Computing and Informatics, Vol. 26, 2007, 469–488

SEMANTIC CO-BROWSING SYSTEM BASED ON CONTEXTUAL SYNCHRONIZATION ON PEER-TO-PEER ENVIRONMENT

Jason J. Jung

Department of Computer Engineering
Yeungnam University
Dae-Dong, Gyeungsan, Korea
e-mail: j2jung@intelligent.pe.kr

Revised manuscript received 20 February 2007

Abstract. In this paper, we focus on a personalized information retrieval system based on multi-agent platform. Especially, they are capable of sharing information between them, for supporting collaborations between people. Personalization module has to be exploited to be aware of the corresponding user's browsing contexts (e.g., purposes, intention, and goals) at the specific moment. We want to recommend as relevant information to the estimated user context as possible, by analyzing the interaction results (e.g., clickstreams or query results). Thereby, we propose a novel approach to self-organizing agent groups based on contextual synchronization. Synchronization is an important requirement for online collaborations among them. This synchronization method exploits contextual information extracted from a set of personal agents in the same group, for real-time information sharing. Through semantically tracking of the users' information searching behaviors, we model the temporal dynamics of personal and group context. More importantly, in a certain moment, the contextual outliers can be detected, so that the groups can be automatically organized again with the same context. The co-browsing system embedding our proposed method was shown 52.7 % and 11.5 % improvements of communication performance, compared to single browsing system and asynchronous collaborative browsing system, respectively.

Keywords: Context-awareness, focused crawling, peer-to-peer

1 INTRODUCTION

The world-wide web (shortly, web) is one of the largest online information space. However, a great amount of information has been overwhelming users on the web. The users need to be recommended and guided to efficiently search for relevant information. In information retrieval (IR) communities, a variety of methodologies have been investigated for dealing with the problem. Representative examples of such methods are information filtering and ranking. The common notion between them is to reduce the searching space for the corresponding users.

As more powerful approaches, computer-supported collaborative work (CSCW) has been adapted into IR systems. It can share not only computational resources (e.g., CPU and memory) but also highly relevant information which was already discovered by other users (e.g., URL of information sources and like-minded people), so that people can expect to improve the performance of filtering and ranking for exchanging the appropriate information and knowledge among all of the end-users. In such collaborative systems, the centralized mediator has to be designed to efficiently coordinate a group of users (or peers) who are closer than others. On the other hand, the personal software agents of the end-users have to be adaptable for personalization applications like e-mail filtering [1] and recommendation [2]. However, on the centralized platform, not only the computational complexity but also privacy-related problems have been caused [3].

Since the basic scheme of focused web crawler was proposed in [4, 5], this work has been regarded as a potential solution to the problem of indexing the exponentially increasing web. Focused crawling is designed to only gather documents on a specific topic, so that the costs for communication can be reduced. Various machine learning methodologies [6, 7, 8] have been applied to estimate the corresponding user's contexts (e.g., searching intentions and preferences).

In this paper, for the purpose of overcoming *information overloading* problem on decentralized computing environment (i.e., without the static centralized facilitator), we focus on *dynamic* self-organization of personal crawlers dispatched from a set of *contextually similar* users. We believe that group organization process can help the crawlers efficiently and proactively communicate among each other. Thereby, the software crawlers have to be able to be aware of up-to-date context about the corresponding users's task.

More importantly, the contextual transition of users should be detected as quickly as possible, so that the groups can be re-organized for real-time cooperation. We refer to context-based group organization as contextual synchronization. In general communication systems, *synchronization* can be defined as the process of making sure that two or more entities contain the same up-to-date information for consistency. Analogically, in case of on-line cooperations between people, we can interpret contextual synchronization as the process of comparing the user's current contexts and categorizing them into the groups in which the most like-minded users are involved. More particularly, the particular moments at which the user groups should be re-arranged can be detected to consider the

association with temporal tendency, e.g., the maximal (or minimal) time duration.

As more practical application, in this paper we introduce a focused collaborative browsing (co-browsing) system which is capable of communicating to share the information (and knowledge) among personal agents during navigating web for searching for a specific information. Additionally, with newly emerging semantic web technologies [9], the collaborative systems have been encouraged to deal with semantic heterogeneity problem, e.g., machine-understandability, caused by communications among software crawlers.

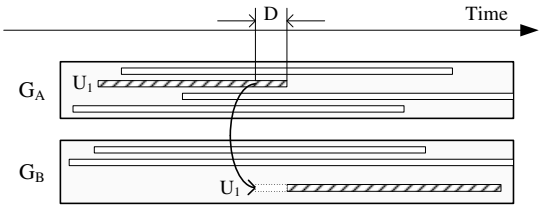

Fig. 1. Contextual synchronization as group switching

This semantics-based co-browsing system on peer-to-peer (P2P) environment has shown two main contributions; *i*) automatic organization of the groups of personal crawlers by comparing the context represented as hierarchical topic paths, and more importantly, *ii*) recognition of temporal dynamics of the corresponding user's context. Thus, when the context of user U_1 in group G_A is changed over time in Figure 1, his personal agent has to be shifted to the more relevant group G_B as quickly as possible. In other words, the delay of detection D should be minimized. Thereby, the contextual transitions while searching information from the web should be detected by analyzing the sequential patterns of user's navigation actions.

This paper is organized as follows. Section 2 presents the semantic modeling of personal context (Section 2.1) and group context (Section 2.2) from user activities and interactions on the web. In Sections 3 and 4, we describe the two-step procedure for detecting contextual transitions from the streaming web accesses by focused crawlers, and self-organization of relevant groups for collaboration. For the practical purpose, Section 4.2 will show the system architecture and processes of co-browsing. Section 5 shows the experimental results we evaluated. In Section 6, we compare the proposed method with the existing studies, and discuss some important issues, respectively. Finally, in Section 7 we draw conclusions of this study.

2 MODELING CONTEXTS

In this section, we formulate the contexts of users $\mathcal{U} = \{u_1, \dots u_{|U|}\}$ and groups $\mathcal{G} = \{g_1, \dots, g_{|G|}\}$. Basically, the context results from a sequence of web accesses $S_i = \{w^0, w^1, \dots, w^t, \dots\}$ (more practically, HTTP requests) by each user u_i, where

w^t is a web access at time t. Web access patterns of a user group g_i aggregated during time interval T is given by a matrix $\mathcal{W}(g_i)$ of which size is $|g_i| \times \max_{u_k \in g_i} |S_k|$ where $|g_i|$ is the number of users in group g_i. Therefore, we can extract two kinds of contextual information.

