Intelligent Network News Reader with Visual User Interface

Hitoshi ISAHARA, Kiyotaka UCHIMOTO and Hiromi OZAKU
Communications Research Laboratory
588-2, Iwaoka, Iwaoka-cho, Nishi-ku, Kobe, Hyogo, 651-2401, Japan

Abstract

We are developing an Intelligent Network News Reader which extracts news articles for users. In contrast to ordinary information retrieval and abstract generation, this method utilizes an "information context" to select articles from newsgroups on the Internet and it displays the context visually. A salient feature of this system is that it retrieves articles dynamically, adapting itself to the user's interests, not classifying them beforehand. Since this system measures the semantic distance between articles, it is possible to refer to the necessary information without being constrained within a particular news group. We finished a prototype of the Intelligent Network News Reader in March 1998 and will complete a final practical version in March 2000.

1 Introduction

Extracting necessary information easily from the bulk of information available throughout the world is crucial for people living in this highly computerized society, and therefore, it is necessary to develop systems which can visually present the selected information necessary to assist people in forming new concepts. A great deal of work on this subject has been done by various researchers, e.g., information retrieval from newspaper articles and message understanding in newspaper articles.

It is not sufficient that this kind of expert system simply imitate the real world. Such systems have to create a richer environment with the visual interface than there is now. This means not simply supplying an imitation of the real world, but actively building a virtual world where the density of information is higher than that of the real world for a given use. In other words, we need information screening for each individual user. Therefore, technology which selects and presents the necessary information will be the key to information retrieval in the future.

As an application of this kind of technology to the Internet, we are developing an Intelligent Network News Reader (HISHO: Helpful Information Selection by Hunting On-line) which extracts news articles for users and which visually displays the structure of articles. In contrast to ordinary information retrieval and abstract generation, this method utilizes an "information context" to select articles from newsgroups on the Internet. We finished our prototype of the Intelligent Network News Reader in March 1998 and will complete a final practical version in March 2000.

In this paper, we discuss how to find topic changing articles in the tree structures of news articles, how to extract topic differences from the thread of articles, and how to indicate this information in the display to help users decide which part of the tree structures of articles they will read.

2 Information Gathering from the Network News

Network news has recently become very popular worldwide, and the number of articles generated every day is increasing rapidly. Also the quality of information in these articles varies widely. This makes the percentage of important information lower and lower.

Many people use or want to use Internet news. However, since it is not possible to read all the articles received, it is difficult to find articles on a specific topic and it is difficult to determine from these articles, which are relevant to one's specific interests, especially where the author of the article neglected to use a suitable subject, i.e., title. This situation, i.e., articles without suitable subjects, often occurs and thus, it is not easy to retrieve information utilizing a simple keyword-based method.

Some research on such problems, i.e., on gathering information efficiently, has been done, but most of the research has been limited to generating abstracts or extracting some topics. However, they are immature and still have many problems. No one, yet, has established a way for the user to tell a news reader what he/she requires.

3 Information Retrieval and News Reader

There is much on-going research in information retrieval. In document retrieval, the key technology is the utilization of keywords, titles, and user defined "key words" (Jacobs, 1992). Full text search is now very fast using some programming techniques. TREC (Text Retrieval Conference) by ARPA includes this kind of approach (Harman, 1994).

One of the targets of the summarization and information extraction domains is to plug information into some templates. MUC (Message Understanding Conference) by ARPA is involved in doing this kind of work (ARPA, 1993).

However, these approaches are not suitable for information retrieval from the network news on the Internet. Therefore, there have been many proposals for network news readers. For example, "Galaxy of News" retrieves sets of information related to one another by adopting a stochastic method to produce a hierarchy of keywords and it presents the results of the search visually, i.e., 3-dimensionally (Rennison, 1994). However, users have to manually chose the articles they want to read.

Another program which assists users in selecting articles they should read is the "Personalized Electronic News Editor" (Sheth, 1994). First, a user instructs the agents who are in charge of information retrieval of his/her preferences. Then, they extract keywords, chose articles using the extracted keyword, and recommend the chosen articles to the user. There is also research being done on the summarization of news articles to help people who read the network news (Sato, 1994). Although this is a very useful research domain, when we think of the actual user needs for a network news reader, these needs are not being met. Users generally want to read the whole article relevant to their interests, and they are not satisfied reading abstracts. Therefore, it is necessary to display not only the summary conceived in terms but the whole relevant section of the original articles.

