Sedis: Agent-based Selective Distribution System Using Hybrid Information Filter

Yusuke Ariyoshi
Human Media Research Labs.
NEC Corp.

Abstract

This paper describes an agent-based selective distribution system (Sedis) that connects a content authors and readers. Sedis has an information-server selection function that selects appropriate digital libraries when an author wants to register his/her work. This system consists of agents at user sites and at digital library sites. When an author wants to register his/her content with a digital library, an agent selects digital libraries that have many readers who may be interested in the content. At the selected library sites, agents then select readers who are interested in the author's content. These agents learn the fields where each library is good at and the interests of readers in order to select appropriate libraries and appropriate readers. For learning and selection, Sedis uses information filtering technique. In researching Sedis, we developed a new filtering method called the "hybrid method," and we are using the method to implement the Sedis system. The hybrid method combines a content-based filtering (CBF) method and a social information filtering (SIF) method. The hybrid method changes the weights of CBF and SIF in the final recommendation based on their reliabilities. Through experiments, it was found that the precision of the hybrid method is higher than that of conventional filtering methods.

Key words: Digital Library, Agent, Information Filtering, Information Distribution System

1 Introduction

In conventional digital library research, the users of digital libraries are referred to as "readers". In the next generation of digital libraries, however, "authors" should be users, too. Accordingly, it is important to provide connections between authors and readers and places where authors and readers can meet and communicate. For instance, a digital library provides services that include registering new content on-line directly from an author to the library, sending lists of new content to library readers when they register, distributing book reviews or lists of favorite writers among library users, and announcing library events.

To achieve connections between authors and readers, technology is needed that connects authors with appropriate digital libraries. In fact, flood of information is a problem for content authors, too. Although authors want to know readers who are interested in their work, it is very difficult to find appropriate digital libraries in which they can register their content. Digital libraries also struggle over how to find good authors who can provide content that will interest their readers.

To solve the problems explained above, we have developed an agent-based selective distribution system called "Sedis." Sedis combines information filtering technology and agent technology to provide links that connect authors, digital libraries, and readers. In researching Sedis, we developed a new filtering method that is combination of CBF (Content-Based Filtering) and SIF (Social Information Filtering), and then used the

