CWI Quarterly Volume 12 (2) 1999, pp. 159 – 172

Advanced Document Management through Thesaurus-based Indexing: the IKEMPP Platform

Dirk Vervenne
BIKIT, The Babbage Institute for Knowledge and Information Technology
J.Plateaustraat 22, 9000 Gent, Belgium
http://www.bikit.org

Electronic thesauri have been identified as strategic instruments for indexing electronic documents (Miller et al, 1990; Salton, 1968). However, one of the main problems of using electronic thesauri, remains the creation and maintenance (Grefenstette, 1994). In this paper, we provide some basic steps towards a solution for this maintenance problem through the IKEM-platform by combining groupware-methods and knowledge-based indexing techniques. IKEM is a client/server-based environment based on two indexing engines which generate automatically keywords, concepts and relation-frames for any electronic document. We present in this paper the basic functions of this platform and illustrate the maintenance issues that are solved by synchronous and asynchronous co-operation tools which support annotation clustering. We end by indicating some future development plans. IKEM is currently patent pending.

1. A lifecycle approach to document management

Today, electronic documents are conceived as 'hybrid containers of digital information': such textual containers can thus contain graphic subcontainers, dynamic spreadsheet tables or even dynamic linked pieces of text (Bielawski et al., 1997). This approach stimulates a more dynamic view to electronic documents which we implemented through a lifecycle with various document-related processes: creation, consultation, indexing, exploitation and archiving (see Figure 1).

It is important that we look at documents from a lifecycle point of view. Since most computers are used for text processing and webpage-design, people will more and more realise that the creation of a document is only the first step in a complex and dynamic lifecycle of the digital document. In this lifecycle, we

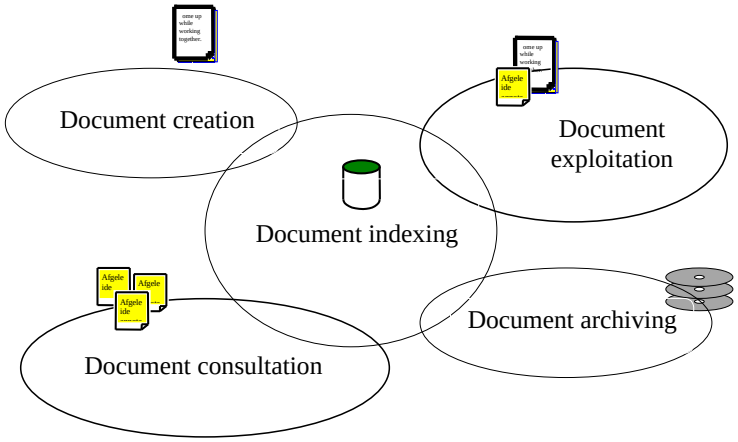

FIGURE 1. An overview of the various phases in the electronic document life cycle

differentiate 5 phases; the creation phase, the indexing phase, the consultation phase, the exploitation phase and the archiving phase. For all these actions, the quality of the indexing process has a large impact on the efficiency of the other processes.

There is a second reason why we consider the indexing process as a crucial phase. Since electronic documents not only contain data and information, but also a lot of strategic knowledge. In our approach, we consider knowledge as interpreted information from an action point-of-view. How can we then grasp the knowledge out of a document and how can we differentiate strategic knowledge form less relevant knowledge? We solve this problem with the IKEM-platform which is a set of document indexing, retrieval and annotation components that can function stand-alone in a LAN/intranet or fully be integrated into existing EDMS-platforms. IKEM has two kernel engines, an electronic thesaurus management system and a relation frame analyser: the combination of both engines with the IKEM web-annotation module provide a first step towards an automatic thesaurus-based indexing system, such as it is considered at the end in MAGRIJN et al., 1997.

2. THE IKEM PLATFORM: BASIC ARCHITECTURE.

The concept of the IKEM document management system is shown in figure 2. Electronic documents are imported into the system via the indexing process. This process generates automatically various index terms such as keywords, concepts and relations, which are assigned to documents as a function of the document content.

