

The Intertek Group

Management Report on
**LEVERAGING UNSTRUCTURED DATA
IN INVESTMENT MANAGEMENT**

The Intertek Group
94, rue de Javel F-75015 Paris
Tel: +33 1/45 75 51 74
www.theintertekgroup.com

Acknowledgements

This report is based on conversations with over 50 persons from financial institutions, content providers, technology vendors, representatives from industry consortia working on establishing XML standards, and researchers from academia and industry. The authors wish to thank all those who contributed by sharing with us their experience and their views.

Information contained in this report is based on the best available sources, but its accuracy cannot be guaranteed. Opinions reflect judgement at the time and are subject to change without notice. The entire contents of this publication are copyrighted by The Intertek Group and may not be reproduced or retransmitted in whole or in part without express permission.

© Copyright The Intertek Group, May 2002 - All Rights Reserved.

Contents

Management summary	5
Information overload?	5
Defining the problem.....	8
1. Functionality.....	11
Core functionality.....	11
Taxonomy generation.....	12
Categorization.....	14
Search and navigation.....	17
Profiling / filtering.....	18
Advanced / complementary functionality.....	18
Visualization.....	18
Summarization.....	19
Automatic real-time translation.....	19
Audio-visual	20
Knowledge extraction.....	20
2. The take-up in finance	21
The business case	21
Generating taxonomies and categorization.....	22
Research portals / business intelligence	23
Integration and hosting services for research portals	24
Automatic handling of correspondence.....	25
Compliance applications.....	25
CRM applications	26
Combining data and text mining	27
3. From unstructured to semi-structured data.....	29
XML: bringing structure to "bags of words"	29
XML dialects in the universe of finance	30
Information-oriented financial MLs	30
XBRL	30
RIXML	31
Other XML standards in finance	32
One standard, no standard, 100,000 standards	34
Implications for "publishers," end users and technology vendors.....	34
From XML standards to the Semantic Web	36
4. Text mining and IT issues	38
Enabled and native XML databases	38
XML and the data model.....	39
XML query languages: one per standard?.....	40
5. The market.....	41
Factors driving the take-up of text mining technology.....	41
Factors holding back wider implementation	41
Structuring of the supply side.....	43
The challenge to vendors.....	44
6. Technology backgrounder	46

Semantics and data structures.....	46
The origins of text mining.....	48
Finding similarity in meaning with statistical techniques.....	50
Generating taxonomies with unsupervised clustering techniques.....	50
Categorizing with supervised learning techniques.....	51
Improving on results with linguistics.....	51
7. Supplier directory.....	52
Autonomy.....	52
ClearForest.....	53
Documentum.....	54
eXcelon.....	54
Factiva.....	55
FactSet Research Systems.....	56
Fast Search.....	57
IBM.....	57
Insightful.....	58
Interwoven.....	58
Inxight.....	59
LexiQuest.....	59
Megaputer Intelligence.....	60
Microsoft.....	60
Multex.....	61
Oracle.....	62
SAS.....	62
Semio.....	63
Software AG.....	63
SPSS.....	64
SRA.....	64
Temis.....	65
Thomson Financial.....	65
Verity.....	66
For more information.....	67
Abbreviations used.....	67
About The Intertek Group.....	68
About the authors.....	68

Management summary

Information overload?

Information is the raw material and analytics the machinery for the "manufacture" and sale of financial products. Times series analysis has benefited from data mining techniques, now extensively used throughout the industry to engineer products, manage risk and profile clients. Textual information has remained largely outside the domain of automatic handling, but this is now changing.

Though a commonplace, financial firms are confronted with the problem of information overload:

- Reuters publishes the equivalent of 3 bibles of (mostly financial) news daily;
- It's estimated that 5 new research documents come out of Wall Street every minute;
- Asset managers at medium-sized firms report receiving up to 1,000 e-mails daily and work with as many as 5 screens on their desk.

Conversely, there is also a dearth of (processed) information. It has been estimated that only one third of the roughly 10,000 US public companies are covered by meaningful Wall Street research; there are thousands of companies quoted on the US exchanges with no Wall Street research. It is unlikely the situation is better relative to the tens of thousands of firms quoted on other exchanges throughout the world. Yet increasingly companies are providing information including financial results on their Web site, adding to the more than 2 billion pages now on the World Wide Web.

Any manual solution to the problem would be costly and ineffective. One widely adopted solution is to simply ignore much of this information, relying on a small number of trusted sources. In the globalized world of finance, this non-solution can prove to be expensive; the technology is now there to help.

A quiet revolution is taking place in the way unstructured information, in particular textual information, is handled. A key aspect of this change: unstructured information is progressively being transformed into self-describing, semi-structured information that can be managed by computers.

Technologies that allow computers to "understand" the content of documents are now widely used in basic functions such as free text search and navigation. This basic functionality is already on the desktop - thanks to web search engines and industry content providers - without any need to know the technology or plan for its implementation. But leveraging the technology throughout the firm to gain an information advantage or to further structure and automate processes such as the investment management process will require much more.

The components in the revolution in handling unstructured data are:

- the development of (XML) standards for tagging textual data, which are taking us from free text search to queries on semi-structured data;
- the development of (RDF) standards for appending metadata, which provide a description of the content of documents;
- the development of algorithms and software that generate taxonomies and perform automatic categorization and indexation;
- the development of database query functions with a high level of expressive power;
- the development of high-level text mining functionality that allow "discovery".

The emergence of standards for the handling of "meaning" is a major development. It implies that unstructured textual information, which some estimates put at 80% of all content stored in computers, will be largely replaced by semi-structured information ready for machine handling at a semantic level. The eXtensible Markup Language (XML) and its Resource Description Framework (RDF) are already a reality. Industry- and application-specific standards are being developed around the general-purpose XML. In finance, standards are being defined to stipulate how the entire universe of

financial information, from time series to analyst and corporate reports and news, will be described. This will greatly facilitate text mining.

The diffusion of XML-based standards and text mining functionality will have consequences on the way the business is managed and, ultimately, on performance:

- Searching large masses of textual data from internal and external sources becomes much more powerful.
- The statistical and logical analysis of textual information (e.g., news, research reports, corporate information) is facilitated, allowing the extraction of meaningful new knowledge.
- The combination of data and text mining is facilitated, allowing, for example, to gauge the impact of events on price and volume.
- The representation of semantic links and visualization technology facilitate the navigation and understanding of masses of documents.
- Tasks related to the control of textual information for compliance to internal or regulatory rules can be automated.
- Communication with clients can be enriched and basic exchanges automated; clients can be profiled and their needs analyzed.
- Ultimately, business processes such as those in investment management will become increasingly structured; automation will be a step closer to realization.

The challenge to buy- and sell-side firms alike is multiple:

- *On the technology side*, it will require the ability to categorize, retrieve and perform queries on textual data, integrating semi-structured data from internal and external sources. Ultimately it will call for the ability to integrate data and text mining in applications such as fundamental research and event analysis, linking news and financial time series.

- *On the business side*, the ability to implement and leverage knowledge management on a firm-wide basis will become increasingly important. This will require a re-evaluation of business processes.
- *As for the skill set*, knowledge workers such as analysts and fund managers will have to be trained to extract maximum benefit from the technology. One problem today is under-usage of the technology already on the desktop.

Text-mining technology will have major implications for individual firms and will likely affect relationships within the industry. But it is unlikely that the technology will have a big effect on financial markets themselves. As occurred with the diffusion of data mining techniques, markets might become slightly more efficient; the shape of price processes might change somewhat, but the global behavior of markets will be largely unaffected.

Defining the problem

The question of machine-understanding came to the forefront with the World Wide Web and the need to search the more than 2 billion pages for content with practically no metadata (i.e., information that identifies the content and context of a document). Initially, brute force techniques were applied; more recently smarter (artificial intelligence) algorithms are being used.

The technology for searching and analyzing textual data is based on the ability of computers to handle the meaning (i.e., semantics) of content. While humans can read and understand texts, computers can not. We are still far from engineering machines that can mimic the human ability to understand and act upon written text; for the coming years (decades by some estimates), the problem of semantics must be sidestepped.

The term "unstructured" when applied to textual data is somewhat of a misnomer: textual data are not collections of random words but have a complex structure that

conveys a meaning too rich for machines to read. Computers require a simple repetitive structure of a hierarchical nature on which they can work sequentially according to the Von Neuman paradigm on which they are still based.

Against the explicit relations required by machines, textual data make implicit reference to a large mass of general knowledge. A sentence such as: “The CEO was forced to step down due to mounting pressure from major shareholders” is easily understood by a human reader because all terms and relationships can be placed within the context of broader knowledge. To enable a computer to perform semantic operations on such a sentence would require the explicit coding in the computer of a vast body of meanings and relationships.

Over the last four decades, the artificial intelligence (AI) community has developed a number of methodologies to explicitly represent knowledge in a computer; among these are systems of rules, frames and semantic nets. These methodologies represent semantics as explicit relationships between objects and are powerful systems for expressing knowledge in a way that machines can handle.

One problem in the explicit coding (or representation) of knowledge is the enormous number of entities and relationships that need to be considered to express an even modest amount of business knowledge. The largest expert systems developed can work only in a limited domain. The London-based fund management firm Pareto Partners, for example, developed an expert system of some 2,000 rules for its Global Bond Allocation Strategy system. Practical problems, including computational costs, put constraints on the semantics we can afford to implement; for the foreseeable future, the coding of knowledge must be limited to the essential.

However, even a limited semantics can be of benefit in handling unstructured data. By appending to every file of unstructured data *metadata* that encode its format and its *essential meaning* in a machine-readable form, it is possible to handle automatically vast amounts of unstructured data. Metadata is a sort of PostIt sticker which carries information on the document's content and context. Standards, namely the XML

standard, were developed to facilitate the metadata encoding of a document description; finance-specific XML languages are creating common definitions of objects in the universe of finance; the Resource Description Framework (RDF) provides a generalized language for describing semantic relationships.

Metadata contain a summary description of the document content - often with reference to a broad taxonomy - sufficient to 1) facilitate searches by human agents and 2) allow a software agent to locate the relevant texts and eventually to perform additional operations. Properly used, even this simplified semantics (i.e., essential meaning) allows easy and effective search and navigation of documents. Within this schema, the outstanding problems are 1) tagging documents, either newly created or legacy, 2) performing searches and 3) extracting insight or knowledge from the documents retrieved.

The technology for handling unstructured data consists of several layers:

- *Basic technology that “reads” an unstructured file, automatically tags the file, adding metadata, and searches and navigates the data.* Taxonomy generation, categorizing, indexing and tagging transform unstructured data into searchable entities on which queries can be performed.
- *High-level functions aimed at extracting knowledge from textual information (and/or sound and images) through semantic analysis.* Functions include summarization and conceptual maps that reduce the complexity of a text by extracting the most relevant concepts. Other functions are more similar to the discovery process in data mining; they might include the discovery of event/price relationships or the detection of credit problems in patterns of corporate news.

This Report begins with an overview of today's technology and then presents some representative implementations in finance. It subsequently discusses how emerging XML standards are changing the scenario, transforming unstructured data into semi-structured data and giving rise to a new set of tools for storing, searching and performing queries on structured text.

1. Functionality

Most of today's text mining solutions address problems associated with the of handling *unstructured* data on a large scale:

- Firms have legacy unstructured documents which some will find the need to categorize;
- News items and analyst reports are not yet being systematically tagged at the source with metadata;
- Today's business web sites are written in HTML which, while not providing a way to describe the content of a text, does accept keyword (but not semantic) searches;
- Communication with regulators, business partners and customers produces a large amount of unstructured data which many firms will want to handle and eventually analyze automatically.

We will see the accent shift progressively from the handling of unstructured data to the handling of *self-describing, semi-structured* data and the associated database and text mining technologies. This will provide the impetus behind a wider adoption of knowledge management inside financial firms.

This Section starts by looking at current core text mining functionality, then more advanced or complementary functionality.

