Discovery of Shared Topics Networks among People
— A Simple Approach to Find Community Knowledge from WWW Bookmarks —

Hideaki Takeda[1,2], Takeshi Matsuzuka[2⋆] and Yuichiro Taniguchi[2]

[1] National Institute of Informatics,
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
takeda@nii.ac.jp
[2] Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-01, Japan
yuichi-t@is.aist-nara.ac.jp

Abstract. In this paper, we propose a system called *kMedia* that can assist users to form knowledge for community by showing shared topics networks (STN) among them. One of the important aspects to know each other is to know topics interested by others and relationship between her/his and others' topics. kMedia can use a simple but effective way to find them. It uses folders in WWW bookmarks as interested topics and can calculate their relations by evaluating similarity of WWW pages under folders. The results are displayed in two ways. One is to show relationship among users by shared topics networks, i.e., a user is connected to the other through both her/his topics and the other's topics that are related to her/his ones. A user can know what kind of relations to others s/he can have, and more precisely know what are counterpart of her/his topics for others. The other way is to show recommended pages for pages in users' bookmarks. Recommended pages are selected from others' bookmarks, and it is the primary result of similarity evaluation among pages by contents. A user can use this result just as recommendation for her/his bookmarked pages or use checking how her/his bookmarked pages are related to others. We tested this system in an experiment with actual bookmark data. Discovery of related topics among users are evaluated as good enough in spite of bad results for recommendation of pages. This result tells that our approach to find common topics among users is effective and practical.

1 Introduction

Although the number of the World Wide Web users is increasing every year and available information is also increasing so rapidly, we are getting frustrated to join WWW networks. Compared to rapid growth of users and contents, our

⋆ Currently Toppan Printing Co.,Ltd.

capacity to access them is almost invariable. We seek better ways to improve our capacity for it. We focus on relationship among people to solve this problem.

Even in the actual society, we already have problems of information flood. Then how can we solve them? One of the key issue to solve them is usage of relationship among people. For example, if one of your friends recommends some TV program to you, you may watch it. Of course just "friend" may not be enough. It depends on what relationship between her/him and you, e.g., a close friend or not, a trustful friend or not, and so on. The most important aspect for relationship for information selection is what and how much they can share interest. We have such knowledge when we are joining a community, e.g., who is the appropriate person to ask this question and how much other members would be interested in a specific topic. This community knowledge can save people from information flood and guide them to access appropriate information in appropriate amount.

In this paper, we propose a system called kMedia that can support users to form community knowledge by showing shared topics networks (STN) among them. A shared topics network is a network where each user can be associated to other user via topics owned by both users. A node represents a user or a topic, and a link represents either a relation between a topic and its owner or a similarity relation between topics of different users. Existence of a path from a user to the other shows that there is a shared topic between them. A shared topic is represented by a pair of topics that are provided by two users. Each path shows how topics of a user can be shared with others, and what are counterpart of her/his topics for others. There can be multiple paths each of which denotes independent shared topic. By viewing this network, a user can know how her/his topics are interested by others and how is relationship between her/him and others as a result.

In this paper, we firstly discuss requirements for community knowledge for information management and propose our method that is to find shared topics networks among people from bookmark data in Section 2. Then we describe our system called kMedia that is implementation of our method in Section 3 and show examples in Section 4. We show results of an experiment to evaluate how this system is suitable for community understanding. We compare our system with other systems in Section 6, and conclude the paper in Section 7.

2 Community Knowledge for Information Management

As we mentioned, relationship among people is one of the important resources for information management. But what relationship is needed for this purpose?

Relations among users should be described in some appropriate level, i.e., not too abstract and not too specific. Most of recommender systems use information objects themselves as shared information among users, e.g., netnews items [12], music[19], WWW pages[1][21], but they are too specific. This approach is desirable to know whether each object is good to share or no, but it is too intricate to understand what is relationship among people. On the other hand,

community support systems like Beehive[7], Babble[5], and Visual Who[4] use more simple relationship. For example, Beehive just calculates active members or not by observing email communication, and Babble shows relations among users in two-dimensional space. Visual Who also use two-dimensional space but show them more dynamically. Such visualization is intuitively easy to understand but it is too abstract to know what kind of relations can be found. In our approach, it is realized as relations between topics of different users. Topics are also intuitively easy to understand and furthermore informative enough to know types or aspects of relations.

