A Method of Information Retrieval using Fuzzy Database

- The Similarity of Texts using the Fuzzy Graph -

Shuichirou Arai Akira Satoh Kensei Tsuchida

Toyo University

Kujirai-nakanodai Kawagoe-shi Saitama Pref.350-8585

Dep. of Information and Computer Sciences,

Graduate School of Engineering, Toyo University

email: arai@sa.cs.toyo.ac.jp

Abstract — There are many sites with databases containing contents such as questions and answers (Qs & As) concerning treatments, problems and repairs for personal computers. These sites play an important role for many novice users of personal computers. In order to use the database, it is important to be able to retrieve relevant material for any given question quickly, easily, and precisely. We propose a new method to achieve this goal more effectively using a fuzzy graph model and its partition tree.

In this paper, we describe the concept of the similarity of texts and fuzzy graph model, its verification, and a case study.

1. INTRODUCTION

There are many internet sites, such as question and answer forums and consultation offices, that used a bulletin board system, but few sites where information is organized.

Therefore, users are obliged to compare their questions with all the texts in database.

The cluster model is effective for its completeness and execution time of information retrieval. The cluster model has the potential for decreasing the possibility of discrepancy between the user's query and the result of the retrieval, by classifying similar texts.

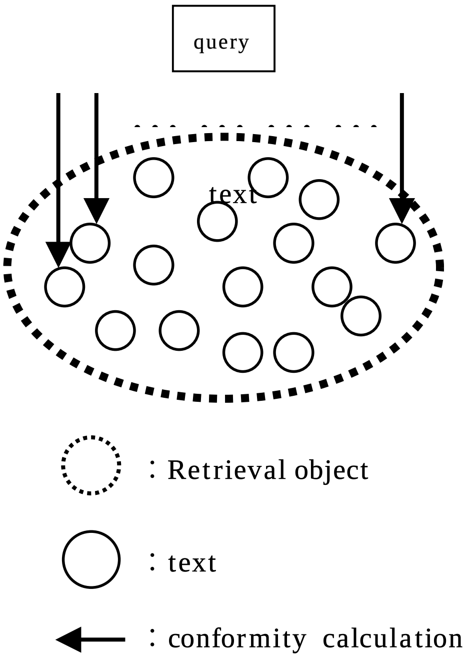

Figure 1. Normal model

It is different from other retrieval models in that classification by similarity of texts about a given retrieval object is necessary before users search the text.

Because queries are usually a word set that users select, it is necessary that the general retrieval method

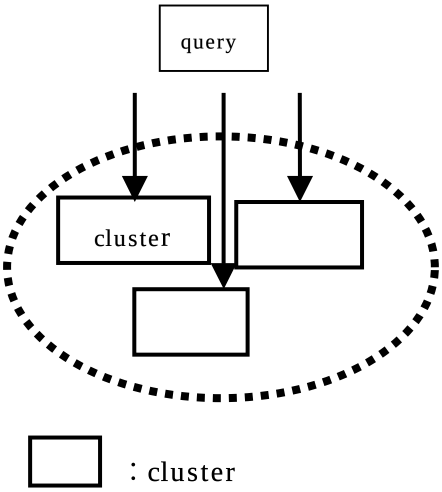

Figure 2. Cluster model

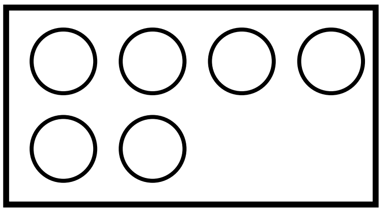

Figure 3. Cluster

convert a text into a word set to match the form of the query. However the co-occurrence relationship between words may be lost to convert a sentence into a word set. There are, of course, retrieval models using co-occurrence relationship between words, as such as similarities between words, but there cases the co-occurrence relationship between words is lost. Because clustering needs only the similarity of texts, it is not necessary to match a text with the form of a query, convert it into a word set and thus lose co-occurrence relationship of the words.

Therefore, we propose a retrieval model that includes the characteristics of a text by using a fuzzy graph that includes co-occurrence relationship between words as arcs, rather than converting a text into a word set.

Many researchers have studied the co-occurrence relationship for the retrieval of texts. Although the co-occurrence relationship between words is usually used according to the similarity of words, we use the co-occurrence relationship to include characteristic of text as the fuzzy membership function in the fuzzy graph.

