Personal Knowledge Publishing Suite with Weblog

Ikki Ohmukai
The Graduate University for Advanced Studies
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan

Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan

Kosuke Numa
Yokohama National University
79-1 Tokiwadai, Hodogaya-ku
Yokohama, Japan

E-mail: i2k@grad.nii.ac.jp

Abstract

We propose a personal knowledge publishing system called Semblog that provide an integrated environment for gathering, authoring, publishing, and making human relationship seamlessly. It enables people to exchange information and knowledge with easy and casual fashion and with a variety of communication levels, e.g, three levels of publishing like checking, clipping, and posting. Semblog extends Weblogs by adding flexible but uniform operations for Weblog sites and entries like aggregation and clipping, and facilities for searching and contacting to other Weblog sites. These are realized systematically because of intensive metadata handling.

Categories and Subject Descriptors

H.4.3 [**Information Systems**]: Communications Applications; H.3.3 [**Information Systems**]: Information Search and Retrieval; I.2.4 [**Computing Methodologies**]: Knowledge Representation Formalisms and Methods—*lightweight metadata*

General Terms

Metadata management

Keywords

Weblog, Semantic Web, Information Distribution

1. Introduction

WWW2004, May 17–22, 2004, New York, NY USA.
ACM xxx.xxx.

We propose a personal knowledge publishing system called *Semblog* with Semantic Web techniques and Weblog tools.

Semblog suites provide an integrated environment for gathering, authoring, publishing, and making human relationship seamlessly to enable people to exchange information and knowledge with easy and casual fashion.

We propose two-layered model of information distribution to clarify how we support the process as shown in Fig.1. It is extended form of ”Activities and Relationships Table” by Shneiderman[11].

The first layer has three elements that concern information handling, i.e., collect, create and donate information. However current web only supports each activities respectively and they are not integrated.

The second layer has also three elements that concerns communication handling, i.e., relate, collaborate and present people. Organization of information in the first layer are performed by the communication process of multiple people in the second layer.

Two layers have different roles but are closely related to each other. It implies that support for information activities requires support for communication activities, and vice versa.

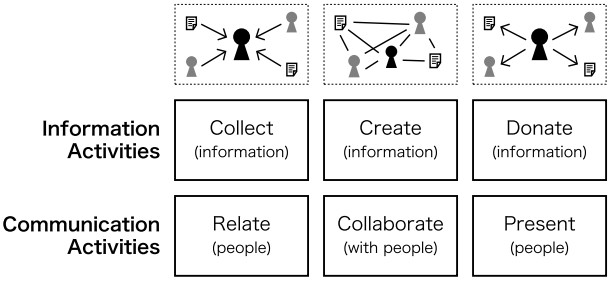


Figure 1. Information and Communication Activities

We focus on Weblog because Weblog is an integrated system for authoring and publishing, while WWW is basically a system for publishing. It means that weblog supports activities in the in-

formation layer integratedly. The next target is to integrate the communication layer and the information layer. Weblog has also a good feature because weblog users tend to refer to each other and form so-called ”weblog communities”. But this process is not supported explicitly by tools.

Semblog suites extend Weblogs by adding flexible but uniform operations for Weblog sites and entries like aggregation and clipping, and facilities for searching and contacting to other Weblog sites. It means that Semblog suites support communication activities as well as information activities

2. Semantic Web and Weblog

2.1 Information Processing with Semantic Web Techniques

We propose a content distribution support system for individuals with Semantic Web techniques. We should consider that the information distribution process does not mean publishing alone but information gathering as a preprocess. In the current web, however, there is no framework to support the whole process of information propagation despite that TBL specify the first world wide web to support both authoring and publishing process equally[3]. Now we could not find practical methods to extract appropriate information from large data source in any other way but search engines. Since we cannot verify an objective validity of a search result, it is hard to say that the engines work effectively at any time.

