

Adaptive Personal Information Filtering System that Organizes Personal Profiles Automatically

Toshiki Kindo Hideyuki Yoshida Tetsuro Morimoto Taisuke Watanabe

Matsushita Research Institute, Tokyo Inc.

3-10-1 Higashimita Tama-ku Kawasaki, 214

JAPAN

Abstract

An adaptive personal information filtering system is presented. The adaptive learning algorithm of the system is derived from the statistical consideration - what keywords are useful for retrieving items in which the user want to see. According to the algorithm, the system organizes a personal profile to reflect the user's interests from an empty profile through user-system interaction - the system displays items and the user returns the answer, "interesting" or "uninteresting". The system gives the user items according to his or her interest estimated using the user's personal profile. The density of the interesting items at the head of the sequence of items sorted by the system is as high as that in the items retrieved using the keyword matching method whose Boolean expression of keywords was made by a researcher who has much experience in patent retrieval. The number of the interesting items included the head of the sorted sequence is larger than that in the retrieved items. A field test is executed to ensure that the system responds to the varied interests of many users. The items used in the test are Japanese newspaper headlines which are supplied at a rate of 600 per day from information providers. Not only researchers and engineers in various technical fields, but also staff members participated in the test as users. The results of the test suggest that the system can be put to practical use.

1 Introduction

With the increasing availability of information, it becomes more important to provide methods of filtering information to reflect the user's needs. There are many difficulties in determining what information the user wants to see. The difficulties are based on the following facts: the words people use do not always clearly reflect the topic or content. A single word has more than one meaning in many documents[Foltz and Dumais, 1992]. Furthermore, not only the user's interest but also the value

of the words used to retrieve items change, i.e. "internet" was a useful key word a few years ago. However, it is no longer useful today.

Many studies have been performed in an attempt to overcome the ambiguity of words. Latent Semantic Indexing is a methods that takes account of the ambiguity for document profiles[Foltz, 1990]. Information Tapestry is a method used to overcome it by means of collaborative filtering[Goldberg *et al.*, 1992].

Furthermore, adaptive learning is discussed as an approach to overcome changes in the user's interest[Asakura *et al.*, 1995]. Unfortunately, the discussed adaptive learning algorithm has many heuristic parameters and its effects are not shown.

In this paper, we present an adaptive personal information filtering system with which users do not need to directly manipulate their profile. The system gives them items sorted in order of interest (priority). We call the present system "information filtering system sorting items in order of user's priority" or simply "INSOP". INSOP has only 2 parameters: one is a parameter to limit the size of user profile and the other is a parameter to control the method of manipulation of rare keywords. In Section 2, we derive an adaptive learning algorithm based on the statistical consideration - what key words are useful for retrieving the items in which only one user is interested. In Section 3, we demonstrate the effects of INSOP. We show that INSOP is a good information filtering system. In Section 4, we show the architecture and browser of the current INSOP system used in our field test.

2 Adaptive Learning Algorithm

In this section, we show the adaptive learning algorithm of INSOP. First, we consider which key words are useful for retrieving items in which the user is interested. Next, we give an adaptive learning algorithm for an ideal case and refine the algorithm for application to actual cases.

2.1 Classification of Key Words

If our subject is to supply a set of key words which are useful for retrieving items in which only one user is interested, then the intrinsic value of each key word depends on whether the key word reflects his or her interest or

not. The semantics of each key word is not important in this context. Using this viewpoint, we classify key words (including Boolean combinations of key words) into 3 groups. The first group includes those whereby if one of them appears in an item, then the user has an interest in the item. We call such key words “positive key words”. The second group is that whereby if one of the key words appears in an item, then the user does not have an interest in the item. We call such key words “negative key words”. The third group includes those which are not useful in determining whether that the user has an interest in the items with the key word or not. We call such key words “neutral key words”.

2.2 Ideal Case

In order to explain the principal idea of INSOP clearly, we consider an ideal case in which we can manipulate items, keywords, and user's inputs by statistical methods.

