Thematically Related Words toward Creative Information Retrieval

Eiko Yamamoto
Graduate School of Engineering,
Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe,
Hyogo, 657-8501, Japan
eiko@mech.kobe-u.ac.jp

Hitoshi Isahara
National Institute of Information and
Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0289, Japan
isahara@nict.go.jp

ABSTRACT

We introduce a mechanism that provides key words which can make human-computer interaction increase in the course of information retrieval, by using natural language processing technology and mathematic measure for calculating degree of inclusion. We show what type of words should be added to the current query, i.e. keywords which previously had been input, in order to make human-computer interaction more creative. We try to extract related word sets from documents by employing case-marking particles derived from syntactic analysis. Then, we verify which kind of related words is more useful as an additional word for retrieval support.

Author Keywords

Natural Language Processing, retrieval support, related words, thematic relation, taxonomical relation.

ACM Classification Keywords

H5.2. INFORMATION INTERFACES AND PRESENTATION (e.g., HCI): User Interfaces – *Natural Language*; H.3.3. INFORMATON STORAGE and RETRIEVAL: Information Search and Retrieval.

INTRODUCTION

Nowadays, we can access huge amount of text data available on the web. The increase of the data quantity causes a paradigm shift for web retrieval. Rhetorically speaking, we can take a walk among the huge text data. The web retrieval supports we need in this novel situation are neither simple query expansion nor our (or someone's) record of previously input keywords, but we need interfaces which interact with people in new ways. What is crucial for such interface is not constructions of interface, i.e. how each part of interface is arranged on the screen, but what information is presented to interact with users.

New ideas pop into one's head when he/she strolls in library, bookstore, or even around town. We need retrieval supports which enable us to expand such creativity. Making computer smarter to automatically extract "correct" retrieval result is one-side way of developing support systems for information retrieval. Seeing the advice provided to a user by computer, how the user achieves next retrieval is one of the most important viewpoints for the future intelligent user interface. We need a technology that enables computer to understand huge text data and make it possible to expand the users' way of thinking.

In this paper, we introduce a mechanism that provides key words which can make human-computer interaction (HCI) during the information retrieval increase, by using natural language processing technology and mathematic measure for calculating degree of inclusion. Concretely, we show what type of words should be added to the current query, i.e. keywords which previously had been input, in order to make HCI more creative.

RELATION BETWEEN WORDS

Many researchers in natural language processing have developed many methodologies for extracting various relations from corpora. Several methods exist for extracting relations such as "is-a" [6], "part-of" [4], causal [3], and entailment [2] relations. Moreover, methods to learn patterns for extracting relations between words have been presented [4, 8]. Such related words can be used to support retrieval in order to lead users to high-quality information. One simple method is to provide additional key words related to the key words users have input. Here we have a question, which is what kinds of relations between the previous key words and the additional word are effective for information retrieval.

As for the relations among words, at least two kinds of relations exist: the taxonomical relation and the thematic

relation [9].[1] The former is a relation representing the physical resemblance among objects, such as, "cow" and "animal," which is typically a semantic relation; the latter is a non-taxonomical relation among objects through a thematic scene, such as "milk" and "cow" as recollected in the scene "milking a cow," which includes causal relation and entailment relation. Taxonomically related words are generally used to query expansion and it is comparatively easy to identify taxonomical relations from linguistic resources such as dictionaries and thesauri. On the other hand, it is difficult to identify thematic relations because they are rarely maintained in linguistic resources.

Most of the previous researches of information retrieval support tended to focus on the improvement of recall by the query expansion. However, our aim is to direct users to the informative information for them by query suggestion. Though some users sometimes do not realize their real intention of retrieval, we would like them to find their hidden needs via interaction in which system shows them suggestive terms.

In this paper, we try to extract related word sets from documents in Japanese by employing case-marking particles derived from syntactic analysis. Then, we compared the results retrieved with words related only taxonomically and those retrieved with words that included a word related non-taxonomically to the other words in order to verify what kind of relation makes human-computer interaction more creative.