Definition 1 (Personal context $\mathcal{C}_\mathcal{U}$)**.** A personal context $\mathcal{C}_\mathcal{U}(u_j)$ of user u_j can be discovered from his web access patterns by using centralized ontologies. Personal context at a certain moment is represented as a set of topical hierarchical paths $\{\langle tp_{root}, \ldots, tp_m \rangle | tp_m \in \Omega(m)\}$ meaning a sequence of concepts from a root concept tp_{root} to concept tp_m in the ontology. We exploit two kinds of semantic labeling processes Ω for a certain web access m (in Section 2.1). Personal context is assumed to be changed over time. Temporal dynamics of user u_j's patterns can be extracted from j^{th} row component in $\mathcal{W}(g_i)$.

Definition 2 (Group context $\mathcal{C}_\mathcal{G}$)**.** A group context is merged from personal contexts of a set of corresponding participants in a same group. Thus, it is given by

$$\mathcal{C}_\mathcal{G}(g_i) = \cap^\diamond_{u_j \in g_i} \mathcal{C}_\mathcal{U}(u_j) \tag{1}$$

where $\cap^\diamond$ is a function returning the principal topical paths. We merge a set of personal context into a group context, by using influence propagation theory. It will be explained in Section 2.2. As a set of personal context are changing over time, group context $\mathcal{C}_\mathcal{G}(g_i)$ also shows temporal dynamics. At a certain moment t_k, $\mathcal{C}_\mathcal{G}(g_i)^{t_k}$ can be extracted from k^{th} column component in $\mathcal{W}(g_i)$, after ordering by timestamps of web accesses.

2.1 Personal Context Based on Conceptualizing Web Accesses

In order to label the HTTP requests by users, we can simply extract the URL information and perform semantic labeling process Ω, which is assigning a set of hierarchical topics (or categories) to the corresponding HTTP request. There are two ways of labeling, which are referred to as *direct* labeling Ω_D and *indirect* labeling Ω_{ID}, depending on whether the web site in question is already registered in the web directories.

For the web sites already registered in the web directory, we can apply direct labeling to them. Direct labeling is a simple querying process which involves looking up the corresponding URLs in the web directory. In order to deal with the drawbacks of the web directory (e.g., multiple attributes and subordination) mentioned in [10], we have to acquire a set of labels which includes all possible paths in order to obtain the desired results.

On the other hand, indirect labeling is used for unregistered web sites. This method is based on link analysis, and involves searching "authoritative" pages about a certain topic on the hyperlinked information space like web pages [11]. We propose a modified HITS algorithm which allows the most similar data to be obtained from the already labeled dataset. The hyperlinked web pages are organized into a directed

graph $G = (V, E)$, where V is the set of nodes representing the web sites, and E is the set of hyperlinks between v_i and v_j. In order to search the most authoritative node of a particular web site, we focus on the outgoing links of that web site. For the given unlabeled web page w, the outgoing and incoming links of graph G can be formulated as the asymmetric adjacency matrix, $O(w)_x^{(d)}$, where $[O(w)]_{ij} = 1$ if $v_i \to v_j$ and $[O(w)]_{ij} = 0$, otherwise. Also, the variable d is the number of iterated expansions, which means the distance from node w. This $O(w)$ is a $|V| \times |V|$ square matrix, where V is the set of nodes within the distance d. Therefore, we can reach some labeled nodes, by repeating this iteration along the outgoing links. If there are more than one labeled node at the same distance, we have to evaluate the incoming degree of these nodes by using the following equation Ω_{ID}

$$\Omega_{ID}(O(w)_x^{(d)}) = \max_{j^\star \in j} \left[\sum_k [O(w)]_{kj^\star} \right] \tag{2}$$

where the $j^{\star\text{th}}$ web sites are labeled. This means that the web sites can be regarded as more authoritative ones, since they are referred to by a larger number of other web sites. In the example shown in Figure 2, the web site m which is requested by the clients is not yet registered in the web directory. The solid arrow lines are outgoing links to other web sites, while the dotted lines are incoming links from other web sites. The web site x belongs to the nearest neighbor category that is registered in the web directory.

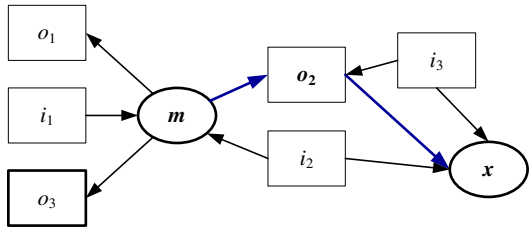

Fig. 2. Indirect labeling of unregistered web site m

Thus, the link matrix of this example is given by

$$O(m)_x^{(2)} = \begin{array}{c} \\ m \\ o_1 \\ o_2 \\ \dots \\ x \end{array} \begin{array}{c} \begin{array}{ccccc} m & o_1 & o_2 & \dots & x \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \end{array} \tag{3}$$

where the distance threshold d is predefined as two. Let the web pages o_3 and x be registered in the web directory. By using Ω_{ID}, the maximum authoritative web page x can be obtained.

Hence, at a given time interval $[t_k, t_k + T]$, we can obtain a sequence of topics representing the personal context $\mathcal{C}_{\mathcal{U}}(u_j)$ of the corresponding user from the matrix $\mathcal{W}(g_i)_{j,[t_k,t_k+T]}$. As combining the temporal patterns of personal contexts of users in the same group, we can semantically estimate the temporal patterns of group context.

2.2 From Personal Context to Group Context by Influence Propagation

As defined in Definition 2, a group context is built by collecting the personal context of participants. Thereby, a given set of personal contexts $\{\mathcal{C}_{\mathcal{U}}(u_j)|u_j \in g_i\}$, the group context $\mathcal{C}_{\mathcal{G}}(g_i)$ is built by the topics representing the most common context among group members g_i. We apply influence propagation scheme to measure a relevance weight R of each topic tp. The super classes of tp on a topic path are assumed to be influenced from the leaf concept tp' with a certain decaying factor λ. It is given by

$$R_{g_i}(tp) = \sum_{u_j \in g_i} \frac{1 - \lambda^{\frac{d_{tp'}}{d_{tp'} - d_{tp}}}}{|\mathcal{C}_{\mathcal{U}}(u_j)|} \tag{4}$$

where d_{tp} is the hierarchical depth of tp on the corresponding topic path. This equation expresses that the more specific topic has the more influential power to group context. The decaying coefficient $\lambda \in [0, 1]$ can control the rate of subsumption influence from the specific topics. The larger λ reduces the influence between topics. More importantly, with high decaying factor, we can expect that topic-based group context eventually is more precise but has low coverage.

