〔公　開〕

ＴＲ－Ｍ－０００３

Effectively-Heterogeneous Information Extraction Toward An Outsider Agent for Supporting A Brainstorming Session

西　本　　一　志
Kazushi　NISHIMOTO

１９９６　２．２１

ＡＴＲ知能映像通信研究所

Effectively-Heterogeneous Information Extraction Toward An Outsider Agent for Supporting A Brainstorming Session

Kazushi Nishimoto

The 2nd departmet
ATR Media Integration & Communications Research Laboratories
knishi@mic.atr.co.jp

ABSTRACT

Conflicts in different concepts are often useful in creating new ideas. Therefore, an outsider's attendance to a brainstorming session is often effective for obtaining a brainwave. Our research goal is to construct an artificial outsider agent. As the first step to the goal, we proposed an outsider model as an information retrieval model for obtaining "effectively-heterogeneous" information, i.e., information having not only evident relevance but also hidden relevance for users and constructed a prototype system. Subjective experiments using the prototype system and a detailed analysis on results confirming that the outsider model can extract information containing "effective-heterogeneousness" are presented.

1. INTRODUCTION

Divergent thinking is one of the important human creative processes. In this process, it is important to collect pieces of information even if their relevance with the problem is not clear at a grance[1]. If someone can find some new unknown relevance among such seemingly heterogeneous information, a brainwave can be obtained[2]. Brainstorming is one of the well-known methods that is often used to support this process in obtaining diverse information[3]. 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. Therefore some supporting methods are necessary and several challenging ones have been attempted [4][5][6].

Our approach for the purpose is construction of an outsider agent. Experience tells us that 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, pieces of information provided by an outsider can be heterogeneous and stimulate the experts' thinking. The goal image is shown in Figure 1. The outsider agent participates a brainstorming session, listens to the expert's opinions and provides several heterogeneous pieces of information.

As the first step toward this goal, we have been researching on a heterogeneous information retrieval method one that would act like a human outsider. Ordinary information retrieval methods

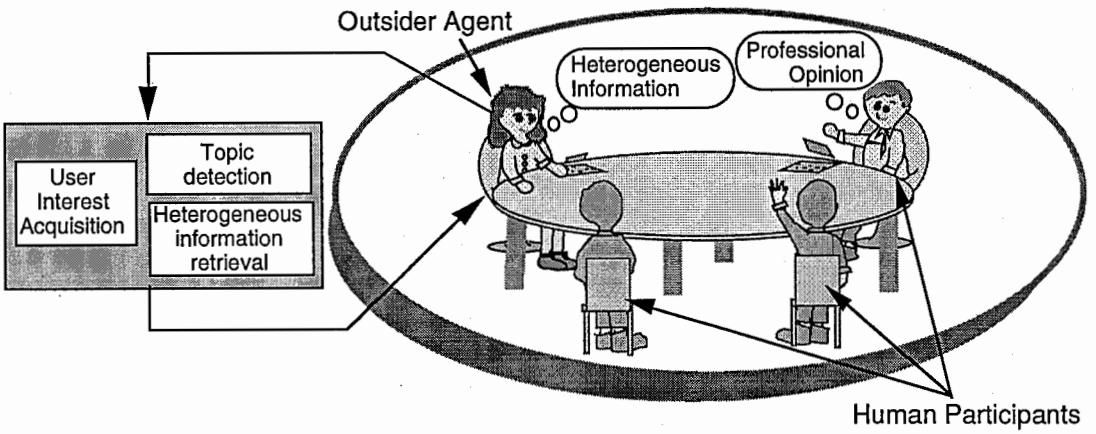

Figure 1. The goal image of supporting a brainstorming session by an outsider agent.

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[7][8].

In this paper, we explain the outsider model and the prototype system, show several subjective experiments for examination of the characteristics of information retrieved by the prototype system in detail and show that the model has the potential to obtain "effectively-heterogeneous" information[9][10].

