Effectively-Heterogeneous Information Extraction To Stimulate Divergent Thinking

Kazushi Nishimoto, Shinji Abe, and Kenji Mase

ATR Media Integration & Communications Research Laboratories
2-2 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan
E-mail: {knishi | abe | mase}@mic.atr.co.jp
Tel: +81 774 95 1442 FAX: +81 774 95 1408

ABSTRACT

Conflicts in different concepts are often useful in creating new ideas. We have proposed an outsider model in which an artificial agent provides "effectively-heterogeneous" information to support human divergent-oriented discussions. Subjective experiments using a prototype system based on the outsider model and a detailed analysis on results confirming that the outsider model can extract information containing hidden relevance, i.e., "effective-heterogeneousness", are presented.

1. INTRODUCTION

Divergent thinking is one of the important human creative processes. Brainstorming is one of the well-known methods that is often used to support this process in obtaining diverse information[1]. However, a team of experts having the same domain of knowledge often share a frame of common fixed ideas; therefore, hardly any information out of the frame is obtained.

Our research goal is to construct an artificial outsider agent that supports the divergent thinking process. Experience tells us that the participation of an outsider to a brainstorming session is effective in obtaining diverse information. Such an outsider has domain knowledge different from the experts and thinks about discussion topics from a different viewpoint. Therefore, information provided by an outsider can be heterogeneous and stimulate the experts' thinking.

As the first step towards this goal, we have been researching a heterogeneous information retrieval method, one that would act like a human outsider. Ordinary information retrieval methods have mainly focused on obtaining information strongly relevant to the query, and therefore have not been able to break the frame of common fixed ideas. This has led to an outsider model that extracts effective-heterogeneous information and a prototype system based

on the model[2].

In this paper, we examine the characteristics of information retrieved by the prototype system in detail and show that the model has the potential to obtain "effectively-heterogeneous" information.

In section 2, we explain the outsider model and the structure of the prototype system. In section 3, we show some experiments and results. In section 4, we discuss the ability of the prototype system in detail.

2. THE PROTOTYPE SYSTEM

2.1 The outsider model

A brainwave is obtained by recognizing new relevance between several seemingly heterogeneous pieces of information[3]. This means that the pieces actually have hidden relevance. We define "effectively-heterogeneous information" as "information having hidden relevance". The outsider model is an information retrieval model for extracting information having some hidden relevance. This model has the following three steps.

(a) Coarse grasping of the meaning: The meaning of a participant's opinion is superficially grasped in this step. This process is realized as follows. A set of keywords is extracted from an opinion O. We call this set the "original meaning set $Go = \{g_1, g_2, ..., g_i, ..., g_{mg}\}$ ", where g_i is one of the extracted keywords and mg is the number of extracted keywords. Here, it is assumed that the set Go can represent the coarse meaning of the opinion although they do not form sentences.

(b) Shallow understanding: An outsider tries to understand the opinion of other participants using domain knowledge different from the others. This can be regarded as re-expressing the original meaning by using a different domain knowledge. This process is realized as follows. First, we prepare an associative dictionary D in the outsider's knowledge domain that is different from the other participants' knowledge domain. By referring the associative dictionary D, associative words sets are obtained from individual keywords of the original meaning set Go. All of the associative words sets are examined and a "re-expressed meaning set Gr" is obtained by extracting words appeared commonly in many of the associative words sets. Consequently, the original meaning set Go is translated to the re-expressed meaning set Gr. The relevance derived from the outsider's knowledge domain is expected to be unnoticeable to the participants.

(c) Extracting relevant information: Based on the result of understanding in the previous step, the outsider retrieves pieces of information from his/her own knowledge. This process is realized as follows. The degree of relevance between the re-expressed meaning set Gr and each article in an article database is calculated, and several articles that have high relevance degree are extracted. As it is appropriate to use a database in the same knowledge domain as a query in a conventional database system, it is also appropriate that the article database of the prototype system is of the same knowledge domain as the re-expressed meaning set Gr, i.e., as the associative dictionary D.