WORD SET EXTRACTION METHOD
In order to derive word sets that direct users to obtain information, we applied the method based on the Complementary Similarity Measure (CSM) which can estimate inclusive relations between two vectors [10]. This measure was developed as a means of recognizing degraded machine-printed text [5].

Estimating Inclusive Relation between Words
We first extract word pairs having an inclusive relation of the appearance patterns between the words by calculating the CSM values. An appearance pattern is expressed a kind of co-occurrence relation by an n-dimensional binary feature vector. Therefore, the dimension of each vector corresponds to a co-occurring word, a document, or a sentence. When $V_i = (v_{i1}, ..., v_{in})$ is a vector for word w_i and $V_j = (v_{j1}, ..., v_{jn})$ is a vector for word w_j, $CSM(V_i, V_j)$ is defined by the following formula:

$$CSM(V_i, V_j) = \frac{ad - bc}{\sqrt{(a + c)(b + d)}},$$

$$a = \sum_{k=1}^{n} v_{ik} \cdot v_{jk}, \qquad b = \sum_{k=1}^{n} v_{ik} \cdot (1 - v_{jk}),$$

$$c = \sum_{k=1}^{n} (1 - v_{ik}) \cdot v_{jk}, \quad d = \sum_{k=1}^{n} (1 - v_{ik}) \cdot (1 - v_{jk}).$$

CSM is an asymmetric measure because the denominator is asymmetric. Therefore, $CSM(V_i, V_j)$ usually differs from $CSM(V_j, V_i)$ exchanged between V_i and V_j. For example, when V_i is 1110010111 and V_j is 1000110110, parameters for $CSM(V_i, V_j)$ are a = 4, b = 3, c = 1, and d = 2, and $CSM(V_i, V_j)$ is greater than $CSM(V_j, V_i)$. According to the asymmetric feature, we can estimate whether the appearance pattern of w_i includes the appearance pattern of w_j. If w_i is "animal" and w_j is "tiger," CSM would estimate that "animal" is a hypernym of "tiger."

Extracted word pairs based on the appearance patterns are expressed by a tuple <w_i, w_j> which is a directed set of words. Tuple <w_i, w_j> represents that $CSM(V_i, V_j)$ is greater than $CSM(V_j, V_i)$ when words w_i and w_j have each appearance pattern represented by each binary vector V_i and V_j. We call w_i the "left word" and w_j the "right word."

Constructing Related Word Sets
We next connected such word pairs with CSM values greater than a certain threshold and constructed word sets.

If we adopt simpler mechanism such as the co-occurrence frequency, which extracts only the co-occurrence relations between words, two tuples extracted from different sentences cannot be merged easily. A feature of our method is that because we use the CSM to calculate the degree of inclusion of appearance patterns between all combinations of words in whole collection of texts, we can connect word pairs consistently. That is to say, we can extract not only pairs of related words but also sets of related words. In other words, our CSM-based method is relevant not only to information within a sentence or a document, but also to information from a wider context. That is, once we obtain two tuples <A, B> and <B, C>, even though the tuples have been extracted from different sentences or documents, we can obtain word set {A, B, C} in order.

Suppose we have tuples <A, B>, <B, C>, <Z, B>, <C, D>, <C, E>, and <C, F>, which are word pairs having greater CSM values than the threshold (TH) in the order of their values. For example, let <B, C> be an initial word set {B, C}. We create a word set as follows.

1. We find the tuple with the greatest CSM value among the tuples in which the word at the tail of the current word set − for example, C in {B, C} − is a left word, and connect the right word of the tuple to the tail of the current word set. In this example, word "D" is connected to {B, C} because <C, D> has the greatest

[1] The taxonomical relation which is, for example, provided by WordNet [1] corresponds to "classical" relation by Morris and Hirst [7], and the thematic relation corresponds to "non-classical" relation.

