ICOT Technical Report: TR-251

TR-251

BASIC METHOD FOR MUTUAL UTILIZATION OF DISTRIBUTED PERSONAL KNOWLEDGE BASES

by

M. OBA and H. ITOH

April, 1987

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1 Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

BASIC METHOD FOR MUTUAL UTILIZATION OF DISTRIBUTED PERSONAL KNOWLEDGE BASES

Masahiro OBA
Hidenori ITOH

Research Center
Institute for New Generation Computer Technology
(ICOT)

21F, Mita Kokusai Bldg.
4-28 Mita 1-Chome, Minato-ku, Tokyo 108
Japan

ABSTRACT

This paper outlines an approach to the distributed knowledge-base system control mechanism, which is being researched and developed in the Fifth Generation Computer System (FGCS) project.

Traditionally, central management of data and knowledge has been widely adopted to share knowledge bases. Another type of sharing and mutual utilization of data and knowledge is where an individual (or group of co-workers) collects data and knowledge necessary for his jobs and for his purpose, describes them using his own words and special words for his job field, and arranges and classifies them based on his views and job. This type of collection of data and knowledge is called a personal knowledge base (PKB). PKBs connected by a local area network are considered as a distributed knowledge base system. The following requirements are necessary for mutual utilization of PKBs.

(1) The content of PKBs created by others must be understood by the users.
(2) Users must be able to integrate PKBs to execute queries and solve problems.
(3) Each PKB must be flexible to cope with the changes in the configuration of distributed knowledge-base systems.

For (1), it is important, first of all, that the words used in PKBs be mutually understood.

Conventionally, the main issues of distributed databases and heterogeneous databases have been management of distributed data, query processing in distributed environments, data-model transformation and data-language translation. However, the verbal problem to mutually understand the words in PKBs has not been considered.

This paper proposes a basic method of control mechanism for PKBs. Each creator of a PKB has his own personal thesaurus (PT) which reflects his objective, view, problem field, and vocabulary. Identifiers for the words in his PT are attached to each record in his PKB. The proposed method enables the creator and users to understand mutually the words in different PTs, to integrate PKBs for problem solving and query processing, and to cope flexibly with system configuration change. The basic concept of the method is word matching, which considers meaning similarity through comparison among PTs, and its architecture based on communication and control.

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

The Fifth Generation Computer Systems (FGCS) project in Japan aims at developing inference and knowledge-base mechanisms to implement a knowledge information processing system. One important research subject is to develop a distributed knowledge-base system, consisting of inference machines, knowledge-base machines and a local area network (LAN) connecting them. Basic components for the knowledge-base machine, such as the relational database system, deductive database system, and distributed database control mechanism, have been developed with the architecture shown in Figure 1[1]. The goal of this study is to develop a communication and control method embedded in a distributed knowledge-base control mechanism.

There are many ways of sharing and disseminating knowledge and data. One of the most frequently observed is that of gathering knowledge and data at one place, where experts (or authority) systematically arrange and classify the knowledge and data so that anyone who wants information can access it easily using formats or procedures defined by the experts. Experts also maintain and manage the gathered knowledge and data. An example of this is a library.

Another way of disseminating knowledge and data is found in daily research or office activities. A person (or a group) collects data, knowledge and materials for his (or their) own purpose, and arranges and classifies them so that he can use them easily. A person who wants to refer to and utilize information held by others may communicate with them to obtain the information that he needs. This form of communication has no predefined format or procedure. By mutually understanding the meaning of the words and concepts for the situation, the user's requirements are transmitted to the person offering the information.

In the computer world, the former type of sharing knowledge and data has been researched and developed as large-scale databases and large-scale knowledge bases.

We aim to implement the latter type. The goal of our research is to realize sharing or mutual utilization of knowledge-bases or databases, each of which are created and maintained by an individual or a group. These knowledge bases or databases are called personal knowledge bases (PKBs). The creator allows some part of each PKB to be used by the public. Users can access and

refer to PKBs for which access is permitted and can integrate the information necessary for their jobs. Examples in the research, development, and office management fields are experimental data collection, bibliographic files, materials and data for decision making, and program tools.

There are many new research and development issues for sharing PKBs, due to the characteristics of PKBs.

The PKB creator's objective, way of forming questions, and way of looking at things form the basis of creation and maintenance of his PKB. Therefore, PKBs give rise to the following characteristics:

* Multiform -- variety of representation language when knowledge is created and variety of structure when data are stored.
* Hetero-substance -- diversity of content (first of all, vocabulary) of stored knowledge and data.
* Independence -- freedom of the PKB creator to modify the body of knowledge and data.

