Personalized Information Delivering Service in Blog-Like Digital Libraries

Jason J. Jung[1,2]

[1] Department of Computer & Information Eng., Inha University
253 Yonghyun-Dong, Nam-Gu, Incheon, Korea 402-751
[2] INRIA Rhône-Alpes
ZIRST 655 avenue de l'Europe, Montbonnot
38334 Saint Ismier cedex, France
j2jung@intelligent.pe.kr

Abstract. With increasing concerns about the personalized digital libraries (e.g., blogs), people need to share relevant information and knowledge with other like-minded users. In this paper, we aim at building a grid environment for information recommendation, in order to support users' information searching tasks. By thoroughly analyzing the social linkage and social interaction patterns, we want to extract the meaningful relationships between the unknown users by co-occurrence analysis. Therefore, social grid environment can be constructed by aggregating a set of virtual hubs discovered from the hidden connections. For implementation and evaluation, we exploit the proposed method to blogosphere. The BlogGrid framework is proposed to provide efficient information pushing service to bloggers without requesting any user intervention.

1 Introduction

Since personalized information spaces (e.g. blogs[1]) were introduced, the usage has been spread to many fields [1,2]. More importantly, each blogger can make explicit connections with other acquaintances (e.g., family members and friends), and take social activities such as information (or knowledge) sharing, responding to answers, and referring to further information [3]. Thus, they can organize communities in a form of socialized information space. On blogosphere, however, we have been facing on two problems. Firstly, a large amount of *information overwhelming to users* means that users are getting much more overloaded to search for relevant information related to a certain topic and, more specifically, the like-minded users. Next problem is *network isolation phenomenon*. Communities tend to be initially organized from the private relations among users. But, the number of members in this community may keeps constant over time, because social interactions between different community members are rarely occurred. We find out the limitation of information flows caused by not only the community policy for protecting the privacy of members but also the topological isolation of social structure.

[1] This is the shorten expression of "weblogs," and this paper uses only "blog".

S. Sugimoto et al. (Eds.): ICADL 2006, LNCS 4312, pp. 163–172, 2006.

In order to solve these problems, we exploit the grid computing paradigm [4] to personal information spaces and users, supporting an efficient framework of information and knowledge sharing between heterogeneous sources. Thereby, each user's behaviors should be captured and analyzed to extract meaningful context (e.g., what topics he is (or they are) interested in and, more exactly, what he has been trying to search for). In particular, we implement a grid environment (called BlogGrid [5]), in order to provide information pushing service on blogosphere. The goal of this service is efficient information diffusion on blogosphere. It means that a certain information should be delivered to the bloggers who have been looking for the information as quickly as possible, regardless of the social distances and topologies.

We exploit co-occurrence analysis between user activities to predict the relationships between the corresponding bloggers. Virtual hubs built by shortest paths are applied to adapting the strength of hub centrality of personal information space.

The outline of this paper is as follows: Section 2 describes modeling user behaviors on blogosphere and analyzing social interactions. In section 3, we explain the main steps of BlogGrid framework. Section 4 addresses implementation with experimental results and some important issues. Finally, in section 5, we draw some conclusions and explain future work.

2 Analyzing Social Interactions Between Users

In order to recognize the relationships between users, we want to model each user's behaviors and measure the distance between them. We can fomulate the behavior of users on social network $\mathcal{SN} = \{U_1, U_2, \ldots, U_N\}$ where N is the total number of users participating in this social network.

2.1 Modeling Behaviors on Blogspace

We assume that each action taken by users have implicit meaning and be able to be applied to extract useful information about their preferences. Also, this information should be used to recognize the relationships between users. We classify the behaviors $\mathcal{B} = \{\mathcal{A}, \mathcal{L}, \mathcal{N}, \mathcal{R}, \mathcal{C}\}$;

1. *Posting articles* $\mathcal{A}$. Users input various information and enrich their own spaces. This action explicitly represents the corresponding user's preferences. A set of articles of U_i is denoted as $\mathcal{A}_i = \{a_1, a_2, \ldots, a_K\}$.
2. *Linking with other blogspaces* $\mathcal{L}$. It is for explicit construction of a social network. More importantly, in order to share information about a particular topic, people can organize a community and actively participate it. We assume this link to be directional. Thus, the neighbors of U_i is represented as $\mathcal{L}_i$ combining two links $\mathcal{L}^{\mathcal{O}}$ (pointing to others) and $\mathcal{L}^{\mathcal{I}}$ (pointed by others).

$$\mathcal{L}_i = \mathcal{L}_i^{\mathcal{O}} + \mathcal{L}_i^{\mathcal{I}} = \{U_p\}_i^{\mathcal{O}} + \{U_q\}_i^{\mathcal{I}} \tag{1}$$

Reversely, Equ. 1 can also be represented as $\{U_m | U_i \in \mathcal{L}_m^{\mathcal{I}}\} + \{U_n | U_i \in \mathcal{L}_n^{\mathcal{O}}\}$.

