RC23311 (W0408-140) August 27, 2004
Computer Science

IBM Research Report

Text Analytics for Asset Valuation

Wlodek Zadrozny
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Text Analytics for Asset Valuation

Wlodek Zadrozny. IBM Research. wlodz@us.ibm.com

Tools for evaluating intangible assets are emerging and will change the investment landscape.

Intangible assets, such as brand value, customer opinions or management quality, constitute 80% of stock market valuation. Moreover, as a percentage, the proportion of intangible assets is increasing (Fig. 1). However, there are few tools for evaluating and comparing intangibles. This situation is about to change: tools for evaluating intangible attributes of value are emerging; they use text analytics and data mining, and exploit information integration to bring together disparate data sources. The coming change could be sudden, because there is a core of a hundred or so attributes used to evaluate intangible assets, and the existing technologies are capable of adequately extracting their values. When this process completes, the new tools and data repositories will allow investors to quickly review company performance with respect to the intangibles in the same way as spreadsheets and balance sheets currently do for the tangibles.

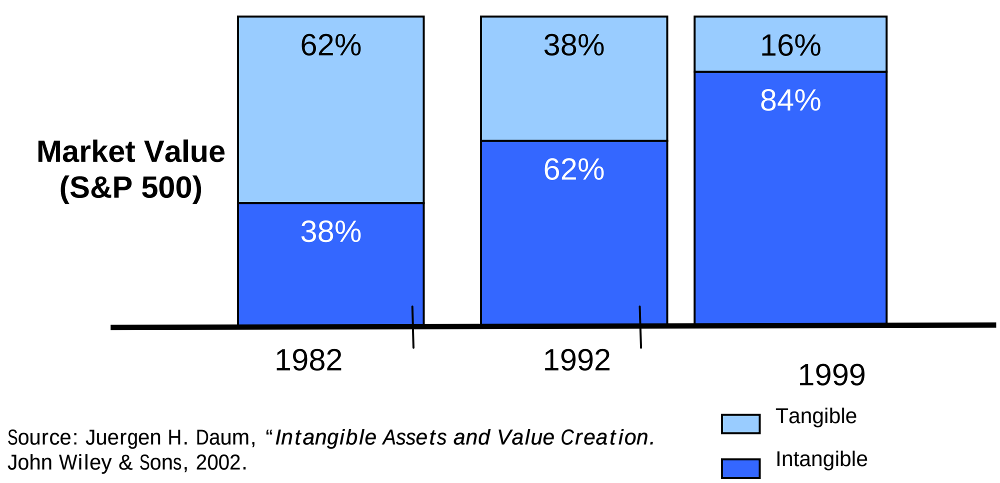

Fig. 1. Intangible assets increase as a percentage of market value.

Currently, any query referencing relative brand positions, quality of workforce or management, and the legal condition of a company would require painstaking research. For example, finding *"companies with a good set of brands, good technical workforce, growing slower than its market and with no poison pill"* would require deciding what constitutes a good brand, researching the brands, the workforce and its technical training, analyzing legal data and linking it with financial, cross-company data. Company filings might contain some of this information, but some might appear in other sources, e.g. news stories. In contrast, any investor can easily ask a query mixing almost any number of a hundred or so standard financial attributes (such as revenue growth rate, sales per employee and debt).

This situation where investment research is a manual and painfully slow process is about to radically change. Text analytics, data mining, information integration and intelligent search will allow analysts or investors to get answers to queries like the example above.

The new process will go roughly as follows: Information integration makes relevant documents and data accessible for search and analysis. Search allows the analysis to focus only on a relevant set of documents. Text analytics can extract meaningful data from documents about a company, its competitors and industry. Data mining can discover statistical correlations between this newly extracted data and standard accounting information.

Automated value analysis is not yet done on a large scale for three reasons:

- Maturity of the technologies
- Solutions have to be custom made and skills are fragmented
- Lack of regulatory requirements

The first two inhibitors are disappearing with the technology progress. Even though a number of companies have been taking advantage of new technologies, this fact has not been noticed by the market at large, because most current solutions are custom made, require very specialized talent, and are not easily replicable. However, emerging services-oriented architectures for text analytics address both the skills shortage and the need for custom made solutions[1].

