnetworking, search, wordpress, socialmedia, seo
97	19.90	Cornwallseo.com, Dave Naylor a UK SEO and Search Marketing, SEOptimise	marketing, digg, socialnetworking, links, webdesign, socialmedia, seo
4202	19.73	Mashable!, The Venture Skills Blog, Google Blogoscoped, Minding the Planet, O'Reilly Radar - Insight, analysis, and research about emerging technologies., Geospatial Semantic Web Blog, AI3:::Adaptive Information	google, search, technology
4200	18.26	Carlos Sanchez's Weblog, Javalobby - The heart of the Java developer community, Code Monkeyism	eclipse, testing, apache, ajax, java, maven
22853	18.22	Dave Naylor a UK SEO and Search Marketing, Mashable!, SEOptimise, The Venture Skills Blog, UMBC ebiquity	google, yahoo, search, wordpress, socialmedia, seo
18546	17.36	Adactio, Mashable!, ReadWriteWeb	music, google, twitter, news, amazon, events, yahoo, identity, apple, facebook, socialnetworks
893	16.97	Cornwallseo.com, SEOptimise, Distilled - Monitoring your Reputation Online	twitter, marketing, links, webdesign, socialmedia, seo

TABLE IV
POSSIBLE INTEREST GROUPS OF *Cloudlands*

Concept	$S(c)$	Users (extents)	Tags (intents)
32772	21.73	Between the Lines, Cloudland	google, web2.0, general, innovation, podcasts
32706	20.61	Webby's World, Cloudland	music, games, google, rss, internet, london, web, web2.0, education, blogs, blog
32719	14.33	Webby's World, Mashable!, Cloudland	music, games, google, internet, web, web2.0, blog
33055	14.20	Minding the Planet, UMBC ebiquity, Cloudland	games, google, rss, internet, web, web2.0, socialnetworks, socialmedia
32707	13.29	Adactio, Webby's World, Cloudland	music, games, google, rss, internet, london, web, web2.0

lower the co-occurrence of users and vice versa. Note that users who share a number of tags may not have similar interests all the time, since they may have different aspects for using tags. In contrast, Fig. 6, based on the significance measure ($S(c)$) presented in the previous section, shows different results. The x-axis refers to the number of users, the y-axis means the number of tags, and the z-axis refers to a significance measure of concepts, which consist of a number of users and tags. The significance scores reflect not only the results of the concept analysis but also the tagging practices of individual users. The plots in Fig. 6 may not be dependent on the number of users or tags, whereas the $S(c)$ computed by (4) has an effect on the locations of the plots.

Surprisingly, some concepts have low significance values or have a significance score of zero, although they are several tags in these concepts. We interpret a low significance score for a concept as follows: The users in that concept have a low incidence of using the tags in the concept, or else, the tags are relatively popular in the individual sense but not in a collective sense. In contrast, many concepts have relatively high scores (e.g., $S_{(C_{4198})} = 33.75$ and $S_{(C_{649})} = 31.31$). The value of $S(c)$ ranges from 0 to 33.75, and the mean of $S(c)$ is 1.31. Table III shows the top ten concepts in the given data set.

B. Findings

The concept analysis technique makes it easy to mine common tags with respect to users in the data set. As discussed earlier, the largest user sets have the smallest sets of shared tags. From a folksonomic point of view, a collection of tags

in a large set of users can be used to identify common interests. In our case, C_{254}, C_{1601}, and C_{319} have 35, 34, and 33 users in the concepts, respectively, and each concept shares the following tags: $A_{254} = \{google, apple\}$, $A_{1601} = \{google, facebook\}$, and $A_{319} = \{google, twitter\}$. Thus, one would naturally assume that the tags *google*, *facebook*, and *twitter* are terms which identify common interests in this data set. However, this may not be the case if the tags are not popular in the context of the whole data set. Additionally, our findings suggest that the amount of users sharing a set of tags in a concept does not always imply that the concept is significant (i.e., that the tag set in the concept identifies common interests). As shown in Table III, the concepts, which are top ten scores in the given data, have less than nine tags. This provides some issues for thought as to why general and large folksonomies such as Delicious or Flickr are not always effective for retrieving the required information (i.e., finding interesting resources/people) based on a selection of tags.