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

2. THE OUTSIDER MODEL AND THE PROTOTYPE SYSTEM

2.1 Definition of effectively-heterogeneous information

When a subject of thinking *T* is given to a person *P*, the whole information space can be classified as followings (see Figure 2).

- Region 1: Upon being given the subject *T*, the person *P* has already recalled information in this region. Boundary *a* is *P*'s recognition limit of relevance when the subject *T* was given.
- Region 2: Given only the subject *T*, the person *P* has not yet recalled information in this region. However, upon being given a piece of information in this region, the person *P* can recognize the relevance of the piece. The outer boundary *s* is *P*'s subjective recognition limit of relevance.
- Region 3: Pieces of information in this region actually have some relevance with the subject *T*. However, the person *P* cannot clearly recognize it even if the pieces of information are given. The outer boundary *o* is objective limit of relevance.
- Region 4: Pieces of information in this region are irrelevant with the subject *T*.

The information in region 2 shows relevance that is known for the person *P* but that was overlooked. Therefore, it is expected such information has effect to directly break *P*'s fixed ideas. Relevance of the information in region 3 is difficult to be clearly noticed by the person *P* even if it is given. However, as a matter of fact such information has some relevance. Therefore, by deeply thinking, studying and finally finding the relevance, it is also expected that such information has effect to break *P*'s fixed ideas.

Consequently, we can conclude that the frame of the fixed ideas can be represented by the boundary a and/or s and providing information in region 2 and 3 is one of the effective methods to break the frame.

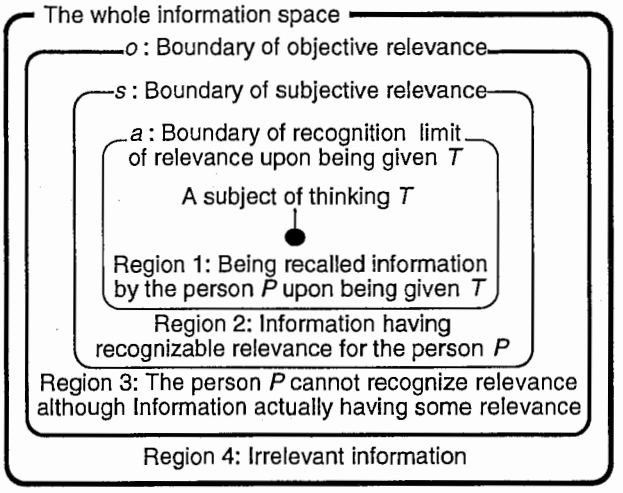

Figure 2: Classification of the whole information space

On the contrary, information in region 4 is completely irrelevant. Therefore, it is impossible to expect that such information effectively affects human thinking. Information in region 1 is basic information to think about the subject *T*. However, it is already within the scope of the person *P*. Therefore, it is also impossible to expect that such information breaks the frame of the fixed ideas.

Based on the above discussion, we define the "effective heterogeneousness" as follows. From the viewpoint of effectiveness to the divergent thinking, there are two kinds of heterogeneousness: "effective heterogeneousness" and "ineffective heterogeneousness". The "ineffective heterogeneousness" is heterogeneousness of the information in region 4, i.e., "irrelevance". On the other hand, the "effective heterogeneousness" is heterogeneousness of the information in region 2 and 3, i.e., "hidden relevance". Below, "heterogeneousness" means "hidden relevance".

2.2 The outsider model

Figure 3 shows the outsider model. This 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 $G_o = \{g_1, g_2, ..., g_i, ..., g_{m_g}\}$ ", where g_i is one of the extracted keywords and m_g is the number of extracted keywords. Here, it is assumed that the set G_o

