Technical Report

Flexible Category Structure for Supporting Document Retrieval and Its Evaluation

Kokoro Nakagawa, Yoshiaki Takata, and Hiroyuki Seki

{kokoro-n,y-takata,seki}@is.aist-nara.ac.jp

December 22, 2000

Graduate School of Information Science
Nara Institute of Science and Technology

Abstract

A method for supporting document retrieval by constructing a flexible category structure is proposed. In this method, a category structure suitable for retrieval by the user is constructed whenever a query is submitted. The method uses categorization viewpoints as *a priori* knowledge, where a categorization viewpoint is a finite set of category names. A set of documents retrieved by initial keywords is decomposed by categorization viewpoints and each decomposition is scored by clearness or entropy. First, the system presents high-scored decompositions, and then the user selects an appropriate decomposition by considering the score. The decomposition can be recursively performed until a category structure of suitable size is obtained.

By using the experimental system which has 68 categorization viewpoints, the proposed method was evaluated with BMIR-J2 test collection. Preliminary experiments show that the set of documents decomposed by the proposed method have higher precision than those decomposed by clustering (K-means). An experiment using human subjects was also performed and we obtained the following results: (1) Quality: The proposed method provides better BMIR-match than both the simple keyword-based method and the clustering method, where BMIR-match is the relative number of documents which BMIR-J2 assumes to be relevant to the topic in the set of documents which the subjects decide to be relevant. (2) Efficiency: Using the proposed method, the subjects can find the same number of relevant documents in shorter time than using the other methods. (3) Usability: The answers given by the subjects to the questionnaire suggest that the proposed method parallels the keyword-based method in usability even for the users familiar with the keyword-based method.

Contents

1 Introduction

As the number of documents increases explosively on the World Wide Web (WWW), more than a hundred Web sites (*search services*) for aiding document retrieval on WWW exist; however, usability problems still remain.

A search engine such as *Alta Vista* provides a user with a list of matching documents relevant to input keywords. However, selecting appropriate keywords which can extract only documents relevant to the user's purpose of retrieval is difficult, especially for users unfamiliar with searching.

A directory service such as *Yahoo!* provides a user with a human-classified collection of Web-site descriptions. This method is simpler than the one by a search engine. However, such a category structure often has too much volume to allow easy browsing. Another problem is that the categorization given by a system is not always suitable for the user's purpose of retrieval. For example, when a user is retrieving some information on official research facilities in Nara which are responsible for environmental problems, a hypothetical path in the category structure would be;

$$\text{root} \rightarrow \textit{Nara} \rightarrow \textit{Environment and Nature} \rightarrow \textit{Institutes}, \quad (1)$$

but the provided category structure does not always have such a path. The relevant information is often scattered around several paths such as:

root → *Science* → *Environmental Engineering*,
root → *Regional* → *Nara* → *Health*,
root → *Society and Culture* → *Environment and Nature* → *Government Agencies*.

In the above case, the user hesitates over which category to select first, and even if the user decides a path after some consideration, she/he will only obtain a subset of documents with poor precision, since relevant documents are scattered around several categories. The problem is that the categorization criterion varies according to the user's purpose of retrieval and one fixed structure cannot cope with this versatility of criteria.

We proposed a method for supporting document retrieval, by providing a *small* and *flexible* category structure adaptable to the user's purpose of retrieval [16]. By providing a *small* category structure, we can avoid a structure which is too large to easily use. On the other hand, *flexibility* means that for each retrieval, a hierarchical

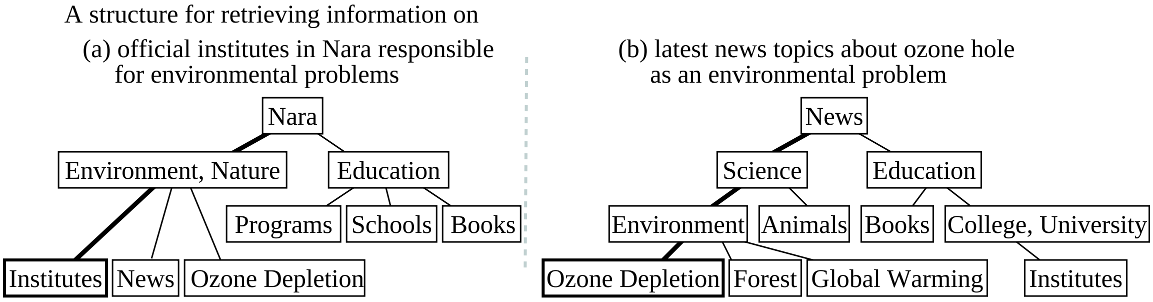


Figure 1: Examples of Category Structures

structure is constructed which contains paths suitable for the user's purpose of retrieval. For the above example, where official research facilities responsible for environmental problems are retrieved, the category structure provides the path shown in (1) (cf. Figure 1 (a)). When the user is retrieving the recent topics of environmental problems, the category structure provides the following path (cf. Figure 1 (b)):

root → *News* → *Science* → *Environment* → *Ozone Depletion*.

This method uses categorization viewpoints as *a priori* knowledge, where a categorization viewpoint is a finite set of consistent category names. A set of documents retrieved by initial keywords is decomposed by categorization viewpoints and each decomposition is scored by clearness or entropy. A user selects an appropriate decomposition by considering the score. The decomposition is recursively performed until a category structure of reasonable size is obtained. As a result, the relevant documents are located in a few categories, which helps the user retrieve relevant documents. Furthermore, other categories related to the user's purpose of retrieval are also located near the categories selected by the user in the category structure.

The rest of this paper is organized as follows: In Section 2, the method proposed in [16] is reviewed. By using BMIR-J2 test collection [8], we conducted several experiments to evaluate the proposed method compared with a simple keyword-based retrieval method and a traditional clustering method (K-means). Section 3 briefly presents the result of the preliminary experiments without human subjects. In Section 4, we describe in detail the experiments which use human subjects. The results are carefully analyzed from the following viewpoints. (1) Quality: Which method provides the best BMIR-match? The BMIR-match is the relative number of documents which BMIR-J2 assumes to be relevant to a topic in the document set which the subjects mark. (2) Efficiency: Which method exhibits the best recall in a fixed amount of time? (3) Usability: Which method do the subjects feel is most useful?

