Document Research based on collaborative provided structural Knowledge

Harald Huber
USU Softwarehaus AG
Spitalhof
D-71693 Möglingen
Hhuber@usu.de

Abstract

Knowledge management has become a question of current interest. Indeed investigating in document bases is often seen as a core technology. In this report, the subject ‚Knowledge Management' is enforced and possibilities to use data processing systems are shown. The primary content describes the idea to provide structural knowledge collaboratively. This is illustrated by the example of a research structure based on semantic networks, which are tended by the users collaboratively.

1 Knowledge management - trend for the turn of the millenium

During the last two years the trend to knowledge management has encountered an euphoric upswing within the European countries. Considering this development independently from claims and interests which consulting companies and software manufacturers contribute to the evolution of such trends, it can be recognised in this evolution that knowledge has become a decisive competitive factor for the refined political economies of the world.

This perception is not new. Peter F. Drucker, the well-known management consultant, determining already in the 50s the expression “knowledge worker”, stated in an 1990 interview: ‚Most people do not consider themselves as part of an economic society anymore but rather as part of a knowledge society' [Rein90].

Conceding knowledge such high importance, however, compulsion results quite automatically to use this knowledge in its best way as well - and for this reason, knowledge management should be an important task, too, - if one only would know precisely, what knowledge management means.

The article ‚Wer weiß, was Wissen ist?'(Who knows what knowledge means?) published in the German weekly newspaper ‚Die Zeit' dated July,16,1998, gibes at this fuzziness of the discussion and German weekly magazine Computerwoche 09/98, too, asserts: , Catchword has career opportunities as a term but less chances as far as realisation within companies is concerned'.

This skepticism is certainly attributed to the circumstance that knowledge is simply something utmost human, something that we reluctantly concede a machine. In most of the conversations where we introduce knowledge management as an issue we are reminded by the company's interlocutors that artificial intelligence, deliberately provided with so many laurels in advance, has not achieved many of its targets too.

Therefore, it is important to define knowledge and knowledge management. We did this in a way that we defined knowledge as 'a perceptible action potential of an intelligent actor'. This means that we define as knowledge an actor's ability (a human being or an organisation) to fix a problem or reach a target. Admittedly, a differentiation between intelligence and knowledge within this definition is only hardly possible, but it has the advantage to express the connection between knowledge and economies very clearly.

In our opinion, knowledge management is ‚every conscious effort to develop, insert, preserve and represent knowledge within enterprises in a better way'. Therefore, knowledge management as a method or technique has many layers. Besides technology, organisational methods are used exactly like approaches influencing both the enterprise and learning culture.

In any case, use of technology always means it either owns knowledge itself and applies it (like fuzzy control systems, which steer blast furnaces or vacuum cleaners,) or supplies working individuals with the required knowledge. In both matters, knowledge has to be filed within the respective system.

In our opinion, there are three models (fig. 1) .

Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98)
Basel, Switzerland, 29-30 Oct. 1998, (U. Reimer, ed.)