Furthermore, PKBs which are connected by LANs have the characteristic of openness. This means that PKBs may be changed by adding or removing inference machines and knowledge-base machines to/from These characteristics result in non-unified and non-systemwide management. For users to access PKBs and integrate the necessary information, distributed control techniques and distributed problem solving method suitable for sharing PKBs must be developed.

Much research and development have been carried out in the area of distributed databases. One topic is management and control for distributed databases[2][3][4]. The main issues of this research are distributed data management, transaction management and commit protocols, and query execution. Another topic is heterogeneous databases[5][6][7]. The main issues of this research are data-model transformation and data-language translation, which are problems of multiform. However, these topics do not deal with hetero-substance.

This paper proposes a basic method for sharing and mutual utilization of PKBs. The emphases of this paper are :

(1) Focus on hetero-substance, independence, and openness. In particular, this paper focuses on verbal problems such as those in hetero-substance. Because it is natural that people use their own vocabulary to create the body of their

PKBs, the verbal problem of different vocabulary is important for mutual understanding of PKBs.

The reason why multiform is excluded is that our system, shown in Figure 1, is based on the relational database and deductive database systems.

(2) Global schema are not adopted. The approach is to share PKBs as they are. For that purpose, it is assumed that each PKB creator has his own personal thesaurus (PT) which represents his conceptual system or view. In order to explain the meaning of the content of knowledge and data, he gives words in his PT to the body of knowledge and data. Verbal problems can be solved by word matching between PTs, considering the meaning of words.

(3) The method and techniques of matching words in PTs should allow the computer system to perform daily human activities. We will develop a communication method on a computer system that follows the model of human activity to understand words and concepts mutually. Accordingly, performance and efficiency are out of the question at present.

Section 2 examines the problems and requirements for sharing PKBs. Section 3 examines the means for the requirements and presents an overview of a framework for the distributed knowledge-base control system. Section 4 presents an overview of functional construction for word matching, which is a central part of the framework. Section 5 gives an example of the word matching process. Section 6 summarizes and concludes this paper.

2. MUTUAL UTILIZATION OF PERSONAL KNOWLEDGE BASES

This section examines the characteristics and problems of PKBs, and introduces requirements based on them.

2.1 Characteristics and Problems of PKBs

Characteristics of PKBs, such as hetero-substance, independence, and openness, emerge from the method of creation and maintenance, and LAN environment.

(1) Hetero-substance

The problem area, research and development phases, purposes, viewpoints, conceptual systems, methods, and theories depend on the creators. Thus, the content of PKBs is different for every PKBs. For example,

(a) Vocabulary -- the same words which appear in different PKBs do not necessarily designate the same object or the same concept.

(b) Content (algorithm, rules)

(2) Independence

The many bodies of knowledge and data are developed at the discretion of their creators. Thus, it may be possible that not all the information needed by a user is in one PKB, but is in several PKBs. In this case, it is necessary to integrate the information from different PKBs.

Besides, because job and study phases, and the way the creator sees the problem are developing and changing, the body of knowledge and data may be modified in form and substance. Thus, the part of the PKB offered to the public is constantly changing.

(3) Openness

The number of inference machines and knowledge-base machines changes. The configuration of the total system, including inference machines, knowledge-base machines, and LANs, is not fixed.

These characteristics result in inconsistency and incoherency of the body of PKBs, as well as non-unified and non-systemwide management.

2.2 Requirement for Mutual Utilization of PKBs

For sharing PKBs, it is necessary to resolve these characteristics.

(1) For hetero-substance

Users must be able to understand the content and meaning of others' PKBs to judge which PKBs are useful for their information request. It is desirable for users to describe a problem or write a query statement in their own vocabulary.

(2) For independence

The whole of each PKB must be adaptable to be integral for problem solving or query processing. Further, this must hold, regardless of any changes in the parts of PKBs offered to the public.

(3) For openness

Flexibility and expansibility for change of configuration of the total system are required.

3. BASIC CONCEPT FOR SHARING PKBs

This section examines the ways to resolve requirements stated in the previous section and integrate them into a communication and control mechanism for a knowledge-base machine.

3.1 Existence of Personal Thesaurus

The body of each PKB depends on the creator's vocabulary. Since it is desirable for users to describe problems or queries using their own words, it is necessary to coordinate the user's words and the PKB creator's words.