Hence, at a certain moment t_k, by merging $\{\mathcal{C}_{\mathcal{U}}(u_1)^{t_k}, \ldots, \mathcal{C}_{\mathcal{U}}(u_{|g_i|})^{t_k}\}$, we can obtain $\mathcal{C}_{\mathcal{G}}(g_i)^{t_k}$. Similarly to the personal context, we also recognize the temporal dynamics of group context by aggregating the merged personal context for the given time interval.

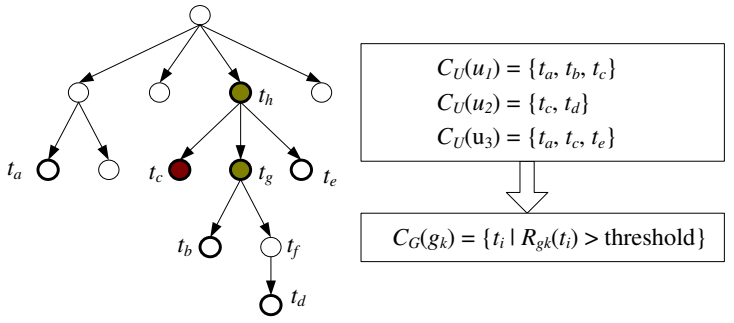

Fig. 3. An example of merging group context

We want to show an example. Assume that a group g_i be organized as three members u_1, u_2, and u_3. The personal context is represented as $\mathcal{C}_{\mathcal{U}}(u_1) = \{t_a, t_b, t_c\}$,

$\mathcal{C}_{\mathcal{U}}(u_2) = \{t_c, t_d\}$, and $\mathcal{C}_{\mathcal{U}}(u_3) = \{t_a, t_c, t_e\}$ (illustrated as bold circles in Figure 3). After the coefficient factor is set to $\lambda = 0.6$, we can compute the relevance weight R of the concepts, as follows.

$$\begin{aligned}
R_{g_i}(t_a) &= 1/3 + 1/3 = \mathbf{0.667} \\
R_{g_i}(t_b) &= 1/3 = 0.333 \\
R_{g_i}(t_c) &= 1/3 + 1/2 + 1/3 = \mathbf{1.167} \\
R_{g_i}(t_d) &= 1/2 = \mathbf{0.5} \\
R_{g_i}(t_e) &= 1/3 = 0.333 \\
R_{g_i}(t_f) &= (1 - 0.6^{4/1})/2 = 0.435 \\
R_{g_i}(t_g) &= (1 - 0.6^{4/2})/2 + (1 - 0.6^{3/1})/3 = \mathbf{0.581} \\
R_{g_i}(t_h) &= (1 - 0.6^{2/1})/3 + (1 - 0.6^{3/2})/3 + (1 - 0.6^{4/3})/2 + (1 - 0.6^{2/1})/3 \\
&= \mathbf{0.852}
\end{aligned}$$

Here, $t'_f = t_d$, $t'_g = \{t_b, t_d\}$, and $t'_h = \{t_b, t_c, t_d, t_e\}$. They have been influenced by their own subconcepts, respectively. We can discriminate which topics are more representative for the group, and which topics might be removed by a certain threshold.

3 DETECTION OF CONTEXTUAL TRANSITION

In this section, we consider that personal context can be changed during co-browsing, and if any contextual transition of the personal context is detected, the group should be re-organized. Thereby, several semantic factors are defined to measure the various relationships between personal contexts of users in a group, and group contexts. Above all, we want to explain how to conceptualize a set of HTTP requests. Then, we have to consider to compute not only the semantic factors in a given time interval but also the distributions of $\mu^{\diamond}$ and $\sigma^{\diamond}$ by using the sliding windows method. Hence, the triggering patterns from these signals are regarded as important evidence. This process should be conducted by following the objective function

$$\min \sum_{g_i \in G} \left| \mathcal{C}_{\mathcal{G}}^{t_{j+1}}(g_i) - \mathcal{C}_{\mathcal{G}}^{t_j}(g_i) \right| \tag{5}$$

where G is a set of user groups. It means that the summation of temporal differences of $\mathcal{C}_{\mathcal{G}}$ of every group should be minimized.

3.1 Semantic Factors

After the semantic labeling process of an arbitrary web request, we have obtained two kinds of context information, $\mathcal{C}_{\mathcal{U}}$ and $\mathcal{C}_{\mathcal{G}}$, which are represented as a set of hierarchical paths of the corresponding topics. We assume that a web page w^i accessed by

user can be categorized to a set of topical paths $\{tp_k^i | tp_k^i \in \Omega(w^i), k \in [1, \ldots, K]\}$ where K is the number of all possible topic paths.

Definition 3 (Semantic distance $\delta^\diamond$)**.** The minimum value among all combinatorial comparisons of two topic sets ($|tp^i| \times |tp^j|$):

$$\delta^\diamond(w^i, w^j) = \min_{m=1,n=1}^{M,N} \frac{\min\left((L_m^i - L_{m,n}^C), (L_m^j - L_{m,n}^C)\right)}{\exp(L_{m,n}^C)} \tag{6}$$

where L_m^i, L_n^j, and $L_{m,n}^C$ are the lengths of tp_m^i, tp_n^j, and the common part of both, respectively. Semantic distance is assigned in the interval $[0, 1]$, and in case of complete matching, it is 0. Exponent function in the denominator is used for increasing the effect of $L_{(m,n)}^C$. Additionally, $\delta^\diamond(w^i, w^j) = \delta^\diamond(w^j, w^i)$.

Definition 4 (Semantic distance matrix $\Delta^\diamond$)**.** From the matrix $\mathcal{W}(g_i)$ aggregating the user web accesses during a given time interval T, we can obtain a sequence of topic sets which are aggregated by either a particular users' web accesses (row components in $\mathcal{W}(g_i)$ for $\mathcal{C}_\mathcal{U}$) or a group context $\mathcal{C}_\mathcal{G}$ merging the column components at each moment. Given a subsequence H from $\mathcal{W}(g_i)$, a semantic distance matrix $\Delta^\diamond$ is represented by

$$\Delta^\diamond(i, j) = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \delta^\diamond(w^i, w^j) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \tag{7}$$

of which size is $|H| \times |H|$, and the diagonal elements are all zero. Also, by the commutative law, it is a symmetric matrix.

Definition 5 (Semantic distance mean $\mu^\diamond$)**.** Semantic distance mean is the average value of upper triangular elements in $\Delta^\diamond$ except diagonal components, and it is given by

$$\mu^\diamond = \frac{2}{T(T-1)} \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \Delta^\diamond(i, j) \tag{8}$$

where T is the time interval, indicating the size of $\Delta^\diamond$. It can measure the semantic consistency of the given context set (including $\mathcal{C}_\mathcal{U}$ and $\mathcal{C}_\mathcal{G}$) extracted from the corresponding set of web requests during co-browsing.