3. *Navigating* $\mathcal{N}$. In order to get relevant information within personal information spaces, people should visit other spaces, by the following two methods;
   - **Random browsing**. A blogger randomly jump into other blogspaces.
   - **Accessing to neighbors on social network**. By referring to the list of neighbors, a blogger easily moves into the blogs of his neighbors. In the same way, he can access to the blogs of his neighbors' neighbors.
4. *Responding* $\mathcal{R}$. A user can respond to a certain information by two ways; *i*) *comment* and *ii*) *trackback*[2]. Moreover, not only a free-text sentence but also numeric rating format (e.g., from 0 to 5) and voting format (e.g., "Yes" or "No") can be applied to reflect the degree of the corresponding user's interests and opinions. Another feature is that the responding can be nested. It means that bloggers can respond to a certain comment already attached to articles. Thus, the responses by user U_i is given as $\mathcal{R}_i = \{r_i^{a_\alpha}, r_i^{r_\beta} | a_\alpha \in \mathcal{A}_j, r_\beta \in \mathcal{R}_k\}$ where $\mathcal{A}_j$ and $\mathcal{R}_k$ mean a set of articles by user u_j and a set of responds by user u_k, respectively.
5. *Categorizing blogspaces* $\mathcal{C}$. Each blog can be labeled with topics, which is represented as hierarchical path on taxonomies. It is represented as $\mathcal{C}_i = \{c_i | c_i \in \mathcal{O}\}$ where $\mathcal{O}$ is a hierarchical taxonomy (e.g., ODP and TopicMap).

2.2 Virtual Hub Generation by Co-occurrence Analysis

Co-occurrence analysis between their behaviors can generate a set of virtual hubs ($\mathcal{VH}$). It is play a role of alternative channels to communicate among potential users. In our previous work [6], we found out that responding behaviors $\mathcal{R}$ has shown more significant effects rather than other ones. For the moment, this paper is considering only responding behaviors $\mathcal{R}$ to measuring co-occurrence between two users. (Of cause, the others' co-occurrence patterns can be easily discovered, as extending according to $\mathcal{B} = \{\mathcal{A}, \mathcal{L}, \mathcal{N}, \mathcal{R}, \mathcal{C}\}$.) Two arbitrary users who commonly respond to a certain article a_β ($\in \mathcal{A}_\alpha$) of personal space of user U_α are regarded as like-minded users. For example, let two users U_A and U_B respond to a certain article in U_C's blog. Even though they are geometrically far away on social network, their preferences are probably similar with each other. In order to make them closer, we formulate some measurements. Simply the (geodesic) distance $\mathcal{D}$ between two users can be measured by

$$\begin{aligned} \mathcal{D}(U_i, U_j) &= shortest\_path(U_i, U_j) \\ &= \min(path(U_i, U_j)). \end{aligned} \tag{2}$$

It is the length of linked path between U_i and U_j, and it is computed by counting all the edges on the shortest path.

Next, we need to recognize the "power" from social linkage patterns in $\mathcal{SN}$, we deploy the hub ($\mathcal{HUB}$) and authority ($\mathcal{AUTH}$) weighting scheme [7] to measure the centralities of every users. Similar to the betweenness [8], hub weight indicates the structural position of the corresponding user. It is a measure of

[2] Movable Type. http://www.sixapart.com/movabletype/.

the influence that people has over the spread of information through the network. On the other hand, authoritative weight is a measure of the information quantity that people occupies. These weights can be iteratively computed by $\mathcal{HUB}_{t-1}^{\langle U_i \rangle} \leftarrow \sum_{U_j : \in \mathcal{L}_i^{\mathcal{O}}} \mathcal{AUTH}_{t-1}^{\langle U_j \rangle}$ and $\mathcal{AUTH}_t^{\langle u_i \rangle} \leftarrow \sum_{U_j \in \mathcal{L}_i^{\mathcal{I}}} \mathcal{HUB}_{t-1}^{\langle U_j \rangle}$ where $\mathcal{AUTH}_0$ is initialized with 1. Additionally, for the "equilibrium" values of the weights, iteration process should be conducted.