Then difficulty lies on how we can identify users' topics. There are many proposals to track users' interest in information filtering field, e.g., WebWatcher[8] and Letizia[14] for WWW browsing, but they are not successful to identify users' interest as topics. There are two methods to capture users' interest on WWW. One is to use machine learning techniques like reinforcement learning and Bayesian network to detect it. The advantage and also disadvantage of this approach stem from assuming "persistence of interest". It is possible to track users' behavior but it is easily confused with big changes or branching of users' interest. Furthermore it is difficult to identify what they are interested because the learned data is just for programs not for users. The other method is to use classification techniques like hierarchical classification., e.g., Scatter/Gather[3] and Webmate[2]. It can show what users are interested more specifically, but its specification is ambiguous and needs efforts to understand because it often shows a (weighted) list of keywords.

We are skeptical about detecting users' interest by text analysis because of a more primitive reason. Most of systems above analyze texts by statistical information and some techniques to highlight importance like TF/IDF method [18]. There is a pit hole to apply these methods to detect users' interest, i.e., it is lack of background knowledge. Important words are often missing or appears very few times in text. For example, suppose that you are collecting pages on animals like pages for elephants, monkeys and so on. But there are probably a few occurrences of word "animal" because a sentence like "elephants are animals" is too common knowledge to describe in text. Applying statistical methods to those pages may produce some other words like "life" and "food" as important words instead.

We here abandon to detect users' interest by computation, but adopt users' own knowledge instead. In our case, it is the folder structure of WWW bookmarks. Names and contents of folders are results of efforts by users to explicate their intension, i.e., a folder name shows what kind of aspect the user are interested in, and contents of the folder show examples what s/he thinks within this category. Although it is restricted to a single hierarchical structure[3], it provides a basic knowledge of each user to use WWW.

Then the left problem is to find relationship among such topics. We regard a relation between topics as having some similarity relation between pages con-

[3] There are some proposals to extend more free structures like lattice structure[11] and bookmarks with comments[13].

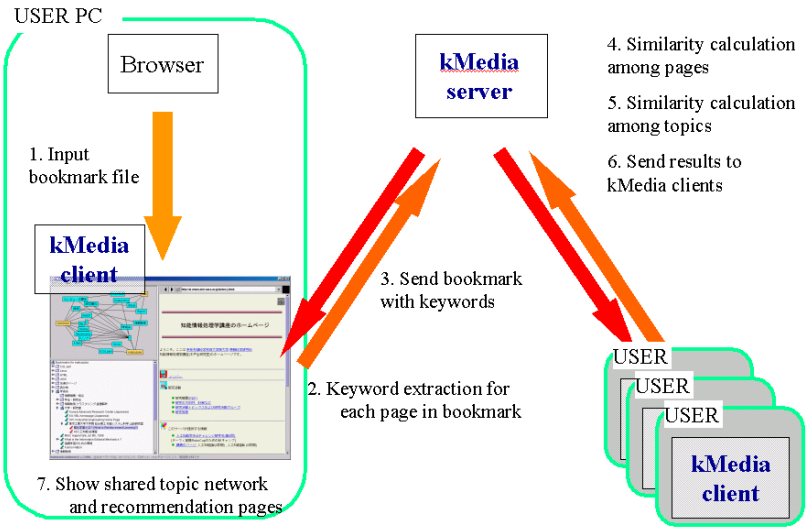

Fig. 1. System Architecture of kMedia

tained by both topics. At this process, we use traditional methods to calculate similarity among texts. The current implemented system just extracts some of most frequent words in each text and determines pairs of similar pages by checking how much such words are shared. As we mentioned above, we do not rely on text analysis so much, i.e., we do not expect high quality of similarity among texts. This similarity just tells that two pages are similar in comparison with other pages. But we expect that the amount of such similarity relations between two topics should suggest some relation between these topics. This expectation is proved by our experiment explained in Section 5.

3 System Overview

kMedia is a client-server system where each client system is provided for a user and a server system is provided for a community. A client works to process a user's bookmark files to extract keywords and show results to the user, and the server to calculate page similarity and determine topic relations (see Figure 1). The client system is implemented for Windows 95/98 with Sun Java2 and IBM XML parser[4], and the server system for FreeBSD/Linux with CGI and Perl 5.005.