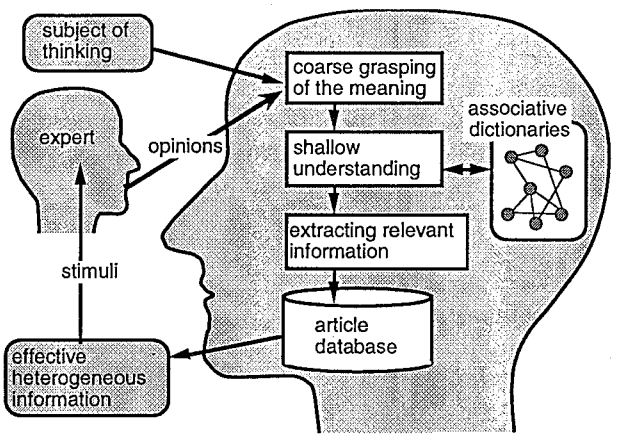

Figure 3: An outsider model

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*.

2.3 Structure of the prototype system

Based on the outsider model, we constructed a prototype system. Figure 4 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.

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.

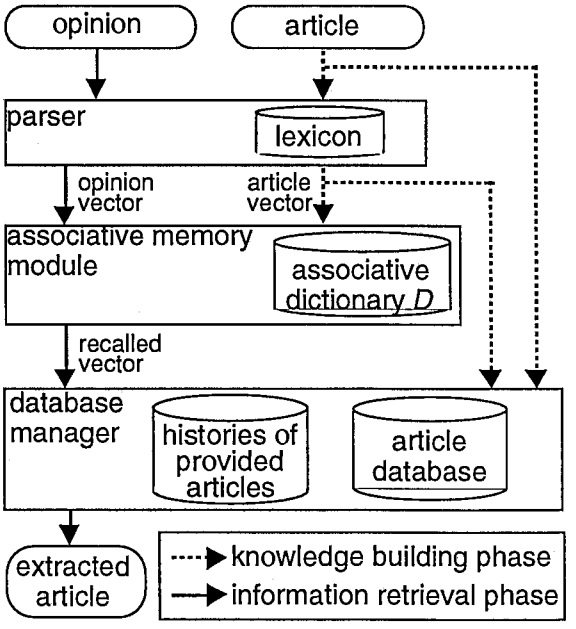

Figure 4: Software structure of the prototype system.

(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;

$$K_j = \left(\delta_1, \delta_2, \delta_3, ..., \delta_i, ..., \delta_{m_T}\right)^t ;$$

$$\delta_i = \begin{cases} 1 & (w_i \in A_j) \\ 0 & (w_i \notin A_j) \end{cases} \quad (1)$$

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.

$$Q = \left(\delta_1, \delta_2, \delta_3, ..., \delta_i, ..., \delta_{m_T}\right)^t ;$$

$$\delta_i = \begin{cases} 1 & (if \ \exists w_i = q_k ; w_i \in W_T) \\ 0 & (otherwise) \end{cases} \quad (2)$$

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;

$$M = \sum_{j=1}^{n} K_j K_j^t \quad (3)$$

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;

$$R = \phi_\theta\left(\phi_{\theta=0}\left(M\right)Q\right) \quad (4)$$

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.

$$x'_{ij} = \begin{cases} 1 \ ; x_{ij} > \theta \\ 0 \ ; 0 \leqq x_{ij} \leqq \theta \end{cases} \quad (5)$$

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;

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

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 Wo is

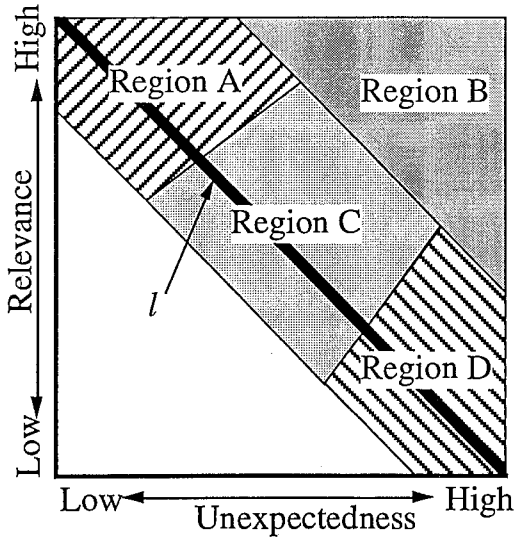

Figure 5: Evaluation results are plotted on this graph

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 strongly 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.