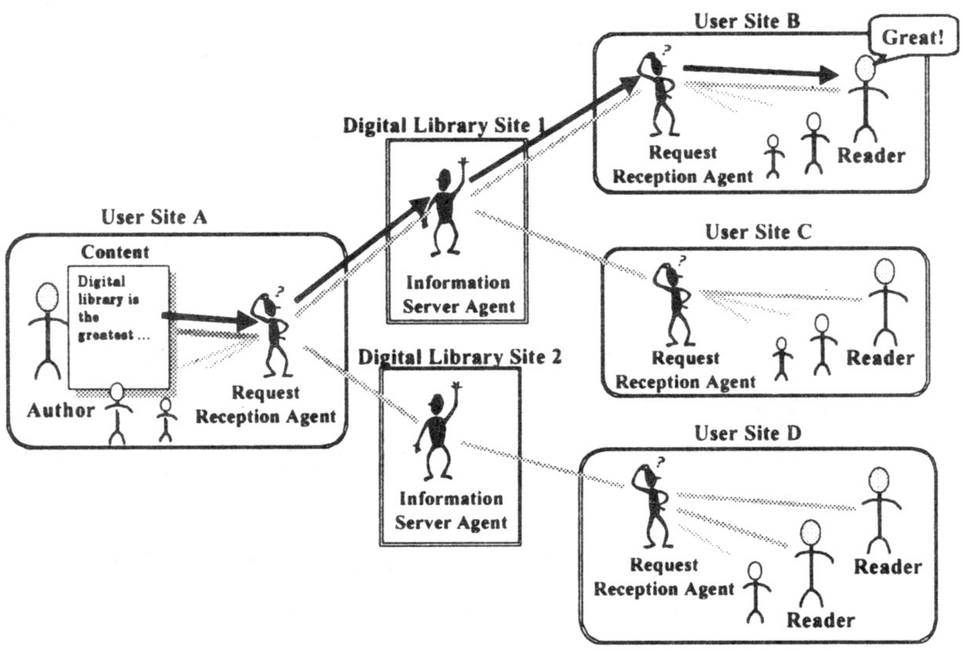

Fig. 1: Agent-based Selective Distribution System

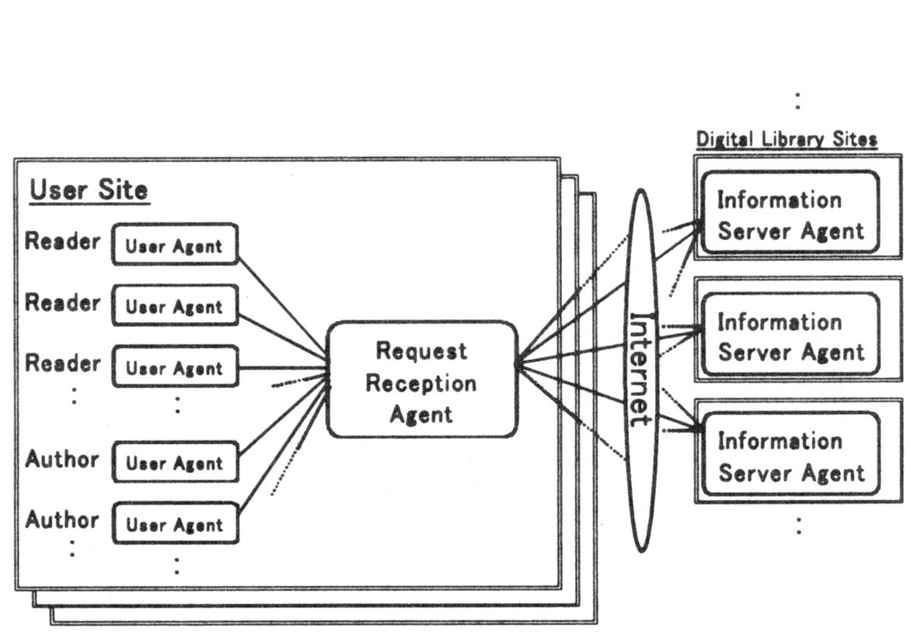

Fig. 2: System Configuration

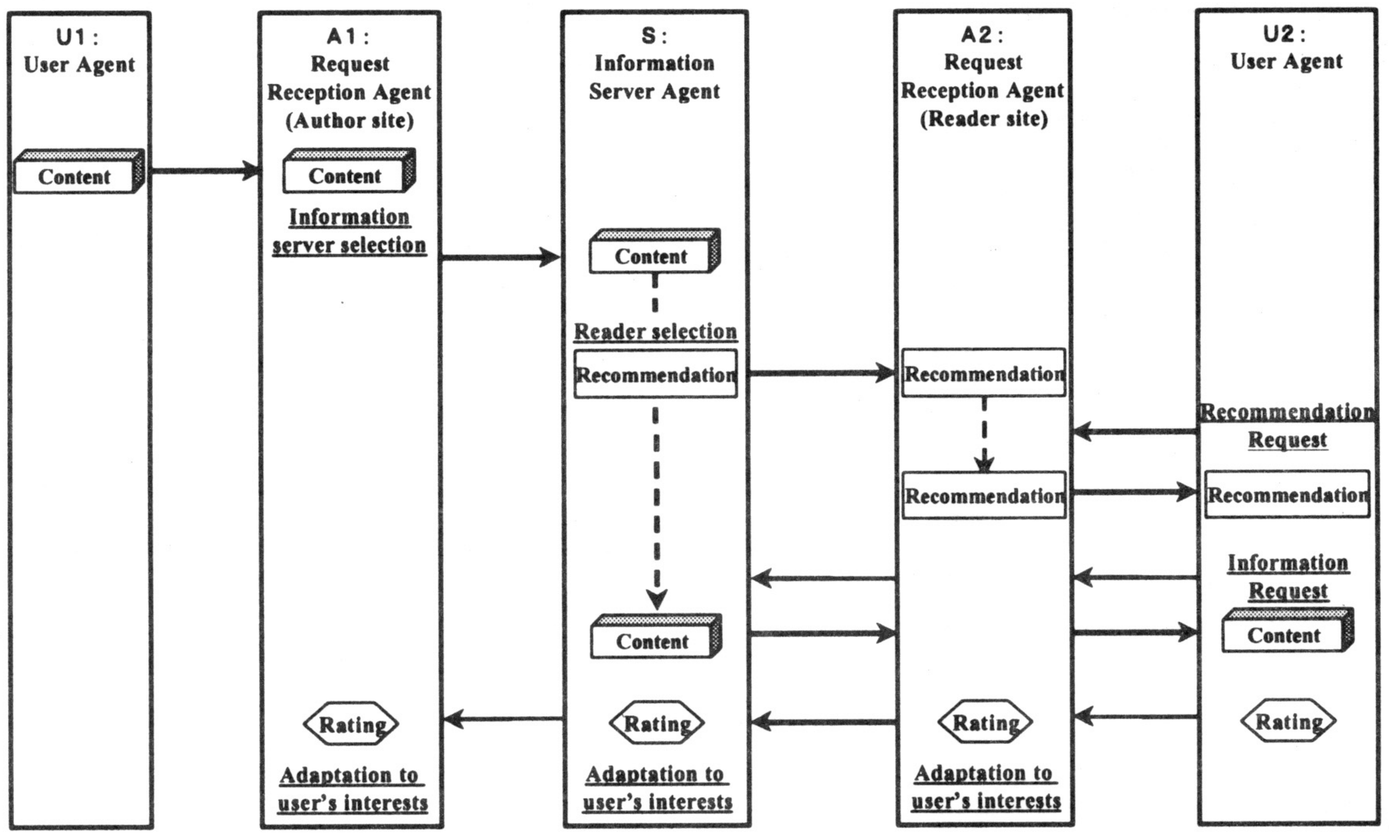

Fig. 3: Information Flow

method to implement the Sedis system.

