Managing the risks from information — through semantic information management

P W Warren and N J Davies

Information is central to the way organisations work — and with the increasing quantity which organisations must manage, the risks from failure to optimally manage information become ever more significant. The typical organisation uses a wide range of heterogeneous information sources, with differing schemata. This means that information cannot easily be gathered for reporting purposes, and significant connections may be undetected. Moreover, it has been estimated that only 20% of corporate information is in ‘official’ database systems. The remaining 80% is unstructured text, i.e. e-mails, intranet pages, documents, spreadsheets and even transcripts of conversations. This creates risks. Unstructured, informal information is often associated with individuals and when individuals leave, although the information may remain, practical access can be lost. Knowledge management tools are needed to prevent this. More dramatically, when faced with litigation or regulatory requests, companies need effective eDiscovery so as to provide, and be seen to be providing, all the relevant documentation. An important tool in integrating heterogeneous information, and in coping with unstructured information is semantic information management. This paper describes the semantic approach to information management and how it can reduce the risks posed by information. It also describes a research project SEKT, led by BT, which is extending the capability of this semantic approach.

1. Introduction

All organisations are managed through information. This must have been true from the earliest times and the earliest civilisations. Modern organisations are characterised by the availability of massive volumes of information, made possible by electronic technology. This apparent advantage brings with it increased risks.

Most obviously, the organisation that does not capture, analyse and disseminate information as effectively as its competitors, will suffer the consequences. The ability to find and share information rapidly and effectively is a clear commercial advantage in all sectors.

So, too, is the ability to retain information. Maintaining the corporate memory bank as employees leave the company is an oft-discussed problem — and one seen as increasingly important given the demographic profile in the western world as the post-war generation of ‘baby-boomers’ contemplates retirement.

To these commercial risks have been added regulatory ones. Organisations which do not disclose all relevant information to regulatory authorities may be seriously penalised. Yet the organisation can only disclose information it knows it has. Information lost on corporate computers can not be disclosed at the appropriate time — but will certainly be revealed if the organisation is subject to a detailed forensic analysis of hard drives prior to a legal hearing.

The last few decades have seen the increasing development and implementation of corporate applications centred on the relational database. Information about customers, products, employees and even competitors is stored in well-organised tables, defined by database schemata. We call this structured data. Since Codd\(^1\) first proposed the relational model, this technology has been highly successful in enabling corporate applications.

Still, the technology has limitations. The structure of the data is defined in database schemata, embedded in the application and hard to change. Different applications, in the same organisation and in collaborating organisations, have different schemata, and this gives rise to the problem of interoperability. Without a common view across heterogeneous data, we cannot easily merge data from different sources (e.g. for reporting purposes), or compare data and detect significant connections.

Even more threateningly, while applications based on relational databases have gained greater and greater

---

\(^1\) For a brief biography of the father of relational database theory, see the article in Wikipedia — http://en.wikipedia.org/Edgar_F._Codd
sophistication, the availability of cheap computing hardware and applications such as word processing and e-mail has seen the massive growth of unstructured data. Documents and presentations on desktop machines, e-mails and associated attachments and intranet pages on corporate servers all contain valuable corporate information. This information is not organised into schema-defined tables; what technology exists for managing this information is almost wholly dependent on textual string processing. Additionally, intermediate between the structured and the unstructured is a significant amount of data stored in various semi-structured and tabular forms, but where well-defined schemata do not always exist. The most notable example of this is the spreadsheet.

Forrester [1] estimates that unstructured data comprises 80% of the information within an organisation. In any case, the precise percentage is hardly important; we are all of us aware that the proportion of unstructured data is large and growing.

The regulatory risks discussed earlier are particularly prominent here. Murphy and Markham [2] quote as an example a $1.4 billion judgement against Morgan Stanley, arising from the latter’s inability to produce requested information. The inference drawn was that Morgan Stanley was deliberately hiding e-mails.

Indeed, the growing use of e-mail is a particular problem. AIIM [3], a non-profit organisation in the electronic content management industry, confirms that e-mail is a central means used for business documentation. Over 70% of respondents to an AIIM survey reported exchanging confidential or sensitive information via e-mail. AIIM found e-mail is being used for critical processes such as contract negotiation, HR discussions and invoice delivery. At the same time, nearly one-third of respondents reported that the Sarbanes-Oxley Act has affected the way their organisation views e-mail — so there is at least some awareness of the risk posed by regulatory legislation.

2. The importance of semantics

The preceding section identified the two major challenges. We need to integrate information held in heterogeneous databases, and we need to manage unstructured data as well as we manage structured. The second of these means that we need to extract structure from the unstructured text. When we have achieved these two challenges, a final step is to combine structured and unstructured information into a co-ordinated view. An understanding of semantics underlies both challenges. Integration across data sources is a long recognised problem. At the syntactic level, e.g. converting between different file formats, it is troublesome but relatively straightforward to overcome. However, when such technical problems have been solved, there remains the harder problem of overcoming semantic incompatibilities. This point has been made by McComb [4] in a book which provides a good introduction to the business benefits of using semantic technology. As mentioned above, pre-existing data silos will typically model the same data in conflicting ways. These types of conflict usually require more extensive semantics-based solutions. Figure 1 shows two examples of semantic conflict that can be found across data sets. These conflicts are very frequent, occurring as a natural consequence of data modeling — whether due to isolated development, changing needs, organisational or structural differences, or simply the different approach of two human data modellers.

This is a problem across applications and departments within individual companies, and even more when companies co-operate, e.g. in joint ventures.

The challenge of unstructured text is, as we have seen in the last section, if anything even more pressing. Here, the traditional approach is based on text-string matching. This may be simply through a user initiating a search, or through text searches embedded in an application. In any case, there are several problems with this approach, which can be divided into four main areas.

- Query construction

In general, when specifying a search, users enter a small number of terms in the query. Yet the query describes the information need, and is commonly based on the words that people expect to occur in the types of document they seek. This gives rise to a fundamental problem, in that not all documents will use the same words to refer to the same concept. Therefore, not all the documents that discuss the concept will be retrieved by a simple keyword-based search. Furthermore, query terms may of course have multiple meanings (query term polysemy). As conventional search engines cannot interpret the sense of the user’s search, the ambiguity of the query leads to the retrieval of irrelevant information.

Although the problems of query ambiguity can be overcome to some degree, for example by careful choice of additional query terms, there is evidence to suggest that many people may not be prepared to do this. For example, an analysis of the transaction logs of the Excite WWW search engine [6] showed that Web search engine queries contain on average 2.2 terms. Comparable user behaviour can also be observed on corporate intranets. An analysis of the queries submitted to BT’s intranet search engine over a four-month