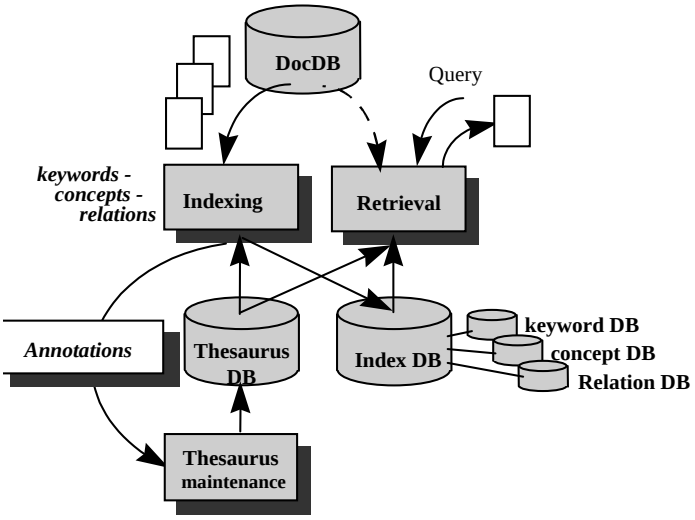

FIGURE 2. IKEM system model: document indexing and retrieval.

The indexing and retrieval processes being used in IKEM, are based on one part on the hierarchical structure of the thesaurus. The thesaurus hierarchy is used in order to create associations between documents and concepts. The use of the thesaurus therefore implies that the associations can also be expressed by terms not explicitly present in the analysed text. Such terms, we call concepts, refer to broader terms of keywords that are found in the document.

Next important part of the IKEM system are causal relations (within IKEM, these relations are referred to as "relation-frames"). The algorithm applied for identification of causal relations uses index terms (the arguments of the relations are keywords). By extracting the causal relations, we define an additional set of (internal) dependencies in the index database, which are the causal links between the index terms. Extraction and management of causal knowledge is a fundamental problem which is important in various domains such as material research, development of knowledge based systems and/or intelligent information filtering (see HAMERLINCK et al., 1995).

One of the most important aspects of the IKEM system is thesaurus maintenance. The indexing process usually identifies in documents not only the keywords, but also high- frequency terms which are not present in the thesaurus. In spite of their relevancy as index terms, those terms are not indicated as keywords because of their absence in the thesaurus. To incorporate those terms in the thesaurus, an additional iteration is needed, which is an update of the thesaurus with those high-frequency terms. This iteration needs to be followed by re- indexing of documents in order to maintain the consistency of the index and thesaurus databases. To support the thesaurus updates in a

multi-user environment, we implemented in IKEM a co-operative annotation system which can be described as an electronic discussion- environment, which clusters thematically all annotations.

In the following sections we discuss all the basic principles of indexing and retrieval in IKEM, the specifications of software modules, and the concept of the IKEM network configuration.

3. Basic principles of indexing and retrieval in IKEM

The most important elements of the indexing and retrieval processes in IKEM are keywords, concepts and causal relations. We present more details about these elements.

3.1. Keywords

The keywords represent a basic level of indexing and retrieval. The indexing process consists in a number of iterations, which are necessary in order to obtain single and composed keywords. The process extracts statistical lists of terms from the analysed document and compares the lists with the thesaurus. As a result of iterations and filtering, a list of terms is generated which are relevant for the document. The quality of the keywords found by this procedure depends on the thesaurus content and on the empirically determined filter values. The latter depends on the kind of application, the type of texts, etc.

The automatically generated keywords are stored in the index database together with tags of the documents. The retrieval process identifies the documents via links between keywords and document tags. The keywords are in this case used as input of the retrieval queries. The thesaurus can in this process serve as a source of hints for keywords.

3.2. Concepts

The second level of indexing and retrieval in IKEM are concepts. The concept terms are indicated automatically by the use of 2 elements: keywords and thesaurus hierarchical relations. Each keyword generates a series of concept terms or associations, which are obtained by traversing the paths between the keywords and the roots of the thesaurus. This implies that specialised terms will generate many associations. This way we create a powerful search system: each generated association will certainly lead to the indexed document. A concept that is related to a given document, is not explicitly in the document but receives its relation to it, since the concept is part of the set of all broader terms of a given keyword.

The associative retrieval process uses the thesaurus as filter. By selecting a concept term in the thesaurus, we obtain via the concept-database a series of documents which have been associated with this term in the indexing process. These documents can further be filtered by choosing specialised terms belonging to the sub-tree of the initial term. This procedure provides the user with

FIGURE 3. IKEM Thesaurus Maintenance Tool.

a possibility to navigate hierarchically along the thesaurus branches to find documents.