CSM value among the three tuples <C, D>, <C, E>, and <C, F>, making the current word set {B, C, D}.

2. This process is repeated until no tuples with a CSM value greater than TH can be chosen.

3. We find the tuple with the greatest CSM value among the tuples in which the word at the head of the current word set − for example, B in {B, C, D} − is the right word, and connect the left word of the tuple to the head of the current word set. In this example, Word "A" is connected to the head of {B, C, D} because <A, B> has a CSM value greater than that of <Z, B>, making the current word set {A, B, C, D}.

4. This process is repeated until no tuples with a CSM value greater than TH can be chosen.

In this example, we obtained the word set {A, B, C, D} beginning with tuple <B, C> as the initial word set {B, C}. In this way, we construct all word sets by beginning with each tuple, using tuples whose CSM values are greater than TH. Then from the word sets obtained, we remove word sets that are embedded in other word sets.

If we set TH to a low value, it is possible to obtain lengthy word sets. When the TH is too low, the number of tuples that must be considered becomes overwhelming and the reliability of the measurement decreases. Consequently, we experimentally set TH.

Extracting Word Sets with Thematic Relation
Finally, we use a thesaurus to extract word sets with a thematic relation. We remove word sets with taxonomical relations from whole word set we extracted, and get the leftover as word sets with thematic (at least non-taxonomical) relations. The heading words in a thesaurus are categorized to represent a taxonomical relationship. If a word set extracted by the CSM-based method demonstrates a taxonomical relation among the words, the words in the CSM-based word set are classified into one category in the thesaurus; that is, if an extracted word set agrees with the thesaurus, we can conclude that a taxonomical relation exists among the words. Through this approach, we remove those word sets with a taxonomical relation by examining the distribution of words in the categories. The rest of the word sets have a non-taxonomical relation — including a thematic relation — among the words. We then extract those word sets that do not agree with the thesaurus, having identified them as word sets with a thematic relation, that is, thematically related word sets.

LINGUISTIC DATA
We extract word sets by utilizing inclusive relations of the appearance pattern between words based on a modifiee/modifier relationship in documents. The Japanese language has case-marking particles that indicate the semantic relation between two elements in a dependency relation, which is a kind of modifiee/modifier relationship.

For our experiment, we used such particles and extracted the data from the documents we gathered.

First, we parsed sentences with the KNP[2]. From the results, we collected dependency relations matching one of the following five patterns of case-marking particles. With A, B, P, Q, R, and S as nouns (including compound words); V as a verb; and <X> as a case-marking particle with its role in parentheses, the five patterns are as follows:

- A <*no* (of)> B
- P <*wo* (object)> V
- Q <*ga* (subject)> V
- R <*ni* (dative)> V
- S <*ha* (topic)> V

Suppose we have a sentence "*Chloe ha Mike ga Judy ni bara no hanataba wo okutta to kiita* (Chloe heard that Mike had given Judy a rose bouquet.)." From this sentence, we can extract five dependency relations between words as follows:

- *bara* (rose) <*no* (of)> *hanataba* (bouquet)
- *hanataba* (bouquet) <*wo* (object)> *okutta* (had presented)
- *Mike* <*ga* (subject)> *okutta*
- *Judy* <*ni* (dative)> *okutta*
- *Chloe* <*ha* (topic)> *kiita* (heard)

From this set of dependency relations, we compiled the following types of experimental data[3]:

- ***NN-data** based on co-occurrence between nouns.* For each sentence in our document collection, we gathered nouns followed by all five of the case-marking particles we used and nouns proceeded by <*no*>; that is, A, B, P, Q, R, and S. For the above sentence, we can gather *Chloe*,

---

[2] A Japanese parser developed at Kyoto University.

