Egocentric Search Method for Authoring Support in Semantic Weblog

Ikki Ohmukai
The Graduate University for Advanced Studies
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
i2k@grad.nii.ac.jp

Kosuke Numa
Yokohama National University
79-1 Tokiwadai, Hodogaya-ku, Yokohama-shi
Kanagawa, Japan
d02hc038@ynu.ac.jp

Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
takeda@nii.ac.jp

ABSTRACT

In this paper we propose egocentric search methods based on the concept of "Information and Communicate Activities Navigation (ICAN)" and an authoring support system for Weblog (blog). ICAN regulates the human activities from a viewpoint of information and communication support. We introduce the idea of "Collect" and "Relate" in the ICAN table into the information retrieval and the search method which uses contents and human relationship produced by daily blogging. Our egocentric methods provide more subjective search result than the conventional engines. We apply the methods to improve the quality of the small contents made with Weblog tools.

Keywords

Social network, egocentric search, Weblog

1. INTRODUCTION

1.1 From ICT to ICA

Computers and networks enrich and facilitate our life so that they now become indispensable for our life. They sometimes enhance our traditional daily activities with their increasing computing and networking power like documenting and communicating with other people, and sometime offer new ways for our activities with new technologies like WWW.

On the other hand, most people become to live with worry that unceasing improvement of computers and networks and installation of new software technologies would change their life and business.

It is not because of such technologies themselves but because of our vision to technologies. We are so eager to develop new technologies that we almost loose the original mission for development of technologies, i.e., technologies just for us. Shneiderman pointed out that we should shift our vision from "old computing" to "new computing". He explained it in his recent book as follows[7]; "The old computing was about what computers could do; the new computing is about what users can do. Successful technologies are those that are in harmony with users' needs. They must support relationships and activities that enrich the users' experiences."

We should shift our focus from information and communication technologies (ICT) to information and communication activities (ICA). We should investigate what are human activities on information and communication and how we can assist people in these activities.

K-CAP2003 Knowledge Markup & Semantic Annotation Workshop '03 Sanibel Island, Florida USA

1.2 Information and Communication Activities

Human activities on information such as collecting information and communication such as contacting to people are only a part of human activities but they become to play an important role more and more in modern life.

They include various kinds of activities. Shneiderman shows a simple and therefore understandable model called ART (Activities and Relationships Table) for them[7]. One axis of the table is activity category, i.e., Collect (information), Relate (Communication), Create (Innovation), and Donate (Dissemination). The other is category of relationship, i.e., Self, Family and friends, Colleagues and neighbors, and Citizens and markets. We agree with relationship categories, while we think that activity categories should be elaborated more because information handling and communication among people are mixed.

To explicate the difference, we propose two-layered model as an extension of his model shown in Fig.1. The first layer has three elements that concern information handling, i.e.,
Collect (information)
Create (information)
and
Donate (information).

It shows user-centered view of life cycle of information. Information is collected, then new information is created based on the collected information, and finally created information is donated to the society for future creation. It should be noted that new information is seldom created from scratch

but created based on existing information.[1]

The second layer has also three elements that concerns communication handling, i.e.,
Relate (people)
Collaborate (with people)
and
Present (people).

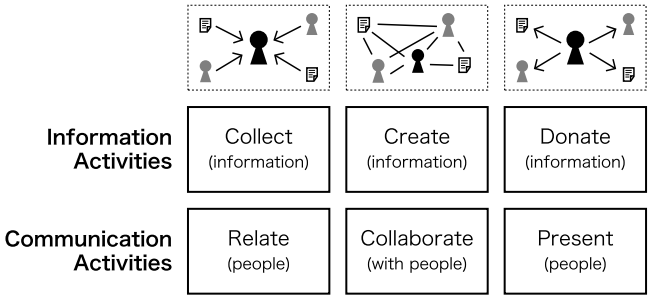

Figure 1: Information and Communication Activities

It is communication-centered view of the above process. People establish relationship with other people, then collaborate with them to create new information, and finally present themselves as donor of new information. Having both information and communication layers is not redundant. What we refer as "information" in the context of computer technologies is stored data in computers, while human is the source of "information" in the broader sense, i.e., human can offer information dynamically. We should consider communication in order to include the function "human as information source". This parallel view of information and communication activities has thus six categories as activities. Ideally all categories should be supported by computers. Some categories like Collect is well investigated, but others are not. In particular, the three categories in the communication layer should be investigated more.