3.3. Causal relations

The causal relations used in IKEM are the cause-effect relations between the keywords or concepts. The process of automatic indexing of causal relations is based in the IKEM system on the concept of frames. In this approach, the frames represent verb-oriented language patterns extracted from natural language. The indexing process compares frames with sentences in the document and generates a table of causal relations: each relation record in this table contains 3 keywords: a cause, an effect and a condition. The relations that are found, are stored in a relations index database, which can be used to analyse links between possible causes and effects.

Given these descriptions of the IKEM-elements, we will now illustrate some features of the IKEM-platform itself, currently containing 6 tools.

3.4. Tool1: Thesaurus Management Tool

The user interface of the Thesaurus Management Tool is shown in Figure 3. This tool is developed as a front-end for the thesaurus database. The most important features and specifications of the tool are the following:

- **A viewer of the thesaurus relations (thesaurus browser):** The thesaurus relations (i.e. broader terms, narrower terms, relates terms and equivalency relations) are presented as clickable lists.
- **Tree view of concept terms:** Next to the linear lists with relations, we use a graphical tree viewer. This viewer is mainly intended to show terms

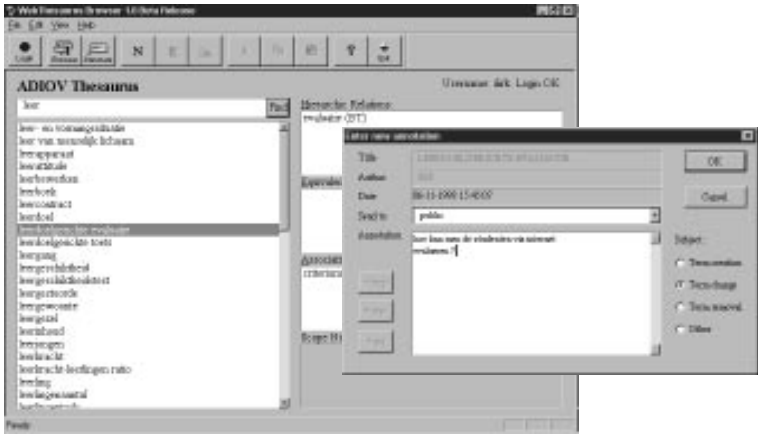

FIGURE 4. Annotation-tool for group-based thesaurus maintenance

ancestors (it shows the position of a term in relation to the root of the thesaurus, including the path of all parent terms).

- **Thesaurus updates:** We incorporated basic functions to insert and remove terms from the thesaurus tree. It should be mentioned that thesaurus updates can only be done by experts (with a status of a Thesaurus Administrator). We developed therefore a system of annotations, which can as be suggestions sent by the users to the Thesaurus Administrator. The users can also consult the update list generated by the parser to create new input for the thesaurus. This update list contains all highly freequent terms which are not in the thesaurus and not in the stop-word list. All thesaurus updates are logged to a thesaurus log file.
- **Annotations (multi-user):** Annotations are in IKEM used as a structured discussion forum regarding a new and/or removal of thesaurus terms. The workflow of annotations and thesaurus updates in a co-operative environment are a subject of on-going research (VERVENNE et al., 1997). The most important issues are synchronisation of thesaurus database with index database and system performance. See Figure 4 as an illustration
- **Conversion to HTML-format:** All thesauri within IKEM can be converted to HTML-format in order to consult the thesaurus with standard web-browsers. All annotation features can function in a Wwweb-environment.
- **Video scope notes:** all keywords can be illustrated by multimedia scope notes such as texts, graphics, digitised video and audio.

3.5. Tool2: Keyword/concept Parser tool

The IKEM parser can be used in an interactive, as well as in a batch mode (see Figure 5). The following functions are provided.

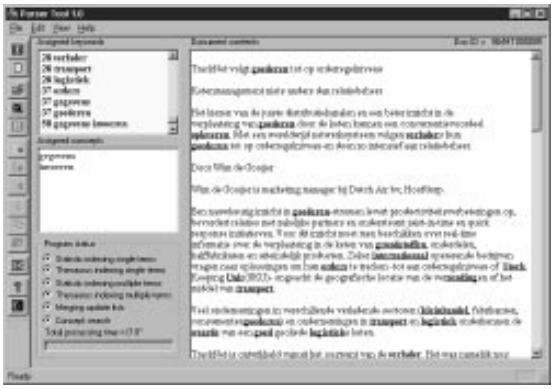

FIGURE 5. IKEM Keyword/concept Parser Tool.