Also, we have to be aware of the following point. There are two types of network news. The first is newswire-like newsgroups, which are similar to the ordinary newspaper and which makes various announcements, such as meetings, job opportunities, and so on. Of course, these newsgroups are very informative, but they do not contain such a large number of articles. The second is newsgroups for discussion among users. This is where people discuss things, the topic for which has been introduced by one of them. Each article in the newswire-like newsgroups is mainly self-contained, therefore, it is possible to retrieve previous articles dealing with the same subject by using simple keyword-based technology.

However, articles in the newsgroups for discussion are neither semantically nor referentially self-contained. The previously mentioned "Personalized Electronic News Editor" and summarization systems are for articles in the newswire-like newsgroups. Our system focuses on assisting the reader of the discussion newsgroups by intelligently screening articles in the network news.

4 Features of Network News

Network news is a good knowledge source and the expectation is that the articles are well organized. The assumption is that related articles should have the same title or be linked together by information in a reference field and that non-related articles should have different titles. However, often this is not true. Recent news reader systems which utilize this kind of information to classify news articles have been misleading. We checked two newsgroups, specifically, fj.life.health and fj.sci.medical. In fj.life.health, we found 525 disjointed parts in 1431 articles over 13 months using their reference fields, and in fj.sci.medical, we found 692 disjointed parts in 1683 articles. For example, in fj.life.health, 209 articles had no relation to other articles explicitly, however, 61 of these had some relation to the others when we checked their content. Also, some parts which involved more than one article were semantically related to the other parts. This indicates that if we use the reference field to find relations between articles, many of these would not be extracted.

The subject field of each article seems informative, however, news writers do not tend to change the subject even if they change the topic of their articles from the former one. We found during our experiment that the subject is not very informative and it is not efficient if a news reader presents all articles with the same subject to users.

Therefore, it is necessary for the Intelligent Network News Reader to have a way of gathering all the relations between articles based on their content. We propose a system which sees the articles in network news as a kind of conversational text, which goes upstream in the flow of topics for articles, considering references, quotations and the relative importance between words and/or sentences, and which extracts and visually displays articles which have user relevant information.

5 Intelligent Network News Reader

We are developing an Intelligent Network News Reader as part of the environment in assisting the growth of human intellectual creativity, focusing on screening technology to raise the density of information. We see this as one application of natural language processing technology as we progress toward a multimedia network society (Nikkei, 1995).

In this paper, we clarify problems which prevent the effective use of information on the network news, and we propose a way of extracting the necessary information by focusing on the consistency of topics in the articles. We also propose a way of displaying the extracted information visually to assist users in reading informative news articles effectively. We also attempt to solve the problems. Our system has the following features: it treats the article in which the user is interested as a key to information retrieval, weighs the relative importance between sentences by using natural language processing technology, and it utilizes heuristics on the features of the network news assuming it to be a kind of conversational text.

Typical usage of this system would be: the user is very busy and cannot keep up with the recent news, he/she gets some free time and takes a look at today's news articles which are extracted from a huge set of unread articles. He/she finds one very interesting article and wants to read all the articles pertaining to that topic, enough to understand whole discussion. So, what should the news reader do to help him/her?

A news reader for busy people needs not only make an abstract of the recent news – since the abstraction process can drop some important information – but to choose the suitable thread to follow on the basis of the content of the news articles. We therefore propose the concept of "information context" defined by the structural distribution of words in the articles. Using this context, the user can follow a suitable thread, even if some articles are lacking a suitable subject.

To make a decision regarding article retrieval or summary generation, it is not enough to give such a system keywords or a title in the subject field of the news. Recently developed news readers classify news items using their subjects, however, since the subject often differs from the contents of the article, many unnecessary articles are extracted by such a simple screening method.