Evaluating and comparing arbitrary intangibles is occasionally described as "accountancy's holy grail"[2]. If we are correct in our assertion that text analytics and data mining enable direct analysis of all attributes of business value, the regulatory requirements likely will follow. (Legal analysis is beyond scope of this report[3]).

Direct analysis of intangibles

In contrast to analyzing intangibles indirectly, e.g. through cash flow analysis or audits,[4] text and data analytics allow the investor direct view of selected intangible attributes of business value. Five examples of such direct analysis and management of intangibles are presented in Fig.2.

---

[1] IBM introduced Unstructured Information Management Architecture (UIMA) as a standard for combining several text analytics engines, for instance classification into a predefined set of categories with subsequent machine translation and information extraction from translated text
www.computer.org/computer/homepage/0303/briefs/r3022.pdf
Similarly, there exist open, XML-based standards, for annotating data with meaning, e.g. RDF, DAML and OIL http://www.daml.org/2001/03/reference.html

[2] E.g. http://thomsonscientific.com/ipmatters/acctecon/8179924/#paul_gosling
[3] Some aspects are discussed in J. Hand and B. Lev (eds.), Intangible Assets: Values, Measures and Risks, Oxford University Press, 2003
[4] See e.g. Harvard Business Review, June 2004, papers by Lev, and Ulrich and Smallwood

Attribute	Technology Used	Description or Comment
Brand perception	Text mining of Internet content and data mining of the results of text processing.	WebFountain[5] is an IBM service mining Internet data to see trends in brand value perceptions and alert companies to emerging problems. It uses text and data analytics and can be combined with internal company data e.g. to measure effects of a marketing campaign.
Management experience and quality	Data analysis	Barr&Siems 1996[6] used statistical data analysis for bank failure prediction. The main variable they measured was "management quality". Their model detected "a bank's troubled status up to two years prior to insolvency using publicly available data."
IT Investment, Joint ventures, Marketing alliances	Text mining[7]	Temis' *Online Miner™* extracts the principal areas of investment of a company, the agreements signed by different companies, etc. from targeted public sites as well as press reports
Compliance (Quality of internal processes)		Inxight's *Smart Discovery* for Sarbanes-Oxley can e.g. provide random sampling for compliance with internal policies such as revenue recognition.
Patents value		At Dow Chemicals, Clearforest's text mining solution helped in "identifying licensing and M&A opportunities around new product development".

Fig. 2. Examples of direct analysis of attributes of business value

We have included in Fig. 2 two solutions focused on management of intangibles to make a point that some intangibles can be best analyzed using company internal data (e.g. quality of internal processes), but approximations of some attributes (e.g. patent portfolio analysis) are possible.

Showing the solutions for managing intangible assets--patent portfolios and quality of internal processes--also makes a point that the very fact of using technology to manage intangibles is sometimes disclosed in SEC filings, press, and executive presentations, e.g. Clearforest[8] reported 30 to 50- fold productivity increases at Dow Chemicals around competitive intelligence and material research. Since it increases productivity, the use of advanced IT technology itself is an attribute worth mining.

We will present a systematic discussion of using technology for analysis of intangibles in one of the next sections. We will observe that text mining can be used to analyze information about dozens of intangibles, and not only the few shown in Fig. 2. The reason is that the same techniques, often involving sophisticated grammatical pre-processing and statistical analysis, can be used for mining arbitrary attributes if the text mining engine has access to appropriate dictionaries of patterns. Similar observations

---

[5] www.almaden.ibm.com/webfountain/ see also www.spectrum.ieee.org/WEBONLY/publicfeature/jan04/0104comp1.html

[6] faculty.smu.edu/barr/pubs

[7] www.temis-group.com/, www.inxight.com/pdfs/SOX_ApplicationOverview.pdf, www.clearforest.com,

[8] www.clearforest.com/Customers/Dow.asp - 22k - Aug 12, 2004

apply to data analysis and data mining. For example, Barr and Siems approach can be extended to cover other industries, because historical data is available in executive profiles databases and can be easily correlated with data on the company. Also, patent analysis can be applied to externally available intellectual property information – even though the example application has to do with management of intellectual property, and not investment analysis.

The path from research focused on only one type of asset to comprehensive investment tools will go through creating technology for aggregating and dissecting results of analysis for many intangible attributes. Current text analytics tools are mature enough to provide stepping stones for this path.