The next section describes the configuration of Sedis. Section 3 presents a new filtering method, and section 4 describes related work. Section 5 discusses future work and provides some conclusions.

2 Development of the Selective Distribution Agent System

Sedis is a de-centralized distribution system that consists of agents at user sites and at digital library sites (Fig. 1). When an author wants to register his/her content with a digital library, an agent selects digital libraries that have many readers who may be interested in the author's content. At the selected digital library sites, agents then select readers who are interested in the author's content, and send the content to agents in the selected user sites.

In order to select the most appropriate libraries for the content, the user-site agents learn the fields where each library is good at. In order to select readers who may be interested in the content, the library-site agents study the interests of readers from readers' ratings and prior content selections.

2.1 System Configuration

The Sedis system consists of information server agents that select readers, request reception agents that select libraries and a control information flow inside each user site, and user agents that act as user interfaces (Fig. 2).

Figure 3 shows the information flow within the system. When a user agent receives some content from an author, the user agent sends it to a request reception agent. This request reception agent then selects libraries suited to the content and registers the content at those libraries. When an information server agent receives the authored content from the request reception agent, it recommends the new content to readers who may be interested.

When the request reception agent at a reader site receives the recommendation, the reception agent sends it to the user agent. Finally, the user agent notifies the reader of the recommendation.

When the reader selects this content from the recommendation list, the body of the content is presented. The reader may then evaluate the content on five levels. When a rating is entered, the evaluation is transferred along the information flow upward so that it is delivered to the request reception agent at the author's site.

2.2 Functions

The request reception agent learns fields where each library is good at in order to select the most appropriate libraries for authored content, and the information server agent learns readers' interests in order to select readers who may be interested in that content. These agents use the hybrid method for selection and learning.

2.2.1 Request reception agent

(1) Information server selection

When new content is registered to the request reception agent, the agent predicts each content score (for the readers) using the hybrid method. Readers at other sites who may be interested in the content are selected, and the content is then registered at the libraries where these readers are registered.

(2) Adaptation to user interests

When the request reception agent receives content ratings by the readers at the readers' and author's sites, the agent changes those readers' interest profiles using the hybrid method.

The agent raises the importance of profile terms that appear in content with high ratings, and lowers the importance of profile terms that appear in content with low ratings.

