

The Use of Cognitive Authority for Information Retrieval in Folksonomy-Based Systems

Roberto Pereira, Sérgio Roberto P. da Silva
Universidade Estadual de Maringá, Departamento de Informática
roberto.uem@gmail.com, sergio.r.dasilva@gmail.com

Abstract

Folksonomy-based systems represent an important initiative to the information organization on the Web. It provides a way for its users to attribute meaning to the content of the information available. However, with the quick increase in the amount of information being published and the number of users categorizing it, an information overload is generated, whose quality cannot be assured. Aiming at reducing this problem, we propose the ascription of cognitive authority to recognize the information quality by qualifying its sources, i.e. its author. We applied this qualification in the information retrieval process in folksonomy-based system and through a simulation process we obtain results that demonstrate the viability of our proposal.

1. Introduction

A huge growth in the volume of on-line information available has been noticed. Studies [1] demonstrate that this amount of electronic information is likely to duplicate annually. This uncontrolled growth causes an information overload as harmful as its absence, which allied with the lack of mechanisms to assure the quality of the retrieved information makes the information retrieval process on the Internet more and more complex and of low quality.

Folksonomy based systems represent an initiative to help the processes of meaning construction and information organization. These systems allow users to categorize information on the web through the categorization of the objects available on the Internet. In these systems, users have active voice and are determining the “aboutness” of the objects they found [2]. However, as the number of users increases, the problem of information overload also appears in those systems once there is no control and no rules in the information categorization process.

In a discussion about folksonomy, Russell [2] approaches the ascription of **cognitive authority** through folksonomy aiming to recognize user’s competences or skills, and to find authorities on any subject matter. According to Russell, this ascription turns users (categorizer of objects) into objects, but in opposition to the categorization that tries to demonstrate “*what an object is*” — its *aboutness*, this categorization will try to demonstrate “*what an object knows*” — its *cognitive authority*.

Considering that the sources responsible for information categorizations in folksonomy-based systems are their own users and grounded by Wilson’s reasoning about cognitive authority [3], we came to conclude that the application of cognitive authority concept in a folksonomy-based system raises the relevance of the results and reduces problems of information overload. To validate our hypothesis, we simulated the ascription of cognitive authority and its use in the information retrieval process. We found out that prioritizing the categorizations accomplished by authorities provides better results than just retrieving information categorized by every entity and presenting them in chronological order.

This paper is organized in 6 sessions. In section 2, a brief basis of the folksonomy technique is exposed. Section 3 approaches the cognitive authority concept and discusses its ascription through folksonomy. Section 4 presents comparisons among our research and some related works. In section 5 we demonstrate the simulation of the information retrieval process. Finally, section 6 presents our conclusions and some directions for future works.

2. Folksonomy

Coined by Thomas Vander Wal [4] in 2004 as a result of the junction of the words “folks” and “taxonomy”, folksonomy represents the categorization done by people — a categorization in which users

work in the attribution of meaning and for the organization of the informational contents [6]. In the technical sense, the term folksonomy [4, 5, 6] represents an emerging technique applied to the categorization of bookmarks, photos, blog posts, physical location or any other object available on the Internet. The whole process is based essentially on the three pivots of folksonomy: the user —who does the categorization—, the object —which is categorized—, and the tags — that make the categorization labeling the object (Figure 1).

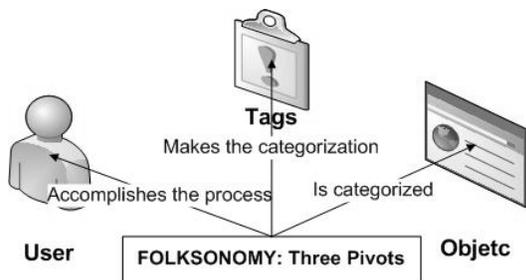


Figure 1. The three pivots of folksonomy.