[3] Japanese case-marking particles define not deep semantics but rather surface syntactic relations between words/phrases; therefore, we utilized not semantic meanings between words, but classifications by case-marking particles. Therefore, the method proposed in this paper is applicable to other languages when a syntactic analyzer that classifies relations between elements, such as subject, direct object, and indirect object, exists for the language. For example, from the output of English parser, we can compile necessary linguistic data, such as *Wo-data* using collocations between verb and its direct object, *Ga-data* from collocations between verb and its subject, *Ni-data* from collocations between verb and its indirect object, and *SO-data* from collocations between subject and object of a verb.

Mike, *Judy*, *bara*, and *hanataba*. The number of data items equals the number of sentences in the documents.

- **NV-data** *based on a dependency relation between noun and verb.* We gathered nouns P, Q, R, and S followed by each of the case-marking particles <wo>, <ga>, <ni>, and <ha> for each verb V. We named them *Wo-data* (with 20,234 gathered data items), *Ga-data* (15,924), *Ni-data* (14,215), and *Ha-data* (15,896), respectively. For the verb *okutta* in the above sentence, the *Wo-data* is *hanataba, Ga-data* is *Mike,* and so on. The number of data items equals the number of kinds of verbs.

- **SO-data** *based on a collocation between subject and object.* We gathered subject Q followed by the case-marking particle <ga> that depends on the same verb V as the object P for each object followed by the case-marking particle <wo>. For the above example, we can gather the subject *Mike* for the object *hanataba* because we have the dependency relations *Mike <ga> okutta* and *hanataba <wo> okutta*. The number of data items equals the number of kinds of objects, where each of them co-occurs with a subject in a sentence and depends on same verb as the subject (4,437).

When we represent experimental data with a binary vector, the vector corresponds to the appearance pattern of a noun. Parameters for calculating the CSM-value correspond to the number of dimensions in each situation. Figure 1 shows images of the appearance pattern expressed by the binary vector for each data item. The number of dimensions equals the number of data items for each experimental data. For *NN-data*, each dimension corresponds to a sentence. The element of the vector is 1 if the noun appears in the sentence and 0 if it does not. Similarly, for *NV-data*, each dimension corresponds to a verb. For *SO-data*, we represent
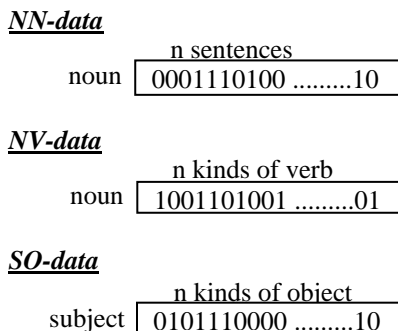
the appearance pattern for each subject with a binary vector whose dimension corresponds to an object.

Therefore, if we calculate the CSM value between Vector A and Vector B, each of the parameters a, b, c, and d used for CSM formula explained above corresponds to the number of each of the following cases:

- The number of dimensions which both Vector A and Vector B have 1 as each element.

- The number of dimensions which Vector A has 1, though Vector B has 0.

- The number of the dimensions which Vector B has 1, though Vector A has 0.

- The number of the dimensions which both Vector A and Vector B have 0.

EXPERIMENT

In our experiment, we used domain-specific documents in Japanese from the medical domain gathered from the Web pages of a medical school. The Japanese documents we used totaled 225,402 sentences (10,144 pages, 37MB).

In applying the CSM-based method, we represented experimental data for medical terms with a binary vector as explained above. We used descriptors in the 2005 Medical Subject Headings (MeSH) thesaurus[4] and translated them into Japanese. The number of terms in Japanese appearing in this experiment is 2,557. We constructed word sets consisting of these medical terms and chose the word sets consisting of three or more terms from them. Figures 2 and 3 show examples of word sets constructed with the CSM-based method. Note that we obtained word sets comprising Japanese medical terms that appear in the Japanese-language medical documents we used. For explanatory purposes, in the following part of this paper we use English terms obtained from the MeSH thesaurus.