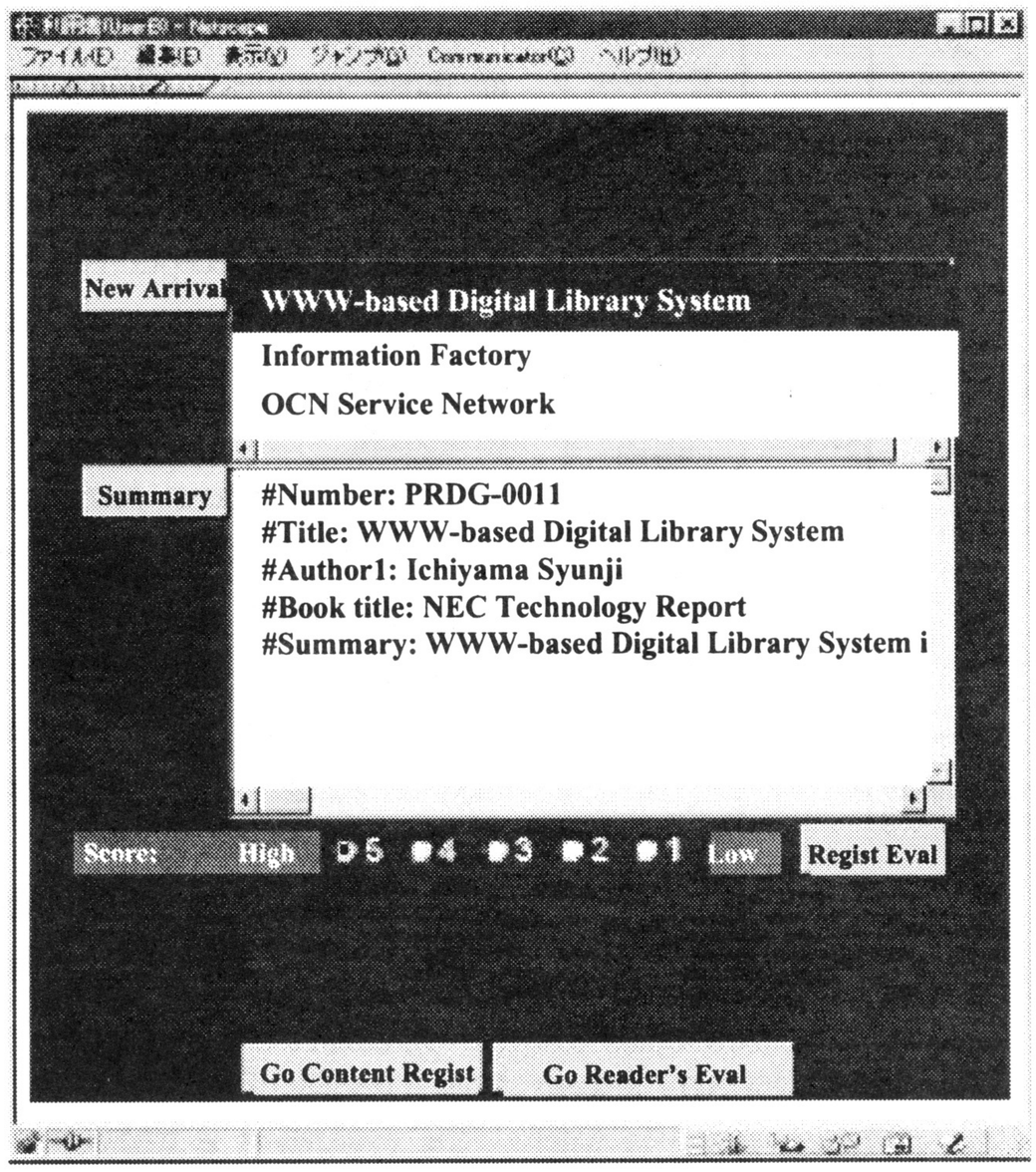

Fig. 4: User Client

2.2.2 Information server agent

(1) User selection

When new content is registered with a library, the information server agent uses the hybrid method to predict each reader's score for that content and only recommends the registered content to those readers who have high-predicted scores.

(2) Adaptation to user interests

When the information server receives reader ratings, the information server agent also changes the stored interest profiles using the hybrid method in the same way as the request reception agent, which was explained earlier.

2.2.3 User agent

The user agent is the user interface of Sedis. The user agent receives content from an author and sends it to a request reception agent. The user agent also displays a recommendation list, which was received from the request reception agent, to a reader and sends reader's ratings to the request reception agent.

The user agent consists of a user client that is a Java applet on a Web browser, and interacts with user. The main body of the agent relays messages between the user client and a request reception agent.

Figure 4 is the user client, which shows the recommendation list and body of the content. If reader selects the content title from the recommendation list (the upper text area), the body of the content appears in the lower text area. Reader can also enter ratings using the radio buttons provided.

3 Development of the Hybrid Method

Using information filtering technology, Sedis achieves 1) the selection of appropriate digital libraries to which authors can register content, 2) the recommendation of content that will interest readers, and 3) the learning of strong fields in each library and the learning of readers' interests for the above selection and recommendation.

There are currently two major types of information filtering methods. One is the CBF (Content-Based Filtering) method, which uses actual content features such as the number of times specific words appear. The other method is the SIF (Social Information Filtering) method (or so-called collaborative filtering), that uses other user ratings of the content. The CBF method can filter information that has not been evaluated previously. SIF, on the other hand, can filter information that contains figures and tables that are too complicated to be analyzed by CBF [1]. To obtain the advantages of both CBF and SIF, the hybrid method uses CBF to predict scores of unrated information and use SIF to predict scores of rated information.

3.1 Conventional Information Filtering Methods

The conventional filtering methods, CBF and SIF, are explained in the following section. This paper considers ratings based on the numeric scores given by users.

3.1.1 CBF

In a typical CBF [2], information content and user's interests are represented by vectors. A vector that represents the information content is called a document vector; it uses the term frequency (or appearance) as the term importance. A vector that represents the interests of user is called a profile.

CBF both predicts scores and learns interests. To predict scores, CBF matches document vectors with user profiles. To learn a user's interests, CBF increases the importance of terms in highly rated information and reduces the importance of terms in poorly rated information (Fig. 5).

3.1.2 SIF

People who have similar interests often recommend useful information to each other. Conceptually, SIF is an automation of the kind of information sharing that is often done by word of mouth. In a typical SIF [3], vectors of user ratings for recommended information represent the