Although many users share some of the same tags, we cannot make a definite decision that they have the same or similar interests. As shown in both Figs. 5 and 6, one can see that a conventional concept analysis is restricted to reflecting users' tagging practices. Based on the significance measure, we can refine a concept and can analyze the hidden structure of the concept. This measure may help us with suggesting interest groups for particular users. For instance, *Cloudlands* is involved in 677 concepts. Although it shares 11 tags with *Webby's World* in C_{32706} and just 5 tags with *Between the Lines (BTL)* in C_{32772}, the $S(c)$ for C_{32772} is higher (i.e., 21.73) than the one (i.e., 20.61) for C_{32706} (see Table IV). Therefore, although
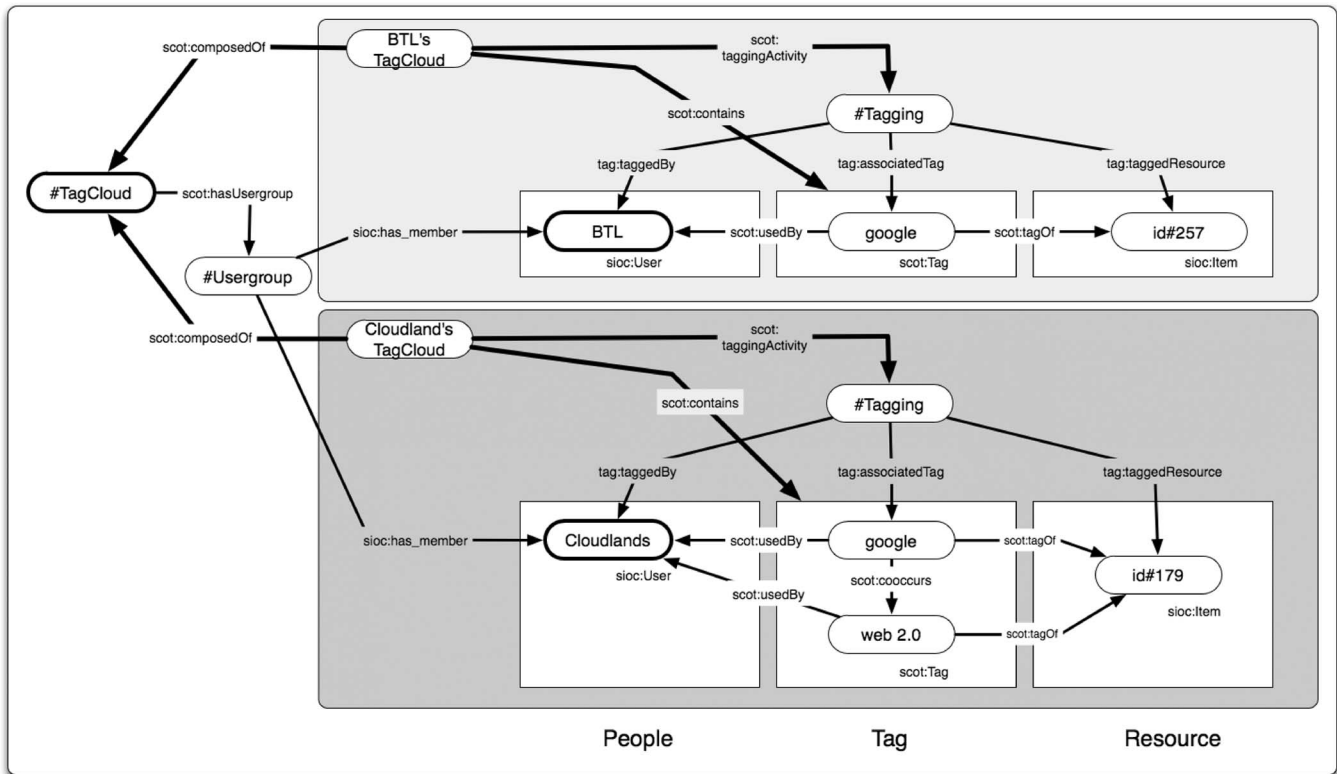
Fig. 7.   *Cloudlands* and *BTL*'s linked tag space. Both users' tag clouds are linked using the `scot:composedOf` property.

Cloudlands and *Webby's World* share more tags, the five tags shared with *Between the Lines* are relatively more important in the data set. Thus, we can refine the suggestion of interest groups for *Cloudlands*. That is, determining interest groups in a given data set may depend on users' tagging practices rather than on the number of tags that they have. However, if one wants to look for users who have a specific set of tags, the set of tags plays an important role for determining the interest groups.