Definition 6 (Semantic distance deviation $\sigma^\diamond$)**.** The standard deviation is calculated by $\mu^\diamond$ and the components from $\Delta^\diamond$. It is formulated by

$$\sigma^\diamond = \sqrt{\frac{2}{T(T-1)} \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \left(\Delta^\diamond(i, j) - \mu^\diamond\right)^2}. \tag{9}$$

It is simply a statistical value measuring the degree of dispersion of the semantic distance values from a given set of topic paths from web requests.

Then, based on temporal dynamics of semantic factors over time, we are expecting to identify the semantically significant transition moments of personal contexts and group context, during co-browsing.

3.2 Two-Step of Contextual Synchronization

This paper has focused on supporting people to efficiently interact with like-minded users whose personal contexts are semantically closer, during online co-browsing. Especially, we want to deal with this problem by detecting contextual transition discovered by the streaming browsing patterns from multiple clients. This detection process is simply expressed by using semantic factors in Equations (6)–(9).

The contextual synchronization process is organized as two steps;

Alarming step. Semantic distance deviation $\sigma^{\diamond}$ of column components in $\mathcal{W}(g_i)$ is applied to capture significant contextual transitions of a particular user u_i. It is a certain change moment t_k when a personal context $\mathcal{C}_{\mathcal{U}}$ is different from the corresponding group's context $\mathcal{C}_{\mathcal{G}}$. Basically, this step is very similar to the outlier detection from a given streaming dataset from various application domains (e.g., financial transactions, environmental and scientific data sources) [12, 13, 14]. In this paper, by using Equation (5), a set of time points t_{Alarm} for generating alarming can be characterized to

$$t_{Alarm} = \left\{ t_k \middle| t_k \in T, |\sigma^{\diamond}(\mathcal{C}_{\mathcal{G}}^{t_k}) - \sigma^{\diamond}(\mathcal{C}_{\mathcal{G}}^{t_{k-1}})| \geq \lambda_{Alarm} \right\} \tag{10}$$

where λ_{Alarm} is the threshold value for alarming that there exist some users u_{Alarm} whose personal contexts are semantically different from the other group members ($\mathcal{C}_{\mathcal{G}}$ transitions). Then, at a given time point t_k, the users $u_{Alarm}^{t_k}$ are simply detected by

$$u_{Alarm}^{t_k} = \left\{ u_j \middle| u_j = \arg_i \max \sum_{h=1}^{|H|} \Delta^{\diamond}(j,h) \right\} \tag{11}$$

where H means the size of the semantic distance matrix $\Delta^{\diamond}$. It finds out and removes the most dissimilar users (the maximal summation of semantic distances $\delta^{\diamond}$) within the group g_i. As removing $u_{Alarm}^{t_k}$ from $\Delta^{\diamond}$, we have to repeat this task, until the temporal difference of group context in Equation (10) is less than λ_{Alarm} (formulated as $|\sigma^{\diamond}(\mathcal{C}_{\mathcal{G}}^{t_k}) - \sigma^{\diamond}(\mathcal{C}_{\mathcal{G}}^{t_{k-1}})| \leq \lambda_{Alarm}$).

Confirming step. In the previous step, we discovered a set of users $u_{Alarm}^{t_k}$ whose personal contexts are "likely" to be changed at time t_k. For confirming step, we want to

- confirm whether the users have shown contextual transitions or not (we will discuss more details about associations between personal contexts and group context in Section 6), and

- discover the specific transition moments of personal contexts of the confirmed users.

Thereby, semantic distance mean $\mu^{\diamond}$ is measured to make sure whether the alarmed users' personal contexts $\mathcal{C}_{\mathcal{U}}(u_i)_{u_i \in u_{Alarm}}$ are changed or not. This can be found out by constructing the semantic distance matrix $\Delta^{\diamond}$ by using the subsequence extracted from each row component of $\mathcal{W}(g_i)$. If the user u_{Alarm} has shown any contextual transitions, we can detect more specific time points t_S of contextual transitions. Similarly to the previous "alarming" step, confirming step for the alarmed users can be characterized as

$$u_{Confirm} = \left\{ \langle u_j, t_{s_q} \rangle \middle| u_j \in u_{Alarm}, CFRM(u_j, t_{s_q}) \geq \lambda_{Confirm} \right\} \tag{12}$$

where threshold $\lambda_{Confirm}$ has to be pre-defined by users, for confirming the contextual transitions of personal context $\mathcal{C}_{\mathcal{U}}(u_j)$ at time t_{s_q}. Function $CFRM$ is given by

$$CFRM(u_j, t_{s_q}) = |\mu^{\diamond}(\mathcal{C}_{\mathcal{U}}(u_j)^{t^{-}_{s_q}}) - \mu^{\diamond}(\mathcal{C}_{\mathcal{U}}(u_j)^{t^{+}_{s_q}})| \tag{13}$$

where $t^{-}_{s_q}$ and $t^{+}_{s_q}$ mean the bisected time intervals $[t_{s_0}, t_{s_q} - 1]$ and $[t_{s_q}, t_T]$, respectively. Hence, a set of time points t^j_S surely is the moment when personal context of the corresponding user is changed.

Each time the streaming web accesses within a group are stored in $\mathcal{W}(g_i)$, the two-step procedure for detecting contextual transitions has to be fulfilled. We employ the semantic distance deviation $\sigma^{\diamond}$ to recognize the dispersion of members in a group, rather than the group context itself. Afterward, if some users would be detected in this step, the confirming step can justify whether their transitions are validated or not, because the semantic distance mean $\mu^{\diamond}$ is useful to measure the semantic cohesion within a certain time interval.

4 ONLINE CO-BROWSING BASED ON SELF-ORGANIZING COLLABORATION

For making online interactions between personal agents more meaningful, we assume that the group should be automatically organized as the people whose context is very similar with each other. After a contextual transition is detected from the users in a certain group, they need to be switched into the more proper groups. To do this, we select N super-peers which can represent N groups, respectively. Coefficient N means the total number of groups, and it should be predefined. The rest of users are determined as to which super-peer is the most proper one by measuring the semantic distance. Basically, this process is similar to k-nearest neighborhood method [15], one of the best known non-parametric approaches, for classifying phenomena based upon observable features.