There is a great hope that the Semantic Web technologies will resolve information overload. According to the manifesto[2], Semantic Web is an environment, which consists of the contents with machine-readable (semantic) tags and the software agents, to realize autonomous information distribution and syndication. Resource Description Framework (RDF)[14] and other ontology definition languages[15] are recommended by W3C as elemental technologies of the Semantic Web and these are now in practical use.

However it is difficult to produce contents with semantic tags because of their complicated syntax and vocabulary. Ordinary people hardly find a merit of semantic annotation because it is a time-consuming task. It is also impossible to annotate the semantic tags to existing enormous information on the Internet. There are some researches about automatic annotation with AI techniques and natural language processing[6] however their effects are still unclear.

In our approach, we use a lightweight metadata format that is RDF Site Summary: RSS[10] to activate the information flow and its activities. RSS is one of the metadata to describe a summary of a web site. It contains general attributions i.e. title and publisher’s name of the web site, and excerpt and updated date of its contents. A number of web sites already publish the RSS, and several applications and services called RSS aggregator are provided based on this trend.

The aggregator collects these RSS from various web site and reform them to show a large amount of contents at a time. There are two types of aggregator, one is standalone applications which are executed on client PCs. The other is aggregation services that run on the internet server and the user access via her/his web browser. The former applications provide rapid browsing of RSS by their flexible user interface and the latter enables the user to access their information wherever she/he is.

Using RSS and the aggregator, it is probably true that the cost of information collection is decreasing, however, finding information set might contain large quantity of noise because of lack of a information extraction process. It is also hard to say that the aggregators support whole process of information distribution since they do not help to create new information.

2.2 Information Creation with Weblog

Recently Weblog (blog) or blogging has come into the spotlight in the World Wide Web[4]. There is no strict definition about Weblog but it is recognized as a web site which consists of miscellaneous notes updated daily[1]. In such sites the authors do not make efforts to knit up these contents and just align them in chronological order. We call these frequently-posted contents as small contents in this paper. Small contents include various subjects including journal, expertise and critique. One of most popular topics is the introductions and comments of the web sites ranging from news sites to the other small contents.

Some Weblog sites attract the attention with their own editorial policy. The authors of Weblog sites reedit the existing web contents by quoting them. Moreover there are new types of Weblogs that criticize the other Weblogs so that these Weblogs are regarded to organize the ”Weblog community”. Now there are millions of weblog sites in the World. It is a surprising number because these people are now active information senders andd distributors as well as information receivers thanks to weblog.

Most of Weblog site uses contents management systems (CMS) called Weblog tools. Weblog tools enable the author to describe and edit the small contents via a web browser and transform the contents form text format to HTML files. These tools are implemented based on MVC (Model / View / Controller) model which is the fundamental concept of web applications. The author defines a view template once then do not have to decorate the contents with various HTML tags. This model decreases the cost of publication remarkably comparing with traditional style which requires local text editor and FTP. This feature contributes abundant production of the small contents. Fig.2 shows typical site with Weblog tool.

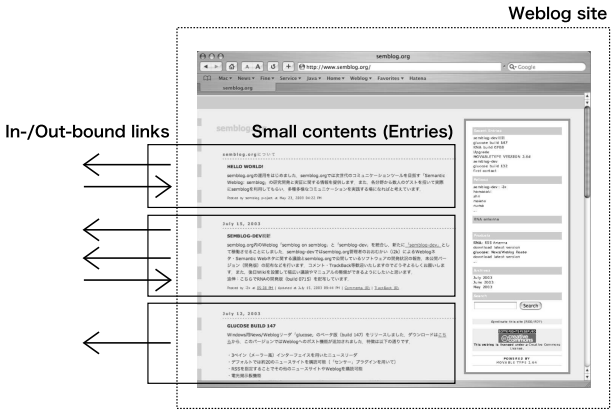


Figure 2. Typical Weblog Site

Weblog tools usually generate RSS automatically. General at-

tributes such as publisher's name are set by the user as a profile. Excerpt and updated date of each content are generated by the tools. Most of distributed RSS generated by these Weblog tools. Main purpose of RSS aggregator is also to browse the Weblog site. Currently the number of news site feeding RSS is expanding.