Related Works

Many approaches to supporting document retrieval by estimating and adapting the user's purpose of retrieval exist. RCAAU[7] performs text data mining on the initial set of documents and provides a user with secondary keywords to obtain a set of documents with higher precision. The system in [1] also provides secondary keywords or phrases for a user. In [1], a word (called *facet*) which has high *lexical dispersion* is assumed to be useful for identifying key concepts related to the initial set of documents, and facets and noun phrases containing a facet are provided as possible subtopics related to the initial query. In [4], the user's log, such as previously input keywords, are analyzed and a query issued by the user is automatically refined to achieve higher precision and recall. These approaches are useful for narrowing down an initial set of retrieved documents. However using these methods, it appears difficult to decompose a document set in a consistent way and/or build a category structure according to each search intention.

Statistical approaches have been used for decomposing a set of documents. For example, [10] uses clustering to construct a category structure. Although [10] aims at constructing a single category structure for a whole set of documents, the method could be used for constructing a category structure for each retrieval if documents are restricted to those retrieved by initial keywords. A statistical approach has the advantage that *a priori* knowledge such as categorization viewpoints in this paper is not needed. On the other hand, there are some drawbacks in this approach. For a user to select a cluster appropriately, sufficient information on each cluster should be provided for the user. However, to automatically associate a title or a summary with a cluster is difficult. In [15], a method for automatically constructing a concept hierarchy without *a priori* knowledge is proposed. Unlike clustering, each node of the concept hierarchy derived by this method is a word or a noun phrase which may simply indicate a concept. However, the method only provides a hierarchy of conceptual words, and cannot be directly used to obtain a set of documents with higher precision in.

Similar to our method, HIBROWSE[11] uses portions of a thesaurus (called *views*) as *a priori* knowledge. A user of HIBROWSE can see categorizations made by several views at once and can choose a keyword in any view to narrow down the temporally retrieved set. Since HIBROWSE was designed for a special purpose (searching medical bibliographies), the selection of views is left to a user and no method for recommending appropriate views is proposed.

Information visualization[3, 13] is another approach which could be consistently incorporated into the above-mentioned methods, including ours.

2 The Method

The proposed method consists of the following two steps: (i) Identify a set of documents which roughly reflect the user's purpose of retrieval. (ii) Construct a category structure depending on the set of documents obtained in (i).

In step (ii), the system uses *categorization viewpoints*, which are built into the system as *a priori* knowledge for constructing a category structure. A categorization viewpoint is a finite set of categories. For example, a categorization viewpoint *Regional* may include *Kyoto*, *Osaka*, and *Nara*, and a viewpoint *Entertainment* may include *Books*, *Games*, *TV*, and *Travel*.

Let

$$\mathcal{S} = \{S_j \mid 1 \leq j \leq m\}$$

be a set of categorization viewpoints where a categorization viewpoint $S_j = (l_j, W_j)$ is a pair of the viewpoint name l_j and a finite set of category names $W_j = \{w_{j1}, \ldots, w_{jk_j}\}$. Each W_j is assumed to be a subset of the set W of all words described in subsection 2.1.

A categorization viewpoint corresponds to a set of brother nodes (the set of children of a node) in the category structure of a directory service (Figure 1). As described in the following, different category structures are constructed depending on which categorization viewpoints are used. The system and a user construct interactively an appropriate category structure by selecting appropriate categorization viewpoints.

2.1 Representation of documents

The *vector space model*[14] is used for defining the similarity coefficient between a document and a category. In this model, each document is represented by a vector which consists of the relevance values of each word to the document. Let

$$W = \{w_1, w_2, \ldots, w_n\}$$

be the set of all words. The feature vector of a document d is

$$c(d) = (c_{w_1,d}, c_{w_2,d}, \ldots, c_{w_n,d}),$$

where $c_{w_i,d}$ is the relevance value of word w_i to document d. The relevance value $c_{w_i,d}$ is defined as the frequency of w_i in d. In a more sophisticated model, a larger weight

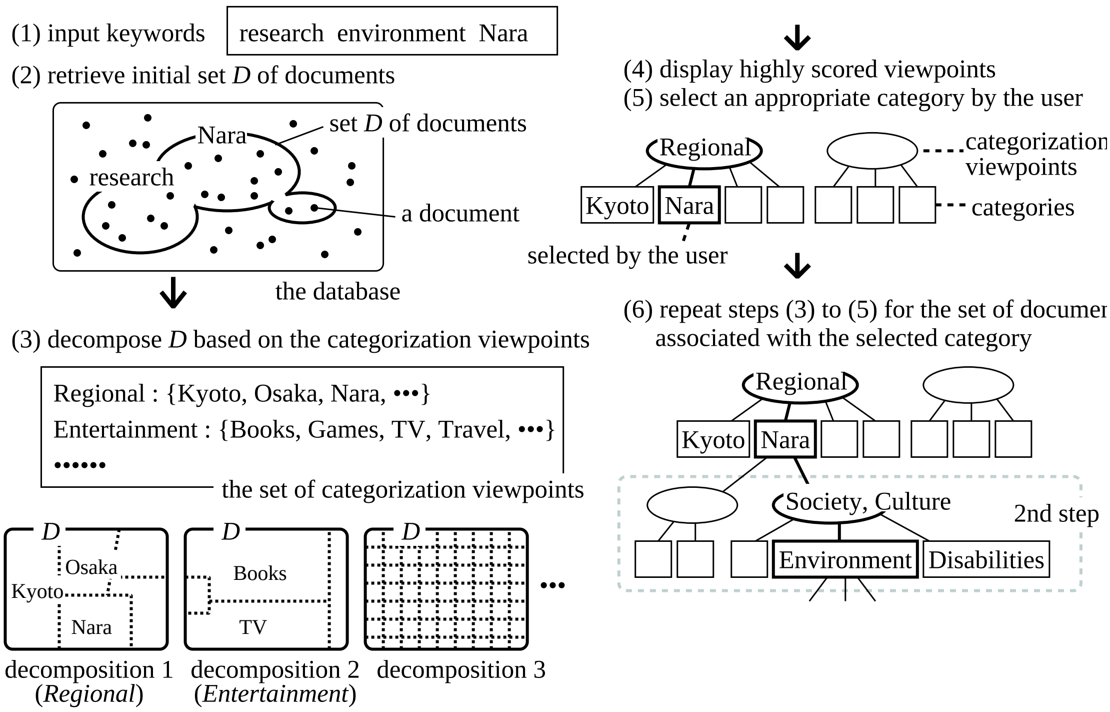


Figure 2: Behavior of the System

is given to a word which is less frequent in the document collection. Also, the word frequency is normalized against the document length[2].