Among the existent folksonomy-based systems, *Delicious* is considered the main reference [2, 6, 7]. It can be defined as a social bookmark manager that allows the URL of a website to be added, categorized with tags, and not just shared with other browsers and machines, but also with other users. Many of the advantages that the application of the folksonomy technique can provide are present in *Delicious*: i) the taxonomy used to carry out the categorization is free and built with social characteristics [8], which offers more flexible expressions and with a lower cost of maintenance; ii) categorization results exist on the surface, visible and useful to users [9] giving an immediate feedback to them; iii) the non-controlled vocabulary used attributes a great flexibility to the categorization process [2, 9]; iv) folksonomy contains individuals' knowledge structures about documents and contents [10] and it can reflect their users' vocabulary [9], which contributes to the "findability" of the information categorized by them.

In the same way that *Delicious* presents several advantages of folksonomy, as the system evolves, the disadvantages of this technique also become evident: i) the non-controlled nature of a folksonomy is fundamentally chaotic and it suffers from problems of organization and ambiguity [5]; ii) categorizations are based on tags, and tags do not work very well in terms of precision and retrieval [7]; iii) the lack of quality control of the categorizations done by the user jeopardize the information reliability [2] — once the only two criteria in which users can lean on are the

collective opinion of the majority of users or their own opinion, there is no way to evaluate, to expose and to recognize the authority of categorizers individually.

The central issue around Folksonomy is that it works because of users, and at the same time, its flaws also happen because of them. However, it does not mean that users are wrong or that they are not categorizing information correctly. What the users are categorizing is appropriated to them or to a small group, but it will not always bring contributions to the collective or to everyone [5]. There are areas in which users are more (or less) appropriate as information sources and it depends on: their capacity, the basis and the competence on the subjects they are categorizing, i.e., in their *cognitive authority*. However, authority is a subjective matter dependent on who grants it [3] in the same way that quality and relevance of results are [12, 13]. Thus, to obtain a better or worse result in an information retrieval process in a system like *Delicious* is more related to the users (sources of categorizations) than to the way the categorizations are accomplished.

3. Cognitive authority

The term "cognitive authority" was coined by Wilson [3] in his book "*Second-hand Knowledge: An Inquiry into Cognitive Authority*" to explain the kind of authority that influences people's thoughts and what they believe. The cognitive authority, as explained by Wilson and as justified by Rieh [11], represents the influence that the authority can cause in the thinking way of an individual, because this individual judges it proper, worthy of credit and trust — in Wilson's words [3] this kind of authority defines "*who knows what about what*".

In a system like *Delicious*, where the number of users is impressive and the amount of produced information is even higher, it is very difficult for users to identify and separate the information that is both interesting and reliable. Rieh & Belkin [11, 12] affirm that the application of cognitive authority concept can aid in the process of information organization and retrieval. It is known that specialists [3] not only use quality documents, but also develop more organized classification schemes and relationship among concepts. So, we can conclude that to recognize its categorizers' abilities, granting cognitive authority to them, can aid in the retrieval process in a folksonomy-based system through the emphasis on information categorized by *who knows what they are talking about*.

Russell [2], when making an approach about folksonomy, mentions Rieh & Belkin's works [11] and suggests that cognitive authority ascription can be

accomplished through folksonomy. Thus, grounded on Wilson’s theory indications [3] that in the recognition of cognitive authorities people have indirect bases and reasons to justify the authority judgment, we can conclude that it is plausible to categorize users-sources recognizing their authorities and, then, to attribute different weights to the information categorized by those authorities in the information retrieval.

3.1. Cognitive authority and quality

Regarding the relationship between cognitive authority and quality, we can say that information quality is **subjective** because it depends on who is making the judgment [12]. Besides, it is also **relative** and **situational**: a) relative, since certain information can be adequate for a specific objective and completely inappropriate for other; b) situational because its judgment can be changed in the course of time, with people’s knowledge evolution and with their needs. These same considerations can be applied to authority [11] that people ascribe to others in their social groups. Our considerations on the context of information retrieval is that as the judgment for the authority ascription to a person has common grounds with the judgment of information quality and relevance [11, 12], from the point of view of who granted this authority, a person considered an authority in certain subject matter tends to have quality information in this subject. This way, we considered that if folksonomy-based system’s users prioritize information coming from people who they granted authority data of better quality and relevance will be presented to them.

