The SIMATIC Knowledge Manager

Mario Lenz[1] and Karl-Heinz Busch[2] and André Hübner[1] and Stefan Wess[1]

[1]TecInno GmbH
Sauerwiesen 2
D-67661 Kaiserslautern
Germany
Tel. +49 6301 606-400
Fax +49 6301 606-409
{lenz,huebner,wess}@tecinno.com

[2]Siemens Automation & Drives AS CS3
SIMATIC Online Support
P.O. Box 48 48, D-90327 Nürnberg
Germany
Tel. +49 911 895-2652
Fax +49 911 895-4212
Karl-Heinz.Busch@nbgm.siemens.de

Abstract

In this paper, we present details about the SIMATIC Knowledge Manager (SKM), a Textual CBR system that uses existing documents, such as FAQs and other user-oriented documentation, and finds the most relevant documents for a given problem description. The major difference of the SKM compared to standard Information Retrieval tools and WWW search engines is that knowledge about the application domain can be brought into play when assessing the relevance of documents. Thus, not only the names of products, devices, and software components can be represented but also their relationships, such as dependencies between a series of products. Furthermore, the structure of the domain can be taken into account thus allowing a clustering of products into categories that express common properties. The SKM employs a case-based approach in that it considers the existing documents as cases and a user's request as a query in the sense of the CBR paradigm. Also, by relying on these documents, a separate case authoring process is avoided which would require a substantial amount of both initial knowledge engineering when setting up the system as well as maintenance while the system is running.

Problem Description

Siemens is selling a wide range of automation systems within its SIMATIC program world-wide. Subsidiaries of Siemens as well as other companies are engaged in repairing and maintaining this equipment. To support technicians when trying to solve problems at the customer's side, Siemens operates a hotline for second level customer support which answers telephone calls. This hotline serves about 65,000 customers world-wide and 85 employees manage approximately 13,000 calls per month.

The hotline has to struggle with two major problems:

- Firstly, there is a huge demand for information from the clients' side (we will refer to both external technicians as well as Siemens internal staff searching for information as *clients* or *users*). Consequently, the hotline staff is always busy and sometimes requests from clients are queued and can only be answered some hours later.
- Secondly, the hotline is contacted again and again because of the same problem due to different clients facing the same difficulties when maintaining SIMATIC components. As the hotline staff itself consists of 60 people, such situations are rarely recognized and, hence, reuse of problem solving knowledge hardly ever occurs.

To overcome these problems, Siemens decided to utilize the increasing popularity of the World Wide Web and to provide information, such as updates of drivers or news about the latest products, via WWW pages. An immediate consequence of that decision was that some kind of system would be required allowing users to search the document collection. This was recognized as a crucial requirement for the success of that strategy because the primary question is whether or not a particular information is given in a set of documents but whether or not users would be able to find it. Due to the expected growth of the collection, a simple folder—oriented categorization very soon would have required a tremendous amount of maintenance and, at the same time, would have limited the benefit for users.

Siemens very soon realized that a standard Information Retrieval (IR) approach [16] would not be appropriate despite numerous tools being available. This has the following reasons:

- The documents frequently refer to names of products, devices, hardware and software components. These sometimes have a kind of code, such as *CP 1473 MAP*, and sometimes consist of a group of words, such as *USER TECHNOLOGY MODULE*. Also, for a single product several names may exist, such as the commonly used name, the correct product identifier, and a code similar to the above. Representing such names in IR tools would be hard if not impossible.
- Products cannot be considered in isolation. Rather, relationships exist among the various products and components which should be taken into account when searching for relevant documents. For example, the above mentioned component *CP 1473 MAP* appears to be highly similar to another component named *CP 1430*. Also, some products may be highly similar because they belong to the same series whereas another

group of products shows completely different properties. A wide range of such relationships exist and demand for means for explicitly representing this type of knowledge.

- SIMATIC is not a single range of products but rather consists of more than a dozen different programs. Some of the products can be used for many programs whereas others are highly specific for a single one. Again, this is a specific type of knowledge that has to be somehow represented in a search engine.
- Although the documents primarily consist of textual descriptions, more structured elements, such as feature values, are widely used, too.

Obviously, all the above remarks indicate that a knowledge-based approach is required which utilizes the various pieces of knowledge about the domain in order to implement an assessment of documents beyond plain keyword matching. To identify appropriate technologies and tools, Siemens in Autumn 1997 started a 3—month trial in which several tools have been tested with respect to their applicability to the task, the expected costs, and the required maintenance while running the system.