Maturing capabilities of text analytics

Text analytics and data mining will allow countless documents to be scanned for information about a company. To be successfully used, the two technologies require integration of many data sources. Of these three ingredients, information integration and data mining are more mature, and large scale solutions are available.

Similarly, large portions of the text analytics capabilities[9] can already be leveraged for investment analysis. For example, as shown on Fig. 3, automated categorization – that is distributing documents into a number of predefined classes -- exceeded human performance around year 2000. Automated categorization is often used in news analysis, automated and semi-automated replies to customer queries, and to speed up search. Where they are deployed, computer categorizers are much more productive than humans. Similarly, text extraction capabilities – i.e. being able to draw a particular data point from text, such as the number of employees-- are good enough to augment or replace humans, as we have seen in a few examples in Fig.2.

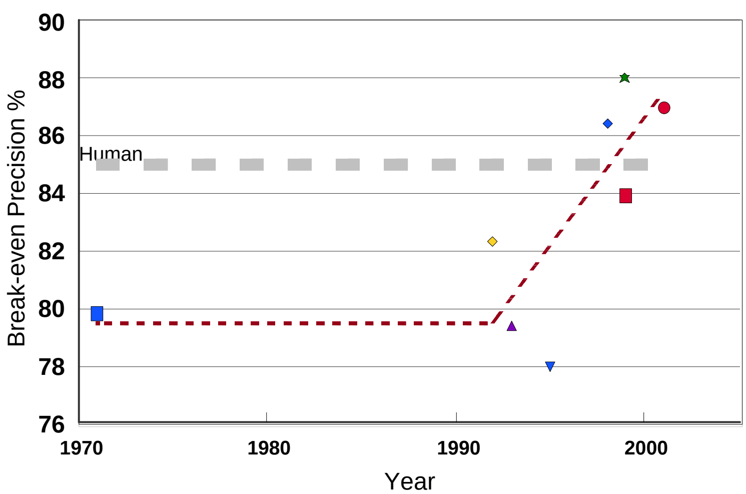

Fig. 3. Automated categorizers exceeded human performance around year 2000 on standardized test sets. (Y-axis shows the accuracy of different categorization methods and a baseline human performance).[10]

---

[9] Recent examples from IT press http://www.computerworld.com/printthis/2004/0,4814,93968,00.html

[10] Source: IBM Research

In addition, less mature technologies like machine translation and extraction of complex relationships are steadily improving. Machine translation for a particular domain (e.g. government information or technical manuals) can be engineered to provide high quality translation that does not require human post-processing. Although we are not aware of using automated translation in pre-processing for data extraction for financial analysis, there have been successful experiments in multilingual information extraction and search in other domains. Some types of financial data might lend themselves better to automated translation, because of restricted vocabulary and style.

Fig. 4 summarizes the status of a few key text analysis technologies relevant for financial analytics. Among them is semantic search[11], that is, the ability to find documents based on their meanings, not the words used. For example, even though "educated workforce" might not be mentioned in words in a filing, a search engine could put it in the index, based on information about the company's cost of internal training. Semantic search will be relevant in reducing the amount of noisy data for automated and human analysis.

Solution Technologies	**Status (average accuracy)**	**Example Financial Analysis Applications**
Categorization	~90% (human accuracy; 100 times faster)	Routing documents to analysts or data extraction modules. Classifying results of search.
Data extraction	~70%	Extracting financial, legal and other data from company filings, news, transcribed broadcasts and meetings.
Relationship extraction	~30-60% (for some data points performance is 90%)	Extracting complex data. E.g. monitoring management changes. Alerts based on events (not key-words).
Machine Translation	Good "gist" translation. Emerging high quality translation.	Access to filings and news in other languages. Pre-processing for data extraction.
Search (key-word)	Very good quality on the Internet. Not so good within enterprises	Finding relevant documents or passages, creating ad-hoc textual databases for further analysis/extraction. Semantic search provides orders of magnitude improvement in search quality.
Search (semantic)	Emerging	

Fig. 4. Status of some key text analytics technologies

As we observed, text analytics and data mining enable direct analysis of attributes of business value. These technologies will drive the revolution in how companies are analyzed, because they will amplify human analysts' capabilities. Such new powers are needed; e.g. in 2003 *McKinsey Quarterly* commented *"analysts urgently need to deliver more relevant, more original, and better-targeted research to justify their cost"*. Text analytics substantially reduces the cost of research and makes it more focused.