3.2. Cognitive authority through folksonomy

Once in a folksonomy-based system it is the users who attributes meaning to the information, our proposal consists of transforming these users (whose accomplish the categorizations) in objects that are prone to be categorized, using a **meta-categorization** process. As users can represent an individual or a group of people, a company or any other organization, we preferred to use the term **entity** to define them. These entities will be categorized with respect to their competences in a process of cognitive authority ascription. However, while the common categorization describes what the object is about, the meta-categorization of an entity describes what the object knows [2]. So, with this process, it is possible to identify and to prioritize information categorized by entities that are considered authorities in the subject matter. Figure 2 shows the cognitive authority

ascription to an entity with the tags: “usability”, “folksonomy” and “design”.

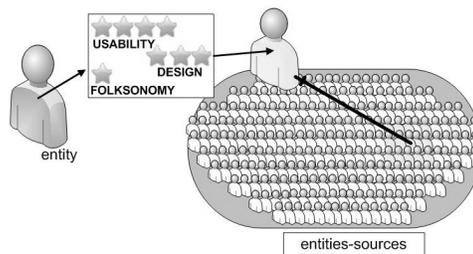


Figure 2. The cognitive authority ascription.

The categorizations for the cognitive authority ascription create a social network between the user and their entities (called **authority network**), identifying and turning explicit the relationships of reliability and recognition already existent in the “real world” [12]. When the authority networks are linked to each other they build a **chain of authority** [2, 14] (see Figure 3).

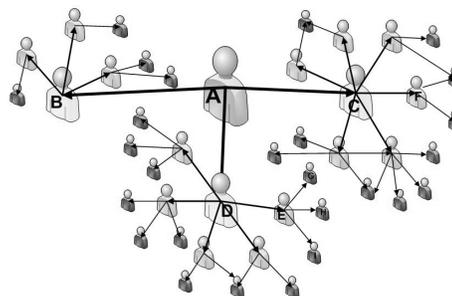


Figure 3. A chain of authority.

The chain of authority is the greatest benefit from the cognitive authority ascription in a folksonomy-based system. It makes possible to extract useful information and to carry out more complex processes, among which we highlighted: i) the discovery of authorities and the recognition of the entities’ competences [2], ii) the sharing of entities’ knowledge and vocabulary, iii) the reduction of problems such as ambiguity, synonymy and polissemly in the information retrieval, iv) the possibility to identify authorities and to prioritize their categorizations for the process of information retrieval, and so on.

As a source of information in our experiments, we choose to use Delicious folksonomy system, since in it the task of finding useful and reliable bookmarks demands an effort that tends to increase as the amount of classified information increases, since in the process of information retrieval it is considered every categorization done by the whole group. The result of the retrieval is presented to the users in reverse chronological order (popularity is also considered as a

sub-criterion). An alternative to this kind of presentation is to order the results based on the categorizations realized by authorities, using the chain of authority to allow the identification of authorities and their differentiation by popularity, weights and proximity (between entity and authorities). These two types of retrieval were used in the experiments.

4. Related works

Although none of the works in the literature that we found are addressed to the use of cognitive authority in information retrieval, many of them possess close focus or they served as a base for we to arrive at our proposal. We emphasize the “Contextual Authority Tagging” project [2] as the closest work to our proposal and as the initial base for our studies and reflections. But, while Russell [3] just proposes the cognitive authority ascription through folksonomy for the recognition of entities’ competences and for the identification of authorities, we get the same results and go beyond proposing the use authorities to improve the results obtained in the information retrieval process in folksonomy based systems.