NN-data

n sentences

noun | 0001110100 10

NV-data

n kinds of verb

noun | 1001101001 01

SO-data

n kinds of object

subject | 0101110000 10

Figure 1. Appearance patterns of a binary vector for a noun in each type of experimental data

---

data - causation - depression - reduction
    - platelet count - bone marrow examination
neonate - patent ductus arteriosus
    - necrotizing enterocolitis
secretion - gastric acid - gastric mucosa
    - duodenal ulcer
skin - atopic dermatitis - herpes viruses
    - antiviral drugs
fatigue - uterine muscle - pregnancy toxemia
water - oxygen - hydrogen - hydrogen ion
person - nicotiana - smoke - oxygen deficiencies

Figure 2. Examples of word sets extracted from *NN-data*

---

[4] The U.S. National Library of Medicine created, maintains, and provides the Medical Subject Headings (MeSH®) thesaurus.

```
latency period - erythrocyte - hepatic cell
snow - school - gas
variation - death - limb
hospitalist - corneal opacities - triazolam
cross reaction - apoptoses - injuries
research - survey - altered taste - rice
environment - state interest - water - meat - diarrhea
rights - energy generating resources - cordia - education
      - deforestation
```

Figure 3. Examples of word sets extracted from *SO-data*

Type of Data	NN	NV			
		Wo	Ga	Ni	Ha
No. of word sets	594	199	62	37	85
No. of agreed word sets (%)	45 (7.5)	58 (29.1)	14 (22.6)	6 (16.2)	7 (8.2)
No. of disagreed word sets	549	141	48	31	78

Table 1. The numbers of word sets that agreed/disagreed with the MeSH thesaurus.

```
skin - abdomen - cervix - cavitas oris - chest [NN]
cardiovascular disease - coronary artery disease
    - bronchitis - thrombophlebitides - flatulence
    - hyperuricemia - lower back pain
    - ulnar nerve palsies - brain hemorrhage
    - obstructive jaundice [NV(Wo)]
extrasystole - bronchospasm - acute renal failure
    - colitides - diabetic coma - pancreatitides [NV(Ga)]
hand - mouth - ear - finger [NV(Ni)]
snake - praying mantis - scorpion [NV(Ha)]
```

Figure 4. Examples of taxonomically related word sets

Then, to obtain the thematically related word sets from the word sets extracted by the CSM-based method, we use the MeSH thesaurus .The MeSH headings are organized into 15 categories and the MeSH trees are hierarchical arrangements of headings with their associated tree numbers, which include information about the category. Notice that some headings are classified into more than one category.

We examined the distribution of terms in the MeSH categories for each word set and extracted word sets that do not agree with the MeSH thesaurus as word sets with a thematic relation. Table 1 shows the number of word sets that agree and that disagree with the MeSH thesaurus. By way of exception, for example, we obtained the word set "tree - forest - orangutan" from *NN-data*. "Tree" is classified into two categories "Organisms (B)" and "Technology and Food and Beverages (J)"; "forest" is classified into "J" and "orangutan" is classified into "B." In this case, we consider that a relation exists between "forest" and "orangutan" via "tree," and we treat this word set as being distributed in one category.

In Table 1 we found that, for *NN-data* and *NV-data*, the ratio of CSM-based word sets that agreed with MeSH thesaurus was between 7.5% and 29.1%. Of the CSM-based word sets, Wo-data provided the highest agreement ratio.

The apparent reason for this is that the object case represented by the case-marking particle *<wo>* restricts nouns more stringently than do the others. Also, comparing the results of *NN-data* and *NV-data*, we found that the word sets extracted from *NV-data* agreed with the MeSH thesaurus to a greater degree than did those extracted from *NN-data*. This suggests that we obtained more word sets having taxonomical relations among words from *NV-data* than from *NN-data*.

