The Intelligent Method of Information Retrieval Based on Self Organized Knowledge Resources

Takayuki MORIMOTO† Takahiro KONDO‡ Katsuhiko SUGITA‡
Daisuke ISHIKAWA‡ Masaya IKEMURA‡ Yuzuru FUJIWARA††
†Faculty of Science, Kanagawa University
‡Graduate School of Science, Kanagawa University
††National Center for Industrial Property Information

Abstract

The global flow of information is being developed at unprecedented speed. However, users can not sufficiently utilize huge amount of information by using conventional systems because their major functions are numerical calculation, symbol matching in information retrieval and deduction. Therefore, further advanced utilization of contents of information are required.

In order to realize such sophisticated utilization, it is necessary to understand meaning and characteristics of information. Therefore, the structuralization is required to represent various semantic relationships in information.

A new representation of such structure is presented and systems for self organized knowledge resources based on semantic relationships have been developed. The method of information retrieval using structuralized knowledge resources is also described.

Keywords: *Semantic Relationship, Conceptual Structure, Semantic Intelligence*

1 Introduction

Computers have achieved higher performance and lower prices, and the global flow of information is being developed at unprecedented speed by Internet. The transmission and utilization of information become more diversified and borderless very rapidly. However, users may not make good use of huge amount of information by using conventional computers whose major functions are numerical calculation, symbol matching in information retrieval and deduction. Therefore, advanced utilization of contents of information is required gradually.

In order to satisfy above requirements, semantic processing and understanding are required. It is necessary to know and to integrate concepts and semantic relationships in a huge amount of stored information for them. The above requirements are solved by semantic analysis of information and the structuralization of organized knowledge resources based on their attributes, characteristics, meaning, and so on. A new representation of such structure and systems for self organized knowledge resources based on semantic relationships are presented, and an intelligent method of information retrieval based organized knowledge resources is also described.

2 Information Retrieval

At the present time which widely spread of information, importance of information retrieval is very high. However, it is very difficult to search relevant information which is satisfied its purpose from a vast information efficiently. Web search engines are typical examples. In many cases, user can find no web page when a restriction of query is strict and a great deal of web pages when it is loose. Re-investigations are demanded in the latter cases.

"Searching contents which are referred to certain terms or concepts" is a query of general information retrieval. However, such a query practically represents to investigate special features, characteristics, and events of them. An actual meaning of a query is "searching contents which are referred to some terms or concepts with certain relationships mutually". Accordingly such relationships may not be denoted by conventional statistical information(i.e. appearance frequency).

In addition, users look for their interests from search results. In many cases, there are too many search results, and it is necessary to suggest some information for a direction of narrowing.

To solve these problems, a new intelligent method of information retrieval using organized knowledge resources based on semantics relationships is proposed.

3 Constructions of Organized Knowledge Resources

Generally, thesaurus, taxonomy, or access file has been used in order to make information adapted for knowledge resources. Relationships are divided into three types in these methods, that is, physical, conceptual and logical one called physical, conceptual, causal structures respectively. Physical structures represent physical origin and storage address. Conceptual structures represent conceptual relationships, i.e. hierarchical and other associative relationships. And, causal structures represent various logical relationships including cause/effect relationships.

Especially, thesaurus is a conventional method to represent conceptual structures, and there are many studies as follows:

- thesauri which are constructed manually[4]
- thesauri which are constructed automatically
  - compiling individual relationships for thesauri using collected documents[8]
  - merging two or more thesauri[1]
  - expert system base (dynamic methods using user information)[10]

However, these thesauri are not sufficient as far as considering contents of information. This is because that thesauri can only represent partial semantic relationships(i.e. simple hierarchical, equivalent, and associative relationships).

In order to represent various semantic relationships, semantic analysis of information and the structuralization of organized knowledge resources based on their attributes, characteristics, meaning, and so on. A model by which multiple hierarchical, overlapping, n-ary, dynamic and relative relationships can be described is devised in order to represent such semantic structures. It is apparent that neither graph model nor hyper graph model has sufficient capability to represent such conceptual structures. We proposed a new representation of such structure called Homogenized Bipartite Model and made a system for self organized knowledge resources based on semantic relationships.[3][9]

Figure 1 shows a system for the structuralization of organized knowledge resources. In this system, equivalent, hierarchical, and various semantic relationships are extracted and integrated to construct thesauri automatically.