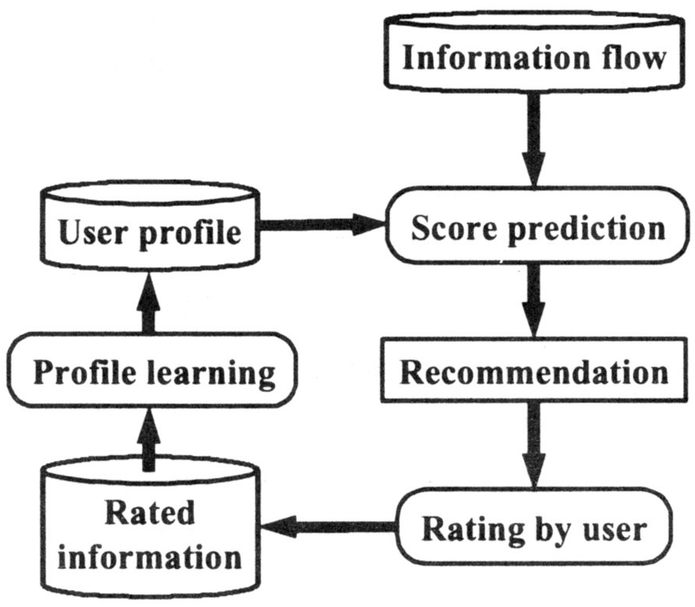

Fig. 5: CBF Configuration

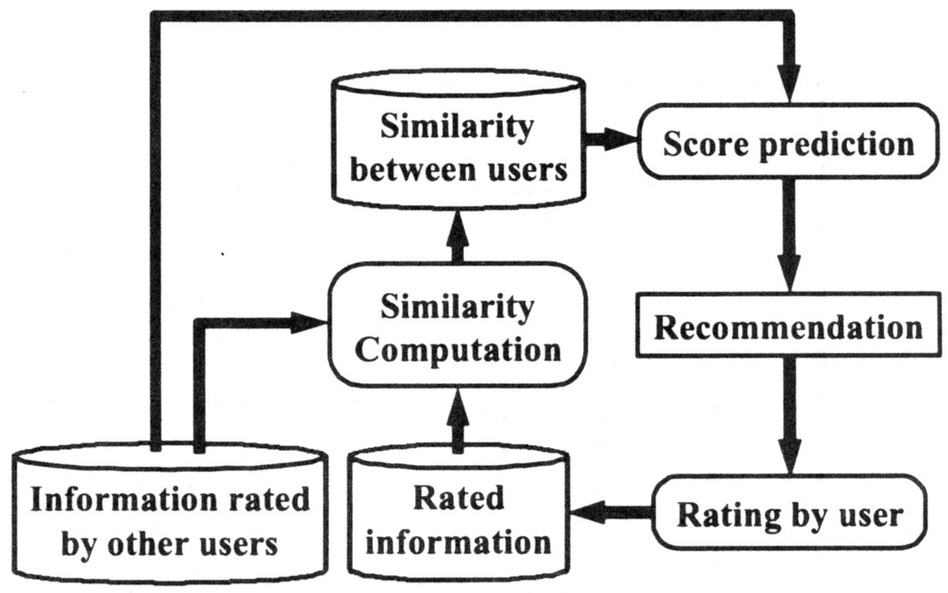

Fig. 6: SIF Configuration

interest of users. SIF consists of a user similarity computation and a score prediction (Fig. 6).

Therefore, the first step of SIF is to measure similarities between past ratings using a correlation coefficient. Then, predictions can be made by computing a weighted average of other user ratings. The similarities between the current user and other users are used as weights. The weighted average is then computed using the ratings of users who have similarities greater than a certain threshold [3].

3.1.3 Comparison

This section compares CBF and SIF, then discusses the problems of these methods. CBF filters information based on the information content (terms), and SIF filters information based on scores given by other people. Because of these differences, the two methods have the following characteristics:

I) Importance of figures and tables:

In CBF, information importance can only be determined by word terms. Therefore, it is difficult to recommend information that includes important information in figures or tables. Here, however, SIF is likely to produce quality recommendations since importance has already been assessed by other users.

II) New information:

CBF is able to filter information that has not been evaluated by other people. SIF, in contrast, can only recommend information that has been evaluated by other users.

III) Unlearned terms:

CBF learns a user's profile only through information that has been rated by the user himself/herself. Therefore, CBF accuracy declines if information contains many unlearned terms that have not previously appeared in rated information.

The ability to filter new information is an advantage of CBF. However, many important new terms (such as "Java" or "XML") appear on the Internet every day. The problem of not being able to understand new terms means that CBF cannot filter these new topics well. This can be disadvantageous for CBF when it is used in information recommendation services [4].

3.2 Hybrid method

CBF can filter information that has not previously been evaluated. SIF, on the other hand, can filter information that contains figures and tables that are too complicated to be analyzed by CBF. The hybrid method was developed to have both the advantages of CBF and SIF. There are two combination stages in this hybrid method. One is a prediction merge which merges predictions, and the other is a learning strengthening which strengthens learning (Fig. 7). The hybrid method improves the filtering performance by using reliability (described below in 3.2.1).

The prediction merge integrates the predictions of CBF and SIF. CBF is used for predicting scores of unrated information. For rated information, a more reliable prediction is selected as the final prediction from SIF's prediction and CBF's prediction. The learning strengthening stage uses SIF results for profile learning in CBF. CBF learns the importance of terms in unrated information using SIF's predicted ratings in addition to the user's real ratings. This strengthened learning uses SIF's predictions, which have a high reliability.

3.2.1 Predicted score reliability

The hybrid method introduces reliability to predicted scores to improve the filtering performance. The reliability of predictions is affected by the amount of data that is used in the predictions.

