Egocentric Search based on RSS

Kosuke Numa[1,2], **Ikki Ohmukai**[1,2], **Masahiro Hamasaki**[1,2], **Hideaki Takeda**[2,1]
[1] The Graduate University for Advanced Studies,
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
{numa, i2k, hamasaki}@grad.nii.ac.jp
[2] National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
takeda@nii.ac.jp

Abstract

We propose an egocentric search method which evaluates the importance of information by the distance between the user and the information. Personal networks around the user can be extracted by using Weblog and RSS. We performed an experiment and found this distance is related to similarity between documents. We applied the method to a Weblog editor to support Weblog authoring.

1 Introduction

In *Semblog* project, we propose a personal knowledge publishing environment for gathering, authoring and publishing information [Ohmukai *et al.*, 2004]. As a part of the project, we propose a new way of search called *egocentric search*, which can support the authoring process by collecting related documents.

When writing a document, we often refer to the related documents written by others. With current search engines (e.g., Google), it is hard to find such subjective and contextual information. Personalized search may find better results, but some sort of user profile is required. Designing profiles is difficult and also maintaining profiles requires much effort to users. In order to solve this problem, we propose egocentric search method. This method uses relationship between the user and the target information on the personal network around the user [Ohmukai *et al.*, 2003]. Our method provides more subjective results than conventional search methods.

2 Egocentric Search

2.1 Concept of Egocentric Search

In our egocentric search method, we consider the Web as a document network, and assume a set of documents which represent a user on the Web. The importance of the target information is measured by the distance between the user and the information on the network.

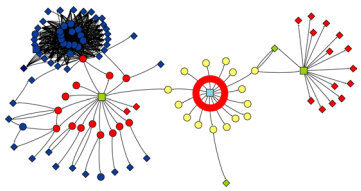
We defined an egocentric network as a local network around a user extracted by tracing hyperlinks from the user's documents. The search target documents are ranked by the distance from the user on the network graph. The search results are shown ordered by the distance.

2.2 Site as Person

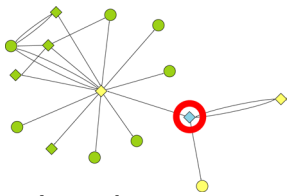
To realize the egocentric search method, we regard a Weblog site as a representation of a person on the Web. Generally, it is difficult to distinguish a personal Web site for each individual user from many Web pages, but it is much easier with RDF Site Summary (RSS) [RSS, 2000]. RSS is becoming popular since Weblog tools create RSS files automatically.

An RSS summary contains a channel element and some item elements. In typical RSS files made by Weblog tools, a channel indicates the Weblog site itself, and each item indicates an individual document which is called as an entry. Accordingly we can find a parent site for each entry. RSS files often include the author information with Dublin Core vocabulary [DCMI, 1999]. Thus, we can find a correspondence between the author and the document.

We extract a personal network from a document network connected by hyperlinks and TrackBacks [Trott *et al.*, 2002]. Based on Weblog and RSS, it is easy to extract a personal network.



(a) Document Network



(b) Personal Network

Figure 1: Example of Obtained Egocentric Network

3 Experiment and Analysis

We performed an experiment to verify a hypothesis that the importance of the information for a user is related to the distance between the user and the information. We extract 4 egocentric networks by crawling around from 4 Weblogs. Figure 1 shows the example of the obtained network. Figure 1(a) stands for a document-based network, and (b) shows a person-based network. The network structures are simplified by binding the documents to the people. The colors of the nodes indicate the distance from the starting Weblog circled in red. Table 1 indicates the numbers of nodes and links in the networks.

	Weblog A		*Weblog B*		*Weblog C*		*Weblog D*	
	Document	Person	Document	Person	Document	Person	Document	Person
Nodes	2966	930	106	15	105	16	3159	734
Links	23053	2260	1120	27	320	27	79174	2097

Table 1: Numbers of Nodes and Links in the Obtained Egocentric Networks

We calculated the document similarity between the user's entry and each document in the distance *n*, and average the values for each distance. This similarity is defined by the similarity of appearance of words. It is calculated by SMART algorithm [Salton *et al.*, 1983]. Figure 2 indicates the correlation between the distance and the average of document similarity. Closer documents to the starting Weblog tend to be scored higher similarity. This fact shows that the proposed method can find suitable information for authoring.