We aim to investigate information and communication activities and support people in the all categories of the activities. We call such support "information and communication activity navigation (ICAN)". It helps people to create new information by guiding information space and human network.

2. WEBLOG AND SEMANTIC WEB

2.1 Weblog and Small Contents

Recently Weblog (blog) or blogging has come into the spotlight in the World Wide Web[3]. 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 that include 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". There are more than 100,000 Weblogs in the United States Weblogs make people to change from information receiver into information sender and distributor.

Most of Weblog site uses the contents management system (CMS) called Weblog tool. 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.

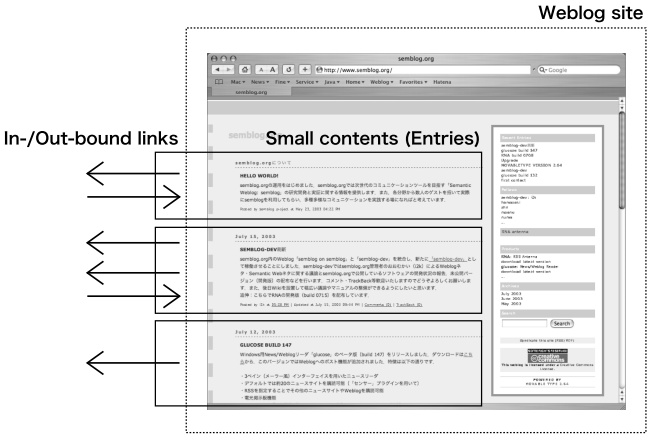

Figure 2: Typical Weblog Site

A huge number of the small contents and citations among Weblog communities are increasing day by day. Some efforts such as topic discovery, trend analysis and content ranking are applied to these large amount of information.

Weblog facilitates to publish the small contents, however, the cost of contents arrangement and classification remains extremely high. Most of Weblog tools are specialized to enhance the convenience of publishing by transforming from text to normal HTML files. There are already billions of HTML files on the Internet so that people are facing troubles both to discover her/his objective articles and to use information effectively. Therefore it is skeptically considered that Weblog will just accelerate this trend.

2.2 Semantic Web

There are great hopes that the Semantic Web technologies will resolve our current condition of information overload. According to the manifest[2], the Semantic Web is an environment, which consists of the contents with machine-readable (semantic) tags and the software agents, to realize

[1]We do not claim that information creation is just combination of existing information. Rather creativity arises with understanding and interpretation of existing information.

autonomous information distribution and syndication. Resource Description Framework (RDF)[13] and other ontology definition languages[14] 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[4] however their effects are still unclear.

In this research we aim to integrate both technologies, Weblog and Semantic Web, to achieve the platform which enables to share, reuse and reedit our small contents. We provide a new function to the contents management systems like Weblog tools and allow a semantic annotation to existing contents semiautomatically. Hereby it is possible to apply the effects of the Semantic Web to all of the contents on the Internet. As a result, links in Weblogs are transformed into semantic annotations so that the Weblog contents and the web contents refered by Weblogs can be worked as Semantic Web.

In this paper we propose the egocentric search methods with relational annotation and description support system for Weblogging as a first stop of our project.

3. CONCEPT OF EGOCENTRIC SEARCH

As mentioned above Weblog tools contribute to increase the amount of small contents. However these tools do not help improve the quality of contents. There is fear that flood of "junk" contents sweeps the Internet as a result.