In CBF, a prediction score is calculated with a document vector and a user profile. In SIF, in contrast, a prediction score is calculated based on user similarities and the scores of other users. Therefore, reliability will increase in proportion to the amount of data used in the predictions. In the hybrid method, resultantly, the

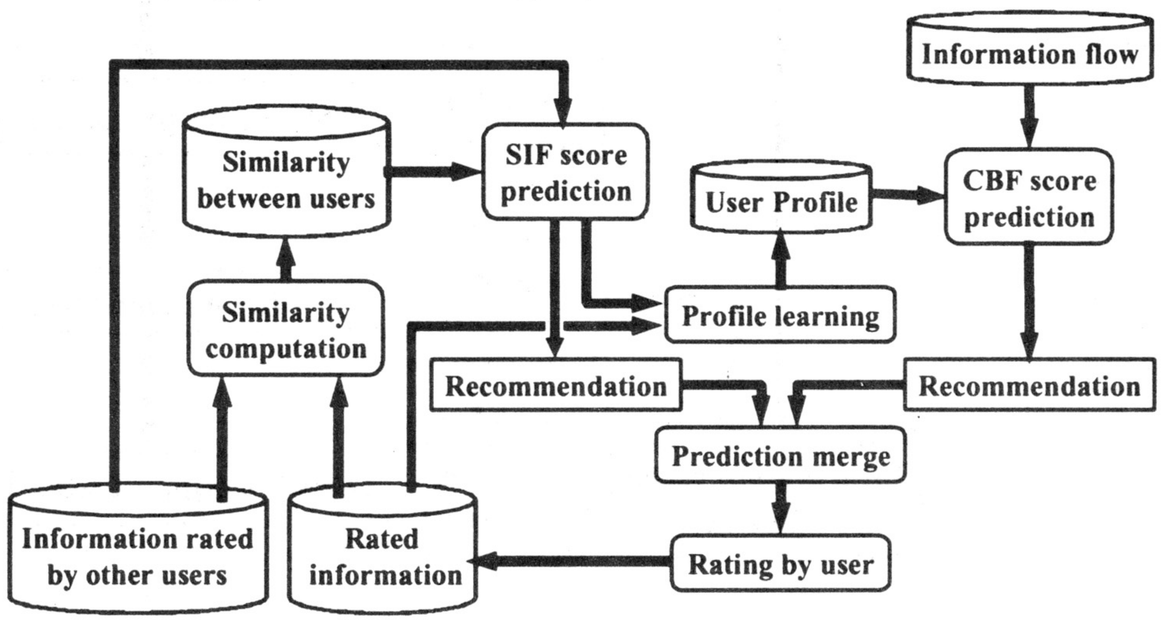

Fig. 7: Hybrid Method Configuration

reliability of predictions is estimated using the amount of data.

3.3 Experimental evaluation

Experimental Data: To obtain efficient data, we ran a patent-clipping service in our laboratory. In this service, user ratings were recorded as integer numbers from 1 to 5. Larger scores indicated stronger interests by users.

In these experiments, there were 45 users, all of whom used the service often. These users provided 10,709 ratings for 3,000 different patents. Table 1 is a histogram of the ratings. In these tests, document vectors were generated to extract nouns (other than numerals) from the titles and the summaries of the individual patents. Table 2 shows the number of terms in the document vectors.

Experimental Methodology: The author uses 10-fold cross-validation in order to compare the individual methods and decide on the value of each parameter. In the 10-fold cross-validation, predictions can be generated for each real recommendation score. Then, the prediction performance is measured by comparing the predicted scores to real scores.

Error Estimation: In the experiments, the reliability is measured by the error value, which is defined by the square of the difference between the predicted rating and the user's true rating. A preliminary experiment was performed to define the error estimate expressions. Errors were estimated with polynomial expressions that were based on the quantity of data used in the predictions.

For CBF, the number of terms in a profile was used for measuring the length of that profile. In addition, the quantity of terms that appeared in both the document vector and the profile was used to measure the length of the document vector.

For SIF, the sum of the similarities between the current user and similar users was used as the number of similar users. The error of the predicted score was estimated using a quadratic polynomial expression, which included the amount of data used in the prediction and the reciprocal of the data amount.

3.3.1 Summary of Results

Figures 8 and 9 are graphs of results from the experiments. The horizontal axis denotes the cut-off rank, i.e., the rank sorted by predicted ratings. The vertical axis represents the precision, i.e., the rate of 4 and 5 scores at the cut–off rank. This represents the quality of a service that a user observes directly.

Learning Strengthening: Figure 8 compares CBF with three learning strengthening methods. One method is CBF without learning strengthening and the other methods are **simple strengthening** (The method chooses higher predicted scores N/2 and lower predicted scores N/2.) and **strengthening with reliability** (First, the method chooses higher predicted scores N and lower predicted scores N. Then, based on the 2N of the predicted scores chosen, the method chooses N.); N is the number of scores rated by the current target user. The results show that strengthening with reliability gives the best precision.