2. CONVERTING TEXT INTO GRAPH

We describe text analyzing using the holistic approach. Specifically, a word is represented as a node on the graph, and the co-occurrence relationship between words is converted into an arc. The value of arcs is defined by the fuzzy membership function.

We define the model used for classifying the text as the fuzzy graph model.

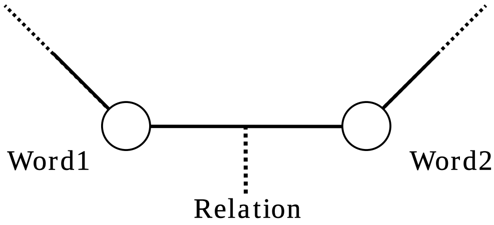

Figure 4. Graph representation

Converting a word set into the fuzzy graph model

Initially, words select, to convert text into the fuzzy graph model using morpheme analysis. In the next phase, we replace words with nodes, and decide the value of each node related. The decision as to the value of each node related requires certain assumptions. In fact, we introduce two postulates concerning distributional hypothesis made by Harris (1968): “ the meaning of entities, and the meaning of grammatical relationships among them, is related to the restriction of combinations of these entities relative to other entities. ”

Text consists of the units of paragraph, sentence and word. Each unit makes one meaning as a unity of the sub constitutive unit. The meaning of each subunit to constitute a unit is connected (Postulates 1). And importance of the meaning of the same level of constitution unit is constant regardless of the number of constitution units (Postulates 2).

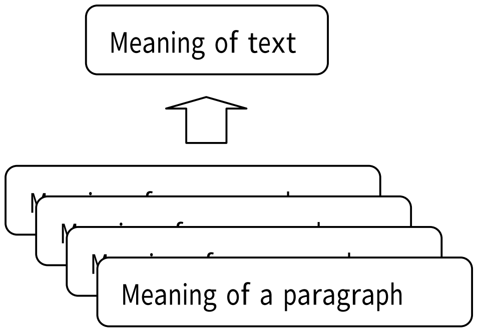

Figure 5. Meaning structure between text and paragraph

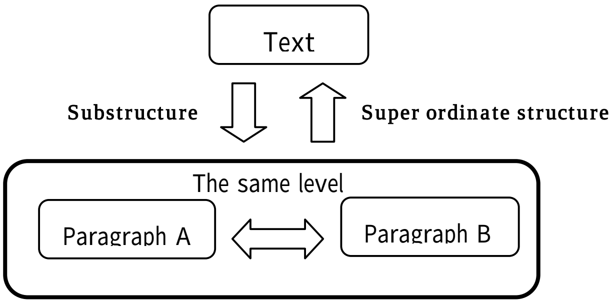

Figure 6. Relation of each unit

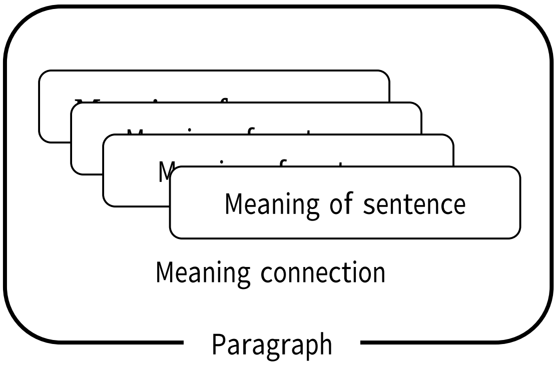

Figure 7. Postulate 1

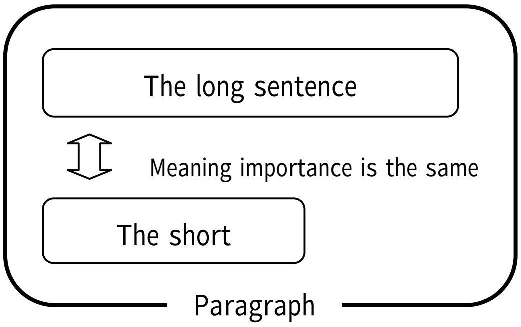

Figure 8. Postulate 2

We convert texts into the fuzzy graph according to these postulates. We define the value in the arc that connects words to constitute the same sentence as a node. Then a sentence S with W_1, W_2, $\cdots$, W_n in order from the top is shown as the graph shown in Figure 9.

