A Cooperative Comprehension Assistant for Intranet-Based Information Environments

Ludger van Elst and Franz Schmalhofer

German Research Center for Artificial Intelligence (DFKI),
University Bldg. 57, Erwin-Schroedinger-Str., D-67663 Kaiserslautern
{elst, schmalho}@dfki.uni-kl.de

Abstract. A cooperative comprehension-assistant is described which represents text documents at three different levels of abstraction (surface, propositional and situational levels of representation). At the *surface level* texts are represented as bags of words without any linguistic structure. For the *propositional level* a definite but evolutionary growing set of predicates and concepts is used for selectively representing the most interesting parts of each text. By a latent semantic analysis the *situational level* is calculated as the third level, which is a parsimonious representation of the complete contents of a document with respect to some previously established frame of reference (metric vector space). Together the three levels provide more flexibility as well as precision for assisting human comprehension, document storage, retrieval and distribution. Furthermore, it is shown how this comprehension assistant is applied for knowledge dissemination in a particular kind of intranet-based information environment (oligo-agent system). The individual and social components of knowledge sharing within such an environment are described by knowledge construction and knowledge integration processes. To take the problem of long-term maintenance into account, the comprehension assistant and the oligo-agent system is embedded in a life-cycle model of cooperative knowledge evolution, consisting of a seeding phase, the system's evolutionary growth and periodical reseedings.

1 Introduction

The successes of the internet as well as intranets have been tremendous. There are now about 320 million pages in the World Wide Web. On any given day a powerful search engine cannot even identify all these pages. Only a relatively small portion (6 million) of these pages can be identified within a day. With intranets there is the similar problem, that it is often quite difficult to find the most relevant knowledge for a particular task at hand. The large amount of electronically available information in combination with the continuously occurring changes of the stored information make the task of finding the best document available quite difficult.

Anyone who has frequently used the search engines on the internet has had the experience that the retrieved information did not satisfy his particular needs although the desired information was indeed available in the net: The retrieved information is often redundant and some portion of it may be quite irrelevant with respect to the particular query. In order to alleviate this problem we have developed a cooperative comprehension assistant for a particular type of multi-agent information system which was termed oligo-agent system [15]. In this paper we describe the comprehension assistant which uses multiple levels of representations as has been suggested by theories of human cognition. Thereafter it will be shown how this comprehension assistant can be utilized for the storage, distribution and retrieval of information in an intranet-based information environment. Finally, a summary will be presented.

2 A Cooperative Comprehension Assistant

2.1 Knowledge Construction and Integration for Document Comprehension

It is well known that one of the core tasks for systems of multiple agents (cf. [1]) is to establish a shared context to allow for collaborative problem solving [9]. This is especially true for human-centered approaches to knowledge management where the information landscapes consist of human *and* artificial agents. Here, a mutual understanding of the structure and contents of the organizational memory is central to enable the different actors to take over their specific responsibilities and achieve an overall performance that is sufficient for the respective needs of the multiple participants [11]. We therefore employ a comprehension assistant within the social coordination region [15]. This assistant was developed according to the principles of a cognitive theory of text comprehension, namely the CI-theory [10], and yields knowledge representations of texts that are therefore relatively similar to the knowledge representations that humans form when they read a text. Hence they are well suited for establishing mutual understanding.

To take the problem of long-term maintenance into consideration, this approach is embedded in a life-cycle model of cooperative knowledge evolution, consisting of a *seeding phase*, the systems *evolutionary growth* and periodical *reseeding* [8].

As it is shown in Figure 1, two types of processes are employed to establish a shared context among the information providers and the information consumers by the comprehension assistant. Whereas the various *knowledge construction processes* are quite general and therefore well suited for all the different participants (perhaps leading to redundant or inconsistent portions of information in some participants' views or with respect to specific purposes), the *knowledge integration processes* yield a more coherent view and are therefore more closely tied to an individual user. The construction processes are based on various specific assumptions for the representation of the documents. This frame of reference is negotiated between the participants in the seeding phase and should stay quite constant during the evolutionary growth where more and more users and documents participate in the system while other documents may be deleted and some users may no longer participate. After a while, the chosen frame of reference,