Here we define notations as follows: w_j is the j -th key word, p^* is the true probability that an item is one in which the user has an interest, and $q^*(w_j)$ is the true probability that an item with the j -th key word is one in which the user has an interest.

The positive (negative) key words satisfy $q^*(w_j) > p^*$ ($q^*(w_j) < p^*$), and the neutral key words satisfy $q^*(w_j) \sim p^*$.

In an ideal case, we can define a statistical quantity to classify key words into the 3 groups as:

$$SKC(q^*(w_j), p^*) = q^*(w_j) \log \frac{q^*(w_j)}{p^*} - (1 - q^*(w_j)) \log \frac{(1 - q^*(w_j))}{(1 - p^*)}. \quad (1)$$

$SKC(\cdot)$ is the “Statistical Key word Classifier”. When the Statistical Key word Classifier receives one of positive, or negative, or neutral key words as an input, it outputs a positive, or negative value, or the value close to 0, respectively. The Statistical Key word Classifier is similar to Kullback-divergence, which is used to measure the “distance” between 2 distributions in statistics,

$$KD(q^*(w_j), p^*) = q^*(w_j) \log \frac{q^*(w_j)}{p^*} + (1 - q^*(w_j)) \log \frac{(1 - q^*(w_j))}{(1 - p^*)} \quad (2)$$

but the sign of its second term is different from that of the Kullback-divergence.

In addition, we define the “Statistical Item Classifier” as

$$SIC(item_a) = \sum_{j=1}^{n_a} SKC(q^*(w_j^a), p^*) \quad (3)$$

where $item_a$ is the a -th item, n_a is the number of the key words of the a -th item, and w_j^a is the j -th key word of them. When the a -th item has many positive (negative) key words, the output of the Statistical Item Classifier is a large positive (negative) value.

The Statistical Key word Classifier and the Statistical Item Classifier work well as tools for classifying items in which a user is interested or not, in the ideal case.

As a result, a system which sorts items in order of the output of the Statistical Item Classifier works as an information filtering system. This is the principal idea of INSOP.

2.3 Actual Case

If we apply the Statistical Key word Classifier and the Statistical Item Classifier to an actual information filtering system, we must estimate the probability that an item with the j -th key word is of interest based on sample items. The estimated probability, $q(w_j)$, is not equal to the true probability, $q^*(w_j)$, in actual cases. The accuracy of the estimated probability depends on the number of sample data. Accordingly, the following problems must be solved: (1) how to treat key words which are too rare to be manipulated by statistical methods, (2) how to initialize the system, (3) how to adapt user's interests and their modifications, (4) how to treat items which have both positive and negative key words. In this section, we provide the solutions to these problems by extending the Statistical Key word Classifier to the “Key word Classifier” for the treatment of both frequent and rare key words and by introducing an adaptive learning algorithm.

First, we extend the Statistical Key word Classifier in order to solve the first problem. The rare key word problem is that the probability estimated on the basis of a small amount of sample data is different from the true value. We consider a typical case where the r -th key word, which is a rare key word whose true probability, $q^*(w_r)$, satisfies $0 \ll q^*(w_r) \ll 1$, appears just once with an interest item. Note that a case in which a new key word, either is rare or not, appears is equivalent to this case. In the real world, there are many such cases. Then, the probability $q(w_r)$ is estimated as

$$q(w_r) = 1 \gg q^*(w_r). \quad (4)$$

As a result, the Statistical Key word Classifier input with the rare (new) r -th key word outputs a large but unreliable value. Here, we remember that the Statistical Key word Classifier input with one of the neutral key words, which are useless to estimate user's interest, outputs the value close to 0. Therefore, if we interpret “unreliable” as “useless”, we are able to introduce the following “Key word Classifier” to treat frequent and rare key words,

$$KC(w_j) = \tanh\left(\frac{c(w_j)}{\alpha}\right) \cdot SKC(q(w_j), p), \quad (5)$$

