Chapter 16

Text mining based knowledge management in banking

K. Lebeth, M. Lorenz & U. Störl
Dresdner Bank AG Frankfurt, University of Bremen, University of Applied Sciences Darmstadt, Germany.

Abstract

This paper presents a text mining based knowledge management (KM) approach. After a short introduction, we describe our idea of an integrated knowledge management infrastructure, using natural language technology as a main building block to enable the sharing of knowledge with as little additional effort as possible. Two tools are presented, which built the main components of the solution. Annotation and indexing of a big textual corpus on the one side and a sophisticated graphic search and retrieval front-end on the other side.

1 Introduction

In the information age, organizations whose primary means of business is information are facing a paradox situation. On the one hand, information is available in ever more amount and quality, on the other hand, the probability that one can utilize it is decreasing caused by a natural lack of human processing capacity, which is not scaling with the supply. As information is to a great extent available in textual, natural language based and as such unstructured or only partially structured form, natural language technology can play an important role in dealing with this information flood.

Apart from this fundamental consideration we can identify several good reasons for the essential importance of KM and especially for a natural language based KM strategy within a large financial service provider. An inherent information need in financial consultancy is owed to the fact, that a well-informed consultant is an essential foundation for the customer's success. One of the most important selling propositions a financial service provider has is his competence, because the products he sells are – at least to a high extent –

WIT Transactions on State of the Art in Science and Engineering, Vol 17, © 2005 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-995-7/16

interchangeable. Moreover the sheer size of the organization, which in the case of Dresdner Bank AG alone adds up to about 40,000 employees, entails that the same information is collected at various places within the company. Knowledge in such an organization is heavily spread and the effort for finding the "right" expert to contact must not be underestimated. Rotation of employees between different entities makes expliciting a crucial concern.

2 The document as a primary source

In a financial services domain, which relies on knowledge as a main business and success factor, the primary means of recording, transporting, and storing knowledge traditionally is the document, be it a letter, an invoice, an internal memo, or whatever form of written documentation one can think. A vague guess at the number of document pages containing valuable information will probably be in the range of billions of pages for big financial service companies.

A core factor of today's systems providing notions of document- or even "knowledge-" management is its dependency on some sort of structural input. Be it some pre-clustering, providing keywords or developing a structure to fit documents into, most commercial systems are not able to generate benefit without manual effort.

3 Knowledge based search

The main issue of most information retrieval in today's working environment therefore is the enormous amount of data to be searched. The engineering progress that has made possible these vast repositories of data has no real correspondent in the semantics of the resulting systems. Nowadays information retrieval primarily means keyword-index based or full-text search on the data (*i.e.* the signal) itself.

Evaluating the statistics of one such retrieval mechanism – the keyword-index based search engine implemented on the Dresdner Bank corporate intranet – has shown that a very high percentage of users are using a single keyword only which leads to a very unspecific and usually huge result set. To identify the relevant items is still subject to manually browsing through the output unless the documents are really well annotated, which usually is not the case.

4 Building up a knowledge management infrastructure

The main idea behind our approach was drawn from the observation, that the typical knowledge worker is not willing or simply does not have the time to do extra work (such as providing and maintaining keywords, sorting documents into pre-existing structures etc.) to facilitate knowledge sharing with colleagues. Therefore we want to provide a set of tools that on the one hand do all the extra work without need of interference or even notice by the knowledge worker, and on the other hand will provide him with an easy to work retrieval tool so he can efficiently find his way through the growing knowledge repository.

An integrated framework built on these tools will enable every employee to use as well as contribute to the companies organizational memory without special skills or knowledge and hence dramatically lower the threshold for knowledge management compared to today's systems.

One of the core requirements for knowledge management is to first of all include those vast document repositories. The automatic creation of a simple syntactic index of electronically available documents is a first solution, which has the advantage of being simple and cheap but has its disadvantage in its "lack of explicitation: knowledge, hidden in the documents, must be exhumed during the consultations" [1] by the user. Therefore the benefit gained by an automatically built index quickly gets lost when the searcher has to read or at least to scan the results manually. Therefore Trichet *et al* [1] propose an orthogonal approach in "conceptualising knowledge in order to abstract a model" but admit, that the major disadvantage is its cost for "modelling is a long and heavy step, that generally requires several experts".

