Managing the Knowledge Contained in Technical Documents

Mario Lenz
Dept. of Computer Science
Humboldt University
D-10099 Berlin, Germany
lenz@informatik.hu-berlin.de

Abstract

In this paper, we present a technique that has successfully been applied for implementing content-oriented search in specific types of documents containing valuable knowledge of a company. The technique, known as *Textual Case-Based Reasoning*, provides mechanisms for integrating virtually any kind of knowledge available. When combined with appropriate business processes, it allows for a reuse of any information stored in the documents and, thus, for an efficient knowledge management with respect to the given documents.

1 Introduction

In today's world, knowledge is one of the most valuable assets of modern companies. This is in particular true for highly industrialized countries with high wages and the need to achieve industrial growth by innovation and additional services which could, in the required quality, not be provided by less industrialized countries. This need for innovation has a number of consequences which, over time, will change the organization of companies:

- Firstly, lifecycles of products are becoming ever shorter. Hence, innovations have to be brought to the market more rapidly in order to take the full benefit and to gain advantage over competitors.
- Secondly, high-tech products require highly skilled employees and often the integration of entire teams of knowledge workers which may even work in distributed places.
- Last but not least, an efficient customer support is getting more and more important. Because of the growing complexity of products, customers are hardly able to handle all problems that might occur with a purchased product. Rather, they expect the sales staff to handle these problems. In particular in business-to-business sales processes, this is becoming more and more crucial as the total cost of ownership of a product is heavily influenced by maintenance costs [LBP+96].

All these issues indicate that a company-wide knowledge management (KM) is required which has to cover the entire production process, including market analysis, product design, manufacturing, sales, and after-sales customer support.

The most important requirement for a successful knowledge management strategy is that well-defined business processes are established which allow for an efficient exchange of information and guarantee that knowledge is available whenever needed. Apart from these business processes, modern information technology can, however, provide useful tools to support various KM processes. In this paper, we will address the problem of managing the knowledge contained in documents. For this, we will clarify what we consider typical *know how documents* (Section 2) and present a technology that has been successfully applied to content-oriented search in these documents (Section 3). We will describe potential application areas and give details of projects performed in industrial

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/

settings (Section 4). Finally, Section 5 will discuss the pros and cons of this approach as well as related work.

In this paper, we address knowledge management problems from an IT point of view. However, we want to emphasize here that even the best technology can be applied successfully only if appropriate business processes are being installed which prepare a fertile ground for the technology itself. In particular we will point out which processes have been essential for the success of the technology when describing the performed projects.

2 Know How Documents

In modern industry, a widespread use of data collections can be observed. Virtually anything that can be recorded is being stored in databases; these provide information about all products, orders, processes and the like via a special query interface (e.g. SQL).

When it comes to document knowledge and experiences, however, human beings very much prefer to express themselves in natural language rather than being restricted to a rigid data format. This has several reasons:

- Firstly, formal languages, which are required for storing information in databases, normally do not provide sufficient flexibility and expressiveness.
- Secondly, if being sufficiently expressive, these languages are hard to handle in particular by a non-expert user who is unwilling to learn a formal language just to query some kind of information system.

Consequently, a huge amount of knowledge is stored in natural language, i.e. in the form of textual documents. Examples for such *experience bases* are collections of Frequently Asked Questions, news group files, handbook, manuals, and program documentations, and informal notes. In the following, we will refer to these documents as *know how documents*.

In fact, the entire World Wide Web consists of textual documents containing useful knowledge. While most of these WWW documents are fairly unstructured, know how documents in the above sense show a number of specific properties:

1. These documents will in most cases discuss problems related to a specific domain. If, for example, a user visits the WWW site of some hardware vendor and reads the FAQs, then it is obvious that these FAQs will describe problems and solutions related to the products of the vendor.
2. Major parts of these documents are given as natural language text. In a FAQ collection, for example, the texts contained in the question and answer sections will describe a problem and a solution, respectively. This information is not available otherwise.
3. In addition to the text, however, the documents typically also contain structured data. The above mentioned FAQ collection will, of course, also contain product names, version numbers, operating systems and so on which are best encoded in an attribute-value representation.
4. Last but not least, these documents usually have some kind of pre-defined internal structure. An FAQ, for example will have a question, an answer, and possibly a title.