$$p = \frac{P^i}{P^i + P^u}, \quad (6)$$

$$c(w_j) = Q_j^i + Q_j^u, \quad (7)$$

$$q(w_j) = \frac{Q_j^i}{Q_j^i + Q_j^u}, \quad (8)$$

where P^i is the appearance count of interesting items, P^u is the appearance count of uninteresting items, Q_j^i is

the appearance count of interesting items with the j -th key word, Q_j^u is the appearance count of uninteresting items with the j -th key word, and α is the parameter, called the “rarity parameter”, which controls the manipulation of rare key words. Accordingly, we define “Item Classifier” as:

$$IC(item_a) = \sum_{j=1}^{n_a} KC(w_j^a). \quad (9)$$

The Key word Classifier outputs a value close to 0 when it receives a key word whose appearance count is less than the rarity parameter α . Namely, the Key word Classifier classifies the key word as a useless key word. Consequently, the unfavorable influence of the rare key words to the output of the Item Classifier is small compared with that to the Statistical Key word Classifier. As a result, the Key word Classifier and the Item Classifier work well as tools for classifying the interesting, neutral, or uninteresting items in actual cases.

This is one solution to the first problem of rare key words.

The definitions of the Key word Classifier and the Item Classifier require that the personal profile of INSOP must be in the form given in Table 1.

Table 1: Format of INSOP personal profile

user name	P^i	P^u
w_1	Q_1^i	Q_1^u
w_2	Q_2^i	Q_2^u
$\vdots$	$\vdots$	$\vdots$
w_K	Q_K^i	Q_K^u

Accordingly, the adaptive learning algorithm of the personal profile is given as follows.

- 1) A user reads an item, $item_a$, with key words, $w_1^a, \dots, w_{n_a}^a$.
- 2) If the user inputs “interesting” to INSOP, then
 - 2-1) $P^i \leftarrow P^i + 1$ and
 - 2-2) search user’s personal profile for the same key word as $w_k^a, k = 1, \dots, n_a$.
 - 2-3) If there is a key word which satisfies $w_j = w_k^a$, then $Q_j^i \leftarrow Q_j^i + 1$.
 - 2-4) Otherwise, add w_k^a into user’s personal profile as the $(K + 1)$ -th key word and set $Q_{K+1}^i = 1, Q_{K+1}^u = 0$.
- 3) If the user inputs “uninteresting” to INSOP, then
 - 3-1) $P^u \leftarrow P^u + 1$ and
 - 3-2) search user’s personal profile for the same key word as $w_k^a, k = 1, \dots, n_a$.
 - 3-3) If there is the key word which satisfies $w_j = w_k^a$, then $Q_j^u \leftarrow Q_j^u + 1$.
 - 3-4) Otherwise, add w_k^a into user’s personal profile as the $(K + 1)$ -th key word and set $Q_{K+1}^i = 0, Q_{K+1}^u = 1$.

Note that we do not need to embed the user’s interest into the personal profile initially. INSOP can be started an empty profile. This is an advantage of INSOP and is also the solution to the second problem (how to initialize the system). There is, however, a disadvantage of the algorithm as follows: the algorithm continually increases the number of key words in the user’s personal profile.

We overcome this disadvantage by adding an algorithm which selects the key words which give the “best” finite personal profile of INSOP. In INSOP, positive and negative key words are useful but neutral key words are useless. We introduce a function which outputs a large positive value for positive or negative key words but which outputs a value close to 0 for neutral key words in order to construct INSOP which selects a finite number of useful key words. Such a function is given by

$$KCH(w_j) = \tanh\left(\frac{c(w_j)}{\alpha}\right) \cdot KD(q(w_j), p). \quad (10)$$

We call $KCH(\cdot)$ the “Key word Checker”. When we design the INSOP system, we must give the limit of the number of key words in the personal profile, N . In order to make the “best” personal profile of INSOP under the limit, we select N key words which give large outputs of the Key word Checker compared to other key words.