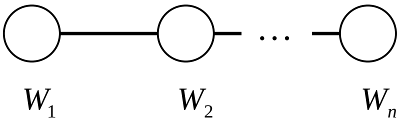

Figure 9. Graph S

3. A CONVERSION METHOD

This section describes the preparation to convert text into a fuzzy graph model

The graph S (Figure 9) is shown as the following matrix. And F is defined as a fuzzy matrix by equation (2).

$$\begin{array}{rc} & W_1\ W_2\ W_3\ \cdots\ W_n \\ \begin{matrix} W_1 \\ W_2 \\ S(i,j)=W_3 \\ \vdots \\ W_n \end{matrix} & \begin{pmatrix} & & \\ & & \\ & s(i,j)_{kl} & \\ & & \\ & & \end{pmatrix} \end{array} \qquad (1)$$

$$F_{ij} = (S(i,j)/n) \qquad (2)$$

Figure 10. Matrix and fuzzy matrix

When graphs of sentences (s(1, 1), s(1, 2), …, s(m, nm)) of text T are defined as matrix (S(1, 1), S(1, 2), $\cdots$, S(m,nm)), fuzzy graph F(T) of text T which has the following structure is defined as the equation(3).

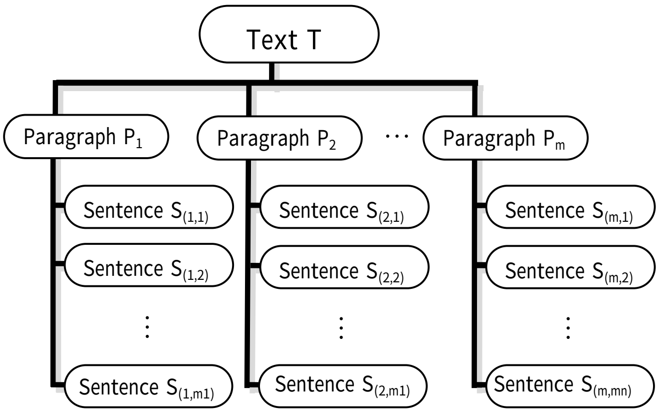

Figure 11. Structure of text T

$$F(T)=\frac{\sum_{i=1}^{m}\sum_{i=1}^{n_m}(F_{ij})}{N} \qquad (3)$$

$$\left(N=\sum_{k=1}^{m}n_k\right)$$

4. FUZZY GRAPH MODEL

We define the degree of similarity between text as follows;

Fuzzy graph $F=(f_{ij}),0\le f_{ij}\le 1,\ f_{ii}=1,1\le i,j\le n,$

Fuzzy graph $Z=(z_{ij}),0\le z_{ij}\le 1,\ z_{ii}=1,1\le i,j\le n,$

f_{ij} means degrees from node i to node j concerned

We define the degree of similarity in the fuzzy graph model as follows:

$$d(F,Z)=1-\frac{2\sum_{i,j}\left|f_{ij}-z_{ij}\right|}{n^2-n}\in[0,1] \qquad (4)$$

5. EVALUATION EXPERIMENT

However, The existing test collection was sorted by hand prior to being sorted automatically according to the clustering algorithm. On the other hand, clustering in this study did not classify text in predetermined categories, and one cluster is equivalent to one category. The number of categories changes dynamically with each increase of the text to be classified.

Therefore it is not suitable to use an existing test collection for the evaluation of our model. As a comparison object, we use the model defined by equation (5) provided by converting text into the set grouped according to frequency of appearance of given words.

We define a degree of similarity for word sets F and Z in the following expression (5), where F and Z are fuzzy matrix depending to frequency of appearance of given words. We call the fuzzy matrix words set model for short.

$$d(F,Z)=1-\frac{2\sum_{i}\left|f_i-z_i\right|}{n^2-n}\in[0,1] \qquad (5)$$

We have used 1382 reported cases (Qs and As) stored in PC beginner consultation as the database for our evaluation. This experiment is to compare Text A (described after) included in Q & A with all other texts using two methods (our fuzzy model and word set model) of calculation. As a result, we arrange texts in the list in order of degree of similarity from highest to lowest as shown in Table 1. The key passage of text and its rating of similarity to the list are represented in this Table.