The *similarity coefficient* $sim(u, v)$ between two vectors u and v is defined as the normalized inner product between u and v; that is,

$$sim(u, v) = \frac{u \cdot v}{|u|\ |v|}$$

where $|u|$ represents the length of vector u.

2.2 The system behavior

A user and the system perform the following steps (in Figure 2):

(1) The user inputs a few keywords to the system. Since these keywords are used for approximately limiting the documents to be searched, the user does not have to select them carefully.

(2) The system retrieves a set of documents which seem to relate to the input keywords from the document database. In the experimental system, documents are retrieved by the disjunctive query of the keywords. The retrieved set of documents is denoted as D.

(3) The system decomposes D into subsets by each categorization viewpoint S_j in the system; for each document $d \in D$, the system decides a category $w_j^{(d)}$ to which d belongs among the categories $W_j = \{w_{j1}, \ldots, w_{jk_j}\}$ of viewpoint S_j. The category $w_j^{(d)}$ is one of the categories in S_j which has the maximum similarity coefficient with document d. The decomposition $(D_{j1}, \ldots, D_{jk_j})$ of a set D of documents by S_j is defined as follows:

$$\begin{aligned} D_{ji} = \{d \in D \mid & \mathit{sim}(c(w_{ji}), c(d)) \\ & = \max_{1 \le h \le k_j} \mathit{sim}(c(w_{jh}), c(d))\}.^{\dagger} \end{aligned}$$

D_{ji} is called the set of documents *associated with* category w_{ji}.

The vector $c(w_{ji})$ is called the *category vector* of w_{ji}. The initial value of $c(w_{ji})$ is defined as $(\delta_1, \delta_2, \ldots, \delta_n)$ where $\delta_k = 1$ if $w_{ji} = w_k \in W$ and $\delta_k = 0$ otherwise. Since $\mathit{sim}(c(w_{ji}), c(d)) = c_{w_{ji},d}/|c(d)|$, decomposition using these initial values is equivalent to identifying the category which has the maximum (normalized) relevance value to each document.

In the decomposition based on the initial value of $c(w_{ji})$, there exist many documents d such that $sim(c(w_{ji}), c(d)) = 0$ with any category w_{ji}, because of the discreteness of the category vectors. Accordingly, we smooth the categorization by redefining the category vector $c(w_{ji})$ as the average of the feature vectors of documents associated with the category w_{ji}:

$$c(w_{ji}) = \sum_{d \in D_{ji}} c(d)/|D_{ji}|\,,$$

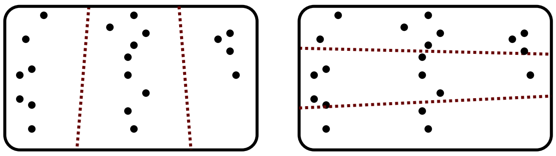
where $|D_{ji}|$ denote the number of documents in D_{ji}.

After decomposition, the system scores each viewpoint for D. The scoring function will be defined in subsection 2.3.

(4) The system shows the user the viewpoints which have a higher score. The system also shows the categories of the shown viewpoints.

---

†For simplicity, if there are more than one h which maximizes $\mathit{sim}(c(w_{jh}), c(d))$, then let d belong to one of D_{jh}.



(a) the decomposition is clear (b) not clear

Figure 3: Clearness of Categorization

(5) The user selects a category among the shown categories. Let D' be the set of documents associated with the selected category. The user also chooses to accept D' as the final result of the session or to further decompose D' recursively into subsets.

(6) When the user chooses to decompose D' in step (5), go to step (3) where $D = D'$.

2.3 Scoring categorization viewpoints

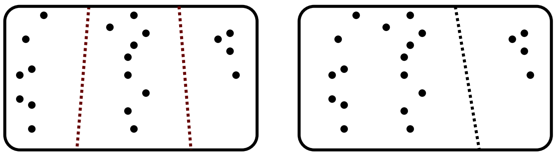
We use two criteria for scoring a categorization viewpoint: the *clearness* of categorization and the *entropy* of categorization. The *score* $e_D(S_j)$ of a categorization viewpoint S_j for a set D of documents is defined as expression (2) when clearness is used, and is defined as expression (3) when entropy is used. The effectiveness of the two criteria is empirically compared to each other in Sections 3 and 4.

2.3.1 Clearness of categorization

We say a categorization of a set D of documents is clear when, for each document $d \in D$, we can clearly decide a category $w_j^{(d)}$ to which d belongs; that is, the similarity coefficient between d and $w_j^{(d)}$ is much larger than the similarity coefficient between d and a category other than $w_j^{(d)}$ (as Figure 3 (a)). Although a categorization viewpoint resulting in high clearness does not always reflect the user's purpose of retrieval, the viewpoint is at least informative to the user since the set of documents associated with each category is clearly separated from each other (Figure 3). The clearness of a categorization viewpoint S_j for a set D of documents is defined as

$$\sum_{d \in D} \mathit{sim}(c(w_j^{(d)}), c(d))/|D|, \tag{2}$$

which is the average of the similarity coefficient between each document and the category to which the document belongs.



the entropy of (a) > the entropy of (b)

Figure 4: Entropy of Categorization

2.3.2 Entropy of categorization

For a set D of documents and a categorization viewpoint S_j, let us define the entropy of S_j for D as

$$H(S_j) = -\sum_{i=1}^{k_j} P_i \log P_i, \tag{3}$$

where

$$\begin{aligned} k_j &: \text{the number of categories in } S_j, \\ P_i &= \frac{|D_{ji}|}{|D|}, \text{ and} \\ D_{ji} &: \text{the set of documents associated with category } w_{ji}. \end{aligned}$$

A categorization viewpoint resulting in high entropy offers an efficient means of narrowing down the documents. If the user selects a category in a viewpoint with higher entropy, the user would obtain a smaller subset of D (cf. Figure 4).

