From Text To Cases: Machine Aided Text Categorization for Capturing Business Reengineering Cases

Catherine Baudin
Price Waterhouse Technology
Centre
baudin@tc.pw.com

Scott Waterman
Price Waterhouse Technology
Centre
waterman@tc.pw.com

Abstract

Sharing business experience, such as client engagements, proposals or best practices, is an important part of the knowledge management task within large business organizations. While full text search is a first step at accessing textual material describing corporate experience, it does not highlight important concepts and similarities between business practices structured or operated differently. Conceptual indexing languages, on the other hand, are high level indexing schemes based on taxonomies of domain concepts designed to provide a common language to describe, retrieve, and compare cases. However, the effective use of these high level languages is limited by the fact that they require users to be able to *describe* cases in terms an often large body of controlled vocabulary. The main challenge to using CBR and data mining technology for accessing and analyzing corporate knowledge is not in designing sophisticated inference mechanisms, but is in *representing* large bodies of qualitative information in textual form for reuse. This knowledge representation task is the process of mapping textual information to pre-defined *domain models* designed by knowledgeable domain experts. We are experimenting with machine aided text categorization technology to support the creation of quality controlled repositories of corporate experience in the business domain.

1. Introduction

Sharing business experience, such as client engagements, proposals or best practices is an important part of the knowledge management task within large business organizations. In the business domain, cases are usually textual documents about business processes, what methodology or technology was key to the success (or failure) of the operation, or what criteria were used to measure the results. While full text search is a first step at accessing textual material describing corporate experience, it does not highlight important concepts and similarities between stories describing business practices structured or operated differently.

The key to effectively capturing and reusing best practices and corporate experience is to be able to describe textual documents in terms of a structured language representing key concepts of the domain. Fully automated text processing techniques such as unsupervised text clustering are not likely to detect key domain concepts in complex textual documents that refer to multiple topics. When processing qualitative information such as business cases, the stories must be mapped to pre-defined *domain models* designed by knowledgeable domain experts. Conceptual indexing languages in the business domain are high level indexing schemes that highlight important aspects of the cases and their similarities. For example, indexing languages such as the International Business Language ™ (Emerson, 1996) or the Process Classification Framework (APQC, 1992), provide a common language to describe and compare business cases.

Unlike applications where cases are described by sets of numerical attributes (price of a PC, size of the hard disk in sales support applications) or by a small set of possible symbolic labels (processor type: pentium_166), mapping text documents to a large controlled vocabulary is not only time consuming but is also a complex decision process that involves human judgment to decide what a document is really about. In particular, business taxonomies can contain several hundred to several thousand concepts.

We are currently investigating techniques and tools to facilitate the mapping between textual information and structured case representations using large indexing taxonomies. This paper summarizes our experience with using machine aided text categorization to support the capture of best practice knowledge for our consulting practice. One characteristic of such a corporate repository is the importance of the accuracy of the stored information and of the way it is represented. For this reason we have been researching the use of text categorization technology (Lewis 96, James et al 95, Borko et al 65) to *support* human indexers rather than as fully automated text classification systems.

In section 2 we briefly describe the use of business process taxonomies and show examples of questions that can be answered using these structured case representations. In section 3 we present a machine aided text categorization tool called the Indexing Assistant that helps information specialists turn textual business stories into structured cases. Section 4 summarizes the key issues and points to directions for future research.

2. Using Business Taxonomies to Describe Business Cases

Conceptual languages, such as the International Business Language ™ (IBL) (Emerson, 1996), provide a common language for the identification of business improvement ideas. This language breaks business processes into activities that are shared by all companies, regardless of industries, enabling the comparison between companies that may be structured or operated differently. For instance, consider the following summarized business stories:

> Case1: "Company X, a petroleum company, had a problem with its purchasing operation and implemented a new order tracking system using an Intranet. This resulted in a 60% improvement in order processing time".
>
> Case2: "Company Y, a high-tech company, implemented a new networking electronic requisition system and found that it reduced the purchase cycle time by 50%".