SO-data is based on a collocation between subject and object; that is, the word sets obtained comprise subjects followed by the case-marking particle *<ga>* that depend on the same verb as the object for each object followed by the case-marking particle *<wo>*. For example, when we have "*ningen* (person) *<ga> hon* (book) *<wo> yomu* (read)," which means, "a person reads a book," and "*nezumi* (mouse) *<ga> hon* (book) *<wo> kajiru* (gnaw)," which means, "a mouse gnaws a book," we estimate the relation between the words *ningen* and *nezumi* with CSM. Therefore, we can surmise that the information we obtain from this data will not agree with a general thesaurus because we do not limit the verbs that subjects and objects depend on. Actually, the word sets we obtained from *SO-data* agreed little with the MeSH thesaurus.

Figure 4 shows examples of taxonomically related word sets, which agree with the MeSH thesaurus, that is, in which all the composing terms in a word set are classified into one category. The symbol in brackets represents the type of data from which each word set was obtained. As the result, we obtained the rest as word sets with a thematic relation, that is, thematically related word sets, which are 847 word sets.

VERIFICATION
In verifying the capability of our word sets to retrieve Web pages, we examined whether our word sets could help limit the search results to more informative Web pages with Google as a search engine. To do this, in our obtained word sets with a thematic relation, we used 294 word sets in which one of the terms is classified into one category and the rest are classified into another category. Figure 5 shows examples of the word sets. The terms with underline indicate ones in a different category.

ovary - spleen - <u>palpation</u> [*NN*]
variation - cross reactions - outbreaks - <u>secretion</u>
                                                    [*NV(Wo)*]
bleeding - pyrexia - hematuria - <u>consciousness disorder</u>
  - vertigo - high blood pressure [*NV(Ga)*]
<u>space flight</u> - insemination - immunity [*NV(Ni)*]
cough - <u>fetus</u>
  - bronchiolitis obliterans organizing pneumonia
                                                    [*NV(Ha)*]

Figure 5. Examples of word sets used to verify

We used the terms that composed such word sets as the key words to input into the search engine and retrieved Web pages. We created three types of search terms from a word set. Suppose the word set is $\{X_1, …, X_n, Y\}$, where X_i is classified into one category and Y is classified into another. The first type uses all terms except the one classified into a category different from the others: $\{X_1, …, X_n\}$, removing Y. The second type uses all terms except the one in the same category as the rest: $\{X_1, …, X_{k-1}, X_{k+1}, …, X_n\}$ removing X_k and Y. In our verification, we removed the term X_k with the highest or lowest frequency among X_i. The third type uses terms in Type 2 and Y, i.e., terms in another category: $\{X_1, …, X_{k-1}, X_{k+1}, …, X_n, Y\}$. When we consider Type 2 as base key words, Type 1 is a set of key words with the addition of one term having the highest or lowest frequency among the terms in the same category; i.e., the additional term X_k has a feature related to frequency and is taxonomically related to other terms. Type 3 is a set of key words with the addition of one term in a category different from those of the other component terms; i.e., the additional term Y seems to be thematically related to other terms.

The retrieval results are shown in Figures 6 and 7 including the results for each the highest frequency and the lowest frequency. The horizontal axis is the number of pages retrieved with Type 2 and the vertical axis is the number of pages retrieved with Type 1 or Type 3 that a certain term X_k or Y is added to Type 2. The circles show the results with Type 1 and the crosses show the results with Type 3. The diagonal line in the graph shows that adding one term to Type 2 does not affect the number of Web pages retrieved.