There are two reasons that the client system performs keyword extraction instead of the server systems. One is to avoid heavy burden on the server system. The other is to make it to personalize the client system. We are planning

[4] We defined syntax by XML for communication between the client and server.

to have favorite lists of words or personalized ontology[20] to reflect personal activities of information management for keyword extraction, and use local files as information sources.

kMedia works as follows; first a user invokes a kMedia client system on her/his personal computer. The client system requires a location of her/his bookmark file if its execution of the client system is the first time for her/him. The client system reads her/his bookmark file and extract keywords for each URL in the bookmark file by fetching pages for these URLs and analyzing their texts. It extracts words from texts except stop words and some common words, then selects some of the most occurred words in them. It composes a bookmark data in which each URL is followed by keywords with occurrence numbers, and sends it to the server system.

The server system first calculates similarity between every pair of pages in the collected bookmark files. Similarity is measured by sums of occurrence of keywords that appeared in the both pages. Pairs of pages of which similarity measurement exceeds a threshold are marked as similarity pairs except pairs of pages from the same users. Then the system calculated similarity between folders. It counts numbers of marked pairs for every pair of folders, i.e., if there is a marked pair of pages and one page belongs to one of the folders and the other page to the other folder, then the pair of folders has one marked pairs of pages. Pairs of folders are marked "found" if the number of marked pairs of pages for them exceeds the threshold. The shared topics network is composed with these marked pairs of folders and folder structures of all users' bookmarks. Finally the server systems returns to client systems the shared topics network and similarity pages concerning to the bookmark of its owner.

4 An example

Figure 2 shows a snapshot of the client system. The top-left windows shows a shared topics network, the bottom-left window her/his bookmark with recommended pages, and the right window a WWW page specified by the user with the bottom-left window.

Figure 3 shows an example of shared topics networks obtained with the actual bookmarks. These bookmarks are brought by members of the artificial intelligence lab. There are three users that are represented as light gray boxes labeled A, B, and C[5]. Each user has several topics represented as dark gray boxes[6]. We can find nine inter-topic relations except root(/) folders, i.e., three relations between A and B: ("computer-related", "free soft"), ("research-related", "Study"), and ("search", "Sarch")[7], five relations between A and C: ("research-related", "Academia"), ("search", "information retrieval"), ("UNIX",

[5] We substituted user names by Alphabets like A, B, and C for privacy.

[6] In this example, we use only the first level of folder structures to simplify the network. And we translated topics' names in Japanese to those in English that are indicated with white boxes in Figure 3.

[7] "Sarch" is a typographical error by User B.

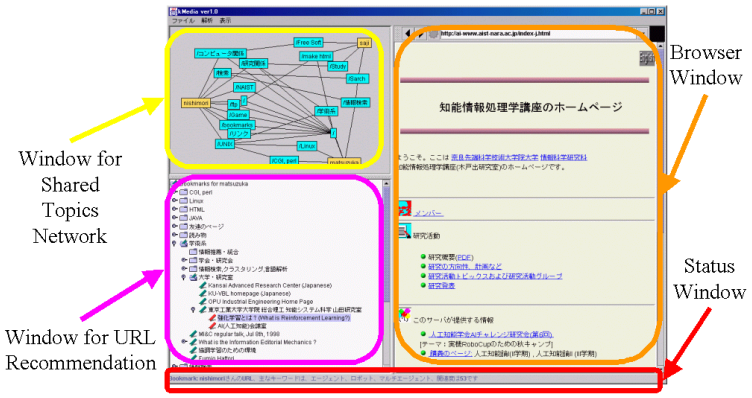

Fig. 2. User Interface of kMedia Client

"Academia"), ("UNIX", "CGI, perl"), and ("UNIX", "Linux"), and one relation between B and C: ("Sarch", "information retrieval").

Some pairs of topics are very common like ("UNIX", "LINUX") and ("UNIX", "CGI, perl"), but some pairs are not. For examples, The combination of ("research-related", "Academia") and ("UNIX", "Academia") is meaningful for a community for computer science research but not for everyone. In other words, users can understand that they are belonging to such community by viewing this relation.