Core functionality

Before texts can be searched (or mined), they must be given structure. One way to give structure to textual documents is to create a taxonomy (or hierarchical classification system) and subsequently categorize and tag documents according to the taxonomy. Structuring textual data at the source or adding structure to an existing corpus of textual data present different problems. Both require general principles on which the structuring will be performed, but while manual intervention might be feasible in structuring small

volumes of textual documents at the source, structuring large volumes of unstructured documents requires automation.

The core - and consolidated - functionality of text mining address these problems, providing technologies for:

- taxonomy generation;
- categorization and indexation;
- search and navigation;
- filtering.

Taxonomy generation

Taxonomies are the foundation of the "filing systems" of unstructured data in the computer. The need for taxonomies (i.e., hierarchical or tree-structure classifications) has grown with the huge amounts of textual information now stored in computers and the development of the web. Taxonomy generation is a major application area for text mining technology; typically part of the larger project of building a corporate search infrastructure, it is the first step towards categorization.

Taxonomies embody knowledge. The classical example comes from the natural sciences, with the taxonomy developed by the 18th century Swedish naturalist Linneus. Many firms have their own proprietary taxonomies, but one problem here is the need to communicate with business partners. There are already some industry-wide standards in finance. The Dow Jones subject codes are considered a de facto standard. The ISO (International Standards Organization) 15022 specification provides a standard set of more than 10,000 data fields for financial information and is a major resource in industry-wide efforts to arrive at XML standards. But assuming one adopts third-party taxonomies, there is still the need to integrate internal documents which might not fit easily into the pre-established classifications.

Per se, creating a taxonomy does not require technology: taxonomies can be built by knowledge workers who define subject codes in function of their judgment. It is the enormous number of "subjects" that need to be classified and the speed at which documents are created (e.g., hundreds of thousands of news stories daily) that creates the need for an automatic approach. Methodologies used to generate taxonomies automatically include rules, statistical analysis and linguistics or various combinations of these. Table 1.1 below schematizes the various methods currently used and their benefits and limitations.

Methods	Description and relative benefits/limitations (B/L)
Rule based	<p><i>Manual:</i> Explicit definition of categorization rules by knowledge workers. <i>Example:</i> ClearForest's Rulebook technology. <i>B/L:</i> Remains the most accurate method of categorization but has the drawback of cost and speed.</p> <hr/> <p><i>Automatic:</i> Rules are established automatically, using AI methodologies such as inductive trees. The automatic discovery of rules can be combined with statistical methods. <i>Example:</i> Verity uses logistic regression to automatically induce rules. <i>B/L:</i> Widely-used, fast method. A plus is the ability to add business rules.</p>
Statistical clustering	<p>Establishes categories by statistical clustering, grouping objects into maximally similar clusters. Requires a measure of similarity that reflects similarity in meaning and a corpus of texts for "training"; new items can then be classified automatically. <i>Standard clustering methods:</i> include k-means, hierarchical clustering, and mixture of distributions. Methods based on graph theory were developed specifically for web use. Bayesian methods and support vector machines (a kind of neural network) also used. <i>Examples:</i> Autonomy uses Bayesian networks; Megaputer and Verity use statistical clustering methodologies. <i>B/L:</i> Fast and tunable, but the quality of the classification depends critically on the size and representativeness of the set on which the learning methods were trained.</p>
Linguistics	<p>Analyzes texts and discovers the relationships among key terms and concepts. <i>Examples:</i> Insightful, Inxight, LexiQuest, Semio and Temis use linguistics to discover taxonomic relationships. <i>B/L:</i> Is a component, not a complete solution. Generally used in conjunction with pattern recognition or statistical methods.</p>

Table 1.1 - Summary of basic techniques used for creating taxonomies and their benefits/limitations.

In practice, most suppliers of tools for generating taxonomies use more than one methodology. For example, ClearForest uses a combination of rules and linguistics, Verity a combination of rules and statistics; Autonomy uses a combination of Bayesian statistics and neural networks, Megaputer a combination of pattern recognition techniques and linguistics; Insightful, Inxight, Semio and Temis use a combination of linguistics and statistics.

There are many possible taxonomies for any given subject, not all equally good. Among the features required of a taxonomy are:

- *Accuracy.* A taxonomy should be able to classify subjects unambiguously; the tension between accuracy and generalization is a big issue in machine learning.
- *Parsimony.* A taxonomy should be able to classify with a minimum of classification steps. Parsimony is also key to generalization and thus accuracy on new items.
- *Expandability.* A taxonomy should be able to include a new entity, expanding intermediate branches without the need to create a new top vertex.
- *Scalability.* A taxonomy should be able to include an arbitrary number of individual items and to be refined without limits.
- *Flexibility.* A taxonomy should be able to handle different domains as part of the taxonomy.
- *Ease of integration.* As many firms already have a taxonomy, a new taxonomy will be most useful if it can be integrated easily with existing taxonomies.

Categorization

The creation of a taxonomy (or a hierarchical set of subject codes) precedes the application of the coding to new items. Once the subject codes have been created, documents are categorized and indexed. Categorization assigns an index or a set of indexes to a textual document, specifying the category (or categories) to which it belongs. Unstructured data are thereby given some structure. Categorization is central to retrieval and distribution. Once a document has been categorized and indexed, it can be searched, queried, and filtered for distribution purposes.

There are many ways to perform categorization. As is the case with building taxonomies, categorization can be done manually. Manual categorization is performed by knowledge workers who assign a textual document to one of a predetermined set of categories inside the taxonomy, typically based on a set of rules. Providers of web search functionality have probably the most arduous categorization task today: they must routinely index the entire World Wide Web to provide up-to-date information on

demand. One reportedly employs more than 1,000 knowledge workers to categorize all the information on the web. More recently, automatic methods are being applied.

Automatic methods of categorization might be separated from or coupled with the methods for generating taxonomies. If coupled, the same system that generates taxonomies is used to categorize. For example, the automatic induction of rules is used to create expert systems that categorize automatically. In the training phase, the system learns the rules from a training set; when the rules are discovered, a machine “reads” new texts and assigns documents to the appropriate category as determined by the rules. Statistical methods work in a similar fashion: a statistical categorizer (such as k-means) learns the taxonomy in unsupervised mode from a training set; the categorizer then classifies any new items.

When uncoupled, statistical and pattern recognition technologies are often used to learn categorization from a set of documents that are already properly categorized. The same system is then used to categorize new items.

In addition to offering systems that generate taxonomies by clustering, most technology vendors also offer the ability to train their categorizers on given categories with learning algorithms such as neural networks and, more recently, support vector machines. Linguistics can be used on top of this to provide functions such as the disambiguation of text.

Table 1.2 below summarizes some of the more widely used methods for automatic categorization.

Methods	Description and relative benefits/limitations (B/L)
Rules	<p><i>Manual:</i> No training set needed but depends on the ability to formulate good rules. <i>Example:</i> ClearForest's Rulebook technology. <i>B/L:</i> Simple and effective.</p> <hr/> <p><i>Automatic:</i> Rule-induction algorithms generate the relevant classification rules on a training set and then apply the rules to the entire set of documents to be categorized. <i>Example:</i> Verity uses logistic regression to induce rules. <i>B/L:</i> Speed; allows the addition / fine tuning of business rules by experts.</p>
Statistical methods	<p>A machine learning approach in which clustering algorithms work on a training set, identifying (without prior "knowledge") clusters on which a taxonomy can be built. <i>Examples:</i> Insightful uses statistical methods for categorization. <i>B/L:</i> field proven methodologies; tunable.</p>
Pattern recognition	<p>A machine-learning approach; methodologies include neural networks and, more recently, support vector machines. <i>Examples:</i> Autonomy and Megaputer use a combination of pattern recognition techniques and statistics. <i>B/L:</i> quality of the categorization depends critically on the size and representativeness of the set on which the learning methods were trained.</p>
Linguistics	<p>Recognizes the basic syntactical structure of text. Often used together with pattern recognition or statistical methods to improve categorization. <i>Examples:</i> Insightful, Inxight, LexiQuest, Semio and Temis developed proprietary linguistic techniques. <i>B/L:</i> Gives a more precise understanding of the text to be categorized, helping to eliminate ambiguities difficult to detect with a purely statistical approach. In general insufficient alone for categorization.</p>

Table 1.2 - Widely used methods for automatic categorization and their benefits/limitations.

The best algorithms available today are estimated to categorize successfully in 75-85% of cases when working on hundreds of training documents in narrow data sets. On a typical intranet application, with a broad range of content and only 10-20 documents in the training set, accuracy in categorization drops to 50-80%. Categorization tools such as Inxight Categorizer have a degree of "introspection" (i.e., a confidence value as to the correctness of a categorization): when the machine is unsure about applying a code, it can turn to a human for help. This "request" for human intervention can be tuned for levels of accuracy, for example, a confidence threshold is set, say 90%; if the system evaluates the degree of confidence in applying a code below this threshold, it sends the problem to a human operator.

Even with automated categorization, human intervention is called. This is typically in the phases of selecting the training set, training the algorithm and reviewing categorization decisions.

Search and navigation

Methodologies to perform searches and navigate through information are central to knowledge management. Search and navigation functionality encompass a set of methods and algorithms for transforming queries expressed in natural or query languages into a standard database query. Rather than read and understand text on the fly, today's solutions work on sets of indexes and metadata; these well-structured data are suitable for standard machine handling.

Products currently available allow searches on key words or phrases with the addition of some syntax (i.e., structure). Today's search technology is widely considered mature, but insufficient; some corporate users are disappointed with retrieval capability. The big limitation is the inability to perform semantic searches unless a proper tagging system is in place (see Section 3 for a discussion of semantic searches). To a certain extent, the problem is more with the lack of tagging (in particular on corporate web sites) than with the technology itself. Search tools from leading vendors can perform semantic searches on corpora of properly tagged documents.

Some new functionality is coming on stream. Verity announced (July 2001) the ability to perform federated searches, extending the concept as the ability to search very large sets of documents from different points of view in parallel.

A perceived advance is the ability to formulate a query in natural language. This requires a natural language interface that transforms a sentence expressed in plain language into a string of codes that a computer can act upon. Inxight's LinguistX Platform, for example, allows the formulation of queries in up to 16 languages. It has been adopted by suppliers of text mining tools including ClearForest and Verity. While natural language query capability might facilitate the formulation of a query, it does not affect the expressive power of the search language and algorithms; this depends on the underlying query language.

Profiling / filtering

Today's profiling technology is considered adequate for business use where the right to access information on corporate intranets is defined by business rules or distribution rights. In broader searches, however, researchers are working on technologies that would automatically personalize filtering in function of the user's interests, matching and ranking search priorities. Example: an intelligent agent would learn a fund manager's areas of interest and, in performing a search, automatically match and rank items retrieved on the web in function of the fund manager's past behavior in searching.

Advanced / complementary functionality

Advanced or complementary functionality include:

- visualization;
- summarization;
- automatic real-time translation;
- audio visual;
- knowledge extraction.

Visualization

Visualization is used to 1) facilitate search and navigation and 2) reveal relationships across a large number of documents. Inxight's Star Tree and Table Lens are examples of the former. Examples of the latter include Autonomy's visualization tools on Classification Server which allow to look at, for example, hot spots and clusters, to gain a high-level view of how a situation develops over time; ClearForest's rule-based technology which is used to extract events or visualize complex relationships; and Megaputer's link analysis capability which is provided on PolyAnalyst, a data mining package with text mining capability. Vendors such as Megaputer believe that visualization will be even more critical in text mining than it proved to be in data mining.

Summarization

Summarization capability is available today on the desktop. To get a feeling for what the basic technology can deliver, click on the Summarization button in Microsoft Word. A widely shared evaluation is that summarization needs more work: today's technology is considered to fall short on coherence, readability and the ability to follow temporal patterns.

Nevertheless firms such as global news and business information content provider Factiva believe that summarization capability in products such as Inxight Summarizer already perform reasonably well when used on clearly written documents on a single theme. Inxight technology is part of the YellowBrix solution for CNNfn, providing real-time summarization on 39,000 words per second (see table 1.3 below).

