Knowledge Management through Content Interpretation

RICHARD R. JONES,
BERNT A. BREMDAL,
CHRISTOPHE SPAGGIARI,
FRED JOHANSEN,
ROBERT ENGELS
CognIT a.s, PB 610, 1754 Halden, Norway

ABSTRACT

The improved performance of computer-based text analysis represents a major step forward for knowledge management. Reliable text interpretation allows focus to be placed upon the content of documents, rather than just the document wrapping, and this helps to emphasise the fundamental difference between knowledge management and document management. It is not uncommon for companies who wish to join the KM band-wagon to re-package existing document management programs with a KM label, even if such programs offer little more than a hierarchical file system and simple key-word search to support content management.

In this paper we present CognIT's "Corporum" text analysis technology that is able to extract automatically the essential context of a given piece of text, and compare it with other texts to test whether they contain any overlap in contextual relevance. Therefore the technology can underpin several key knowledge management areas, including advanced search and retrieval, multi-dimensional text classification, meta-tagging, auto-summarising, portal building, business intelligence, site surveillance etc. Performance is sufficient on a ordinary desktop PC to analyse 100s - 1000s of texts per hour.

The technology contains two essential elements. Firstly, the content of a given text is analysed thoroughly, and the contextual knowledge it contains is encapsulated automatically in the form of a detailed semantic net (ontology). The second element is then able to compare this knowledge representation with any other texts retrieved, using a neuro-fuzzy analysis method. This allows an estimate to be made of the document's relevance with respect to the original text, and also enables a justification of the analysis to be provided to the user in the form of a brief textual explanation outlining the relevance of the document.

The current implementation of the technology is optimised for the English language. At present it is only able to interpret contextual relevance rather than intentionality. The system is not foolproof, but generally has a performance broadly comparable to a smart teenager. Ongoing research & development has already identified areas of major improvement.

INTRODUCTION

In much of the literature published within the rapidly growing field of Knowledge Management (KM), most emphasis has been placed on the potential benefits of KM from a managerial perspective. In this sense, Knowledge Management as a discipline has so far focused more upon “management” than on “knowledge”. This paper describes the philosophy, implementation and performance of a specific piece of commercial technology that extracts and manipulates knowledge, so that it can be represented in a way in which some of the benefits of KM already identified can actually be realised.

DOCUMENT MANAGEMENT vs. KNOWLEDGE MANAGEMENT

The obvious need for software tools that can support the demands of KM has lead several software manufacturers to re-package their existing product line, and market it strongly as a KM solution. For example, KM as currently championed by Microsoft consists of pre-existing back-office products re-wrapped in a front-end that fundamentally provides file management but little more. Functionality for information retrieval is still limited to simple keyword-based search already present in the operating system. It is important that a clear distinction is made between Document Management and Knowledge Management - there is a tendency for companies with existing Document Management Systems (DMS) to push their products under the KM banner. Most Document Management Systems regard each document as a single entity, and the *content* of the document is not considered (i.e. content is not “managed”). Even a

DMS in which meta-information about each document can be added (such as title, author, category, keywords) should be regarded as offering KM functionality only at the most trivial level.

As an analogy, consider a typical modern-day postal service: this is a highly developed, optimised global system that is capable of delivering letters from sender to addressee quickly and reliably. The system treats each letter as the base entity that it must handle. But a postal service would have no meaningful function if its users were unable to read or write - the value of the system lies in the content of the letter/document (and particularly, the flow of data/information from the sender to addressee). The surrounding system is very important, but functionally is just the way in which the content is packaged and delivered. Knowledge Management must consider the content of documents, and not just treat documents at the "document-wrapper" level.

The internet represents a further example of the difference between document and content management. The rapid growth of the world wide web during the last decade was dependant upon the accepted TCP/IP, HTTP, URL and HTML protocols. These protocols, and the physical infrastructure that utilises them, represent a huge, effective document management system. However, the whole web would be impotent without the functionality provided by search engines such as Alta Vista, Infoseek, Yahoo, Excite etc. Search engines such as these are good, simple content management tools that have provided the bare minimum level of content management needed for the internet to function. The technology that we discuss in this paper uses artificial intelligence methodologies (from both "hard" and "soft" disciplines) to provide sophisticated content management through intelligent internet/intranet search.