Another work is the Fringe Contacts Project [15], which embodies the development of an application for make it possible for workers to characterize their colleagues through the attribution of tags, aiming to reduce the maintenance cost of the professional relationships in an organization. There are other works related to the trust in social networks, as [13], that approach related subjects, however, they do not apply folksonomy technique and do not use the concepts of the cognitive authority theory.

5. The simulation of authority ascription

Goldback in [19] says that “...naturally occurring networks take a long term to gain the large number of users, and the topological properties are fixed...”, concluding that simulation is a viable alternative for studying such systems. Thus, to validate our hypothesis that “the application of the cognitive authority concept in folksonomy-based systems elevates the precision of the recovered information and softens the impact of the information overload in those systems”, we conducted a simulation experiment that is presented in this section.

5.1. Methodology

To make it possible to do an accurate analysis with the generated data, the simulation process needs to

follow a well defined methodology and should respect an established group of rules. The steps executed for the simulation done are presented in Figure 4.

The generation of the main pivots of a folksonomy-based system (entities, objects, and tags) takes place in the three initial steps. The 1° step consists in defining the set of tags used during the simulation process. Those tags represent the vocabulary of terms that the entities use to the objects categorizations and which will be used also for the cognitive authority ascription. The set of tags defined in this experiment has 250 tags.

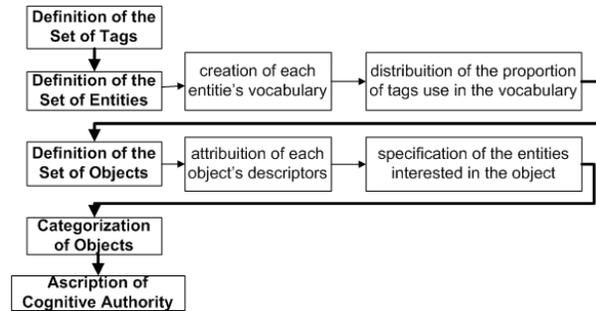


Figure 4. Steps of the simulation process.

In the 2° step, the set of entities (i.e., users) is defined. Entities are used for the creation of categorization for the objects and for the cognitive authority ascription. The defined set for this simulation has a population of a 100 entities and, for each entity it was generated a vocabulary composed by 20 tags randomly selected from the set of tags. It was attributed a differentiated usage proportion to the tags, to produce the long tail behavior of tags observed in real systems — to generate this distribution it was applied a variation of the Pareto Principle [16].

The 3° step embodies the creation of the set of objects. Objects represent the information that can be categorized by entities for future retrieval or for any other reason. The set is composed by 1000 objects and for each object it was attributed 10 tags randomly selected from the set of tags to be used as object descriptors. It was selected a quantity between 0 and 20 entities to be related to the object, indicating entities to which the object would be recommended in the information retrieval process.

In the 4° step the categorization of objects happens. The procedure consists in building a relationship, modeled by a tuple {entity, tags, object}, representing a process in which folksonomy-based system’s users categorize an object attributing tags to it. For this simulation, each entity has, on average, 95 objects categorized with a random quantity of tags (minimum 2, maximum 5) — the objects are randomly selected among the 1000 generated objects. The choice of tags

obeys the differentiated distribution which generates the long tail, and it prioritizes tags that are present in the entities' vocabulary as well as the ones present in object's descriptors.

The last step consists of a procedure in which each entity grants authority to a random number of other entities (between 1 and 10). The authority is ascribed through 5 tags randomly selected from the entity's vocabulary that receives the authority (terms located in the long tail are excluded). Each tag is associated to a weight which is measured by a value between 1 and 5 stars, which represents the distinction of authority levels. The generated relationship is represented by a unique tuple {entity1, tags, entity2} between "entity1" and "entity2", indicating that entity1 granted authority to entity2 with the set of tags "tags".