Although these cases use a different vocabulary (purchasing system vs. Requisition system) and refer to companies in different industries, they are similar in that they both refer to the same business processes (orders and purchasing processes), and that they used similar enablers to reengineer the processes. We would like to produce the following conceptual description of case1 (where the attributes in bold are common to both case1 and case2).

Type: Engagement
Client: X
Project Lead: John Foo
Processes: Order Materials/Supplies,Receive Materials/Supplies
Enablers: EDI (Electronic Data Interchange)
Measures: Purchase Orders (Volume/Frequency)
Text:<the textual representation of the document>

Such a conceptual description is the key to retrieving similar cases that might be missed by a search based on exact keywords in a query.

In addition to retrieving information stored *explicitly* in a form of text documents, such structured representations can also be used to extract *implicit* information from the case repository by drawing inferences on case attribute relationships. Examples of implicit information that can be inferred from structured cases like case1 are:

- Searching for experts in different process areas within a company: Who has expertise in re-engineering the order processing function? (John Foo because he has been associated with several engagements in this area).
- Who in the firm has a business relationship with company X?
- What are the main enablers that are currently used to reduce order processing time? (trend analysis).
- How much experience do we have in using EDI for reengineering the order processing function?

Answering these types of questions from raw text is extremely difficult. However, these answers are relatively easy to infer from a set of well structured cases such as the one shown in our example.

In reality of course, textual business stories are rarely as concise and focused as the two examples presented above. They can be large documents describing the operation of several business processes over a period of several months. Mapping text to cases is an indexing task that requires both a deep knowledge of a particular classification system and a good understanding of the key issues discussed in the documents. For these reasons, in many companies, documents are structured (indexed) manually by knowledgeable corporate librarians rather than by authors or by people who route documents for sharing. This is a time consuming process. In our experiments for instance, an information specialist trained in using process taxonomies to describe business reengineering documents manually indexes 7 to 18 documents per hour.

The next section briefly presents the Indexing Assistant, a tool that facilitates the creation of repositories of business cases that can be then exploited by CBR and data mining tools.

3. From Text to Cases: Machine Aided Indexing of Business Reengineering Stories.

The process of turning a document into a structured business case involves mapping the text of the document to a taxonomy of pre-defined business concepts. The Indexing Assistant uses text categorization technology to help information specialists choose appropriate descriptors (or categories) from the taxonomy. To support the decision process, it allows the user to investigate the evidence supporting the system's recommendations. The Indexing Assistant is currently used as a front end to a repository containing more than 6000 business cases shared through a Lotus Notes database.

Figure 1 illustrates part of this knowledge management process. Document X is a text describing some activity (e.g. a consulting engagement with a client). X is first routed to an information specialist for approval. The Indexing Assistant then helps the information specialist select attributes from a controlled vocabulary in order to generate a structured representation of the document content. The result is a structured representation of the text containing descriptors representing what the business story is about (what processes were used, what methods were used and what measures are used to assess the results). Personal and company names are not part of the taxonomy, but can be added separately as the byproduct of automated workflow (for example, when a document is created in Lotus Notes the name of the author is automatically added as a descriptor of the document).

The saved document is then ready to be processed for retrieval using full text search, category based retrieval (show engagements that use "knowledge sharing" as an enabler), similarity-based retrieval (CBR), or more general data mining techniques.