For coordination, this paper adopts the method of translating the user's words into the PKB server's words, and vice versa. In this paper, it is assumed that each PKB creator has his own personal thesaurus (PT) in his knowledge-based system, as a basic apparatus for translation. The user's words are allocated to the PKB creator's words by matching the user's and the creator's PTs.

3.1.1 Personal Thesaurus

The PT is constructed in the following way (Figure 2). PKB creators undertake to attach the following two items to each body of knowledge and record of data in their PKBs according to their objectives, views and notions, using their own words.

(1) Keywords [Classification words] -- keywords, indices, or abstract concepts which characterize and signify the content of knowledge and data.

(2) Relationships between keywords [classification words] -- expressed as knowledge in the form of rules, for example.

If keywords are collected from the entire PKB created by an individual and if they are structured, the structure may represent the conceptual system, viewpoint, and notions of that person, and can therefore be regarded as his personal thesaurus. An example of PT is shown in Figure 3.

It is assumed that each person has his own PT and that the storage model and retrieval method for the PT is unified between all PKBs in this distributed knowledge-base system.

3.1.2 Matching PTs

The words used in the user's problem or query are restricted to those in his PT. Access to a PKB is executed using words in the creator's PT. Therefore, the words in the user's PT must be closely associated with those in the creator's PT.

The basic method for this association is to compare and correlatethe user's PT with the creator's PT so as to match words in each PT considering the similarity of meaning. This method is called PT matching (Figure 4).

PT matching enables the following:

(1) Keywords in a user's query are allocated to suitable keywords in meaning to access the creator's PKB.

(2) Keywords attached to the body of the creator's PKB are converted into suitable keywords in the user's PT. The user can understand the content of the creator's PKB in his own words.

3.2 Framework for Problem Solving and PT Matching

Where creators manage and maintain PKBs at their own discretion, the content and storage structure of the PKBs, and keywords and structure of PTs may change as the creator's research phase, the way he sees the problems, or interests change.

Problem solving in distributed systems is performed by information exchange among related sites and access to PKBs in each site. The way of problem solving should be independent of those changes. Further, it is desirable that all the related sites for problem solving work in a totally integrated way.

For these requirements, two frameworks are employed: one for problem solving and one for PT matching (Figure 5). The purpose of the framework for PT matching is for the framework for problem solving to execute problem solving without consciousness of verbal distinction.

3.2.1 Framework for PT Matching

This is a framework to execute PT matching by communication and control in a distributed processing environment. This framework is independent of problem solving type and can be used commonly to the variety of problems. All the related sites for PT matching have this common framework in knowledge-base machines.

A framework for PT matching on a site interchanges predefined messages with other sites. The messages are used for information exchange, control of other sites' activities, etc. On receiving the message, the framework invokes execution of procedures dependent on the message. The procedures include those for PT matching. Message interchange and procedure execution by the framework perform PT matching.

Section 4 discusses the construction of this framework.

3.2.2 Framework for Problem Solving

This is a framework to carry out problem solving efficiently by communication and control in a distributed processing environment. The problem-solving framework should be problem-type-oriented. Several problem-solving methods and frameworks have been developed, one of the most famous of which is Contract Net[8]. All the participant sites for the problem solving embed the same framework in knowledge-base machines.

The function of this framework is similar to that of the framework for PT matching. The procedures of this framework do not include those for PT matching, but those for tasks dependent on this framework. Message interchange and procedure execution by this framework perform problem solving and query processing.

3.3 Common Communication and Control Mechanism

To support the configuration change and system growth of the distributed PKB system, flexibility and expansibility require that there be no modification to the existing PKBs and no new programming for the expansion.

This requirement can be resolved by embedding common communication and control mechanisms in each knowledge-base machine. The mechanism operates the functions of the frameworks described above.

3.4 Basic Configuration for Sharing PKBs

The discussion so far has focused on the basic configuration, the overview of which is shown in Figure 6.

The framework of problem solving and that of PT matching jointly make tasks progress. The course of problem solving and PT matching is shown using an example taken from query translation. On the user's site, the framework of problem solving (psf) passes the user's query to the framework of PT matching (PTmf). Of course, the query is written in the user's vocabulary. The PTmf on the user's site transfers the query to the PTmf on the server's site. Then, the PTmfs on both sites work together for PT matching. Occasionally, the PT matching may be carried out accurately with the co-operation of other sites. The PTmf on the server's site translates the user's words into the server's words and hands over the query to the psf on its site.

4. Architecture for PT matching

This section describes the basic architecture of the framework for PT matching.

4.1 Prescription for the Framework

The PT matching framework prescribes four conceptual components.