C-TRAN (**C**onstrained **Tran**sitive Closure) :
extraction based on bilingual relations in glossaries, equivalent (synonym) and hierarchical relationships (terms represented a super-ordinate and a subordinate concepts)[2]

SS-KWEIC (**S**emantically **S**tructured **K**ey **W**ord **E**lements **I**ndex in Terminological **C**ontext) :
extraction of hierarchical and associative relationships using modified relations in terminological contexts.[5]

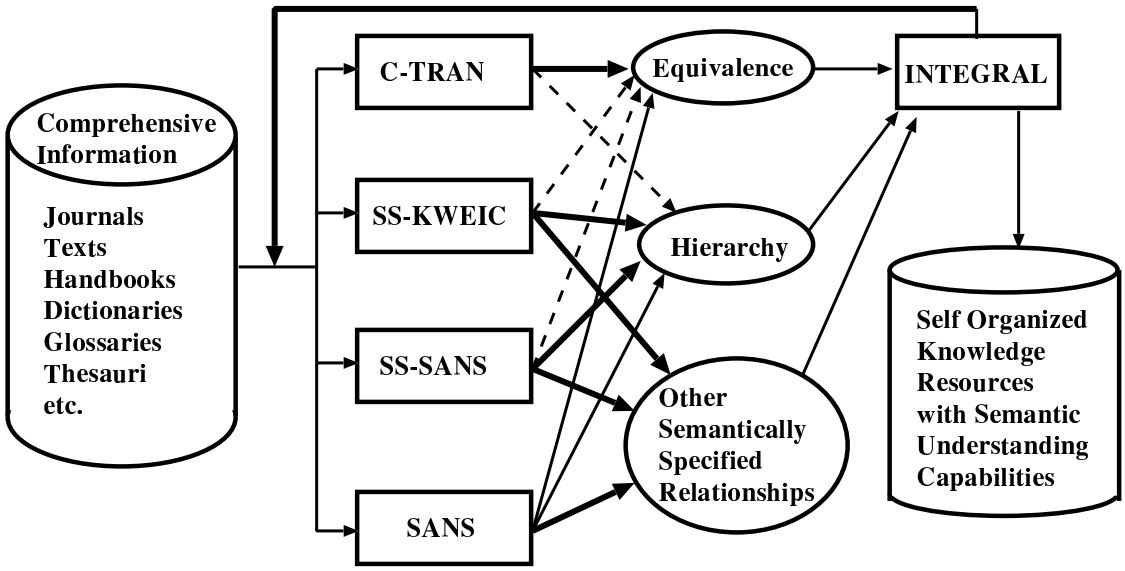

Figure 1. Self organized knowledge resources based on semantic relationships

SS-SANS (**S**emantically **S**pecified **S**yntactic **An**alysis of **S**entences) :
extraction of various semantic relationships based on syntactic templates and definitions of terminologies[7][9]

SANS (**S**emantic **An**alysis of **S**entences) :
semantic analysis of contents

INTEGRAL : building the general structure of a conceptual

4 Homogenized Bipartite Model

The Homogenized Bipartite Model (HBM) was developed in order to describe semantic relationships between conceptual structures. HBM is an extended Hypergraph, and recursive and nested structures can be described in this model. Relations which can be represented by HBM and conventional graph models in Table 1.

HBM is formulated as follows:

$$E \subseteq 2^V \quad (1)$$
$$V = V \cup E \quad (2)$$
$$E = E \cup V \quad (3)$$
$$L \rightarrow E \cup V \quad (4)$$
$$V : a\ set\ of\ vertices$$
$$E : a\ set\ of\ edges$$
$$L : a\ set\ of\ labels$$

The formula (1) represents that many-to-many relations can be described, and it is the same with Hypergraph. Recursive and nested structures are allowed by the formulas (2) and (3) respectively. By the integration between the formulas (2) and (3), nodes(V) and links(E) are homogenized. Table 2 shows the potency of representation in HBM and conventional models, and Table 3 shows structures and characteristics of relationships which can be represented in them. (As a