To summarize, know how documents show some specific properties which allow them to be called *semi-structured* in the sense that textual information is mixed with more structured forms of representations and both are arranged in a fairly npre-defined manner in such a document. Thus, these documents can be clearly distinguished from other textual pieces, such as novels or WWW pages in general.

In addition, know how documents usually contain substantial parts of the knowledge assets of a company. Of course, this knowledge is much harder to manage than, for example, product specifications contained in a database where a well-defined logic can be applied to retrieve the relevant information.

Traditionally, techniques from Information Retrieval (IR) or even a simple keyword search are applied for this task, such as in the widely used Internet search engines. These approaches have a number of shortcomings which are due to the fact that they provide generic techniques which are applicable in virtually any domain. On the one hand, this allows for a wide-spread usage. On the other hand, domain-specific knowledge can hardly be integrated into such approaches. For example, knowledge of the following kind can hardly be considered in such a system:

- Domain specific concepts and meanings associated to various terms: When speaking about printer problems, the concept "jam clearly describes a problem with the paper feeder and thus disambiguity is avoided.
- Relationships between various concepts: Two different printers may be highly similar because they are both inkjet printers; on the other hand, they substantially differ from any laser printer.
- Structure of the domain: Printer problems can usually be divided into installation problems, printing problems and the like, which for quite different problem types.

- Structure of documents: FAQs on the WWW site of a printer manufacturer will most likely have the above discussed title-question-answer structure.

In the following, we will present a technology that is able to deal with these problems. It directly utilizes techniques from Case-Based Reasoning and, hence, is able to consider knowledge in the above sense when searching for specific information. As both experimental evaluations as well as applications show, this approach allows for much better results than traditional IR techniques.

3 Textual CBR

In this section, we will briefly explain the basic ideas of Textual CBR and sketch the CBR-Answers system which has been developed for this purpose and utilized in a number of projects. More detailed descriptions can be found in [LB97, LHK98].

3.1 From Documents to Cases

Case-Based Reasoning (CBR) is concerned with the reuse of episodic knowledge in similar problem situations. More precisely, the general CBR process can be described as a cycle consisting of four phases [AP94], namely

RETRIEVAL of cases similar to the current problem

REUSE of the retrieved cases for the current problem, including adaptation

REVISION of the proposed solution w.r.t. correctness

RETAINING new knowledge into the knowledge base

For solving specific problems, a CBR system may utilize knowledge which usually belongs to one of the following knowledge containers [Ric98]:

a) the case base, i.e. the collection of all cases

b) the vocabulary used to describe cases

c) the similarity measure used to compare cases during RETRIEVAL

d) the adaptation model used during REVISION

For the purpose of this paper, we assume that documents are available which can serve as a starting point for building the case base (a). Further assuming that adaptation (d) is of limited use only in this type of application, we have to solve three major problems:

1. How can documents be converted into cases?

2. How to determine an appropriate vocabulary for case representation?

3. How to assess similarity of cases?

The first question we will address is the vocabulary (2): For this, a careful knowledge acquisition is required during which features important in that domain, specific terms, names of devices and products etc. are identified. The result of this process is a concept dictionary containing all the relevant terms including their contexts. Note that for the construction of this dictionary, not only statistics about keywords may be considered but also techniques from Natural Language Processing, such as stemming information, and virtually any kind of information available for the particular domain (e.g. product databases).

In terms of KM, this concept dictionary is closely related to an ontology for a specific application [O'L98, FES98]. However, the expressiveness is limited in that highly sophisticated inferences are not supported. On the other hand, obtaining such a dictionary will be possible with less effort than usually for ontologies.