http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-13/

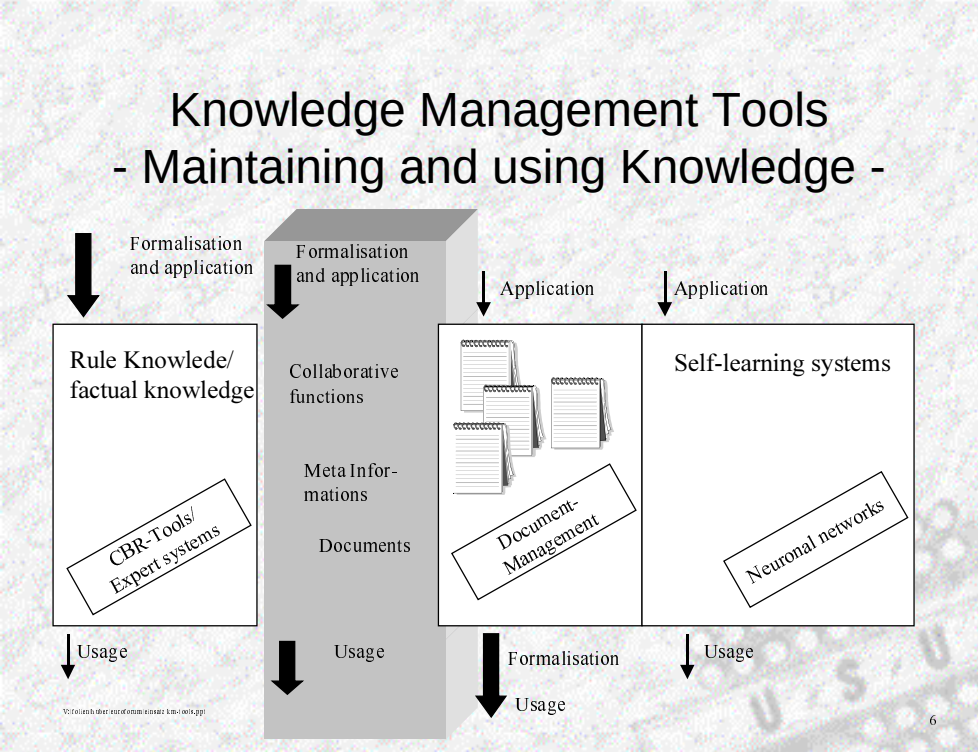

Fig. 1: Possibilities to deposit knowledge

One possibility consists of the fact that the system itself obtains required knowledge for itself which means, it is a self-learning system. Practitioners often overestimate this approach. It is proceeded on the assumption that a system very often has to analyse repair reports for example thus it will surely recognise which knowledge field-technicians will use[1].

In our opinion during the next 10 to 15 years, undoubtedly no system will be launched onto the market that is able to acquire knowledge from non-structured documents by itself.

Thereby only the choice of a different possibility remains, such as the knowledge formalisation and its insertion into a correspondingly formal system, for example a 'Case Based Reasoning Tool'.

Nowadays these systems are used quite successfully in technical systems diagnosis for example, and quite efficient medical diagnostics systems were shown as well.

Even these systems, however, have boundaries that simply result from the formalisation quantity that arises during knowledge deposition. Rule-based tools in help desks, supporting IT users remained quite unsuccessful. The reason is cost expenditure, which is required to deposit knowledge within the system. Additionally, it is simply not able to be provided economically and reasonably due to rapid systems development and its joined evolution of new mistakes and problems.

Herewith the only possibility remaining is to deposit knowledge in a form of documents, which is internal practice within companies. Thereby the formalisation amount is dropped. Nevertheless, expenditures to find required knowledge in a document base increase.

At last, two techniques are offered in the area of investigating documents in document bases. On one hand, a possibility to categorise documents in a way thus they can be found simply by means of key words. In this case, it is indifferent whether it is so-called "Catchwording" or genuine classification. On the other hand, it is a full-text search possibility. Both prove to be only few effective in usage.

Classification by categories or ‚Catchwording' of document bases is challenged by two problems:

- Users will not accept a structural classification as it represents an additional expense in a stage in which they have to fight with motivation problems anyway. It is difficult enough already to

[1] There are some systems in developement, which are demonstrating the possibilities and limits (e.g. [Hah98], [App93])

motivate employees to deposit their knowledge in one of these document bases. If learning a classification system is additionally required one has to be concerned about catchwords that have been employed, while storing a document - it often leads to renouncing to deposit documents. Usually maintenance of a classified document base only performs if at least one employee is responsible for editorial advising of the document base.

- The second problem is visible in the low flexibility of such a classification. It is not possible to simply re-work a classification completely if more than 1,000 documents are supervised in one document base already. The risk, however, exists in a way that this classification ages rapidly and does not contain current terms for example. Furthermore, its acceptance decreases.

Full text search, too, has to struggle with difficulties. Additionally, it presupposes a high standard of the user‘s knowledge and this is not only for the service of the search machine, for example like full-text search syntax.

Nevertheless, knowledge of typical terminology, usual terms and also homonyms and synonyms is required. Most of the times this knowledge cannot be expected from the majority of users although there might be expertise-owners in special fields.

2 Structural, collaboratively well-kept knowledge as a solution approach

Since 1996, USU AG are engaged in the of knowledge management subject. In order to address the problems stated above we defined two basic ideas:

We attempt to combine knowledge, motivation and functional business process with each other. That means that we try to form functional business processes or tasks in a way thus it means a personal advantage for the employee to deposit his knowledge in a system. We do not only try to attempt to motivate the employees for a medium-term period via incentives or similar structures but motivate them to benefit from depositing knowledge.

For this reason, an own analysing method was developed which allows construction of such business processes saving the knowledge of employees.