Table 1. Comparison between graph models

Models	Relations				
	Binary	Many-to-many	Overlap	Nest	Relative
Graph	○	×	×	×	×
Hyper Graph	○	○	○	×	×
HBM	○	○	○	○	○

Table 2. Potency of information

Type	Model	Potency	Bipartite	
Set	S = V	N	S = (V, $\emptyset$, $\emptyset$)	
Tree	S = (V, E)	2N	S = (V, E, L)	$E \subseteq V \times V$
Graph	S = (V, E)	N^2	S = (V, E, L)	$E \subseteq 2^V$
Hyper Graph	S = (V, E')	2^N	S = (V, E', L)	$E \subseteq 2^{2 \cdots V}$
HBM	S = (E", E", L)	$2^{2 \cdots N}$	S = (V, E", L)	$L \subseteq (V) \times (E, E', E'')$

Table 3. Correspondence between information structures and described semantic relationships

Type of structure	relation	characteristics(semantic relationship)
Set	–	
Tree	binary	classification : hierarchy
Graph	binary	multiple inheritance, etc.
Hyper Graph	n-ary	partial sharing, duality, etc.
HBM	n-ary	nested structure, modality, relativity, etc.

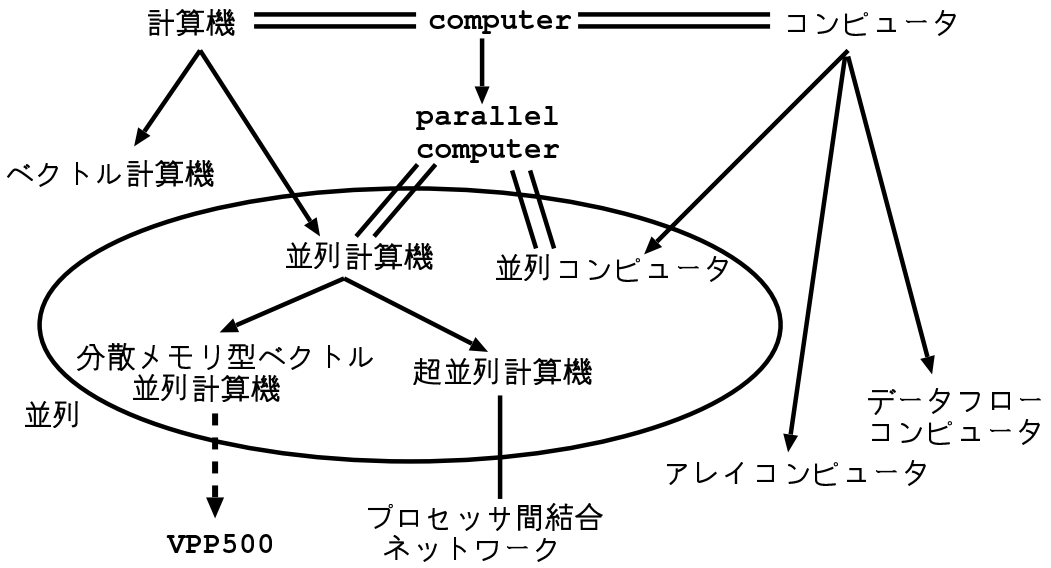

Figure 2. Example of Organized Knowledge Resources Based on HBM

matter of course, lower structures in Table 3 include upper structures.) It is clearly that HBM is more capable than other models.

Figure 2 shows an example of organized knowledge resources based on HBM. In this figure, the solid line directional arrow represents hierarchical relationship. Two folds of solid lines represents equivalent relationships, and the wavy line directional arrow shows inclusive relationships, One solid line and circle expresses association relationships. A difference of above associative relationships is that the circle is based on the key word "並列" and solid line is extracted from a context by SS-SANS. The associative relationship between "超並列計算機" and "プロセッサ間結合ネットワーク" shows not only that these terms are included in the contents but also that there is/are contents which concerned with them mutually.