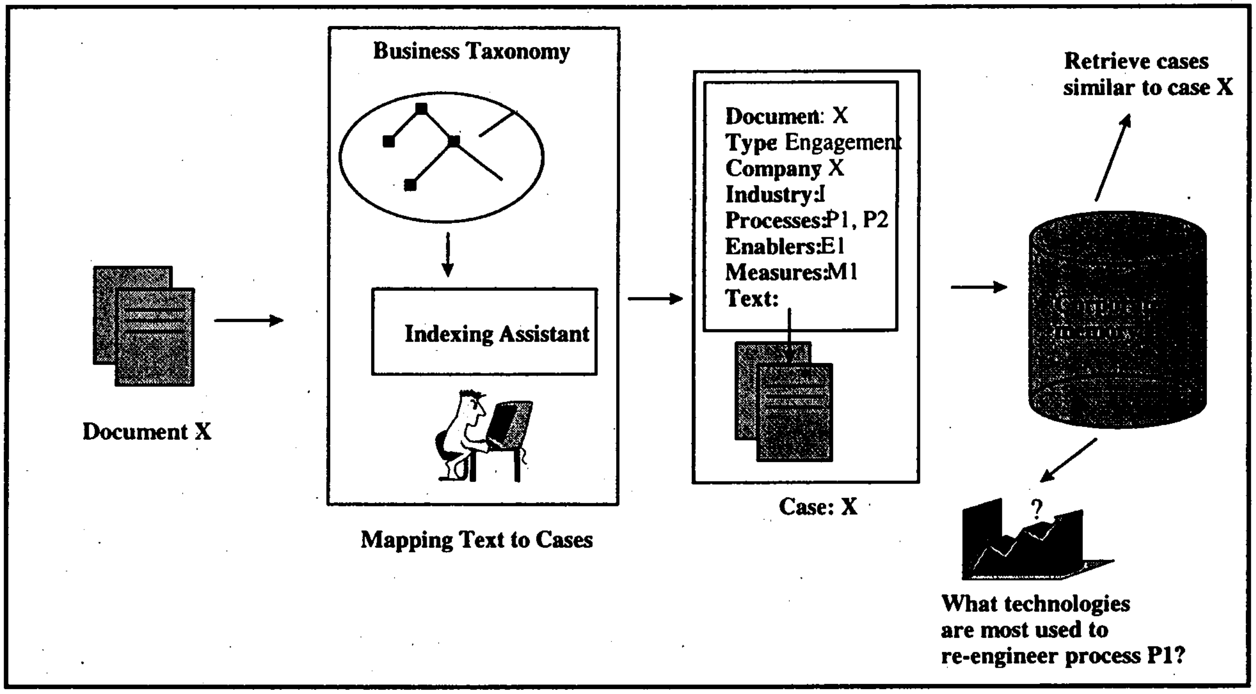


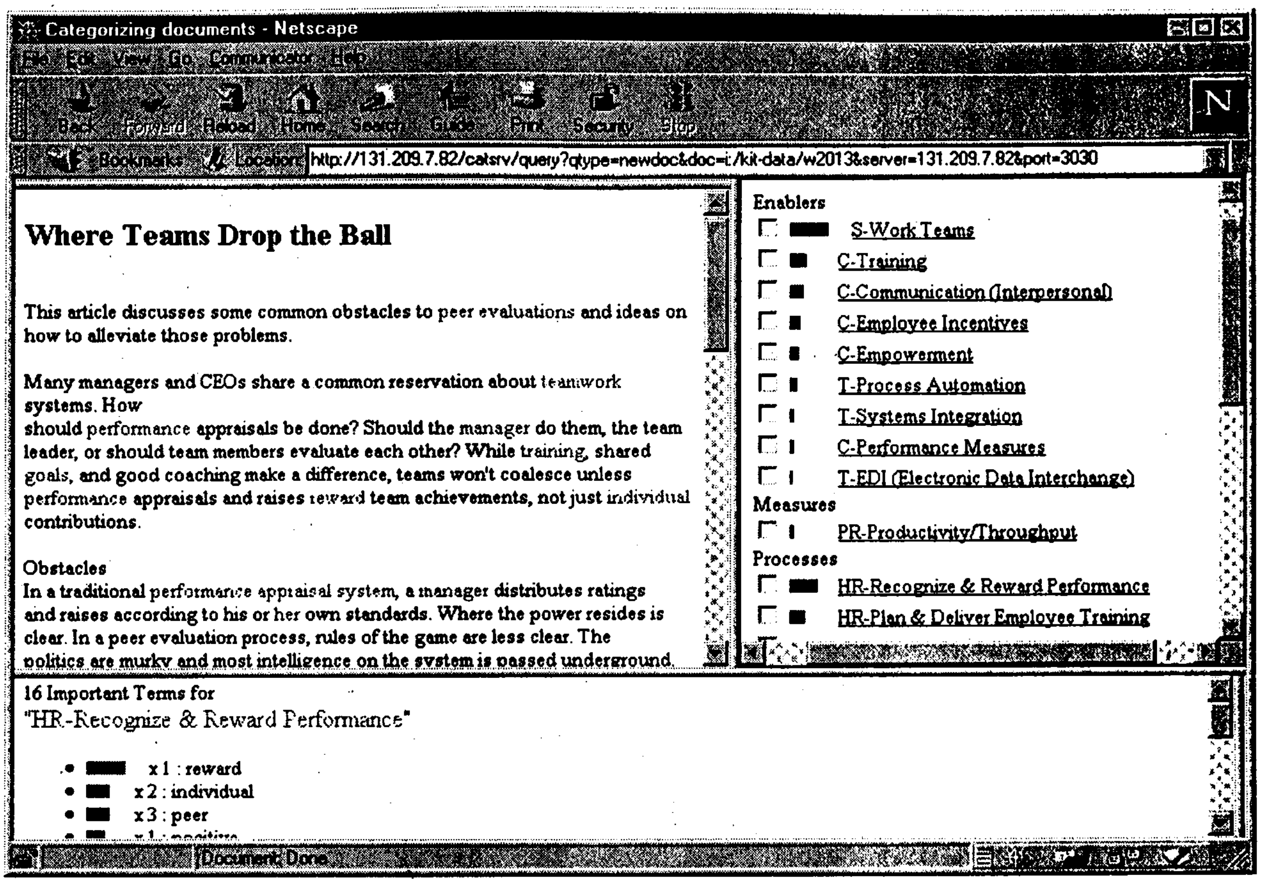
Figure 1: Structuring and Processing Cases for Knowledge Management