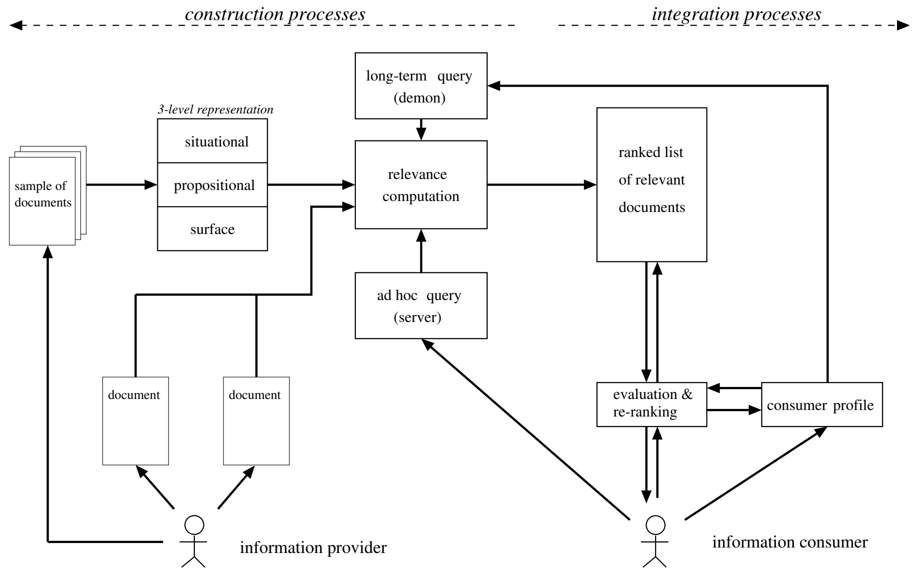

Fig. 1. Knowledge construction and knowledge integration processes for three-level representations as a means for accomplishing mutual understandings among information providers and information consumers

i.e. the "blueprint" of the organizational memory, may no longer be appropriate and the new representational assumptions must be negotiated (reseeding). Next we will describe the three-level representations of text documents with their respective construction processes in more detail.

2.2 Three-Level Representations

The most important task of the comprehension assistant is to represent text documents on a) a surface, b) a propositional and c) a situational level. While the surface level is quite word-oriented, the two other levels utilize more abstract representation spaces (discrete spaces for the propositional level and continuous spaces for the situational level).

surface level: On this level, documents are seen as "bags of words". Hereby we enforce an abstraction from the linguistic structure of the text as well as from concrete semantics of the words. Technically speaking, the representation space is a highly dimensional vector space where the dimensions are all the words that occur in all documents. Documents and queries are vectors in this space and the relevance of a document with respect to a specific query is approximated by the similarity of the two vectors (e.g. the cosine). Most search engines that are in use nowadays are based on this vector space model (e.g. [13]) and the problems like low precision and recall values due to polysemy or synonymy are well known. However, the method works very fast and can easily be understood by novice users as it doesn't require a deeper theory. Furthermore, a good portion of practical problems can be solved with the assistance of this level, as the application of WWW search engines shows. The only arrangements that have to be made in the seeding phase is the specification of parameters like stop word lists, the concrete weighting schema and similarity function. Much past research has be done in this area [5].

propositional level: This level of representation abstracts from the specific words to diminish some of the problems of the surface level. On this level, documents are represented in a discrete representation space that is based on a fixed vocabulary which has to be negotiated in the seeding phase. On the basis of a pre-defined list of propositions which may consist of concept terms and/or predicate argument terms, each sentence will be searched for these specific propositions. Technically speaking, this amounts to searching for specific terms or combinations of terms in a sentence. Thereby synonyms will be matched to the same concept or proposition. Such propositions are called CL-Propositions which stands for "controlled language propositions". These CL-Propositions have the following basic properties:

- CL-Propositions are predicates that contain no variables.
- Predicates must be elements of a controlled language, i.e. elements of an emerging ontology.
- The terms of the controlled language as well as strings can be used as arguments.
- CL-Propositions can be related to various segments of a document, e.g. sentences or paragraphs. Here, the choice of the controlled language is pivotal.