For this vision to be realized, in addition to the quality of the technologies, there must be some agreement on the directions of analysis.

[11] Maas http://www.w3c.org.il/events/semWebEvent/maas.pdf discusses an approach to semantic search

Approaches to valuation of intangible assets focus on similar sets of attributes
At the surface, different approaches to asset analysis focus on categories of value that sound similar, but seem to view assets from different perspectives; e.g. tangible vs. intangible; intellectual capital, human capital, structure capital, customer capital etc[12]. Some academic discussions may focus on macroeconomic aspects of intellectual capital. This might seem remote from the practical task of asset evaluation in mergers and acquisitions. Also, there are few relevant accounting standards.

For example, in the Value Chain Scoreboard™, Lev[13] considers three groups of intangibles: *Discovery/Learning*, *Implementation* and *Commercialization*. Each group, in turn, is divided into three or four subgroups focused on a specific topic. Thus *Discovery/ Learning* focuses on *Internal Renews, Acquired Knowledge* and *Networking*; *Implementation* contains topics of *Intellectual Property, Customers* and *Employees*; and *Commercialization* deals e.g. with *Bottom Line*. These topics in turn are divided into themes (attributes). For example, for *Networking* these are *R&D Alliances*, *Joint Ventures, Supplier and Customer Integration*. *Bottom Line* consists of *Productivity Gains, Online Supply Channels, Earnings/Cash Flows, Value Added*, and *Cash Burn Rate*. Altogether, there are about three dozen specific attributes.

Ballow et al. in an Accenture report[14] list about sixty types of assets in two dimensions: Tangible vs. Intangible and Traditional Accounting Assets vs. Intellectual Capital Assets. Several books on quantifying the economic value of intangible assets for mergers and acquisitions make their own -- partly overlapping, but distinct – taxonomies.[15]

Thus, in contrast to tangible assets, there appears to be no common standard on how to measure intangibles. Therefore, intangible assets valuation could seem too immature to benefit from automated or even semi-automated analysis.

However, one level deeper, if we disregard their taxonomies, all these approaches deal with similar *attributes of value*. They include profits from new business, training per employee, customer loyalty, management quality, quality of internal processes etc. In total, there seem to a list of a hundred-or-so attributes covering all the approaches. (A comprehensive list is presented in Fig. 5).

Having only a hundred or so types of intangibles to measure means that, in fact, there is agreement on what constitutes a valuable asset.

---

[12] See http://www.sveiby.com/articles/IntangibleMethods.htm for a list of pointers for measuring intangible assets
[13] "Intangibles: Management, Measurement, and Reporting" Baruch Lev. Brookings Institution Press 2001
[14] J.J.Ballow, R. J. Thomas and G. Roos "Future Value: The $7 Trillion Challenge" http://www.accenture.com/xd/xd.asp?it=enweb&xd=services%5Csba%5Chotidea%5Cvalue.xml
[15] See e.g."Valuation of Intellectual Property and Intangible Assets" Gordon V. Smith, Russell L. Parr. J. Wiley and Sons. 2000. "Valuing Intangible Assets" Robert F. Reilly, Robert P. Schweihs. McGraw-Hill. 1998. "Intangible Assets and Value Creation". Jurgen Daum. J. Wiley and Sons. 2003.

Most intangibles can be automatically analyzed

We have analyzed a comprehensive list of intangibles to assess the possibility of extracting them with existing tools for data mining and text mining. Keeping in view how they can be used for investment analysis or in mergers and acquisitions, we looked at how difficult it would be to extract their values when appropriate textual and numerical data are available. We concluded that 40 to 60% of attributes of value can be extracted and analyzed using existing commercial technology; another 15 to 25% by existing technology in development; with the remainder split between those that can be extracted and analyzed using advanced research tools and the ones that require human insight. This surprising observation leads us to believe that comprehensive solutions amplifying several times human abilities to analyze and compare intangible assets of companies can be built now using text analytics, data mining and information integration.