(1) Protocol

The protocol specifies the possible interaction between PT matching frameworks not only on the sites which participate in problem solving and query processing but, if necessary, those which are not participating. It encodes information that specifies the possible actions and interactions of each site. It also encodes information necessary for the decisions that guide the control of the site.

(2) Message processing sequence

On receiving a message, each site invokes actions specified by the massages and may send responses. The message processing sequence specifies the order of invocation of actions for the messages of the protocol. Actions are

categorized into two types: those for PT matching and those which are independent of PT matching.

(3) Way of PT matching

Correlating PTs gives rise to verbal meaning and classification structure problems. Knowledge of how to correlate PTs is required. Actions for PT matching are specified by the knowledge.

(4) Message processing procedures

During PT matching, there are some tasks other than PT matching itself, e.g., total decision making and synchronization of execution. These tasks specify the actions independent of PT matching.

4.2 Structure of the Framework

The PT matching framework is structured as shown in Figure 7.

Knowledge sources are grouped into: message processing sequences (KS1), message processing procedures (KS2), and PT matching knowledge (KS3). These are introduced in the previous section.

Memories store data as follows:

(1) The message history memory (MHM) contains messages received from, sent to and waiting to send to other sites.

(2) The query memory (QM) contains the user's queries as they are, i.e., conditions in the queries are described by the user's words.

(3) The other sites' thesauri memory (OTM) contains the parts of others' PTs.

(4) The word matching memory (WMM) contains a table of words correlated together.

(5) The schedule memory (SM) contains the invocation sequence of PT matching knowledge and message processing procedures.

The function of processing units are as follows.

(1) The communication manager (CMM) controls message transfer and reception.

(2) The scheduler (SCH) plans the invocation sequence of PT matching knowledge and message processing procedures based on the priority of received messages. It may change the stored sequence or delete part of the sequence.

(3) The inference (INF) mechanism executes the PT matching knowledge.

(4) The message procedure execution unit (MPE) carries out the execution of message processing procedures.

5. EXAMPLE OF PKBs SHARING

This section shows an example of mutual utilization of PKBs and gives a rough idea of PT matching.

The example is bibliography retrieval. A bibliography database is assumed to exist in someone's PKB. It is also assumed that a server of a bibliography database has the PT shown in Figure 3, and a person who wants to access to it (user) has the PT shown in Figure 8. The user is looking for a paper about a "knowledge-acquisition-support-system" which focuses on the "completeness-of-knowledge-base" using "meta-knowledge". The words with quotation marks are used in the user's query statement for retrieval keywords.

The system runs as follows:

* The CMM on the user's site transfers the message to the CMM on the server's site. The query is encoded in the message.
* The CMM on server's site receives the message and puts it into the MIIM.
* The SCH plans the appropriate processing sequence for the message.
* According to the sequence, the query is extracted and stored into the QM, the MPE executes the message processing procedure, and the INF executes the PT matching knowledge.
* These executions may require some information about the user's PT, e.g., "knowledge-representation-function" as an upper concept of "meta-knowledge", and "integrated-AI-system-building-environment" as an upper concept of "knowledge-representation-function." These words and the upper-lower relation are stored in the OTM.
* The result of PT matching execution may be

  knowledge-acquisition-support-system .vs. knowledge-acquisition
  completeness-of-knowledgebase .vs. consistency
  meta-knowledge .vs. meta-knowledge .

The result of matching is stored in the WMM.

* Based on the result, the user's keywords in the query are translated into the server's keywords. A query suitable for the server's site is generated.
* The KBMS performs retrieval using the new query. Selected records are organized into a message for the user site, which is stored in the MIIM.

* The CMM on the server's site transfers the result message to the CMM on the user's site.

This example considers only two sites. However, the system can be enhanced to cover more than two sites.

6. CONCLUSION

This paper focused on a distributed knowledge-base system, whichis composed of personal knowledge bases (PKBs) connected by LANs. A PKB is created and maintained by an individual according to his objective, view, and job field. For the mutual utilization of PKBs, the following requirement must be fulfilled.

(1) The content of PKBs created by others must be understood by users.

(2) Users must be able to integrate PKBs to execute queries and solve problems.

(3) Each PKB must be flexible for the change of configuration of the distributed knowledge-base system.

For (1), it is important, first of all, that the words used in PKBs be mutually understood.

This paper proposed a basic method for communication and control mechanism of the distributed knowledge-base system. It also proposed an architecture for the framework of correlating words.

We are now examining the protocol and the way of correlating words into details in order to implement the system.