Taking this into account the benefit of a semantically enriched search mechanism becomes obvious. Our concept combines the low initial effort to be taken by an indexing approach with several methods of automated implicit model building and a retrieval mechanism which smoothly integrates into the individual desktop workspace. On the retrieval side we use the term correlation measure to extract some semantic information out of large corpora of documents that are already available in the intranet or other large document repositories. This information is then used to draw a semantic network (see figs. 1 and 2) to enable associative search and retrieval of documents. On the other hand we enable the automatic enrichment of documents, using term extraction methods and term frequency analysis to provide the most important keywords that might describe a given document as meta-information for easier indexing and retrieval.

5 Integrating principles

The common driving principle found in the various approaches is our notion of "minimal invasive knowledge management", which may be summed up as an IT-driven knowledge management infrastructure, which smoothly integrates itself in the workspace and workflow accustomed by each individual knowledge worker without any need to learn new tools, change habits or even take notice of its functionality [2].

Therefore the whole knowledge management process must be detached into the background and run automatically without the need for interference by the worker. Commonly used applications like word-processors, e-mail, or presentation graphics are the primary means of work for expliciting knowledge. Thus, providing integration into these standard applications can easily collect knowledge. Ideally introducing a new item to the knowledge base must not cause any effort in addition to the usual daily work.

On the retrieval side a semantically rich presentation layer is provided which incorporates more information in an earlier step within the retrieval process.

6 Modules

Several modules, implementing respective parts of the concept have been implemented or are currently in development. The main building blocks to form a knowledge management tool-set are described in this chapter.

6.1 Term extractor

To support the intelligent indexing part of our framework, a natural language term extractor is used. Currently we use the Insight Discoverer Extractor by Temis SA (www.temis-group.com). This tool uses linguistic analysis methods to extract tagged terms from given documents driven by a task specific grammar. Based on this component two high level applications have been developed in Dresdner Bank: the Knowledge Net and an Automated Metatagging Engine. These tools are being used for different aspects of the previously described indexing and semantical enrichment of documents:

6.2 Knowledge Net

The aim of the "Knowledge Net" (kNet) application is to improve the usability of the search interface and the quality of search results. It provides a general architecture for knowledge retrieval with text mining technologies by automatically generating a browsable semantic network of terms extracted from the document repository without any need for expensive human modelling or supervision [3].

Figure 1: kNet Example: Search Result for "Dresdner Bank".

The indexing engine is based on a term correlation grammar, which extracts pairs of named entities and calculates two correlation measures: a *syntactic* distance measure shows an untyped relation between two terms within a document or document part; the second, *paradigmatic* measure is drawn from statistical analysis of context similarities.

On the retrieval side the kNet is a visual search interface that implements our idea of an implicit knowledge model. Based on the two correlation measures a semantic network is drawn, showing the terms as nodes and weighted, coloured edges representing the two types of correlation. Figure 1 shows the search result for "Dresdner Bank" kept in autumn 2001. There is a strong *syntactic* relation between Dresdner Bank and Allianz – not a surprise considering the fact that Dresdner Bank has been taken over by Allianz in spring 2001. Now it is very easy for the user to restrict the search to documents containing the term "Dresdner Bank" as well as the term "Allianz", just by one mouse click – adding "Allianz" to the search phrase.

Moreover this example shows another feature of the kNet – the ability to extract named entities, *e.g.* company names like "Dresdner Bank" or persons names.