5.2. Information retrieval

The simulations *by itself* consists of the accomplishment of three steps which were executed for each one of the 100 entities, resulting in approximately 25.000 executions, they are: i) the objects categorized by each one of the 250 tags used in the simulation are retrieved, following the chronological retrieval process; ii) for each retrieval process executed, the order that the recovered objects appear among the results is verified, and also who are the entities which categorized them; and iii) it is identified which entities are authorities in the subject and what is the weight of those authorities, then, the order of the objects are rearranged according to the amount of authorities which categorized them, considering the authorities' weights.

5.3. Results analysis

The data obtained with the simulation of the retrieval of all objects by every tag demonstrates all possible positions that those objects can appear when executing retrieval for each one of the tags. It also give us a basis to identify, for each one of the 1000 objects, the tags that best retrieve a certain object using the chronological retrieval and also in the process where authorities are prioritized. For example, suppose we want to retrieve *Object993*. The best position it appear in the chronological retrieval is 15th position when retrieved using the tag *learning*. For the same tag, but using the prioritization by authorities it appears in the 13th position. On the other hand, it appears in the 128th position using the tag *book* in the chronological process, but it appears in the 1st position using the prioritization by authorities'. Thus, in this experiment

the focus is on the position that the object appears in the retrieval process not in the tag.

Table 1. Chronological retrieval process.

Position	Chronological (objects)	(%)	Authority (objects)	(%)
1 st — 10 th	251	25.1	244	24.4
11 th — 20 th	247	24.7	235	23.5
21 st — 30 th	182	18.2	170	17.0
31 st — 40 th	151	15.1	144	14.4
41 st — 50 th	88	8.8	86	8.6
51 st ...	81	8.1	119	11.9

Table 1 demonstrates the results obtained with the tags that best retrieve the objects in the chronological retrieval. We can see that, by this process, exactly 25.1% of the objects have, at least, one tag that places them among the first 10th positions. Considering the first 20th positions this percentage raises to almost 50% of the objects. Thus, according to the simulations, approximately half of the set of objects can be found among the first 20th results. On the other hand, if the authorities' categorizations are prioritized in the retrieval process, using the same tags, the percentage of objects found between the first 10th and 20th positions is reduced to approximately 24 and 48%, respectively. This reduction is not significant and in this case it happens because in the chronological retrieval the tags belonging to the long tail are those that best position the objects in the retrieval, and they are seldom used by authorities in the target subject. The most important finding of this results is the fact that the use of the authorities prioritizations does not make the retrieval process worst that the conventional retrieval process.

Table 2. Authorities' retrieval process.

Position	Chronological (objects)	(%)	Authority (objects)	(%)
1 st — 10 th	125	12.5	890	89.0
11 th — 20 th	91	9.1	94	9.4
21 st — 30 th	94	9.4	14	1.4
31 st — 40 th	88	8.8	1	0.1
41 st — 50 th	100	10.0	1	0.1
51 st ...	502	50.2	-	-

Table 2 shows data where the retrieval prioritizes categorizations done by authorities. It also takes into account the best position where the object appears. In this case, 89% of the objects can be found among the first 10th results. This percentage rises to 98% of the objects considering the first 20th results. Thus, in retrieval which prioritizes information categorized by

authorities only 2% of all objects do not possess any tag that positions them among the first 20th results.

When considering the positions in which these objects are placed in the chronological retrieval, using the same tags as above, a big difference is seen, since only 21.6% of the objects are presented among the first 20th results. Unlike what was exposed in Table 1, this difference in the objects' positioning is very significant and it happens because the objects presented on the best positions are those categorized by the greatest number of entities, considering its weighting — with tags used many times and that do not belong to the long tail. Thus, in the chronological information retrieval the same tag used to categorize several objects hardly will maintain a good objects positioning.