We explore new ways of authoring support for blogging. Most of the contents on Weblogs are closely related to the other contents from different sites. Therefore it is important for Weblog authors to know similar or related contents to currently describing contents.

We illustrate the procedures of authoring support according to the ICAN table discussed before[2]. We assume the user usually refers and comments on several contents on the Internet in her/his Weblog site. These activities associate the user's contents with the other contents. We define these activities as "Collect" of information. We also consider these linkages not only as relationships between each contents but relations between the authors who have these contents. This fact corresponds to "Relate" in the ICAN table.

Thus the contents and human network are built around the user with these Collect and Relate activities.

In case of describing a new content, that is the "Create" activity, the authoring support system will retrieve the contents and human network close to the new content. The closeness among the small contents is calculated not as the score of semantic similarity but as the distance from the user on her/his network. We call these search methods as "Egocentric search"[3]. The user will add a new link onto the authoring content and polish the content out with the search results. Iteration of these processes may improve the quality of each content in Weblogs.

4. IMPLEMENTATION

4.1 System Architecture

We implemented authoring support system for Weblogs with our proposed method as shown in Fig.3. The system consists of three modules as follows.

- Weblog tool

  We use ready-made Weblog tool as an infrastructure of our system. In a number of Weblog tools released recently we introduce Movable Type system[8] which is one of the most popular tools. Movable Type can communicate CGI (Common Gateway Interface) programs via MetaWeblog API[11] based on XML-RPC protocol[10].

- Editor

  We developed an Editor interface as Web application. It can connect the Weblog tools and the Cache Database and execute the egocentric search methods. We will explain it in detail in following section.

- Cache Database (DB)

  Cache DB stores all contents which is linked in the user's Weblog.

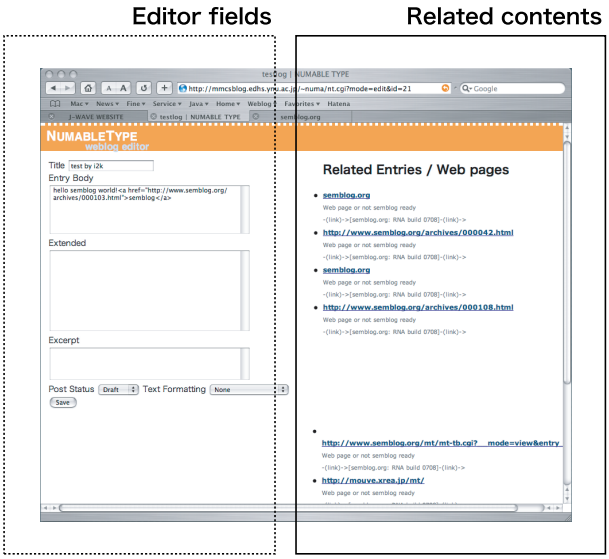

Figure 3: Snapshot of Proposed System

[2]Donate and Present are achieved as the original functions of Weblog that are one of most excellent functions among other information publishing tools.

[3]The word "egocentric" is borrowed from Social Network Analysis[12]. In Social Network Analysis, sociocentric and egocentric network analysis provide two distinctive views for network where the former concerns the nature of the whole network and the latter is focused in network of individuals. Search engines like google shares the same view with the former, and our approach with the latter.

4.2 Extension of RSS

Most of Weblog tools generate a RSS (RDF Site Summary[4])[5][6] file automatically. RSS is an XML-based metadata format for describing an abstract of Web pages. Basic elements of RSS are shown as follows:

- <channel>

  <channel> element contains information of entire Web site such as name of Web site and author.

- <item>

  <item> element describes metadata of a content in a Web site. In a single Weblog site there are multiple contents (entries) so that their titles and update times are described in the item elements correspond to them.