The intangible attributes can roughly be grouped into five categories, depending on the type and availability of solutions required to mine them. The easiest is the case when information is already available, for instance profit per employee can easily be computed from reported financial data. In other cases, as seen in example solutions in Fig. 2, text extraction can be used. A more difficult case occurs when information must first be gathered and organized. This is a WebFountain-like solution. Finally, there are two cases where analysis is not possible: either it would require access to confidential company data or it must be driven by human insights (e.g. to judge structural appropriateness of a company).

Fig. 5 lists ninety common intangible attributes of value. They have been derived from a list of about two hundred attributes from several sources[16] by removing repetitions, and in a few cases were grouped by topic for conciseness. Some redundancy was preserved though to avoid too much abstraction, and point out that when evaluating a class of assets, the focus might be on different attributes, e.g. “Software”, “Systems” and “Technology Purchase”, which could even be extracted by different text analysis engines.

The attributes differ in their degree of specificity: “Innovation” vs. “Frequency of repeat orders”. The latter should be more easily measured. The former can be benchmarked after some measurement criteria are established, for instance, percent of revenue from new product and services[17]. Only a few of them, e.g. “Structural Appropriateness”, imply the need for judgment, and have no obvious way of measuring without building an elaborate model of the enterprise.

Fig. 5 also provides a visual representation of our analysis. We have color coded sets of attributes depending on the degree of difficulty and accessibility of data. It can be readily seen that only a small set of attributes requires human insight or specialized benchmarks

[16] In addition to the sources quoted above e.g. ”Introducing the new Value Creation Index Geoff Baum, Chris Ittner, David Larcker, Jonathan Low, Tony Siesfeld, and Michael S. Malone” , 04.03.00 www.forbes.com/asap/2000/0403/140_3.html.
J Liebowitz provides several lists of intangible assets in http://organik.kmworld.com/upload/6/315/3344/developing%20KM%20metrics%20for%20measuring%20IC.pdf

[17] (cf. e.g. p.511, Hand & Lev, op.cit.)

for assessment. Also, there is a group of attributes that could be easily mined if reported, such as retention or “Win/Loss” index. Obviously there are cases, like “Joint Ventures” when information is both available from a database (e.g. Thomson Financial), and presumably extracted by an analyst, and amenable to automated extraction.

Ability to attract talented employees. Losing talent.	Competence: Index, Turnover,	Employee experience	Investments in Internal Structure	Online Revenues	Productivity Gains	Records and Drawings (ie. Proprietary databases - WZ)	Software	Technology Purchase
Access Rights	Competence -Enhancing Customers	Environmental Performance	IT Acquisition	Online Supply Channels	Professionals Turnover.	Regulatory imposts	Stickiness and Loyalty Traffic Measures	Top management experience
Alliances	Credit ratings	Eyeballs (usage raffic)	IT Development	Organic Growth.	Profit per Employee.	Relative Pay.	Stranded assets	Top management quality
Borrowing capacity	Customer Acquisition Costs	Formal alliances (e.g. JVs, supply agreements)	Know-how	Organization Enhancing Customers.	Profit per Professional, customer,	Research and Development	Strength of stakeholder support including opinion leaders	Tradability of facilities
Brand (investment, stature, support)	Customer Contracts	Formalized processes	Leases	Organizational reputation	Proportion of Big Customers.	Retention	Structural appropriateness	Training and Education Costs.
Capabilities	Customer Loyalty, Satisfaction	Franchise Agreements	Licence agreements	Patent/Know-how Royalties	Quality of corporate governance	Revenue Growth by Segments	Subscriptions	Undrawn facilities
Cash burnout rate	Devoted Customers Ratio	Frequency of Repeat Orders.	Market Potential/Growth	Patents, Trademarks, Copyrights	Quality of earnings	Revenues from Alliances	Supplier/ Customer Integration	Value Added per Employee.
Clinical Tests, FDA Approvals	Diversity	Informal processes	Market Share/Growth	Plant flexibility	Quality of processes, products or services	Reverse Engineering —Spillovers	Support Staff Turnover.	Value Added per Professional.
Codified knowledge	Employee loyalty	Innovation	Marketing Alliances	Plant infrastructure	Quality of supply contracts	Right to tender, right to compete, right to design	Systems	Values/Attitudes Index
Communities of Practice	Employee Training	Investment in IT	Mastheads	Plant modernity	R&D Alliances/Joint Ventures	Rookie Ratio.	Tacit knowledge	Win/Loss Index.