During this period, it also became obvious that the initial idea of letting the hotline staff use the tool would not be feasible. The reason for this is that the employees running the hotline are highly skilled and, hence, would make use of such a system only in rare circumstances if particularly difficult problems have to be solved. Of course, building a system for these situations is also highly difficult. Instead, Siemens decided to design the system for use by the clients with the objective of achieving a call avoidance at the hotline due to clients solving their problems at least partially in a self-service manner.

Application Description

System Architecture

As a result of the above described trial, Siemens decided to utilize the CBR-ANSWERS technology developed by tecInno and to build a customized version called the SIMATIC Knowledge Manager (SKM). Using the SKM basically consists of three different phases: Encoding knowledge about the domain, building an index, and running the system for answering information requests. In the following, we will describe the latter two while the first phase is described in Section 2.2. The overall architecture of the system is shown in Figure 1.

The index is constructed in an offline process during which a given document collection is first converted by a *Preprocessor* to an internal *case document* format that has been specified in order to abstract, for example, from graphical elements in the original documents and to have a unique encoding of characters. In a second step, these *case documents* are analyzed by means of a *Case Parser* and the *index* is built. During this process, knowledge about the domain, for example about

- relevant terms and concepts, names of products etc.
- relationships between the various products
- the structure of the domain

is taken into account. The *index* then is based on the model of Case Retrieval Nets [7] and can be considered as a case memory containing the references to the documents as well as the encoded knowledge in a *compiled* form.

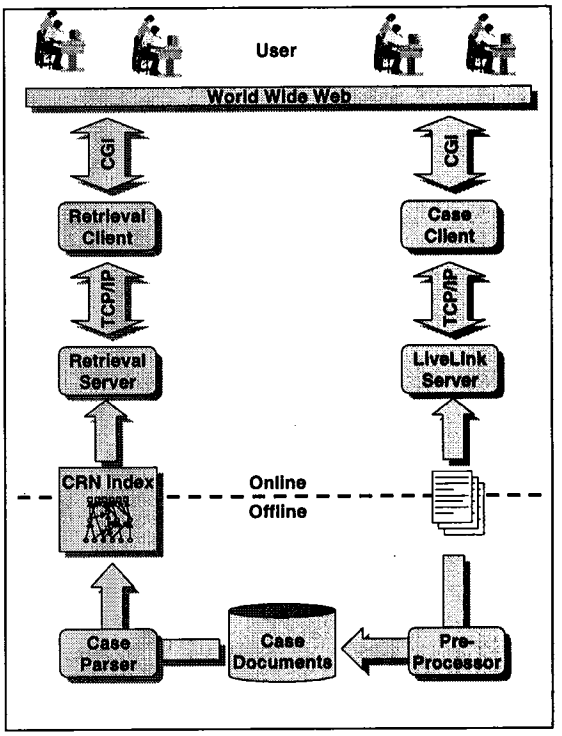

Figure 1: Overall architecture of the SKM

The online process is started by users contacting the system, for example via the WWW as shown in Figure 1. Standard CGI script technology is used to prepare the queries, and the *Retrieval Client* then connects to the *Retrieval Server* via a TCP/IP protocol. This server performs the actual document retrieval and returns a list of *document identifiers* ranked according to relevance to the query. An additional service, the *Case Client*, is used to contact a *Case Server* in order to obtain the actual documents for display. In the SIMATIC Knowledge Manager, the latter service is provided by a *LiveLink Server* that has already been in use independent of the SKM. For *LiveLink*, the document identifier then consists of a number uniquely referring to a single document in its latest revision.

Figure 1 and the above description correspond to the internet version of the system. For the two other versions (see Section 2.3), the architecture is similar. The major difference in the CD-ROM version are that users access the system via a Java-GUI rather than the WWW and that Windows DLL calls are being used instead of the described communication protocols.

Knowledge Representation

The major form of knowledge representation is by means of a so-called *dictionary* in which all the relevant terms and their relationships are specified. Each relevant concept is encoded as an *Information Entity* (IE) which can be activated by a set of keywords and phrases. A very important aspect of that representation is that multiple languages can be covered by a single dictionary in that an IE does not have single set of keywords associated but rather keywords for every supported language. As an IE is a kind of *symbol*, the internal representation will not depend on the language. Currently, the system is available for German and English.