Viewing this network as relationship among users, we can find that A and B have the same interest on computer science research, while A and C have the same interest on UNIX matters. Furthermore three users have a common interest on search. From this result, they may think that A has more knowledge on computer than others because both B and C are links to A with computer-related topics.

Figure 4 shows recommendation of URLs based on the shared topics network. Recommendation is done for marked pairs of pages. In this figure, two pages are recommended to a single page[8]. Pages for "what is reinforcement Learning" and "AI meeting room" are recommended to the originally bookmarked page "Yamada Lab., TITECH"[9].

5 An Experiment

We tested our system by a simple experiment to evaluate its performance. The main objective of the experiment is how our proposed method is useful to identify users' relations. We asked three persons to submit their bookmark files to

[8] bookmarked and recommended pages are represented with different colors, i.e., green and red respectively.

[9] `http://www.ymd.dis.titech.ac.jp/` where Ref. [16] is developed.

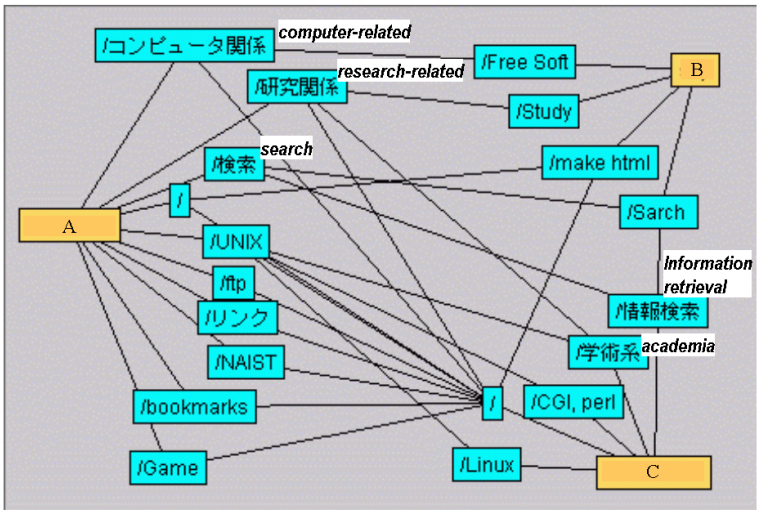

Fig. 3. Window for Shared Topics Network

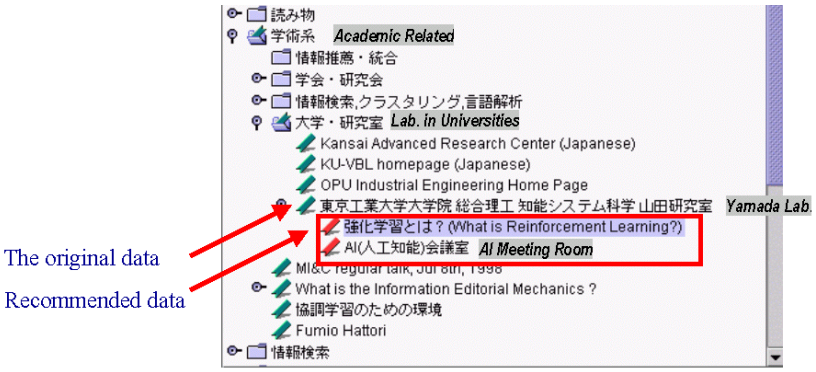

Fig. 4. Windows for Recommendation

the system, and asked subjective evaluation to each recommended page and each generated inter-topic relation by ranking 5 to 1 (5 is the best, and 1 is the worst). Criteria for evaluation is how suggested pages and inter-topic relations are acceptable according to their own bookmark. Table 1 shows data for submitted bookmark files and results generated by the system.

Evaluation for recommended pages is shown in Table 2(a) and evaluation for inter-topic relations is shown in Table 2(b). These tables show clearly different results. The highest scored rank for page recommendation is Rank 1 (the worst),

	User A	User B	User C
No. of Bookmarked Pages	376	278	297
No. of Analyzed Pages	263	185	240
No. of Topics	13	17	5
No. of Recommended Pages	345	513	454
No. of Shared Topics	10	10	3
Rate of Shared Topics/Total Topics	0.77	0.58	0.6