- **Automatic generation of keywords and concepts:** The analysis is performed automatically by an algorithm that selects the keywords and concepts by combining statistical data with thesaurus filtering. For the moment, this implementation does not include any structural information related to documents (document meta-knowledge). In the future we will integrate such structural knowledge in the parser (SGML- and/or HTML- and/or RTF "recognisers").
- **Thesaurus-based:** The keywords as well as concepts are thesaurus terms.
- **Generation of update list for the parser:** the parser creates a list of relevant keywords, that are not in the used thesaurus, but which could be considered as candidate terms, given their frequency. This list can be consulted by the thesaurus manager are mailed automatically.

3.6. Tool3: Visual Query Builder

This tool permits the enduser to consult and retrieve all indexed documents (see Figure 6). It provides following facilities:

- **A visual tool for retrieval of indexed documents:** Search for indexed documents via keyword queries and concept queries.
- **Features of queries:** The queries are launched with 'point-and-click'. The user chooses keywords from a visual thesaurus viewer implemented as hierarchical tree structure. He/she can browse through the thesaurus and launch queries with a single button click. Each time a query is started, the viewer shows in response the number of documents associated with the queried term (independently for keywords and concepts). By clicking on the button "Get Documents", one obtains document tags relevant for the selected thesaurus term.

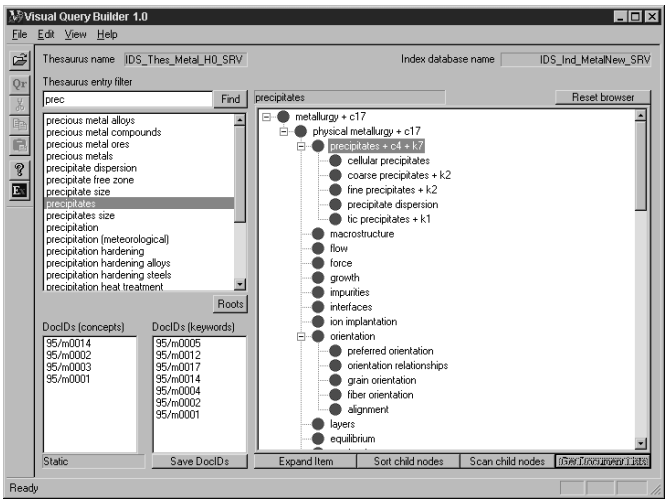

FIGURE 6. IKEM Visual Query Builder.

3.7. Tool4: Relations Analyzer Tool

With this tool, strategic knowledge is extracted from the documents. At this moment, all relations deal with causal knowledge (see Figure 7). We take an example. Suppose following sentence is part of a document:

> *"The present study has been undertaken to examine the effect of chemical composition on recrystallization behavior and r-value in Ti-added ultra low carbon sheet steel"*

Then the IKEM Relation Analyzer tool will identify the following structure

Causal relation =	*the effect of*
Cause =	*chemical composition*
Effect(s) =	*recrystallization behavior*, and *r-value*
Context =	*carbon sheet steel*

The tool has following functions:

– **Extraction of causal relations.** The function of the Relations Analyzer Tool is automatic extraction of causal relations. The algorithm we employ is based on the concept of frames. Details of frames and a summary of the algorithm are presented in KACZMARSKI, et al., 1996. The tool scans documents to identify sentences and expressions containing causal relations. The sentences found are extracted and compared with a frame dictionary in order to select arguments of relations. This way we extract causes, effects and conditions embedded in natural language. The tool generates a table