ACKNOWLEDGEMENTS

We would like to thank Dr. K. Fuchi for giving us the opportunity to carry out this research.

REFERENCES

[1] Itoh, H. et. al. : KBMS PHI(1)-Basic Configuration of Distributed Knowlwdge System-,Proc. Information Processing Society of Japan, March 1986 (in Japanese).

[2] Williams, R., et. al.: R*: An Overview of the Architectere, Proc. Int. Conf. on Database Systems, pp.1-27 (1982).

[3] Rothnie, J.B., et. al.: Introduction to System for Distributed Databases (SDD-1), ACM Trans. on Database Systems, Vol.5, No.1, March 1980.

[4] Epstein, R.S. and Stonebraker, M. :Analysis of Distributed Database Processing Strategies, Int. Conf. on Very Large Data Bases, October 1980.

[5] Demurjian, S.A. and Hsiao, D.K. : The Multi-Lingual Database System, Proc. Third Int. Conf. Data Engineering, pp.44-51 (1987).

[6] Katz, R.H. and Wong, E. : Decompiling CODASYL DML into Relational Queries, ACM Trans. on Database Systems, Vol.7, No.1, March 1982.

[7] Rosenberg, R.L. and Landers, T. : An Overview of MULTIBASE, Distributed Data Bases, H.-J. Schneider, ed., North-Holland Publishing Company, 1982.

[8] Smith, R.G. : A Framework for Distributed Problem Solving, UMI Research Press (1981).