4.1 Group Organization by Super-Peers

Originally, a super-peer is a node in a P2P network that operates both as a server to a set of clients, and as an equal in a network of super-peers [16]. This super-peer user is playing a role of monitoring and controlling the rest of members in the same group. Some of super-peer networks are allowing super-peer redundancy, meaning that each peer can be connected with several different super-peers, at the same time. These networks, in fact, have shown better performance with respect to cost effectiveness and reliability [17]. In this paper, however, we assume that each group should select only one user as a super-peer. We believe that in case of dynamic and real-time organization of super-peer networks, single super-peer network will outperform the redundancy one, in terms of scalability. We will show experimental results in Section 5, and discuss this issue in Section 6.2.

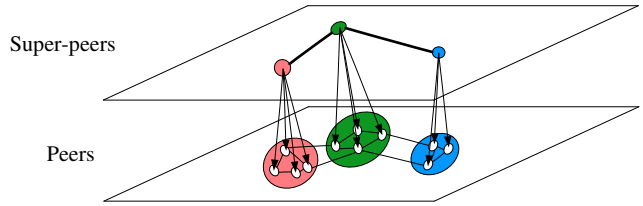

Fig. 4. Bipartite super-peer network

In order to build a bipartite super-peer network shown in Figure 4, we have to assign the most *central* user as a super-peer of the corresponding group. In other words, the super-peer user's personal context is most similar to the corresponding group's context $\mathcal{C}_{\mathcal{G}}$. Thus, a set of the super-peer users $SupPeer$ is selected by

$$SupPeer_{\mathcal{G}} = \left\{ u^k_{Sup} \middle| g_i \in \mathcal{G}, \min_{u^k_{Sup}, u_j \in g_i} \sum_{j=1}^{|g_i|} \delta^{\diamond}(\mathcal{C}_{\mathcal{U}}(u^k_{Sup}), \mathcal{C}_{\mathcal{U}}(u_j)) \right\} \tag{14}$$

which means that the super-peer users must be in the most middle of the group as minimizing the semantic distances between other members. This selection process is conducted when i) the initial group organization is configured and ii) the summation of semantic distances between personal context of members in a same group is over the threshold λ_{Alarm}. The size of a set of super-peer users is $|SupPeer_{\mathcal{G}}| = |\mathcal{G}|$.

Now, we want to explain the automated group re-organization process based on the users whose contextual transition has been detected. It is simply based on combination of the objective functions Equations (5) and (14). Let a user u_i be in a group g_j. At time t_f, contextual transition of his personal context $\mathcal{C}_{\mathcal{U}}(u_i)$ from the group context $\mathcal{C}_{\mathcal{G}}(g_j)$ is confirmed ($u_i \in u^{t_f}_{Confirm}$), his new group is decided by

$$\arg_{u^k_{Sup} \in g_k} \min_{g_k \in \mathcal{G}} \delta^{\diamond}(\mathcal{C}_{\mathcal{U}}(u_i), \mathcal{C}_{\mathcal{G}}(g_k)) \tag{15}$$

where $|\mathcal{G}|$ is the total number of groups in super-peer network. It can search for the most relevant group g_k, meaning that the group context $\mathcal{C}_{\mathcal{G}}(g_k)$ is closest to user $\mathcal{C}_{\mathcal{U}}(u_i)$. With the information, the super-peers of the corresponding group g_k are surely regarded as the representative context of the corresponding group context. Hence, u_i can join the most relevant group $\mathcal{C}_{\mathcal{G}}(g_k)$, by making connection to the super-peer user u^k_{Sup}.

4.2 Blackboard-based Co-Browsing System

The users who detected their context transitions should be re-organized to the most relevant group by Equation (15). Thus, they can get all information about the group members' browsing patterns through blackboard module. Figure 5 shows the architecture of our proposed system.

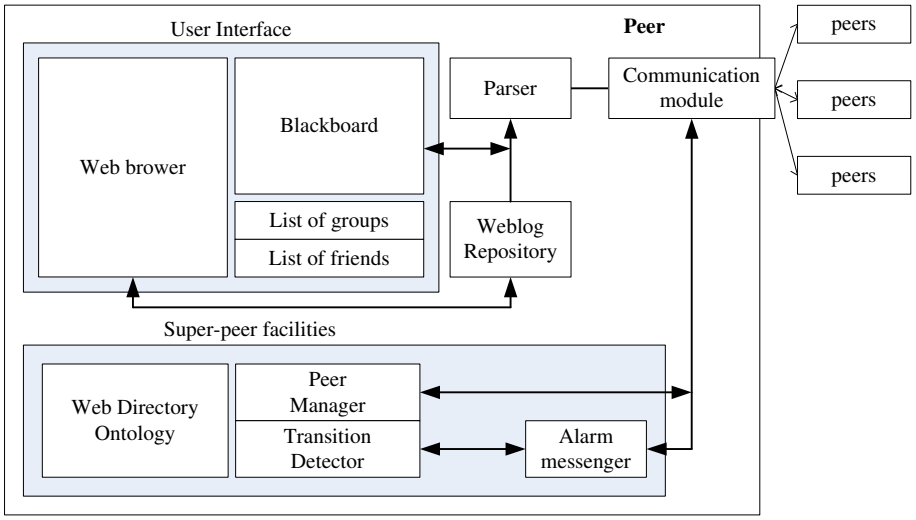

Fig. 5. System architecture based on blackboard module

User interface is simply composed of three frames for web browser, blackboard, and lists of friends/groups. On the super-peer network environment, all users utilize homogeneous system, except the facilities for super-peers.

We exploit the blackboard system to share information among only the users within a certain group g_i. Of course, there are many methods to visualize the information to improve user intuition, but we exploit the simple text-based interface. Simply each user's actions occurred during web browsing are announced into the interface of the group members. More importantly, the information about which users are newly joined in and discarded from his group is also announced.

The information on blackboard can help users access to the candidate web pages that are potentially relevant to the searching context. Also, blackboard system is

capable of parsing their URLs, counting their frequencies, and finding out the most popular ones.

5 EXPERIMENTAL RESULTS

In order to implement the proposed co-browsing system, the development specification is mainly divided into two parts; *i*) for peer modules (e.g., graphic user interface (GUI) module and web browser), we exploited Borland Delphi[1], and *ii*) for super-peer modules (e.g., peer manager and alarm manager), JXTA API libraries[2] was being applied.

We conducted simulations to evaluate the performance of communications on the proposed method for co-browsing system on P2P environment. Three groups G_A, G_B, and G_C are organized by 30 users (ten users in each group), and then we collected the web logs dataset by letting these users to browse the testing bed space in the fixed personal context. After cleansing the collected dataset by preprocessing scheme proposed in [18], we prepared the testing dataset, which is composed of 4 610 web pages labeled by 28 categories from ODP[3].