Type	How the attribute can be analyzed
DB	Existing databases. Data available. Often proprietary (e.g. from Factiva, Thomson or Bloomberg)[18]
TE	Text Extraction from textual databases, for instance Edgar. No crawling required. Predictable data sources.
WF	WebFountain-like solution: crawl Internet, index, extract and analyze. Possibly accessing proprietary web sources. Difficulty lies in the high noise ratio, large amount of data and unpredictability of sources.
P	No clear way of assessing without access to company proprietary information. Some attributes are could be assessed if disclosed.
I	No clear way of using TE or WF. Driven by human insight (but possibly with SW support)

Fig. 5. Most intangibles can be analyzed using existing or emerging technologies.

Note an intangible attribute can be extracted in more than one way, e.g. analysis of business news for marketing alliances can be enhanced by mining the Internet for details,

[18] See e.g. SDC Platinum ™ product from Thomson (thomson.com)

progress reports etc. However, as a rule, the natural starting point is in predictable textual streams e.g. news, company filings and analysts reports. If some information exists in databases (e.g. Thomson Financial), but can be significantly and easily enhanced by text extraction we also classify such problem as text extraction. For “Revenue growth by segments” or “Investment in IT”, for instance, more important than actual number is the type of investment and expectation of its impact. In other cases, like measuring “Employee loyalty”, we assume that the problem can be approximated by using text extraction from existing databases or internet sources (vault.com, monster.com, local paper job changes announcements), even though no one to our knowledge has attempted to do so. Obviously, if companies disclosed such data in their filings the problem would be easier and conclusions more reliable (however a sophisticated statistical model can derive reliable conclusion from noisy data).

The small size of the list of intangibles makes it easy to make the point that, if there exists a solution for assessing a particular attribute, then other attributes, either appearing in the same sources or conceptually similar, can also be assessed. Thus, we will not provide separate recipes for mining ninety intangible attributes, but rather refer the reader to Fig. 2, for inspirational examples dealing with more difficult cases.

How to build a system for analyzing intangibles

So far, we have presented arguments that the number of intangibles is small and that with the help of text and data analytics they can be gleaned from documents. Now, we will present reasons why building a system capable of answering queries on intangibles is possible.

First, recently attempts have been made to create software tools helping with management of intangibles (e.g. d-xyfer.com). This show that a system focusing on evaluation of a comprehensive set of attributes can be made rigorous, even if data feeds are manual or come from a database or a spreadsheet.

The second reason is that the problem of feeding data into such a system is solvable. Namely, as we argue below, service oriented architectures will conquer the solution complexity, and permit the data feeds to be built incrementally.

UIMA[19]– Unstructured Information Management Architecture -- developed at IBM Research can be viewed as a blueprint for putting together many of the required modules. UIMA specifies the interfaces that different modules must adhere to, but otherwise they are viewed as interchangeable. Therefore, they can come from different vendors and be easily replaced as the quality of text analyzers, search engines and available knowledge sources improves.

[19] See footnote 1. Fig. 6 source: www.computer.org/computer/homepage/0303/briefs/r3022.pdf

The modules shown in Fig. 6[20] are clearly relevant to the task of analyzing intangibles: search helps create collections of documents for analysis or directly answers specific queries; collection processing can be used to find trends in company filings; and application libraries can provide semantic or taxonomical knowledge needed by text analyzers.

Clearly, with the huge number of document sources, the task of putting together an application covering all intangibles is formidable. But it can be attempted, in stages.