Given this dictionary, a *parser* can be implemented which automatically extracts the concepts contained in a text and builds a case representation of the document (1). Once cases are represented this way, they can be compared in terms of the sets of concepts they contain. This, however, is not based simply on the intersection of two concept sets but also takes into account similarities between different concepts — as always in CBR. Thus, documents are ranked with respect to how many concepts they contain that are closely related to the concepts expressed in a specific query.

3.2 The CBR-Answers System

The CBR-Answers system is the result of our research and development activities in the field of Textual CBR. It has been developed in cooperation with TecInno GmbH, Kaiserslautern, who have a special expertise in developing and marketing CBR products in general. Figure 1 gives an overview of the overall architecture of the system. It directly implements the ideas explained in the previous section in a Client-Server architecture. The system itself works in two different phases:

- During *pre-processing*, the available documents are parsed, the concepts according to the given concept dictionary are extracted, and a case base is constructed where each case represents a particular document. Once pre-processing is finished, the *Retrieval Server* receives an update command and loads the new case base. This process can be performed off-line.

- During online *retrieval*, the users enter their questions in a WWW form and a *Retrieval Client* is invoked which connects to the *Retrieval Server* which handles the request and retrieves the most

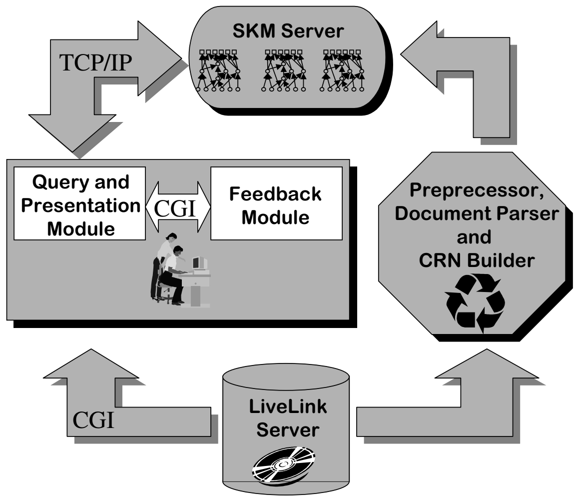

Figure 1: Overall Architecture of the CBR-Answers System as implemented for the SIMATIC Knowledge Manager

relevant documents. An additional module, the *Case Client*, may be used to display the documents in an appropriate format and to retrieve additional information, access rights etc. from a *database*.

4 Application Areas and Projects

In this section, we will briefly describe two potential application areas of Textual CBR as well as industrial projects performed. Further details on these projects can be found in [LHK98]

4.1 Hotline Support: SIMATIC Knowledge Manager

Siemens AG is selling so-called SIMATIC products world-wide. To support technicians when trying to solve problems at the customer's side, Siemens already had installed an efficient customer support hotline. To further improve this service, Siemens evaluated the CBR-Answers system. This evaluation revealed that there would have been only a limited benefit when installing such a system directly to support the hotline. The reason for this is that Siemens has a highly qualified hotline staff who would only consider using a tool in about 30% of all requests, that is only for the very difficult questions which are, of course, also hard to handle with Textual CBR techniques.

However, as a result of this evaluation, the scenario of an Automatic Hotline has been developed. The idea here is that external technicians are referred to a set of FAQs and related documents before they consider calling the hotline. In order to efficiently find related information in the huge amount of documents available, CBR-Answers is being used as an intelligent search engine. The major goal on behalf of Siemens is to achieve a *call-avoidance*, i.e. hotline staff should only be contacted in case of really difficult problems that have not been solved before.

In March 1998, an Internet version of the SIMATIC Knowledge Manager was installed at Siemens, in April 1998 a CD-ROM was available for customers, and in June 1998 an extended Intranet version was implemented. The system currently handles approximately 7,500 German documents; work is going on to support further languages.

4.2 In-house Knowledge Management: FALLQ

LHS AG is market-leader in developing a customer care and billing system for cellular telephone service providers. In principle, LHS delivers just one, though highly complex, software product. In practice, however, every customer (i.e. telecommunications company) demands some specific features which are not available for competitors. Consequently, there are many different versions and releases which have to be maintained in parallel. Due to the rapid growth of the telecommunications industry, LHS is permanently employing new staff also for the system development and customer support groups. Furthermore, a lot of distributed project teams work for different customer releases.