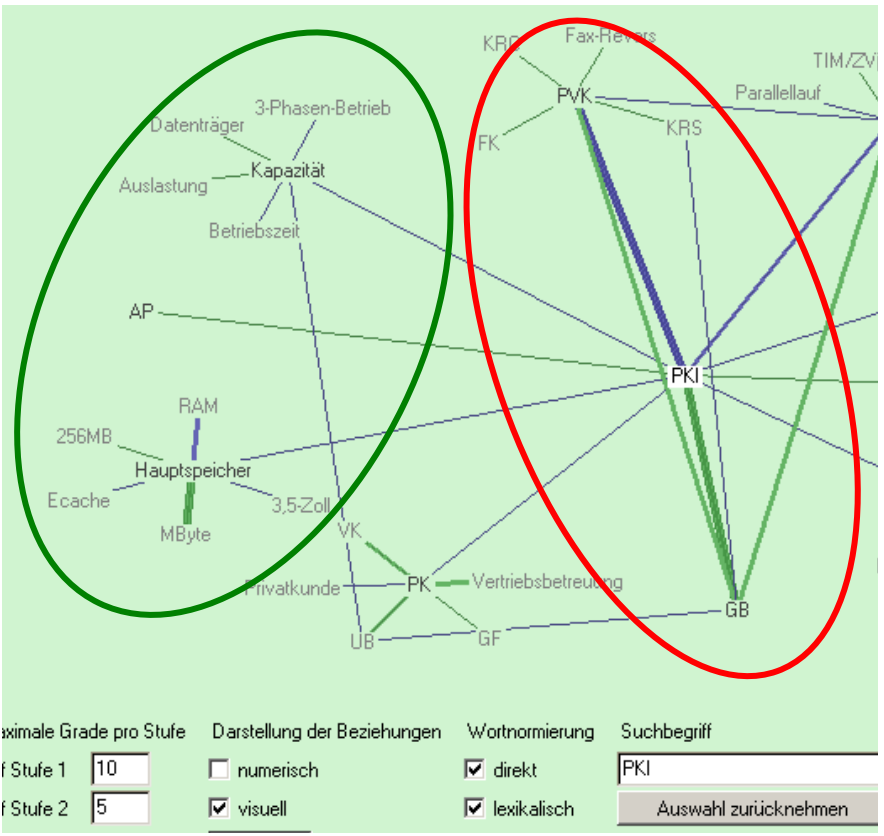

Figure 2: kNet Example: Search Result for "PKI".

Furthermore, this kind of graphical representation indicates if terms occur in different typical contexts. As an example – the search result for the term "PKI" – is shown in fig. 2. Even though the term "PKI" mostly stands for Public Key Infrastructure it has two completely different meanings inside Dresdner Bank:

There is a strong *syntactic* relation between "PKI" and "GB". In this context "PKI" stands for "Private Kunden Inland" (private customers) and "GB" means "Geschäftsbereich" (business area). So the solution is: "PKI" is the name of a business area of Dresdner Bank. There is also a strong *syntactic* relation between "GB" and "PVK" and a strong *paradigmatic* relation between "PKI" and "PVK". However, there are not many documents containing "PKI" as well as "PVK" there must be a correlation between PKI and PVK. PVK is the former name of the business area PKI. So you can detect interesting correlations between terms and use this information for a more successful search, *e.g.* including "PKI" as well as "PVK" in your search phrase to find all relevant documents – independent from the current name of the business area.

However, there is also a relation between "PKI" and such terms like "Hauptspeicher" (main memory) "RAM" and "MByte". What does this mean? The solution is: "PKI" is also the name for an in-house application of Dresdner Bank and there exist documents describing the necessary computer configuration parameters for this application. Again there is the possibility to describe the search more precisely by including relevant terms and later restrict the search to the relevant context.

While the kNet (Currently the kNet is being further developed to a commercial solution by H.A.S.E. GmbH www.h-a-s-e.org) improves the quality of search results on the client side there are also possibilities to improve the quality, *i.e.* the semantic richness of documents on the "backend" side. One possibility is the use of meta information. In the next section we describe a prototype, which enables the automated extraction of keywords for documents of almost all formats. Both concepts are orthogonal but can (and should) be used together.

6.3 Automated metatagging engine

Assigning meta information, *e.g.* topic or keywords to documents helps in identification of relevant documents in search and retrieval processes. However, assigning keywords and topics to documents as meta information is a very painful task. On the one hand, the attempt to define keywords for a document confronts the author with a cognitive effort that consumes too much time and energy. On the other hand, the assigned keywords seem to be sensitive to subjective mood, varying on the situation of the author etc. Oftentimes, such keywords tend to be too general, thus their usefulness is extremely low. Furthermore, a unified linguistic basis seems to be difficult to establish (*e.g.* some keywords appear preferably in their plural form, others in singular, etc.).

To address these problems an automated metatagging engine was implemented based on another specialized grammar for the term extractor which provides a set of statistically selected lemmatized keywords for a given

document. At present we evaluate different algorithms using text frequency combined with document frequency as statistical measure as well as clustering and categorizing based approaches.