A keyword method can be useful when one knows what information he/she wants, or when one precisely knows the hierarchy of keywords, e.g., a thesaurus. When one is in the process of forming a new concept from his/her basic concept, it is not possible to chose a suitable keyword. The human conception process begins from the basic stage, passes into the thinking stage with the process of extracting related news, and clarifies its target and/or its result. Our intelligent news reader is expected to improve the efficiency of the retrieval of network news, and is also expected to be a tool for assisting some intelligent activities by humans.

Here, the key to retrieval is not the keyword or titles which are decided by the users based on their own intuition, but the relevant article itself. In other words, this system allows information retrieval through the use of ambiguous keys. It is not necessary for the user to enter any concrete keywords or titles of articles. He/she simply needs to point to the article which he/she is interested in. The system will find (almost) all related articles.

We are developing the system in JAVA language which is one of the most popular languages capable of handling visual images on the screen.

The system works as follows:

1. A user finds an article which fits his/her interests.
2. The HISHO system makes a *reference tree* (RT) and sets a *family tree* obeying the user's selected article. It checks the article's relation inside the FT.
3. The system checks the relation between the FT and other RTs.
4. The system displays the relevant articles which fit the user's interests by using a graphic interface.

When the user activates the system, it automatically creates a tree structure of articles in a news spool by using their "References" field, then it refines these tree structures using the "Subject" field. We call these structures reference trees.

The user begins to read articles and finds a news article which fits his interests. The selected article is called the *focus article* (FA). An RT including the FA is called a family tree.

When the user selects the FA, HISHO starts to find the FT and extract features of the FA. Sometimes an RT has a lot of articles. In that case, it is possible that the RT includes several topics. So, HISHO identifies a topic-changing article in the FT.

The feature of the FT is calculated by using the score of terms in articles of the FT. The terms in the FA add the special score. The system calculates features of the RTs including the FT and gathers similar RTs. Related articles are extracted using the feature value from articles not connected by "Subject"

and/or "References" field. It means that HISHO can gather similar RTs even if those belong to different news groups from the original news group.

HISHO gathers the articles which are related to the article selected by a user. After calculating the relevance that is checking the topics, HISHO categorizes some articles in time order, and gives the user the end result by using a graphic interface.

The salient feature of this system is that it retrieves articles dynamically, that is adapting to the user's interests, without classifying them beforehand. Since this system measures the semantic distance between articles, it is possible to refer to the necessary information without being constrained within a particular newsgroup.

6 Visualization of Articles

Our aim is to allow users to clearly grasp the stream of discussion in discussion-type newsgroups when they are shown articles by our system. The summarization of articles is an efficient means of outlining a discussion. However, it is hard to convey the stream of discussion by using only summarization. We are developing a system that can help users to read smoothly by showing them structuralized articles instead of summaries.

We can divide streams of discussion into three kinds of groups by paying attention to the transition of topics. The first is a stream where the topic does not shift from first to last, the second is a stream where the topic shifts halfway, and the third is a stream where several topics are discussed in a certain article and then each topic is discussed respectively. Further, the attitudes of contributors, e.g., proposal, approval, opposition, supplements and so on, are reflected in each stream. In this paper, we discuss a method of presenting articles so that it is easy for users to grasp the stream of discussion, and we defer dealing with the individual attitudes of contributors of the articles towards the discussion.

We assume the following structure which is easy for users to understand:

- Parts where the topic shifts or branches are tagged.
- The difference between topics is represented by keywords.

If we structure articles in this manner, users can catch the changing topic points and the topic branching points and they can easily grasp the difference between topics, enabling them to have a clear grasp of the stream of discussion. A distinctive feature of our method is that when users read a certain article, they can grasp the outline of the articles which follow.

So far, several methods of visualizing archives by using keywords have been proposed. These methods were applied to discussion-type newsgroups and the WWW (Yabe et al., 1997; Arita et al., 1995). In these, articles where the same topic is discussed are located nearer than those where a different topic is discussed, and the topics are visualized by representing keywords. Those methods have an advantage in that users can easily grasp what kinds of topics are being discussed as a whole and which articles those topics are discussed in. However, these methods do not deal with the stream of discussion. Our prototype system can extract the stream of discussion as an RT (Isahara et al., 1997), and it can indicate the article region where the same topic is discussed by identifying the changing topic. That is to say, when a user is interested in a certain article, the system can designate the article region that he should read next. Furthermore, in our method, topic branching can automatically be identified, and the difference between topics discussed in articles can be represented by using keywords, so that articles can be shown to users as those being easy to understand.