Now, we propose a novel similarity (or distance) measurement based on the "hub" weights of neighbors. By using the response r^{a_β} co-occurred by a group of users U_{CO}, the hub similarities $\mathcal{S}_{\mathcal{HUB}}$ is represented as a square matrix of which size is $|U_{CO}| \times |U_{CO}|$, and each element is given by

$$\mathcal{S}_{\mathcal{HUB}}(i,j) = \frac{\mathcal{D}(U_i, U_j)}{\sum_{U_k} \mathcal{HUB}^{\langle U_k \rangle}} \times \frac{1}{\exp(|U_{CO}|)} \tag{3}$$

where $U_k \in \mathit{shortest\_path}(U_i, U_j)$ and $U_i, U_j \in U_{CO}$. If $\mathcal{D}(U_i, U_j)$ is less than two (e.g., the diagonal elements), $\mathcal{S}_{\mathcal{HUB}}(i,i)$ is zero. We assume that the longer distance and the lower hub weights between two users indicate the higher similarity. In the second term, we express that the higher number of co-occurring users exponentially decays the hub similarity among them. The $\mathcal{VH}$ among the corresponding users, therefore, is generated, as shown in Fig. 1. Then, as

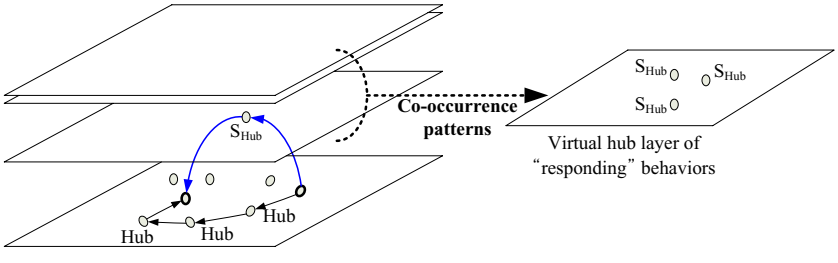

Fig. 1. Measuring hub similarities with common responses and aggregating with others

accumulating the $\mathcal{VH}$'s generated by all co-occurred response patterns, we can construct the $\mathcal{VH}$ layer ($\mathcal{VHL}$). In case of this paper considering only patterns from "response" behaviors, it can be denoted as $\mathcal{VHL}_{\mathcal{R}}$. This is simply based on matrix concatenation tasks, but we have to focus on the duplications. Here, for adaptability, $\mathcal{S}_{\mathcal{HUB}}$ between the repeated pairs of users U_i and U_j have to be reinforced twice by

$$\mathcal{S}_{\mathcal{HUB}}(i,k)' = \mathcal{S}_{\mathcal{HUB}}(i,k) \times \begin{cases} 1+\Delta & \text{if } k = j; \\ 1 - \frac{\Delta \times \mathcal{S}_{\mathcal{HUB}}(i,k)}{|U_{CO}|-1} & \text{otherwise;} \end{cases} \tag{4}$$

and $\mathcal{S}_{\mathcal{HUB}}(k,j)'$, which can be obtained by exactly opposite way to Equ. 5. Δ is the coefficient of learning rate in $(0,1]$, and $U_{CO} = U_{CO_1} \cup \ldots \cup U_{CO_N}$. Of cause,

instead of the adaptation scheme based on *Hebbian* learning that we apply for simplicity, the other various learning methods are exploitable.

3 Information Delivering on BlogGrid

Each information should be pushed (more exactly, notified) to the users who are interested in and searching for it. A facilitator agent has to conduct this service, and it is simply composed of two main processes; i) ranking-based organization of user group, and ii) propagation of relevant information. Thereby, when a response behavior $\mathcal{B}_i$ (e.g., in this paper, $r_i \in \mathcal{R}_i$) by user U_i is detected, this facilitator should predict the candidate users U_x who is potentially close to U_i by ranking $\mathcal{S}_{\mathcal{HUB}}(i, x)$ from the $\mathcal{VHL}$. As measuring the hub similarity between two users by the given response patterns, we can infer their relationship and make a decision whether they should be participated in a same article together or not. Especially, this task should be transparently done, which means it has to be performed without user's intervention and realization.