End user	Real-time summarization capability
CNNfn	<i>Inxight Summarizer</i> technology (as part of YellowBrix's news content solution) provides real-time summarization on 39,000 words/second, allowing to tailor summaries on a functional basis for distribution to subscribers.

Table 1.3 - An example of real-time summarization on news items.

One limitation to a greater use of summarization by content providers is the contractual constraint they are under in terms of manipulating the content in third-party news or research reports.

Automatic real-time translation

While not strictly speaking a functionality of text mining, automatic real-time translation is complementary. For some content providers, automatic translation is an important issue. Factiva cites the need to ensure that language is not a barrier to information. They believe that there is an argument for "quick-and-dirty" translations that can flag important news. Factiva has evaluated automated translation machines and concluded that state-of-the-art technology is a huge aid as a productivity tool but that it is not yet good enough to replace humans.

Audio-visual

The technology for indexing and searching audio-visual information is there; suppliers include Convera. There are some reported applications on the sell side, in equity research, for example, where archived analyst reports and audio-visual footage are run alongside real-time news feeds. Content providers to the buy-side report that there is some interest in footage, but no firm requests yet. Despite the potential, the business case remains elusive for all but the biggest players: costs and bandwidth are among the problems.

Knowledge extraction

Most text mining technologies have a common objective: to extract knowledge from information. In principle, this is the problem of how we make sense of individual facts, formulate knowledge and make forecasts. In practice, it boils down 1) how to access and navigate information and 2) how to draw conclusions for decision making.

At a basic level, knowledge extraction coincides with smart searches on semi-structured data, performing queries such as: Retrieve all the items on mergers and acquisitions in the last 12 months valued at more than X euros. At a higher level, the objective is to compress information, reducing it to its essential. Summarization is one possible solution, but it implies that the computer is able to rank the importance of sentences in a corpus of documents. Given our current understanding of semantics, this task can only be partially performed. Visualization offers another solution. Using a combination of linguistics and statistical approaches, the content of documents is summarized in graphical displays, which visualize entities and their associations. This capability, available in a number of text mining solutions (see above), helps to grasp the content of documents without the need to read all the texts. Conceptual maps allow zooming in on texts in function of relevancy.

2. The take-up in finance

The business case

In the present environment of back-to-the-basics and cost cutting, new technologies are having to work hard to prove their worth. The business benefit of managing unstructured data depends on the time horizon:

- *In the short-term, the technology can deliver cost savings.* A return-on-investment (ROI) can be easily demonstrated in applications such as 1) categorization (this is of greatest interest to large "publishers"), 2) handling e-mails (though the ROI depends on scale; few buy-side firms have the volume to justify this) and 3) search (studies have put at 15-35% the amount of time knowledge workers such as analysts or fund managers spend searching for information).
- *In the medium-term, the technology enables knowledge management.* This requires management vision on how to organize decision making processes, a problem made difficult in finance by the lack of an agreed-upon "science". A key benefit of text mining will be the ability to further structure processes, creating a layer of shared knowledge in the organization.
- *In the long-term, the technology will allow the integrated handling of databases and knowledge management.* It is likely that, on a large scale, unstructured (textual) data will be replaced by structured or semi-structured data. Databases of numerical and textual data will be integrated either in a native way or through middleware. At this point, the cost of non-adoption will be high.

The ability to handle unstructured data is already a reality from a technology point of view; the diffusion of standards will make it compelling from a practical standpoint. Persons involved in standardization efforts estimate that within several years XML standards will be widely adopted in finance. The question will then be how to leverage the technology to the full.

Let's now look at some representative implementations in finance.

Generating taxonomies and categorization

Generating taxonomies and categorization are major application areas for text mining technology. Big publishers have been among the first to take it up; for vendors of automatic categorization tools, this represents a large though already crowded market.

For content aggregators and providers, the volume of textual data to categorize is formidable: Reuters publishes the equivalent of 3 bibles of news daily; Factiva, a Dow Jones-Reuters joint venture handles more than 400,000 news items daily. Over the last two years, pushed by the volume of content to categorize and the requirement to do so in real-time, major content providers to the industry have adopted automatic solutions for categorization. Table 2.1 below presents some examples.

Content provider	Implementation
Factiva	Uses <i>Inxight Categorizer</i> to automatically code and categorize >400,000 news items daily in real time. Two years ago, Factiva was manually coding and categorizing >50,000 news articles daily.
Thomson Financial	Thomson Financial Research (TFR), the data engine and production facility behind Thomson Financial Products, is using <i>ClearTags</i> from <i>ClearForest</i> to identify and automatically tag entities, facts and events in research reports, producing richly-tagged XML files. When completed in Q2 2002, will allow drill down on the content. Replaces a predominately manual process for indexing hundreds of thousands of research reports yearly.

Table 2.1 - Examples of how content providers are using text mining technology to automate categorization.

Sell-side and independent research firms (and the research teams at large buy-side firms) are also collectively large publishers, producing, according to some estimates, about 10,000 analyst reports daily. A group of large institutions has formed a consortium to address the problem of taxonomies at an industry-wide level (see Section 3, RIXML).

"Publishers" are not the only ones adopting automatic categorization. Many large financial institutions have automatic categorization projects running, employing

dynamic indexing tools from firms such as Autonomy or Verity. These applications are typically part of a larger project for building an intra- or extranet search infrastructure.

There is also a commercial offering of (customizable) off-the-shelf taxonomies. Content provider Factiva licenses its 300,000 company codes and Factiva Intelligent Indexing. The latter is a taxonomy of 1,500 subject codes for classifying industries, geographic regions and news topics pre-trained with 150,000 news stories. The firm's consulting teams will customize taxonomies and map Factiva codes to established internal taxonomies. Other content providers have or are preparing to offer similar services.

One factor that is slowing down a wider take-up of categorization inside financial firms: "you can't just throw the job to a machine and say 'categorize'". Categorization requires formulating an explicit view of the world and organizing the information flows. Another problem is the lack of standards. While many of the larger firms are going ahead without standards - or in some cases forming consortia to establish standards - lack of firmly established standards is considered an obstacle by smaller firms.

Research portals / business intelligence

As publishers of research, independent research firms and analysts at buy- and sell-side institutions have a problem of business intelligence and distribution. For the biggest firms, the problem is seen as one of IT infrastructure. Most of the large sell-side firms have turned to suppliers such as Autonomy or Verity to implement research portals; the latter counts among its clients many of the top 10 Wall Street firms.

Most of today's solutions have the functionality required for converting unstructured data into semi-structured data, be it legacy data or unstructured text from other sources. This functionality includes the ability to:

- discover taxonomies of corpora of texts by using clustering algorithms;
- organize existing corpora of unstructured data, imposing structure based on a taxonomy;

- search semi-structured databases with multidimensional criteria;
- search semi-structured data from third parties;
- perform business intelligence by crawling web sites;
- profile customers and site users;
- perform higher-level text mining functions, including the creation of semantic nets and conceptual maps;
- filter incoming and outgoing information.

For smaller firms, the problem is posed in terms of business tools. Rather than build proprietary research portals, these institutions typically rely on the search functionality incorporated in the offering of content providers or, in wider searches, on the search functionality available directly on the web. The former typically offers semantic search capability; the latter is essentially limited to ad hoc keyword searches.

Things are changing: the universe of assets is now global; asset management is moving towards real-time risk and performance measurements; the objectivity of sell-side research has been questioned. One result is a growing demand for one-stop shopping, with research, news feeds, real-time and historical prices and analytical tools all from one supplier. This is forcing content providers to enlarge their offering and move up the value chain. Moving up the value chain will include the ability to offer higher-level text mining capability and combined text and data mining functionality. Another result is a growing demand for integration services.

Integration and hosting services for research portals

Rather than turn directly to technology vendors (not one supplier of search engines that we talked to for this report was able to cite an implementation on the buy side), small and medium-sized institutions are turning to the business intelligence or content providers for building - and often hosting - proprietary research portals.

Major content providers such as Factiva, FactSet, Multex or Thomson Financial now offer to integrate third-party and proprietary data. Behind their integration offering is

experience in creating taxonomies and access to the infrastructure and text mining technology that the large financial firms typically implement in-house (see table 2.2 below for representative implementations).

Provider	Representative implementations of integrated research portals
FactSet	For <i>Asahi Life Asset Management</i> , a Tokyo-based money management firm, integrated the firm's internal research reports (Japanese pdf documents) with the third-party news and research sources currently available in FactSet's end user applications; allows Asahi's FactSet subscribers to share internal information through FactSet.
Multex	Builds (and hosts) research portals which commingle third-party and internal content. Representative applications include: <i>Baring Asset Management</i> , in an internal research portal project to give the firm's fund managers and analysts access to internal and broker research using the firm's industry classifications and subject sets; and <i>Scottish Widows Investment Partnership</i> , in an investment research facility project that will tie together internal and external broker research and integrate data feeds.

Table 2.2 - Representative examples of content aggregators providing bespoke solutions for buy-side research portals, commingling internal and external research and data.

Automatic handling of correspondence

Compliance applications

The automatic handling of incoming and outgoing electronic mail is widely identified as a major application area for text mining. Text mining methods are being applied to scan and understand incoming and outgoing e-mails and attachments for various purposes including archiving and conformity to corporate policy. Products such as SRA's Assentor are used by 85 securities firms, insurance companies, retail and institutional brokerages in the United States to identify and flag e-mail content that might raise legal or compliance issues and to provide centralized data storage for e-mail related documents (see table 2.3 below). This product uses natural language processing (NLP) technology (as opposed to key word or key phrase analysis) to understand texts.

End users	Implementation
85 US securities firms, insurance companies, and retail and institutional brokerages	Electronically scans, understands, indexes and archives electronic mail and attachments entering or leaving a firm, using <i>Assentor</i> from <i>SRA</i> . Objective: identify and flag e-mail content that might raise legal or compliance issues and provide centralized data storage for e-mail related documents.

Table 2.3 - Example of text mining in scanning incoming and outgoing correspondence for compliance purposes.

Applications such as scanning incoming and outgoing e-mails for compliance or regulatory purposes are more typical of the States where there is a greater need to quickly access such data to provide audit or legal support and respond to regulatory requirements.

CRM applications

Though many evaluate that text mining technology is not sufficiently mature to take humans out of the CRM (customer relationship management) loop all together, the technology is being used to categorize and route correspondence to appropriate persons for handling or to do analysis for marketing purposes. While this study did not identify implementations on the buy or sell sides of investment banking, we are beginning to see implementations in insurance and retail banking where the volume and nature of correspondence make it relatively easy to demonstrate a sufficiently rapid return on investment (see table 2.4 below) One problem with text mining solutions in CRM applications is the difficulty in dealing with the "dirty" texts (e.g., texts with symbols, typo mistakes, bad grammar) typical of correspondence; today's technology is not yet able to cope with this.

End user	Implementation
Liberty Mutual (insurance)	Categorize thousands of incoming e-mails daily and route to the appropriate specialist for handling, using <i>Megaputer's TextAnalyst</i> .
Standard Chartered Bank of Hong Kong	Categorize customer correspondence for satisfaction studies in a beta test application using <i>SAS's Text Miner</i> scheduled for release in the summer of 2002.

Table 2.4 - Implementations of text mining in CRM applications.

There is some expectation that the real potential for text mining in CRM is in the automatic handling of incoming calls, routing them to the right person, but this requires

additional progress in voice-to-text technology. The potential for text mining in CRM is not, however, universally shared; there is some feeling that automating customer service may have already gone too far.

Combining data and text mining

Major areas for combined text and data mining are fundamental analysis, the analysis of events on prices and volumes traded, and CRM. We are beginning to see some big text mining applications in fundamental analysis. In one such implementation using Insightful's InFact text mining solution, a major investor information service is ingesting large numbers of corporate documents such as SEC filings, press releases, analyst reports and annual reports in various formats, and asking questions on these. The application combines analyses on numerical data, using the statistical data analysis program S-PLUS.

