5-2008

Extraction of Word Set for Increasing Human-Computer Interaction in Information Retrieval

Eiko Yamamoto
Kobe University, eiko@mech.kobe-u.ac.jp

Hitoshi Isahara
National Institute of Information and Communications Technology, isahara@nict.go.jp

43F. Extraction of Word Set for Increasing Human-Computer Interaction in Information Retrieval

Eiko Yamamoto
Kobe University
eiko@mech.kobe-u.ac.jp

Hitoshi Isahara
National Institute of Information and Communications Technology
isahara@nict.go.jp

Abstract

We present a mechanism that provides word sets which can make human-computer interaction more active in the course of information retrieval, with natural language processing technology and a 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 with taxonomical and non-taxonomical relations 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 and makes human-computer interaction more fruitful.

Keywords

Retrieval Support, Related Words, Thematic Relation, Taxonomical Relation.

1. Introduction

These days, 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 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 are crucial for such interface are 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 present a mechanism that provides keywords which can make human-computer interaction (HCI) during the information retrieval more active, with 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.

2. 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" (Hearst, 1992), "part-of" (Girju, 2006), causal (Girju, 2003), and entailment (Geffet & Dagan, 2005). 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 keywords related to the keywords users have input. Here we have a question, which is what kinds of relations between the previous keywords 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 (Wisniewski & Bassok, 1999).[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.

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.

3. 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 (Yamamoto et al. 2005). This measure was developed as a means of recognizing degraded machine-printed text (Hagita & Sawaki, 1995).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 as an n-dimensional binary feature vector. 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 as follows:

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

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

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

[1] The taxonomical relation which is, for example, provided by WordNet (Fellbaum, 1998) corresponds to "classical" relation by Morris and Hirst (2004), and the thematic relation corresponds to "non-classical" relation.

CSM is an asymmetric measure. Therefore, $CSM(V_i, V_j)$ usually differs from $CSM(V_j, V_i)$ exchanged between V_i and V_j. According to the asymmetric feature, we can estimate whether the appearance pattern of w_i includes the appearance pattern of w_j.

Extracted word pairs are expressed by a tuple $<w_i, w_j>$, where $CSM(V_i, V_j)$ 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. Then, we connected word pairs with CSM values greater than a certain threshold and constructed word sets. A feature of the CSM-based method is that it can extract not only pairs of related words but also sets of related words because it connects their word pairs consistently.

We connect word pairs with CSM values greater than a certain threshold and constructed word sets. 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 in the order of their values. For example, let <B, C> be an initial word set {B, C}. First, we find the tuple with the greatest CSM value among the tuples in which the word C at the tail of the current word set is the left word, and connect the right word behind C. In this example, word "D" in <C, D> is connected to {B, C}, making the current word set {B, C, D}. This process is repeated until no tuples can be chosen. Next, we find the tuple with the greatest CSM value among the tuples in which the word B at the head of the current word set is the right word, and connect the left word before B. This process is repeated until no tuples can be chosen. In this example, we obtain the word set {A, B, C, D}.

Finally, by using a thesaurus, we identify the word sets which all words taxonomically relate, that are, which agree with the thesaurus. As the rest of the word sets have a non-taxonomical relation among the words, we identify them as word sets with a thematic relation.

4. Linguistic Data

We extract word sets by utilizing inclusive relations of the appearance pattern between words based on a modifiee/modifier relationship in Japanese 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 A <*no* (of)> B, P <*wo* (object)> V, Q <*ga* (subject)> V, R <*ni* (dative)> V, and 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; *bara* (rose) <*no* (of)> *hanataba* (bouquet), *hanataba* <*wo* (object)> *okutta* (had presented), *Mike* <*ga* (subject)> *okutta*, *Judy* <*ni* (dative)> *okutta*, and *Chloe* <*ha* (topic)> *kiita* (heard).

[2] A Japanese parser developed at Kyoto University.

From this set of dependency relations, we compiled the following types of experimental data.

- ***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*, *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*, *Ga-data*, *Ni-data*, and *Ha-data*, 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.

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.

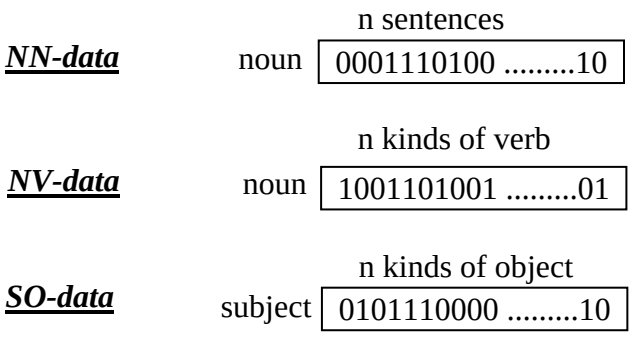

Figure 1: Appearance patterns of a binary vector for a noun in each type of experimental data

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 2: Examples of taxonomically related word sets

5. Experiment

In our experiment, we used domain-specific documents in Japanese from the medical domain gathered from the Web pages of a medical school. We used the Japanese documents (10,144 pages, 225,402 sentences) and compiled *NN-data*(with 225,402 gathered data items), *Wo-data*(20,234), *Ga-data*(15,924), *Ni-data*(14,215), *Ha-data*(15,896), and *SO-data*(4,437). We translated descriptors in the 2005 Medical Subject Headings (MeSH) thesaurus 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. We then identified the word sets which all composing words taxonomically related, by using the MeSH thesaurus and obtained the rest as word sets with a thematic relation, which are 847 word sets. Figure 2 shows examples in which all the terms in a word set are classified into one category, that is, taxonomically related word sets. The symbol in brackets represents the type of data from which each word set was obtained.