One should note that the proposed approach can be considered as an intuitive method to mine the relationships between users and tags. Both the users and the tags in the data set, however, overlap in a large number of concepts. For example, the users in Table IV use some of the same tags despite them being in different concepts. Therefore, we need to integrate these concepts by using other similarity measures.

VIII. DESCRIBING TAGGING ACTIVITIES

Some form of explicit relationships is needed in order to represent tagging activities in the clusters. The beginning of this section notes that interlinked structures can be represented by aggregating tagging activities which are described as decentralized activities by different users. A tagging is composed of three entities (i.e., users, tags, and resources), and these entities have their own properties. Users may have a number of tagging activities composed of arbitrary relationships between tagging entities. As these users work constantly in their tagging events, the properties (e.g., occurrence of tags, co-occurring tags, etc.) of tagging entities should be updated.

In order to represent tagging activities for a cluster, tagging instances from users in the cluster must be aggregated and represented. For example, as shown in Table IV, C_{32772} consists of two users *Cloudlands* and *Between the Lines* and their tagging activities, and these relationships between tagging entities must be explicitly specified. The SCOT ontology is to design for describing folksonomies or collective tagging activities. This ontology, which extends Newman's model, introduces some approaches for representing these features.

A. Personomy

An individual tagging event is represented in *tag:Tagging* with some properties (i.e., *tag:associatedTag*, *tag:taggedBy*, *tag:taggedOn*, and *tag:taggedResource*) in Newman's ontology. This class is mapped to the `sioc:User` and `sioc:Item` to describe taggers (i.e., *Cloudlands* and *Between the Lines*) and the resources (blog posts). SCOT introduces some approaches defining collective and aggregated properties of tagging activities. For example, the *scot:taggingActivity* describes a relationship between *scot:TagCloud* and *tag:Tagging*. Thus, all tagging events for a user are collectively linked to an instance of the TagCloud class. Multiple tags in tagging events are aggregated to one unique *scot:Tag* if the names of the tags coincide. At the same time, occurrences of the tags are updated via two properties: *scot:ownAFrequency* and *scot:ownRFrequency*. Tagger (user) information is represented using Semantically Interlinked Online Communities (SIOC). *scot:Tag* is linked to *scot:TagCloud* via the *scot:contains* property.

```
PREFIX scot: <http://scot-project.org/scot/ns>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
SELECT ?whom, ?tag
WHERE {
    ?tc scot:hasUsergroup ?ug .
    ?ug sioc:has_member ?whom .
    ?tc scot:contains ?tag .
    ?tag scot:ownAFrequency ?frequency.
    FILTER regex(str(?tag), "web")
    && (?frequency > 30)
}
```

Fig. 8. SPARQL query to extract all the users using the tag *web* that the frequency is over 30.

B. Folksonomy

A TagCloud class in SCOT aggregates all tagging instances with their relevant information. At this level, tagging entities are represented with their collective features underlying their relationships. SCOT introduces the *scot:composedOf* property to link multiple TagClouds. With this property, *Cloudlands* and *Between the Lines*'s personomies can be interlinked. The tagging information of both users can be stored within one tag cloud (i.e., folksonomy) and simultaneously interlinked between them. This approach shows how a user-centric folksonomy for representing the interests of small groups or communities can be created in SCOT. It is also possible to adopt this example across sites or resources. Although both users have tagged different resources, the entities involved in the tagging activities are explicitly linked to each other, and the structure of the tagging data is consistent for sharing and reuse. Fig. 7 shows the merged folksonomy from both *Cloudlands* and *Between the Lines*'s personal tag clouds (i.e., personomies).

Fig. 7 shows the merged tag space from both *Cloudlands* and *BTL*'s personal tag clouds. This approach shows how a user-centric folksonomy to represent the interests of small groups or communities can be created by combining RDF vocabularies such as SIOC [7]. Although our example is limited to Weblogs, it is also possible to extend this approach across sites or applications. For instance, different users have tagging data on Flickr and Delicious. Our method can be adopted to create their customized folksonomy.