A quantitative understanding of the impact of news on stock prices has been the subject of academic studies for years and a number of financial firms are reportedly doing such analyses. At the academic level, a well-publicized application is Eanalyst. Developed by Victor Lavrenko and fellow researchers at the University of Massachusetts, Eanalyst predicts prices from streams of news and streams of price data. The ability to combine mining operations on data and text will likely represent a huge potential for discovering pockets of predictability in asset prices. It will, for example, allow taking contrarian views on the implications of corporate news that might be interpreted in naive ways.

The advent of XML standards is expected to give a serious push to event analysis. MDDL (Market Data Definition Language), originally foreseen for transport, is to deliver a glossary of agreed-upon normalized standards for referring to "things," including events. The expectation is that by removing conflicting meanings, linking events to time series will be facilitated.

An application that combines text and data mining in product definition and pricing comes from the insurance sector. Liberty Mutual is using Megaputer's PolyAnalyst 4.5

to automatically mine text and perform statistical analysis on claims. The objective is to optimize the firm's offering in function of the thousands of claims filed daily.

Table 2.5 below summarizes the implementations discussed above.

End user	Implementation
Investor information services firm	Use <i>Insightful's</i> linguistics-based <i>InFact</i> text mining solution to ask questions on a large corpus of corporate documents, e.g., SEC filings, press releases, analyst and annual reports. The application combines analyses on numerical data using the statistical data analysis program <i>S-PLUS</i> .
Un. of Massachusetts	Software program <i>Eanalyst</i> predicts prices from streams of news and streams of price data.
Liberty Mutual (insurance)	Use <i>Megaputer's PolyAnalyst 4.5</i> text mining and link analysis capabilities to extract meaningful relationships and analyze statistical trends. Objective: optimize insurance offerings from the thousands of claims recorded daily by claim processing operators.

Table 2.5 - Implementations combining text and data mining.

3. From unstructured to semi-structured data

XML and the Semantic Web

XML: bringing structure to "bags of words"

The basic standard for handling unstructured documents and their semantics is the eXtensible Markup Language (XML), developed by the World Wide Web Consortium (W3C) in a project initiated in 1996. An evolution of the Standard Generalized Markup Language (SGML), XML is destined to replace the HyperText Markup Language (HTML), which has become a standard for the Web since its introduction in 1990.

XML is similar to HTML insofar it uses tags (i.e., words enclosed in brackets <...>) and attributes, though the use of tags and attributes is different in the two languages. XML is a language for describing documents in a standard way; it does not prescribe how documents should be formatted but how computers should read them. Its tree structure facilitates the addition of layers of metadata (i.e., information that identifies the content and context of a document) to the underlying data. Among XML components, XML Schemas help developers define the structure of their XML-based formats.

The Resource Description Framework (RDF) is an XML format in which the coding of semantics is formulated. It supports resource descriptions and metadata. RDF is the basis of the Semantic Web, tomorrow's web as envisioned by Tim Berners-Lee, inventor of today's web (1989), director of the W3C and researcher at the Laboratory for Computer Science at the Massachusetts Institute of Technology (MIT).

Computers require a full explicit representation of knowledge. The basic idea behind the Semantic Web is to attach to each web page a representation of the meaning (i.e., semantics) of its content which computers can "read". The idea is not limited to the web: computers might handle the semantics of any collection of documents to which a representation of content has been appended. RDF provides both a standard for this representation and a minimal domain-independent taxonomy (or hierarchical

classification); it prescribes basic rules in making reference to individual things, collections of things and so on.

In addition to specifying the basic semantics, XML and RDF define the conditions for searching. In fact, searching a collection of XML documents with their RDF descriptions requires XML-specific technologies. There are implications for IT: databases, data models and query languages must be XML-compliant (see Section 4).

XML dialects in the universe of finance

There are numerous initiatives established with the objective of defining XML standards in various areas of finance. Some are related to textual documents, others to combinations of textual and market data. We begin with a review of the former.

Information-oriented financial MLs

The emergence of domain-specific XML standards was widely cited as a factor in the take-up of text mining functionality. Buy- and sell-side firms alike mentioned putting projects on hold until it became clear which of the competing efforts would emerge as the de facto standard.

Among the XML standards for handling textual information of immediate interest to the buy and sell sides are XBRL (Extensible Business Reporting Language) and RIXML (Research Information Exchange Markup Language).

XBRL

XBRL (formerly XFRML) is an electronic language for financial reporting. The objective is to provide an XML-based framework for creating and automatically exchanging and analyzing (across all software formats) financial reporting information including but not limited to regulatory filings. In addition to the balance sheet and statements of income, equity and cash flows, an XBRL-based financial statement will include notes and the accountant's report. The EBRL consortium has more than 140 international members, including the large audit firms, exchanges and regulatory

authorities, credit rating agencies, buy- and sell-side institutions and the content providers Multex, Reuters and Thomson Financial.

XBRL specifications were released in 2000. A major complexity is the need for different XBRL taxonomies given the different accounting standards. The first completed taxonomy, containing thousands of items, was the US GAAP (Commercial & Industrial). Adoption of EBRL has begun in the States. In December 2001, Edgar Online, a provider of financial information derived from the US Securities & Exchange Commission data, announced XBRL Express, a public repository for XBRL-tagged company financial statements. XRBL Express is intended to serve as a single-source database for analysts, investors, news organizations and others. In the UK, EBRL reporting is foreseen for 2003.

Microsoft, a charter member of the EBRL consortium and among the early adopters of the XML-based framework for its 2001 fiscal year reporting, built what they call the XBRL Builder to automatically transform financial data into XBRL. Among other tasks, the software builds and maintains taxonomies.

In commenting on the future of XBRL, the consortium mentions a recent study by AIMR indicating that analysts prefer to obtain information from web sites even when other sources exist, citing ease of use. Other factors that will influence the take-up of EBRL are the accountants (the big ones are firmly behind it), suppliers of accounting packages (they will have to put EBRL-type tags in their accounting software) and the stock exchanges and regulators. In an effort to gain a competitive advantage, some of the world's exchanges are considering moving to "continuous" (e.g., quarterly) reporting and an EBRL-reporting format.

RIXML

RIXML is an open specification which provides a structure for classifying investment and financial research, regardless of the format or medium. It consists of a standard (completely externalized) attachment (or metadata) for indicating the nature of the content. The goal is to enable sell-side research publishers to tag research content with

enough information for buy-side firms to rank and sort reports and to provide filtering and personalization criteria across research publishers.

Originated by major buy- and sell-side firms, the RIXML consortium now counts among its associate members third-party content aggregators and service providers including Multex, ResearchSummary, TheMarkets.com and Thomson Financial. ResearchSummary recently announced that they would be adopting XML technologies to facilitate comparative research for investment professionals.

The top-level tag in a RIXML document will be "Product", a globally unique identifier. Further tagging will include source (the publisher and analyst who produced the note), content (the title of the document, type of file, etc.), context (e.g., that it is a research note for external buy-side analysts covering IBM), and legal requirements if any (e.g., copyright information, disclaimers).

Tagging reports at the source will require a retooling of processes within sell-side research departments. In the present business context, with constraints on spending, it is likely that this retooling will be slowed down; however, the expectation is that by mid 2003, consortium members will be structurally compliant with the RIXML standard. Asset managers wanting to adopt RIXML will have to review their procedures and systems.

Other XML standards in finance

In addition to the above XML-based standards, there are a number of initiatives in defining transaction-oriented financial MLs or data mining MLs. Some of these will be of interest as we begin to effectively combine data and text mining. Briefly these include:

- *MDDL (Market Data Definition Language), an XML-based language for posting statistical data on the web.* MDDL aims to map all market data into a common language and syntax to facilitate the interchange and processing of data sets. Version 1.0 released in November 2001 provides an interchange format and

common data dictionary on the fields needed to describe common equities, indices and mutual funds; version 2.0 scheduled for mid year 2002 release will cover corporate and US municipal bonds. Future plans include development of a common language for defining corporate events and an MDDL-specific query language. Among the founding members are data vendors Bloomberg, Dow Jones, Financial Times and Reuters, major sell-side firms, and Fidelity. Efforts are now being coordinated by the Financial Information Services Division (FISD) of the Software & Information Industry Association (SIIA). Estimates are that it will take two years to move from specifications to products.

- *FpML (Financial products Markup Language), an XLM-based protocol for e-commerce in complex financial products such as over-the-counter (OTC) derivatives.* FpML was integrated into the International Swaps and Derivatives Association (ISDA) in January 2002.
- *NewsML, an XML model for packaging news objects of any media type and adding metadata.* NewsML was developed by the International Press Telecommunications Council whose members include Dow Jones and Reuters. For the moment, NewsML is not foreseen as a tool for consumers but future development does not exclude a browser interface.
- *PMML (Predictive Model Markup Language), an initiative of the Data Mining Group.* The objective is to provide a way to define predictive models and share models between compliant vendors' applications.

There are a number of other efforts, some of them regional in character. One such one is Funds XML, an initiative of investment management firms in the German-speaking countries. The objective is to create an XML schema for defining funds-related content to enable automating the distribution and collection of organizational, structural and historical information on investment funds. Still in its set-up stage, the organization may open to non-German-speaking participants.

The above-mentioned domain-specific standards define their own taxonomies but are open to firm-specific extensions. Additional information and specification details can be found on the respective web sites.

One standard, no standard, 100,000 standards

With so many efforts underway to establish XML standards, the problem of competing standards is causing some confusion. While participants in the various initiatives evaluate that there is not much overlapping, integration is considered an area of concern. Efforts are now being made to coordinate the development of some XML languages (e.g., EBRL, NewsML, RIXML and MDDL). The central data repository ISO 15022, which provides a standard set of more than 10,000 data fields for financial information and about 100 messages for financial transactions, is seen as a useful integration point.

An issue being debated is whether the industry should work towards one big all-encompassing format (a solution seen to encourage market acceptance) or simply use the ISO central repository for understanding the underlying definition of terms, mapping terms from one standard to another. Putting a time frame to resolving the issue is difficult, but a new version of ISO 15022 which will for the first time include XML components is foreseen for this year.

Implications for "publishers," end users and technology vendors

There is no doubt that the XML standardization process is here to stay. We are now building an all-encompassing technology that will allow the seamless flow of transactions, data, news and analyst reports and the interaction of a vast number of market participants on a global basis. This has been made possible by an understanding of the limited intelligence available in today's computers and a huge and clever (XML) coding effort.

The benefit of XML standards will be the ability to take notice of more information. By calling the same thing with the same name, market participants will be able to perform better searches and do more with the data, performing, for example link and event analysis. In today's difficult business environment, with the number of analysts on the investment banking side dropping and the number of firms being researched likewise dropping, XML standards hold the promise of allowing market participants to use more sources of information efficiently.

Adopting XML standards will require some effort and cost:

- *Publishers* will need desktop tagging tools, typically found in content management systems from suppliers such as Documentum and Interwoven.
- *Content aggregators* will need to tag whatever documents are not tagged on arrival, before aggregation.
- *End users*, used to performing free text searches, will have to adapt to the structure of documents, performing searches on different fields.
- *Technology vendors* will see 1) the growing importance of content management functionality on the desktop and 2) complementary to free text search capability, the need to provide search capability on different fields as defined by XML-tagged data. Most suppliers of text-mining technology can already handle XML documents in addition to today's more common formats such as pdf and HTML.

A challenge to the traditional suppliers of text-mining tools will be to offer functionality beyond automatic categorization and indexing. Discovery tools such as conceptual maps, link analysis and visualization tools which provide a higher level of analysis will be of growing importance. It is likely that we will see a more wide-spread use of data and text mining in discovery in areas such as fundamental research and event analysis. Many suppliers are already putting the accent on advanced functionality. There is some expectation that XML will prove to be a "seminal" development for text mining tools.