6. 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 3 shows examples of the word sets. The terms with underline indicate ones in a different category.

We used the terms that composed such word sets as the keywords 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 $\{X_1, \ldots, X_n, Y\}$, 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: $\{X_1, \ldots, X_n\}$, removing Y. The second type uses all terms except the one in the same category as the rest: $\{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n\}$ 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: $\{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n, Y\}$. When we consider

Type 2 as base keywords, Type 1 is a set of keywords with the addition of one term having the highest or lowest frequency among the terms in the same category; i.e., the additional term has a feature related to frequency and is taxonomically related to other terms. Type 3 is a set of keywords with the addition of one term in a category different from those of the other component terms; i.e., the additional term seems to be thematically related to other terms.

ovary - spleen - palpation [NN]

variation - cross reactions - outbreaks - secretion [NV(Wo)]

bleeding - pyrexia - hematuria - consciousness disorder - vertigo - high blood pressure [NV(Ga)]

space flight - insemination - immunity [NV(Ni)]

cough - fetus - bronchiolitis obliterans organizing pneumonia [NV(Ha)]

Figure 3: Examples of word sets used to verify

The retrieval results are shown in Figures 4 and 5 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 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 4, 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 keywords is crucial to retrieving informative pages, i.e., such terms are informative terms themselves. Constantly, in Figure 5, 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. 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, among the word lists obtained in our experiments, “latency period - erythrocyte - hepatic cell,” where “erythrocyte” is in a different category, retrieved pages related to “malaria.” If we input “latency period” and “hepatic cell,” Google retrieves too much information related to “hepatic trouble” in the 385 pages of retrieval results. By using this word list extracted with our method, that is, adding “erythrocyte,” users can retrieve only the relevant and necessary pages, similar to the approach of a professional who knows that patients experience hepatic trouble during the latency period for malaria. The number of retrieval results is 181 pages. In the results, the page related to malaria is the seventh, though it does not appear within the top 100 of the results without “erythrocyte.”

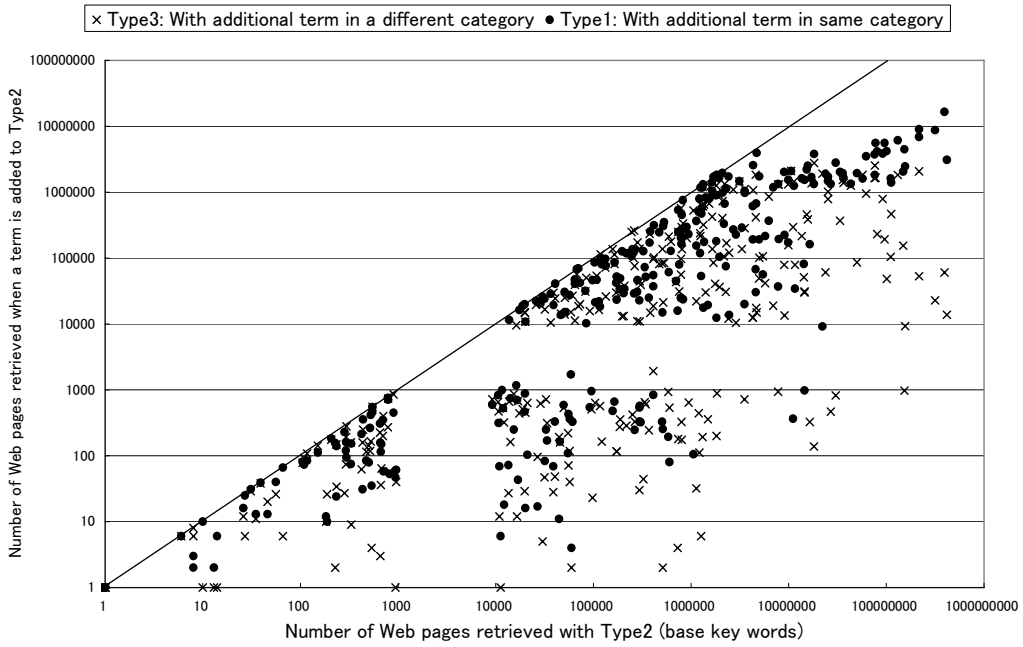

Figure 4: Fluctuation of number of Web pages retrieved by adding the high frequency term in same category (Type 1) and the term in a different category (Type 3)