Along with the established $\mathcal{VHL}$, the relevant pieces of information should be pushed actively. Information pushing service proposed in this paper is remote and synchronous because this is based on web-based blogging system and information is promptly propagated according to the participant's interests extracted from his own behaviors. Figure 2 shows the whole system architecture.

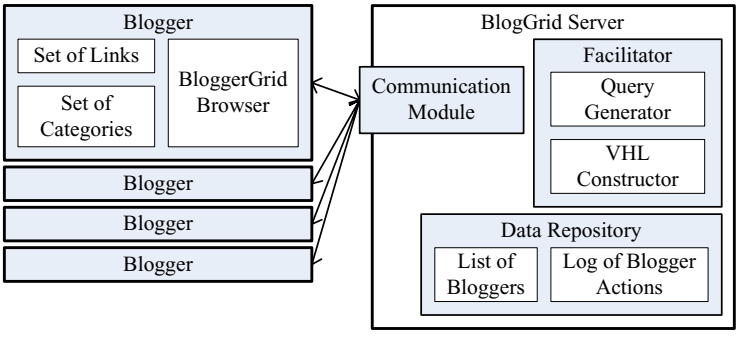

Fig. 2. System architecture of BlogGrid

It consists of two main parts, which are a facilitator located between the users and the client-side blogspace browser that communicates with the facilitator. We embed autonomous and proactive agent module into this system. Every communication between agents is conducted, regardless of user's interventions. Also, while browsing blogspaces to search information, users can be "implicitly" recommended from the facilitator in the following two ways:

- *By querying specific information for the facilitator.* After the information about a particular topic is requested, the facilitator can determine who has the maximum weight value of that topic by scanning his yellow pages.
- *By broadcasting new information of like-minded bloggers from the facilitator.* Every time a user responds a new comment, this fact, after normalization, is sent to the facilitator. Users thereby can obtain information related to the common concepts in their own preferences from neighbors.

Each blogger needs personal agent module. This agent initializes and manages the preference of the corresponding blogger based on blogspace repository. Through personal agents' reporting responding activities of bloggers, the facilitator agent can automatically generate queries and recommendations.

4 Experimental Results

For evaluating the performance of adaptive socialization on BlogGrid, we randomly divided the students into two groups, named *Alpha* (75 students) and *Beta* (75 students). Only students in a group Beta could use BlogGrid browser. The students within the same group were able to make linkages with their acquaintances for constructing social networks, access to the other blogspaces, and share information with each other for ten weeks (from 14 February, 2005 to 23 April, 2005). We monitored all bloggers' activities in these two groups with respect to the following issues;

- Information propagation and convergence patterns on social network
- Network traffic on blogspaces

First experiment is to recognize the particular patterns for information propagation on social network. We measured the duration of each article from the moment firstly posted on social network to the finally posted in any blogs. In figure 3, we are told that information propagation on the *Beta* group is much more dynamic rather than the *Alpha*. The *Alpha* needed approximately 54 days to be settled. In contrast, it took only 34 days for any articles to be propagated and converged. As a result, BlogGrid made about 37% of time cost saved through efficient recommendation. Next test is measuring the network traffic of bloggers (*Alpha* and *Beta*) during ten weeks. It proves the performance of BlogGrid's information pushing service. To do so, we counted the total number of HTTP requests related to URL's of our testing bed. As shown in Fig. 4, after 21 days, the network traffic of *Beta* group was kept to be lower than that of *Alpha*. Specially, after 34 days, we have seen that the number of HTTP requests in *Beta* was only 61% of *Alpha*'s.

Another evaluation is the degree of user satisfaction about BlogGrid's recommendation in group *Beta*. It is a measurement for how accurately BlogGrid recommended information to each blogger. We measured the ratio of the number of posts by BlogGrid's recommendation to the total number of posts. With human evaluation by interviewing with bloggers, we verified that