While standards such as XML help in giving a more consistent structure to data, making it easier to manage and normalize, standards are only a starting point, not the entire solution. Suppliers of text-mining technology note that more than standards are required to get around problems such as specificity, idea differencing or reconciling different tagger schemes. For firms like Autonomy, this calls for technology that conceptualizes concepts as well as tags.

From XML standards to the Semantic Web

Finance requires a semantic understanding. In the language of computers, this means making explicit all relationships between terms. While making the whole universe of knowledge explicit is impossible with today's tools, we are able to encode small domains of knowledge. In banking and finance, a reduced semantics might be contained in the ISO 15022 central data repository.

The idea of the Semantic Web is have all data on the web defined and linked in such a way as to allow automated handling by computers across applications and platforms. This is made possible by encoding a small amount of semantic information in each document and in each web site. Although the amount of semantic information encoded in each entity is small, the semantics encoded across a large universe of entities makes a huge difference. The clever intuition behind the Semantic Web is that a small amount of semantic information widely distributed will allow to perform complex intelligent tasks.

While the technology for the Semantic Web is still immature, a semantic web for finance would not have to cover the whole universe. There is some expectation that we might see a semantic web for finance before the big all-encompassing project is realized.

A semantic web for finance would allow searching the entire web for information, querying the metadata and performing comparative research - a big advance with respect to what is possible to do with keyword searches. This functionality is already available (in part) on proprietary research portals and on streams of news and research

from companies such as Factiva, FactSet, Multex or Thomson Financial. A semantic web for finance would make this functionality available on a much larger set of information across the web.

4. Text mining and IT issues

Enabled and native XML databases

From the IT perspective, the processing of unstructured data is essentially a question of database management. The tasks are to:

- transform unstructured textual data into semi-structured data by extracting the basic structure from the texts, tagging, and adding metadata;
- organize and manipulate the semi-structured data;
- (eventually) perform sophisticated operations on subsets of semi-structured and structured data.

The handling of semi-structured databases is a new chapter in the evolution of database technology. Standard database technology requires that the structure of data be well defined. First came the hierarchical structures in the 1960s. These were followed in the 1970s by the relational databases where relations between entities are represented as tables. The 1980s saw the introduction of the object-oriented paradigm, based on defining abstract objects with various mutual or hierarchical relationships between them plus a flow of messages. Object-oriented databases offer an increased level of flexibility in representing a fixed set of relationships.

Now semi-structured database technology allows abandoning the constraint of a fixed data structure thanks to its ability to handle *partially structured, dynamically changing* data. The key was to make data self-describing and extensible. Well-structured descriptive metadata provide the essential information on the data, i.e., its content and context; data can now change dynamically, updating their own description. This evolution of database technology mirrors the evolution from data processing to knowledge processing - though the handling of knowledge has been reduced to the handling of structured data.

Whether or not a native XML database is required will depend on the type of data stored. A number of native XML databases have been engineered specifically to handle XML documents. Among these are eXcelon's eXtensible Information Server (XIS) and Software AG's Tamino. Most databases, however, are XML-enabled through middleware that can translate XML structures into, for instance, relational structures. This is the case with databases such as Microsoft's Access 2002 and SQL Server 2000 and Oracle 8i and 9i.

By and large, the decisive factor in choosing the optimal database and data warehousing structure will be the predominance of the type of data, structured or textual. In the first case, it is likely that the role of the XML structures will be limited to that of a convenient transport facility, suggesting the adoption of XML-enabled databases. In the latter case, XML will be at the heart of the application and native XML databases will likely be preferred.

The handling of semi-structured data is ultimately a question of boiling text management down to more classical database concepts. In time, just as data mining works on a relational database, text mining will work on tags and metadata. Information will be extracted from tags and put in XML(-enabled) databases. Text can be fields in a structured database. We will likely see the appearance of new tools for searching structured data that understand different fields in a data store and use the information, make sense out of it, and treat fields differently. Research and development on searching tools is under way on many different fronts. Companies such as the Norwegian FastSearch offer comprehensive solutions for searching structured and semi-structured data.

XML and the data model

Will the adoption of XML standards require an XML data model? The XML data model is inherently different from relational database (RDB) data models: it is a hierarchical tree structure of nodes with strict rules of inheritance. However, XML is a flexible high-level structure that can be used to integrate different data structures and data models. Though XML has its own data models, any object-specific data model can be

integrated. XML is effectively used as a glue to integrate different systems - including legacy and web-based systems - into a single IT structure. A database of data from a clickstream analysis package can, for example, be integrated with other databases of customer related information.

XML query languages: one per standard?

Finding the right structure to represent the many relationships with a bearing on financial semantics is only half of the problem of handling semi-structured data. Searching the database with an appropriate query language is the other half. From a knowledge management point of view, defining the appropriate query language is key: the ability of the system to retrieve the required information depends on the expressive power of the query language. The World Wide Web Consortium (W3C) has stipulated the specifications for a general XML query language, called XML Query. Independent of any specific implementation of XML databases, XML Query plays a role similar to SQL.

There is some expectation by persons involved in specifying the various XML standards in finance that at least some of these standards will require their own query language. Queries on market data, for example, are not the same as queries on corporate reporting; an MDDL-specific query language is envisioned. In financial applications, a key requirement is the ability to implement jointly full text and numerical searches in queries such as: "Retrieve all mentions of share buy backs in sector X whose total value is within the bounds of \$Y-Z million." It is likely that the integration of XML query languages with existing databases will be handled by middleware.

5. The market

Factors driving the take-up of text mining technology

Throughout the report, mention has been made of factors driving the take-up of text mining technology. In their generality, they can be summarized as follows:

- The amount of unstructured data with business value produced and stored in corporate computer systems is growing rapidly, making their handling a compelling task.
- The amount of information on the economy, finance, and individual firms available in electronic format on the web (including via content aggregators) is growing rapidly; mining this data *will* give an information advantage.
- Text mining will enable knowledge management, allowing to better structure business processes such as asset management.
- The arrival of (XML) standards will facilitate text mining operations, including the ability to combine text and data mining in applications such as identifying business opportunities in investment banking, fundamental analysis and event analysis, and CRM.

Factors holding back wider implementation

Before mentioning factors holding back a wider implementation of text mining technology in finance, several considerations should be made:

- Most large sell-side firms have already invested heavily in the technology, creating central data repositories and research and e-business portals;
- On the buy-side, text mining functionality is already on the desktop, in the services offered by content providers. The question on the buy side is whether there is the need to manage textual data at the firm-wide level, performing higher-level functions or exploring other sources of information such as corporate web sites.

These considerations aside, users and vendors alike evaluate that take-up of text mining technology has been slow. Among the reasons cited are the following:

- *There is a reappraisal going on regarding technology in general.* Many firms have been "burnt" on the internet hype and are now taking a more realistic look at what technology can do. Nevertheless, while some point to shortcomings of today's text mining technology others believe that the technology has improved tremendously over the past few years.
- *In today's difficult business environment, the accent is on cost containment.* Demonstrating return on investment (ROI) is considered key. Vendors are putting the accent on tools for measuring ROI. In areas such as categorization or the automatic handling of incoming e-mails, this is relatively easy (though these applications are sensitive to scale); in more "amorphous" applications such as knowledge management, measuring ROI is less easy, though the benefits might be compelling. There is not much expectation, now that the technology hype is over, that spending on technology will pick up significantly in 2002.
- *Knowledge management presents additional problems.* Knowing where to start can be disconcerting; a top-down approach and hard thinking about the business is required. Knowledge management also comes up against a human barrier: the unwillingness to share information, considered a differentiator. Nevertheless, there are some on-going pilot projects.
- *Lack of standards.* The flurry of activity in establishing XML standards in finance has been somewhat disorienting for on-lookers, but considerable progress is being made in areas such as business reporting with EBRL and analyst reports with RIXML. In addition, attempts are now being made to coordinate standardization efforts, eventually adopting the ISO 15022 central data repository as an integration point. There is some expectation that once standards have been specified in detail, we might see a snowball effect in terms of take-up of text mining technology.

- *The basic technology is now far ahead of its deployment.* Vendors will be putting more effort into raising the visibility of the technology.
- *Perhaps paradoxically, the capability of today's search technology is also a limiting factor.* To grasp higher-level, nontrivial things, search technology would require the ability to handle semantics (as opposed to relying on statistics to count words or phrases). Doing this in English is still not possible; other languages are even less researched. XML standards are bringing an at least partial solution to the problem.

Structuring of the supply side

While the market has been slower in developing than many had hoped, technologies for handling unstructured and semi-structured data are becoming mainstream. This is drawing new players into what some consider an already crowded market given the revenues generated. Technology and content vendors alike are extending their offerings in an attempt to cover an ever wider range of needs. Alliances and consolidation will continue.

Vendors fall into the following categories:

- *Suppliers of online web search engines.* Suppliers of consumer-oriented online web search engines such as Google are now targeting corporate users.
- *Technology infrastructure suppliers.* The core of the market is represented by firms such as Autonomy and Verity. These provide the technology infrastructure to power portals and corporate IT systems.
- *Niche text (and data) mining suppliers.* These include ClearForest, Megaputer, Semio and Temis. Some are now expanding into the broader context of semi-structured data.
- *Suppliers of statistical analytical packages.* Recent months have seen the entrance of SPSS (with the February 2002 acquisition of the linguistics-based text mining specialist LexiQuest) and Insightful (ex-MathSoft, with the April 2002 introduction

of its in-house engineered linguistics-based text mining tool InFact). SAS's entrance is scheduled for mid 2002 with the introduction of their Text Miner, an add-on to their Enterprise Miner data mining suite.

- *Suppliers of content management software.* Often working in partnership with the search engine vendors, enterprise content management vendors are extending the functionality of their products. Interwoven's March 2002 announced acquisition of XYZFind will enhance its retrieval capability and give it access to native XML database technology.
- *Database management suppliers.* Oracle announced its text mining functionality as part of Oracle9i in May 2001. Microsoft, which already has some (hidden) data mining capability on SQL2, is a potential big player in the future: it is devoting resources to text mining technology and has plans to introduce the ability to store and manage both structured and unstructured data in a future release of its SQL Server database.
- *Content providers offering taxonomy creation, the integration of third-party and proprietary data and search and analytical functionality.* These include Factiva, FactSet, Multex and Thomson Financial. They will be adding value to their offering, enhancing data and text mining capability, typically supplied by the specialized technology vendors.

Users will select vendors in function of where they place the accent (e.g., document management, search functionality, integrated content and analytical capability) and the available budget.

The challenge to vendors

Challenges confronting vendors include:

- *The rapid pace of change.* Perhaps the biggest challenge, in the space of a few years vendors have seen the accent shift from the need to handle unstructured data to developments in XML standards which are transforming unstructured data into semi-structured data.

- *The need to add value in terms of functionality.* Related to the above, vendors will have to distinguish their offering with new functionality - without exceeding the market's ability to effectively employ the functionality.
- *Continued investment in R&D.* Generating steady profits remains a challenge here as in many technology sectors presently, but the technology for handling unstructured and semi-structured data will continue to require investment.

6. Technology backgrounder

Semantics and data structures

Structured data are data organized in a set of elementary components linked by relationships. A computer can easily access each field in a file of structured data. A set of invoices, for instance, can be structured in different fields: date, company, amount, product, and so on. Using database software, a computer can drill down to items or sets of items that respond to logical or numerical criteria to the level of the lowest field.

Initially, structures for storing and retrieving data were hierarchical: data were mirrored in a hierarchical set of indexes. Then came relational data structures, where data are indexed by sets of relations represented as tables. More recently, object-oriented technology introduced the notion of logical objects that are linked by relationships and/or pass on messages.

Unstructured data, on the other hand, are seen by a computer as a sequence of characters without internal structure. The computer can access the entire block, not the individual parts. Letters or memos, for instance, are retrieved as single objects. A first level of structuring unstructured data is achieved by appending to unstructured data information describing the data. Data thereby become self-describing and can be dynamically changed and extended. *Self-description* and *extensibility* are the keys to semi-structured data. Though clearly useful, this structuring tells us nothing about the content of the document. One would like to be able to access and manipulate documents based on their content.