Basically, CL-propositions are normal forms of selected sentences. By the application of some prototypes of CL-propositions and a thesaurus that maps the words of the documents to the controlled language selected sentences may be represented in this controlled language. The number of prototypical CL-propositions may be at first small, but with the use of the system it may incrementally grow to larger numbers (for a more detailed description see [7]).

situational level: The situational level of a document will be generated as a vector in a latent semantic representation space [12]. Alternative situation models of a text may thereby be constructed by using alternative semantic spaces. For example, a text which describes a computer science department of an American university may be represented in the space of academic department descriptions of all academic disciplines in the whole world or in the more detailed representation space of computer science departments in North America. The selection of a representation space thus determines the scope and the frame of reference [4] for the interpretation of a text.

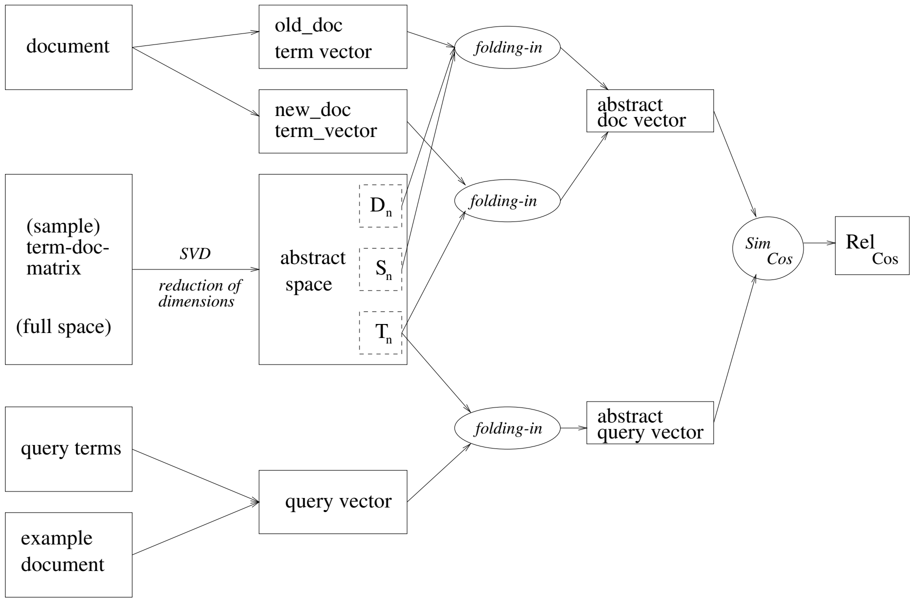

Fig. 2. Technical specification of the construction processes at the situation level

Figure 2 summarizes the technical description of the situation level. Of central importance for the situational level is the abstract space which is shown in the center of the figure. This abstract space is obtained from the standard term-document matrix [13] by a singular value decomposition [6] which results in a reduced but still high dimensional semantic space that is represented by three matrices (D, S, T). These matrices contain the row and the column entities (matrices T and D) of the original term-document matrix as well as scaling values (diagonal matrix S) in such a way that by multiplying D, S and T the original matrix is approximated. Here, an abstraction is achieved by the reduction of dimensions. As is shown at the top region of Figure 2, old as well as new documents can subsequently by represented by abstract document vectors (*folding in*). As is shown in the bottom region of Figure 2, any user query (in the form of some sequence of search terms of the length i, $1 \leq \mathrm{i} \leq \mathrm{m}$, or in the form of some example document) may also be folded into the abstract representation space and thereby yield an abstract query vector. For any given query, the most relevant documents (on the basis of the abstract semantic space) may now be determined by some similarity measure (e.g. Sim_{Cos}, the angle between the abstract document vectors and the abstract query vector) which is defined in the abstract space. The three levels thus allow a user to search for documents in a way that is adjusted to his prior domain knowledge.

We will now explain the usefulness of the three levels of representations by an example: Suppose, there is an intranet with a web-page for each computer science course that has been offered by any instructor worldwide. Furthermore, there are requests by a large number of international students for acquiring some specific computer science training. As the terminology is quite varied across different countries, a broker that uses only the surface forms of the course descriptions and the student request will not necessarily find the best match. For example, in some contexts, computer science, informatics, and electrical engineering may indeed be treated as synonyms. (When a university does not have a computer science department, a true computer science course may be offered by the department of electrical engineering).