The second idea is by far more complex. In order to explain this approach two examples should be illustrated at first:

- As mentioned above classified structures age rapidly. The reason for this can be that many persons maintain documents themselves. However, on the contrary only a central administrator keeps up classified structures in most cases. As the latter is mostly unable to keep up with the documents growth and to analyse them on new topics or approaches, classified structures simply have to age as the development of documents maintenance expenditure, well-kept by many users, does not keep up.

- Looking at a user who is investigating in the Internet for any subject. After receiving a report of 10000 hits from the respective search machine, the user keeps on optimising his search-string until the number of responses is small enough in order to be analysed at an available time. After having found, the documents he was looking for, actually two working results were achieved. On the one hand, the found document and on the other hand, research knowledge, very pictorial in a form of full-text search-string. A document is kept, printed for example, but the search-string, however, gets lost.

The second basic idea consists of the fact that we try having the structural knowledge not accomplished by one administrator only but by many users. Structural knowledge for example is understood as categorising or any other semantically significant structure which knowledge represents around structures of reality or included documents.

Up to the present USU AG have realised this approach into two basic functions, first in process modeling and second in text search. The case of text search shall be demonstrated here as an example of this technology idea. Within its tool ’Value Base’ USU AG have realised basic ideas described here as well as the technology of search inquiries in semantic networks, which will be outlined afterwards. Value Base is based upon Lotus Notes and is offered on the market by USU AG and during 1999 it will also be offered based on Intranet technologies.

2.1 Description of Technique

Text search consists of the following basic modules:

- One document base with full-text search containing documents, that are to be investigated.

- A surface which is able to represent a semantically significant structure.

- Moreover, a logic that connects this semantically significant structure with the documents.

We decided to use a semantic network [Rei91] for representing structural knowledge[2].

In this case, edges and nodes play a double part. Certainly, they represent knowledge around structures of reality or documents but simultaneously they include a semantic network connection with the documents. For this purpose, a full-text search-string is assigned to every node and a rule to each edge in order to connect the respective nodes.

This can be seen in the following example (fig. 2):

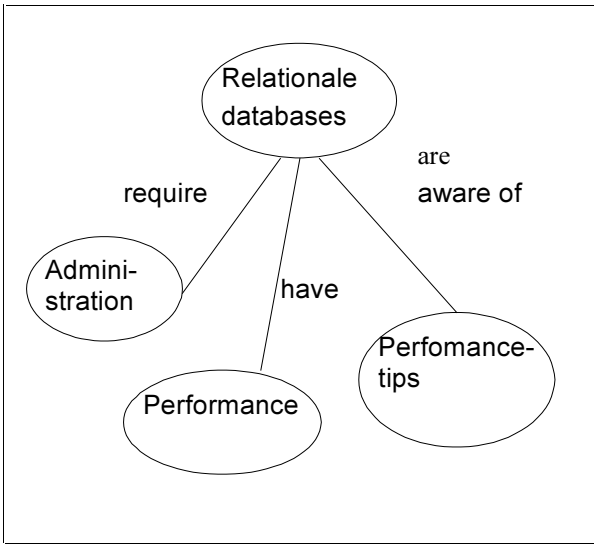

Fig. 2: Net semantic example

The above-presented semantic network can be assigned to the following search elements (the searching example employs a kind of syntax that is feasible within Alta Vista):

Administration: {(Systems* near management) or administrat* or monito*}

Require: AND

Have: AND

Relational databases: {(Relational near data base) or (data base near table) or (relational) or (DB/2 or....)}

Performance: {Performa* or Tuning or Throughpu*}

[2] RDF [RDF97] uses a similar approach, but the binding between structure and documents is based on tags and therefore requires administrational effort

Of course, an edge is able to include quite more complex combination rules than simple AND-rules as shown here.

Hence, semantic networks describe the structure of subordinate documents. If a user would like to investigate, he employs structural knowledge, deposited in the network; in order to make his interrogation available. This means that he selects from the interrogations of the network, which would fit the formulation of his question.

If a user himself for example is interested in relational data base administration, he selects both of the two appropriate nodes and edge accomplishing the generated Query.

The advantage of this technology certainly consists of the fact that a user encounters a great number of reusable interrogations and that users, having only low research know-how are able to investigate considerably better with such prepared structures than without it.

The most essential advantage, however, is that a high degree of reusability will be achieved. Supposing that a user indeed is searching for administration hints for relational databases. For this purpose, he will combine a node called ‚Relational Databases' and ‚Administration' via the corresponding edge but, his investigation is not successful.