The key is semantics, the science that studies meaning and how words and phrases convey meaning. In its technological aspects, semantics assumes that there is a world composed of objects and that these objects have relationships amongst themselves; it limits its exploration to how a given set of (possibly complex) symbols represent a set of objects and their relationships. This allows tailoring semantics to the objects of

interest. If, for example, our area of interest is financial products or corporations, we will limit our considerations to how these can be represented.

To structure documents in function of their meaning requires:

- a complete specification of the objects, of their meaning, and of their relationships;
- a complete specification of the set of symbols (i.e., language) used in the documents;
- a specification of the relationships between the language and the objects.

This is clearly a brute-force approach based on the ability to completely specify sets of objects and their relationships. In limited domains such as finance this might be relatively easy; in vast domains it might be impossible.

In practice, we have reduced semantics to database technology, but at the cost of building a database of *explicit* relationships. From a mathematical and technological standpoint, this is a formidable problem. Our everyday language is not explicit in this mathematical sense; it assumes a vast number of implicit relationships which would have to be made explicit.

Explicitly coding semantics requires special methods. Over the last forty years the artificial intelligence (AI) community has developed a number of ways to code knowledge (i.e., to represent the relationships in the set of objects). These representations of knowledge are typically extensive, not intensive - that is to say, they identify concepts with sets of objects. Example: the concept "red" is identified with the set of red things. In this way every concept is identified with relationships.

The AI tools for coding knowledge include:

- *Semantic nets*, which represent knowledge as a set of relationships between objects in a way similar to relational databases;

- *Frames*, which, while similar to semantic nets, are organized around nodes that have many slots that represent properties;
- *Sets of rules*, whose rules are logical expressions that define properties.

While in principle these tools can represent any sort of knowledge, the coding of all the knowledge implicit in, for example, business processes, has proved a daunting task. As a consequence, only small domains of knowledge have been coded and represented in systems such as expert systems. One such system developed by the London-based fund management firm Pareto Partners for its Global Bond Allocation Strategy system includes some 2,000 rules. Presently, the solution to structuring unstructured data is being solved by confining the domain of knowledge (or semantics) to the minimum. Still, this minimum implies a vast effort of coding.

The origins of text mining

There is no unified, all-encompassing technology for text transformation and text mining. Most commercial software consists of a suite of algorithms, each algorithm solving one aspect of the problem. Technologies for the basic functions of generating taxonomies, categorization, indexation and search are widely considered to be mature; technologies related to the higher-level functions such as profiling, summarization and discovery are less mature.

The basis on which all text-mining software works is the discovery that *many cognitive functions can be boiled down to symbol manipulation*. The 17th century German philosopher Leibnitz was perhaps the first to clearly state the possibility of reducing cognition to symbol manipulation. In a statement that was to become famous, Leibnitz suggested that future philosophical disputes would be settled by computing.

At the end of the 19th century it was believed that all logic and mathematics could be reduced to syntactical symbol manipulation. Kurt Goedel's discovery (1927) that there are fundamental limits to the possibility of reducing logic to the manipulation of

symbols came as a surprise. The theorem of Goedel opened a new era in mathematical thinking.

The development of computers in the 1950s marked a change in direction of research on the automation of cognitive processes. Attention shifted from pure speculation on the fundamental nature of cognition to the practical endeavor of building machines that exhibit intelligence. The field of AI was born. The conceptual stage was set by the English mathematician Alan Turing, one of the fathers of modern computing and famous for breaking the German encryption machinery during World War II. Turing's definition of an intelligent machine - one able to converse and answer questions like a human being - underlines the basic difficulty we still have in defining machine intelligence: still now we can ultimately benchmark intelligent behavior only against human behavior.

Progress in AI was slow. Scientists rapidly discovered that very simple routine tasks involved searching an intractably infinite space, the so-called frame-problem. The feasibility of "doable" tasks was hampered by a lack of computing power. The neural network approach, which seemed so promising, was (temporarily) abandoned in favor of expert systems following the criticism of MIT's Marvin Minsky.

The 1980s witnessed a number of breakthroughs. One of these was the understanding that a lot of intelligent behavior, in particular learning, could be mimicked as optimization in a probabilistic framework. Neural networks are an example of this approach which is, however, much more general. For example, speech recognition (used in voice-to-text technology) is based on optimizing a stochastic phenomenon. A large number of small "intelligent" applications resulted in areas such as database technology and control systems. This bottom-up approach to machine intelligence as a set of techniques for solving specific problems has changed the perception of AI in the scientific and business communities.

Text-mining is not a technology that allows a machine to read and understand texts but a set of functions that can be performed - more or less successfully - on texts. The

notion that a machine can read and understand a general text is presently out of reach - it is not even well-defined - but a computer can perform a number of manipulations on texts that would otherwise require an intelligent human being. These functions might simplify and speed-up routine business tasks such as document retrieval but might also help in finding relationships hidden in large bodies of text or between textual and other data.

Finding similarity in meaning with statistical techniques

In the last two decades, we have learned that important similarities in the *meaning* of two texts can be revealed by looking for statistical similarities in the *sequences of words* in the text. Far from being obvious, this was in a sense a genuine scientific discovery: similarities of meaning in texts are reflected in similarities in the statistics of the relative sequences of words. This empirical discovery underpins the use of statistics in categorizing and retrieving texts.

The result of statistical tests is a probabilistic representation of a text and a measure of the similarity / dissimilarity between given texts. Though similarity measures can also be used to implement searches directly, statistical techniques based on similarity are more commonly used to automatically generate taxonomies, categorize, and create automatic indexing systems.

Generating taxonomies with unsupervised clustering techniques

Automatic categorization and indexing are based on the notion of clustering. Similarity enables clustering. Given a similarity relationship, and thus a distance function, it is possible to cluster together objects that are close to each other. In text mining, this means clustering together texts on the same topic or sharing the same meaning.

Given a similarity / dissimilarity matrix (i.e., a matrix that quantifies the similarity / dissimilarity between texts), various clustering algorithms can be used. Trained on a given corpus of texts, the clustering algorithm establishes the relevant categories. Categories might then become templates for categorizing any new incoming texts.

Categorizing with supervised learning techniques

Supervised learning techniques learn from examples. In text mining, they are used to learn categorization from given examples. Once the categorization is learned, supervised learning techniques generalize it over a set of unclassified texts for proper classification.

Supervised learning techniques include neural networks and statistical frameworks such as regression techniques and Bayesian frameworks. Important theoretical advances (thanks largely to the work of Vapnik and Chervonenkis) have been made over the last twenty years, yielding modern learning theory. Vector support machines - a type of neural network derived by the application of modern learning theory - are proving to be useful techniques in supervised indexing and categorization applications.

Improving on results with linguistics

Statistical methods applied to text mining greatly benefit from any knowledge on the general structure of texts. Many text transformations are obvious. For instance, phrase delimiters such as commas and dots fix the basic structure of a text. Stemming (or lemmatization as it is often called) allows to use only the unique root of each word and not its many variations. Example: category, categorizing, categorization have all the same stem (or lemma). Computational linguistics determines the syntactic form of a phrase, distinguishing nouns, verbs and syntactical relationships.

Current linguistic approaches are based on thesauri and dictionaries to capture synonyms and other general relationships between key words. Texts are analyzed with linguistic tools to extract lemmas, define syntactical relationships, resolve ambiguities, eliminate synonyms, and so on. A number of key relationships between terms are then extracted using a variety of approaches based on linguistics and / or statistics. What can be done today with the algorithms, the knowledge basis and the computing power available for industrial and commercial use is to extract a set of key relationships between terms. This is still far from a real understanding of text, but it is helpful in searching a knowledge database or navigating large sets of documents.

7. Supplier directory

N.B.: The companies listed below are representative of firms with a technology or service offering for handling unstructured and semi-structured data. This does not constitute an exhaustive list; we regret omissions.

The information provided is based on conversations with the industry and vendor documentation, but its accuracy cannot be guaranteed. Please refer directly to the company concerned for any questions or additional information.

AUTONOMY

HQ: Cambridge, England
Tel: 44 (0) 1223/448 000
www.autonomy.com

PROFILE

Software company providing infrastructure technology that automatically processes and organizes semi-structured and unstructured information.

Public company with fiscal 2001 revenue of \$52.6 million.

OFFERING

Core (language-independent) technology is THE DYNAMIC REASONING ENGINE (DRE) based on pattern-matching algorithms and probabilistic Bayesian inference; provides the platform for automatic categorization and (XML) tagging, hyperlinking, retrieval of unstructured information and user profiling. Products include PORTAL-IN-A-BOX, an information portal which provides a single entry-point to a firm's information, resources and expertise; AUTONOMY ANSWER, an add-on product to automate responses to customer queries; and ACTIVEKNOWLEDGE, a package that provides real-time recommended links to internal or external information related to documents under preparation.

CLIENTS

Include Baring Asset Management in an automated e-mail response system; Danske Bank in a customized implementation of Portal-in-a-Box for the real-time automatic aggregation and delivery of internal and external information sources (including analyst and research reports, stock exchange and regulatory information) with automated tagging; Dresdner Kleinwort Wasserstein in a trading risk management application on historical (unstructured) data and real-time feeds; and Zuercher Kantonalbank in an internet search project that delivers automatically hypertext-linked information to employees.

CLEARFOREST

HQ: New York, NY, USA
Tel: 1 212/432 15 15
www.clearforest.com

PROFILE

Software company providing a suite of content enhancement and analytical tools for knowledge extraction and information delivery.

Privately-held company whose cofounder and chief scientist is Dr. Ronen Feldman, a pioneer in the field of text mining.

OFFERING

The CLEARFOREST suite of categorization and knowledge extraction tools includes CLEARRESEARCH, an enterprise research application that presents a single-screen view of complex inter-relationships; cleartags, an auto-tagging platform using semantic / linguistic information, statistical categorization and structural analysis technologies; CLEAREVENTS, a web-based events monitoring and real-time notification application; CLEARsight, a browser-based portal solution that provides visualization (interactive maps) of complex relationships and includes drill-down functionality; CLEARCHARTS, an interactive graphic application that depicts trends in real time, matching quantitative and qualitative data; CLEARSTUDIO, a wizard-based environment that enables users to adapt the technology to specific vertical markets using ClearForest's rulebook technology and integrating the natural language processing software engine LinguistX Platform from Inxight; and CLEARLAB, a technical development environment.

CLIENTS

Include Thomson Financial Research, using ClearTag 4.0 to identify and automatically tag entities, facts and events in 100s of 1000s of research reports yearly, producing richly-tagged XML files.

DOCUMENTUM

HQ: Pleasanton, CA, USA
Tel: 1 925/600 68 00
www.documentum.com

PROFILE

Enterprise content management provider, offering XML-enabled tools to automate the production, exchange and personalization of content.

Public company with fiscal 2001 revenue of \$185.7 million.

OFFERING

Includes DOCUMENTUM 4i WEBCONTENT MANAGEMENT EDITION which automates workflows and lifecycles of the web content management process and DOCUMENTUM 4i EBUSINESS PLATFORM which automates the translation, localization and personalization of content. PORTAL INTEGRATION PACK, which includes a set of embeddable application components (portlets) that deliver content management capabilities to portals, facilitates the integration of content and content management capability with portal applications. Recently announced acquisitions will enhance content management and delivery and extend capability to video, audio and image.

CLIENTS

Include UBS Warburg, using Documentum 4i eBusiness Platform (together with Verity classification and search software) to manage unstructured data for some 27,000 users worldwide; Morgan Stanley, using Documentum as a content repository and web content management solution with a template-based authoring environment; and Barclays Global Investors, using Documentum 4i WebContent Management Edition to automate workflows and lifecycles of the web content management process.

EXCELON

HQ: Burlington, MA, USA
Tel: 1 781/674 50 00
www.exceloncorp.com

PROFILE

Provider of database management software for distributed applications built using XML and Java.