Next, we show that the rarity parameter α is related to the adaptive speed of INSOP. The properties of the Key word Checker are summarized in Table 2.

Table 2: Relations between input key word and the Key word Checker output

input	output
positive	large
negative	large
rare/new	small $\sim$ large
neutral	zero

The outputs of the Key word Checker following input of rare or new key words depend on the rarity parameter. If we give the rarity parameter a value close to 0, then the outputs become very large and INSOP manipulates rare and/or new key words as useful key words compared to positive and negative key words. As a result, INSOP retains new or rare key words whose appearance counts are small and releases old key words - positive or negative key words - whose appearance counts are large. As the rarity parameter increases, rare and new key words become useless and are easily released but positive and negative key words are held fast in INSOP. This shows that we can control the adaptive speed of INSOP according to our requirements using the rarity parameter. This is the solution to the third problem (how to adapt the user’s interests and their modifications) and also that of the first problem (how to treat rare key words).

The remaining problem is how to treat items which have both positive and negative key words. We can determine the characteristic of such items but we have no direct solution for this problem yet. The characteristic is that the output of the Item Classifier is close to 0 but

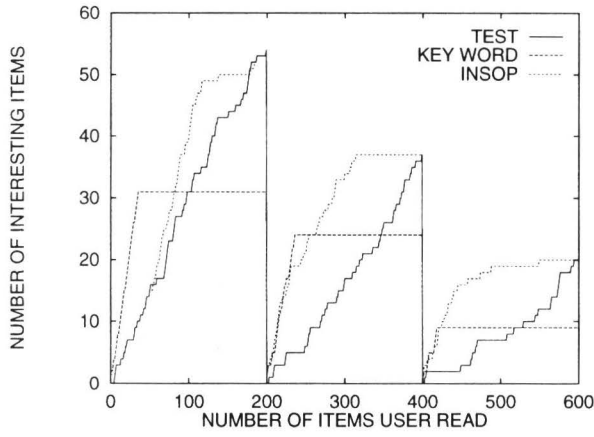


Figure 1: The distribution of items in which the testee has interests

the sum of the outputs of the Key word Checker is large when we input such an item. Accordingly, we designed the current INSOP system which displays the output of the Item Classifier as “needs” and the sum of the Key word Checker outputs as “reliance” in order to help users to identify such items.

3 Performance of INSOP

In this section, we demonstrate the performance of INSOP based on experiments conducted in collaboration with several researchers who cooperated as testees. All of them are highly experienced in patent retrieval. We set the limit of the number of key words in the personal profile to $N = 750$ in this experiment.

We obtained test sequences of items which were headlines of Japanese newspapers provided by information providers and the Boolean combination of key words for retrieving to the items in which each testee had interest as follows.

- 1) The tester requested each testee to mark the items in which they had interests and executed the interface program which displays one item without key words and receives the testee’s input - “interesting” or “uninteresting”.
- 2) The testee read 600 items and input his answers.
- 3) After the testee had input all the answers, the tester stopped the program and requested the testee to produce the Boolean combination of key words to retrieve the items which he marked “interesting”.

In order to simulate a case where each user received three sets of items, we divided each test sequence into 3 short test sequences including 200 items.

In Figure 1, we show typical results of the experiments to evaluate the performance of INSOP. The horizontal axis indicates the number of items which the testees read. We refer to the number as m . The vertical axis indicates the number of interesting items including the read items. Three sequences of items are shown in Figure 1. The

first sequence corresponding to a solid line is the test sequence. The second sequence corresponding to a dashed line is that of items retrieved by the key word matching method using the Boolean combination of key words set by the testee. The third sequence corresponding to a dotted line is that of items sorted by INSOP.

The solid line shows that the distribution of the marked items in the short test sequence is approximately uniform.