Prediction Merge: Figure 9 compares three prediction merge methods after learning strengthening (strengthening with reliability). **Simple merge** does not consider estimated errors. It is done prior to SIF. The approach only chooses a CBF prediction if SIF cannot predict a score. **Switch merge** chooses a prediction that has a low estimated error. **Blend merge** blends the CBF prediction and the SIF prediction using errors as weights. The final prediction is calculated with the following expressions: (predicted score of SIF * estimated error of CBF + predicted score of CBF * estimated error of SIF) / (estimated error of CBF + estimated error of SIF). The results show that switch merge gives the best precision.

In our experiments, the number of users was less than that needed to ensure the good prediction reported in [3].

Table 1: Histogram of ratings

Sum	Ratings				
	1	2	3	4	5
10,709	5,165	1,997	1,547	1,270	730

Table 2: Number of terms

Documents	Terms		
	Mean	Max	Min
3,000	32.58	66	10

However, the results of our experiments show that the hybrid method has an excellent performance, even though the filtering performance of SIF alone was not ideal.

4 Related Works

This research focuses on the technology that provides connections between authors and readers through digital libraries. The technology that connects readers with digital libraries has been researched. For example, a cross-domain search system has been developed. It helps user to search many digital libraries to find content that fits their interests[5]. Recommendation systems[1] that have been developed also help digital libraries to offer content that fits users interests. However, in these previous research works, the users of digital libraries have not included "authors". Sedis combines information filtering and agent technology to provide links that connect authors, digital libraries and readers. In the Sedis, authors can find appropriate digital libraries in which they can register their content, and digital libraries can get contents that will interest their readers.

Concerning other filtering methods, Fab of Stanford University [6] also uses a combination of CBF and SIF. Fab is a Web page recommendation system and performs collection and selection in two stages. In the collection stage, each collection agent collects pages on the topics it covers. The central router and selection agent control the selection stage. The central router matches pages collected by the collection agent with user profiles and sends them to the appropriate user selection agents. The selection agent presents pages that have not yet been read by a user. When a user gives a rating of one of seven (a grade), the selection agent uses that rating to update the user's profile, and in turn, the collection agent updates the topic profile. These agents use CBF to select information. In addition, pages with high ratings are recommended to similar users. In other words, Fab uses a simple combination of CBF and SIF. This method can be called prediction merge without reliability.

The hybrid method used by Sedis uses both the results from SIF, for learning the importance of words in a CBF user profile, and the merging recommendation results from CBF and SIF. This learning strengthening enables the method to do highly accurate recommendations. That is, with learning strengthening, the accuracy is improved in predicting the importance of words, and this means that highly accurate recommendations can be achieved. This is because word importance can be calculated if words appear in content, which means that learning strengthening can be applied even to unknown words that do not appear in rated content.

5 Conclusions

This paper describes an agent-based selective distribution agent system (Sedis). Sedis provides links between a content author and readers, which will be an important service in the next generation of digital libraries. The system combines information filtering technology and agent technology to provide links that connect content authors, digital libraries, and readers. In researching Sedis, we first developed a hybrid information filtering method, and then used that method to develop the Sedis system. This hybrid method uses both content features and rated data.

The Sedis system consists of information server agents that are located at digital library sites, request reception agents that are located at user sites, and user agents that act as user interfaces. The request reception agent learns fields