WORKING DEFINITION OF KNOWLEDGE AND EXPERTISE

For the purposes of this paper we have adopted a broad definition of knowledge as being "familiarity gained by experience (of person, thing, fact); person's range of information; theoretical or practical understanding (of language, subject)" [1]. We also favour a pragmatic definition of expertise, which we can consider here as "the practical application of knowledge in solving specific problems or tasks".

KNOWLEDGE TYPOLOGY

Nonaka and Takeuchi [2] have discussed the difference between tacit and explicit knowledge. In this paper we are primarily concerned with *explicit* knowledge [2] that is represented in the form of electronic text documents (as discussed by Beckman [3]).

In today's commercial world, many workers spend considerable amounts of time writing documents directly related to their work experiences, in which they typically record some of the following:

- raw data gathered by experimentation, hands-on experience, and/or observation
- description of a particular event or specific case
- interpretation of data
- beliefs, guesses, hunches, heresay
- insight, ideas, theories, opinions
- conclusions, summaries, recommendations, judgements, proposed actions

It is a central assumption in our approach to KM that such documents are actually able to convey an explicit representation of at least some of the knowledge and expertise of their authors. This is a common-sense assumption - the author of a business document intends to convey specific information to the reader, and although the quality of written documents does vary (i.e. some authors manage to convey meaning more effectively than others), in practice, documents are never random collections of unrelated words! In fact, according to Feldman [4], up to 80% of a company's explicit knowledge is typically stored in the form of text documents, rather than in formally structured databases. This emphasises that the corpus of documents within a company represents an extremely valuable repository of knowledge, and that knowledge sharing between employees can occur as long as we are able to access and utilise the content of the texts.

KNOWLEDGE CAPTURE

In addition to sharing knowledge that already exists within the finite scope of a company intranet, technology that supports intelligent content management also has enormous potential for the capturing of "new" knowledge from outwith the intranet, by searching on the (almost infinite) world wide web. By "new" knowledge in this sense, we really mean pre-existing knowledge (from remote sources) that previously is unknown to those within the company, until we locate, analysed and retrieve it from the internet and incorporate it into the

company's body of knowledge (or "Corporate Memory"). We postulate that within most business domains there are huge amounts of relevant information readily available and awaiting capture on the internet, an assertion that is strongly supported by our analysis of the content of several million documents related to a number of diverse business domains, including the following:

- aerospace development
- petroleum exploration and production
- shipping industry
- micro-processor technology
- medical and health services
- financial, economic and political news
- Knowledge Management
- and many others.

CORPORUM AND CONTENT MANAGEMENT

"Corporum" is a KM product line based on CognIT's content interpretation technology. The underlying core technology consists of two main parts:

- functionality based on natural language processing that can automatically extract the essential information from a piece of text, and represent this in the form of a semantic net (a basic ontology).
- functionality based on neuro-fuzzy reasoning to compare the semantic nets with other texts to test for any overlap of interest. This is the basis of sophisticated text search.

The format and presentation style in which a specific Corporum-based product presents the results of search/comparison to the end-user, depends upon the exact purpose of the product. In other words, the same technological core components have been wrapped into quite different system architectures, with a corresponding range of different graphical user interfaces. One such product is the Corporum Business Intelligence Portal [5], which is designed to allow large areas of the internet firstly to be searched and filtered for relevant information, and then subsequently to be monitored for ongoing changes and updates.

AUTOMATIC GENERATION OF SEMANTIC NETS

This core Corporum technology utilises natural language processing (NLP) strategies to analyse automatically a given piece of text in order to extract the most essential themes and concepts present in the text. The analysis performed by the system enables such concepts and their inter-relationships to be represented in the form of a semantic net (or ontology, *sensu lato*).

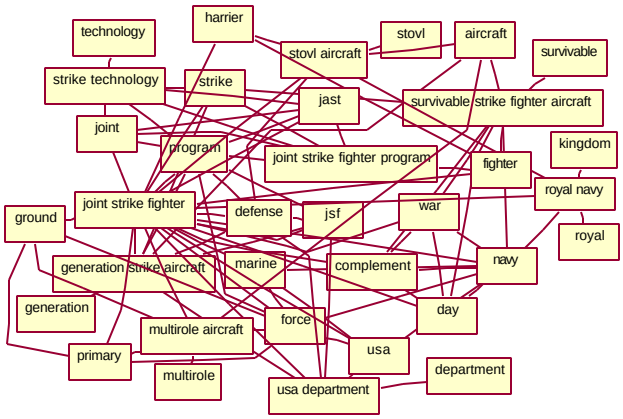
For example, when presented with the following text:

"Joint Strike Fighter Program.