3. Purpose of Semblog

We propose a personal publishing system with Semantic Web techniques and Weblog tools. The system will support whole process of information distribution which includes gathering and authoring. Furthermore it connects information distribution process of different people seamlessly.

We define the level of interest of information gathering and publishing i.e., "check", "clip" and "post". Our system provides different ways to distribute information depend on the interest level.

"Check" means that the user routinely browses particular web sites and information sources. The user does not know "what" is described at those site exactly but she/he know "what kind" of contents included in these sites. We assume that these "what kind" knowledge are important for information distribution so that our system supports the user to present her/his interest by publishing the list of URIs she/he always accesses.

"Clip" makes shortcut to a content to which the user have strong interest among various contents of "Checked" sites. Our system automatically publishes the list of "clipped" contents. We suppose that "clip" link presents stronger interest of the user than "check" link because it points individual content directly. In addition, contents of "checked" links are changing momentary but those of "clipped" links are persistent (called "permalink").

Online bookmark systems were developed for the same objective. These systems make a backup of "bookmark" or "favorite" of the user's web browser and she/he can obtain the list from anywhere. The user can also make the list to be public or not so the list will be a web content. However that bookmark is only a static list of URIs and is lacking functions for change in time. Therefore these online bookmark systems are not useful for everyone except the user.

Conventional bookmark systems do not distinguish the URIs between to a site and a content. It is necessary to switch the way to distribute these links because they have different level of interest as described above.

"Post" means that the user quotes content, adds some comment, and publishes it as a new information. In that case the user presents not only strong interest but her/his personal opinion. In our system, "post" type of information publication is made by the Weblog tools.

Providing different levels for publishing reduces the psychological burden of publishing. In standard BBSs, it is said that lurkers show feeling of belonging to commuities in sprite of their reluctant behavior[5]. This is important to involve ordinary people into the cycle of information gathering and publishing.

The all process are realized by metadata handling therefore it is open to other systems and applications. It enables seamless connection of information distribution of different people.

4. Semblog Platform

Fig.3 shows the system architecture of Semblog platform. Semblog system consists of the two types of extended RSS aggregator, semblog applications and conventional Weblog tools. Each module exchanges data in RSS format and communicates with XML-RPC protocol[13] for dynamic invocation. We use Movable Type system as Weblog tool[12], which supports RSS and XML-RPC call.

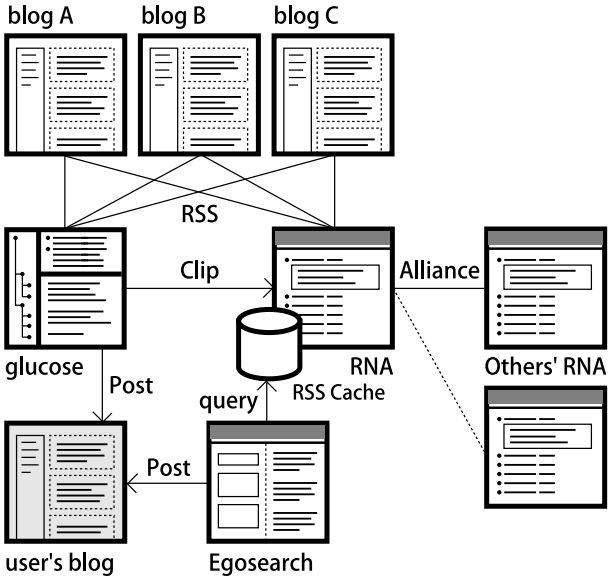


Figure 3. System Architecture

4.1 RNA: RSS Aggregation Service

RNA is an extended RSS aggregator described with Perl CGI. Fig.4 shows a snapshot of RNA. The user puts this script to her/his own web server and operates it. Basic function and interface of RNA are shown below.