As shown in Figure 6, most crosses fall further below the line. This graph indicates that adding a search term related non-taxonomically tends to make a bigger difference than adding a term related taxonomically and with high frequency. This means that adding a term related non-taxonomically to key words is crucial to retrieving informative pages, i.e., such terms are informative terms themselves.Table 2 shows the number of cases in which term in different category decreases the number of hit pages more than high frequency term. Then, we found that most of the additional terms with high frequency contributed less than additional terms related non-taxonomically to
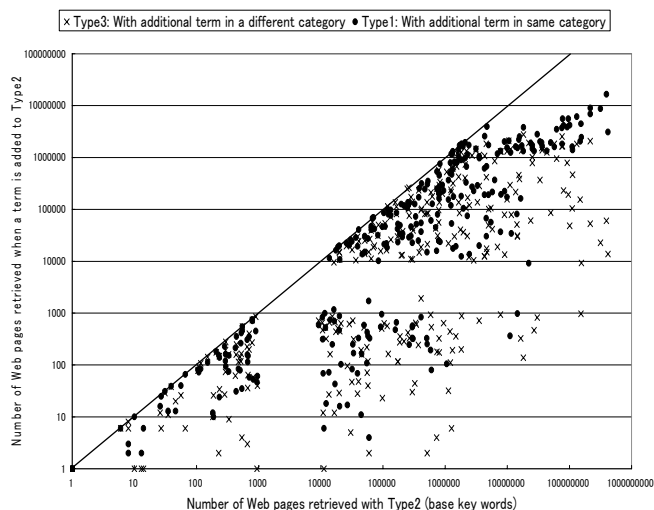

Figure 6. Fluctuation of number of Web pages retrieved by adding the high frequency term in same category (Type 1) and the term in a different category (Type 3)

Type of Data	NN	NV			
		Wo	Ga	Ni	Ha
Number of word sets for verification	175	43	23	13	26
Number of cases in which Type 3 defeated Type 1 in retrieval	108	37	15	12	18

Table 2. Number of cases in which the term in different category decreases the number of Web pages retrieved more than the high frequency term.
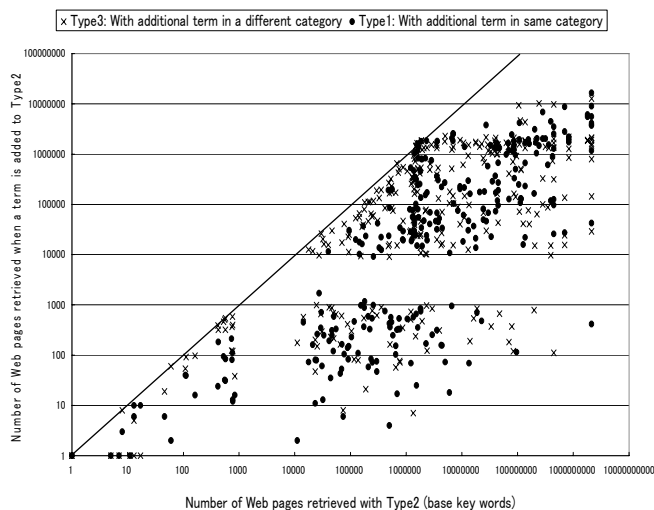

Figure 7. Fluctuation of number of Web pages retrieved by adding the low frequency term in same category (Type 1) and the term in a different category (Type 3)

Type of Data	NN	NV			
		Wo	Ga	Ni	Ha
Number of word sets for verification	175	43	23	13	26
Number of cases in which Type 3 defeated Type 1 in retrieval	61	18	7	6	13

Table 3. Number of cases in which the term in different category decreases the number of Web pages retrieved more than the low frequency term.

decreasing the number of Web pages retrieved. This means that, in comparison to the high frequency terms, which might not be so informative in themselves, the terms in the other category — related non-taxonomically — are effective for retrieving useful Web pages.

Constantly, in Figure 7, most circles fall further below the line. This indicates that adding a term related taxonomically and with low frequency tends to make a bigger difference than does adding a term with high frequency. Certainly, additional terms with low frequency would be informative terms, even though they are related taxonomically, because they may be rare terms on the Internet. Thus, the taxonomically related terms with low frequencies are quantitatively effective for information retrieval as the non-taxonomically related terms.