2.2 Structure of the prototype system

Based on the outsider model, we constructed a prototype system. Figure 1 shows its software structure and the process flow. The system has two process phases: knowledge building phase and information retrieval phase.

In the knowledge building phase, we first prepare articles in the knowledge domain that the system should have. Each article is input into the parser. After the parser analyzes an article, it generates an article vector for the article. The article vector is input into the associative memory module and the module generates/renews the associative dictionary *D*. On the other hand, the database manager registers each article together with its article vector to an article database. By this process, the system knowledge (i.e., the associative dictionary and the article database) which depends on the knowledge domain of the prepared articles is constructed.

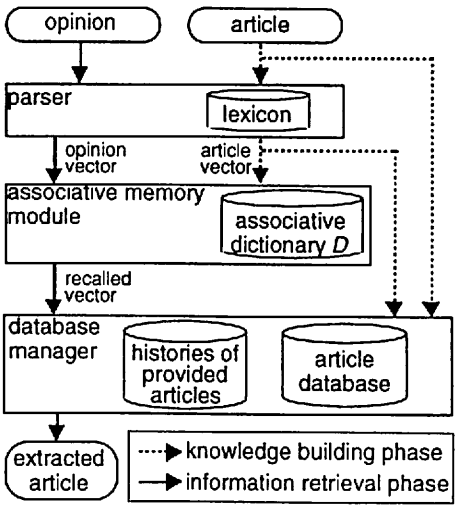

Figure 1. The software structure of the prototype system.

In the information retrieval phase, an input into the system is an opinion of a participant. The parser analyzes the opinion and generates an opinion vector. This vector corresponds to the original meaning set *Go*. Using the opinion vector and the associative dictionary *D*, the associative memory module recalls a certain keywords vector. This recalled vector corresponds to the re-expressed meaning set *Gr*. The database manager calculates the degree of resemblance between the recalled vector and the article vector of each article stored in the article database and an article with a high degree of resemblance is provided as the output of the system.

The details of each module are explained below.

(a) Parser

This module morphologically analyzes the input text (i.e., articles and opinions) to extract nouns and unknown-part-of-speech-words as keywords by the appearing order in the text. Even if a word repeatedly appears in a text, the word is employed as a keyword only once. Then, a keywords vector (i.e., article vector or opinion vector) is generated as follows.

In the knowledge building phase, where n is the number of articles to be memorized, an article vector K_j of an article A_j ($j=1 \sim n$) is denoted by the following notation;

$$K_j = (\delta_1, \delta_2, \delta_3, \ldots, \delta_i, \ldots, \delta_{m_T})^t ; \quad \delta_i = \begin{cases} 1 & (w_i \in A_j) \\ 0 & (w_i \notin A_j) \end{cases} \qquad (1)$$

where m_T is the total number of keywords obtained from the n articles (Even if a certain keyword is included in plural articles, it is counted only once). w_i is the i-th keyword of the total keyword set $W_T = \{w_i; 1 \leq i \leq m_T\}$. Therefore, the

keyword w_i that corresponds to δ_i whose value is 1 is considered as one of the keywords from the article A_j. "X^t" denotes the transposition of a vector X.

In the information retrieval phase, using an opinion keywords set $W_O = \{q_1, q_2, q_3, ..., q_k, ...\}$ obtained from an input opinion O, an opinion vector Q is generated as follows.

$$Q = (\delta_1, \delta_2, \delta_3, ..., \delta_i, ..., \delta_{m_T})^t \; ; \quad \delta_i = \begin{cases} 1 \; (if \;\; \exists w_i = q_k \; ; \; w_i \in W_T) \\ 0 \; (otherwise) \end{cases} \tag{2}$$

This vector corresponds to the original meaning set Go.

The number of δ_i whose value is 1 in both the article vectors and the opinion vectors is restricted to under m_u (constant) at most.

(b) Associative memory module

Associatron[4] was applied to the associative memory method. From this, in the knowledge building phase, n article vectors are memorized as follows;

$$M = \sum_{j=1}^{n} K_j K_j^t \tag{3}$$

where M is an associative memory matrix describing cooccurrent relations between individual keywords and corresponds to the associative dictionary D.