Evaluation results are plotted on a graph shown in Figure 5.

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 6, 7, 8 and Table 1 show the evaluation results. Figure 6 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 7 shows histgrams of frequency at each degree of relevance and unexpectedness for the three algorithms after the first quick evaluation. It also shows the average frequency of the direct algorithm and the random algorithm. Figure 8 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.

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

where TD_α is the total difference of the algorithm α, D_{ij} is the relevance degree after deep thinking for article j by subject i, and R_{ij} is the relevance degree of the first quick evaluation for article j by subject i.

4. DISCUSSION

4.1 Evaluation Policy

As we discussed in section 2, for the purpose to stimulate human divergent thinking and to support human creativity, it is necessary to extract information in region 2 and 3 of Figure 2.

Generally speaking, it is difficult to notice hidden relevance clearly and it is felt vaguely. Therefore, most of the articles having hidden relevance with the opinion are evaluated as having moderate relevance as well as moderate unexpectedness. Hence, region C of Figure 5 corresponds to region 3 of Figure 2. If such hidden relevance of an article is noticed as soon as an article was provided, the article is evaluated as having not only high relevance but also high unexpectedness at the same time and is plotted in far-upper-right region of line l of Figure 5, which is denoted as (Relevance + Unexpectedness) = 10. Hence, region B of Figure 5 corresponds to region 2 of Figure 2.

On the contrary, articles whose relevance people have already known are evaluated as having high relevance and low unexpectedness. Therefore, region A of Figure 5 corresponds to region 1 of Figure 2. Entirely irrelevant articles are evaluated as having low relevance and high unexpectedness. Therefore, region D of Figure 5 corresponds to region 4 of Figure 2.

Consequently, we can conclude that the

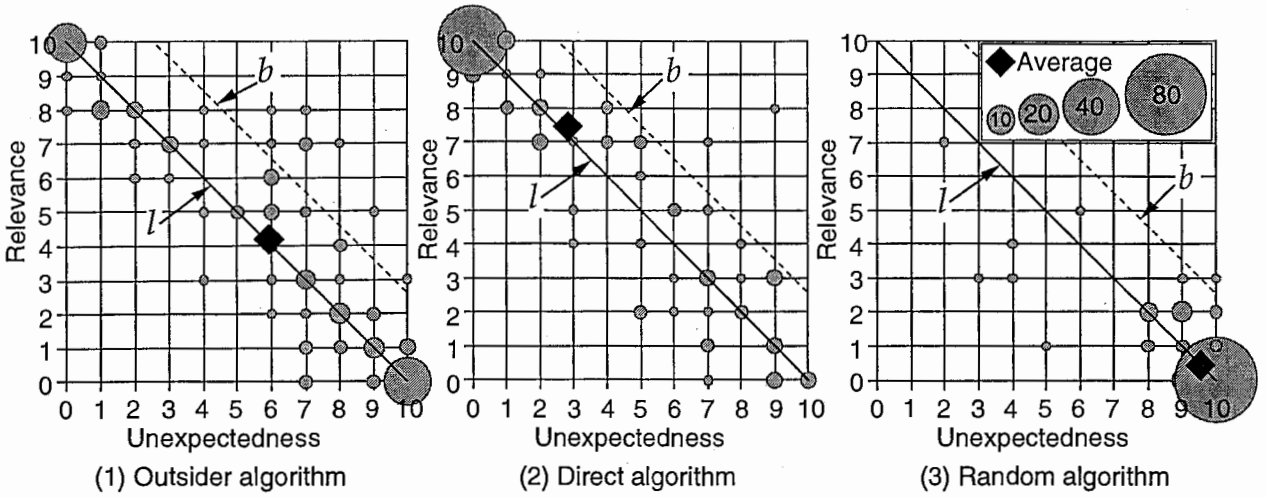

Figure 6: Scatter diagrams of the evaluation results of all articles by all of the subjects for the three algorithms after the first quick evaluation