Therefore, he will examine the search-string of ‚relational databases' and will detect that the Microsoft SQL-server is not contained within the relational database listing, yet. He then will accomplish this search-string accordingly and, in fact, might be able to find the demanded document.

Minutes later, a different user is searching for hints to optimise relational data base performance.

This one will pick up a different combination of nodes and edges and will combine‚relational databases' and ‚performance'. The interesting issue herewith is, that the second employee is able to refer to the research knowledge of the first one as also the second is employing the extended search-string that considers the Microsoft SQL-server.

By this example, the basic idea of technology can be seen, but on the other hand the combination between

motivation and knowledge, is shown too. During his text storage activities, the user is not motivated to integrate the new database ‚SQL-server', that possibly is contained in his text, into the catchword system as well. In the case of research, however, this required motivation is available and exactly this motivation can be used.

3 Practice example

At an IT department of a large insurance company, supplying necessary security is a non-trivial problem. It is true that the practice-relevant basics of safety technology are not necessarily difficult to learn. The company‘s internal practice, however, is bristled with a vast quantity of details that cause difficulties. At the announcement time for a new virus for specific operating systems, in a very large enterprise it is already difficult enough to find, whether and where a specific operating system is used at all. Definitely it becomes much more difficult if legal norms change or new hacker technologies are reported. Therefore, it is essential to determine the extent of danger for one’s own data processing division.

Even broad IT users rarely have more than one safety agent who would be able to turn 100% to this subject. As this employee of course is not able to control neither the complete infrastructure nor the applied technologies and the existing legal regulations totally, he is always depended on the support of the disciplines. These again do not have enough time to dedicate themselves