In the information retrieval phase, recalling is done from the opinion vector Q by using the associative memory matrix M as follows;

$$R = \phi_\theta (\phi_{\theta=0} (M) Q) \tag{4}$$

where R is a recalled vector and corresponds to the re-expressed meaning set Gr. ϕ_θ is the quantizing operator which quantizes each element, i.e., x_{ij} of a matrix X, by a threshold θ. In other words, the operation $X' = \phi_\theta(X)$ is defined as the following equation.

$$x'_{ij} = \begin{cases} 1 \; ; \; x_{ij} > \theta \\ 0 \; ; \; 0 \leq x_{ij} \leq \theta \end{cases} \tag{5}$$

The value of θ of the outer ϕ_θ in equation (4) is determined to restrict the number of elements whose value is 1 in the recalled vector R to less than m_u for every recalling.

(c) Database manager module

In the knowledge building phase, this module registers each input article A_j along with its article vector K_j to an article database.

In the information retrieval phase, this module calculates the degree of resemblance r_j between the recalled vector R and each article vector K_j ($j=1 \sim n$) as follows;

$$r_j = \frac{K_j^t \cdot R^t}{\sum_{\delta_i \in R} \delta_i} \times \frac{K_j^t \cdot R^t}{\sum_{\delta_i \in K_j} \delta_i} \tag{6}$$

where the operator " $\cdot$ " denotes the inner product of the vectors.

This module also has a history containing the list of articles already extracted as outputs. By referring to it, the system can always provide a new article to participants and avoid the used articles.

3. SUBJECTIVE EXPERIMENTS AND THE RESULTS

We conducted subjective experiments to evaluate the ability of the prototype system in obtaining effectively-heterogeneous information. The employed subjects were members of our laboratory. Therefore, they could be regarded as "same-domain" experts. The number of subjects was 24. The knowledge of the prototype system was generated from articles of "Gendai-yougo no Kiso-chishiki 93 (A Japanese dictionary of contemporary vocabularies in 1993)" by Jiyuu Kokumin Sha Co. The number of memorized articles was 10406 and the total number of keywords, i.e. m_T, was 37502.

We prepared three experimental systems with the following algorithms:

(1) Outsider algorithm: This is the prototype system described in section 2.

(2) Direct algorithm (Conventional retrieval algorithm): The prototype system without the shallow understanding step (the associative memory module) is equivalent to this. That is, an opinion keywords set W_O is directly used to retrieve the article database.

(3) Random algorithm: Articles randomly extracted from the article database.

By comparing pieces of information extracted by algorithm (1) with the other two algorithms, we could evaluate the ability of the prototype system.

We used the introduction part of an engineering paper as an opinion. This paper discusses the teleconference system that has been researched at our institute. Therefore, all of the subjects were quite knowledgeable about the contents. Five articles for each algorithm were extracted. The input opinion and a total of fifteen extracted articles were given to the subjects by concealing the algorithms that extracted the articles.

At first, the subjects were instructed to compare the opinion and each article quickly, and then perform evaluation from the following two viewpoints;

(a) Relevance: To what degree were the input opinion and the extracted article relevant? 0: No relevance; 10: Very strong relevance.

(b) Unexpectedness: To what degree was it unpredictable for you that such an article was provided from the opinion? 0: Able to sufficiently predict; 10: Completely unable to predict.

After the first evaluation, we related the following condition to the subjects. "You are discussing the teleconference system with your colleagues and an outsider. One of your colleagues states the input opinion as a personal opinion and after that the outsider gives articles as relevant opinions to your colleague's opinion. By considering this situation, to what degree were the opinion and the articles relevant? 0: No relevance; 10: Very strong relevance. Think deeply, if needed."