6.1 Structuralization of Articles

Figure 1 shows a conceptual image of the structuralized articles. The tree represents a series of discussions. By using the "References" information each article has, we can easily relate the articles in the tree structure. The tags, "TCA" and "TBA", indicate that the topic changes and branches respectively from each tagged article. Our method can correctly identify changing topics and topic branching through evaluating the difference in keywords between articles. The keywords confirming identification are those that represent the difference in topics. Therefore, we extract these keywords and display them as Figure 1 shows. In the articles within the ellipse, the same topic is discussed.

In the following section, we first define topic-changing and topic-branching articles, and in Sections 6.1.2 and 6.1.3 we describe the basic idea of our methods in identifying these articles.

6.1.1 Topic-changing Articles and Topic-branching Articles

Users in discussion-type newsgroups have discussions with each other in the form of articles. Each article contributed to network newsgroups has "References" information, which is a list of related articles and is much like a list of cross-references. By using this information, we can easily relate the articles in a tree structure (reference tree).

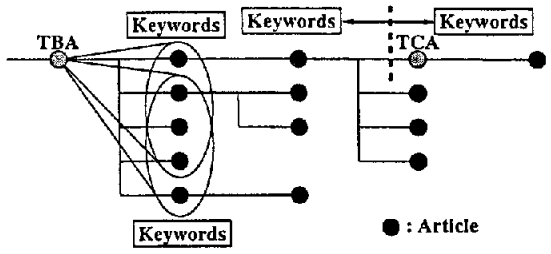


Figure 1: Conceptual Image of Structuralized Articles.

In this tree, a subordinate article is a reply to or comment on the more highly ranked article. The tree branches off at articles which are replied to or commented on by several contributors.

The greater the length of the path and the more branches the tree has, the higher the probability of topic-changing and topic-branching. In this paper, we call an article in which the topic changes a *topic-changing article* and one in which the topic branches a *topic-branching article*.

6.1.2 Method of Identifying a Topic-changing Article

If the topic does not change, in a series of articles, a lot of the same words tend to be used in all the articles. If the topic changes, on the other hand, it is expected that words different from those in previous articles will be used after that turning point. Our system identifies topic-changing articles by looking for the transition in the frequency of words (Uchimoto et al., 1997).

We utilize the following distinctive features to identify topic-changing articles.

Feature 1 At a topic-changing article, the ratio of keywords never seen in the previous articles to all keywords in the article is higher than the ratio in the previous article.

Feature 2 When we split articles into two groups at a topic-changing article, keywords chosen in one group tend to appear frequently in that group, and less frequently in the other group.

We extract keywords which conform to Feature 2 from identified topic-changing articles, and utilize them as the keywords to present to users.

It is impossible for our system to split a sentence into words correctly, since it does not use dictionaries. So instead of using words, our system uses *keywords*. A keyword consists of strings of kanji, e.g., "構文解析", or strings of kanji, katakana, letters, and/or numbers, e.g., "n グラム" or "イタリア料理". We assume that nouns represent features of an article better than verbs, adjectives, and so on do, and that most of the nouns in articles consist of strings of kanji or strings of katakana, letters and/or numbers followed by hiragana. If we cut the strings of hiragana from the text, what is left will be either nouns or arbitrary strings without hiragana. When that remainder consists of only one kanji character and is not followed by a cue word, e.g., a function word "は (ha)", "が (ga)", "を (wo)", "として (to-shite)", we eliminate it, because it will not be a noun but a verb stem or an adjective stem. We regard these hiragana-free strings as keywords.

6.1.3 Method of Identifying a Topic-branching Article

When several topics are discussed in a certain article, as is often the case, each topic is discussed respectively at each branch extending from the article. However, each topic is not always clearly discussed at each branch, but several topics are often discussed at several branches. When this happens, the article region where the same topic is discussed overlaps the others as the left branch of Figure 1 shows. Therefore, in the clustering of articles that branch from a certain article, the articles are allowed to belong to several clusters. Our method compares pairs of articles and classifies articles whose topic is the same into the same cluster. If several clusters are produced by clustering, our method presumes that the topics branch at that branching article. For example, we assume that five articles $A_1 - A_5$ branch from article A_0 as Figure 2 shows. When our method compares pairs of articles and identifies the two articles indicated by the open circle (○) in the Table of Figure 2 as articles where the same topic is discussed, the results of clustering can be presented as shown at the right of Figure 2.