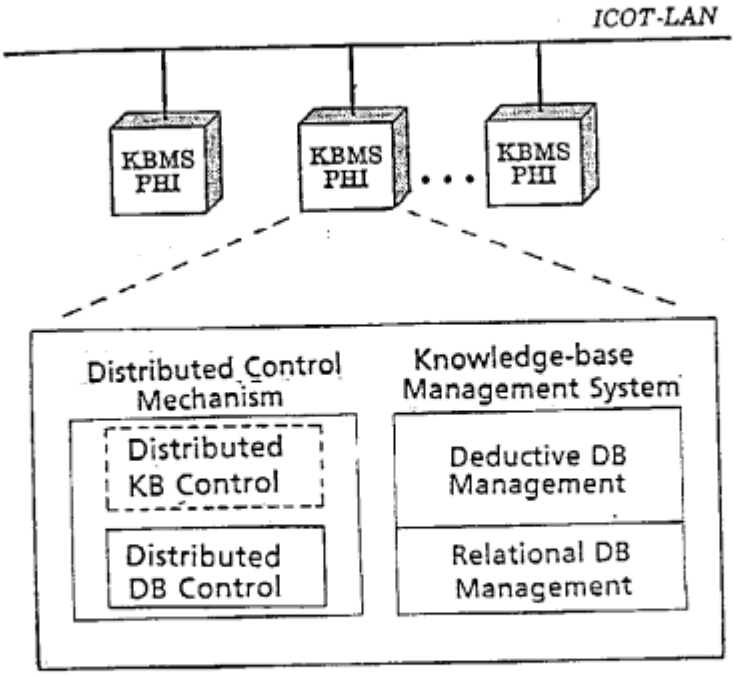

Fig. 1 Configuration of Distributed Knowledge-base System

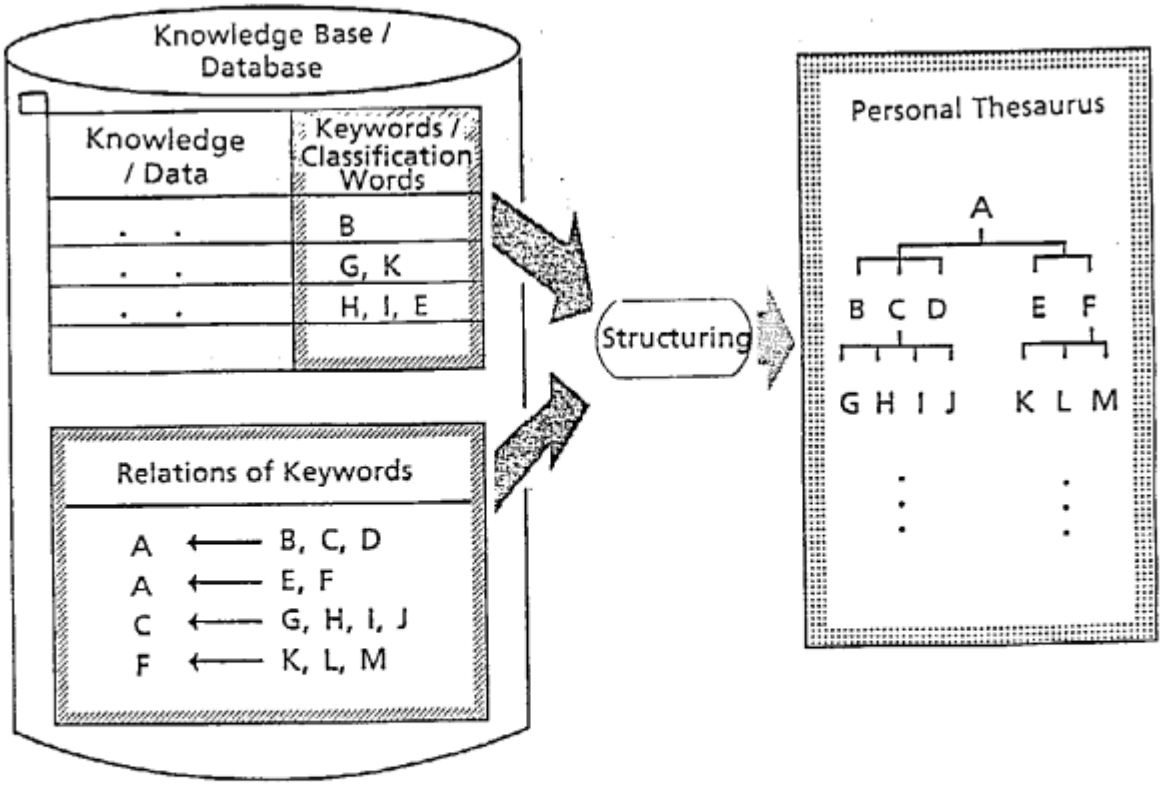

Fig. 2 Construction of Personal Thesaurus

PERSONAL THESAURUS 1

Artificial-Intelligence
- Artificial-Intelligence-Application
  - Common-Application-System
    - Diagnosis-Expert / Operation-Expert
    - Natural-Language-Processing
    - Speech-Recognition / Image-Recognition
    - Software-Development-Support
    - Computer-Aided-Education
  - Industrial-Application-System
    - Government / Banking / Industry / Physical-Distribution / Service / Medical
- Framework
  - Hardware
    - Main-Frame / Personal-Computer / Inference-Machine / Knowledge-base-Machine
  - AI-Building-Environment
    - AI-Basic-Language
      - LISP
      - PROLOG
    - Support-Software
    - Expertsystem-Building-Tool
      - "EXCORE"
        - Knowledge-Representation
          - Frame / Logic / Rule
        - Knowledge-Base
          - Factorial-Knowledge / Rule / Meta-Knowledge
        - Knowledge-Base-Management
        - Knowledge-Editor / Knowledge-Debugger
        - Working-Stage
        - Inference-Engine
        - Explanation-Mechanism
        - User-Interface
- System-Characteristic
  - Expansible
    - Stand-Alone-Type
      - AI-Building-Environment
      - Intellectual-Interface
    - Hybrid-Type
      - Program / DB-Reusable-Interface
  - Integrable
    - Network-Connection / Distribution / Object-Oriented
- AI-System-Construction-Method
  - Knowledge-Acquisition
  - Knowledge-Analysis · Arrangement
    - Consistency
  - Knowledge-Base-Structure-Design
  - Knowledge-Base-Creation
  - Development · Test

Fig. 3 Example of Personal Thesaurus (1)

