Information Classification and Navigation Based on 5W1H of the Target Information

Takahiro Ikeda and **Akitoshi Okumura** and **Kazunori Muraki**
C&C Media Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216

Abstract

This paper proposes a method by which 5W1H (who, when, where, what, why, how, and predicate) information is used to classify and navigate Japanese-language texts. 5W1H information, extracted from text data, has an access platform with three functions: episodic retrieval, multi-dimensional classification, and overall classification. In a six-month trial, the platform was used by 50 people to access 6400 newspaper articles. The three functions proved to be effective for office documentation work and the precision of extraction was approximately 82%.

1 Introduction

In recent years, we have seen an explosive growth in the volume of information available through online networks and from large capacity storage devices. High-speed and large-scale retrieval techniques have made it possible to receive information through information services such as news clipping and keyword-based retrieval. However, information retrieval is not a purpose in itself, but a means in most cases. In office work, users use retrieval services to create various documents such as proposals and reports.

Conventional retrieval services do not provide users with a good access platform to help them achieve their practical purposes (Sakamoto, 1997; Lesk et al., 1997). They have to repeat retrieval operations and classify the data for themselves.

To overcome this difficulty, this paper proposes a method by which 5W1H (who, when, where, what, why, how, and predicate) information can be used to classify and navigate Japanese-language texts. 5W1H information provides users with easy-to-understand classification axes and retrieval keys because it has a set of fundamental elements needed to describe events.

In this paper, we discuss common information retrieval requirements for office work and describe the three functions that our access platform using 5W1H information provides: episodic retrieval, multi-dimensional classification, and overall classification. We then discuss 5W1H extraction methods, and, finally, we report on the results of a six-month trial in which 50 people, linked to a company intranet, used the platform to access newspaper articles.

2 Retrieval Requirements In an Office

Information retrieval is an extremely important part of office work, and particularly crucial in the creation of office documents. The retrieval requirements in office work can be classified into three types.

Episodic viewpoint: We are often required to make an episode, temporal transition data on a certain event. For example, "Company X succeeded in developing a two-gigabyte memory" makes the user want to investigate what kind of events were announced about Company X's memory before this event. The user has to collect the related events and then arrange them in temporal order to make an episode.

Comparative viewpoint: The comparative viewpoint is familiar to office workers. For example, when the user fills out a purchase request form to buy a product, he has to collect comparative information on price, performance and so on, from several companies. Here, the retrieval is done by changing retrieval viewpoints.

Overall viewpoint: An overall viewpoint is necessary when there is a large amount of classification data. When a user produces a technical analysis report after collecting electronics-related articles from a newspaper over one year, the amount of data is too large to allow global tendencies to be interpreted such as when the events occurred, what kind of companies were involved, and what type of action was required. Here, users have to repeat retrieval and classification by choosing appropriate keywords to condense classification so that it is not too broad-ranging to understand.