Figures 2 ,3 and Table 1 show the evaluation results. Figure 2 shows scatter diagrams of the evaluation results of all articles by all of the subjects for the three algorithms after the first quick evaluation. Figure 3 shows how many articles increased the degree of relevance by more than one after deep thinking. Table 1 shows the total increase in the degree of relevance for each algorithm. The total increase of an algorithm α is calculated by the following equation.

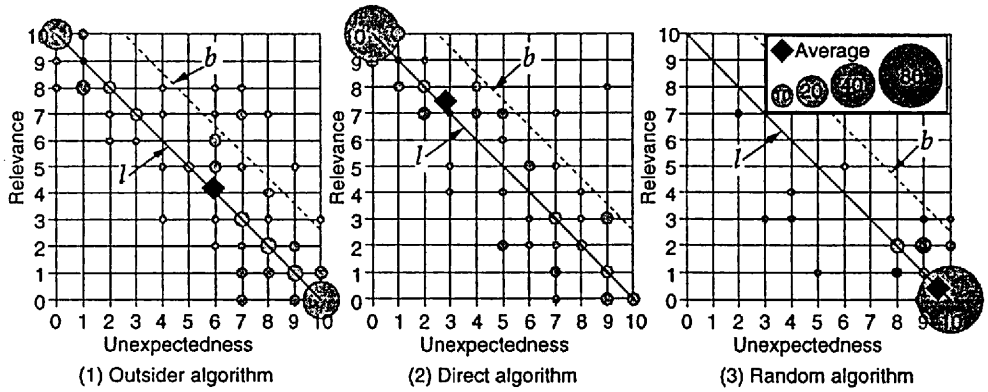

Figure 2. Scatter diagrams of the subjective evaluation results for three algorithms after the first quick evaluation.

$$TD_\alpha = \sum_i \sum_j (D_{ij} - R_{ij}) \qquad (7)$$

where $TD\alpha$ is the total difference of the algorithm α, Dij is the relevance degree after deep thinking for article j by subject i, and Rij is the relevance degree of the first quick evaluation for article j by subject i.

4. DISCUSSION

4.1 Evaluation Policy

For the purpose of stimulating human divergent thinking and supporting human creativity, obtaining moderately relevant/moderately unexpected information and highly relevant/highly unexpected information is necessary.

Generally speaking, it is difficult to notice hidden relevance clearly and it is felt vaguely. Therefore, most articles having hidden relevance with the opinion are evaluated as having moderate relevance as well as moderate unexpectedness. However, in the convergent thinking process, if someone clearly understands the relevance of such an article, it becomes one of the seeds of a new idea. If such hidden

algorithm	Outsider	Direct	Random
TDa	107	54	81

Table 1. Total increase in relevance degree after deep thinking.

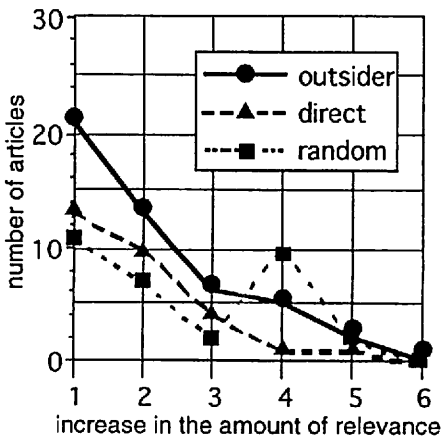

Figure 3. Number of articles whose relevance increased by more than one after deep thinking.

relevance of an article is noticed upon an article being provided, the article is evaluated as having not only high relevance but also high unexpectedness at the same time. In this case, it is expected that the frame of the subject's fixed ideas is quickly broken.

On the contrary, it is impossible for articles whose relevance people already know to stimulate divergent thinking. Such articles are evaluated as having high relevance and low unexpectedness. It is also impossible for entirely irrelevant articles to effectively stimulate divergent thinking. Such articles are evaluated as having low relevance and high unexpectedness.

4.2 Characteristics Of The Outsider Model

Based on the experimental results and the evaluation policy, we discuss the characteristics of the outsider model.

(A) Ability to obtain moderately relevant and moderately unexpected articles.