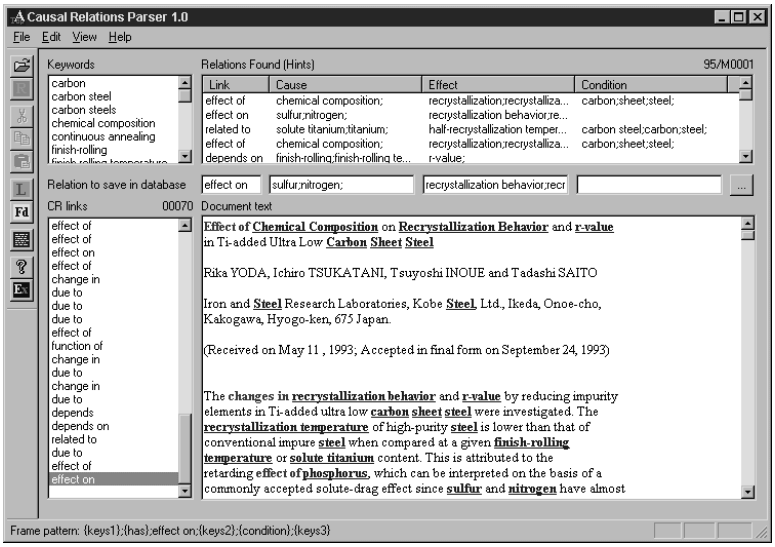

FIGURE 7. IKEM Relations Analyzer Tool

of relations found per document, which can further be 'manually' evaluated by the user and saved in the relations database.

- **Updates of frames:** The frame dictionary employed can be extended. Also, an annotation system for frames seems to be a relevant future extension.
- **Input for thesaurus:** Components of non-complete frames can be used as input for the thesaurus.
- **Relation conditions:** The relation analyzer database contains also tables to store conditions of relations. This can be either simple strings or things as complicated as a series of process parameters.

3.8. Tool5: Relations Query Tool

This tool is an easy-to-use query window for the causal relations, extracted from a set of document (see Figure 8). It has the following features.

- **Retrieval of causal relations and association of relations with documents:** The function of the tool is to retrieve relations and documents from relations database. The components of the relations are displayed in 3 lists ("cause-effect-condition" in Figure 7). The fourth list shows document tags. The user can click on terms in lists to see whether the terms are related to other terms and/or documents.
- **Test of hypotheses (Boolean 'AND' queries):** A hypothesis is created if one expects a link between cause X and effect Y. Such relations can be

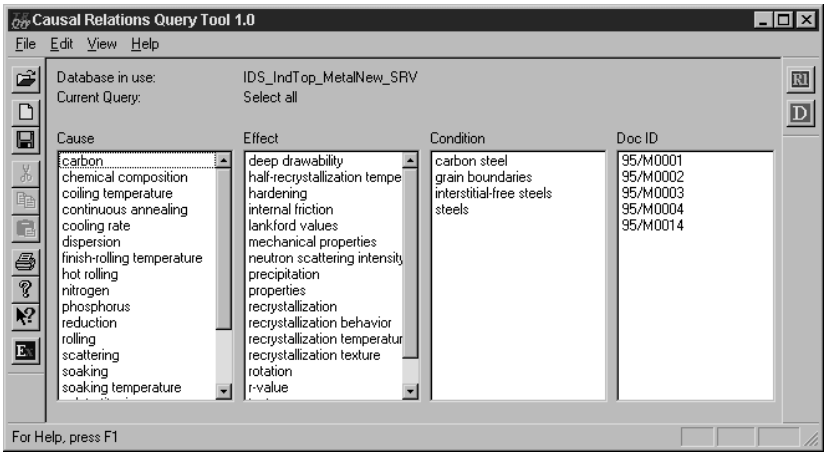

FIGURE 8. IKEM Relations Query Tool illustrating an metallurgic application.

tested by the use of the Boolean 'AND'-operator. This query is launched automatically if a clink in the 'cause' column is followed by a click in the 'effect' column (or vice versa). In response shows the tool the documents which confirm the hypothesis. In the future we will extend this type of queries with cross-references. The tool will indicate also the place in the document where the hypothesis was found.

3.9. Tool 6: IKEM Web Annotator

The most important function of the IKEM Web Annotator is synchronous and asynchronous support for information exchange between users via the Wwweb-network about the results of the IKEM-tools (see Figure 9)

Basic features:

- The application is based on TCP/IP and can therefore be used in LAN or via Internet.
- Broadcast mode - more than 2 users can participate in information exchange sessions.
- It has an asynchronous version through the annotation-mode and also a synchronous version through a chatter-option.