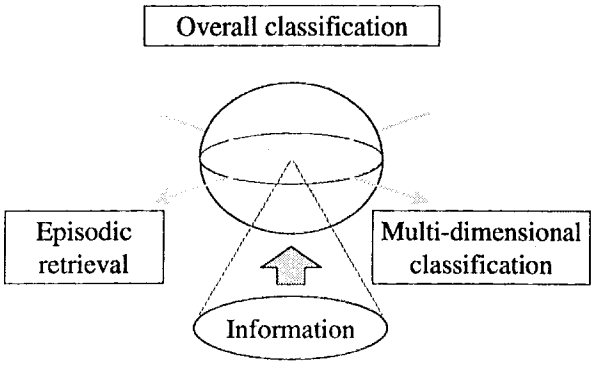

Figure 1: 5W1H classification and navigation

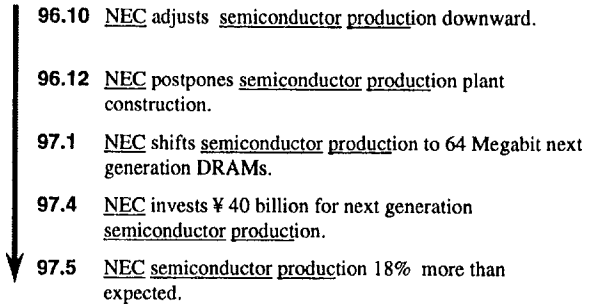

Figure 2: Episodic retrieval example

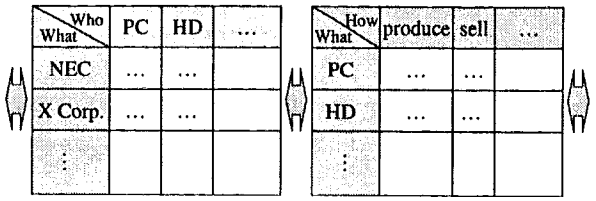

Figure 3: Multi-dimensional classification example

3 5W1H Classification and Navigation

Conventional keyword-based retrieval does not consider logical relationships between keywords. For example, the condition, "NEC & semiconductor & produce" retrieves an article containing "NEC formed a technical alliance with B company, and B company produced semiconductor X." Mine et al. and Satoh et al. reported that this problem leads to retrieval noise and unnecessary results (Mine et al., 1997; Satoh and Muraki, 1993). This problem makes it difficult to meet the requirements of an office because it produces retrieval noise in these three types of operations.

5W1H information is who, when, where, what, why, how, and predicate information extracted from text data through the 5W1H extraction module using language dictionary and sentence analysis techniques. 5W1H extraction modules assign 5W1H indexes to the text data. The indexes are stored in list form of predicates and arguments (when, who, what, why, where, how) (Lesk et al., 1997). The 5W1H index can suppress retrieval noise because the index considers the logical relationships between keywords. For example, the 5W1H index makes it possible to retrieve texts using the retrieval condition "who: NEC & what: semiconductor & predicate: produce." It can filter out the article containing "NEC formed a technical alliance with B company, and B company produced semiconductor X."

Based on 5W1H information, we propose a 5W1H classification and navigation model which can meet office retrieval requirements. The model has three functions: episodic retrieval, multi-dimensional classification, and overall classification (Figure 1).

3.1 Episodic Retrieval

The 5W1H index can easily do episodic retrieval by choosing a set of related events and arranging the events in temporal order. The results are readable by users as a kind of episode. For example, an NEC semiconductor production episode is made by retrieving texts containing "who: NEC & what: semiconductor & predicate: product" indexes and sorting the retrieved texts in temporal order (Figure 2).

The 5W1H index can suppress retrieval noise by conventional keyword-based retrieval such as "NEC & semiconductor & produce." Also, the result is an easily readable series of events which is able to meet episodic viewpoint requirements in office retrieval.

3.2 Multi-dimensional Classification

The 5W1H index has seven-dimensional axes for classification. Texts are classified into categories on the basis of whether they contain a certain combination of 5W1H elements or not. Though 5W1H elements create seven-dimensional space, users are provided with a two-dimensional matrix because this makes it easier for them to understand text distribution. Users can choose a fundamental viewpoint from 5W1H elements to be the vertical axis. The other elements are arranged on the horizontal axis as the left matrix of Figure 3 shows. Classification makes it possible to access data from a user's comparative viewpoints by combining 5W1H elements. For example, the cell specified by NEC and PC shows the number of articles containing NEC as a "who" element and PC as a "what" element.

Users can easily obtain comparable data by switching their fundamental viewpoint from the

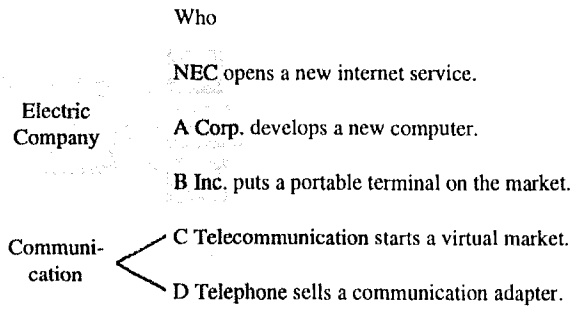

Figure 4: Overall classification example

"who" viewpoint to the "what" viewpoint, for example, as the right matrix of Figure 3 shows. This meets comparative viewpoint requirements in office retrieval.