The dashed line stops to increase at $m = 36, 238$ and 421. This shows that the key word matching method retrieves 10%~ 20% items. About 60% marked items are included in the retrieved items. Consequently, the density of the marked items in the items retrieved using the key word matching method is about 3 times as high as that in the short test sequences. This rate is consistent with the value of 1.5 ~ 2.5 reported by ref.[Foltz and Dumais, 1992]. Note that the key word matching method misses several of the marked items (about 40%). Namely, it is difficult to create a Boolean combination of key words that includes all our personal interests, which are indefinite, even if we have much experience in information retrieval. This result suggests that the key word matching method is not useful for retrieving the marked items in this test task.

However, based on this result, we are not suggesting that the key word matching method is useless for all types of information retrieval or filtering. In general, retrieval methods, into which the key word matching method is classified, in which personal profiles are directly manipulated by users, are still useful in cases where users understand the exact subjects to retrieve, i.e. patent retrieval.

The dotted line shows the performance of INSOP. In the experiment, adaptive learning of the personal profile is executed at $m = 50, 100, 250, 300, 450$ and 500. The order of the first 50 items presented by an INSOP system whose personal profile is empty is the same as the order of the test sequence. At $m = 50$, INSOP executes the first adaptive learning of the personal profile and sorts the remaining 150 items into the first short sequence in the order of the testee’s degree of interest estimated using the learned personal profile. As a result, the dotted line begins to go upward and deviates from the solid line at $m = 50$. This indicates that INSOP moves interesting items forward and uninteresting items backward. INSOP executes the second learning and sorting of the remaining items in the first short sequence at $m = 100$. At $m = 200$, INSOP sorts all the items in the second short sequence and executes the adaptive learning and sorting of the remaining items in the short sequence at $m = 250$.

The density of interesting items at the head of the sorted sequence is as high as that in the retrieved items using the key word matching method. The number of interesting items including in the head is larger than the number including in the retrieved items.

These graphs show that INSOP is a good information filtering system which sorts items in order of interest-

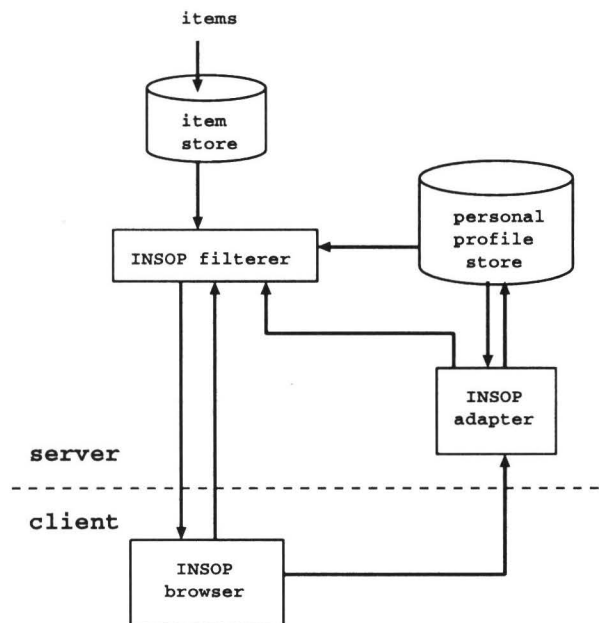


Figure 2: The flow of items through the current INSOP system

ing, neutral and uninteresting items compared to the key word matching method.

4 INSOP in Field Test

In this section, we demonstrate the current INSOP system in our field test. First, we explain the architecture of the current INSOP system. Next, we show how to use the INSOP browser to receive the INSOP service, and we present the results of the field test.

4.1 Architecture

Figure 2 shows the flow of items through the following major architectural components of the current INSOP system.

- *Item store.* Provides storage for items sent by information providers.
- *Personal profile store.* Provides storage of a personal profile.
- *INSOP filterer.* Sorts items in order of user's interest.
- *INSOP browser.* Provides the user with an interface for accessing INSOP services. This includes facilities for displaying items and receiving user inputs.
- *INSOP adapter.* Organizes user's personal profile according to the adaptive learning algorithm described in the previous section.