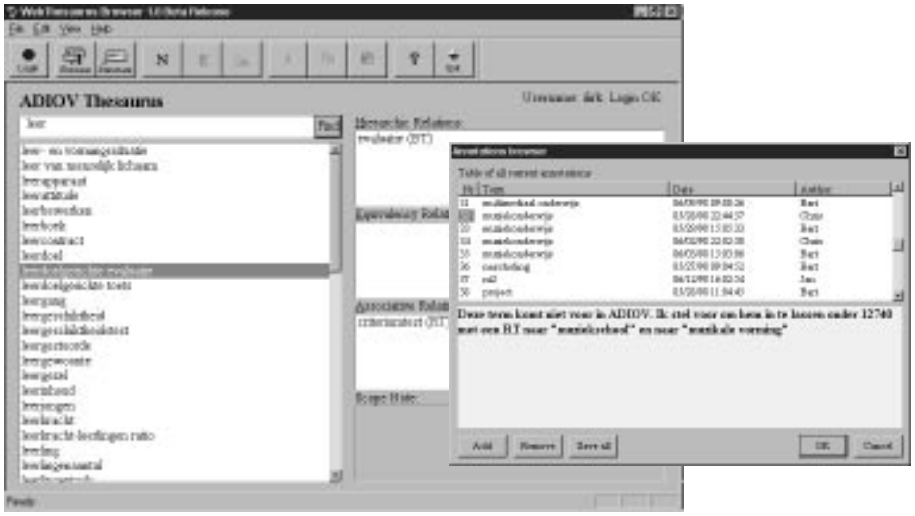

FIGURE 9. The IKEM Web annotator which clusters all annotations. The screendumps illustrate the maintenance of the ADIOV-thesaurus.

4. IKEM network configuration and integration with existing platforms

The network configuration of IKEM is presented in figure 10. It consists of Win 95/Win NT platforms provided with 32-bit ODBC and Winsock. The computers connected to the network have the following functions:

- IKEM server computer,
- IKEM client computers,
- IKEM Thesaurus and System Administrator site[1].

The IKEM-platform is a modular set of tools which can be used as such in a network or they can be integrated with existing document management environments or groupware-tools which can benefit of the IKEM-advantages. Figure 11 illustrates how IKEM can be integrated in a Lotus Notes/Domino™ environment. The extracted keywords, concepts as well as the causal relations are here implemented as 'Lotus views' on the documents; these have been imported automatically by an external call to IKEM-modules which operate in the background.

5. Conclusions

One of the most important advantages of the IKEM-platform is automatic indexing on the level of keywords, concepts and casual relations. The tools

[1] The Administrator site can also be installed on the server computer

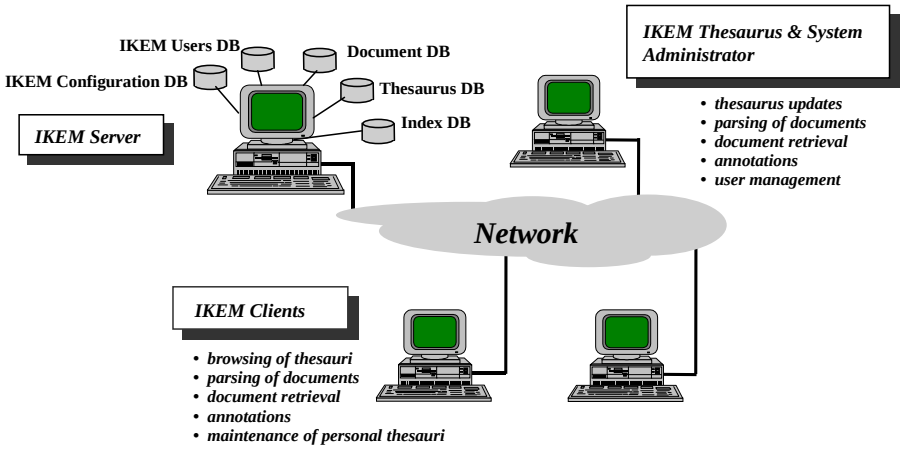

FIGURE 10. IKEM network configuration and basic functions (client - server - IKEM administrator)

allow thus for consistent analysis of large amounts of documents. The system can also be extended to batch processing. The quality of obtained indexing terms depends however on the thesaurus used. We therefore devoted much attention to maintenance of the thesaurus. This implies that the processes of indexing and updates play an important role within the system.

The IKEM system is implemented as a multi-user distributed system. The IKEM network consists of a database server and a number of client and/or IKEM Administrator sites. It has been illustrated that the IKEM-toolset can be integrated in existing groupware environments.