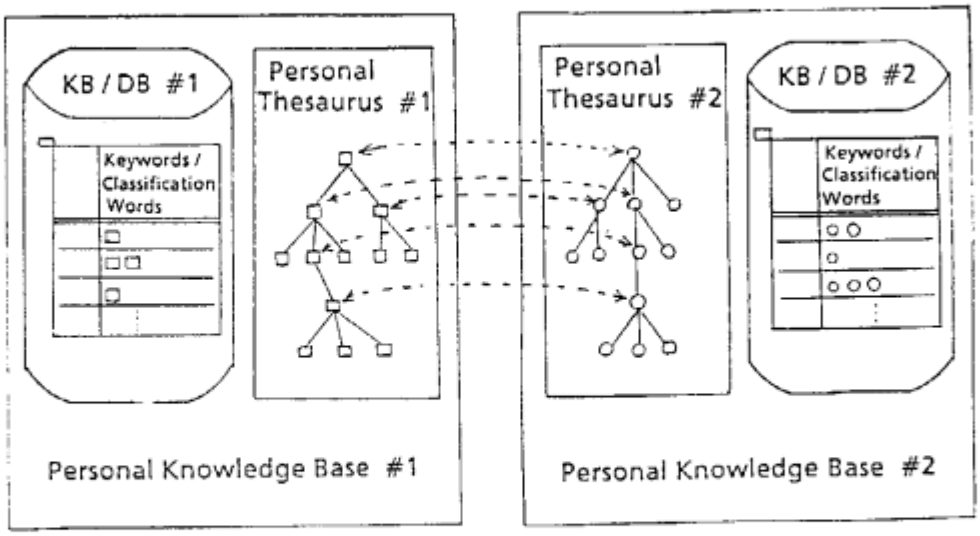

Fig. 4 Matching of Personal Thesaurus

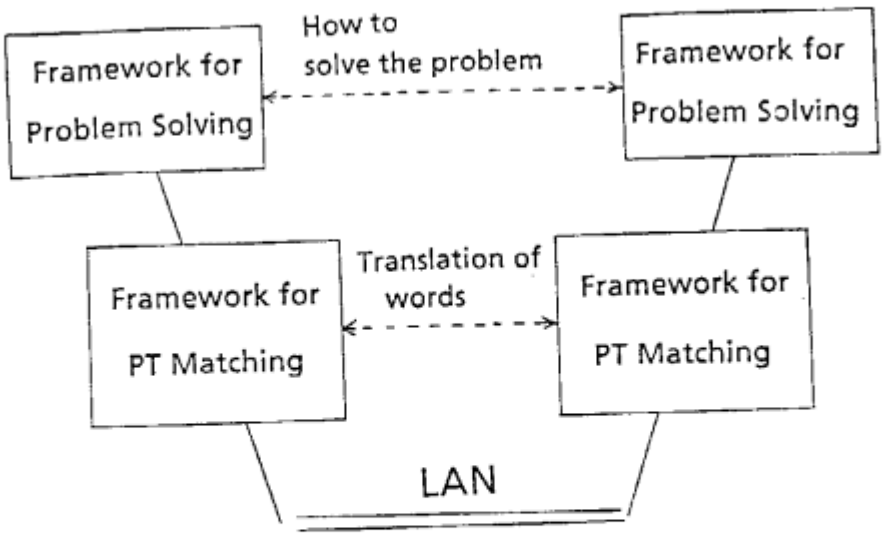

Fig. 5 Frameworks for Problem Solving and PT Matching

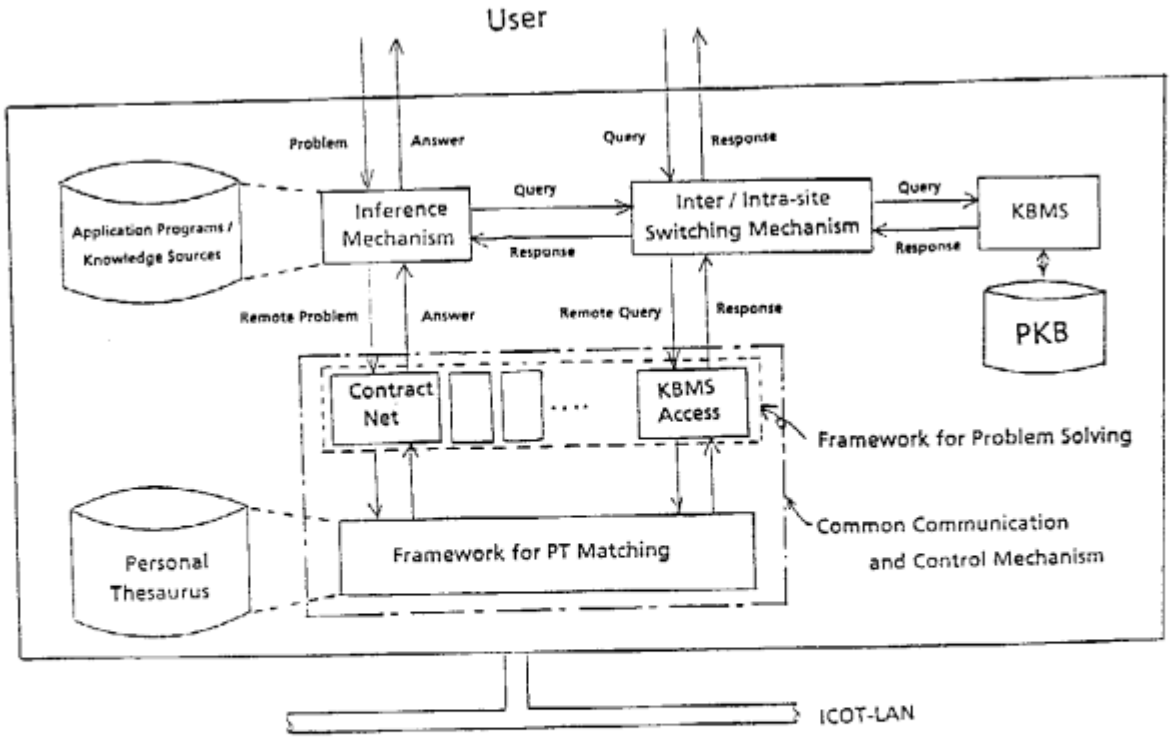

Fig. 6 Basic Configuration for Sharing PKBs

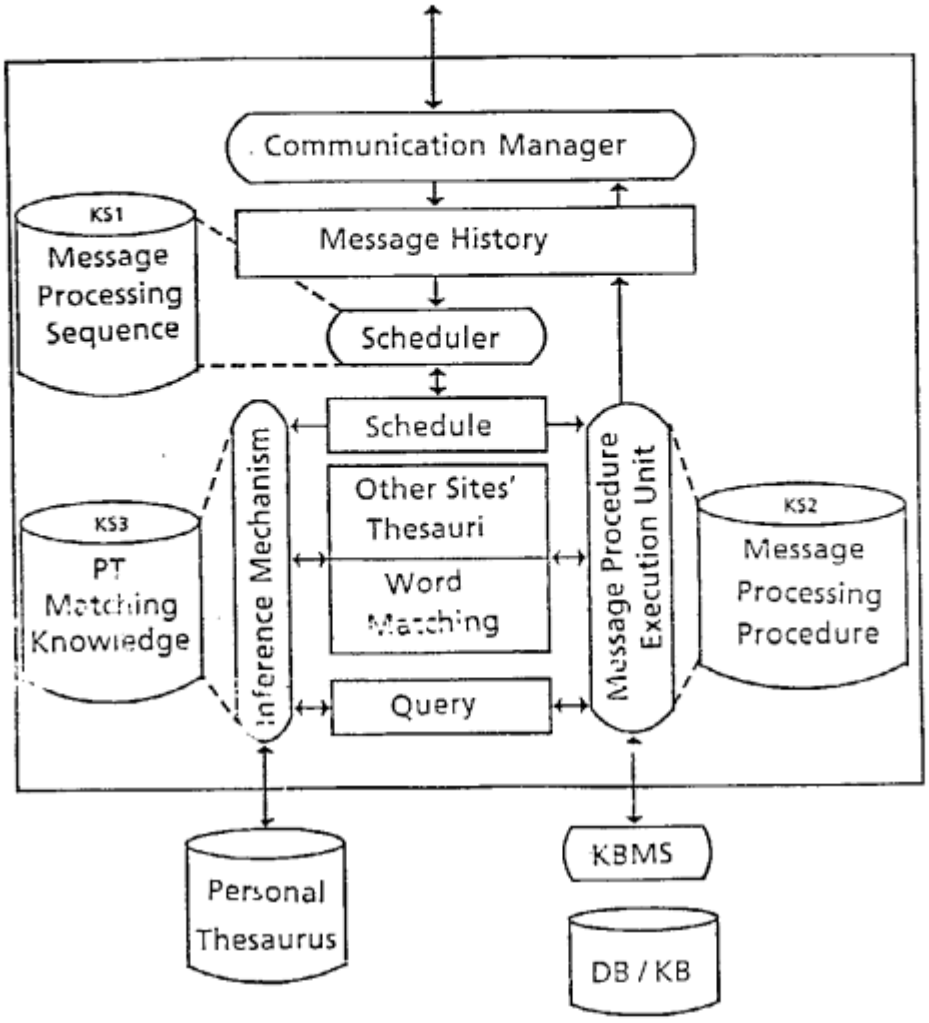

Fig. 7 Configuration of PT Matching Mechanism

PERSONAL THESAURUS 2

Subject-of-AI
- Practical
  - Integrated-AI-system-Building-Environment
    - Knowledge-Representation-Function
      - Meta-Knowledge
- Basic
  - Knowledge-base-Building
    - Knowledge-Acquisition-Support-System
      - Automated-Knowledge-Acquisition-System
  - Knowledge-base-Management
    - Completeness-of-Knowledgebase

Fig. 8 Example of Personal Thesaurus (2)