3 Preliminary Experiments

3.1 Overview

Our prototypic system consists of two components: the database constructor and the query processor. The database constructor pre-processes the document database, and the query processor responds to a user's query. The system uses a keyword-based text retrieval system Freya[6] for constructing the *inverted index*[5], which is a data structure to retrieve documents containing specified keywords. The system also uses a Japanese morphological analysis system ChaSen[9] for identifying the set W of words (in subsection 2.1). The query processor was implemented as a CGI program.

To evaluate the method, we defined the following four experimental systems and we conducted the two experiments described below.

- System A: this system uses clearness for scoring categorization viewpoints.
- System B: this system uses entropy for scoring categorization viewpoints.
- System C: this system does not use categorization viewpoints but decomposes a set of documents by clustering.
- System D: the keyword-based retrieval system Freya.

System C is for comparing our method which uses *a priori* knowledge with statistical methods for decomposing a set of documents. We consider K-means method[17], which is one of the simplest but most useful methods. In the experiments, we let the number of clusters = 10. If the centers of the clusters do not converge after 100 iterations, then terminate the iteration and exit.

In these experiments, we did not use a human subject, but assumed an ideal user who always selects a categorization viewpoint with the highest score and categories with which a set of documents with high precision are associated.

Ten topics out of the BMIR-J2 test collection[8] were used in the experiments, which are shown in Table 1. BMIR-J2 is a test collection (which includes 5,080 articles) for evaluation of information retrieval systems, based on the Mainichi Shimbun Newspaper CD-ROM '94 data collection. BMIR-J2 was constructed by the SIG Database Systems of the Information Processing Society of Japan, in collaboration with the Real World Computing Partnership.

We prepared a set of categorization viewpoints (including 68 categorization viewpoints and 1,723 categories) by referring to several category structures served by exis-

Table 1: Topics Used in the Experiments

Topic No.	Name of topic *1	The number of relevant documents
i	Agricultural Chemicals	26
ii	Beverages	60
iii	Tax Reduction	302
iv	Nuclear Weapons	98
v*2	The Educational Industry	17
vi	Trend of Stock Prices	37
vii	Discount Brokers	36
viii	Motion Pictures	21
ix	Exports from Southeast Asia to Japan	59
x	Harassment against Women in Civil Employment	80

*1: Every topic is originally written in Japanese.

*2: This topic is replaced by "The Video Recorders (33)" in the experiments in Section 4.

tent directory services.

3.2 Experiment 1

This basic experiment is used to compare precision-recall among Systems A, B, C and D. For each topic q in the test set (Table 1) and for each $r = 0.1, 0.2, \ldots,$ and 1.0, perform the following (1) to (4) to evaluate Systems A and B.

(1) Let *Result* = $\emptyset$. Retrieve the set D_0 of documents by a few keywords for topic q and let $D = D_0$. Calculate the score (clearness for System A and entropy for System B) of each categorization viewpoint for D.

(2) Choose S_j with the highest score among the categorization viewpoints $S_1, S_2, \ldots, S_m$ ($m = 68$). Let $W_j = \{w_{j1}, w_{j2}, \ldots, w_{jk_j}\}$ be the set of categories of S_j and let D_{ji} be the set of documents associated with category w_{ji} ($1 \leq i \leq k_j$).

(3) Choose category w_{jh} with the highest precision among W_j which has not yet been chosen. If $|D_{jh}| \leq 30$ then let *Result* = *Result* $\cup D_{jh}$. Otherwise let $D = D_{jh}$

and go to (2).

(4) Repeat (2) and (3) until (the recall of *Result*) $\geq r$. Output the precision and the recall of *Result*.

Let *Rel*(q) denote the set of relevant documents for a topic q. The precision and recall of *Result* are defined as follows:

$$\textit{precision} = \frac{|\textit{Result} \cap \textit{Rel}(q)|}{|\textit{Result}|}, \quad \textit{recall} = \frac{|\textit{Result} \cap \textit{Rel}(q)|}{|D_0 \cap \textit{Rel}(q)|}.$$

To evaluate System C, instead of step (2), the K-means method is performed. The obtained clusters are referred to as $D_{j1}, D_{j2}, \ldots, D_{jK}$. To evaluate System D, we first obtain the ranked list of documents retrieved by a few keywords for each topic, and then calculate the precision–recall pairs according to the method defined in [18].

Results

The average precision-recall curves of ten topics are shown in Figure 5. In the figure, the curves labeled with Clearness, Entropy, Clustering and Keyword are those for Systems A, B, C and D, respectively. In Figure 5, we see that the results for Systems A and B are better than those for Systems C and D (on the average). On the other hand, the results for System B (entropy) is slightly better than those for System A (clearness).

3.3 Experiment 2

In this experiment, the correlation between the categorization viewpoint score and the precision of the retrieved documents is evaluated for Systems A and B. The outline of the experiment is as follows :

- Retrieve the set D of documents by a few keywords. Calculate the score (clearness or entropy) of each categorization viewpoint for D. For each categorization viewpoint $S_j (1 \leq j \leq m)$, repeatedly add the set of documents associated with the category of highest precision to *Result*$_j$ until (the recall of *Result*$_j$) $\geq$ the threshold.
- Calculate the correlation coefficient of the score of S_j and the precision of *Result*$_j$ with $1 \leq j \leq m$.