The Joint Strike Fighter (JSF) Program is the next generation strike aircraft from the USA Department of Defense. The Joint Strike Fighter will be used by the US Navy, US Air Force, US Marines, and allies of the USA. The Joint Strike Fighter Program was previously called the Joint Advanced Strike Technology (JAST) Program.

For the US Navy, the Joint Strike Fighter will be a first day of war, survivable strike fighter aircraft to complement the F/A-18E/F. For the US Air Force, the Joint Strike Fighter will be a multirole aircraft (primary-air-to-ground) to replace the F-16 and A-10. For the US Marines, the Joint Strike Fighter will be a STOVL aircraft to replace the AV-8B and F/A-18. For the United Kingdom Royal Navy, the Joint Strike Fighter will be a STOVL aircraft to replace the Sea Harrier."

… the system automatically creates the following schematic semantic net:



The semantic net is represented in binary form internally in the system – we show them graphically here to demonstrate the internal capability of the core technology. During creation of the net nodes and links are assigned certain characteristics, but for the sake of clarity these are not shown here.

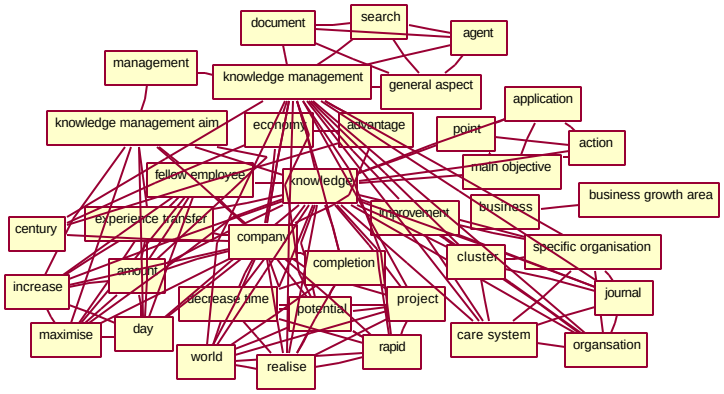
As a further example, the system can represent the following short description of KM with the semantic net shown below:

"This is an intelligent agent that can search for documents related to general aspects of Knowledge Management.

Corporate Knowledge Management aims to maximise the re-use of knowledge throughout a company, and to increase the amount of knowledge sharing and experience transfer between fellow employees during their day-to-day business. Knowledge Management is a rapid growth area, as companies around the world have started to realise the enormous potential of knowledge reuse to help to reduce project costs and decrease time to completion. Effective Knowledge Management will help

companies to achieve competitive advantage in the dynamic global economy of the 21st century.

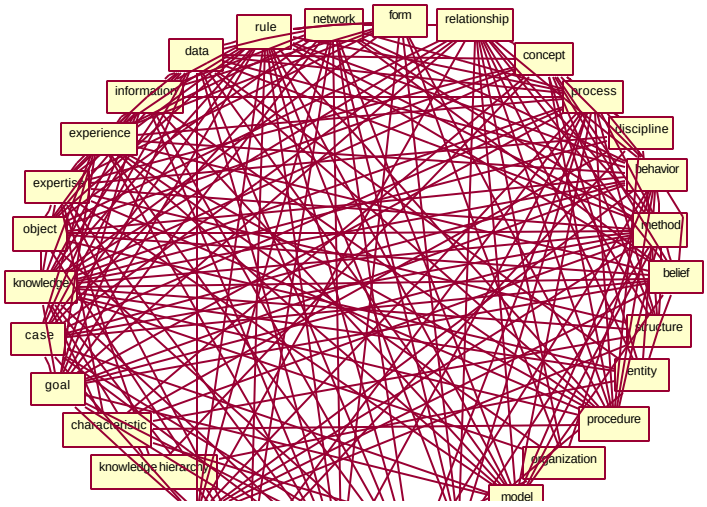
The "Knowledge Management" Journal has defined KM as "the organisation and improvement of a care system of knowledge within a specific organisation or cluster. Its main objective is to



enable an optimal application of knowledge at the point-of-action".