Public company with fiscal 2001 revenue of \$49.2 million.

OFFERING

Provides EXTENSIBLE INFORMATION SERVER (XIS), a native node-level XML database management system with an engine for storing and transforming dynamic collections of XML across the firm; STYLUS STUDIO, a development environment for XML-to-XML mappings; and BUSINESS PROCESS MANAGER, an XML-based business document rules engine.

FACTIVA

HQ: New York, NY, USA
Tel: 1 609/627 20 00
www.factiva.com

PROFILE

A Dow Jones-Reuters company providing global news (from some 8,000 sources worldwide), business information and corporate indexing solutions through its web sites and content integration solutions.

OFFERING

FACTIVA INTELLIGENT INDEXING, a coding system to enhance keyword searches; major indexing categories include companies (>300,000), geographies (>370), industries (>740) and news topics (>430); uses Inxight Categorizer for indexing and Inxight Thing Finder for extraction; can be licensed for use in proprietary portal systems. FACTIVA CONSULTING assists in creating customized taxonomies and in mapping the Factiva-Inxight system to established taxonomies.

Other services include FACTIVA PUBLISHER, which allows the delivery of global news and business information to corporate internets and portal desktops; FACTIVA SELECT, a customizable news feed that enables pulling filtered XML content from Factiva content for proprietary XML-based knowledge management systems; and FACTIVA.COM, a web-based service providing personalized tools for researching (Verity-powered searches) and monitoring news and business information, drawing from Factiva content covering 118 countries and including some 8,000 publications, 8,500 key web sites and thousands of company profiles.

FACTSET RESEARCH SYSTEMS

HQ: Greenwich, CT, USA
Tel: 1 203/863 15 00
www.factset.com

PROFILE

Provider of online content and analytics for the investment management and investment banking communities.

Public company with fiscal 2001 revenue of \$176.7 million.

OFFERING

Online service which combines more than 100 databases of global financial and economic information (and increasingly real-time news) as well as tools to download, combine and manipulate the data for investment analysis; (Verity-based) search capability (by keyword, data, company ticker, industry sector or subject code) and some data + text analytical capability available. DATA WAREHOUSING suite's recently announced DATA CENTRAL lets clients create, update, manage and query current and historical proprietary and commercially available data on FactSet's online system.

Also offers an integration service which commingles proprietary and third-party news, research and data to create proprietary knowledge management systems.

CLIENTS

The online content service counts over 830 corporate subscribers. Integration service clients include the Tokyo-based money management firm Asahi Life Asset Management, in an application integrating the firm's internal research reports with FactSet third-party news and research sources, allowing Asahi's FactSet subscribers to share internal information through FactSet.

FAST SEARCH

(Fast Search & Transfer ASA)

HQ: Oslo, Norway
Tel: 47/23 01 12 00
www.fastsearch.com

PROFILE

Search technology company providing search, real-time content matching and filter technologies for internet and enterprise applications.

Public company founded by researchers from the Norwegian Institute of Technology. Fiscal 2001 revenue of \$36.1 million.

OFFERING

Includes FAST DATA SEARCH, a unified customizable search engine interface that aggregates data from databases, web and file servers, XML sources and other document types for real-time search and information retrieval; and FAST REAL-TIME FILTER, to dynamically monitor information and deliver personalized alert-based information.

Also: ALLTHEWEB.COM, an internet search engine using an index (updated every 9-12 days) with a match to web content including multimedia files; includes automatic search tips based on artificial intelligence algorithms.

CLIENTS

Include Reuters News Distribution Service, using FAST Data Search to perform real-time content filtering and matching on thousands of news stories (whole as opposed to headlines) daily for customized distribution.

IBM

Lotus Knowledge Management Unit (LKMU)

HQ: Armonk, NY, USA
HQ LKMU: Cambridge, MA
www.ibm.com

PROFILE

IT hardware, software (including the Lotus Knowledge Management Unit) and service company.

Public company with fiscal 2001 revenue of \$85.9 billion.

OFFERING

LOTUS DISCOVERY SYSTEM includes two components designed to operate as stand-alone products or an integrated solution: LOTUS K-STATION, a portal for organizing, managing and accessing information and LOTUS DISCOVERY SERVER, a search engine that builds taxonomies and allows key phrase search and browsing of hierarchical topic trees.

IBM also offers INTELLIGENT MINER FOR TEXT, part of the INTELLIGENT MINER family which includes DB2 INTELLIGENT MINER FOR DATA. Intelligent Miner for Text includes text analysis tools for feature extraction, clustering, summarization and categorization. Complementary products include the TEXT SEARCH ENGINE, NETQUESTION SOLUTION and the WEB CRAWLER PACKAGE.

INSIGHTFUL

Formerly MathSoft Data
Analysis Products Division

HQ: Seattle, Washington, USA
Tel: 1 206/802 12 40
www.insightful.com

PROFILE

Software company providing statistical data analysis, data mining and predictive analysis software, including the statistical data analysis package S-PLUS, the optimization tool NUOPT, the web-based analytic delivery tool STATSERVER and the recently announced (Feb. 2002) desktop predictive analysis package INSIGHTFUL MINER DESKTOP EDITION.

Public company with fiscal 2001 revenue of \$17.4 million.

OFFERING

Entered the text mining arena with INFAC (April 2002), a stand-alone text mining solution for information retrieval based on semantic and syntactic methods. InFact handles text, image and numerical data, producing direct (and when not possible analyzed / synthesized) answers (as opposed to lists of documents) in the form of text, charts, images, maps, tables or other; includes automated, incremental indexing techniques. Text mining results can be delivered into S-PLUS frames for statistical representation.

CLIENTS

Include a large Wall Street investor information service, using InFact to generate analyzed /synthesized answers to questions, drawing on and associating information found in hundreds of documents (e.g., SEC filings, press releases, analyst reports and annual reports).

INTERWOVEN

HQ: Sunnyvale, CA, USA
Tel: 1 408/774 20 01
www.interwoven.com

PROFILE

Software company providing enterprise content management (ECM) software under the Interwoven 5 ECM platform.

Public company with fiscal 2001 revenue of \$202.7 million.

OFFERING

INTERWOVEN 5 ECM PLATFORM includes TEAMSITE for content management; METATAGGER, whose business rules engine automatically generates taxonomies based on existing collections of content and categorizes content according to one or multiple taxonomies; and TEAMPORTAL for integrating content management into corporate portals.

Recently expanded product offering reinforces Interwoven's position in the search arena. March 2002 announcement extended its CONTENT DISCOVERY FRAMEWORK (which includes an XML content repository) to enhance information retrieval functionality thanks to the acquisition of XYZFind Corporation whose XZYFIND technology includes a native XML database with schema-independent architecture.

INXIGHT

HQ: Santa Clara, CA, USA
Tel: 1 408/969 72 00
www.inxight.com

PROFILE

Software company providing language-intelligent tools for analyzing, organizing and navigating information in (currently) 16 languages on the internet and enterprise networks.

A privately-held company spun off the Xerox Palo Alto Research Center, with patented technology in visualization and linguistics.

OFFERING

Core building block is LINGUISTX PLATFORM, a natural language processing software engine used to analyze massive text repositories, performing automatic language identification, tokenization, stemming, part-of-speech tagging and noun / phrase extraction. INXIGHT CATEGORIZER automatically classifies unstructured data and text into pre-defined categories. Two frameworks: METATEXT SERVER and VIZSERVER. METATEXT SERVER automatically extracts and indexes metatext. Services within the framework include: INXIGHT SUMMARIZER, a software engine that creates abstracts of online documents and customizes content into a summary size relevant to a user's needs; INXIGHT CONCEPT LINKER, a dynamic application that analyzes search results and organizes information based on related concepts; INXIGHT THING FINDER, a content indexing product that identifies and extracts entities such as people, places and things within unstructured text; and SIMILARITY FINDER, an application that finds documents similar in content. VIZSERVER is a server application for the deployment of the visualization technologies STAR TREE, for navigating and visualizing hierarchical information collections and TABLE LENS, which provides graphical displays of tabular data (over 100 columns and 65,000 rows of data) for exploring data sets.

CLIENTS

Include Deutsche Bank using Star Tree to visualize multidimensional financial data; YellowBrix/CNNfn using Inxight Summarizer to provide real-time summarization on 39,000 words/second; and Factiva using Inxight Categorizer to automatically code >400,000 news items daily and deliver customized news to subscribers.

LEXIQUEST

See SPSS

MEGAPUTER INTELLIGENCE

HQ: Bloomington, Indiana, US
Tel: 1 812/330 01 10
www.megaputer.com

PROFILE

Software company providing tools for data mining (POLYANALYST), text mining and web data analysis (WEBANALYST).

A privately-held company founded by researchers in the artificial intelligence group at Moscow State University.

OFFERING

Text mining products include: TEXTANALYST, a tool for summarizing, navigating and clustering documents based on linguistic and neural network technologies; and MEGASEARCH, a fuzzy logic, natural language query tool for document retrieval. Complementary products include the data mining application POLYANALYST whose V. 4.5 (released April 2002) includes link analysis functionality and tools for the (supervised and unsupervised) semantic analysis of natural language text in relational databases, allowing the analysis of structured data and text in a single system.

CLIENTS

Include Liberty Mutual which uses TextAnalyst to categorize and route incoming e-mails; and PolyAnalyst 4.5 text mining and link analysis capabilities to extract relationships and analyze statistical trends to optimize insurance offerings.

MICROSOFT

HQ: Seattle, Washington, USA
www.microsoft.com

PROFILE

Software company providing the flagship WINDOWS family of products, OFFICE, MICROSOFT SQL SERVER 2000 and EXCHANGE 2000 SERVER, and the .NET ENTERPRISE SERVER family for building internet-based commercial applications.

Public company with fiscal 2001 revenue of \$25.3 billion.

OFFERING

SHAREPOINT PORTAL SERVER, for creating portals out of the box; offers content indexing, search and document management. At the foundation of SharePoint Portal Server is WEB STORAGE SYSTEM, which provides a mechanism to store and retrieve semi-structured data. A future release of its SQL SERVER database will store and manage both structured and unstructured data.

MULTEX

HQ: New York, NY, USA
Tel: 1 212/607 25 00
www.multexusa.com

PROFILE

Provider of web-based investment information (including research reports, real-time estimates data, fundamental information) and technology solutions (including information hosting) for buy- and sell-side institutions, retail brokerages and corporations.

Public company with fiscal 2001 revenue of \$96.9 million.

OFFERING

MULTEXNET is an online service providing real-time investment research, company fundamentals and events information; its "Company Page" produces a holistic view of publicly traded firms with one click (i.e., a single corporate identifier), using Multex's INTELLIGENT SYMBOLOGY technology to search. INTELLIGENT CATEGORIZATION enhances searches on non equities data. MULTEXEXPRESS is an ASP service that allows buy- and sell-side firms to customize and brand Multex content for online delivery to employees and/or clients, eventually integrating internal research.

CLIENTS

Provides the technological infrastructure and web hosting services for TheMarkets.com, common web site of nine of Wall Street's biggest investment banks, offering commingled equity research and other information. Develops and supports (and in many cases hosts) research portals; clients include buy-side firms Baring Asset Management, Scottish Widows Investment Partners and Tilney Investment Management and, on the sell-side, ING Barings. Services typically performed include the tasks of classification, indexation, tagging, navigation and search. Clients for MultexNET include DuPont Capital Management and Zurich Scudder Investments.

ORACLE

HQ: Redwood Shores, CA, US
www.oracle.com

PROFILE

Software company providing database, server, business applications and decision support tools, including the flagship relational database server ORACLE.

Public company with fiscal 2001 revenue of \$10.8 billion.

OFFERING

A text engine running inside the Oracle9i relational database server, ORACLE TEXT provides specialized text indexes for full text retrieval using keyword searching, context queries, Boolean operations, linguistic features and pattern matching. Complementary products include: ORACLE ULTRA SEARCH, a web-based application built on top of Oracle Text that crawls multiple data repositories, gathering and storing index information in an Oracle Text index; native XML datatype (XML TYPE) to facilitate the storage of XML content directly into the database and XML PATH SEARCHING, a mechanism for specifying complex search queries in XML documents.