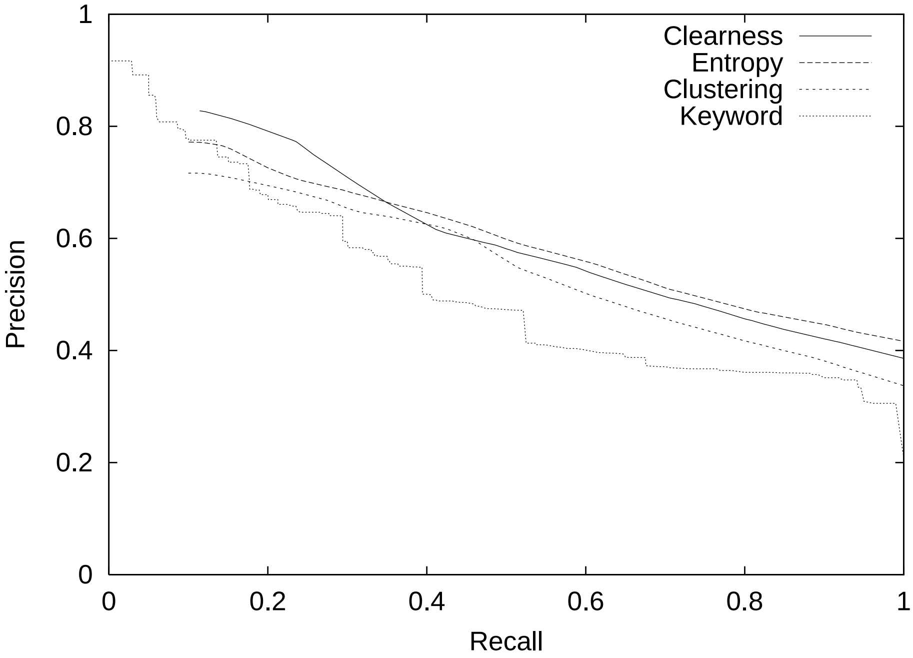


Figure 5: Average Precision-Recall Curves

Results

The results for Systems A and B are shown in Table 2. For System A, the precision and the score (clearness) have a positive correlation for 8 or 9 topics. For System B, the precision and the score (entropy) have a positive correlation for all topics.

For each of the categorization viewpoints S_j with the top ten scores, the number of categories which are needed to satisfy (the recall of $Result_j$) $\geq r$ is examined. The average number of such categories among ten topics is also shown in Table 2, where the average numbers for System A are less than those for System B. The two-way analysis of variance with independent variables, the score (clearness or entropy) and the topic, results in a significant difference between clearness and entropy in the case of $r = 0.5$ and $r = 0.8$ ($p < 0.05$). Since it is desirable for a user to choose fewer categories to achieve the same recall, from a usability point of view, this result suggests that clearness is a better criterion of the score than entropy.

Table 2: Averages of Correlation Coefficients between the Score and the Precision

Scoring criterion	r *1	The number of topics	$\overline{R}$ *2	The number of categories *3
Clearness	0.2	9	0.496	2.2
	0.5	8	0.595	4.1
	0.8	8	0.646	7.3
Entropy	0.2	10	0.596	2.3
	0.5	10	0.675	4.9
	0.8	10	0.684	8.7

*1: The threshold of recall

*2: The average of the correlation coefficients between the score and the precision.

*3: The average number of categories which are needed to satisfy the recall $\geq r$ on each of the highest-scored ten categorization viewpoints.

4 Experiments with Human Subjects

4.1 Overview

We conducted experiments using human subjects to evaluate Systems A, B, C, and D used in the preliminary experiments. Each subject was asked to retrieve documents relevant to a specified topic in a limited amount of time. Which system a specified subject should use for a specified topic was determined so that each topic–system pair was assigned to five different subjects.

Search Topics

The same ten topics as in the preliminary experiments (Table 1) were used, except the topic "The Educational Industry" was replaced by "The Video Recorders" since the number of documents relevant to the former topic was relatively small. We will refer to each topic by the topic number shown in Table 1 .

Subjects

Twenty subjects participated in this experiment. All of them were graduate students in information science, and they were paid 1000 Yen ‡ for their participation. The attributes of the subjects are summarized as follows :

- age : the latter half of the 20s – the first half of the 30s
- experience in personal computers : 2 – 18 years
- experience in Windows98 : 1 – 7 years
- experience in search services on WWW : 1 – 7 years

All subjects mainly use keyword-based search services for information retrieval. Hence we conjecture :

> most of the subjects firmly establish user-interface of a keyword-based retrieval system as the mental model[12] for an information retrieval.

Equipment

To operate the systems, the subjects used Microsoft Internet Explorer version 5.0 on DELL Dimension XPS D300 PC and Microsoft Windows98 Second Edition. Every operation of the subject was recorded as screen video by Lotus ScreenCam97.

‡One dollar is approx. 105 Yen.

Procedure

- Several days before the experiment

  (1) Documents : The subjects were given a document which described the purpose and overview of the experiment and an operating manual of the four systems.

  (2) Prior exercise : The subjects were asked to retrieve documents for a sample topic by using the same systems as used in the experiment.

- Just before the experiment

  (1) Oral explanation : The experimenter explains to each subject the purpose and overview of the experiment using the distributed documents. At the same time, the subject is shown ten topic–system pairs for which she/he should perform the experiment. For example, subject No.1 has to perform the experiment in order of (topic iii - System C), (vii-C), (ii-B), (x-B), (vi-B), (iv-D), (viii-D), (i-A), (ix-A), and (v-A).

  (2) Rehearsal : The subject is allowed to retrieve a topic specified for exercise using any of the four systems.

- At the experiment
  The subject performs retrieval for each topic–system pair during ten minutes, or until the number of marked documents exceeds twenty. The subject starts the retrieval by inputting keywords, and marks the documents which are regarded as relevant to the topic.
- After the experiment (questionnaires)
  The subjects are asked about the usability of and the satisfaction with each of the four systems as well as good points and/or problems of each system.

The other conditions of the experiment are as follows :

- The experimenter observes the behavior of the subject during the experiment while the subject is banned from talking with the experimenter.
- All the operations of the subject are recorded in two ways : screen video recording software on PC and video camera shooting from the rear of the subject.
- To cancel the effect of the characteristics of the subjects, the set of topic–system pairs is designed to be different from one subject to another.