For every pair of IE a relationships may be defined which, at run time, is converted to a degree of similarity. Currently, relationships such as *is part of*, *is generalization of* and various levels of similarities (ranging from *not similar* to *synonymous*) can be used. This knowledge can be defined by using the CBR-ANSWERS KNOWLEDGE MANAGER, a Java-based tool that can be considered as an authoring kit for the actual CBR-ANSWERS run-time system. A snapshot of that system is shown in Figure 2, a node in that graphical representation corresponds to an IE, the keywords and phrases indicate when this IE is present in a document, and the net indicates relationships to other IEs.

This kind of knowledge representation is in some sense similar to semantic nets and the use of ontologies. However, because of pragmatic reasons, we did not utilize the full power of these formalisms:

- Firstly, building an ontology requires a substantial amount of knowledge engineering [4,12] which did not seem acceptable with respect to the available man power.
- Secondly, querying an ontology often requires a kind of formal language in order to benefit from the reasoning power of that formalism. Such a formal language, however, would not be appropriate when the system is being used by external clients.
- Thirdly, reasoning capabilities, such as checking the correctness of some statement or whether the overall model is consistent, did not seem realistic for this highly complex domain.

Consequently, we restricted ourselves to the described form of dictionaries which can be considered as weak forms of ontologies.

In addition to the set of IEs and their relationships, valuable knowledge also exists that can best be encoded by means of a taxonomy of feature values. For example, each document in the SIMATIC domain belongs to a particular *topic*. The knowledge about the relationships between all topics can best be represented by means of two structures:

- Firstly, there is a taxonomy of topics based on which a similarity model can be build.
- Secondly, *a consistency matrix* encodes which topics are related in the sense that documents about one can be helpful in the context of the other.

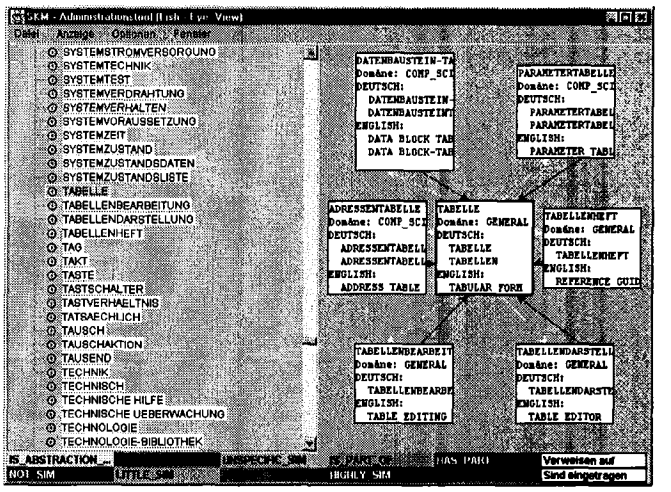

Figure 2: Snapshot of the CBR-ANSWERS KNOWLEDGE MANAGER. On the left, a part of the IE dictionary is displayed in alphabetical order (in German here); on the right, relationships between the IEs are visualized and can be edited. By clicking on neighboring nodes, one can navigate through the net of IEs.

In order to provide a close integration into other systems running at Siemens, two more components have been which allow for a close interaction between the SKM and other tools that clients are already used to:

- Clients often refer to products using so-called MLFBs. These are numbers each describing a specific product or group of products. As related products have a common prefix in a MLFB, relationships between products can be obtained.
- Furthermore, an integration with the product catalogue has been implemented in such a way that whenever a user submits a query and some product name(s) can be identified, then a link to the online catalogue is generated thus allowing users to obtain detailed information about those products.

System Environment

The SIMATIC Knowledge Manager currently runs in three different versions:

1. An internet version is accessible without restrictions and mainly used by external technicians. For this, the SKM runs as a server in a Unix environment and CGI scripts are used for managing the clients' sessions.
2. An intranet version runs in a similar environment except that an extended set of documents is used and access is only granted to internal staff of Siemens.

3. For an offline version, the SKM can also be started from a CD-ROM that is updated four times a year. Here, a Java-based GUI is used and the SKM itself is available as a DLL running under 32 bit Windows.