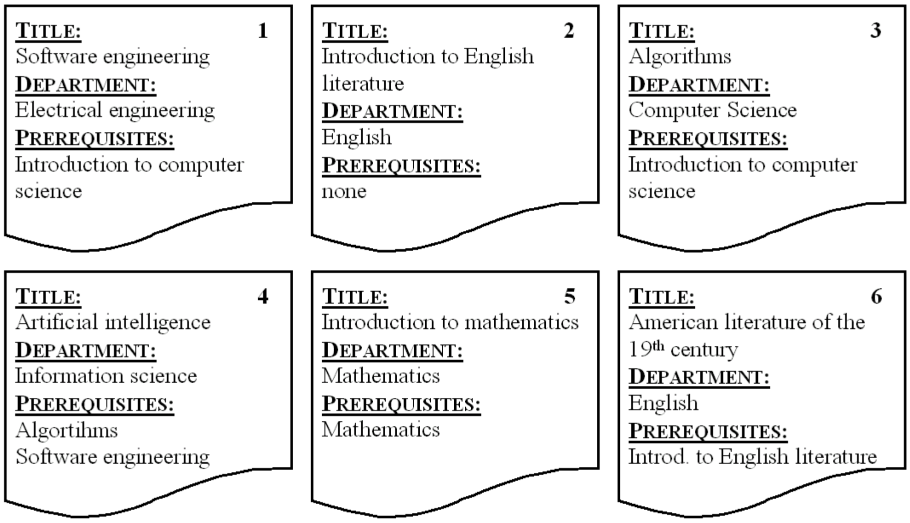

Fig. 3. Six hypothetical web-pages about academic classes

As a minute example, imagine that there are n=6 courses offered, where each course is described by a web page, as is shown in Figure 3. Furthermore, there are high school graduates who are interested in studying artificial intelligence and do not yet have much knowledge about this field. When using a search engine with the search key "artificial intelligence" they would find only page 4. When using the search key "computer science", they would find page 3. With such a surface oriented search, the users are literally getting what they are requesting.

With the three-level representation and the described comprehension assistant, on the other hand, these users may find what they really wanted, even when they are not knowledgeable enough to explicitly request it. In the information environment of the oligo-agent system for course brokering, the propositional level would encompass concepts like "engineering", "humanities" and "natural sciences". At the propositional level, "electrical engineering", "computer science" and "information science" would therefore all be mapped onto "engineering". In the latent semantic space of the situational level, all web-pages about science courses would be represented by vectors in the same vicinity of the metric vector space. Web pages about humanity courses would similarly build a cluster, however, at a quite different location of the latent semantic space.

With the comprehension assistant, the high school graduates would again start out by using "artificial intelligence" as search key. With the retrieved page 4, there would also follow the propositional unit "engineering", which would subsequently retrieve the pages 1 and 3, but not page 5 (nor pages 2 and 6), because mathematics is not mapped onto "engineering" at the propositional level. However, when the search is extended to the situational level, all pages about science and engineering are found (1, 3, 4, 5), because with the latent semantic analysis these pages cluster in the same region of the multidimensional space, just like all humanity courses cluster in some other region of the space of university courses. Courses 2 and 6 are therefore not retrieved at the situational level.

In summary, if a user has little or no knowledge of what he is really searching for, he may use some key terms and the surface level is then applied as in any other search engine. As retrieval result he will, however, also obtain the propositional and situational representations of those documents as well. In subsequent searches he can now use those propositional or situational levels of the documents that most closely correspond to his interests as a search key. The search for relevant text documents over time thus advances from a superficial level (surface search) to a deeper level of understanding (latent semantic space) and thereby to a mutual understanding between user and system representations.

Figure 1 also shows how the comprehension assistant can function as a tool for mutual understanding and cooperative comprehension among information providers and information consumers. This is accomplished by the various knowledge construction and knowledge integration processes which are depicted in Figure 1.

The different roles of the participating agents can lead to quite conflictive goals. For example can the structure of an information archive that is "optimal" for the librarian be very different from the structure that is "optimal" for the distributors. Therefore, the various frames of reference (or representational assumptions) that build the basis of the comprehension assistant must be negotiated between the agents. An initial coordination of the individual and the global concerns of the different users is already achieved at the definition time of all potential communications within the intranet. This coordination is accomplished by a more or less representative sample of documents from which a three-level representation and the associated relevance scores are established. Information-identification is initiated by long-term queries as well as ad hoc queries. The most relevant information (i.e. the right documents) become explicitly specified (information identification time). Integration processes [7] form consumer-centered views (e.g. individual re-ranking) and allow an information consumer to converse about his individual relevance rankings of the various selected documents on the basis of the three-level representation (cf. [16]). Figure 4 shows the particular interface which is used for conversing about the individual relevance of documents. Through the various sliders and buttons, the parameters of the integration processes (e.g. the weighting of different topics or the desired level of abstraction) can easily be manipulated and the effects can be seen immediately.