Moreover, conceptual structures based on HBM can be used for major thinking functions such as induction, analogical reasoning, (analogical) abduction. The mechanisms of such functions are as follows:

Let $C = (V, E)$ be the universe of concepts, and $C_r = (V_r, E_r), C_s = (V_s, E_s), C_c = (V_c, E_c)$, where r, s, and c designate reference, sample, and common substructures respectively. The mechanism of induction:

$$\begin{aligned} C_c &\subseteq C_{si} \cap C_{sj} \cap \cdots \cap C_{sn} \cap C_r \\ C_{s'} &= (V_{s'}, E_{s'}) \\ i.e.\ \ V_r &= V_c \cup \delta V_r \\ E_r &= E_c \cup \delta E_r \\ V_{s'} &= V_s \cup \delta V_r \\ E_{s'} &= E_s \cup \delta E_r \end{aligned} \tag{5}$$

The mechanisms of analogical reasoning and (analogical) abduction:

$$\begin{aligned} C_c &\subseteq C_s \cap C_r \\ C_{s'}(V_{s'}, E_{s'}) &= C_s(V_s, E_s) \\ &\quad \cup \delta C_r(\delta V_r, \delta E_r) \\ i.e.\ \ V_r &= V_c \cup \delta V_r \\ E_r &= E_c \cup \delta E_r \\ V_{s'} &= V_s \cup \delta V_r \\ E_{s'} &= E_s \cup \delta E_r \end{aligned} \tag{6}$$

5 Systems of Information Retrieval

5.1 Concept

Information retrieval systems consist of two parts. One is structuralization of knowledge resources, and another is search process. Knowledge resources are organized with the following procedures using the system shown in Figure 1.

1. extracting terms and their semantic relationships from contexts
2. partitioning terms into basic words based on terminological contexts using "JUMAN"[6]
3. structuralizing terms according to their semantic relationships

Each term and each semantic relationship have their own source information. Search processes navigate various semantic relationships, and show search

results and information which are associated with queries.

Prototype system is implemented and consists of hierarchical and equivalent relationships at present.

5.2 Example

Figure 3 shows a part of search results whose query is “並列コンピュータ”. We use a test collection called NTCIR-2 provided by National Institute of Informatics and “Information Processing Dictionary(ISBN:427407742X)” published by Ohmusha as input data.

In this figure, indents represent level of hierarchy. Directional arrows represent a direction of hierarchy from upper concept to lower concept, and an equal sign represents equivalent relationships. The term surrounded “『』” shows a query. In addition, a “gakkai-j-XXXXXXXXXX” shows source information of terms.(omit it by using “...” when more than three sources of information correspond)

6 Parallelization

Our systems are implemented by programming language C. For efficiency of processes and to treat enumerate knowledge resources, we implement to parallelize our systems using MPI.

The items listed bellow are processes of parallelized structuralization. These processes are executed on master-slave system.

1. Slaves construct conceptual structures using their own input data, and send their own information of them to master.
2. Master totals them and broadcasts new information for distribution of knowledge resources.
3. Slaves send and receive conceptual structures using received information, and re-construct them.

Conceptual structures are distributed on average based on hierarchical relationships by above processes. Theses processes are executed whenever new knowledge resources are inserted. Figure 4 shows the example of above structuralization.

In our system, conceptual structures are distributed slave processes. Therefore, information retrieval consists of searching queries, totalization of conceptual structures, and re-construction of them. The following items are the processes of information retrieval.

1. Master broadcasts a query to slaves.
2. Each slaves receives it and does search processes.
3. Slaves send conceptual structures which are associated with the query when it is matched.
4. Master receive and re-construct them.

コンピュータ : gakkai-j-0000341470 gakkai-j-0000342371 gakkai-j-0000343309 ...
→ノートブック型コンピュータ : gakkai-j-0000342168
→携帯型コンピュータ : gakkai-j-0000348016
→脳型コンピュータ : gakkai-j-0000342489
→『並列コンピュータ』 : gakkai-j-0000340080
→超並列コンピュータ : gakkai-j-0000345205
＝計算機 : gakkai-j-0000343562 gakkai-j-0000344216 gakkai-j-0000345141
→アナログ計算機 : gakkai-j-0000343604
→ 64 ビット計算機 : gakkai-j-0000342940
→ベクトル計算機 : gakkai-j-0000342091
→メッシュバス計算機 : gakkai-j-0000342093
→モーバイル計算機 : gakkai-j-0000342157
→分散共有計算機 : gakkai-j-0000342728
→移動型計算機 : gakkai-j-0000342159
→ SMP 型計算機 : gakkai-j-0000343419
→並列計算機 : gakkai-j-0000340082 gakkai-j-0000340084 gakkai-j-0000340086 ...
→仮想並列計算機 : gakkai-j-0000345206
→クラスタ型並列計算機 : gakkai-j-0000342073
→分散メモリ型並列計算機 : gakkai-j-0000342206
→超並列計算機 : gakkai-j-0000340098 gakkai-j-0000342135 gakkai-j-0000342136 ...