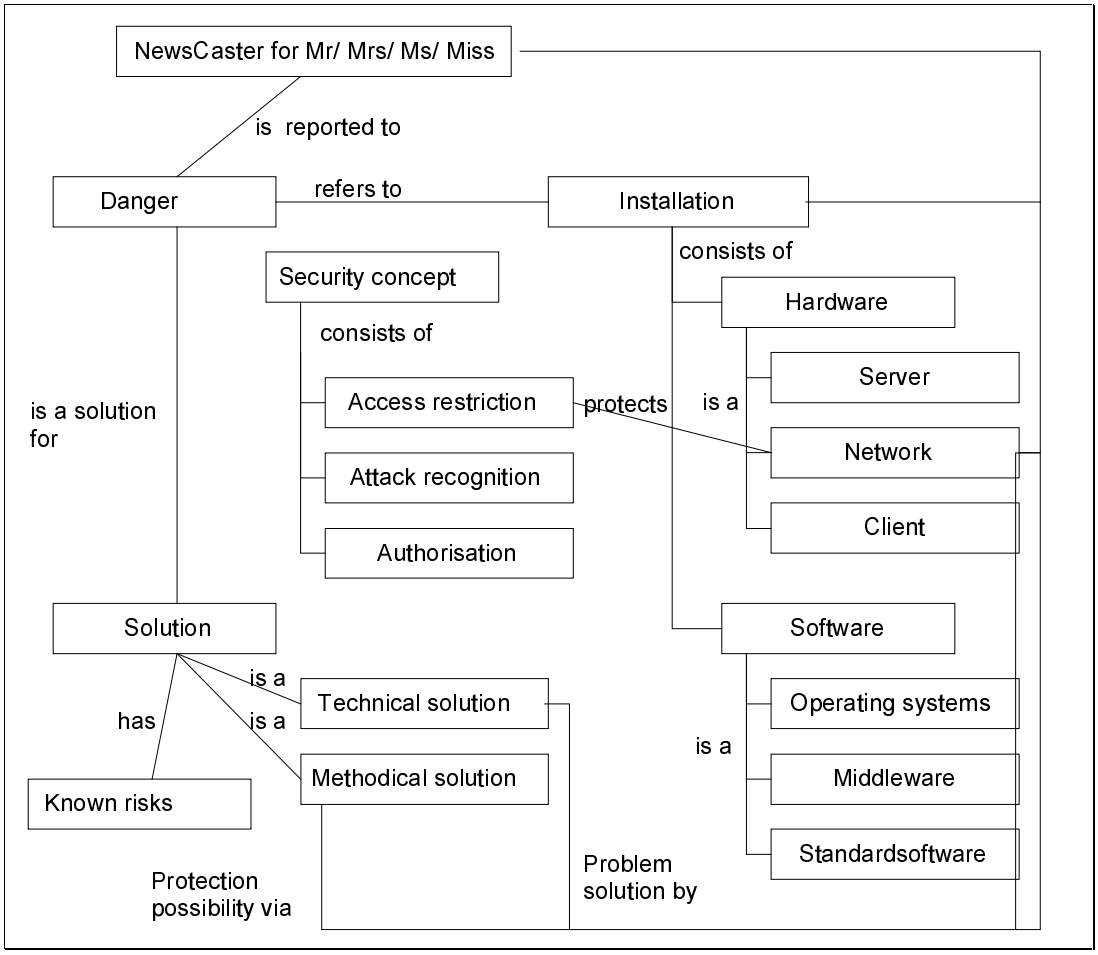

Fig. 3: Semantic network ‘Security’

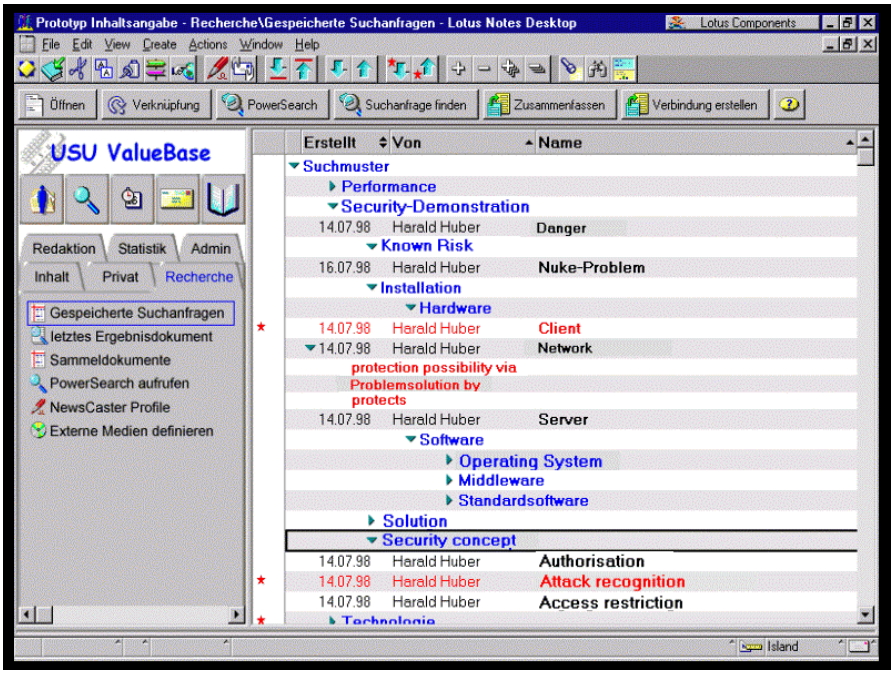

Fig.4: Semantic Network in ValueBase

wholly to this subject alone and therefore, they are often only available after a certain delay for information. Exemplary, the process of determining a special risk will be accomplished in a way that the IT manager for instance found an article describing an attack upon the CIA home page, adding Mickey Mouse' ears to this site. Naturally, he immediately will ask his safety agent whether something like this could happen in his own company as well. As he is not familiar with something of that kind, the agent has to contact the experts.