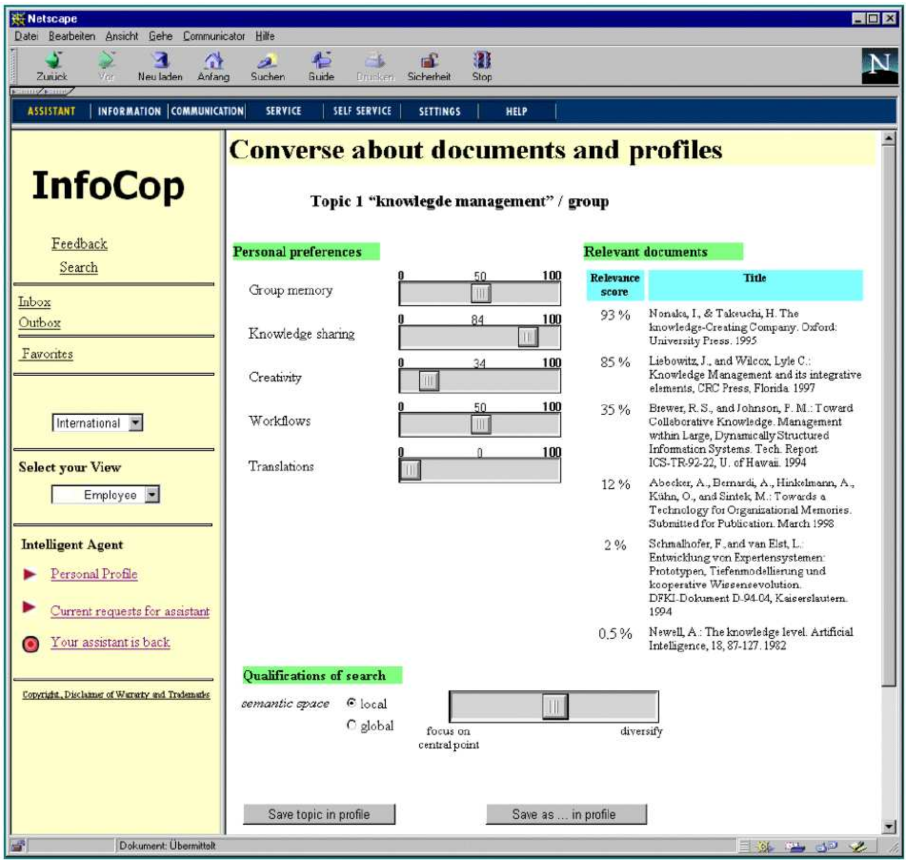

Fig. 4. A dynamic query interface for conversing about the personal relevance of documents

It should be noted that the comprehension assistant does in no way replace the human comprehender. Quite to the contrary, the comprehension assistant and its human users are cooperating with one another and thereby enhance but also somehow change the human comprehension process. This is comparable to the changes which are occurring in text composition when an author switches from a manual typewriter to a word-processor with spelling and grammar checkers. As is well known, certain qualities of the task execution are modified by using the new technology [2].

3 Embedding in a Multi-Agent System

Figure 5 shows how the described comprehension assistant is embedded in an oligo-agent system. An oligo-agent system is a special kind of multi-agent system, where 1) there is only a small number of different types of agents and 2) the agents are defined by their social responsibilities for achieving and maintaining certain qualities of the whole agent-system and the particular information environment in which they operate (cf. [15]).