3.3 Overall Classification

When there are a large number of 5W1H elements, the classification matrix can be packed by using a thesaurus. As 5W1H elements are represented by upper concepts in the thesaurus, the matrix can be condensed. Figure 4 has an example with six "who" elements which are represented by two categories. The matrix provides users with overall classification as well as detailed sub-classification through the selection of appropriate hierarchical levels. This meets overall classification requirements in office retrieval.

4 5W1H Information Extraction

5W1H extraction was done by a case-based shallow parsing (CBSP) model based on the algorithm used in the VENIEX, Japanese information extraction system (Muraki et al., 1993). CBSP is a robust and effective method of analysis which uses lexical information, expression patterns and case-markers in sentences. Figure 5 shows the detail on the algorithm for CBSP.

In this algorithm, input sentences are first segmented into words by Japanese morphological analysis (Japanese sentences have no blanks between words.) Lexical information is linked to each word such as the part-of-speech, root forms and semantic categories.

Next, 5W1H elements are extracted by proper noun extraction, pattern expression matching and case-maker matching.

In the proper noun extraction phase, a 60 000-word proper noun dictionary made it possible to indicate people's names and organization names as "who" elements and place names as "where" elements. For example, NEC and China are respectively extracted as a "who" element and a "where"

```
procedure CBSP;
begin
  Apply morphological analysis to the sentence;
  foreach word in the sentence do begin
    if the word is a people's name or
        an organization name then
      Mark the word as a "who" element and
          push it to the stack;
    else if the word is a place name then
      Mark the word as a "where" element and
          push it to the stack;
    else if the word matches an organization
        name pattern then
      Mark the word as a "who" element and
          push it to the stack;
    else if the word matches a date pattern then
      Mark the word as a "when" element and
          push it to the stack;
    else if the word is a noun then
      if the next word is が or は then
        Mark the word and the kept unspecified
            elements as "who" elements and
            push them to the stack;
      if the next word is を or に then
        Mark the word and the kept unspecified
            elements as "what" elements and
            push them to the stack;
      else
        Keep the word as an unspecified element;
    else if the word is a verb then begin
      Fix the word as the predicate element of
          a 5W1H set;
      repeat
        Pop one marked word from the stack;
        if the 5W1H element
            corresponding to the mark
            of the word is not fixed then
          Fix the word as the 5W1H element
              corresponding to its mark;
        else
          break repeat;
      until stack is empty;
    end
  end
end
```

Figure 5: The algorithm for CBSP

element from the sentence, "NECが中国で半導体を生産。(NEC produces semiconductors in China.)"

In the pattern expression matching phase, the system extracts words matching predefined patterns as "who" and "when" elements. There are several typ-

Table 1: The results of evaluation for "who," "what," and "predicate" elements and overall extracted information.

	"Who" elements			"What" elements			"Predicate" elements			
	Present	Absent	Total	Present	Absent	Total	Present	Absent	Total	Overall
Correct	5423	71	5494	5653	50	5703	6042	5	6047	5270
Error	414	490	904	681	14	695	55	296	351	1128
Total	5837	561	6398	6334	64	6398	6097	301	6398	6396
Precision	92.9%	12.7%	85.9%	89.2%	78.1%	89.1%	99.1%	1.7%	94.5%	82.4%

ical patterns for organization names and people's names, dates, and places (Muraki et al., 1993). For example, nouns followed by 会社 (Co., Inc. Ltd.) and 大学 (Univ.) mean they are organizations and "who" elements. For example, 1998 年 4 月 18 日 (April 18, 1998) can be identified as a date. "When" elements can be recognized by focusing on the pattern for 年 (year), 月 (month), and 日 (day).

For words which are not extracted as 5W1H elements in previous phases, the system decides its 5W1H index by case marker matching. The system checks the relationships between Japanese particles (case markers) and verbs and assigns a 5W1H index to each word according to rules such as が is a marker of a "who" element and を is a marker of a "what" element. In the example "A 社が製品 X を発売 (Company A sells product X.)," company A is identified as a "who" element according to the case marker が if it is not specified as a "who" element by proper noun extraction and pattern expression matching.