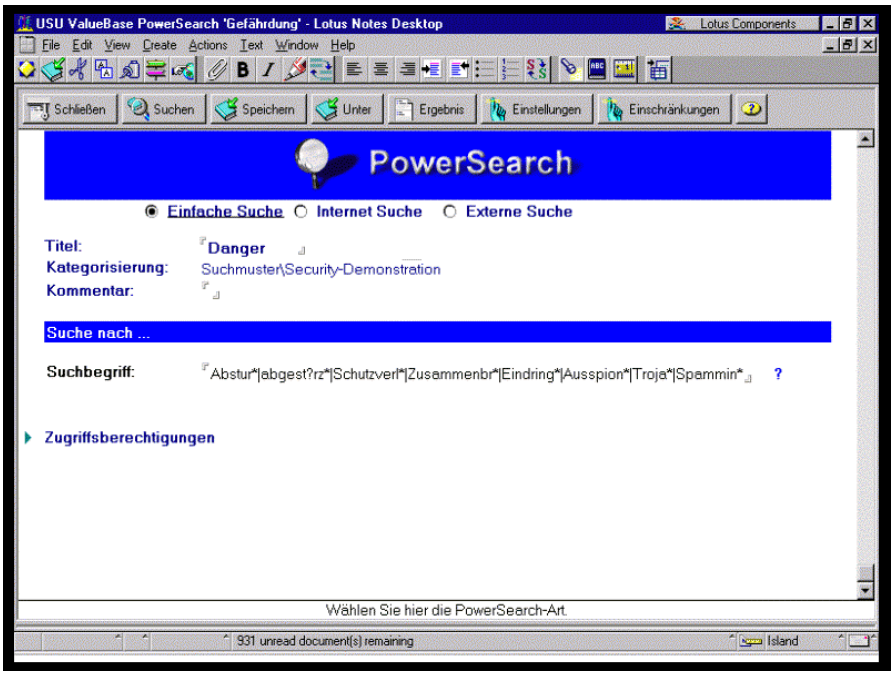

Fig. 5: Node 'danger' as full-text search-string

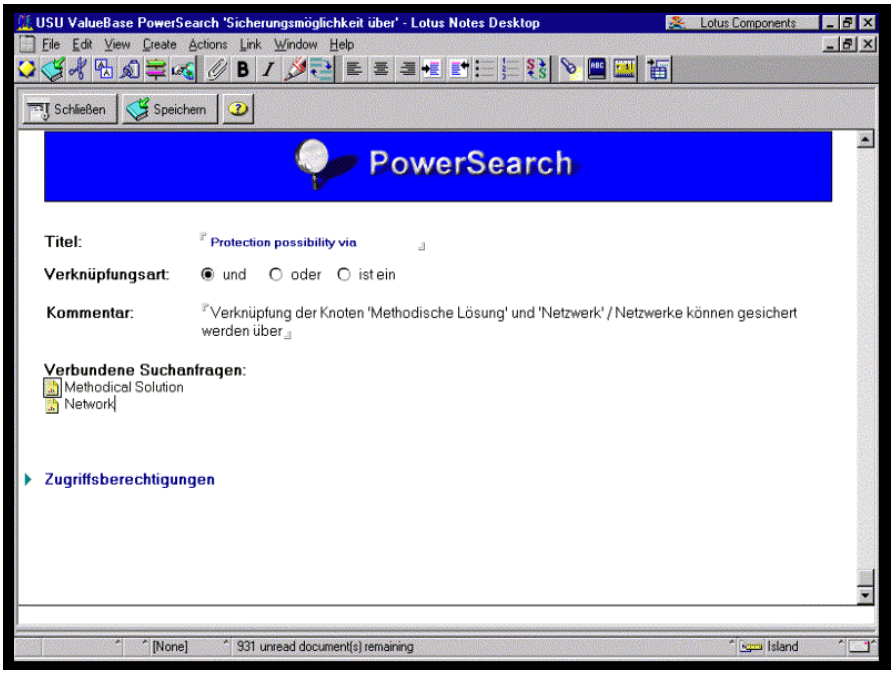

Fig. 6: Edge between ‘methodical solution’ and ‘network’

At a first step, it is not simple to find out who is well versed in this area. Indeed, at first, he will ask the colleagues who provided the own home page, but these inform him that the firewall administrators are responsible for the safety of such attacks. These in turn report to him that safety in this sense might be indeed a task of the firewall. However, in the reported case it would be likely that the authorisation structure and the general administration on the Internet server could have been the problem. Hence, again he will be with the colleagues who look after the home page.

Having this information, the latter will send him further to the system administration of the servers