The current INSOP system uses a client/server model. The client(INSOP browser) is installed into over 50 personal computers (PCs) and operates as an application software on each PC whose OS is Windows95. The server

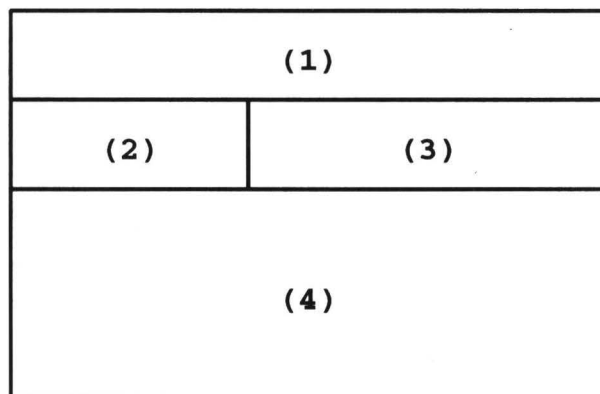


Figure 3: Four areas of the INSOP browser

(INSOP filterer/adaptor) operates as a daemon process on several UNIX machines. The clients and the servers send data (sorted items, user's input etc.) to each other according to our original protocol based on the TCP/IP.

4.2 Browser

If we execute the INSOP browser, the browser requests for us to input a server name, our user IDs, and our passwords. After receiving them, the browser opens the window in which there are 4 areas as shown in Figure 3. In area (1), the name of the opened file or the folder is shown. In area (2), the trees of the files and folders are shown according to the same format as Microsoft Explore. In area (3), accessible files or folders are shown. Area (4) displays items with the two values, "needs" and "reliance", defined at the end of Section 2.

We open an item folder and select a file using a PC mouse in the same manner as with Explore. The items included in the file are displayed in area (4) (we show a copy of the browser at Figure 4 where Japanese items are displayed). The bold frame indicates the item to which we should respond using the mouse buttons - its left button for "interesting" and its right button for "uninteresting" and is moved using the mouse. If we mark an item with "interesting" or "uninteresting", then the color of the left block in area (4) where "needs" and "reliance" are displayed changes from white to dark or light gray, respectively. If we attempt to open a new file after marking items using the buttons, the INSOP browser sends the set of our inputs and the key words with the items to the INSOP adapter. The INSOP adapter organizes our personal profiles according to the adaptive learning algorithm. Then, the INSOP filterer, based on our personal profiles, sorts the items of the new file and sends the sorted items to the INSOP browser. The INSOP browser displays them in area (4).

The calculation time from the start of the adaptive learning to the sorting items is less than 1s in the cases where the number of items is 100. The actual waiting time, depending on the load levels of the network and the server, is 1~5s on the local area network (LAN) of Matsushita Research Institute, Tokyo Inc. (MRIT).

4.3 Field Test

We executed the field test of the current INSOP system on the LAN. The purpose of the test was to ensure that INSOP responds to the various interests of many users. The items used in the test were headlines of Japanese newspapers which were supplied by information providers at a rate of about 600 per day. Over 50 users participated in the test including not only researchers and engineers in material, mechatronics, image processing, speech technology and mobile communication but also staff members.

We asked a set of questions to the users regarding their impression of INSOP and the sorting ability 3 times: when the INSOP filterer was operating (case 1), when the we stopped the INSOP filterer (case 2), and when the INSOP filterer was operating again (case 3).

Table 3 shows the result of a questionnaire on the current INSOP system.

Table 3a Impression of INSOP

	excellent	normal	poor
case 1	7%	50%	39%
case 2	0%	4%	35%
case 3	11%	68%	18%

Table 3b Sorting ability of INSOP

	excellent	normal	poor
case 1	0%	79%	21%
case 2	0%	0%	14%
case 3	14%	75%	11%

Furthermore, over 90% of users answered, "INSOP is useful reading the headlines of Japanese newspapers."