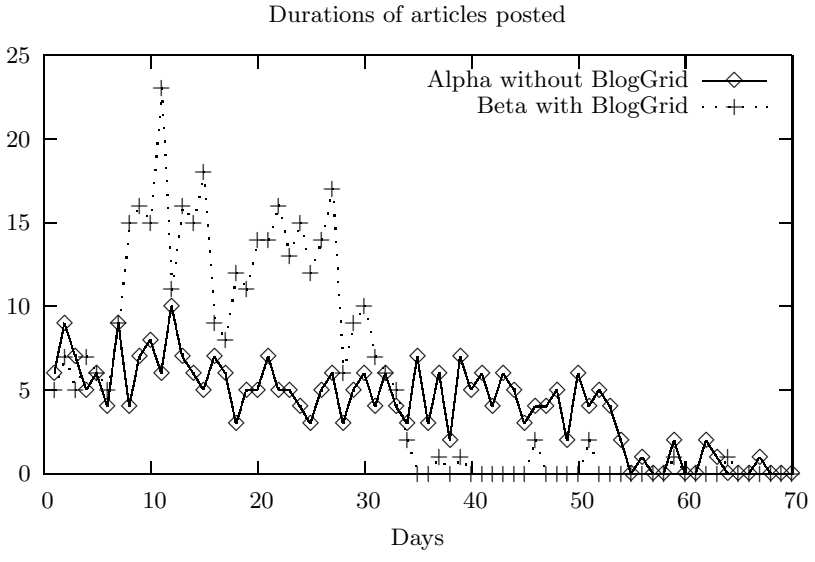

Fig. 3. Experimental results of the durations of each article posted on blogspace

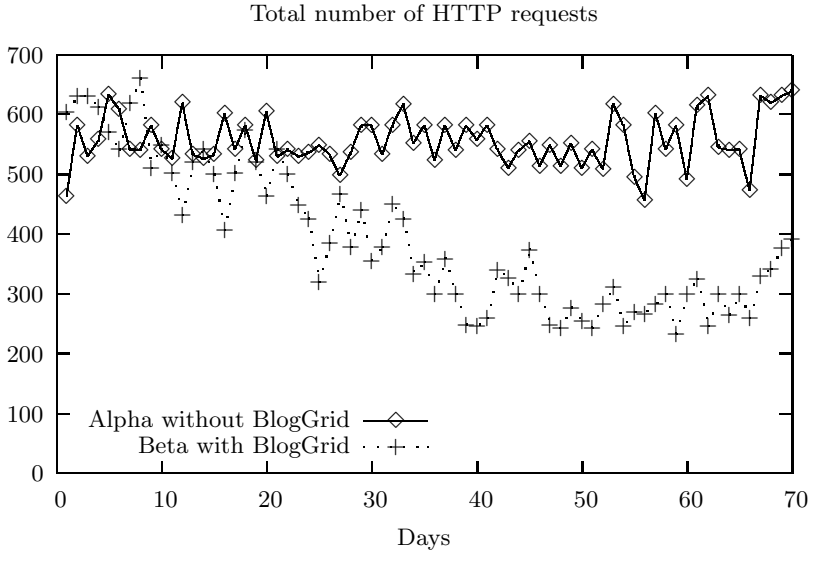

Fig. 4. Experimental results of the length of traversal paths of bloggers

– 47 bloggers, who is 62.67% of total bloggers in *Beta*, were fully satisfied,
– 16 bloggers, who is 21.3%, were partially satisfied,
– 8 blogger, who is 10.67%, was partially unsatisfied, and
– 4 blogger, who is 5.33%, were fully unsatisfied

with BlogGrid's recommendations. About 84% of bloggers rated BlogGrid as a useful system to them. This means that they were been effectively helped and guided by BlogGrid. On the other hand, only 16% of bloggers complained some inconveniences such as

- receiving irrelevant information and
- installing stand-alone application.

Particularly, we found out that most users who judged BlogGrid to be negative have shown very diverse preferences. Also, relatively many users had an opinion about software installation.

5 Discussion and Related Work

We have claimed that social grid computing environment can be organized for efficient information dissemination on personal information space, and it can be adaptive over time.

So far, several RDF languages have been developed to build social networks. Such languages are FOAF (Friend Of A Friend)[3] and SIOC (Semantically-Interlinked Online Communities) [9]. While these are based on user's explicit assertions, our proposed system is based on implicit relations. From this difference, our system can detect the relationships between potentially similar users. In other words, network isolation problem can be dealt with.

In order to maximize the spread of influence on a certain social network, game-theoretical approach has been proposed. Similar to [10], we have been attempting on discovering the more influential person (or node) from social network. More particularly, on our system, the influential weight, meaning the power on social network, can be incrementally updated over time.

In addition, we have to discuss the scalability of this system. We have tested three different number of members (25, 50, and 75). The time duration was measured as shown in Table 1. As the number of participants increases, delivering a certain information shows better performance. It proves that BlogGrid was well-organized user communities, even the number of people was larger.

Table 1. Scalability testing

Number of users	25	50	75
Average duration (days)	62.3	56.2	54.3
Ratio	-	90.21%	87.16%

Finally, we want to mention privacy issues while propagating information on distributed environment. A variety of work has been introduced to measure the reputation (or trust) weight on users [11]. In our case, we also adopt this kind approach on each hub node on social network.