Figure 4. RNA: Snapshot

- RSS registration and loading

  The user should register RSS in a configuration interface. She/he once enters the URIs of RSS, RNA obtains these

RSS files via HTTP. The user is able to classify each RSS using categories. The list of registered sites is transformed into RSS and it can be used by the other applications. RNA can also import or export an OPML file which is standard format for the bookmarks.

- Building RSS tree

  RNA parses multiple RSSs from various sites and builds a ”global” RSS tree from each RSS tree. The global tree stores all information obtained by RNA. Next RNA recomposes the global tree into several sub trees according to rule templates. In default setting, RNA reproduces three types of sub tree i.e., list of contents sorted by chronological order and updated contents in each site. The user can describe own template script to generate a new sub tree.

- RSS redistribution

  Generated sub trees are distributed as new RSS and are visualized by server / client side XSLT engine using XSL stylesheet. RNA can also transform these RSS into HTML with its own template system. With these templates, the user is easily able to customize an output style rather than XSLT since the syntax of template is similar to HTML.

- Content clipping

  The user can register a content to a clip list in one click. Clipped contents are stored in ”clipped” RSS tree and it is published like the other RSS. Each element of RSS tree is changing momentary but those of ”clipped” tree are persistent.

- Updating

  It is necessary to get RSS and build trees occationally since its contents are changeable with update of its distributor. RNA can update periodically by a cron interface of the server. Update interface can be called both manually and remotely by XML-RPC message the Weblog tools automatically generated. The latter feature enables the author of a Weblog to notify her/his new content to RNA.

- TrackBack tracing

  RNA queries each registered Weblog tool for that each content has TrackBack links (reverse link provided by Weblog tools) or not. If exist, RNA extracts these links and adds them to RSS tree.

- Sanitizing and caching RSS

  RNA checks syntax of acquired RSS and corrects them if they are not valid. Currently three versions of RSS are proposed i.e., 0.91, 1.0 and 2.0. RNA converts all versions into 1.0, which is based on RDF model. Every RSS is cached and decomposed to attach to each content as a fine-grained metadata.

”Antenna” services are already provided to check update of registered sites. Conventional antenna obtains HTTP LastModified header of each URI and sorts the site list by chronological order. Using RNA the user can get richer information than the antenna since RSS contains not only updated time but its excerpt and so on. RNA also enables the user to edit information in the contents level such as clip and TrackBack tracing.

4.2 Glucose: Standalone RSS Aggregator

Glucose is also an extended RSS aggregator but a standalone program for Windows. Fig.5 shows a snapshot of Glucose. Different from orthodox aggregator, Glucose is developed to support information distribution process by coordination with RNA. Main functions and interfaces of Glucose are shown below.

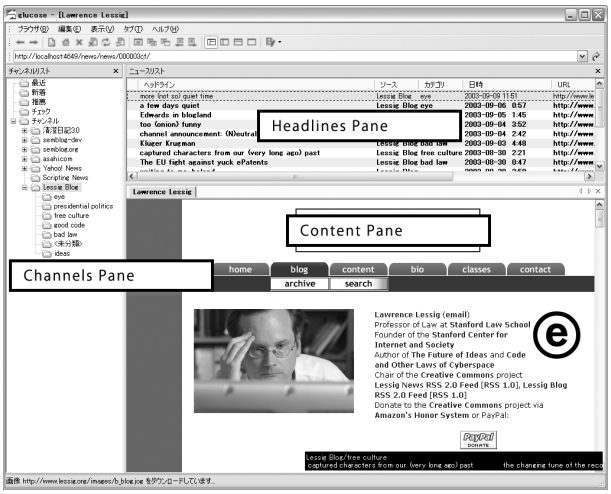


Figure 5. Glucose: Interface

- RSS registration

  Like in RNA, the user registers URIs of RSS or OPML site list. Glucose can access several news sites without RSS by ”sensor” script which extracts articles and converts them to RSS.