The following subsection describes the Indexing Assistant in more detail and shows how the system interacts with human indexers to help them turn text documents into structured business cases.

The Indexing Assistant

The Indexing Assistant integrates a trainable text categorization tool as in (Lewis 96, James and al 95, Borko and al 65). The input to the training component is a set of examples categorized manually, the output of this training phase is a dictionary of weighted term/category associations that can then be used to automatically categorize new documents.

We trained the categorization component on a set of two thousand manually indexed cases representing best practices, engagements with clients, and benchmarking studies. After the system has been trained, the categorizer takes a document and presents the user with a weighted list of candidates. The user can then select a category to see the terms in the document that are highly correlated with the category before selecting or rejecting the system's recommendations. At the end of this process the categories selected by the user are added to the case representation.


Categorizing documents - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Document: Done

http://131.209.7.82/catsrv/query?qtype=newdoc&doc=i:/kit-data/w2013&server=131.209.7.82&port=3030

Where Teams Drop the Ball

This article discusses some common obstacles to peer evaluations and ideas on how to alleviate those problems.

Many managers and CEOs share a common reservation about teamwork systems. How
should performance appraisals be done? Should the manager do them, the team leader, or should team members evaluate each other? While training, shared goals, and good coaching make a difference, teams won't coalesce unless performance appraisals and raises reward team achievements, not just individual contributions.

Obstacles
In a traditional performance appraisal system, a manager distributes ratings and raises according to his or her own standards. Where the power resides is clear. In a peer evaluation process, rules of the game are less clear. The politics are murky and most intelligence on the system is passed underground.

Enablers
- □ ■■■ S-Work Teams
- □ ■■ C-Training
- □ ■■ C-Communication (Interpersonal)
- □ ■ C-Employee Incentives
- □ ■ C-Empowerment
- □ ■ T-Process Automation
- □ ▮ T-Systems Integration
- □ ▮ C-Performance Measures
- □ ▮ T-EDI (Electronic Data Interchange)

Measures
- □ ▮ PR-Productivity/Throughput

Processes
- □ ■■■ HR-Recognize & Reward Performance
- □ ■■ HR-Plan & Deliver Employee Training

16 Important Terms for
"HR-Recognize & Reward Performance"

- ■■■ x1 : reward
- ■■ x2 : individual
- ■■ x3 : peer

Figure 2: Indexing Assistant interface

Figure 2 illustrates part of a typical interaction with a user. The text of the document being indexed is shown in the top left frame.

Document Categorization: The system performs a global analysis of the document content by categorizing the text. The top N (20 shown here) ranked categories are presented in the top right window and are grouped into the major divisions of the taxonomy. The strength of each recommendation is indicated by the width of a slider.