Having generated meta information about document content, the gained information has to be stored and managed appropriately. With respect to the format, our aim was to use non-proprietary standards, which are supported by different search and retrieval engines. Beside the established meta tags for HTML documents we decided to use XML-based Semantic Web Technologies, *i.e.* RDF (www.w3.org/RDF) and Dublin Core Metadata Initiative (www.dublincore.org) for all document formats. Moreover, for some document formats (*e.g.* MS Word) there are also possibilities to store meta information inside the document in the proprietary meta data format.

To achieve the most painless workspace integration the engine is implemented as a Java-based web service with several clients, one of which – as an example – is invoked by saving a document in the commonly used word-processor. This client is .Net/C#-based whereas other clients, *i.e.* for batch processing of huge amounts of documents are written in Java [4]. Even though there are some teething troubles, web services seem to be a great opportunity for easy and efficient enterprise application integration in certain use cases.

One next step is the integration of the automated metatagging engine with the content management system used in Dresdner Bank to support the content authors to improve the quality of meta tags used within the Intranet without additional effort and following to improve the quality of search results.

7 Conclusion and future work

To build an enterprise wide organizational memory can either be an expensive and lengthy task for experts with the risk to fail by lack of acceptance or it can only cover the indexing and retrieval of documents which in turn makes the task of knowledge extraction uneasy and time consuming.

In our work we proposed a “medium approach” that on the one hand provides linguistically founded indexing tools and a visual based retrieval interface to integrate pre-existing document repositories, as they are always found in large companies. On the other hand – the knowledge creation side of the process – it relieves the burden of annotating from the author by automated keyword extraction.

These tools enable a nearly maintenance free indexing without previous model building or continuous supervision by expensive experts. Natural language technology is the main enabling factor, which together with an intuitive graphical visualization makes the described tools a valuable instrument in a knowledge intensive working environment.

As a next step we are working towards integrating the described tools into one “knowledge workbench” to establish a closed circuit from knowledge generation to retrieval within the already established toolset of office packages and intranet applications.

This idea of a "knowledge workbench" together with the vision of automatically extracting ontologies from a users own data filing structure to enable knowledge sharing has led to a partnership in the EU-funded project SWAP (Semantic Web and Peer-to-Peer). The main objective of SWAP is to successfully combine Peer-to-Peer solutions with Semantic Web technologies by means of emergent semantics [5, 6].

Thereafter it might be interesting to investigate community building and the use of a reputation mechanism as a stimulus for higher quality input, which we hope will increase the value of the emerging organizational memory.

References

[1] Trichet, F., Leclère, M. & Tixier, B. Capitalizing and Sharing Know and Know-How: an approach based on a Task/Method Knowledge-Based System, chapter 3, pages 31–40. In Dieng-Kuntz, R. & Matta, N. (editors). *Knowledge management and organizational memories*. Kluwer Academic Publishers Group, Dordrecht, The Nederlands, 2002.

[2] Cavar, D. & Kauppert, R. Strategien für die Implementierung IT-basierter KM-Lösungen: Minimal invasive Systeme. In Prange, C. (editor) *Organisationales Lernen und Wissensmanagement - Fallstudien aus der Unternehmenspraxis*. Gabler Verlag, 2002.

[3] Lebeth, K. Semantic Networks in a Knowledge Management Portal. In Proc. *KI/ÖGAI 2001: Advances in Artificial Intelligence, Joint German/Austrian Conference on AI*, Vienna, Austria, LNCS, Springer Verlag, 2001.

[4] Brandt, S., Cavar, D. & Störl, U. A real live web service using semantic web technologies: Automatic generation of meta-information. In Proc. *On the Move to Meaningful Internet Systems 2002: DOA, CoopIS, and ODBASE: Confederated International Conferences DOA, CoopIS, and ODBASE 2002,* Irvine, CA, October 2002.

[5] Erig, M., Tempich, C. & Staab, S. SWAP: Ontology-based knowledge management with peer-to-peer technology. submitted to: *4th European Workshop on Image Analysis for Multimedia Interactive Services.*

[6] SWAP Consortium: The Project. Online source: http://swap.semanticweb.-org/public/theproject.htm#Objectives (status 2004-12).