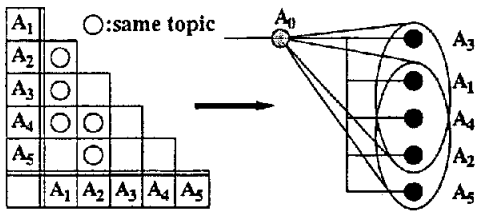


Figure 2: Example of Clustering Articles

We utilized the following distinctive features to determine whether the topic discussed in the articles was the same or not. Two branches where the same topic is discussed tend to quote the identical

parts from the branching original article (**Feature 3**) and to have a lot of common words (**Feature 4**). Concretely, when the proportion of the same quoted part is high or the proportion of the common words between two articles is high, our method determines that the same topic is discussed in those two branches.

Our method weights the keywords in each article according to the positional information and keyword frequencies in articles occurring before and after the article, and it uses keywords whose score is above a given threshold. Concretely, keywords are weighted using the following heuristics:

- Keywords used in sentences which are not quoted from the previous article are more important than those used in the quoted sentences. In particular, keywords used in a sentence next to the quoted sentences are the most important because a contributor tends to write what he wants to say in such a place.
- Keywords also used in articles before and after the article are important because such keywords often represent the central topic discussed in the stream.

Our system detects quoted sentences by investigating the correspondence of sentences between two articles related to each other such as a parent-child relationship (Uchimoto et al., 1998).

6.1.4 Experiment and Evaluation

We constructed RTs from about 10,000 articles in two discussion-type newsgroups, e.g., fj.life.health and fj.living. From these RTs we selected 20 RTs which consisted of about 400 articles with topic-changing articles. We applied our methods after cutting the headers and footers from the articles.

In order to evaluate our methods, we also had the topic-changing and topic-branching articles identified by human subjects. They identified topic-changing articles and topic-branching articles by actually reading the articles. We selected these as target articles, and compared the output of our system with the target articles. The results are listed in Table 1.

Our system could correctly identify 18 topic-branching articles, and nine of these had more than three branches. Our system could correctly identify six of the nine. We used the following criterion for topic-changing articles; When articles the system identifies are the same as the target article, or adjacent to a target article, the system is judged to be correct (Uchimoto et al., 1997).

Table 1: Results.

	Recall	Precision
Topic-branching article (TBA)	18/22 (78%)	18/23 (82%)
Topic-changing article (TCA)	20/35 (57%)	17/18 (94%)

6.2 Actual Example and Discussion

In the experiment in Section 6.1.4, There were 35 target articles for topic-changing articles and 22 for topic-branching ones. Out of these target articles, our system could correctly identify 17 articles and 18 articles respectively. Incidentally, there were three articles where the topic changed and branched.