Uses of AI Technology

Textual CBR

As already mentioned above, the SIMATIC Knowledge Manager applies case-based approaches to document management. In particular, results from Textual CBR have been utilized which suggest how a CBR model for this type of task should look like [6,10]. Following the knowledge container model [13,14], four containers are of interest:

- The *case base* consists of the document collection. More precisely, cases are interpreted as *views* on the documents. I.e., the case base exists only temporarily (while building the index) and internally to the system.
- The *vocabulary* is encoded by means of the IE dictionary as described in Section 2.2.
- The *similarity model*, too, is partially contained in the IE dictionary but is enriched by taxonomies.
- *Adaptation knowledge*, on the other hand, is not used.

The latter point already indicates a limitation compared to the full CBR process model in that the SKM is restricted to the retrieval task and no adaptation or integrated learning are performed within the system.

The retrieval itself is considered as an information completion process [1] as compared to more traditional problem solving methods, such as classification or diagnosis. The key idea of information completion is that there is no *a priori* distinction between a problem description and a solution in a case. This has a number of consequences with respect to the design of the CBR system, including the format of case representation and the kind of case memory that can be used for retrieval purposes. Concerning the latter, the SKM heavily relies on the model of Case Retrieval Nets [7] which allow for the implementation of information completion processes and provide both

- efficiency thus allowing the application to deal with thousands of documents (i.e. cases)
- flexibility in the sense that a case memory can be built despite the fact that the cases only show very little structure compared to, say, feature vectors.

Shallow NLP

As we are dealing with textual documents and want to represent the content of these, it seems straightforward to use techniques originating in Natural Language Processing (NLP). With respect to state-of-the-art tools in that area we see the following principle problems:

1. We have to deal with natural language in a technical domain. This implies that in virtually any document new terms will occur that have not been described before. Thus, a lot of techniques requiring a *complete* dictionary are definitely not applicable.
2. The SKM has to handle large amounts of text efficiently.
3. The documents often do not contain properly structured sentences but only phrases or tables. Following some kind of grammar about English or German probably does not help very much in such situations.
4. Even if one could solve the above problems, the question arises what to do with the result of a parsing process of some NLP tool. Any statement in a document can be paraphrased such that the same contents is expressed by means of completely different grammatical structures. Thus, one would have to decide whether two parse structures (trees etc.) are about a related topic or not – a decision for which the actual parse structure will be of little help only.

Nevertheless, we applied some kind of *shallow NLP* in that we used part-of-speech tagging methods [17] in particular when building the initial model. By means of such techniques, we were able to automatically sort out a huge amount of the vocabulary, such as determiners and other auxiliary words which are useless for the intended task.

Information Extraction

To a very limited extend, the SKM also applies techniques known from Information Extraction: When parsing a document, the text is scanned for specific expressions which should be represented as attribute-value pairs. Mostly, these expressions correspond to physical measures, such as `220 Volt`. For this, rules can be defined which are fired by a trigger (here `Volt`) and check the surroundings of the trigger for a specified expression, such as a number. This corresponds to the task of named entity recognition in Information Extraction systems [3,15].

Application Use and Payoff

The first versions of the SKM, namely the internet and intranet versions, went online in March 1998, a first CD-ROM followed in April 1998. In December 1998, version 3.0 has been delivered. The current version supports two languages, namely English and German. For both, the same knowledge model is being utilized. This is possible since documents are represented by sets of IEs – which are independent of the underlying language.

Figure 3 shows a snapshot of the CD-ROM version which is shipped by Siemens three times each year. Besides the documents on the CD, this version also uses so-called *CD2Web* technology: If the user of the system is online, it checks for up-to-date documents on the web and, thus, combines the advantages of both media: fast access of the CD and latest versions from the WWW.

The CD-ROM contains about 26 MB textual data in approximately 3,500 HTML files per language. On the internet, about 150 files are added per language per month. Internal users can access additional documents so the

volume of the database is approximately 80 MB in about 10,000 documents per language (as of December 1998).
The usage of the SKM on the CD cannot be measured yet but the usage of the HTML version on the internet-server is promising (Figure 4). The number of sessions increased from 1,500 in August 1998 to 3,000 in November 1998 while during the same period the number of calls at the hotline remained on the same level. The figures for the offline CD-ROM version are not included here.