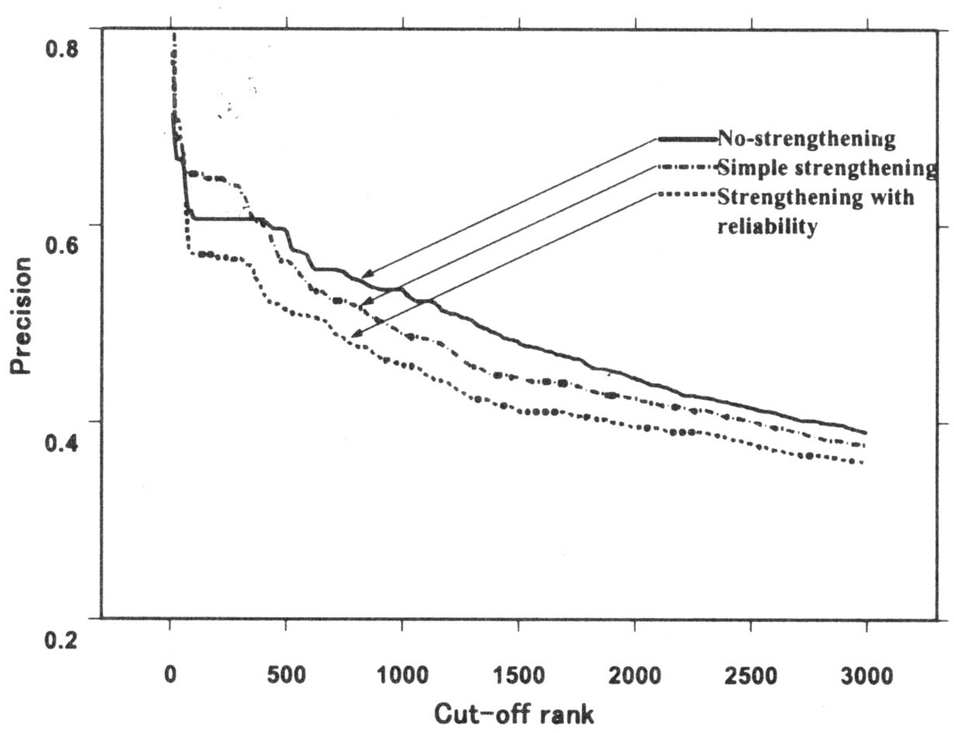

Fig. 8: Learning strengthening

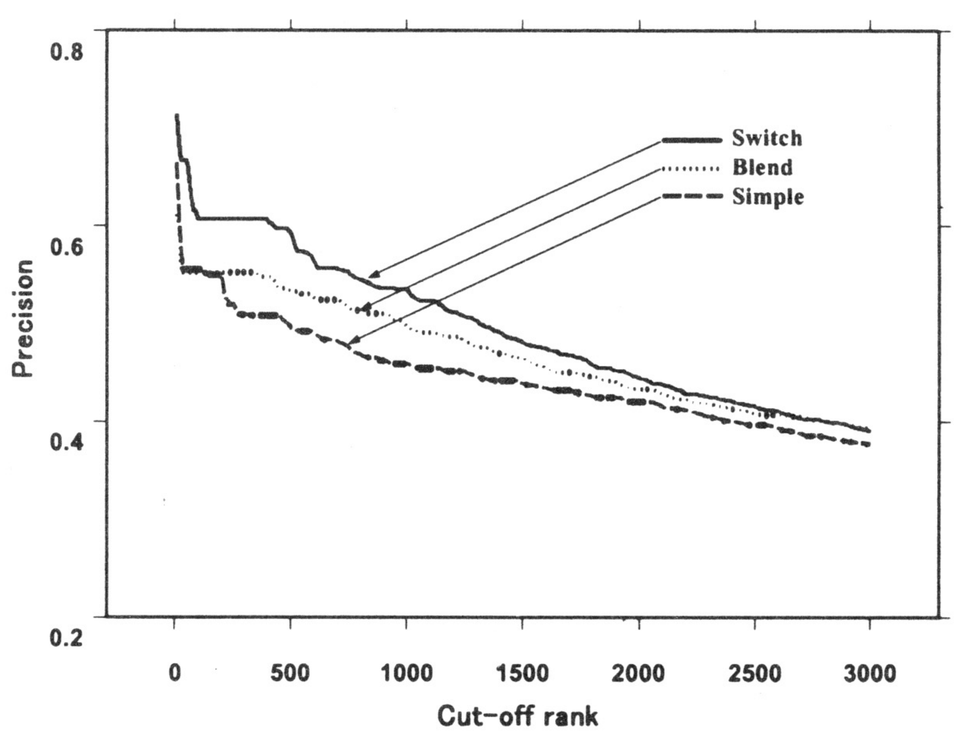

Fig. 9: Prediction merge

where each library is good at in order to select the most appropriate libraries for registering content, and the information server agent learns readers' interests in order to select readers who may be interested in the content. These agents use the hybrid method for selection and learning.

The hybrid method is a combination of CBF and SIF. The hybrid method estimates the reliability of CBF's prediction and SIF's prediction based on the amount of data that is used in these prediction processes. Then, a more reliable prediction is selected as the final prediction. Through experiments, it was found that the precision of the hybrid method is higher than that of conventional filtering methods.

Acknowledgments

A part of this research and development was funded by the Next-Generation Digital Library Project under the auspices of the Japan Information Processing Development Center (JIPDEC).

References

[1] "Special Section: Recommender Systems," CACM, Mar 97, Vol. 40, No. 3, pp. 56-89, 1997.

[2] Atsuyoshi Nakamura. et al., "Learning personal preferences by on-line prediction algorithms," *IJCAI-97 poster session abstracts*, p. 23, 1997.

[3] Upendra Shardanand, Pattice Maes, "Social Information Filtering: Algorithms for Automating "Word of Mouth," Proc. of CHI'95, pp. 210-217, 1995.

[4] Yusuke Ariyoshi, Shunji Ichiyama, "An information filtering method combining content-based filtering and social-information filtering," *Proc. of IEICE 8th Workshop on Data Engineering*, pp. 49-54, 1997.

[5] Hidekazu Yanagimoto, et al., "Development of a Cross-Domain Search System With Concurrent Bidding Agents," Digital Libraries, No. 13, pp. 53-65, 1998.

[6] Marko Balabanovic, Yoav Shoham, "Fab: Content-Based, Collaborative Recommendation," CACM, Vol.40 No.3, pp. 66-72, Mar 1997.