5 Conclusion

In this paper, we presented an adaptive personal information filtering system, INSOP, whose personal profile is

automatically organized from an empty profile by means of user-system interaction. The INSOP system is a good information filtering system which sorts items in order of interesting, neutral, and uninteresting items. The results of the experiment and field test suggest that INSOP can be put to practical use.

Acknowledgments

We thank Electric Library Co. which allows us to demonstrate the INSOP system using the headlines of Japanese newspapers, Mr. Natsuki Oka, Mr. Seiji Kishimoto and Mr. Tatsuro Itoh for their useful comments, and the members of MRIT who cooperated with us as testees.

References

- [Foltz and Dumais, 1992] Peter W. Foltz and Susan T. Dumais. Personalized information delivery: an analysis of information filtering methods. *Communications of the ACM*,35(12):51-60, December 1992.
- [Foltz, 1990] Peter W. Foltz. Using Latent Semantic Indexing for information filtering. *Proceedings of the ACM Conference on Office Information Systems*, pages 40-47, Boston, April, 1990.
- [Goldberg et al., 1992] David Goldberg, David Nichols, Brain M.Oki, and Douglas Terry. Using collaborative filtering to weave an information TAPESTRY. *Communications of the ACM*,35(12):61-70, December 1992.
- [Asakura et al., 1995] Takayoshi Asakura, Koji Kida, Hiroyuki Tarumi, and Toshiaki Miyashita. Agent based information filtering *Information Processing Society of Japan SIG Notes* 95-20-7:49-55, March 1995.

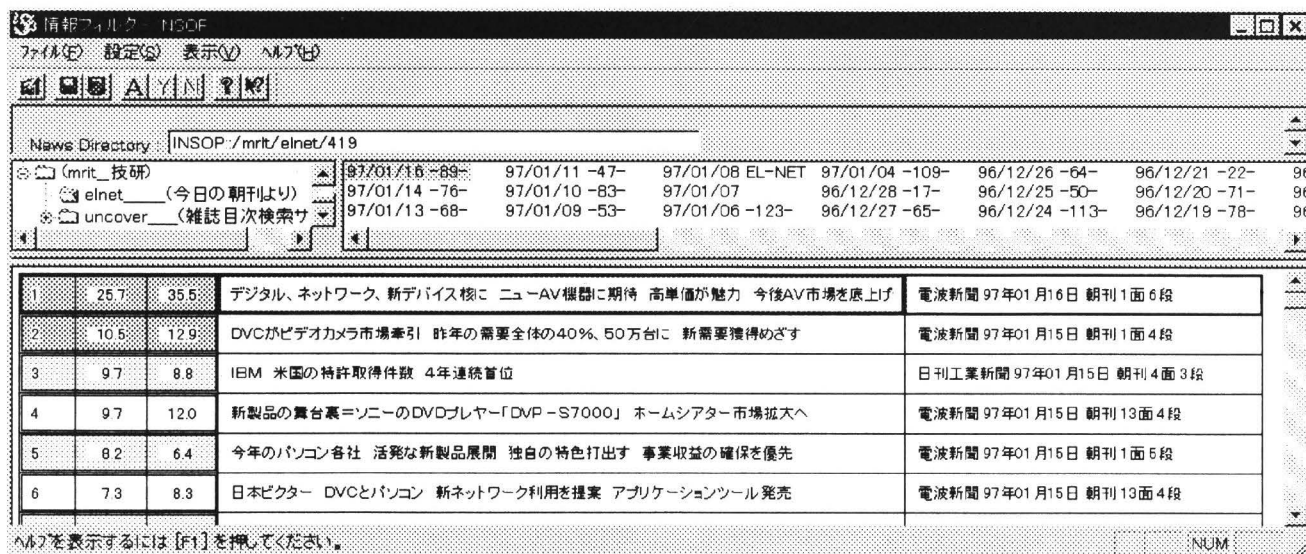


Figure 4:INSOP browser in the field test.