Being able to produce automatically this kind of semantic net represents enormous time savings - the system typically takes a few milliseconds to analyse a text, whereas we usually needed many hours to complete the task when we prepared ontologies manually (which we did in order to validate the accuracy of system performance). An additional complication that we encountered during our validation process was that knowledge representations made by human experts were usually based on subjective decisions involving knowledge of a very tacit nature. Not surprisingly, making this knowledge explicit often proved difficult. In contrast, the NLP rules and heuristics used by Corporum are expressed *explicitly* in the form of coded algorithms, and the system performs predictably in a way that is understandable to those of us who are well acquainted with the detailed implementation of the code.

By automatically analysing large (or very large) texts it is possible to build semantic nets of



considerable size. By way of example, we have analysed the paper by Beckman presented in these conference proceedings [3]. The resultant ontology contained over 800 thematic nodes, which were inter-related by several thousand links. Visualising the whole ontology is not practical (in paper format), but the system allows us to focus upon the themes that are most central to the overall subject matter of the text. A sub-ontology showing 30 of the most important concepts is depicted below. The pervasive “many-to-many” relationships between the nodes shows clearly that these nodes play a important part in filling the knowledge space represented by the whole ontology. Concepts of less central importance tend to have much fewer links, and lie at the peripheries of sparse clusters of nodes.

An important limitation in the currently available version of the Corporum software is that in general only *contextual* information will be extracted from the text. At present, any intentionality expressed within the text is not captured. For example, when analysing the sentence; “Margaret Thatcher is a human”, the system identifies two concepts (firstly “Margaret Thatcher”, and secondly “human”) and understands the relationship between the two. However, the system identifies the same two concepts, but no significantly new information, when confronted with a sentence that is contextually similar, but which has a radically different meaning, such as: “I do not believe that Margaret Thatcher was human”.

Our NLP core technology as used today in the commercial version of Corporum is able to produce reasonably good semantic representations of the knowledge content of a text, but the process can clearly be improved in many ways. At present, the system performs better with English texts than texts written in other languages. So far we have optimised automatic ontology generation more for speed than accuracy - that the ontologies generated automatically can be successfully used for text search is partially due to the robustness of the algorithms used by the second core technology, "Search using semantic nets", outlined in the following section.

SEARCH USING SEMANTIC NETS

Once the system has created a semantic representation of a text, we can analyse the content of further texts to test for contextual overlap. To do this we use a hybrid neuro-fuzzy strategy which is rapid, robust and performs well. For commercial

reasons we do not describe this method in more detail, but summarise some of the advantages of the approach as follows:

- the relevance of a text (in relation to another) is measured by the degree of contextual overlap, and can be expressed in fuzzy terms, rather than the boolean approach used by most existing search engines.
- the system can justify its ranking of relevance in several ways (none of which are supported using pattern-matching search strategies), by:
  - providing an free-text explanation of the amount and nature of contextual overlap
  - showing the central themes present in each text analysed
  - presenting the most important sentences within a document
- texts can be classified and grouped automatically in multi-dimensional, dynamic hierarchies according to knowledge content, forming the basis of a flexible corporate memory.

In short, we try to analyse, manipulate, filter and store the knowledge content of a text, rather than merely searching for the presence or absence of a particular keyword, key-phrase, or bit-pattern.

SUMMARY

- Many companies tend to generate and store large amounts of written text documents as part of normal work processes. Such documents encapsulate a significant amount of a company's knowledge in an explicit form.
- For most companies, the world wide web represents an additional source of knowledge that also can be of enormous commercial value.
- Retrieving, analysing, filtering and storing the *content* of documents within a company intranet and on the internet is a key strategy to unleashing the value of the knowledge contained in such documents.
- CognIT's "Corporum" technology is able to extract and represent the knowledge content of text documents by automatically generating a semantic network of concepts contained in each document.
- The system can then use the semantic network generated automatically, as the basis for sophisticated text search tools that enable knowledge management through content interpretation.

REFERENCES

[1] Sykes, J.B. ed. *The Oxford Dictionary of Current English*. Clarendon Press. 1978.

[2] Nonaka, I. and Takeuchi, H. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press. 1995.

[3] Beckman, T. Meta-Knowledge and Other Knowledge Dimensions. *Proceedings from the 3rd AI and Soft Computing Conference*. IASTED. 2000.

[4] Feldman, R. Text Mining: Theory and Practice. 4th World Congress on Expert Systems - Application of Advanced Information Technologies. ITESM Mexico City Campus. 1998.

[5] http://www.corporum.com/

KNOWLEDGE FUSION

Often into best practise documents, summarising the companies accepted method of how to perform a particular work procedure.