Figure 3. Example of Search Results

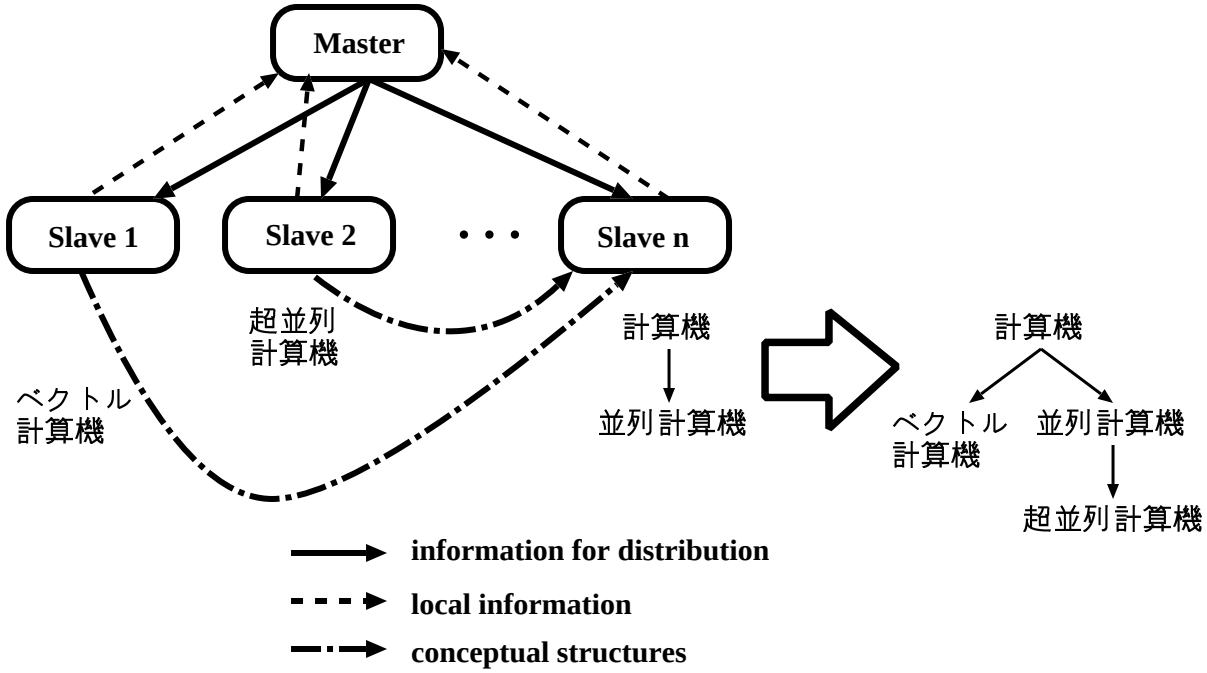

Figure 4. Parallel Structuralization of Organized Knowledge Resources

When users need more information, above processes are executed again using a new query. Figure 5 shows the example of information retrieval whose queries are "計算機" and "コンピュータ".

Prototype of above parallel system which supports only hierarchical relationships is completed.

7 Conclusions and Future Works

Structuralization of organized knowledge resources shows usefulness to realize advanced functions for semantic contents of information as new functions of computer which are demanded for the global flow of information which is being progressed at unprecedented speed. In this paper, we report the utilization of structuralized knowledge resources based on semantic relationships for information retrieval.

There are two future works. One is to implement more complex and flexible conceptual structures. Another is to expand the parallelization of our system.

Acknowledgment

This study used test collections called NTCIR-1 and NTCIR-2 provided by National Institute of Informatics.

References

[1] R. Forsyth and R. Rada. *Machine Learning-Applications in Expert Systems and Information Retrieval*. England:Ellis Hornwood Series in Artificial Inteligence, 1986.