By looking at the average value in Figure 2, the following overall characteristics of each algorithm are easily recognized ;

- The direct algorithm extracts highly relevant and lowly unexpected articles.
- The random algorithm extracts very lowly relevant and very highly unexpected articles.
- The outsider algorithm extracts moderately relevant and moderately unexpected articles.

The difference in relevance and unexpectedness between the direct algorithm and the outsider algorithm and between the random algorithm and the outsider algorithm were significant by t-test. Thus, moderately relevant and moderately unexpected articles can be obtained by the outsider algorithm.

(B) Ability to obtain highly relevant and highly unexpected articles.

It has conventionally been expected that most of the results will scatter near line l in Figure 2. However, as we mentioned above, it has also been expected that some results might scatter in the high relevance and high unexpectedness area, i.e., the far-upper-right region of line l. The distance between line l and line b is $\overline{d} + 2\sigma$, where $\overline{d}$ is the average of distances between line l and all of the evaluation results and σ is the standard deviation . In the upper-right region of line b, there are eight points in Figure 2 (1), two points in Figure 2 (2) and only one point in Figure 2 (3). It has statistically been expected that there will be 2.2% the amount of data, say 2 or 3 points on average in each diagram if we assume a normal distribution and there are two or three times as many points in Figure 2(1). It is difficult to make a clear conclusion with only a small amount of data. However, the results suggest that the outsider model can obtain better highly relevant and highly unexpected articles compared with the other algorithms.

(C) Ability to obtain articles having hidden relevance.

In Figure 3, the increase in the relevance degree after deep thinking by the outsider algorithm is larger than that of the others at most of the points. The outsider algorithm achieved the best results in terms of the total increase as shown in Table 1. The random algorithm has the largest margin of relevance. Therefore, the random algorithm is potentially able to achieve the largest

increase. However, the fact that the outsider algorithm had the largest increase, where the increase of relevance derived from finding the hidden relevance, supports our conclusion that articles obtained by the outsider algorithm have more hidden relevance than articles of the other algorithms.

The shallow understanding step of the outsider model takes its relevance from a different viewpoint of the original opinion. Articles are retrieved not only by keywords originally included in the input opinion but also by associated words. Therefore, the articles include not only direct relevance to the opinion but also different relevance. Such different relevance is felt as heterogeneousness by the subjects. Although it is difficult for many of the subjects to clearly recognize the hidden different relevance at first, some of the subjects do notice the hidden relevance after deep thinking. Consequently, we can conclude that the outsider algorithm has the ability to obtain articles having hidden relevance, i.e., "effective heterogeneousness".

5. CONCLUSION

Using the prototype information retrieval system based on the outsider model, we conducted subjective experiments to evaluate the system's capability of obtaining "effectively-heterogeneous" information. It is important to note that this effective heterogeneousness is not irrelevance, but rather hidden relevance. The effectively-heterogeneous information can be expected to stimulate the human divergent thinking process. Comparing the prototype system based on the outsider algorithm with the direct algorithm and the random algorithm, we obtained the following results:
(a) Moderately relevant and moderately unexpected articles can be obtained with the outsider algorithm.
(b) There is a high possibility of extracting highly relevant as well as highly unexpected articles with the outsider algorithm.
(c) The outsider algorithm has a high capability of obtaining information having hidden relevance, i.e., "effective heterogeneousness". The shallow understanding step of the outsider model is the main contributing factor for this.

References

[1] Osborn, A. Applied Imagination: Principles and Procedures of Creative Thinking, Scribner's, New York, 1963.
[2] Nishimoto, K., Abe. S., Miyasato, T. and Kishino,F. A system supporting the human divergent thinking process by provision of relevant and heterogeneous pieces of information based on an outsider model, proc. of the eighth Int. Conf. of Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, 1995; June: 575-584.
[3] Kawakita, J. Hassou-hou, Chukou shinsho (in Japanese), 1967.
[4] Nakano, K. Associatron - A Model of Associative Memory, IEEE Trans. on S.M.C., SMC-2,3, pp.381-388, (1972).