In addition, the SCOT ontology can be utilized by SPARQL, the query language for Semantic Web data. This ontology does not support the significance measure, which we proposed, directly. However, using Semantic Web technologies such as SPARQL queries, we can get minimal information to compute the significance of tagging data. Fig. 8 shows some typical example that can be posed on the SCOT and SIOC ontologies in the SPARQL query language. This query retrieves all users using the tag *web*.

IX. CONCLUSION AND FUTURE WORK

Mining clusters through tagging data has become quite popular. However, the majority of social sites providing tagging functionalities do not allow the reuse of user-generated data across other sites or applications. Under these circumstances, it is not easy to aggregate user interests from multiple sources.

In this paper, we have presented a novel approach that employs FCA and a significance measure based on the weight of the tags in the given data set. Our approach can be used to suggest new social relationships within a small-size group based on the users' interests by analyzing the tagging practices of individual users. We acknowledge that it is not straightforward to build a general-level folksonomy for the given data. However, interest groups can be built via a unique and complete lattice of formal concepts that takes into account a group of users and the tags that they assign.

In order to construct a more generic community for a given data set, we have to consider ways of linking or integrating among concepts. In the future, we will try to carry out the building of large community-level folksonomies by adopting the approaches presented in this paper. In addition, we also intend to improve the significance measure of a concept by taking into account co-occurrence relationships between individual tags. This might result in further refined tagging clusters which can consequently be employed to build interest-centric folksonomies.

REFERENCES

[1] Resource Description Framework (RDF): Concepts and Abstract Syntax, Technical report, W3C, Feb. 2004. (accessed Dec. 22, 2010). [Online]. Available: http://www.w3.org/TR/rdf-concepts/

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: Membership, growth, and evolution," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 44–54.

[3] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, OWL Web Ontology Language Reference, Technical report, World Wide Web Consortium, 2004. (accessed Dec. 22, 2010). [Online]. Available: http://www.w3.org/TR/owl-ref/

[4] T. Berners-Lee, Semantic Web Roadmap. World Wide Web: Design Issues, 1998. (accessed Dec. 22, 2010). [Online]. Available: http://www.w3.org/DesignIssues/Semantic.html

[5] T. Berners-Lee, Linked data. World Wide Web: Design Issues, Jul. 2006. (accessed Dec. 22, 2010). [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[6] A. Berry, J.-P. Bordat, and A. Sigayret, "A local approach to concept generation," *Ann. Math. Artif. Intell.*, vol. 49, no. 1–4, pp. 117–136, Apr. 2007.

[7] U. Bojārs, J. G. Breslin, V. Peristeras, G. Tummarello, and S. Decker, "Interlinking the social Web with semantics," *IEEE Intell. Syst.*, vol. 23, no. 3, pp. 29–40, May/Jun. 2008.

[8] J. G. Breslin and S. Decker, "The future of social networks on the Internet: The need for semantics," *IEEE Internet Comput.*, vol. 11, no. 6, pp. 86–90, Nov./Dec. 2007.

[9] K. K. Cetina and U. Bruegger, "The market as an object of attachment: Exploring postsocial relations in financial markets," *Can. J. Sociol./Cahiers Canadiens de Sociologie*, vol. 25, no. 2, pp. 141–168, 2000.

[10] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing tags over time," in *Proc. 15th Int. Conf. WWW*, 2006, pp. 193–202.

[11] J. Engeström, Why some social network services work and others don't or: The case for object-centered sociality, Apr. 2005. (accessed Dec. 22, 2010). [Online]. Available: http://www.zengestrom.com/blog/2005/04/why_some_social.html

[12] L. Freeman and D. White, "Using Galois lattices to represent network data," *Sociol. Methodol.*, vol. 23, pp. 127–146, 1993.

[13] B. Ganter and S. O. Kuznetsov, "Stepwise construction of the Dedekind–MacNeille completion (research note)," in *Proc. 6th ICCS*, 1998, pp. 295–302.

[14] B. Ganter and R. Wille, "Conceptual scaling," in *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*, F. Roberts, Ed. New York: Springer-Verlag, 1989, pp. 139–167.

[15] S. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198–208, Apr. 2006.

[16] T. Gruber, "Ontology of folksonomy: A mash-up of apples and oranges," *Int. J. Semantic Web Inf. Syst.*, vol. 3, no. 2, pp. 1–11, 2007.

[17] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *Proc. 16th Int. Conf. WWW*, 2007, pp. 211–220.