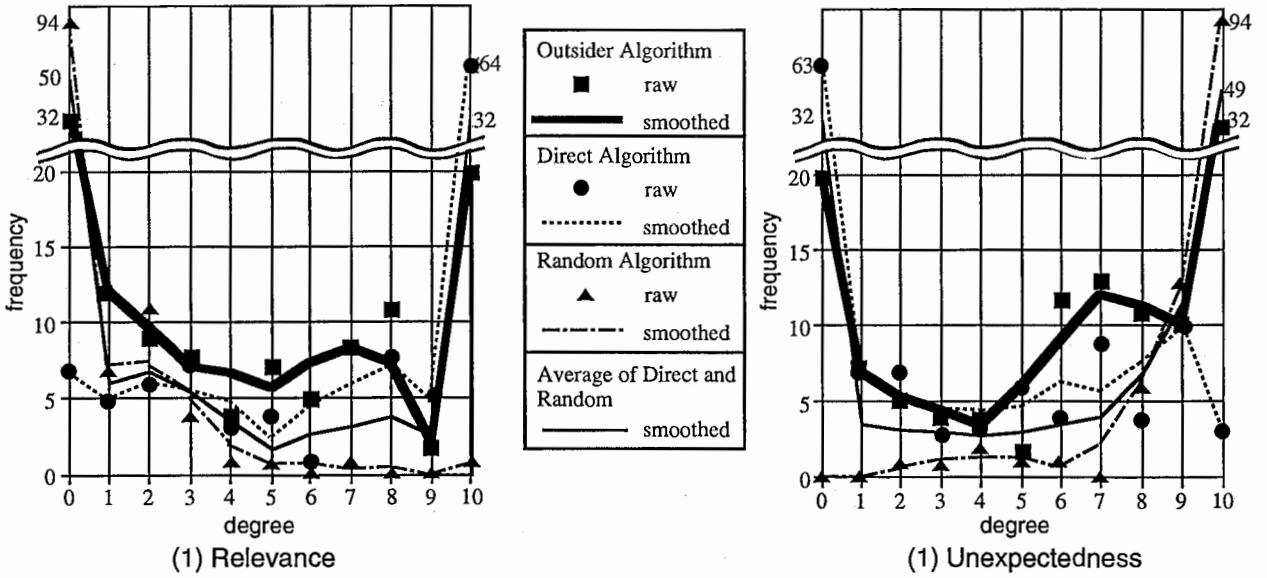

Figure 7: Histgrams of frequency at each degree of relevance and unexpectedness for the three algorithms and the average of the direct algorithm and the random algorithm.

algorithm which extracts many articles in region B and C of Figure 5 is needed.

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 6, 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.

The distribution of evaluation results in the Figure 6(1) seems to be able to be obtained by the simple combination of the other two algorithms. However, Figure 7 shows that the outsider algorithm obtained more pieces of information in moderately relevant and

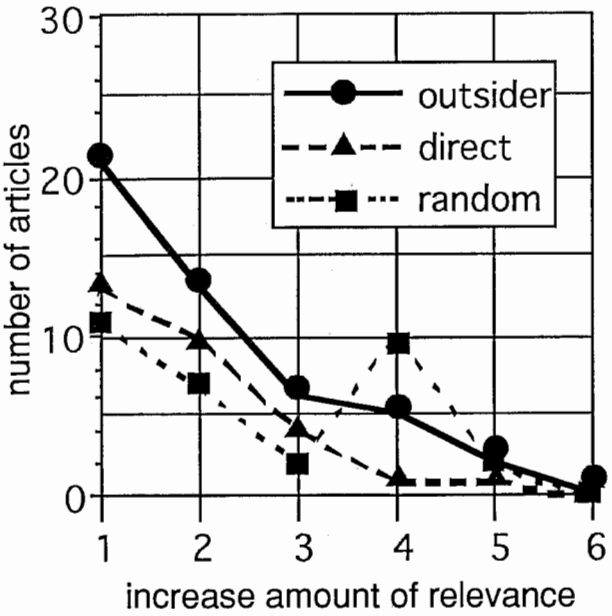

Figure 8: Increase amount of relevance after deep thinking for the three algorithms.