Table 1. Bookmark Data and Generated Results

	Good		Average		Bad
	Rank 5	**Rank 4**	**Rank 3**	**Rank 2**	**Rank 1**
User A	29	30	27	77	182
User B	94	86	73	185	75
User C	66	90	88	88	122
Total	189	206	186	350	379

(a) Subjective Evaluation for Recommended Pages

	Good		Average		Bad
	Rank 5	**Rank 4**	**Rank 3**	**Rank 2**	**Rank 1**
User A	3	3	2	0	2
User B	3	3	2	2	0
User C	2	0	1	0	0
Total	8	6	5	2	2

(b) Subjective Evaluation for Inter-Topic Relations

Table 2. Results of Evaluation

and the average is 2.6. The highest scored rank for inter-topic relations is Rank 5 (the best), and the average is 3.7. Discovery of related topics among users are evaluated as good enough in spite of bad results for recommendation of pages.

This result tells that our approach to find common topics among users is useful and practical. High average of evaluation for inter-topic relations means that it is useful to identify shared topics, and difference between page recommendation and inter-relation averages indicates that it can work even though text analysis methods are not sufficient.

6 Discussion

As we mentioned, the result of the experiment is interesting because two types of information generated from the same data shows different effects for users. The reason seems to lie on concept formation process of users. There can be three

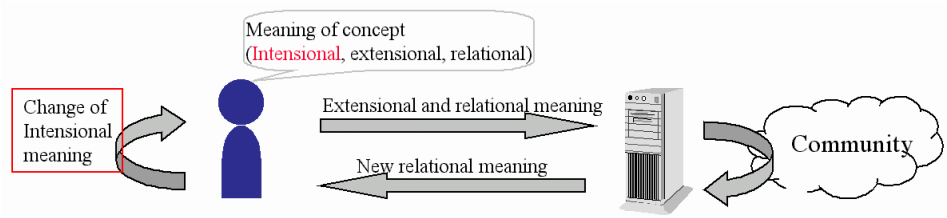

Fig. 5. Cycle of Concept Formation

types of meaning of concepts, i,e., *intensional* meaning that shows essential attributes, *extensional* meaning that shows the domain of objects, and *relational* meaning that describes relations to other concepts. Providing a folder and pages belonging to it is to define extensional meaning by its belonging pages and relational meaning by the folder structure whereas intensional meaning remains implicit. Page recommendation is then a question whether the recommended page can be within the extension of the concept or not. It is usually the severe question. On the other hand, proposal of inter-topic relations is a question whether addition of relational meaning of the concept is acceptable or not, which is more moderate question. Accept of the proposal in turn should cause modification of its intensional meaning that is kept in mind all the times. Considering that we need better understanding of each other in community, this change is preferable because they can integrate their knowledge more. We believe that supporting of such cycle of concept formation is the essential function of "intelligent" community support systems (see Figure 5).

7 Related Work

Kautz *et al.*[10] emphasized importance of people relations for WWW and have done primary work for finding people relations, i.e., their system called *Referral Web* can find people by analyzing bibliography database. Our aim is very similar to them, but we realized it differently. The benefit of our approach is to identify topics shared by users.

CommunityBoard[6][15] is another way to explicate users' relations by using topics. This system can show dynamics of interest on topics, i.e., who and when initiates and participates discussion for topics. But no discovery of topics are supported because topics themselves in this study are already shared by participants.

There are many bookmark-based WWW systems, e.g., bookmark-agent[16] and Webtagger[11]. Siteseer[17] also uses folder structures in bookmarks, but aims are different from us. It seems to aim large-scale social filtering, i.e., it uses folder information to decide recommended pages not either folders themselves

nor people themselves. LOUIS[22] is another recommender system based on not "folder" but "label" that is extension of "folder" idea, but it still remains URL recommender. There are the other types of usage of bookmarks, e.g., direct integration of bookmark files of multiple users by multi-tree[23] and integration by comments[13], but they do not care relations among people directly.

Grassroots[9] proposed to use "folder structure" as a basic structure to organize information and people. Grassroots approach is excellent but imposes a heavy charge to every user because it requires to unify a folder structure of various activities from information storing to information and even folders for other users. Our system can ease this problem by finding topic relations among users automatically.