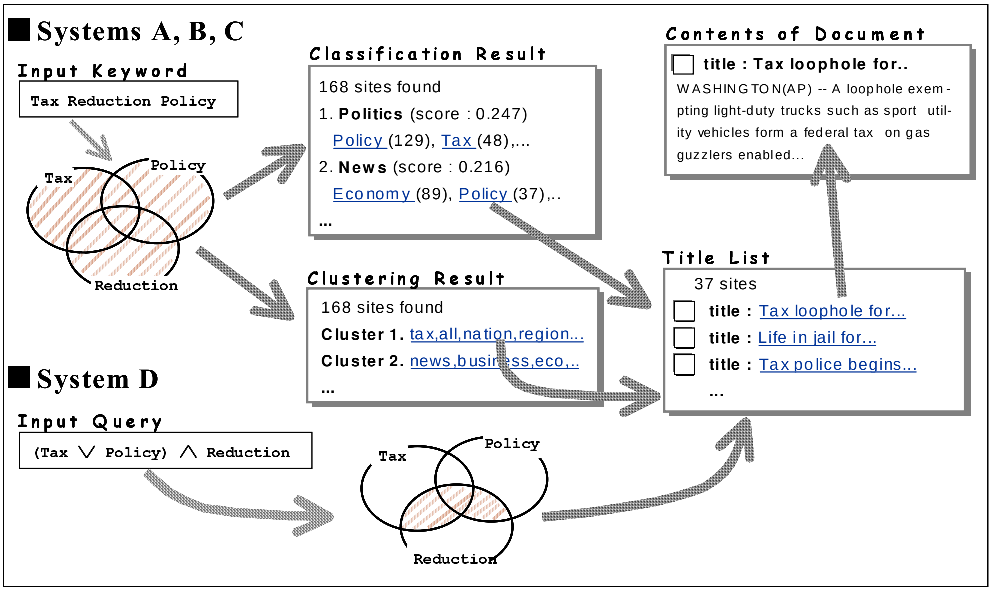


Figure 6: User Interface

- To cancel the effect of learning, the order of using the systems is randomized among the subjects.

How to Mark Documents (User Interface)

Figure 6 shows the user-interface of the experimental systems.

- Systems A, B, and C list the titles of documents in the set D' when the subject accepts D' as the final result in step (5) in Section 2.2 (**title list** screen in Figure 6) . In System C (clustering), each cluster is labeled with ten words which appear most frequently in the set of documents associated with the cluster.
- In System D, the title list of the document set retrieved by a query is displayed when the subject inputs the query (**title list** screen).
- Contents of a document are displayed when the subject selects the document title out of the title list on a screen (**contents of document** screen) .
- The subject can mark the documents which are regarded as relevant to a given topic on either **title list** screen or **contents of document** screen.

Measured Value

Let $Title(q)$ be the set of documents which appear on the **title list** screen during the retrieval for a topic q. We regard $Title(q)$ as the set of documents retrieved by the subject. $Title(q) = \emptyset$ at the beginning of the retrieval for topic q, and $|Title(q)|$ increases in two ways: The subject opens a new **title list** screen [§] , or the subject scrolls down a **title list** screen. Let $Mark(q)$ be the set of documents on which the subject marks for a topic q. $Mark(q)$ is a subset of $Title(q)$. Let $Rel(q)$ be the set of relevant documents provided by the BMIR-J2 test collection. The *precision*, *recall* and *BMIR-match* for a topic q are defined as follows:

$$precision = \frac{|Title(q) \cap Mark(q)|}{|Title(q)|}, \quad recall = \frac{|Mark(q)|}{|Rel(q)|},$$

$$BMIR\text{-}match = \frac{|Mark(q) \cap Rel(q)|}{|Mark(q)|}.$$

4.2 Results and Discussions

(a) BMIR-match (Quality)

The BMIR-match of every topic and the average BMIR-match of ten topics are shown in Table 3. As the BMIR-match becomes higher, the relative number of relevant documents provided by BMIR-J2 increases in the set of documents marked by the subject. In other words, high BMIR-match implies a high quality of the set of documents marked by the subject.

In Table 3, the BMIR-match of the proposed systems (System A and B) are on the average higher than for Systems C and D. System B has the highest value, followed by Systems A, D, and C (though the BMIR-matches of the last three are almost equal).

When we examine the results for each topic, System B provides a higher BMIR-match than Systems C and D for seven of the ten topics. On the other hand, System D exceeds both Systems A and B for two topics (topics iii and iv). These two topics are rather "simple" ones to which it is easy to find the relevant documents, since they have many relevant documents provided by the BMIR-J2 test collection (Table 1).

(b) Precision (Quality)

The bivariate distribution of search time (kiloseconds) and precision for all topics is shown in Figure 7. High precision means that the subject retrieved few irrelevant

[§] $|Title(q)|$ does not increase if the opened **title list** screen has already been opened.

Table 3: BMIR-match

	System A	System B	System C	System D
Topic i	0.947	0.984	0.958	0.982
Topic ii	0.968	0.938	0.908	0.883
Topic iii	0.990	0.990	0.980	1.000
Topic iv	0.580	0.458	0.560	0.726
Topic v	0.829	0.907	0.872	0.817
Topic vi	0.471	0.606	0.578	0.505
Topic vii	0.537	0.516	0.476	0.479
Topic viii	0.929	0.900	0.607	0.719
Topic ix	0.651	0.727	0.667	0.600
Topic x	0.835	0.900	0.963	0.890
Average	0.774	0.793	0.757	0.760

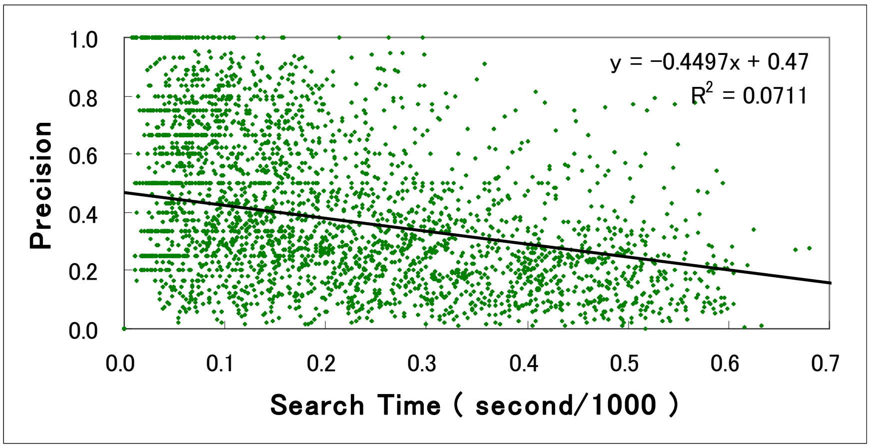


Figure 7: Search Time and Precision

documents. Thus, we examine the average of precision during the retrieval for each topic, as shown in Table 4.