After the evaluation of a prototypical Textual CBR implementation, LHS decided to apply this approach for in-house knowledge management. For example, it is often the case that highly similar requests by different customers are handled by different staff without knowing from each other. This happens, for instance, when fixing bugs. On the other hand, LHS has a well-defined regime of documenting all processes (bug fix reports, customer requests etc.) in pre-defined electronic forms. Hence, it seemed straightforward to apply Textual CBR for making the knowledge contained in these documents available to other staff.

After the prototype had successfully passed internal testing processes, a first Intranet version has been installed at LHS in Summer 1997. The system currently handles approximately 45,000 English documents of various types; negotiations about further improvements and extensions of the systems are going on [LB97].

5 Discussion

5.1 Relations to Knowledge Management

As discussed in the above sections, Textual CBR may provide useful techniques for managing the knowledge contained in what we called know how documents. According to the two scenarios we discussed, the main benefits are:

- The development of an inventory of knowledge is supported by the dissemination of appropriately styled documents. Note that, according to our experiences in particular from the SIMATIC Knowledge Manager, having a tool at hand to retrieve the relevant documents will drastically increase willingness to invest in writing such documents.
- Users are enabled to query the system(s) at any time and, in the case of Internet applications, all over the world. Thus, information is provided whenever needed.
- For in-house applications, re-inventing knowledge is avoided by sharing and reusing the information stored during earlier problem solving episodes.

However, Textual CBR techniques have to be integrated into well-defined business processes which assure that the techniques can efficiently be applied. In both application projects sketched in Section 4, for example, broad document collections have already been available as well as clear specifications about *who* has to write *what* kind of documents under *which* circumstances, and *how* these documents should be structured. In that sense, Textual CBR may be seen as a technique useful for supporting a more general knowledge management strategy.

5.2 Comparison to Standard IR

As we have discussed in Section 2 already, Textual CBR techniques allow the integration of any type of knowledge available in a domain and, thus, can be used to implement content-oriented document search strategies which perform much better than traditional IR. However, this benefit does not come for free:

- Firstly, the application of Textual CBR requires the focus on a specific domain, for example the products of a particular company. In contrast, IR is generally applicable and can also be used for searching the entire WWW — which is not the case for Textual CBR.
- Secondly, in order to benefit from the knowledge one has to perform a careful knowledge acquisition process which may be supported by appropriate tools and may utilize all knowledge sources available for the domain. However, the domain expert will always be required as a last resort for answering open questions. Hence, the entire process of knowledge acquisition requires a substantial amount of man power.

5.3 Benefits of the Knowledge Layers

In order to be able to evaluate the contributions of each of the above discussed knowledge layers to the retrieval results, we performed a number of tests in which we varied the amount of knowledge considered during document searching and determined *precision* and *recall* which, however, have been slightly changed compared to the standard IR definition in order to make them applicable for the hotline scenario. This evaluation revealed that, as expected, each layer further improves the performance of the system. In particular, a significant improvement compared to a pure keyword-based search could be observed. Details of the performed experiments are beyond the scope of this paper but have been described elsewhere [LHK98].

5.4 Related Work

Recently, quite a number of projects have been launched which address the problem of handling textual documents by means of knowledge-based techniques, in particular CBR. In the following, we will discuss some of these.

The **FAQFinder** project [BHK+97] also tries to apply CBR technology, in combination with other techniques, to document retrieval. In particular, FAQFinder's goal is to answer natural language questions by retrieving these from FAQ files from USENET news groups. FAQFinder also uses a thesaurus (namely, WordNet) to base its reasoning on a semantic knowledge base and assumes that documents are given in a semi-structured format (namely, as questions-answer (QA) pairs).