- Three-pane interface

  Glucose has three pane interface. The left pane shows ”RSS Channels” which is subscribed by the user. The upper right pane indicates the headline list of contents including title, updated time, source and category. The lower right pane shows original contents.

- TrackBack tracing

  Glucose can extract TrackBack links from each content. Obtained links are shown below corresponding entry in headline pane like ”Re:” in an emailer.

- Posting to Weblog

  With a Weblog editor interface in Glucose, the user can post an entry to her/his Weblog if she/he has strong interest for content. This interface uses XML-RPC protocol.

- RNA clipping

  The user can post content to the clip list of own RNA using XML-RPC.

- P2P content recommendation

  Glucose runs as a P2P servant and the users automatically construct P2P network. In this network, several contents

read by someone are distributed randomly and the other user may receive these contents as a recommendation. The user can clip or post it to her/his own system and can also register URI of its RSS to the site list.

We distribute RNA and Glucose in our web site (`http://www.semblog.org/wiki/?en`). About 1000 users downloaded RNA and over 10000 users downloaded Glucose from September 2003.

5. Application on Semblog Platform

Using functions on Semblog platform, we develop a new type of recommendation and retrieval systems.

5.1 RNA Alliance

Each RNA has XML-RPC interface that can send and receive its data dynamically. RNA alliance is a content recommendation system based on cooperation of multiple RNA.

We use Friend Of A Friend: FOAF metadata to identify each RNA. FOAF is RDF-based metadata format for describing human relationship. Besides the basic elements such as name, email and URI of the user, FOAF provides a statement that user X knows user Y.

The current version of RNA can generate FOAF data. RNA also has an interface for FOAF management to extend social network easily. We call this method as "FOAF TrackBack".

First the user X enters an RNA URI of the user Y in her/his own FOAF manager. The manager X asks the manager Y to acquire the FOAF data of Y, and writes "X knows Y" link in its FOAF. The manager Y records "Y isKnown X" link in its FOAF and notifies to the user Y. If the user Y agrees, her/his manager registers "Y knows X" link. Repeating this process, a personal network of the user is constructed. Following recommendation methods are performed in the network.

5.1.1 Collaborative Recommendation

Collaborative recommendation is based on difference of registered sites or clips among multiple RNAs. At first it calculates similarities: S_i between the user's RNA: R_0 and each RNA on the personal network: $R_1, \ldots, R_n$ by following formula:

$$S_i = \frac{C_i}{N_0 + Ni}$$

where N_i is the number of the registered site of R_i, and C_i is the number of common site in R_0 and R_i. Each RNA has the list of URIs: $R_i = \{u_0, \ldots, u_k\}$.

The system gives recommendation score: $V(u)$ to each URI by following formula:

$$V_i(u) = \begin{cases} S_i & \text{if } u \in R_i \\ 0 & \text{if } u \notin R_i \ (i = 1, \ldots, n) \end{cases}$$

$$V(u) = \frac{\sum_{i=i}^{n} V_i(u)}{n}$$

This score is used for recommendation to R_0's user if URI u is not included in R_0.

The system shows the list of recommended URIs sorted by the score. The user can add these URI to her/his own "check" list.

5.1.2 Categorical Similarity and Recommendation

The user of RNA can categorize the sites and clips she/he registered. Using this category, relationship of *interest* among users can be identified and recommended. We apply a recommendation method based on a categorical similarity to this objective[9].

5.1.3 Relational Filtering

Relational filtering method realizes access control of information using the categorized social network. By merging the personal network of every user, a large human network like a small world is constructed. This method extracts multiple communities from this network and enables information sharing in single community[7]. The user can manage her/his contents which have various level of disclosure through an unified interface.

5.2 Egocentric Search

Egocentric search provides subjective search which collects and evaluates collect information for each user[8]. We think that it is suitable for handling of the small contents like Weblog entries because connection is usually more subjective than other WWW pages.