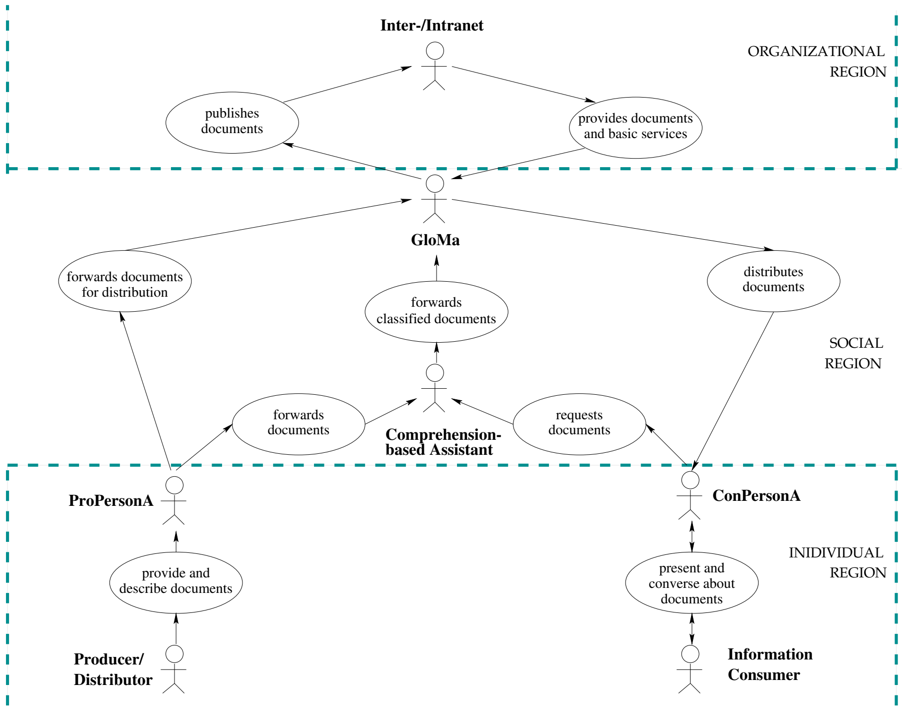

Fig. 5. The structure of the oligo-agent system for distribution and comprehension assistance within the organizational memory

We will now explain how the comprehension assistant is utilized for the various distribution- and comprehension-based tasks which have to be performed within an intranet. At first we have to explain some additional specifications. Repositories of potential information consumer groups and document groups as well as templates of distribution and interests-lists are held in a relational database system (RDBS). Based on these repositories, specific distribution and interests-lists are constructed (or chosen) and refined. These tasks also specify information as to when the information should be delivered. A demon acts upon this information. Whenever a *specific distribution* or *interests list* reaches its *distribution time* this demon computes the delivery information which basically consists of a table of (user, document info)-pairs. This table is processed by a standard delivery mechanism (e.g. e-mail), sending the document notifications to the user's inbox. The delivered information is filtered by the consumer agent (ConPersonA) and thus categorized individually. Each category has *presentation time* information which is used by a ConPersonA-demon to initiate presentation via the special user interface that is shown in Figure 4 and allows for individualized ranking and selection of documents.

How the *document providers* and *information consumers* of the intranet can employ the distribution and comprehension assistant to improve their consensual understanding of documents in the organizational memory is shown in the form of a use-case diagram (see Figure 5). A document provider (who may be either an author or a distributor) can publish a document in the organizational memory in one of two ways. He may have his own ideas as to where in the intranet (i.e. the virtual library) the document should be stored and to whom it should be distributed. In this case he would directly forward the documents and/or distribution list to GloMa, the global manager that has access to and maintains a representation of the organizational memory. Alternatively, he can call on the comprehension assistant, that will then make suggestions as to where the document is to be stored and who should be informed about its publication. In either case, GloMa will eventually store the newly published information in the organizational memory and distribute the information via the personal agents to the potential information consumers.

In addition to this *push-oriented approach* to knowledge management, the integrated assistant also provides a sophisticated *information pull* solution that is based on content and meta-content descriptions of the documents as well as user interests. With the personal agent and the described dynamic queries interface (see Figure 4), a user can specify his interests to the comprehension assistant, that will in turn communicate to GloMa which will then provide the specific documents.

4 Summary

The described cooperative comprehension-assistant that is to be applied for knowledge dissemination in an organization (i. e. the right documents to the right people at the right time) can be more generally characterized through the following design specifications and accomplishments:

1. Based on theories and models of human text comprehension [10], [14], a fully automated comprehension assistant with *three levels of representations* is used to achieve mutually shared representations between the oligo-agent system and their users. With increasingly more people applying this system, the mutual understanding may be shared among increasingly more people and increasingly more documents.