As introduced in [19], in order to evaluate our co-browsing framework, we computed *precision* and *recall* indices measuring the ratio of the matched results (e.g., contextual transition), in both cases (i.e., single browsing and co-browsing).

$$recall = \frac{\text{Matched results}}{\text{Retrieved (or detected) results}} \quad \text{and} \tag{16}$$

$$precision = \frac{\text{Matched results}}{\text{Targeted results}} \tag{17}$$

First, in order to evaluate the detection of contextual transitions of each user, we generated 30 synthesized sequences, including totally 583 contextual transitions, by randomly intermixing the fragments which are randomly segmented from the web log dataset. We examined how exactly the transitions could be detected with respect to two measurements, *precision* and *recall*. Additionally, *F1*-value is computed by $\frac{2recall \times precision}{recall + precision}$ for combining these two measures.

We evaluated the procedures of alarming step, as changing the threshold λ_{Alarm}. Figure 6 depicts the experimental results of contextual transition detection in alarming step (Equations (10) and (11)). We obtained in average $F1 = 0.52$, and when $\lambda_{Alarm} = 0.4$, the best performance ($F1 = 0.544$).

In case of confirming step, Figure 7 shows the experimental results, as changing $\lambda_{Confirm}$. Average performance was $F1 = 0.72$, and when $\lambda_{Confirm} = 0.6$, the maximum results have been shown ($F1 = 0.759$).

We empirically uncovered the best threshold values $\lambda_{Alarm} = 0.4$ and $\lambda_{Confirm} = 0.6$. The threshold level of confirming step seems slightly more critical, because it is

[1] Borland Delphi. `http://www.borland.com/`.

[2] JXTA API. `http://www.jxta.org/`.

[3] Open Directory Project, `http://www.dmoz.org/`.

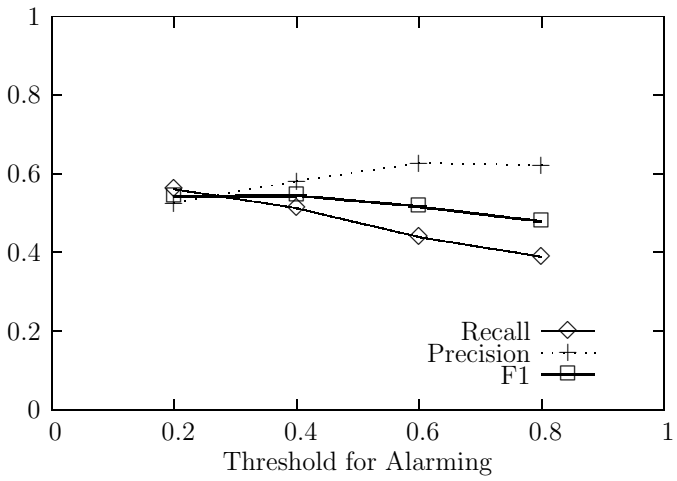

Fig. 6. Detection of contextual transitions in alarming step over λ_{Alarm}

for the personal context. Generally, the average performance of confirming process is about 37 % higher than that of alarming step.

Second, we evaluated the performance of communications by group organization. This proves the efficiency of online co-browsing, rather than single browsing or basic co-browsing systems (i.e., without contextual synchronization). While users in G_A browsed without any collaboration, G_B and G_C was under co-browsing. G_C was the only users provided the group re-organization process based on detecting contextual

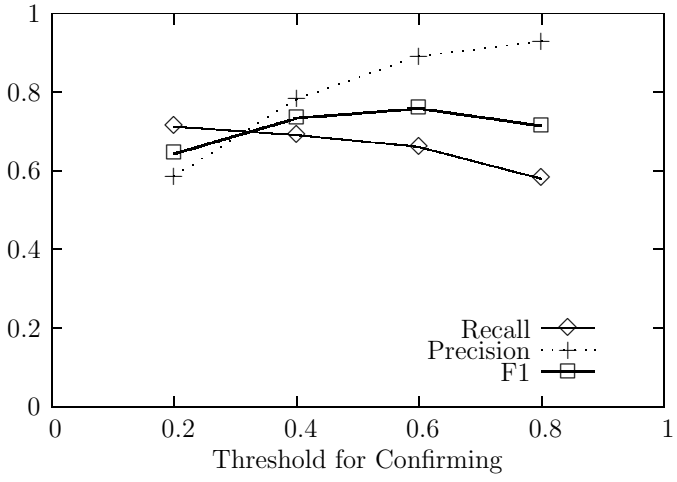

Fig. 7. Detection of contextual transitions in confirming step over $\lambda_{Confirm}$

transitions. We monitored the performance of information searching tasks in three groups over time, by comparing the topics extracted from the retrieved information with the topics the users selected as their interests before experiments.

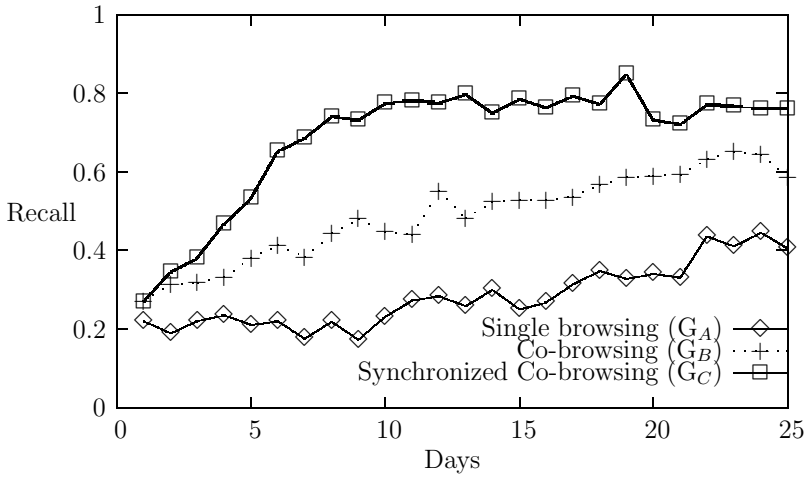

Fig. 8. Performance of information searching from three groups with *recall*

As shown in Figure 8, G_B with synchronized co-browsing has shown the best *recall* results (on the 9$^{\text{th}}$ day, about four times higher than G_A, and on the 7$^{\text{th}}$ day, 79 % higher than G_B). It means that synchronized co-browsing can support the users, particularly in the early stage.