The users daily write and post contents to their Weblog sites with the editor interface (Fig.8). Egocentric search interface scans a content that the user is posting. If this content contains a hyperlink, the editor acquires whole content and RSS of the link, and constructs an entry network around the user's content (Fig.6(left)). This network indicates not only relations of contents but also human relationships because all entries on Weblog are owned by some authors. Our method organizes a document network into person-based network (Fig.6(right)). Each path of the personal network is weighted relatively to the frequency of citation. Once the user cites some site as a topic in the new content (entry), the editor interface performs egocentric search and shows the result.

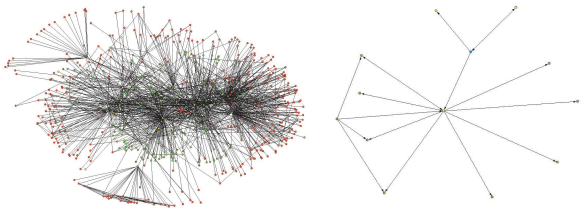


Figure 6. Document-based(left) and Person-based(right) Network

- Relative Chain Search

  Relative chain search returns the contents which is directly linked with the entry cited by the authoring content. This model is based on a simple model but consequently it seems most trustful. (Fig.7(a))

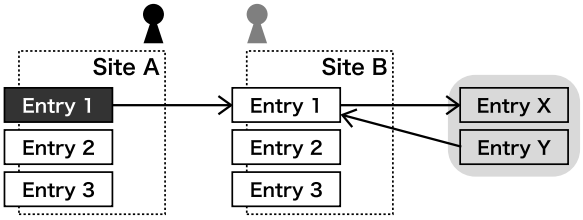
- Co-citation Search

  Co-citation search discovers the entries that link the same contents as the authoring entry links to. Co-citation entries are retrieved from the RSS cache and the search result contains the weight of authors. (Fig.7(b))

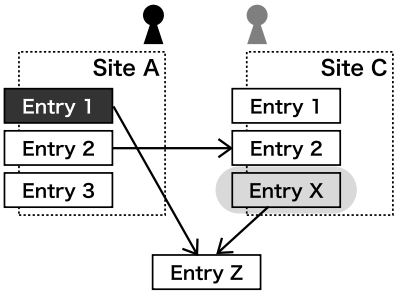
- Keyword Search

  Keyword search method picks up the entries by keyword matching from the RSS cache. Unlike the conventional search engines, our method targets only related sites around the user's Weblog. (Fig.7(c))

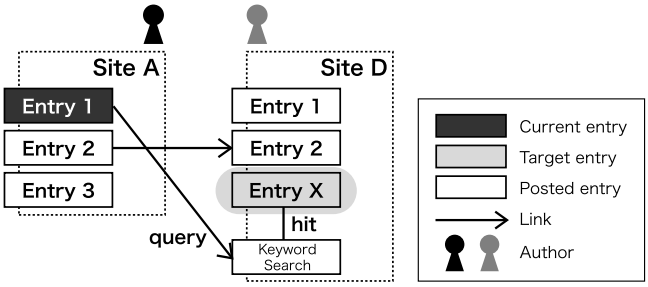
The user read search results by these methods and can append the link of some helpful contents to contents currently describing. This process may enrich the user's content and change the search result of the system.



(a) Relative Chain Search



(b) Relative Co-citation Search



(c) Relative Keyword Search

Figure 7. Egocentric Search Methods

6. Conclusion

In this paper we propose a personal publishing system with Semantic Web techniques and Weblog tools. We use a lightweight metadata format like RSS to activate the information flow and its activities. We define three level of interest of information gathering and publishing i.e., "check", "clip" and "post" and provide several way to distribute information depend on the interest level. Our system called Semblog platform consists of two types of extended content aggregator and information retrieval / recommendation applications. The system will support not only content publishing process but also information gathering and authoring processes synthetically.