[2] Y. Fujiwara. The model for self structured semantic relationships of information and its advanced utilization. *International Forum on Information and Documentation*, vol.1 9(2):8–10, 1994.

[3] Y. Fujiwara and Y. Liu. The homogenized bipartite model for self organization of knowledge and information. *IFID*, 2(1):13–17, 1998.

[4] A. Ghose and A. Dhawle. Problems of thesaurus construction. *Journal of the American Society for Information Science*, pages 211–217, 1997.

[5] J. Lai, H. Chen, and Y. Fujiwara. An information-base system based on the self-organization of concepts represented by terms. *International Journal of Terminology*, vol. 3(2):313–334, 1996.

[6] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html.

[7] H. Sano and Y. Fujiwara. Syntactic and semantic structure analysis of article titles in analytical chemistry. *J. Inf. Sci. Principles and Practice 19*, pages 119–124, 1993.

[8] D. Soergel. Automatic and semi-automatic methods as an aid in the construction if indexing language and thesauri. *International Classification*, 1(1):34–39, 1974.

[9] T. Morimoto, T. Maeshiro, and Y. Fujiwara. Extraction of semantic relationships among terms to construct organized knowledge resources. In *Proc. of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pages 459–465, 1999.

[10] U. Guntzer, and et al. Automatic thesaurus construction by machine learning from retrieval sesseions. *Information Processing and Managament*, 25(3):265–273, 1989.

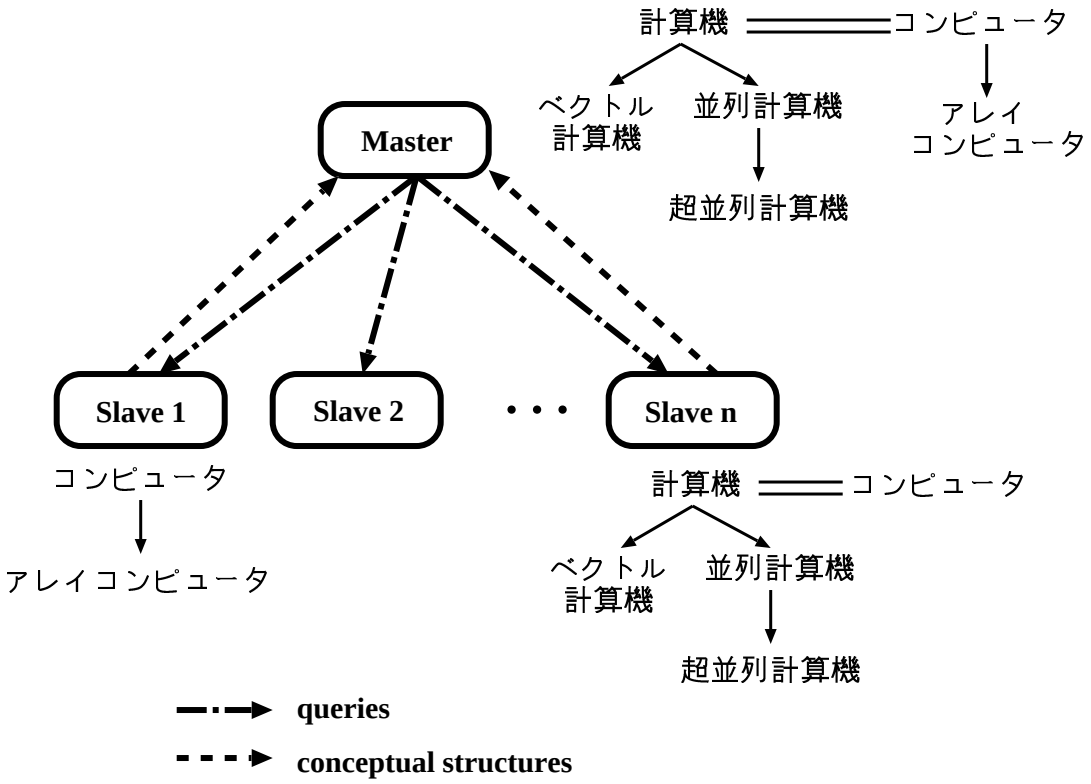

Figure 5. Parallelization of Information Retrieval