algorithm	Outsider	Direct	Random
TDa	107	54	81

Table 1: Total increase of the degree of relevance between post-deep thinking and pre-deep thinking for the three algorithms.

moderately unexpected region (from 2 to 8 degree) than both the other two algorithms and the average of them.

Thus, moderately relevant and moderately unexpected articles can effectively 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 6. 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 6 (1), two points in Figure 6 (2) and only one point in Figure 6 (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 6(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 8, 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 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

As the first step to create an outsider agent for supporting human divergent thinking process, especially for supporting a brainstorming session, we proposed an outsider model and constructed a prototype system for obtaining heterogeneous information. Using the prototype information retrieval system, 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.

Acknowledgement

I would like to thank Dr. K. Habara, Chairman of the Board of ATR Media Integration & Communications Research Laboratories, and Dr. R. Nakatsu, President of ATR Media Integration & Communications Research Laboratories, for giving me the opportunity of my research. I would like to thank Dr. K. Mase, Head of the second department of ATR Media Integration & Communications Research Laboratories, Dr. F. Kishino, Head of the Artificial Intelligence Department of ATR Communication Systems Research Laboratories, Dr. S. Abe, Senior Researcher of ATR Media Integration & Communications Research Laboratories, and Mr. H. Watanabe, Researcher of ATR Human Information Processing Research Laboratories, for giving me many valuable advices. I would like to thank the members of the ATR Media Integration & Communications Research Laboratories and the ATR Communication Systems Research Laboratories for taking part in the experiments.

References

[1] Kunifuji, S.:"*A Survey on Creative Thinking Support Systems and the Issues for Developing Them*", Journal of Japanese Society for Artificial Intelligence, Vol.8, No.5, pp. 552-559, 1993.(in Japanese)

[2] Kawakita, J.: "*Hassou-hou*", Chukou shinsho, 1967.(in Japanese)

[3] Osborn, A.: "*Applied Imagination: Principles and Procedures of Creative Thinking*", Scribner's, New York, 1963.

[4] Orihara, R.: "*Trends of Systems Supporting Divergent Thinking*", Journal of Japanese Society for Artificial Intelligence, Vol.8, No.5, pp. 560-567, 1993.(in Japanese)

[5] Young, L. F., "*The Metaphor Machine: A Database Method for Creativity Support*", Decision Support Systems, Vol.3, No.4, pp.309-317, 1987.

[6] Sumi, Y., Ogawa, R., Hori, K., Ohsuga, S. and Mase, K.: "*A Human Communication Support Method by Visualizing Thought Space Structure*", The Trans. of the Institute of Electronics, Information and Communication A, Vol. J79-A, No.2, pp.1-10, 1996.

[7] 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 Intl. Conf. of Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, pp. 575-584, 1995.

[8] 西本一志, "連想記憶を用いた異質性を含む情報の抽出手法の検討", ATR Technical Report TR-C-0119, 1995. (in Japanese)

[9] Nishimoto, K., Abe, S. and Mase, K.: "*Effectively-heterogeneous information extraction to stimulate divergent thinking*", proc. of Intl. symp. of Creativity and Cognition 2 (to be published), 1996.

[10] Nishimoto, K., Abe, S., and Mase, K.: "An Outsider Agent For Supporting A Brainstorming Session", proc. of ATR MIC Workshop on Intelligent Agents, (1996).

[11] Nakano, K.: "*Associatron - A Model of Associative Memory*", IEEE Trans. on S.M.C., SMC-2,3, pp.381-388, (1972).