Similarity of texts is evaluated in the following ways:
A: The text addresses the same problem
B: The texts are somewhat connected with each other
C: The texts have nothing to do with each other

Text A is a question sent by user concerning the sales of PCs via overseas mail order sites. As the results, we found that more two Texts of 15th, and the 28th show a similar case (ranking A). Why will order of these candidates which address the same case have fallen? We reached the conclusion that the cause depends on a difference of the user resident in Japan uses a PC in Japan and the user resident in foreign countries may use a PC abroad.

Table 1. Result by comparison of graph

Candidate	Key passage	Rating
1	The method for installing a Japanese application on to a PC purchased abroad	A
2	The method for installing the Japanese-language version of the OS to a PC purchased abroad	A
3	Problems occurring as a result of having installed the Japanese-language version of the OS in a PC purchased abroad	A
4	Problems concerning Japanese language input and Japanese-language display in a PC purchased in the U.S.A.	A
5	Problems relating to voltage when using a PC purchased in the Canada in Japan	A
6	Problems related to upgrading the OS	C
7	Problems related to specific models of PCs purchased abroad on which the Japanese-language version of the OS was installed	B
8	Problems related to upgrading the OS	C

Table 2. Result by comparison of word set

Candidate	Key passage	Rating
1	Problems related to upgrading the OS	C
2	Problems related to upgrading the OS	C
3	Problems resulting from changing the video card	C
4	The problem by the change of video card	C
5	Methods for eliminating crashes in computers purchased second-hand	C
6	The method for installing the Japanese-language version of the OS on a PC purchased abroad	A
7	Problem related to sound cards in home-made PCs acquired by auction	C
8	Problems related to direct import sound cards	B

Table 3. Result by comparison of a graph

Candidate	Key passage	Rating
1 5	Methods of Japanese language input with a PC purchased in the U.S.A.	A
2 8	Methods for installing the Japanese language version of the OS to a PC in England.	A

6. CONCLUDING REMARKS

We have proposed a fuzzy graph model that is effective for its completeness and execution time of information retrieval in database such as Qs & As and consultation offices. In this modeling, we have adopted two suppositions: the meaning of a subunit that constitutes a unit has a close relationship to one of another subunits in the same unit, and the importance of the meaning of the same level of constitutive unit is constant regardless of the number of constitutive units.

We have succeeded in building a model that represents the text with fuzzy graph. We achieved this by using co-occurrence relationships between specific words of a sentence in the text. And this model have been verified by using texts in the database of the site containing contents such as questions and answers (Qs & As) concerning treatments, problems and repairs for personal computers.

Although we have some problems to be solved for our evaluation method, we have proved the superiority of the fuzzy graph model by comparing with simple retrieval model (word set model) that is built depending upon word sets.

An issue to be addressed in the near future is improvement of the modeling and clustering of texts by using partition tree method depending on the degrees of similarity among texts. Another items is application to the other databases such as general internet sites.

REFERENCES

[1] Sunouti haruo, Yamasita hajime: Fuzzy information analysis to approach the human sciences, Kyouritsh Ltd(1995)

[2] Personal Computer User's Association: PC beginner consultation office
URL:http://www/pcua.or.jp/

[3] Harris, Zelig S. Mathematical Structures of Language. New York: Wiley(1968)

[4] Tsunenori Ishioka, Masayuki Kameda : Document Retrieval Based on Words' Cooccurrences, the Algorithm and Its Application. Journal of JASAS Vol.28, No2 (1999)

[5] Tsuneaki Kato : Similarities in Natural Language based Information Retrieval. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics vol.13, No.5, pp.436-444 (2001)

[6] Sadaaki Miyamoto, Teruhisa Miyake: On Fuzzy Information Retrieval. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics vol.3, No.1, pp.15-26 (1991)

[7] Kazuhiro Kazama, Harada Masanori: Advanced Web Search Engine Technologies. Journal of The Japanese Society for Artificial Intelligence vol.16, No.4, pp.503-508 (2001)

[8] Mitsuru Oda: Relevant Document Gathering based on the Immune System. Document of The Japanese Society for Artificial Intelligence vol.56, pp.87-92 (2003)

[9] Donald Hindle: NOUN CLASSIFICATION FROM PREDICATE-ARGUMENT STRUCTURES. Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics p.268-275 (1990)