FAQFinder differs from our projects in so far as it does not focus on a specific domain. Instead, it applies a two stage process:

- In the first step, a shallow analysis, mainly of the keywords contained in the query, is used to infer the most likely news groups related to the request.
- After the user has decided on one of the presented news groups (i.e., after s/he selected a topic to focus on), a more sophisticated analysis of the related FAQ file starts to compare the contained QA pairs with the query entered by the user.

In some sense, this interactive scenario relates to the focus on a specific domain in that the user confirms the topic suggested by FAQFinder in the first stage. With the CBR-Answers project, we have focussed still further on a specific domain in that each application has been designed specifically for a particular domain. For example, in technical areas, a lot of terms exist that would hardly be represented in general purpose thesauri, such as WordNet. Also, a careful knowledge engineering process has been undertaken to employ domain-specific knowledge for similarity assessment. This would not be possible in the scenario of FAQFinder, where the semantic base (i.e., WordNet) is the same for all news group topics.

The **Spire** system [DR97] uses a completely different approach for dealing with textual cases. Based on the observation from IR that people have problems in formulating *good queries* to IR systems, the idea behind Spire is to use a combination of CBR and IR technology:

- Given a user's request, a HYPO–style CBR module is used to analyze this request semantically and select a small number of relevant cases representing text documents.
- The most relevant cases from the first stage are then used to pose a query to the Inquery retrieval engine.
- After having retrieved relevant documents, cases are used again to extract the most relevant passages from the documents.

Compared to the projects described in this paper, this is a completely different approach in which CBR is in fact used as an interface to IR. Here, the usage of domain knowledge is limited to the cases suggesting good indices for the IR system. It cannot be used for similarity assessment, etc.

From the knowledge management community, projects dealing with building ontologies and using these for providing content-oriented search facilities are closely related. As already discussed above, the concept dictionaries required for Textual CBR can be seen as ontologies, too. However, building complete ontologies requires much more effort than constructing the simpler concept dictionaries for Textual CBR. Also, ontological systems, such as the OntoBroker [FES98] require that WWW documents be annotated with categories according to the underlyintg ontology. From our experiences, this is unrealistic in so far as in many applications tens of thousands of documents exist and nobody could afford doing this. Finally, making use of the power of ontological systems requires highly complex query languages whereas our focus is on letting the user express his information need in natural language.

References

[AP94] A. Aamodt and E. Plaza. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[BHK+97] R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg. Question Answering from Frequently Asked Question Files. *AI Magazine*, 18(2):57–66, 1997.

[DR97] J. Daniels and E. Rissland. What You Saw Is What You Want: Using Cases to Seed Information. In [LP97], pages 325–336.

[FES98] D. Fensel, M. Erdmann, and R. Studer. Ontobroker: The Very High Idea. In *Proceedings of the 11th International Flairs Conference (FLAIRS-98)*, Sanibal Island, FL, 1998.

[LB97] M. Lenz and H.D. Burkhard. CBR for Document Retrieval - The FAllQ Project. In [LP97], pages 84–93.

[LBBSW98] M. Lenz, H.D. Burkhard, B. Bartsch-Spörl, and S. Wess. *Case-Based Reasoning Technology – From Foundations to Applications.* Lecture Notes in Artificial Intelligence 1400. Springer Verlag, 1998.

[LBP+96] M. Lenz, H.D. Burkhard, P. Pirk, E. Auriol, and M. Manago. CBR for Diagnosis and Decision Support. *AI Communications*, 9(3):138–146, 1996.

[LHK98] M. Lenz, A. Hübner, and M. Kunze. Textual CBR. In [LBBSW98], chapter 5, pages 115–138.

[LP97] D. Leake and E. Plaza, editors. *Case-Based Reasoning Research and Development, Proceedings ICCBR-97*, Lecture Notes in Artificial Intelligence, 1266. Springer Verlag, 1997.

[O'L98] D. O'Leary. Using AI in knowledge management: Knowledge bases and ontologies. *IEEE Intelligent Systems*, 13(3):34–39, 1998.

[Ric98] M.M. Richter. Introduction. In [LBBSW98], chapter 1, pages 1–16.