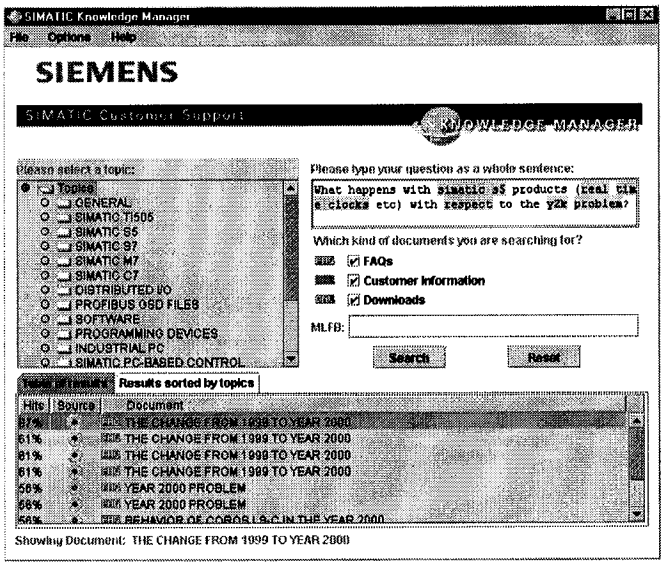

Figure 3: The CD-ROM version of the SKM. In the query field, the concepts that have been recognized by the system are marked.

With the increased usage of the SIMATIC Knowledge Manager, the benefits of the system will increase, too. Siemens expects to reach the turnover in 1998 [2]. The main benefits of the system are (depending on who is using the system):

- call-avoidance from customers
- call-avoidance from Siemens staff in the regional bureaus and other parts of the company
- reducing the call duration time in the hotline itself.

These are benefits that lead to an indirect reduction of costs for Siemens in that the costs for the hotline support do not grow. Other soft benefits are:

- Customers and Siemens staff have fast access to the provided sources of knowledge.
- The service is available 24 hours all over the world.
- The offline CD-ROM solution can be used in places where internet access is not available.
- The service is easy to use even by clients who are not used to search with internet search engines.

To summarize, the use of the SIMATIC Knowledge Manager and the knowledge base gives Siemens a lead on the support of its products and helps stabilizing its market share world wide.

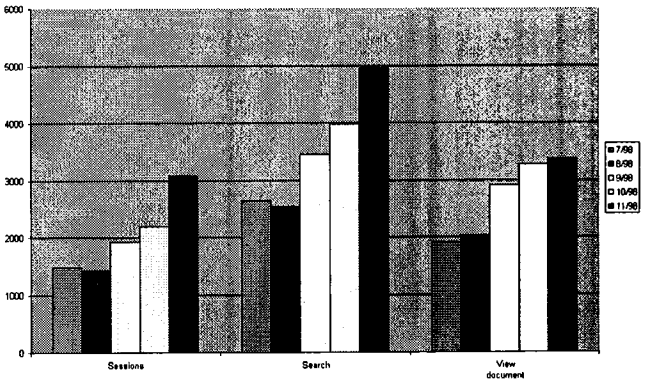

Figure 4: Usage of the SKM on the internet [2].

Application Development and Deployment

As mentioned in Section 1, Siemens performed a trial to figure out the appropriate tools in 1997. After that, a prototype of the SKM has been developed in December 1997 which then was tested again. The development of the prototype lasted about one month, the development of the first full version took another 3 months. As the SIMATIC Knowledge Manager is the result of a cooperation between tecInno and Humboldt University, Berlin, major parts of the implementation have been performed by university staff and students. Therefore, an estimation of the overall manpower required for building the system is hard. This is even more true since the development of the system has been tightly coupled with research in the area of Textual CBR. Also, the result of that work is the CBR-ANSWERS tool which is applicable in other circumstances, too – in fact the core of the system has already been used in related projects [8]. Nevertheless, we estimate approximately

- 4 man months for the development of the system itself (i.e. the CBR-ANSWERS server resp. the DLL)
- 1 man month for implementing the described customization for the Siemens environment
- 1 man month for building an initial model (i.e. the dictionary) based on a given document collection
- 3 man months for the development of the CD-ROM solution, including the Java-GUI and additional features

The development of the system greatly benefited from the cooperation with Siemens and an intensive communication between system developers and hotline staff. When launching the project, an initial two-day workshop was held during which the main features of the system have been worked out. Also, at this workshop a kind of common vocabulary has been defined which during the development of the system very much simplified the communication between the involved partners. Two more one-day workshops during the development phase helped clarifying open issues.
Another advantage for the deployment of the system has been that for the telephone hotline a certain infrastructure was already available at Siemens which could be reused for

the SKM. For example, clients were already used to the online product catalogue and a CD-ROM providing the latest information had been shipped before. Even more important, clients when calling the hotline and, as explained above, the answers to their requests may be substantially delayed. By using the SKM, in contrast, clients get an immediate response. Both facts, obviously, help very much in motivating people to use the system.