The most important aspects of future extensions are summarised below:

- Automation of the thesaurus maintenance.
- User Modelling through interest profiles.
- Extension of the annotations system (also input from indexing) via workflow concepts
- Integration of speech technology (voice controlled system, use of voice recognition for input of thesaurus terms and/or 'dictated' annotations).

ACKNOWLEDGMENT The authors of this article wish to thank the IWT (Het Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie) for financial support.

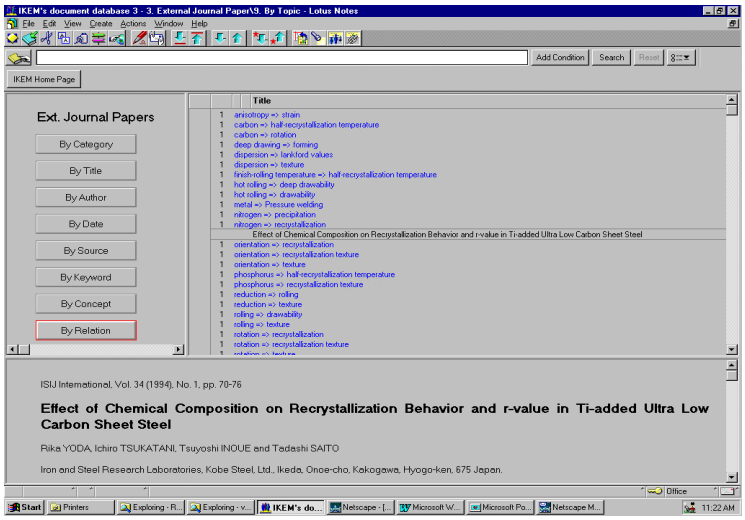

FIGURE 11. The integration of IKEM-tools within LotusNotes^{TM} platform.

REFERENCES

1. BIELAWSKI L. & BOYLE J. (1997). *Electronic Document Management Systems: a user centered approach for creating, distributing and managing online publications*, Prentice Hall PTR, NJ.
2. GINSBERG A., (1993). A unified approach to automatic indexing and information retrieval *IEEE Expert*, pp. 46–56.
3. GREFENSTETTE G. (1994). *Explorations in automatic thesaurus discovery*, Kluwer Academic Publishers.
4. HAMERLINCK F., VERVENNE D., VERHEYEN M., ARENTS H., DILEWIJNS J., VANDAMME F., BOGAERTS W. (1995). IKEM, an interactive knowledge based system for discovering and validating metallurgic knowledge, *Proceedings of ISS-Conference: 37th conference of mechanical working and steel processing*, Canada.
5. KACZMARSKI P., VERHEYEN M. and VERVENNE D. (1995). *Developing a document indexing tool for an Electronic Document Management System*, BIKIT Library Bulletin.
6. KACZMARSKI P., VERHEYEN M. and VERVENNE D. (1996). IKEM Relaties: indexering en retrieval, Technisch Rapport *IKEM – Werkpakket Relaties*.
7. MAGRIJN, H., PONTZEN, S.A.TH.M., RIESTHUIS, G.J.A., SCHIPPER J.D., WIJNANDS G.J. (1997). *Woordsystemen: Theorie en praktijk van thesauri en trefwoordsystemen*. Den Haag: NBLC.
8. MILLER G., BECKWITH R., FELLBAUM C., GROSS D., MILLER K. (1990).

Five papers on WordNet. CLS Report 43. Cognitive Science Lab. Princeton Univ.
9. Salton G. (1968). *Automatic information organisation and retrieval*, McGraw-Hill Book Company.
10. Spark Jones K. (1971). *Automatic Keyword Classification and information retrieval*, London, Butterworths.
11. Vervenne D., Kazmarski P., Verheyen M., Vandamme F. (1997). Strategic management of metallurgical knowledge through intelligent processing of electronic documents *Proceedings of IPMM'97*, Brisbane, Australia.
12. Vervenne D., Vandamme F. (1997). *Knowledge management based on groupware, workflow and document archiving: towards intelligent intranets*, Phenix series, Communication & Cognition, Gent.
13. Vervenne D., Vandamme M., Verheyen M., Van Craen R., Vandamme F. (1995). *Virtual Browsers for navigation and annotation of multifaceted thesauri*, contribution to COST-14 workshop CSCW-VR computer-supported co-operative work with virtual reality, Stockholm, 20-22 april.