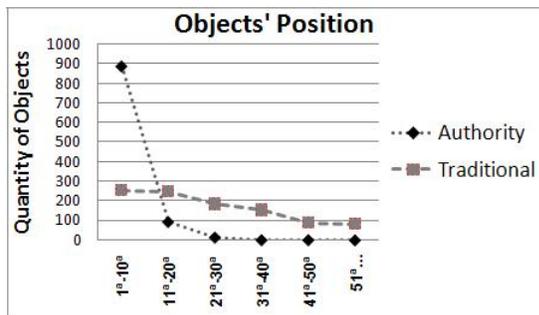


Figure 5. Objects positioning in the retrieval process experiments.

In Figure 5, we present a graph for the number of the objects in each partition of the positioning for the chronological retrieval, shown in Table 1, and for the retrieval considering authorities, shown in Table 2. So, we can conclude that the prioritization of authorities' maximize the chances of an object to appear among the first results of a retrieval giving in a great advantage over the chronological retrieval process. This conclusion can be supported by the great difference shown in the number of objects that are not retrieved by the chronological process (74,9%) against the number of objects that is lost in the authorities' prioritization process (11%), considering the 10th initial positions. The first 10th positions are very important since in a research carried out by iProspect [17] it is shown that most of the users consider only the first results presented in an information retrieval process. Consequently, useful results that are not presented in the first positions will never be found by the users in the chronological process.

Besides the analysis exposed above, it is also possible to carry out analysis on the **chain of authorities** and to extract information, as already mentioned by Russell [2], about “Who is an authority

concerning topic Y?”, and “What is user X an authority on?”. Simultaneously, there are other results related to the chain of authorities that should be exposed. Figure 6 demonstrates the distribution of authorities on an example chain.

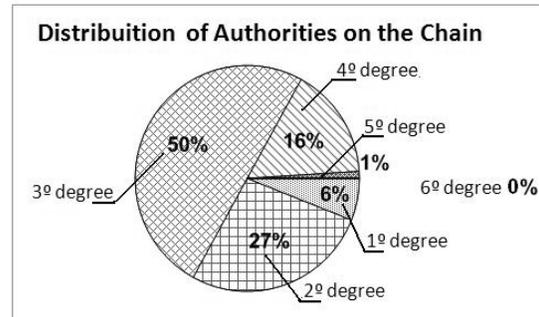


Figure 6. Authorities on the chain.

According to the percentage of the first degree of separation an entity possesses, on average, six direct authorities (their network of authorities). The graph of Figure 6 also demonstrates that any entity is linked to approximately 83% of other entities until the 3rd degree of separation and that only 1% of the entities are beyond the 4th degree — there are no occurrences of entities whose degree of separation is higher than the 5th degree. This data demonstrate that an entity is relatively close to most of their authorities, a result that also agrees with the studies of Yu [18] on the degrees of separation in social networks.

At the moment of the data generation, the objects of interest of each entity were defined. Thus, identifying the distribution of authorities on the chain is also possible to identify the distribution of each entity's objects of interest.

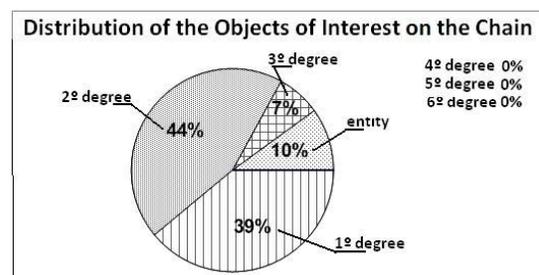


Figure 7. Objects on the chain.

The graph of the Figure 7 shows that, on average, 83% of the objects of interest of each entity are found up to the 2nd degree of separation on the chain. Comparing those results with data shown in Figure 6, we noticed that 33% of the authorities are placed up to the 2nd degree of the chain. Thus, it determines that 33% of authorities did the categorization of 83% of the

objects of interest of each entity. These results do not allow us to affirm that common entities do worse categorizations, but they demonstrate that authorities that are close to entities categorize most of their objects of interest. Therefore, we can say that authorities' categorizations presents better contributions to the process of information retrieval.