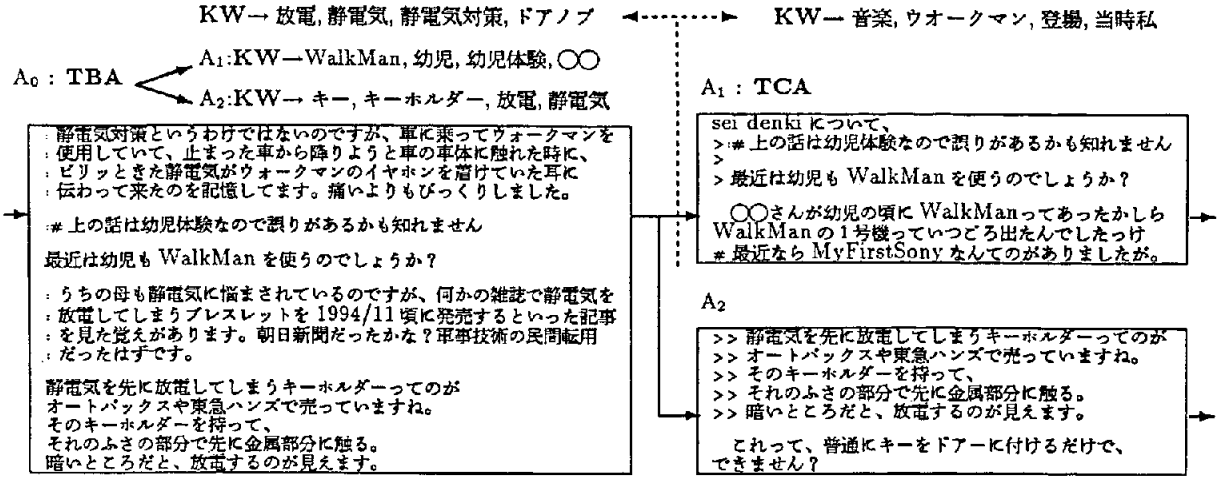
We structuralized actual articles using the output of our system and extracted keywords. Figure 3 shows part of the structuralized articles, and it shows the top four keywords according to their scores at the topic-changing and topic-branching articles. The discussion topic was "静電気(static electricity)" until article A_0. Then, the topic branched and changed in the article. The topic changed to "Walk Man" in article A_1, and in article A_2, the topic about static electricity was discussed throughout.

We want to evaluate our method in the near future using psychological experiments. We need to investigate whether the presented keywords are useful for users to grasp the stream of discussion or not, and we need to estimate the number of keywords our system should present to users.

7 Conclusion and Future Directions

In real articles in network news, writers do not always make suitable references to former articles. They might refer to all of a former article or only talk about a small part. Or, they might talk about a topic which is mentioned in the former articles which is not referred to by their article. It is necessary to develop a powerful and precise retrieval system to solve, among others, the following problems:

1. The addition of a better visual man-machine interface, users can more easily find where the information they need is.
2. The development of heuristics to define the differences in weight of general and domain-specific terms.
3. The improvement in calculation of semantic features of sentences and articles.



(**KW→** : The keywords that the user who reads the article A_0 should refer before reading next articles
TBA : Topic-branching article, **TCA** : Topic-changing article)

Figure 3: Actual Example

We have finished testing our prototype, and we are now studying the results in order to develop a practical system which will be open to the public. We intend to research the problems above to improve a practical model of an Intelligent Network News Reader.

This project is partially funded by the Advanced Information Technology Program(AITP) of the Information-technology Promotion Agency(IPA), Japan.

References

H. Arita, T. Yasui, and S. Tsudaka. 1995. Information strolling through automatically organized information space. *IPSJ-WGNL*, NLC95-17. (in Japanese).

ARPA. 1993. *Proceedings of Fifth Message Understanding Conference (MUC-5).*

M. Sato. et al. 1994. An implementation of automatic digesting on the netnews. In *The 49th Annual Convention IPS Japan.*

D. Harman. 1994. Overview of the second text retrieval conference (trec-2). In *NIST Special Publication 500-215*, pages 1–20.

H. Isahara, H. Ozaku, and K. Uchimoto. 1997. Intelligent network news reader for discussion type news groups. *IPSJ-WGNL*, NL119-3. (in Japanese).

P. S. Jacobs. 1992. *Text-Based Intelligent Systems.* Lawrence Erlbaum Associates.

Nikkei. 1995. Intelligent news retrieval system. Nikkei Computer, 10, 2. (in Japanese).

E. Rennison. 1994. Galaxies of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of UIST 94.*

B. Sheth. 1994. *NEWT: A Learning Approach to Personalized Information Filtering.* MIT Masters Theses.

K. Uchimoto, H. Ozaku, and H. Isahara. 1997. A method for identifying topic-changing articles in discussion-type newsgroups within the intelligent network news reader hisho. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 375–380.

K. Uchimoto, H. Ozaku, and H. Isahara. 1998. Structurization of network news articles using keywords. *Proceedings of The Fourth Annual Meeting of The Assosication for Natural Language Processing.* (in Japanese).

J. Yabe, S. Takahashi, and E. Shibayama. 1997. Visualizing semantic content and relationships. In *Proceedings of The 14th Annual Meeting of Japan Society for Software Science and Technology*, pages 129–132. (in Japanese).