The RSS, which is the unified regular description format for Web sites, is now propagating from Weblogs to enterprise sites so that people can incorporate various contents into her/his Weblog using the RSS called "Content Syndication". However the RSS is for describing content relation in single site, not for inter-site relationship. Therefore we extend the concept of the RSS to describe metadata like inter-site relation. We call this metadata "RDF Content Summary (RCS)" that is annotated to every entry in Weblog. The RCS uses following modules in addition to the elements of the RSS.

- <semblog:outlink>

  <semblog:outlink> module represents an ordinary hyperlink. "semblog" indicates the XML namespace we originally defined. Instance URI of the element is extracted from <a href> tag in a HTML document.

- <semblog:inlink>

  <semblog:inlink> module shows the URI which is provided as a reverse link (also called "TrackBack"[9]) by several Weblog tools. For example, the author of Weblog B publishes an entry 1 and pings to the entry X in Weblog A, then the system of Weblog A recognizes this message and appends the URI of entry 1 to the entry X. This type of reverse link is regarded as a metadata or an annotation of entry X.

As just described, the RCS maintains both link information of related contents and metadata of entry itself. Furthermore the Movable Type system can generate equivalent RCSs for each entry simply with template. Fig.4 shows an example of RCS file.

4.3 Search Methods

In this section we explain the search methods in the circumstance described previously.

The users daily write and post the small contents to their Weblog sites with the Editor program. The Editor scans the text strings of these contents each times they are posted. If content contains a hyperlink, the Editor acquires whole content and RCS of the link and store it in the Cache DB. Then the Editor extracts hyperlinks from stored contents and constructs a entry network around the user's contents

[4]RSS is also a acronym of "Rich Site Summary" or "Really Simple Syndication".

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:semblog="http://www.semblog.org/ns/semblog/0.1/"
  xmlns="http://purl.org/rss/1.0/">

<channel rdf:about="http://www.semblog.org/i2k/archives/000046.html">
<title>Expantion of RSS</title>
<link>http://www.semblog.org/i2k/archives/000046.html</link>
<description>RSS must be extented to describe the relation among ...</description>
<dc:language>ja</dc:language>
<dc:creator>i2k</dc:creator>
<dc:date>2003-07-18T22:36:23+09:00</dc:date>
<items>
<rdf:Seq>
<rdf:li rdf:resource="http://www.w3.org/2001/sw/" />
<rdf:li rdf:resource="http://www-kasm.nii.ac.jp/~numa/mt/archives/000161.html" />
</rdf:Seq>
</items>
</channel>

<!-- out-bound link -->
<item rdf:about="http://www.w3.org/2001/sw/">
<link>http://www.w3.org/2001/sw/</link>
<semblog:outlink>http://www.w3.org/2001/sw/</semblog:outlink>
</item>

<!-- in-bound link (trackback) -->
<item rdf:about="http://www-kasm.nii.ac.jp/~numa/mt/archives/000161.html">
<title>ba-log has accepted</title>
<link>http://www-kasm.nii.ac.jp/~numa/mt/archives/000161.html</link>
<description>Today our proposal about "ba-log" has accepted by...</description>
<dc:subject>misc</dc:subject>
<dc:creator>numa</dc:creator>
<dc:date>2003-07-20T15:08:17+09:00</dc:date>
<semblog:inlink>
  http://www-kasm.nii.ac.jp/~numa/mt/archives/000161.html
</semblog:inlink>
</item>