6. Conclusions

The results of our simulations demonstrate that to prioritize categorizations carried out by authorities contributes to a larger number of objects been presented among the first results of the retrieval. This makes it possible for objects condemned to appear among the last results to be located more easily.

Thus, besides improving the objects positioning, considering that contents categorized by authorities possess more reliability and quality to who grants the authority, we can affirm that, for those who grants or agrees with the entities' authority, the contents presented among the first results in the information retrieval are more reliable and of better quality, since at least the categorization sources can be identified and analyzed. In this way, we demonstrate that our hypothesis is valid.

Also, the construction of the chain of authorities allowed us to obtaining useful information such as: the possibility of searching for authorities, the recognition of competences, the way authorities are distributed, the more popular authorities, and possible recommendation of authorities and contents.

Currently, we are working on the *CAW project* (Cognitive Authority on the Web) for the development of a folksonomy-based system which makes the cognitive authority ascription possible to better evaluate its capabilities. This system is being projected to work with information categorized in other systems (i.e. *Flickr* and *Delicious*) and it will be available on the Internet. Besides the technical questions, social, ethical and human-computer interaction aspects are being thoroughly considered in the project, as they are crucial for its success and its acceptance.

7. References

- [1] Lyman, P. How Much Information? University of California, USA, 2003. <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003>
- [2] Russell, T. Contextual Authority Tagging: Cognitive Authority Through Folksonomy. University North Carolina, 2005.
- [3] Wilson, P. Second-hand Knowledge: An Inquiry into Cognitive Authority. Westport, Greenwood Press, 1983.
- [4] Wal, T. V. Folksonomy <http://www.vanderwal.net/random/entrysel.php?blog=1622>. [Jan. 12, 2008]
- [5] Mathes, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. University of Illinois. 2004.
- [6] Shen, K. and Wu, L. Folksonomy as a Complex Network. Department of Computer Science. Shanghai, Fudan University, 2005.
- [7] Riddle, P. Tags: What are They Good For? School of Information, University of Texas, 2005. http://www.ischool.utexas.edu/~i385q/archive/riddle_p/riddle-2005-tags.pdf.
- [8] Ohmukai, I. A Proposal of Community-based Folksonomy with RDF Metadata. Galway : 4th International Semantic Web Conference (ISWC2005), 2005.
- [9] Sturtz, D. N. Communal Categorization: The Folksonomy INFO622: Content Representation. Dec. 2004.
- [10] Wu, Harris. Harvesting Social Knowledge from Folksonomies. 17th Hypertext and hypermedia, 2006.
- [11] Rieh, S. Y. and Belkin, N. J. Interaction on the: Scholars' Judgement of Information Quality and Cognitive Authority. 63st ASIS Annual Meeting, 2000, pp. 83-87.
- [12] Rieh, S. Y. Cognitive Authority. Medford, NJ: Information Today, Theories of information behavior: A researchers' guide, 2005. pp. 83-87.
- [13] Golbeck, J. and Hendler, J. Inferring Binary Trust Relationships in Web-Based Social Networks. ACM Transactions on Internet Technology, New York, 2006.
- [14] Matsuo, Y. et al. Finding Social Network for Trust Calculation. Japanese National Institute of Advanced Industrial Science and Technology (AIST). 2004.
- [15] Farrell, S. and Lau, T. Fringe Contacts: People-Tagging for the Enterprise. IBM Almaden Research Center, WWW2006: Edinburgh, Scotland, 2006.
- [16] Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. Statistical Mechanics. Contemporary Physics 46, 323-351, May 29, 2006.
- [17] iProspect. iProspect Search Engine User Attitudes. Marketing Firm. 2007. <http://www.iprospect.com/premiumPDFs/iProspectSurveyComplete.pdf>. [October, 15, 2007]
- [18] Yu, B. and Munindar, P. Singh Searching Social Networks [Journal]. International Conference on Autonomous Agents, Melbourne, Australia : ACM, 2003.