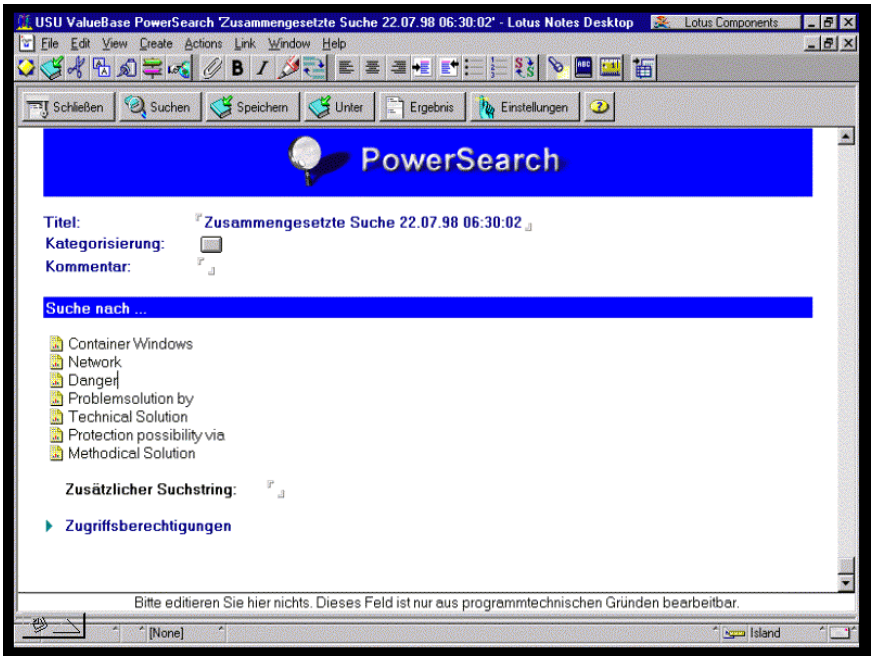

Fig. 7: Selected search inquiries

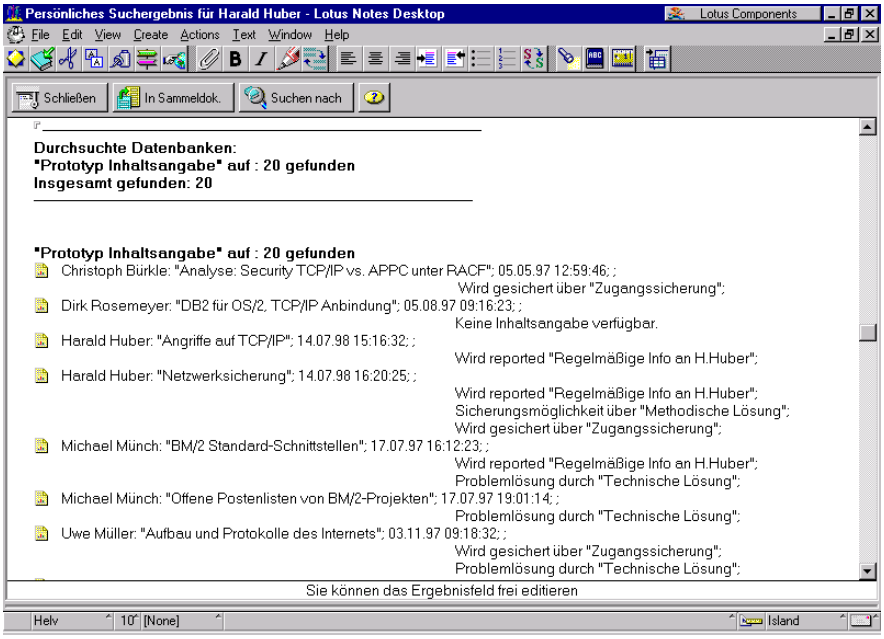

Fig. 8: Search result with summary

who take care for the operating system there. Possibly, he may find a competent contact person there.

A quite realistic scenario! In addition, formulating this kind of questions will occur repeatedly every three to six months and have to be answered laborious again, as nobody anymore will be able to remember problems that already have been solved due to the great number of cases. Therefore, it is reasonable to create feasibility with recourse to a document base in order to examine problems and risks, which already have been solved and now can be relieved from judging evaluated risks once again.

For this purpose, a semantic model has been provided by us which roughly represents the connections of the security management (fig. 3).

(These sections are descended from a demonstration project and, therefore, are not complete. Nevertheless to the author's opinion it will satisfy the purpose for supplying the principle).

The network depicted here is represented through ValueBase by several hierarchy trees within ValueBase due to the userinterface-functions available in Lotus Notes but also following the user's behaviour. (fig. 4).

Semantic network nodes are represented as documents (i.e. "Danger" or "Network") and the hierarchy is realised by putting the documents in Lotus Notes folders on several levels. Nodes (fig. 5) consist of the node-name ('Titel') which gives the node a semantic meaning and a fultext-searchstring ("Suchbegriff". "*"and "?" are wildcards, "|" means logical OR).

Nodes (search inquiries) are connected via edges, which represent combination rules and relationship respectively (fig. 6).

In ValueBase, these edges are realised using Lotus Notes Documents as well. These edges consist of their own name ("Titel"), the combination rule ("Verknüpfungsart", which could be "Und" (Logical AND), "Oder" (Logical OR) und "ist ein" (Is A)) and document links to the mentioned documents which are connected by this edge (i.e. "Methodical Solution" and "Network").