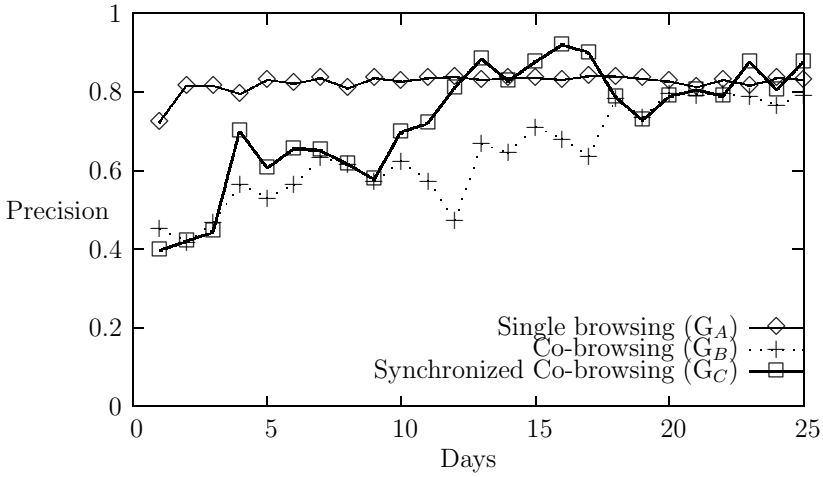

Fig. 9. Performance of information searching from three groups with *precision*

With respect to *precision*, in Figure 9 we found out that co-browsing systems finally have shown the converged result over 80 % precision level, even though, in the initial stage, single browsing has shown the best performance. In case of single browsing, the users put their preferences into the corresponding personal crawlers.

Communications	G_A	G_B	G_C
Group communications	-	13 441	12 325
Web accesses	6 232	3 662	3 285
Ratio	-	58.76 %	52.71 %

Table 1. Evaluation of the performance of communications

Generally, Table 1 shows the final results of three group members' browsing for four weeks. G_C in online co-browsing has shown only 53 % web access with helping each other according to the context. Compared with G_B, our proposed method has slightly improved by 11.5 %.

6 DISCUSSION AND RELATED WORK

In this paper, we have proposed the collaborative personal agents on heterogeneous web. Basically, the focused crawler systems, also known as topical (topic-driven) crawlers, are rather accessible to most of web spaces systematically. We first need to compare single ontology-based platform with multiple ontology-based platform. Second, more importantly, we want to discuss the dynamics of personal and group contexts on super-peer network.

6.1 Comparison Between Single and Multiple Ontologies

In this work, we exploit a single centralized ontology, i.e., web directory. It makes the semantic heterogeneity problem to be automatically solved, because every resource is annotated (or labeled) by referring to the single ontology.

However, we have to consider a platform which is providing multiple ontologies. Each information source (or system) can build its own ontology. This sort of ontologies might be domain-specific and cause semantic heterogeneity problem. In our case, the users' agents can be embedded with their personal interests. Also, they can possibly edit their own personal ontologies. Thereby, ontology alignment (or mapping) methods have been proposed. Recently, Shvaiko and Euzenat explained the classification method of ontology alignment (or matching) [20] and ontology mapping algorithms [21]. Several alignment methodologies have been introduced. Since Dieng and Hug proposed an algorithm for matching conceptual graphs using terminological linguistic techniques and comparing superclasses and subclasses [22], Euzenat developed *T-tree* to infer the dependencies between classes (bridges) of different ontologies sharing the same set of instances based only on the "extensions of classes" [23]. Additionally, *FCA-merge* uses formal concept analysis techniques to

merge two ontologies sharing the same set of instances while properties of classes are ignored [24]. Meanwhile, *Cupid* is the first approach combining many of the other techniques. It aligns acyclic structures taking into account terminology and data types (internal structure) and giving more importance to leaves [25]. Especially, we have studied ontological mediator framework for sharing semantic information between personal crawlers [26].

6.2 Discussing Dynamics of Personal Contexts and Group Context

We have assumed that the context during web browsing can be changed. In most collaboration systems, the context is coupled and related with each other *i*) between people, *ii*) between groups, and *iii*) between a person and a group. For supporting the online collaborations between users, we realized that capturing (and comparing) the dynamics of contexts is more important than representing (and comparing) the contexts themselves.

There have been two well-known P2P networks such as Napster[4] and Gnutella[5]. Especially, we exploit the super-peer network scheme like Napster-style. A pure P2P network is a "degenerate" super-peer network where cluster size is one. While it means that every node is a super-peer with no clients, super-peer networks such as KaZaA[6] use the heterogeneity of peers to their advantages. Also, with regard to the computational complexity, the overhead of maintaining an index at the super-peer networks is small in comparison to the savings in query cost this centralized index allows [16]. Furthermore, Xiao et al. proposed dynamic super-peer network based on dynamic layer management [27].

7 CONCLUDING REMARKS AND FUTURE WORK

Social browsing has been emerged in various collaborative systems. In this paper, we have proposed online co-browsing system of which characteristics are spatially remote and temporal synchronous. It is capable of detecting the contextual transitions of users in a group, so that they are efficiently shifted into the relevant group communications. As main contribution of this paper, most importantly, we propose tracking the contextual dynamic of the groups while co-browsing, rather than modeling the consensual context of the groups.

However, this system has still many problems that have to be dealt with in future work. We modified *Levenshtein* edit distance [28] to measure the hierarchical path-labeled web pages. In order to support more general users, we obviously consider various semantic annotation methods [29] to compare the relationships between them. Another issue is topology in P2P network. Because, as mentioned in [11], the hyperlinked environment has various topological features such as authorities

[4] Napster. `http://www.napster.com/`.

[5] Gnutella. `http://www.gnutella.com/`.

[6] KaZaA. `http://www.kazaa.com/`.

and hubs, we have to think over about selection process of super-peers. Finally, more globally, by using grid-computing paradigm [30], we also image the social grid environment, providing k-redundant super-peer networks [16].

Acknowledgement

This research was conducted as part of National IT Ontology Infrastructure Technology Development project funded by Ministry of Information & Communication of Korean government.

REFERENCES

[1] Sahami, M.—Dumais, S.—Heckerman, D.—Horvitz, E.: A Bayesian Approach to Filtering Junk E-mail. In: Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, 1998.

[2] Middleton, S.—Shadbolt, N.—de Roure, D.: Ontological User Profiling in Recommender Systems. ACM Transactions on Information Systems, Vol. 22, 2004, No. 1, pp. 54–88.

[3] Ziegler, C.-N.—Lausen, G.: Paradigms for Decentralized Social Filtering Exploiting Trust Network Structure. In: Z. T. R. Meersman (Ed.), Proceedings of the International Conference on Cooperative Information Systems (CoopIS '04), Vol. 3291 of Lecture Notes in Computer Science, Springer-Verlag, 2004, pp. 840–858.