5W1H elements followed by a verb (predicate) are fixed as a 5W1H set so that a 5W1H set does not include two elements for the same 5W1H index. A 5W1H element belongs to the same 5W1H set as the nearest predicate after it.

5 Information Access Platform

5W1H information classification and navigation works in the information access platform. The platform disseminates users with newspaper information through the company intranet. The platform structure is shown in Figure 6.

Web robots collect newspaper articles from specified URLs every day. The data is stored in the database, and a 5W1H index data is made for the data. Currently, 6398 news articles are stored in the databases. Some articles are disseminated to users according to their profiles. Users can browse all the data through WWW browsers and use 5W1H classification and navigation functions by typing sentences or specifying regions in the browsing texts.

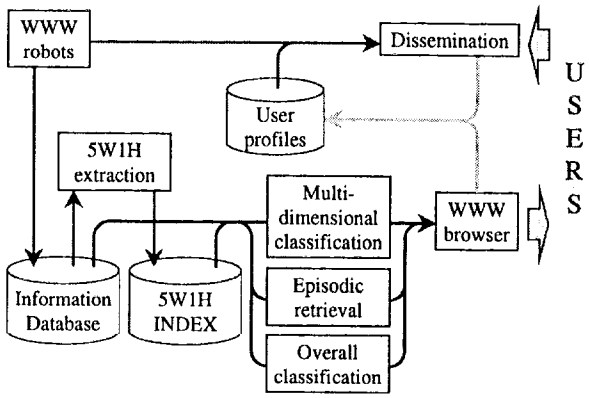

Figure 6: Information access interface structure

5W1H elements are automatically extracted from the typed sentences and specified regions. The extracted 5W1H elements are used as retrieval keys for episodic retrieval, and as axes for multi-dimensional classification and overall classification.

5.1 5W1H Information Extraction

"When," "who," "what," and "predicate" information has been extracted from 6398 electronics industry news articles since August, 1996. We have evaluated extracted information for 6398 news headlines. The headline average length is approximately 12 words. Table 1 shows the result of evaluating "who," "what," and "predicate" information and overall extracted information.

In this table, the results are classified with regard to the presence of corresponding elements in the news headlines. More than 90% of "who," "what," and "predicate" elements can correctly be extracted with our extraction algorithm from headlines having such elements. On the other hand, the algorithm is not highly precise when there is no corresponding element in the article. The errors are caused by picking up other elements despite the absence of the element to be extracted. However, the errors hardly affect applications such as episodic re-

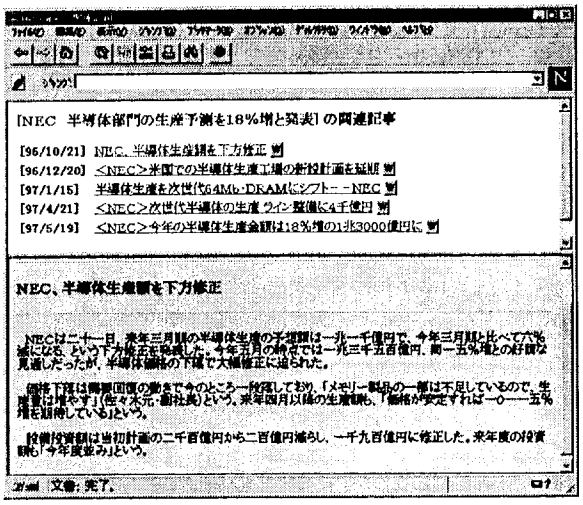

Figure 7: Episodic retrieval example (2)

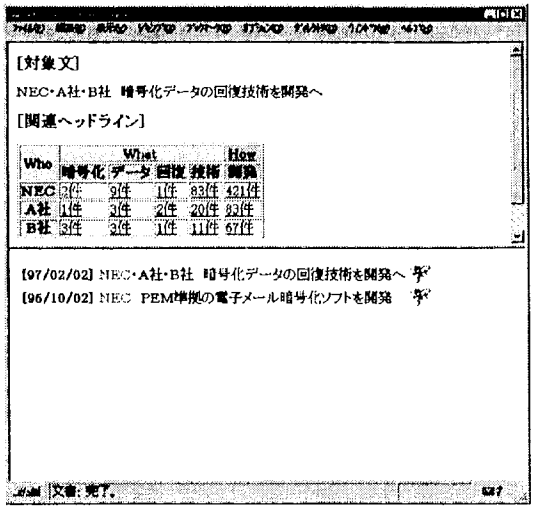

Figure 8: Multi-dimensional classification example (2)