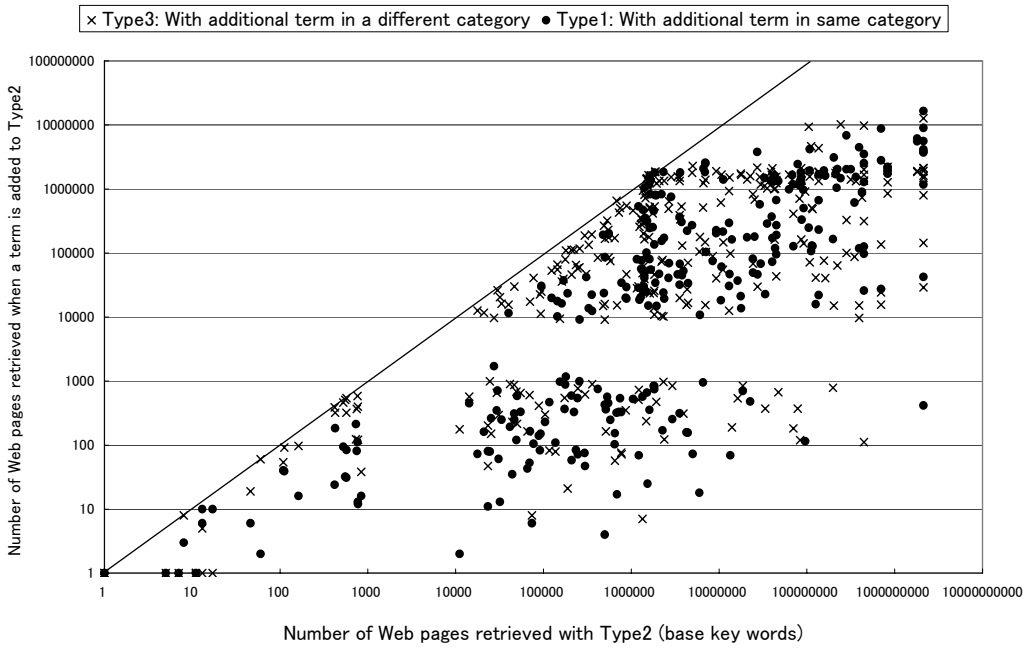

Figure 5: Fluctuation of number of Web pages retrieved by adding the low frequency term in same category (Type 1) and the term in a different category (Type 3)

7. Remarks

Note that in this paper, we use Japanese documents as the input. Because we are using only syntactic information output by the Japanese parser, our mechanism is also applicable to other languages such as English. Japanese case-marking particles define not deep semantics but rather surface syntactic relations between words/phrases; namely, 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.

Also note that in this paper, we use medical domain documents as the input. We have experience of analyzing texts in aviation domain. The application of our method to other languages and other domains are described in (Yamamoto et al. 2008).

8. Conclusions

We presented a mechanism that provides keywords which can make human-computer interaction (HCI) more active, with natural language processing technology and a 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 and support information retrieval.

We extracted the 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 in order to support procedure of information retrieval.

As for the future directions of this work, one of most crucial issues is evaluation. Though precision and recall by human subjects are popular measure of evaluation of information retrieval technology, human subjects usually have no intention to retrieve information and can not make decision based on their own needs and interests. We will evaluate the effectiveness of our method from human-centered viewpoints.

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

Fellbaum, C., “WordNet: An electronic lexical database”, Cambridge, Mass.: The MIT Press, 1998.

Geffet, M. and Dagan, I., “The distribution inclusion hypotheses and lexical entailment”, Proceedings of ACL 2005, 107-114, 2005.

Girju, R., “Automatic detection of causal relations for question answering”, Proceedings of ACL Workshop on Multilingual Summarization and Question Answering, 76-114, 2003.

Girju, R., Badulescu, A., and Moldovan, D. “Automatic discovery of part-whole relations”, Computational Linguistics, 32(1): 83-135, 2006.

Hagita, N. and Sawaki, M., “Robust recognition of degraded machine-printed characters using complementary similarity measure and error-correction learning”, Proceedings of SPIE – The International Society for Optical Engineering, 2442: 236-244, 1995.

Hearst, M. A., “Automatic acquisition of hyponyms from large text corpora”, Proceedings of Coling 92, 539-545, 1992.

Morris, J. and Hirst, G., “Non-classical lexical semantic relations”, Workshop on Computational Lexical Semantics, Human Language Technology Conference of the NAACL, 2004.

Wisniewski, E. J. and Bassok, M., “What makes a man similar to a tie?”, Cognitive Psychology, 39: 208-238, 1999.

Yamamoto, E., Kanzaki, K., and Isahara, H., “Extraction of hierarchies based on inclusion of co-occurring words with frequency information”, Proceedings of IJCAI2005, 1166-1172, 2005.

Yamamoto, E., Isahara, H., Terada, A., and Abe, Y., “Extraction of informative expressions from domain-specific documents”, Proceedings of LREC2008, 2008 (will appear).