Maintenance

Maintenance of the SIMATIC Knowledge Manager covers three different aspects:

Firstly, the system itself has been improved steadily over the last months. This primarily concerns the integration of the system (as explained with the online product catalogue) as well as some functionality improving the usability, such as an extended GUI. This improvement, of course, implied further implementation and has been carried out by the developers of the system at TecInno.

Secondly, new documents have to be included in the system. For this, a set of tools have been provided which can be used in the offline process of index generation as described in Section 2.1. In principle, a fully automatic update of documents could be performed. However, this is not really needed and due to the existing environment at Siemens it is simpler to let a member of the hotline start this process. Also, the process described in Section 2.1 is only an approximation of the actual process in so far as an intermediate step has been omitted which has been introduced in order to allow for an incremental update process during which only the modifications on the documents are considered.

Thirdly, the IE dictionary representing the model of the entire system needs maintenance, be it because it has only been partially defined before or because new products have to be integrated in it. This model maintenance is performed by Siemens staff, i.e. by one member of the hotline who invests about half a day per week.

Currently, research is being carried out aiming at a further support in particular of the latter aspect of maintenance. The objective here is to analyze a document collection and to figure out both

- relevant terms that should be represented in the system
- as well as relationships between these terms.

The results so far are promising [5,9] but these techniques will only be used in a semi-automatic manner, i.e. suggestions will be generated about which a (human) expert will have to decide.

References

[1] Burkhard, H.-D.: *Extending some Foundations of CBR.* in: [11], Chapter 2

[2] Busch, K.-H.: *Customer Support for Siemens Products on the Internet and CD-ROM.* Invited talk at EWCBR-98

[3] Cowie, J. and Lehnert, W.: *Information Extraction.* Communications of the ACM, 39(1), 1996

[4] Fensel, D.; Erdmann, M.; Studer, R.: *OntoBroker: The Very High Idea.* Proceedings of the 11th International Flairs Conference, 1998

[5] Hübner, A.: *Semi-automatische Wissensanreicherung im textuellen fallbasierten Schließen*, Master's thesis, Humboldt University Berlin, 1999

[6] Lenz, M. and Ashley K.D. (Eds.): *Proc. of the AAAI Workshop on Textual CBR*, AAAI Press, 1998

[7] Lenz, M. and Burkhard, H.-D.: *Case Retrieval Nets: Basic Ideas and Extensions.* in: Görz, G. and Hölldobler, S. (Eds.): *KI-96: Advances in Artificial Intelligence*, Spriner Verlag, LNAI 1137, 1996

[8] Lenz, M. and Burkhard, H.-D.: *CBR for Document Retrieval - The FAllQ Project.* in: Leake, D. B. and Plaza, E. (Eds.): *Case-Based Reasoning Research and Development*, Springer Verlag, LNAI 1266, 1997

[9] Lenz, M. and Glintschert A.: *On Texts, Cases, and Concepts.* to appear in: Proc. XPS-99, Springer Verlag, 1999

[10] Lenz, M.: *Defining Knowledge Layers for Textual Case-Based Reasoning.* in: Smyth, B. and Cunningham, P. (Eds.): *Advances in Case-Based Reasoning*, Springer Verlag, LNAI 1488, 1998

[11] Lenz, M.; Burkhard, H.-D.; Bartsch-Spörl, B.; Wess, S.: *Case-Based Reasoning Technology – From Foundations to Applications.* Springer Verlag, LNAI 1400, 1998

[12] O'Leary, D.E.: *Using AI in Knowledge Management: Knowledge Bases and Ontologies.* IEEE Intelligent Systems, 13(3), 1998

[13] Richter, M.M.: *The Knowledge Contained in Similarity Measures.* Invited talk at ICCBR-95

[14] Richter, M.M.: *Introduction.* in: [11], Chapter 1

[15] Riloff, E. and Lehnert, W.: *Information Extraction as a Basis for High-Precision Text Classification.* ACM Transactions on Information Systems, 12(3), 1994

[16] Salton, G. and McGill, M.: *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983

[17] Schmid, H.: *Probabilistic part-of-speech tagging using decision trees.* Technical Report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, Nov 1995