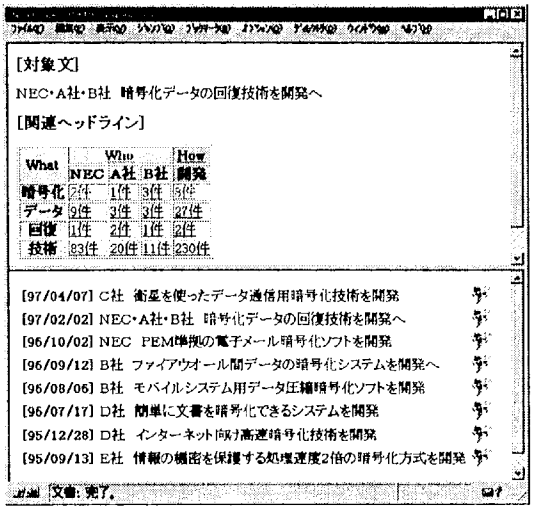

Figure 9: Multi-dimensional classification example (3)

trieval and multi-dimensional classification because they only add unnecessary information and do not remove necessary information.

The precision independent of the presence of the element is from 85% to 95% for each, and the overall precision is 82.4%.

5.1.1 Episodic Retrieval

Figure 7 is an actual screen of Figure 2, which shows an example of episodic retrieval based on headline news saying, "NECが半導体の生産を予定より18%増 (NEC produces 18% more semiconductors than expected.)" The user specifies the region, "NEC が半導体の生産 (NEC produces semiconductors)" on the headline for episodic retrieval. A "who" element NEC, a "what" element 半導体 (semiconductor), and a "predicate" element 生産 (produce) are episodic retrieval keys. The extracted results are NEC's semiconductor production story.

The upper frame of the window lists a set of headlines arranged in temporal order. In each article, NEC is a "who" element, the semiconductor is a "what" element and production is a "predicate" element. By tracing episodic headlines, the user can find that the semiconductor market was not good at the end of 1996 but that it began turning around in 1997. The lower frame shows an article corresponding to the headline in the upper frame. When the user clicks the 96/10/21 headline, the complete article is displayed in the lower frame.

5.1.2 Multi-dimensional Classification

Figures 8 and 9 show multi-dimensional classification results based on the headline, "NEC・A 社・B 社 暗号化データの回復技術を開発へ (NEC, A Co., and B Co. are developing encoded data recovery techniques.)." "Who" elements are "NEC, A Co., and B Co." listed on the vertical axis which is the fundamental axis in the upper frame of Figure 8. "What" elements are "暗号化 (encode), データ (data), 回復 (recovery), and 技術 (technique)." A "predicate" element is a "開発 (develop)." "What" and "predicate" elements are both arranged on the horizontal axis in the upper frame of Figure 8. When clicking a cell for "who": NEC and "what": 暗号化 (encode), users can see the headlines of articles containing the above two keywords in the lower frame of Figure 8.