8 Conclusion

In this paper, we discuss how relation among people should be explicated to facilitate information exchanging and proposed a system called kMedia that can show shared topics networks for this purpose. Our method to identify shared topics networks is simple and effective. We use folder structures as topics and identify inter-topic relations by analyzing texts associated to the folders. This combination of human knowledge and automatic discovery of relations works well. Even discovered relations between pages are not sufficient, discovered relations between topics can be acceptable. It seems that people can re-define or extend meaning of their own categorization represented as folders, on the other hand they cannot change meaning of pages. In this sense, our method is appropriate to find relation among people, because finding relations among people is not finding exactly shared information or interest but finding possibility of sharing of information or interest.

References

1. Marko Balabanovic and Yoav Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.
2. L. Chen and K. Sycara. Webmate: A personal agent for browsing and searching. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems, AGENTS '98*, pages 132 – 139, 1998.
3. Douglass R. Cutting, David R. Karger, Jan O. Pederson, and John W. Tukey. Scatter/Gather: A cluster-based approach to browwing large document collections. In *15th Annual International ACM/SIGIR Conference*, pages 318–329, 1992.
4. Judith S. Donath. Visual who: Animating the affinities and activities of an electronic community. In *ACM Multimedia 95*, 1995.
5. Thomas Erickson, David N. Smith, Wendy A. Kellogg, Mark Laff, John T. Richards, and Erin Bradner. Socially translucent systems: Social proxies, persistent conversation, and the design of "babble". In *CHI*, pages 72–79, 1999.
6. F. Hattori, T. Ohguro, M. Yokoo, S. Matsubara, and S. Yoshida. Socialware: Multiagent systems for supporting network communities. *Communications of ACM*, 42(3):55 – 61, 1999.

7. B. A. Huberman and M. Kaminsky. Beehive: A system for cooperative filtering and sharing of information, 1996. ftp://parcftp.xerox.com/pub/dynamics/beehive.html.
8. T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *IJCAI-97*, 1997.
9. Kenichi Kamiya, Martin R´oscheisen, and Terry Winograd. Grassroots: A system providing a uniform framework for communicating, structuring, sharing information, and organizing people. In *Proceedings of The 6th International World Wide Web Conference (WWW-6)*, 1997.
10. Henry Kautz, Bart Selman, and Mehul Shah. Referral web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.
11. Richard M. Keller, Shawn Wolfe, James R. Chen, Joshua L. Rabinowitz, and Nathalie Mathe. A bookmarking service for organizing and sharing urls. In *Proceedings of The 6th International World Wide Web Conference (WWW-6)*, 1997.
12. Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):76–87, 1997.
13. Wen-Syan Li, Quoc Vu, Divyakant Agrawal, Yoshinori Hara, and Hajime Takano. Powerbookmarks: A system for personalizable web information organization, sharing, and management. In *Proceedings of The 8th International World Wide Web Conference (WWW-8)*, 1999.
14. Henry Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of IJCAI-95*, pages 924–929, 1995.
15. S. Matsubara, T. Ohguro, and F. Hattori. Communityboard: Social meeting system able to visualize the structure of discussions. In *Proceedings of the 2nd International Conference on Knowledge-based Intelligent Electronic Systems (KES'98)*, pages 423–428, 1998.
16. M. Mori and S. Yamada. Bookmark-agent: Information sharing of urls. In *Poster Proceedings of The 8th International World Wide Web Conference (WWW-8)*, 1999.
17. James Rucker and Marcos J. Polanco. Siteseer: Personalized navigation for the web. *Communications of the ACM*, 40(3):73–75, 1997.
18. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.
19. Upendra Shardanand and Patti Maes. Social information filtering: Algorithms for automating "word of mouth". In *CHI*, pages 210–217, 1995.
20. Motoyuki Takaai, Hideaki Takeda, and Toyoaki Nishida. Knowledge sharing and organization by multiple ontologies. In *Proceedings First International Workshop on Strategic Knowledge and Concept Formation*, pages 73–84, 1997.
21. Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. PHOAKS: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.
22. Hideo Umeki and Takehiko Yokota. Louis - a labeling-based recommender system for web resources and communities of interest. In *Poster Proceedings of The 8th International World Wide Web Conference (WWW-8)*, 1999.
23. Kent Wittenburg, Duco Das, Will Hill, and Larry Stead. Group asynchronous browsing on the world wide web. In *Proceedings of The 4th International World Wide Web Conference (WWW-4)*, 1995.