SAS

HQ: Cary, North Carolina, US
Tel: 1 919/677 80 00
www.sas.com

PROFILE

Software company providing information delivery and data analysis and data warehousing software, including ENTERPRISE MINER, a suite of integrated data mining tools.

World's largest privately-held software company.

OFFERING

Mid 2002 will be introducing TEXT MINER as an add-on to the ENTERPRISE MINER suite of data mining tools. January 2002, announced that Text Miner would be enhanced with Inxight's natural language text analysis solution LinguistX Platform and Inxight's Thing Finder for identifying and extracting key content from documents.

CLIENTS

Include Standard Chartered Bank of Hong Kong, in a (beta test) CRM application to categorize customer correspondence.

SEMIO

HQ: San Mateo, CA., USA
Tel: 1 650/638 33 30
www.semio.com

PROFILE

Software company providing automated content categorization and indexing software based on proprietary linguistics techniques.

A privately-held company founded by Claude Vogel, professor of semiology at the University of Leonardo da Vinci in Paris.

OFFERING

Includes SEMIOTAGGER, which uses a linguistic analysis process to extract key concepts and automatically build taxonomies or custom content categories; SEMIOTAXONOMY, a web-based taxonomy viewer; SEMIOMAP, a 3D graphical viewer for discovering relationships between documents; and SEMIOSKYLINE, an interface for browsing documents and unstructured data categorized and indexed by SemioTagger. Industry-specific terminology and categorization supplied with templates.

SOFTWARE AG

HQ: Darmstadt, Germany
Tel: 49 (0) 6151 92 1669
www.softwareag.com

PROFILE

A system software provider supplying data management and e-business technology, focused on the XML standard. Founded 30 years ago with the launch of the hardware-independent ADABAS database.

Public company with fiscal 2001 revenue of Euros 588.5 million.

OFFERING

Includes TAMINO XML SERVER, a data management platform based on XML and other open standard internet technologies, providing native XML storage capability; and XML Server's X-QUERY language based on the XML Path Language (XPath). Other products: ENTIREX, a middleware product for integrating platforms and applications within the enterprise and beyond (e.g., the web).

SPSS

SPSS/LexiQuest

HQ: Chicago, Illinois, USA

Tel: 1 312/651 30 00

www.spss.com

PROFILE

Software company providing predictive analytics and data and text mining tools, including the data mining tool CLEMENTINE. Entered the text mining arena in the February 2002 with acquisition of LEXIQUEST, a Paris-based supplier of linguistics-based text mining software.

Public company with fiscal 2001 revenue of \$187.4 million.

OFFERING

LexiQuest suite of text mining tools includes LEXIQUEST MINE, a discovery tool that reads 250,000 pages of text/hour and presents results using graphical maps; LEXIQUEST CATEGORIZER, a linguistics-powered automatic document categorization and taxonomy management tool; LEXIQUEST Guide, a natural language information retrieval application with add-on modules for creating industry-specific dictionaries of terminology and proper names; and LEXIQUEST RESPOND, a natural language automatic response system for intra- and extranet users.

CLIENTS

Include BNP Paribas, in a corporate search portal application for some 70,000 employees.

SRA

HQ: Fairfax, Virginia, USA

Tel: 1 703/803 15 00

www.sra.com

PROFILE

Software company providing IT services and solutions to US federal government organizations and businesses.

A privately-held company with fiscal 2001 revenue of \$313 million.

OFFERING

Includes NETOWL, a multilingual information extraction engine that creates conceptual indexes; and ASSENTOR, an e-mail screening and archiving system which uses NetOwl technology.

CLIENTS

Some 85 US securities firms, insurance companies, retail and institutional brokerages use Assentor to automatically categorize electronic mail for compliance purposes.

TEMIS

HQ: Paris, France
Tel: 33 1/58 56 48 01
www.temis-group.com

PROFILE

Software start-up founded by (ex-IBM) text mining experts, providing extraction, categorization and clustering software in a multilingual context.

OFFERING

The INSIGHT DISCOVERER suite includes: INSIGHT DISCOVERER EXTRACTOR, which extracts entities and relationships using first a morpho-syntactical tagger and subsequently a set of grammatical rules and thesauri (stored in Skill Cartridges); INSIGHT DISCOVERER CATEGORIZER, a server with supervised and unsupervised classification methods; and INSIGHT DISCOVERER ANALYST, which visualizes text mining results. Pluggable SKILL CARTRIDGE modules provide knowledge extraction rules and thesauri adapted to domains, products or technologies. Linguistic component provided by XELDA, a linguistic engine from MKMS (Multilingual Knowledge Management Solutions), a Xerox document company.

CLIENTS

Include Dresdner Bank in an intelligent information retrieval/extraction system for the bank's employees, allowing navigation in the bank's 10s of 1000s of pages of internal documentation.

THOMSON FINANCIAL

HQ: New York, NY, USA
www.thomsonfinancial.com

PROFILE

Part of the Thomson Corporation, Thomson Financial provides investment information (including FIRST CALL and I/B/E/S), analytics (including VESTEK/QUANTEC), and technology solutions to the financial community. The latter include custom-configured web-based solutions.

Revenues in the range of \$2 billion.

OFFERING

THOMSON ASP SOLUTIONS GROUP builds custom intra- and/or extranet solutions including commingled proprietary and third-party research, analytical tools and search engine capability.

CLIENTS

Include the brokerage firms Gruntal & Co., Scott & Springfellow and Legg Mason, the independent equity research and trading boutique C.L. King & Associates, and the securities firm H.C. Wainwright, with custom configured intra- and/or extranet solutions.

VERITY

HQ: Sunnyvale, CA, USA
Tel: 1 408/541 15 00
www.verity.com

PROFILE

Technology company providing business portal infrastructure software including technology for full-text search, classification, categorization and personalization.

Public company with fiscal 2001 revenue of \$145.0 million.

OFFERING

VERITY 2K ENTERPRISE portal infrastructure software provides the following functionality: classification, using business rules to edit the definitions of categories and sub-categories; the creation of taxonomies either manually or generated automatically from clusters produced by K2 Enterprise; search, including full-text (fuzzy and concept extraction) search, category drill-down, parametric search (which combines factual queries and data mining) and federated search (i.e., single query brings back results from multiple sites and information sources). Other features include adaptive relevancy ranking, document recommendation, expert location and support for 26 languages.

CLIENTS

Include many of the top 10 Wall Street firms, to build a central repository for equity research; and the online investment research service supplier FactSet who is using Verity technology for indexing, parsing, storage and retrieval.

For more information

General references in text mining, see wfan@umich.edu.

Ontologies (taxonomies), for an introduction to ontologies, see www.SemanticWeb.org.

RDF, for a formal description of the Model-Theoretic semantics for RDF, see the working draft of RDF Model Theory available on the W3C Web site www.w3.org.

Research projects, for information on the European Union Sol-Eu-Net text and web mining project, see soleunet.ijs.si.

Semantic Web, see World Wide Web Consortium (W3C), www.w3.org.

XML, for a review of XML in finance, see "The Role of XML in Finance", Anthony B. Coates (tony.coates@reuters.com); Originally presented at the *XML 2001 Conference*.

For more information on XML and database issues, see Ronald Bourret, www.rpbourret.com.

Abbreviations used

AI, Artificial Intelligence

AIMR, Association for Investment Management & Research

ASP, Application Specific Provider

CRM, Customer Relationship Management

EBRL, Extensible Business Reporting Language

ECN, electronic communications network

FISD, Financial Information Services Division

FpML, Financial products Markup Language

GAAP, Generally Accepted Accounting Principles

HTML, HyperText Markup Language

IAS, International Accounting Standards

ISDA, International Swaps and Derivatives Association

ISO, International Standards Organization

MDML, Market Data Markup Language, replaced by MDDL

MDDL, Market Data Definition Language

MIT, Massachusetts Institute of Technology

ML, Markup Language

NewsML, News Markup Language

PMML, Predictive Model Markup Language

RDF, Resource Description Framework

RIXML, Research Information Exchange Markup Language

SGML, Standard Generalized Markup Language

SIIA, Software & Information Industry Association

STP, straight-through processing

W3C, World Wide Web Consortium

XML, eXtensible Markup Language

About The Intertek Group

The Intertek Group is a Paris-based firm that provides research, consulting and training on advanced IT and modeling techniques in the financial services sector and industry at large.

About the authors

Sergio Focardi

E-mail: interteksf@aol.com

A founding partner of The Intertek Group, Sergio Focardi consults and trains on modeling and high-performance computing.

Sergio has been a guest lecturer at Yale University (on risk management), at the Engineering Faculty of the University of Genoa, Italy (on financial engineering) and at the University of St. Gallen, Switzerland (on technological innovation). He has published numerous articles and co-authored the books *Modeling the Markets: New Theories and Techniques* (F.J. Fabozzi Associates, 1997) and *Risk Management: Framework, Methods and Practice* (F.J. Fabozzi Associates, 1998) and contributed Chapter 3, "The Changing Framework and Methods of Investment Management", to the *Handbook of Portfolio Management*, editor Frank J. Fabozzi (F.J. Fabozzi Associates, 1998).

Sergio has 25 years experience in technical and scientific computing. Prior to founding The Intertek Group, he held various positions at Digital Equipment Corporation and Data General and more recently was Managing Director, Italy at the supercomputer company Control Data. Sergio holds a degree in Electronic Engineering from the University of Genoa and a post-graduate degree in Communications from the Galileo Ferraris Electrotechnical Institute (Turin). He is a co-founder of CINEF, the Interdisciplinary Center for Economic and Financial Engineering at the University of Genoa.

Caroline Jonas

E-mail: intertekcj@aol.com

A founding partner of The Intertek Group, Caroline Jonas heads client research projects for end users and technology vendors, as well as Intertek field research projects including the recent four-part survey *Quantitative Methods in Asset Management*.

Caroline has authored numerous research reports and co-authored the books *Modeling the Markets: New Theories and Techniques* (F.J. Fabozzi Associates, 1997) and *Risk Management: Framework, Methods and Practice* (F.J. Fabozzi Associates, 1998).

Caroline has 25 years experience in high-tech research and marketing. Prior to founding The Intertek Group, she was Senior Consultant, Europe with the Palo Alto-based high-tech marketing consulting firm Regis McKenna. In this capacity, she was responsible for pan-European projects for firms such as Apple, Hewlett Packard and National Semiconductor. Caroline holds a degree in Political Science from the University of Illinois, Urbana-Champaign.

For additional copies

Additional copies can be ordered using this form or from The Intertek Group web site (www.theintertekgroup.com).

Please send me the management report *Leveraging Unstructured Data in Investment Management* (65 pages):

- via e-mail
 hard copy via post.

Price per copy:

- Euro: 195.-
 France only: Euro: 233.22 (includes TVA at 19,60%)

Discussion sessions

Half-day discussion sessions with an author: Euro: 1,145.- (exclusive of travel costs). For more information, please e-mail (info@theintertekgroup.com) or call +33 1/45 75 51 74.

Name _____
 Company _____
 Address _____
 City/Postal code _____
 Country _____
 Title/position _____
 Business phone _____
 E-mail _____

Billing address if different from the above:

Address _____
 City/Postal code _____

Payment

- Bank transfer
 The Intertek Group, account n. 00020061055, Société Générale, bank code 30003, agency code 3497, international SWIFT code SOGEFRPP
 Enclosed is a check for Euro _____
 Payable to The Intertek Group
 Please invoice me / my company.

To avoid VAT charges, companies in EU member states (except France) are asked to supply identifying numbers (BTW, FPA, IVA, MOMS, MVST, TVA, VAT):

Please send your order with remittance to:

The Intertek Group
 94, rue de Javel F-75015 Paris
 Tel: +33 1/45 75 51 74 www.theintertekgroup.com