When clicking on the "What" cell in the upper

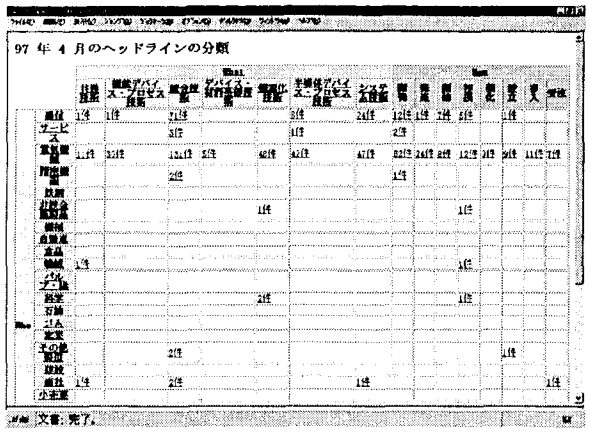

Figure 10: Overall classification for 97/4 news

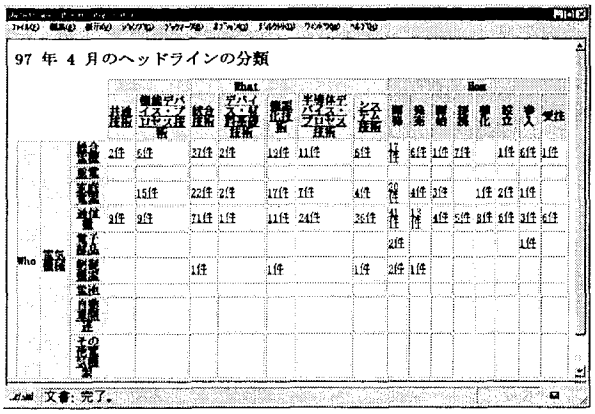

Figure 11: Overall sub-classification for 97/4 news

frame of Figure 8, the user can switch the fundamental axis from "who" to "what" (Figure 9, upper frame). By switching the fundamental axis, the user can easily see classification from different viewpoints. On clicking the cell for "what": 暗号化 (encode) and "predicate": 開発 (develop), the user finds eight headlines (Figure 9, lower frame). The user can then see different company activities such as the 97/04/07 headline; "C社 衛生を使ったデータ通信用暗号化技術を開発 (C Company has developed data transmission encoding technology using a satellite)," shown in the lower frame of Figure 9.

In this way, a user can classify article headlines by switching 5W1H viewpoints.

5.1.3 Overall Classification

Overall classification is condensed by using an organization and a technical thesaurus. The organization thesaurus has three layers and 2800 items, and the technical thesaurus has two layers and 1000 technical terms. "Who" and "what" elements are respectively represented by the upper classes of the organization thesaurus and the technical thesaurus. The upper classes are vertical and horizontal elements in the multi-dimensional classification matrix. "Predicate" elements are categorized by several frequent predicates based on the user's priorities.

Figure 10 shows the results of overall classification for 250 articles disseminated in April, 1997. Here, "who" elements on the vertical axis are represented by industry categories instead of company names, and "what" elements on the horizontal axis are represented by technical fields instead of technical terms. On clicking the second cell from the top of the "who" elements, 電気機械 (electrical and mechanical) in Figure 10, the user can view subcategorized classification on electrical and mechanical industries as indicated in Figure 11. Here, 電気機械 (electrical and mechanical) is expanded to the subcategories; 総合電機 (general electric) 重電 (power electric), 家庭電器 (home electric), 通信機 (communication), and so on.

6 Current Status

The information access platform was exploited during the MIIDAS (Multiple Indexed Information Dissemination and Acquisition Service) project which NEC used internally (Okumura et al., 1997). The DEC Alpha workstation (300 MHz) is a server machine providing 5W1H classification and navigation functions for 50 users through WWW browsers. User interaction occurs through CGI and JAVA programs.

After a six-month trial by 50 users, four areas for improvement become evident.

1) 5W1H extraction: 5W1H extraction precision was approximately 82% for newspaper headlines. The extraction algorithm should be improved so that it can deal with embedded sentences and compound sentences.

Also, dictionaries should be improved in order to be able to deal with different domains such as patent data and academic papers.

2) Episodic retrieval: The interface should be improved so that the user can switch retrieval from episodic to normal retrieval in order to compare retrieval data.

Episodic retrieval is based on the temporal sorting of a set of related events. At present, geographic arrangement is expected to become a branch function for episodic retrieval. It is possible to arrange each event on a map by using 5W1H index data. This would enable users to trace moving events such as the onset of a typhoon or the escape of a criminal.

3) Multi-dimensional classification: Some users need to edit the matrix for themselves on the screen.

Moreover, it is necessary to insert new keywords and delete unnecessary keywords.

7 Related Work

SOM (Self-Organization Map) is an effective automatic classification method for any data represented by vectors (Kohonen, 1990). However, the meaning of each cluster is difficult to understand intuitively. The clusters have no logical meaning because they depend on a keyword set based on the frequency that keywords occur.