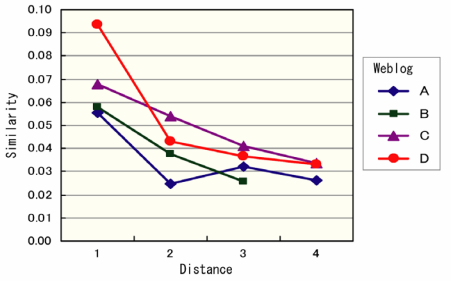


Figure 2: Relation between Distance and Averages of Similarity

4 Implementation on Weblog Editor

We implement our egocentric search method on a Weblog editor, which finds and shows related contents to the currently writing document by the method. Figure 3 shows the system architecture. The system consists of 3 parts: (1) crawler which extracts the egocentric network, (2) cache database which stores the surrounding documents, and (3) editor which recommends related contents and posts the edited document to the user's Weblog site.

Figure 4 shows the snapshot of the system. The editing field is placed on the left-hand side, and the recommendation field on the right-hand side. The user reads the recommended contents and can make hyperlinks to them if she/he likes. This action will change her/his egocentric network. Therefore, new search results will be shown. It is expected that the user can brush up the document and the network incrementally by repeating this process.

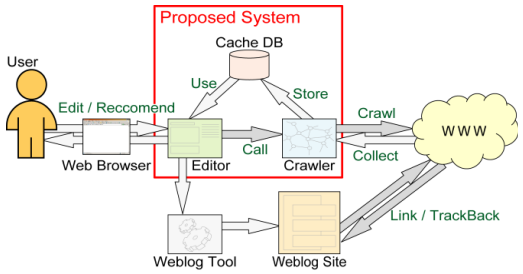


Figure 3: System Architecture

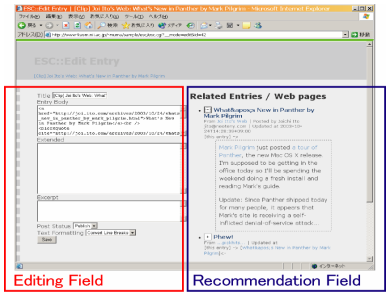


Figure 4: Snapshot of the Weblog Editor

5 Conclusion

This paper presented the new way of search for Weblog authoring. We proposed the egocentric search method, which evaluates the importance of the information by the distance between the user and the information. We assume a Weblog site as a representation of its author described with RSS metadata. Personal relationship can be extracted by tracing hyperlinks and TrackBacks. The experimental results showed that our method can find suitable information. We applied the method to the Weblog editor.

References

[Ohmukai *et al.*, 2004] I. Ohmukai, H. Takeda, M. Hamasaki, K. Numa, S. Adachi. Metadata-driven Personal Knowledge Publishing. In *Proceedings of 3rd International Semantic Web Conference (ISWC2004)*, 2004 (to appear).

[Ohmukai *et al.*, 2003] I. Ohmukai, K. Numa, H. Takeda. Egocentric Search Method for Authoring Support in Semantic Weblog. *Workshop on Knowledge Markup and Semantic Annotation (Semannot2003)*, 2003.

[RSS, 2000] RDF Site Summary 1.0 Specification Working Group. RDF Site Summary (RSS) 1.0. http://web.resource.org/rss/1.0/spec, 2001.

[DCMI, 1999] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://dublincore.org/documents/dces/, 1999.

[Trott *et al.*, 2002] B. Trott, M. Trott. TrackBack Technical Specification. http://www.movabletype.org/docs/mttrackback.html, 2002.

[Salton *et al.*, 1983] G. Salton, M. J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, 1983.