Category Evaluation: Categories can be compared by obtaining list of terms in the text that are high indicators of these categories. In this example the system suggests that the "Recognize & Reward Performance" process and the "Work Team" or "Training" enablers might be good categories to describe the business problem in the document and the method used to solve it. When the user selects "Recognize & Reward Performance" from the list of suggested categories, the system displays a list of terms in the text that relate to this category (bottom frame). In this case, the system displays terms such as *reward*, *individual* and *peer evaluation* that are high indicators of the selected process. These terms are highlighted in the text to help the user assess the recommendation in the context of the story. This kind of interaction with the system allows the user to elicit information about the relationship between document and categories, and helps provide the justification required to make categorization decisions.

Category Selection: The user accepts a set of system recommendations for describing the document by checking the boxes associated with one or more categories. In this example the resulting case might be coded as Enabler: *Training*, Process: *Recognize & Reward Performance* to indicate that this document is about solving a peer evaluation problem by training employees. Most documents in our application are associated with several categories.

Although the resulting case representation is too abstract to be useful as a stand-alone description of the document content, it is a powerful way of describing a story when it is associated with the text document. We are currently running experiments with professional corporate librarians to measure the impact of using such case structuring aid on indexing time, accuracy and consistency. Our preliminary results indicate that the assistive system improves indexing performance in all three aspects.

4. Conclusion

An essential part of knowledge management in large organizations involves sharing and reusing textual information. Using conceptual languages such as business taxonomies to describe information not only facilitates the retrieval of text documents but also uncovers implicit knowledge by detecting patterns in sets of documents (see examples in section 2). A key factor for being able to reuse qualitative information, such as lessons learned during consulting engagements with clients or best practice business reengineering stories, is to map text documents to conceptual languages based on domain models *designed by knowledgeable domain experts*. In our application, such domain models (i.e. a controlled vocabulary) are the main drivers for detecting key concepts in a text and for generating normalized case representations of business knowledge. However, because this "translation" process is time consuming and requires extensive training to be performed accurately, there is a need for tools that support the formalization of free text into structured cases.

We have described a method that uses text categorization technology to facilitate the creation and maintenance of case representations of business experience. Our approach is to first provide high quality normalized representations of textual information that can then be processed by CBR or data mining tools that operate on structured cases. This pre-processing approach to knowledge management is adequate for creating relatively small (in the thousands rather than in the millions) quality controlled collections of text. However, extremely large collections (e.g. the Web) cannot be pre-structured in this way, even with the help of assistive methods such as the Indexing Assistant. We are looking into trading some representation accuracy and using text categorization in a fully automated way to analyze text "on the fly" for processing larger text collections (Rissland & Daniels 95).

In addition to text categorization, there are other text processing methods for generating case representations from text. We are also investigating how to integrate text categorization with *information extraction* techniques (MUC-6, 1996) that can identify precise entities (people names, company names, quantitative values) and their relationships.

References

Allan, James, Lisa Ballesteros, James P. Callan, W. Bruce Croft and Zhihong Lu. 1995. "Recent experiments with INQUERY," in Fourth Text Retrieval Conference (TREC-4).

Borko, Harold, 1964. Research in computer based classification systems. In International Study Conference on Classification Research, 2^{nd} ed, Munksgaard, Copenhagen, 1965.

Emerson, J.,C. 1996. "Price Waterhouse's Knowledge View: Setting the Stage For Serving Clients In The Year 2000", in Emerson's professional services review. March/April issue.

Larkey, L.S. and W.B. Croft, 1995. "Combining classifiers in text categorization," in Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, Zurich, Switzerland.

Lewis, D.L., Schapire,R.E., Callan, J.P., Papka, R. 1996. Training algorithms for linear text classifiers, in Proceedings of the 19th Ann. Intl. ACM SIGIR Conf. On Rsch. And Dev. In Information Retrieval.

MUC-6, 1996. Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufmann.

Porter, Michael,E., 1990. "Competitive Advantage". The Free Press. Collier Macmillan Publishers London.

Rissland, E.L. and Daniels, J.J. (1995). "Using CBR to Drive IR." In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), 400-407. Montreal, Canada.