Table 3 shows the number of cases in which term in different category decreases the number of hit pages more than low frequency term. In comparing these numbers, we also found that the additional term with low frequency helped to reduce the number of Web pages retrieved, making no effort to determine the kind of relation the term had with the other terms. Thus, the terms with low frequencies are quantitatively effective when used for retrieval.

However, if we consider contents of the results retrieved with Type 1 and Type 3, it is clear that big differences exist between them. For example, consider "latency period - erythrocyte - hepatic cell" obtained from *SO-data* in Figure 3. "Latency period" is classified into a category different from the other terms and "hepatic cell" has the lowest frequency in this word set. When we used all the three terms, we obtained pages related to "malaria" at the top of the results and the title of the top page was "What is malaria?" in Japanese. With "latency period" and "erythrocyte," we again obtained the same page at the top, although it was not at the top when we used "erythrocyte" and "hepatic cell" which have a taxonomical relation.

As we showed above, the terms with thematic relations with other search terms are effective at directing users to informative pages. Quantitatively, terms with a high frequency are not effective at reducing the number of pages retrieved; qualitatively, low frequency terms may not effective to direct users to informative pages.

CONCLUSION

We introduced a mechanism that provides key words which can make human-computer interaction (HCI) increase, by using natural language processing technology and mathematic measure for calculating degree of inclusion. We showed what type of words should be added to the current query, i.e. keywords which previously had been input, in order to make HCI more creative.

We extracted related word sets from documents by employing case-marking particles derived from syntactic analysis. Then, we verified which kind of related word is more useful as an additional word for retrieval support. That is, we found the additional term which is thematically related to other terms is effective at retrieving informative pages by comparing the results retrieved with words related only taxonomically and those retrieved with words that included a word related non-taxonomically to the other words. This suggests that words with a thematic relation can be useful to make the HCI more active.

As for the future directions of this work, one of most crucial issues is evaluation. We will evaluate the effectiveness of our method from human-centered viewpoints, possibly by human judgement.

In the future, we can understand the contents of huge text data with higher natural language processing technology and develop a system which makes it possible to expand the users' ways of thinking.

REFERENCES
1. Fellbaum, C. *WordNet: An electronic lexical database*. Cambridge, Mass.: The MIT Press, (1998).
2. Geffet, M. and Dagan, I. The distribution inclusion hypotheses and lexical entailment. In *Proc. ACL 2005*, (2005), 107-114.
3. Girju, R. Automatic detection of causal relations for question answering. In *Proc. ACL Workshop on Multilingual summarization and question answering*, (2003), 76-114.
4. Girju, R., Badulescu, A., and Moldovan, D. Automatic discovery of part-whole relations. *Computational Linguistics, 32(1)*, (2006), 83-135.
5. Hagita, N. and Sawaki, M. Robust recognition of degraded machine-printed characters using complementary similarity measure and error-correction learning. In *Proc. SPIE – The International Society for Optical Engineering, 2442*, (1995), 236-244.
6. Hearst, M. A. Automatic acquisition of hyponyms from large text corpora, In *Proc. Coling 92*, (1992), 539-545.
7. Morris, J. and Hirst, G. Non-classical lexical semantic relations. Workshop on Computational Lexical

Semantics, In *Proc. Human Language Technology Conference of the NAACL*, (2004).

8. Pantel, P. and Pennacchiotti, M. Espresso: Leveraging generic patterns for automatically harvesting semantic relations In *Proceedings of ACL 2006*, (2006), 113–120.

9. Wisniewski, E. J. and Bassok, M. What makes a man similar to a tie? *Cognitive Psychology, 39*, (1999), 208-238.

10. Yamamoto, E., Kanzaki, K., and Isahara, H. Extraction of hierarchies based on inclusion of co-occurring words with frequency information. In *Proc. IJCAI2005*, (2005), 1166-1172.