2. *Individual* and *global concerns* are coordinated in an oligo-agent system with shared responsibilities. While the individual agents are client-based, the global agent runs on an application server.
3. GloMa uses an RDBS to keep and maintain repositories that are used for the definition of user profiles (information consumers), document groups and distribution and interests-lists. Thereby it is possible to use the available information, consisting of document attributes, document contents and organizational structures as a whole. Concerning the *distribution and interest identification tasks*, GloMa maintains a continually updated representation of the *information environment* at the *relevant level of abstraction*.
4. *Decoupling* of *definition time*, *information-identification time* and *presentation time* allows distribution to be determined either by the *individual* or as a *common responsibility* of administrator, author, distributors and information consumers. These responsibilities may concern *one-shot distributions* as well as *periodical repetitions* of some general distribution specification.

From an application point of view, the functions that are provided by the comprehension assistant and its embedding in the oligo-agent system can be compared to the Fishwrap system [3]. The information consumer gets a personalized view of the information environment. This view consists of individual aspects that are based upon a semantic document analysis as well as upon organizational aspects (distribution lists), in which a distributor determines the portion of information that is delivered to the consumer. However, unlike an individualized newspaper, the proposed system has a very sophisticated representation about the ripeness and expiration time for relevant information. This allows for a different "date of publication" for each "line" of the newspaper. The combination of these elements leads to a flexible tool for handling different aspects of information distribution and information utilization which are the core problems of document and knowledge exploitation in the knowledge society.

References

1. Ambite, J. L., and Knoblock, C. A.: Agents for Information Gathering, *IEEE Expert, September/October*, 1997.
2. Boy, G. A.: Cognitive Function Analysis. *NATO-ONR Workshop Presentation, Washington D.C.*, 1997.
3. Chesnais, P. R., Mucklo, M. J, and Sheena, J. A.: The Fishwrap Personalized News System. *http://fishwrap-docs.www.media.mit.edu/docs/dev/CNGlue/cnglue.html*, 1997.
4. Clancey, W. J.: *Situated Cognition: On human knowledge and computer representations.* New York: Cambridge University Press, 1997.
5. DARPA (Ed.): *Proceedings of the 6th Message Understanding Conference (MUC-6), Columbia, Maryland, November 6-8, 1995.* Morgan Kaufmann, San Francisco, 1996.
6. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshmann, R.: Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391–407, 1990.

7. van Elst, L.: Ein kooperativer Informationsassistent zum gemeinsamen Verstehen von Textdokumenten (*An information assistant for the cooperative comprehension of text documents*). Master Thesis, Department of Computer Science, University of Kaiserslautern, 1998.
8. Fischer, G., McCall, R., Ostwald, J., Reeves, B., and Shipman, F.: Seeding, Evolutionary Growth and Reseeding: Supporting the Incremental Development of Design Environments. In *Proceedings of ACM CHI94 Conference on Human Factors in Computing Systems, 2, p. 220*, 1994.
9. Fischer, G., and Redmiles, D.: Enhancing Indirect Long-Term Collaboration with Intelligent Agents. NSF-Grant IRI-9311839, Boulder.
10. Kintsch, W.: *Comprehension: A Paradigm for Cognition.* Cambridge University Press, 1998.
11. Krauss, R. M., and Fussell, S. R.: Mutual knowledge and communicative effectiveness. In J. Galegher, R. Kraut, C. Egido (Eds): *Intellectual Teamwork: Social and technological foundations of cooperative work, pp. 111–146.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1990.
12. Landauer, T.K., Foltz, P.W., and Laham, D.: Introduction to Latent Semantic Analysis. *Discourse Processes*, 1998.
13. Salton, G.: *The Smart Retrieval System Experiments in Automatic Document Processing.* Prentice Hall, 1971.
14. Schmalhofer, F.: *Constructive Knowledge Acquisition: A Computational Model and Experimental Evaluation.* Mahwah, N.J.: Lawrence Erlbaum Associates, 1998.
15. Schmalhofer, F., and van Elst, L.: An oligo-agent system with shared responsibilities for knowledge management. In D. Fensel, and R. Studer: *Proceedings of EKAW '99*, Springer-Verlag, 1999.
16. Williamson,C., and Shneiderman, B.: The Dynamic HomeFinder: Evaluation dynamic queries in a real-estate information exploration system. *Proceedings of the ACM SIGIR'92 Conference, Copenhagen, Denmark, (June 1992), pp. 338–346*, 1992.