To research the database, the user has to pick up the section of the network, that is relevant for individual question forming.

To search all documents with information about danger (in a security management sense), the user picks up the node 'danger'.

To search information about danger in networks, the user can pick up the nodes "danger" and "network". To define an explicit combination-rule, it was possible to select an edge. If no edge is defined, a default-rule was used.

Let us make an example. If an employee reads in the newspaper that Microsoft operating systems can only be used in a network by taking a high risk, he would try to find all the relevant information about MS-operating systems and security-problems. He however wants to do everything which is necessary to optimise the network, therefor he is interested in possible solutions.

So the node MS-operating systems, network, technical and methodical solutions and the additional edges where picked up (fig. 7).

3.1 Former Experiences

Former experiences show that expectations linked with this technology will come true in fact. At least in smaller groups of about 20 to 30 people search inquiries will develop further more via the research activity. Actually, the probability to find documents increases subjectively – substantial tests have yet to come.

The willingness of the users to provide an efficient search inquiry increases perceptibly as the reusability becomes clear to them very quickly

4 Further developments

Nowadays system is not a very high sophisticated one, but due to our opinion the system uses very interesting basic ideas:

- the representation of the structure of the documents via a semantic network
- and the binding of this network with the documents via a fulltext-searchstring.

The combination of that two ideas represents an utmost powerful base.

It gives us the ability to use knowledge in the network in order to interpret the documents as you can see in fig. 8.

In this example we are using knowledge in the network in order to generate small summaries of the document's content.

This functionality has already been realised as a prototype and will become generally available by the next release.

A subsequent development step consists of the fact that we automatically improve the semantic structure network due to information contained in the semantic network and statistical data on the occurrence of specific catchwords.

Supposing that the following network construction (fig. 9) exists.

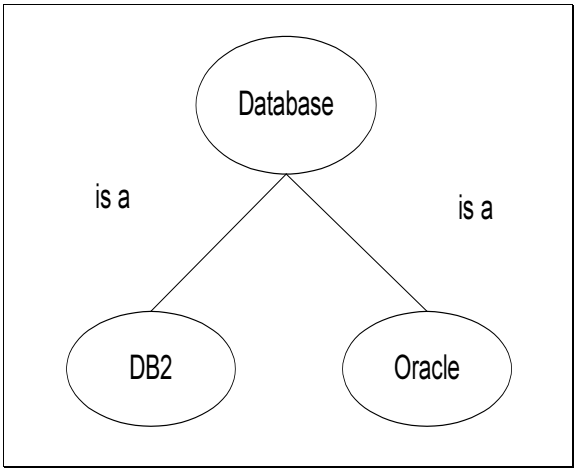

Fig. 9: Example for a further semantic network development

According to the above mentioned figure, the system should formulate the following proposition after having located the term 'Sybase':

"Sybase is a database"

5 Final word

To our opinion, collaborative of structural knowledge development is a future technology. Limited contemporary systems ability to identify connections and similarities makes the human being necessary as an essential element. On the contrary, the machine is able to offer a provision of a structure, which allows individuals to cooperate also on the structural knowledge level. Further, use of motivation by system functionality will be added.

This report represents only an example of the basic idea; nevertheless, it also shows the evolutional potentials of this technology if it is taken as an independent element.

References

[Rein90] E. Reingold. Facing the 'New and Dynamic'. Interview with Peter F. Drucker in TIME Magazin, Jan.22 1990

[App93] D. E. Appelt, J. R. Hobbs, J. Bear, D. Israel, M. Tyson. FASTUS: a finite-state processor for information extraction from real-world text. *IJCAI'93 – Proc. Of the 13th International Joint Conference on Artificial Intelligence*. Vol. 2 (pp 1172-1178). Chambery, France, August 28 – September 3, 1993. San Mateo, CA: Morgan Kaufmann.

[Hah98] U. Hahn, M. Romacker. Automatische Erzeugung von medizinischen Wissensbasen durch partielles Textverstehen: Befundtextanalyse im MEDSYNDIKATE-System. *KI Künstliche Intelligenz*, June 1998 (pp 65 – 72), Bad Ems, Germany. ScienTec Publishing

[RDF97] World Wide Web Consortium. Resource Description Framework (RDF) Model and Syntax. Version 1, 02-Oct-1997. http://www.w3.org/TR/WD-rdf-syntax-971002

[Rei91] U. Reimer. Einführung in die Wissensrepräsentation. Stuttgart, Germany. Teubner Verlag 1991