[3] http://www.foaf-project.org/

6 Concluding Remarks and Future Work

As many information systems have been concerning about the paradigm of personal information space, efficient information delivering function is desperately necessary to them. We were motivated to enhance this primitive blogspace with cooperative computing methodologies. This paper proposes BlogGrid system to help bloggers and optimize efficiency of information diffusion, by analyzing and recognizing user activity and interests on social space. Most importantly, we propose an adaptive socialization method by adjusting "Hub" weighting. Therefore, bloggers were able to get information pushing service with BlogGrid's recommendation, which consists of mainly two steps; *i*) organization of virtual communities and *ii*) information pushing by facilitator. This system can conduct users overcoming the local minima by putting the effect of simulated annealing into the social network.

We developed BlogGrid system by using Blojsom[4] libraries and Borland Delphi software[5]. BlogGrid browser's graphic user interface can visualize the communications and relationships between bloggers. More importantly, the facilitator of BlogGrid is capable of

1. investigating the factors which influence relationships,
2. drawing out implications of the relational data, and
3. making recommendations to improve communications between people.

Through two main experimentations and one human evaluation, we have shown that a relatively large part of bloggers were satisfied with this service.

For short discussion and future work, we have used quite a poor size of testing bed for experimentation, so we will invite more bloggers and exploit virtual user models. Also, we are considering the topological features of social networks mentioned in [7], for recognizing more accurate relationships between bloggers. Moreover, we are considering to applying natural language processing methods to free-text articles posted on blogspaces. Similar to [12] and [13], we also plan to apply semantic information like ontologies. We are expecting it will be very powerful to interoperate between heterogeneous blogspaces. Moreover, we have to select the most appropriate domain to which BlogGrid should be applied. Various personalized or e-bussiness applications such as e-learning domain are possibly potential. Practically, BlogGrid framework will be embodied with RDF site summary (RSS)[6] service for the usability of users.

Acknowledgement. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD). (KRF-2005-214-D00347).

References

1. Dekker, A.: A category-theoretic approach to social network analysis. Electronic Notes in Theoretical Computer Science **61** (2002) 21–33
2. Rosenbloom, A.: The blogosphere. Communications of ACM **47**(12) (2004) 31–33

[4] http://wiki.blojsom.com/wiki/display/blojsom/About+blojsom.
[5] http://www.borland.com/us/products/delphi/index.html.
[6] RDF site summary (RSS). http://web.resource.org/rss/1.0/.

3. Nardi, B., Schiano, D., Gumbrecht, M., Swartz, L.: Why we blog. Communications of ACM **47**(12) (2004) 41–46
4. Zhuge, H.: Semantics, resource and grid. Future Generation Computer Systems **21**(1) (2004) 1–5
5. Jung, J.J., Ha, I., Jo, G.: BlogGrid: Towards and efficient information pushing service on blogspace. In Zhung, H., Fox, G., eds.: Proceedings of the International Conference on Grid and Cooperative Computing (GCC 2005). Volume 3795 of Lecture Notes in Computer Science., Springer-Verlag (2005) 178–183
6. Jung, J.J.: Constructing virtual communities with social interactions on blogspaces. Computing and Informatics **submitted**(x) (xxxx) xxx–xxx
7. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5) (1999) 604–632
8. Freeman, L.: Centrality in social networks: Conceptual clarification. Social Networks **1** (1979) 215–239
9. Breslin, J.G., Harth, A., Bojars, U., Decker, S.: Towards semantically-interlinked online communities. In Gómez-Pérez, A., Euzenat, J., eds.: Proceedings of the 3th European Semantic Web Conference. Volume 3532 of Lecture Notes in Computer Science., Springer-Verlag (2005) 500–514
10. Kempe, D., Kleinberg, J., Éva Tardos: Maximizing the spread of influence through a social network. In: Proceedings of the ACM SIGKDD. (2003)
11. Wang, Y., Vassileva, J.: Trust-based community formation in peer-to-peer file sharing networks. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), 20-24 September 2004, IEEE Computer Society (2004) 341–348
12. Cayzer, S.: Semantic blogging and distributed knowledge management. Communications of ACM **47**(12) (2004) 47–52
13. Jung, J.J.: An application of collaborative web browsing based on ontology learning from user activities on the web. Computing and Informatics **23**(4) (2004) 337–353