</rdf:RDF>
```

Standard RSS element
Reverse link
Originally defined for RCS

Figure 4: Example of RDF Content Summary

with these link information. This network indicates not only relations of contents but also human relationships because all entries on Weblog are owned by an author. Each path of the human networks is weighted relatively to the frequency of citation.

Once the user cites some site as a topic in the new content (entry), the Editor program performs three types of egocentric search and shows the result.

- 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.5(a))

- Relative Co-citation Search

  Relative co-citation search discovers the entries that link same contents as the authoring entry links to. Co-citation entries are retrieved from the Cache DB and the search result contains the weight of authors. (Fig.5(b))

- Relative Keyword Search

  Relative keyword search picks up the entries by keyword matching from the Cache DB. Different from the conventional search engines, our method targets only related sites around the user's Weblog. (Fig.5(c))

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

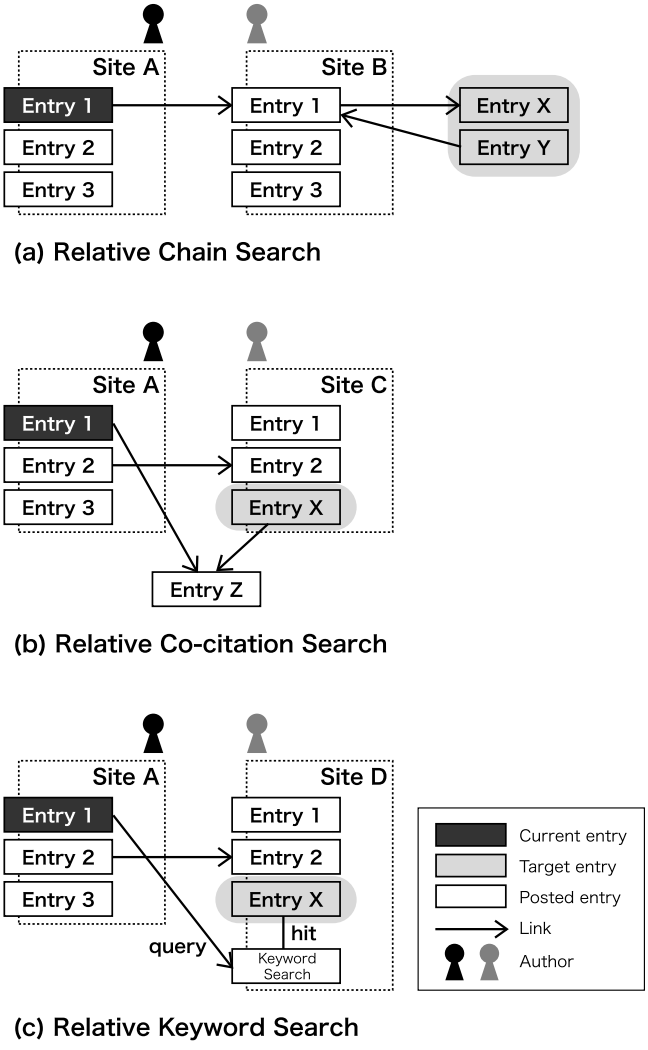

Figure 5: Egocentric Search Methods

5. CONCLUSIONS

In this paper we propose egocentric search methods based on the concept of "Information and Communicate Activities Navigation (ICAN)" and an authoring support system for Weblog (blog). ICAN regulates the human activities from a viewpoint of information and communication support. We introduce the idea of "Collect" and "Relate" in the ICAN table into the information retrieval and the search method which uses contents and human relationship produced by daily blogging. Our egocentric methods provide more subjective search result than the conventional engines. We apply the methods to improve the quality of the small contents made with Weblog tools.

We will develop the metadata format and the extensions of RSS moreover to represent the relationships among the contents explicitly. We will also provide an egocentric search method based on a pure P2P model, which does not depend on the cache DB. Future model may create a foothold of the Semantic Web for "the rest of us". We will take an experimental proof of our system with large Weblog communities in the near future.

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

[4] 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.

[5] B. Hammersley. *Content Syndication with RSS*. O'Reilly & Associates, 2003.

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

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

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

[9] B. Trott and M. Trott. TrackBack Technical Specification. `http://www.movabletype.org/docs/mttrackback.html`, 2002.

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

[11] UserLand Software. MetaWeblog API. `http://www.xmlrpc.com/metaWeblogApi`, 2002.

[12] B. Wellman. An Egocentric Network Tale: Comment on Bien et al. *Social Networks*, 15:423–436, 1993.

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

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