Scatter/Gather is clustering information based on user interaction (Hearst and Pederson, 1995; Hearst et al., 1995). Initial cluster sets are based on keyword frequencies.

GALOIS/ULYSSES is a lattice-based classification system and the user can browse information on the lattice produced by the existence of keywords (Carpineto and Romano, 1995).

5W1H classification and navigation is unique in that it is based on keyword functions, not on the existence of keywords.

Lifestream manages e-mail by focusing on temporal viewpoints (Freeman and Fertig, 1995). In this sense, this idea is similar to our episodic retrieval though the purpose and target are different.

Mine et al. and Hyodo and Ikeda reported on the effectiveness of using dependency relations between keywords for retrieval (Mine et al., 1997; Hyodo and Ikeda, 1994).

As the 5W1H index is more informative than simple word dependency, it is possible to create more functions. More informative indexing such as semantic indexing and conceptual indexing can theoretically provide more sophisticated classification. However, this indexing is not always successful for practical use because of semantic analysis difficulties. Consequently 5W1H is the most appropriate indexing method from the practical viewpoint.

8 Conclusion

This paper proposed a method by which 5W1H (who, when, where, what, why, how, and predicate) information is used to classify and navigate Japanese-language texts. 5W1H information, extracted from text data, provides an access platform with three functions: episodic retrieval, multidimensional classification, and overall classification. In a six-month trial, the platform was used by 50 people to access 6400 newspaper articles.

The three functions proved to be effective for office documentation work and the extraction precision was approximately 82%.

We intend to make a more quantitative evaluation by surveying more users about the functions. We also plan to improve the 5W1H extraction algorithm, dictionaries and the user interface.

Acknowledgment

We would like to thank Dr. Satoshi Goto and Dr. Takao Watanabe for their encouragement and continued support throughout this work.

We also appreciate the contribution of Mr. Kenji Satoh, Mr. Takayoshi Ochiai, Mr. Satoshi Shimokawara, and Mr. Masahito Abe to this work.

References

C. Carpineto and G. Romano. 1995. A system for conceptual structuring and hybrid navigation of text database. In *AAAI Fall Symposium on AI Application in Knowledge Navigation and Retrieval*, pages 20-25.

E. Freeman and S. Fertig. 1995. Lifestreams: Organizing your electric life. In *AAAI Fall Symposium on AI Application in Knowledge Navigation and Retrieval*, pages 38–44.

M. A. Hearst and J. O. Pederson. 1995. Revealing collection structure through information access interface. In *Proceedings of IJCAI'95*, pages 2047–2048.

M. A. Hearst, D. R. Karger, and J. O. Pederson. 1995. Scatter/gather as a tool for navigation of retrieval results. In *AAAI Fall Symposium on AI Application in Knowledge Navigation and Retrieval*, pages 65–71.

Y. Hyodo and T. Ikeda. 1994. Text retrieval system used on structure matching. *The Transactions of The Institute of Electronics, Information and Communication Engineers*, J77-D-II(5):1028–1030.

T. Kohonen. 1990. The self-organizing map. In *Proceedings of IEEE*, volume 78, pages 1059–1063.

M. Lesk, D. Cutting, J. Pedersen, T. Noreault, and M. Koll. 1997. Real life information retrieval: commercial search engines. In *Proceedings of SIGIR'97*, page 333, July.

T. Mine, K. Aso, and M. Amamiya. 1997. Japanese document retrieval system on www using dependency relations between words. In *Proceedings of PACLING'97*, pages 290–215, September.

K. Muraki, S. Doi, and S. Ando. 1993. Description of the veniex system as used for muc-r. In *Proceedings of MUC5*, pages 147–159, August.

A. Okumura, T. Ikeda, and K. Muraki. 1997. Selective dissemination of information based on a multiple-ontology. In *Proceedings of IJCAI'97 Ontology Workshop*, pages 138–145, August.

H. Sakamoto. 1997. Natural language processing technology for information. In *JEIDA NLP Workshop*, July.

K. Satoh and K. Muraki. 1993. Penstation for idea processing. In *Proceedings of NLPRS'93*, pages 153–158, December.