In Table 4, we see that the proposed systems (System A and B) are defeated on the average, by both Systems C and D. System D has overwhelmingly the highest value,

Table 4: Average of Precision at the Termination of Retrieval

	System A	System B	System C	System D
Topic i	0.277	0.182	0.348	0.352
Topic ii	0.214	0.413	0.252	0.393
Topic iii	0.586	0.492	0.627	0.844
Topic iv	0.435	0.262	0.446	0.433
Topic v	0.296	0.161	0.121	0.083
Topic vi	0.235	0.558	0.258	0.536
Topic vii	0.277	0.155	0.320	0.236
Topic viii	0.105	0.087	0.156	0.124
Topic ix	0.187	0.103	0.177	0.304
Topic x	0.212	0.191	0.249	0.484
Average	0.282	0.260	0.295	0.379

followed by Systems C, A, and B with narrow margins.

When we examine the results for each topic, at least one of A and B outperforms System D in five topics. These five topics (topics ii, iv, v, vi and vii) can be thought "difficult" for the following reasons: (1) These topics have relatively few relevant documents provided by the test collection. (2) The topic names of these topics are not suitable for initial keywords; if the subject used those topic names as initial keywords, she/he gets few relevant documents and many irrelevant documents.

From the above results, we reach the following observations:

> The traditional keyword-based method is useful if a topic has many relevant documents, or if the subject can obtain many relevant documents using the topic name for an input keyword. On the other hand, the proposed method is useful for a "difficult" topic which has few relevant documents. In particular, the topic name is not suitable for a keyword.

(c) Recall (Efficiency)

The bivariate distribution of search time (kiloseconds) and recall for all topics is shown in Figure 8. The regressive line of recall on search time is also shown in Figure 8 with the determination coefficient R^2. In this figure, due to the big gap among ten

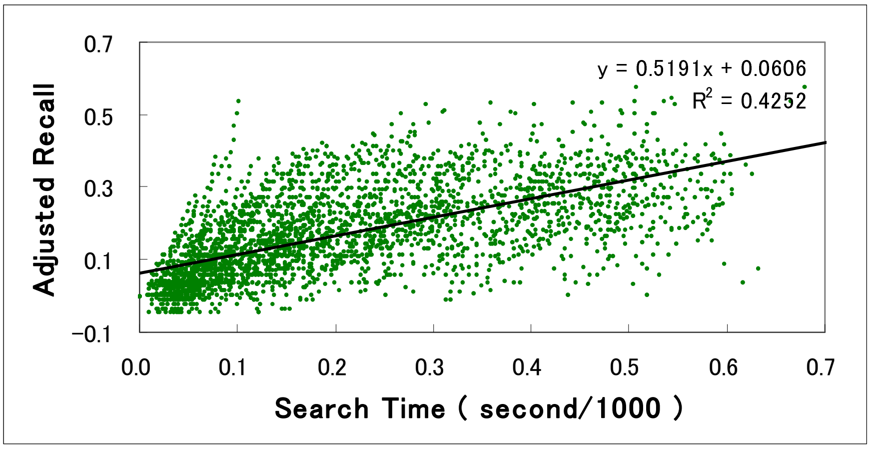


Figure 8: Search Time and Adjusted Recall

topics, we adjusted the original data as follows. Let the regressive function of recall calculated from all data of all subjects be

$$Recall = a + b \times time \tag{4}$$

and let the regressive function of recall calculated from all data of a topic q be

$$Recall = a_q + b_q \times time. \tag{5}$$

The adjusted recall $Recall'$ is defined as follows :

$$Recall' = \frac{b}{b_q}(Recall - a_q) + a. \tag{6}$$

The longer time the subject spends for retrieval, the more documents the subject marks, that is, the higher recall the subject has. Hence, the regressive line of recall on the search time generally shows the upward slant line to the right. Furthermore, the big slope (high regressive coefficient) of the regressive line is desirable since it means that the subject can find more relevant documents in a shorter time.

Table 5 shows the regressive coefficient b_X, intercept a_X, and determination coefficient R^2 for each System X:

$$Recall' = a_X + b_X \times time. \tag{7}$$

Table 5: Regressive Analysis

	System A	System B	System C	System D	all
a_X	0.0373	0.0528	0.0372	0.0914	0.0606
b_X	0.6617	0.5029	0.5263	0.4904	0.5191
R^2	0.5426	0.4436	0.5188	0.3020	0.4252

Table 6: Ranking by the Subjects

	System A	System B	System C	System D
Usefulness (0 – 10)	6.50	6.30	4.85	6.65
Satisfaction (0 – 10)	5.75	5.65	4.85	6.35
Overall Ranking (1 – 4)	2.65	2.65	1.75	3.10

System A has the biggest slope, followed by Systems C, B, and D. The proposed method (A and B) is more efficient than the keyword-based method (D) from this point of view.

(d) Answers to the questionnaires (Usability)

Comparison of the four systems : The subjects were asked to evaluate the usefulness of and the satisfaction with each system using an eleven rank grading (0 – 10). The higher rank indicates the better impression the subject felt. The subjects were also asked to arrange the four systems from the easiest to use (rank = 4) to the most difficult (rank = 1). Table 6 shows the average of these three rankings answered by all subjects.