[18] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: Search and ranking," in *The Semantic Web: Research and Applications*, vol. 4011, *Lecture Notes in Computer Science*, Y. Sure and J. Domingue, Eds. New York: Springer-Verlag, Jun. 2006, pp. 411–426.

[19] H.-L. Kim, S. Decker, and J. G. Breslin, "Representing and sharing folksonomies with semantics," *J. Inf. Sci.*, vol. 36, no. 1, pp. 57–72, Feb. 2010.

[20] H.-L. Kim, S. Decker, and J. G. Breslin, "Social Semantic Cloud of Tags: Semantic model for folksonomies," *Knowl. Manage. Res. Pract.*, vol. 8, no. 3, pp. 193–202, Sep. 2010.

[21] H.-L. Kim, S. H. Hwang, and H. G. Kim, "FCA-based approach for mining contextualized folksonomy," in *Proc. ACM SAC*, 2007, pp. 1340–1345.

[22] H.-L. Kim, A. Passant, J. G. Breslin, S. Scerri, and S. Decker, "Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces," in *Proc. 2nd IEEE Int. Conf. Semantic Comput.*, Aug. 2008, pp. 315–322.

[23] H.-L. Kim, S. Scerri, J. G. Breslin, S. Decker, and H. Kim, "The state of the art in tag ontologies: A semantic model for tagging and folksonomies," in *Proc. Int. Conf. Dublin Core Metadata Appl.*, Sep. 2008, pp. 128–137.

[24] D. G. Kourie, S. Obiedkov, B. W. Watson, and D. van der Merwe, "An incremental algorithm to construct a lattice of set intersections," *Sci. Comput. Program.*, vol. 74, no. 3, pp. 128–142, Jan. 2009.

[25] S. Kuznetsov, "A fast algorithm for computing all intersections of objects in a finite semi-lattice," *Autom. Docum. Math. Ling.*, vol. 27, no. 5, pp. 11–21, 1993.

[26] S. Kuznetsov and S. Obiedkov, "Comparing performance of algorithms for generating concept lattices," *J. Exp. Theor. Artif. Intell.*, vol. 14, no. 2/3, pp. 189–216, Apr. 2002.

[27] C. Lindig, "Fast concept analysis," in *Working With Conceptual Structures—Contributions to ICCS 2000*. Aachen, Germany: Shaker Verlag, Aug. 2000, pp. 152–161.

[28] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, flickr, academic article, to read," in *Proc. 17th Conf. Hypertext Hypermedia (HYPERTEXT)*, 2006, pp. 31–40.

[29] B. Martin and P. Eklund, "From concepts to concept lattice: A border algorithm for making covers explicit," in *Proc. 6th ICFCA*, 2008, pp. 78–89.

[30] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Web Semant.*, vol. 5, no. 1, pp. 5–15, Mar. 2007.

[31] E. Norris, "An algorithm for computing the maximal rectangles in a binary relation," *Rev. Roumaine Math. Pures Appl.*, vol. 23, no. 2, pp. 243–250, 1978.

[32] L. Nourine and O. Raynaud, "A fast algorithm for building lattices," *Inf. Process. Lett.*, vol. 71, no. 5/6, pp. 199–204, Sep. 1999.

[33] U. Priss, "Formal concept analysis in information science," *Annu. Rev. Inf. Sci. Technol.*, vol. 40, no. 1, pp. 521–543, 2006.

[34] E. Quintarelli, "Folksonomies: Power to the people," presented at the ISKO Italy-UniMIB Meeting, Milan, Italy, Jun. 24, 2005.

[35] S. Rettberg, All together now: Collective knowledge, collective narratives, and architectures of participation, in *Proc. Digital Arts Culture Conf.*, Copenhagen, Denmark, 2005. (accessed Dec. 22, 2010). [Online]. Available: http://retts.net/documents/cnarrativeDAC.pdf

[36] C. Roth, "Co-evolution in epistemic networks—Reconstructing social complex systems," *Struct. Dyn.: eJ. Anthropol. Relat. Sci.*, vol. 1, no. 3, pp. 1–160, 2006.

[37] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. 17th Int. Conf. WWW*, 2008, pp. 327–336.

[38] G. Smith, *Tagging: People-Powered Metadata for the Social Web (Voices That Matter)*. Berkeley, CA: New Riders Press, Dec. 2007.