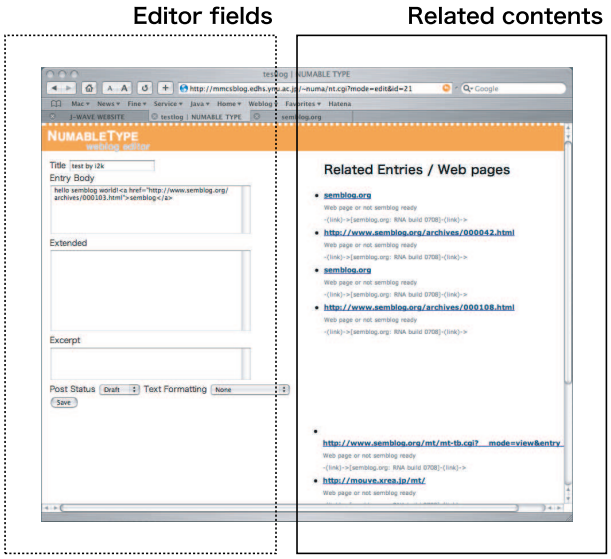


Figure 8. Snapshot of Proposed System

7. Additional Authors

Additional authors: Masahiro Hamasaki (The Graduate University for Advanced Studies) and Shin Adachi (Waseda University).

8. REFERENCES

[1] E. Aimeur, G. Brassard, and S. Paquet. Using Personal Knowledge Publishing to Facilitate Sharing Across Communities. *Workshop on (Virtual) Community Informatics, Held in conjunction with the Twelfth International World Wide Web Conference (WWW2003)*, 2003.

[2] T. Berners-Lee. A roadmap to the Semantic Web. `http://www.w3.org/DesignIssues/Semantic.html`, 1998.

[3] T. Berners-Lee. *Weaving the Web*. HarperCollins, 1999.

[4] R. Blood. *We've Got Blog: How Weblogs are Changing Our Culture*. Perseus Publishing, 2002.

[5] B.Nonnecke and J.Preece. Shedding light on Lurkers in Online Communities. *Ethnographic Studies in Real and Virtual Environments: Inhabited Information Spaces and Connected Communities*, pages 123–128, 1999.

[6] S. Dill, N. Eiron, D. Gibson, and et al. SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation. *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)*, 2003.

[7] I.Ohmukai and H.Takeda. Social Scheduler: A Proposal of Collaborative Personal Task Management. *Proceedings of Web Intelligence (WI2003)*, 2003.

[8] I.Ohmukai, K.Numa, and H.Takeda. Egocentric Search Method for Authoring Support in Semantic Weblog. *Workshop on Knowledge Markup and Semantic Annotation (Semannot2003), Held in conjunction with the Second International Conference on Knowledge Capture (K-CAP2003)*, 2003.

[9] M.Hamasaki and H.Takeda. Find better friends? – re-configuration of personal networks by the neighborhood matchmaker method –. In *The International Workshop on Semantic Web Foundations and Application Technologies (SWFAT)*, pages 73–76, 2003.

[10] RDF Site Summary 1.0 Specification Working Group. RDF Site Summary (RSS) 1.0. `http://web.resource.org/rss/1.0/spec`, 2001.

[11] B. Shneiderman. *Leonardo's Laptop: Human Needs and the New Computing Technologies*. MIT Press, 2002.

[12] Six Apart. Movable Type. `http://www.movabletype.org/`, 2003.

[13] UserLand Software. XML-RPC Specification. `http://www.xmlrpc.com/spec`, 1999.

[14] World Wide Web Consortium (W3C). Resource Description Framework (RDF) Model and Syntax Specification. `http://www.w3.org/TR/REC-rdf-syntax`, 1999.

[15] World Wide Web Consortium (W3C). OWL Web Ontology Language Overview. `http://www.w3.org/TR/owl-features/`, 2003.