Using the three evaluations, System D was on average the best, Systems A and B were worse than System D with narrow margin, and System C was evidently the worst. This result suggests that the proposed method using a priori knowledge for classification is more useful for subjects than the clustering method is.

On the other hand, there is a gap between System D and Systems A and B. Notice that almost all the subjects use a keyword-based retrieval system for document retrieval for at least one year in a real life, while they have used the proposed method for only a few hours. Hence, we expect the proposed method to not be inferior to the keyword-based method in usefulness, allowing that the subjects firmly establish user-interface of a keyword-based retrieval system as the mental model for an information retrieval system.

Comparison between clearness and entropy : Regarding Systems A (clearness) and B (entropy), the subjects were asked to evaluate whether the displayed (ten highest) categorization viewpoints were appropriate for a topic, using an eleven rank grading (0–10). On average, System A is slightly superior to System B (System A : 6.00, System B : 5.35).

Half of the subjects were not aware of the differences between Systems A and B since they did not retrieve the same topic using both systems. However, we obtained the following opinions from a few subjects, which point out the distinction between Systems A and B:

- System B was better than System A, since the categorization viewpoints displayed by System B were often different from the earlier ones when the set of documents was recursively decomposed (step (3) in subsection 2.2).
- System A was better than System B, since the high-ranked categorization viewpoints by System A seemed to be suitable for the topics.

Considering the above results (a) – (d), we reached the following observations :

> In both the numerical results (a) – (c) and the impressive result (d), there are few differences between Systems A and B. From these experimental results, it is difficult to conclude which system is superior.

We should try to compare the clearness and entropy in more detail. For example, the subject should use both Systems A and B for the same topic. To improve the proposed method, it might be better to provide a user with either of these two methods, according to the preference of the user or the situation of the retrieval process.

5 Conclusion

In this paper, we proposed a method for supporting document retrieval using a flexible category structure. We also discussed experimental results using by a prototypic system based on the proposed method.

In the preliminary experiments, the results showed that the proposed method provided higher precision than both the clustering method and the keyword-based method. It was also shown that scoring based on entropy provided slightly better precision than scoring based on clearness.

In the experiments with human subjects, the results showed that the proposed method allows a user to efficiently obtain a high-quality set of documents, and that the proposed method is useful for “difficult” topics. Furthermore, according to the subject’s impressions, the usability of the proposed method is not inferior to the keyword-based method, allowing that the subjects firmly establish user-interface of a keyword-based retrieval system as the mental model for an information retrieval system. On the other hand, scoring based on clearness provides the subjects a slightly better impression than scoring based on entropy. However, an advantage was pointed out for each criterion by the subjects.

As a future study, we should try to compare the scoring criteria (clearness and entropy) in more detail, and try to find a method of providing a way to select appropriate scoring criteria based on categorization to make the system more useful for retrieval.

Acknowledgments

The authors sincerely thank Associate Professor Shin Ishii of Nara Institute of Science and Technology, for providing much instructive advice on the criteria for scoring the categorization viewpoints. The authors also thank Masahide Iwasaki for his assistance and Professor Dee A. Worman of Nara Institute of Science and Technology for carefully reading and editing the manuscript.

References

[1] Anick, P. G. and Tipirneni, S.: The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking, in *Proc. SIGIR '99*, pp.153–159, 1999.

[2] Dreilinger, D. and Howe, A. E.: Experiences with Selecting Search Engines Using Metasearch, *ACM Trans. Information Systems*, Vol. 15, No. 3, pp.195–222, 1997.

[3] Fishkin, K. and Stone, M. C.: Enhanced Dynamic Queries via Movable Filters, in *Proc. CHI '95*, pp.415–420, 1995.

[4] Golovchinsky, G.: Queries? Links? Is There a Difference?, in *Proc. CHI 97*, pp.407–414, 1997.

[5] Grossman, D. A. and Frieder, O.: *Information Retrieval: Algorithms and Heuristics*, pp.134–142, Kluwer Academic Publishers, 1998.

[6] Harada, M.: *Freya version 0.92*, 1998, `http://odin.ingrid.org/freya/`.

[7] Kawano, H. and Hasegawa, T.: Data Mining Technology for WWW Resource Retrieval, in *IPSJ SIG Notes*, DBS108, pp.33–40, 1996.

[8] Kitani, T., et al.: BMIR-J2 – A Test Collection for Evaluation of Japanese Information Retrieval Systems, in *IPSJ SIG Notes*, DBS114, pp.15–22, 1998.

[9] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Imaichi, O. and Imamura, T.: Japanese Morphological Analysis System ChaSen Manual, Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology, 1997.

[10] Pirolli, P., Shank, P., Hearst, M. and Diehl, C.: Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection, in *Proc. CHI 96*, pp.213–220, 1996.

[11] Pollitt, A. S.: The Key Role of Classification and Indexing in View-based Searching, in *Proc. 63rd IFLA General Conf.*, 1997.

[12] Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T.: *Human-Computer Interaction*, Addison-Wesley, 1994.

[13] Robertson, G. G., Card, S. K. and Mackinlay, J. D.: Information Visualization using 3D Interactive Animation, *Comm. ACM*, Vol. 36, No. 4, pp.57–71, 1993.

[14] Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, in *Proc. Hypertext '96*, pp.53–65, 1996.

[15] Sanderson, M. and Croft, B.: Deriving Concept Hierarchies from Text, in *Proc. SIGIR '99*, pp.206–213, 1999.

[16] Takata, Y., Nakagawa, K. and Seki, H.: Flexible Category Structure for Supporting WWW Retrieval, in *Proc. ER2000 Conference Workshop on the World Wide Web and Conceptual Modeling (WCM2000)*, Salt Lake City, to appear.

[17] Tou, J. T. and Gonzalez, R. C.: *Pattern Recognition Principles*, pp.89–97, Addison-Wesley, 1974.

[18] Voorhees, E. M. and Harman, D. K.: Evaluation Techniques and Measures, in *The Seventh Text REtrieval Conference (TREC 7)*, pp.A–1, National Institute of Standards and Technology (NIST), 1998.