[4] Bra, P. M. E. D.—Post, R. D. J.: Information Retrieval in the World-Wide Web: Making Client-Based Searching Feasible. Computer Networks and ISDN Systems, Vol. 27, 1994, No. 2, pp. 183–192.

[5] Chakrabarti, S.—van den Berg, M.—Dom, B.: Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. Computer Networks, Vol. 31, 1999, No. 11–16, pp. 1623–1640.

[6] Chakrabarti, S.—Punera, K.—Subramanyam, M.: Accelerated Focused Crawling Through Online Relevance Feedback. In: Proceedings of the 11th International Conference on World Wide Web (WWW '02), ACM Press, New York, NY, USA, 2002, pp. 148–159.

[7] Menczer, F.—Pant, G.—Srinivasan, P.: Topical Web Crawlers: Evaluating adaptive algorithms. ACM Transactions on Internet Technology, Vol. 4, 2004, No. 4, pp. 378–419.

[8] Pant, G.—Srinivasan, P.: Learning to Crawl: Comparing Classification Schemes. ACM Transactions on Internet Technology, Vol. 23, 2005, No. 4, pp. 430–462.

[9] Berners-Lee, T.: The Semantic Web. Scientific American, Vol. 285, 2001, No. 5, pp. 34–43.

[10] Jung, J. J.: Collaborative Web Browsing Based on Semantic Extraction of User Interests With Bookmarks. Journal of Universal Computer Science, Vol. 11, 2005, No. 2, pp. 213–228.

[11] Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, Vol. 46, 1999, No. 5, pp. 604–632.

[12] Puttagunta, V.—Kalpakis, K.: Adaptive Methods for Activity Monitoring of Streaming Data. In: M. A. Wani, H. R. Arabnia, K. J. Cios, K. Hafeez, G. Kendall (Eds.), Proceedings of the 2002 International Conference on Machine Learning and Applications (ICMLA 2002), CSREA Press, 2002, pp. 197–203.

[13] Papadimitriou, S.—Brockwell, A.—Faloutsos, C.: Adaptive, Unsupervised Stream Mining. The VLDB Journal, Vol. 13, 2004, No. 3, pp. 222–239.

[14] Papadimitriou, S.—Sun, J.—Faloutsos, C.: Streaming Pattern Discovery in Multiple Time-Series. In: K. Böhm, C. S. Jensen, L. M. Haas, M. L. Kersten, P.-Å. Larson, B. C. Ooi (Eds.), Proceedings of the 31st International Conference on Very Large Data Bases (VLDB), ACM, 2005, pp. 697–708.

[15] Cover, T.—Hart, P.: Nearest Neighbor Pattern Classification, IEEE Transactions in Information Theory, Vol. 13, 1967, No. X, pp. 21–27.

[16] Yang, B.—Garcia-Molina, H.: Designing a Super-Peer Network. In: U. Dayal, K. Ramamritham, T. M. Vijayaraman (Eds.), Proceedings of the 19th International Conference on Data Engineering (ICDE 2003), IEEE Computer Society, 2003, pp. 49–60.

[17] Androutsellis-Theotokis, S.—Spinellis, D.: A Survey of Peer-to-Peer Content Distribution Technologies. ACM Computing Surveys, Vol. 36, 2004, No. 4, pp. 335–371.

[18] Jung, J. J.: Semantic Preprocessing of Web Request Streams for Web Usage Mining. Journal of Universal Computer Science, Vol. 11, 2005, No. 8, pp. 1383–1396.

[19] Srinivasan, P.—Menczer, F.—Pant, G.: A General Evaluation Framework for Topical Crawlers. Information Retrieval, Vol. 8, 2005, No. 3, pp. 417–447.

[20] Kalfoglou, Y.—Schorlemmer, M.: Ontology Mapping: The State of the Art. Knowledge Engineering Review, Vol. 18, 2003, No. 1, pp. 1–31.

[21] Shvaiko, P.—Euzenat, J.: A Survey Of Schema-Based Matching Approaches. Journal on Data Semantics IV, Vol. 3730, 2005, pp. 146–171.

[22] Dieng, R.—Hug, S.: Comparison of "Personal Ontologies" Represented Through Conceptual Graphs. In: Proceedings of the 13th European Conference on Artificial Intelligence, Brighton, UK, 1998, pp. 341–345.

[23] Euzenat, J.: Brief Overview of T-Tree: The Tropes Taxonomy Building Tool. In: Proceedings of the 4th ASIS SIG/CR workshop on classification research, Columbus, USA, 1994, pp. 69–87.

[24] Stumme, G.—Maedche, A.: FCA-Merge: Bottom-Up Merging of Ontologies. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI '01), Seattle, USA, 2001, pp. 225–230.

[25] Madhavan, J.—Bernstein, P.—Rahm, E.: Generic Schema Matching Using Cupid. In: Proceedings of the 27th International Conference on Very Large Data Bases (VLDB '01), Roma, Italy, 2001, pp. 48–58.

[26] Jung, J. J.: Ontological Framework Based on Contextual Mediation for Collaborative Information Retrieval. Information Retrieval, DOI: 10.1007/s10791-006-9013-5 (in press).

[27] Xiao, L.—Zhuang, Z.—Liu, Y.: Dynamic Layer Management in Superpeer Architectures. IEEE Transactions on Parallel and Distributed Systems, Vol. 16, 2005, No. 11, pp. 1078–1091.

[28] Levenshtein, I.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Cybernetics and Control Theory, Vol. 10, 1996, No. 8, pp. 707–710.

[29] Jung, J. J.—Lee, K.-S.—Park, S.-B.—Jo, G.-S.: Efficient Web Browsing with Semantic Annotation: A Case Study of Product Images in E-Commerce Sites. IEICE Transactions on Information and Systems E88-D, Vol. 5, 2005, pp. 843–850. http://ietisy.oxfordjournals.org/cgi/content/abstract/E88-D/5/843.

[30] Zhuge, H.: Semantics, Resource and Grid. Future Generation Computer Systems, Vol. 21, 2004, No. 1, pp. 1–5.

Jason J. Jung is a full-time lecturer in Yeungnam University, Korea, since September 2007. He was a postdoctoral researcher in INRIA Rhône-Alpes, France in 2006, and a visiting scientist in Fraunhofer Institute (FIRST) in Berlin, Germany in 2004. He received the B. Eng. in computer science and mechanical engineering from Inha University in 1999. He received M. Sc. and Ph. D. degrees in computer and information engineering from Inha University in 2002 and 2005, respectively. His research topics are knowledge engineering on social networks by using machine learning, semantic Web mining, and ambient intelligence.