[39] J. F. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley, Jul. 1983, ser. Systems Programming Series.

[40] G. Stumme, R. Wille, and U. Wille, "Conceptual knowledge discovery in databases using formal concept analysis methods," in *Proc. 2nd Eur. Symp. PKDD*, vol. 1510, *LNAI*, J. M. Zytkow and M. Quafofou, Eds., Heidelberg, Germany, 1998, pp. 450–458.

[41] T. A. Tilley, "Tool support for FCA," in *Proc. 2nd ICFCA. Concept Lattices*, vol. 2961, *LNAI*, P. Eklund, Ed., 2004, pp. 104–111.

[42] P. Valtchev and R. Missaoui, Similarity-Based Clustering Versus Galois Lattice Building: Strengths and Weaknesses. (accessed Dec. 22, 2010). [Online]. Available: http://www.latece.uqam.ca/publications/valtchevmissaouipaper.pdf

[43] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[44] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," in *Ordered Sets*, I. Rival, Ed. Dordrecht, The Netherlands: Reidel, 1982, pp. 445–470.

Hak-Lae Kim received the B.S. degree in business administration and the M.S. and Ph.D. degrees in management information system from Dankook University, Yongin, Korea, and the Ph.D. degree in computer science from the National University of Ireland, Galway (NUI Galway), Ireland.

He was a Researcher with the Digital Enterprise Research Institute, NUI Galway, and his research efforts focused on creating social spaces based on tagging activities. He is currently a Senior Engineer with the Media Solution Center, Samsung Electronics Company, Ltd., Suwon, Korea. He is also the Director of the Web Science Workgroup in Korea. He is the Founder of the Social Semantic Cloud of Tags ontology and project, aims to provide semantic relations for tagging data, and has implemented the int.ere.st Web site for sharing tagging data.

John G. Breslin received the B.E. degree (with first-class honors) in electronic engineering and the Ph.D. degree from the National University of Ireland (NUI) (recently known as University College, Galway), Galway, Ireland, in 1994 and 2002, respectively.

He is a Lecturer of electrical and electronic engineering with the College of Engineering and Informatics, NUI, and a Researcher with the Digital Enterprise Research Institute (DERI), a Semantic Web research institute at NUI. He is also an Associate Researcher with the National Centre for Biomedical Engineering Science, NUI. He is a Coleader and the Founder of the Social Software Unit at DERI. He is also the Founder of the Semantically Interlinked Online Communities project (Wikipedia article), which aims to interlink online communities with semantics. He is a coauthor of the book *The Social Semantic Web* (Springer, 2009) and has over 100 peer-reviewed academic publications. His research interests include the Social Semantic Web and health care applications.

Dr. Breslin is the Tutorials and Workshops Chair for the forthcoming 2011 International Conference on Weblogs and Social Media and has served on over 75 committees for workshops, conferences, or journals.

Stefan Decker received the M.S. degree (with distinction) in computer science from the University of Kaiserslautern, Kaiserslautern, Germany, in 1995 and the Ph.D. degree (with distinction) in computer science from the University of Karlsruhe, Karlsruhe, Germany, in 2002.

He is a Professor with the National University of Ireland, Galway, Ireland, where he is also the Director of the Digital Enterprise Research Institute and the Leader of the Semantic Web Cluster within the institute. He published around 80 papers as books and journal, book, conference, and workshop contributions. His current research interests include the Semantic Web, metadata, ontologies and semistructured data, Web services, and applications for digital libraries, knowledge management, information integration, and peer-to-peer technology.

Dr. Decker co-organized around 35 scientific workshops and conferences and has edited several special issues of scientific journals.

Hong-Gee Kim received the B.S. degree in psychology from Korea University, Seoul, Korea, and the M.S. and Ph.D. degrees in artificial intelligence and philosophy from The University of Georgia, Athens.

He was a Professor of MIS in a business school. He is an Associate Professor of medical informatics and the Director of the Biomedical Knowledge Engineering Laboratory with Seoul National University, Seoul. He is also an Adjunct Professor with the College of Engineering and Informatics, National University of Ireland, Galway, Ireland. His primary research efforts are in building real-world applications by using ontology and semantic technologies in the e-health domain. He is currently leading various research projects concerning a semantic terminology server for electronic health record systems, a semantic search engine on a cloud computing environment, and a social semantic digital library system. He has published approximately 100 academic papers.