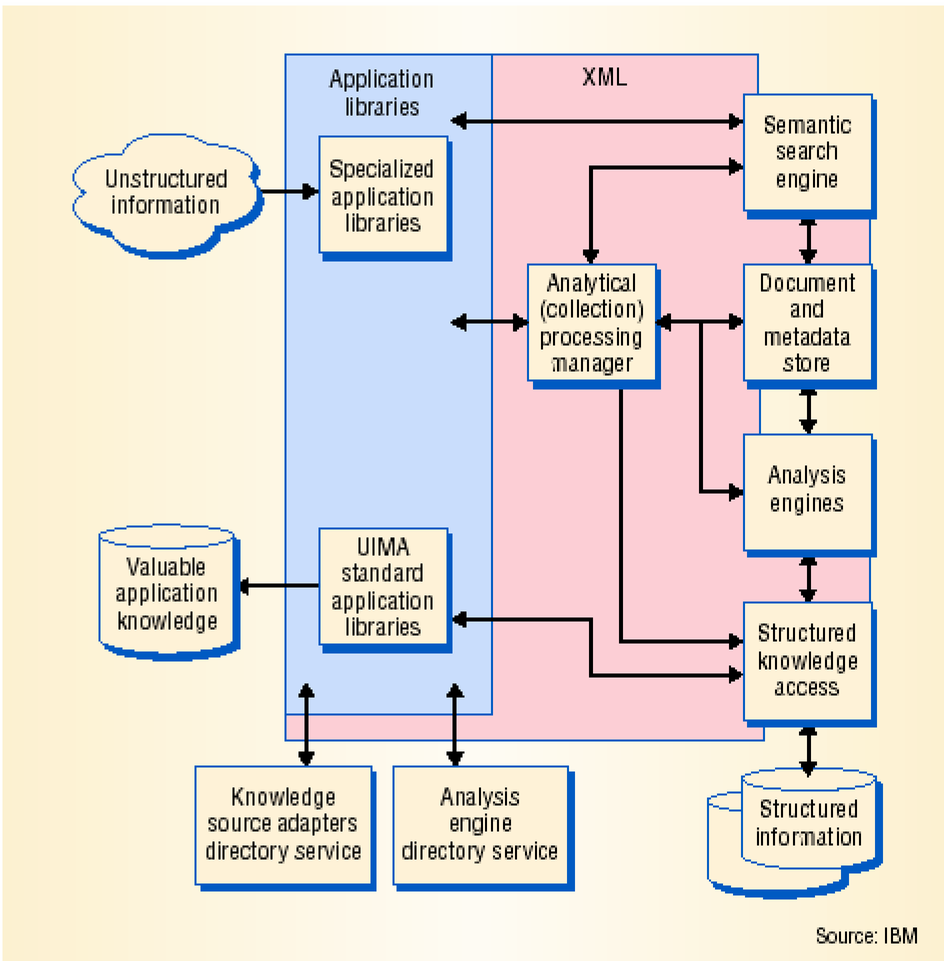

Fig. 6. UIMA – Unstructured Information Management Architecture

There are already systems in operation that integrate large amounts of information from many sources, e.g. news, warranty claims, databases etc. to provide early warning capabilities for car manufacturers.[21] Furthermore, the integration can be applied first to

---

[20] Fig. 6 source: www.computer.org/computer/homepage/0303/briefs/r3022.pdf
[21] http://www.sas.com/news/feature/24mar03/ibm.html describes an IBM and SAS early warning system

existing structured databases[22], before using UIMA-like solution to link them to unstructured data. These would be the natural stages for data integration.

Leveraging text and data analytics for analyzing intangibles can also happen in several stages. It can start with automating data gathering. Or it can address some immediate opportunities:

- Adding a new capability to an existing service. For example
  - Semantic search combined with a database on corporate governance and standard financial data can answer a question whether Bio-tech companies with no poison pills can be overvalued shortly after FDA approvals.
  - Text analytics added to the alert system can provide alerts targeting a legal position of a company or a change of a mutual fund management.
- Adding new, selected new data points by industry, geography or topic. For example
  - Banking might be easier to analyze than Pharma.
  - US and Canada might be more relevant than Japan and Korea. Or vice versa. Introduction of a new accounting standard in Europe might make it a better target.

Further down the road might be an ad-hoc analysis of arbitrary collections of data, which might be required to find *"companies with a good set of brands, good technical workforce, growing slower than its market and with no poison pill."* However, even this type of request can be tackled in semi-automated fashion. For example, taking advantage of data integration, one might first create a list of companies satisfying the last three conditions, and then crawl and mine the Internet to further restrict the list to companies. The crawling could be selective (say, www.consumerreports.org and selected business magazines), and the text mining restricted to positive and negative opinions. Tools for both already exist.

With the agreement on what constitutes a valuable asset, with the emergence of advanced software for information integration, and the maturing of text and data mining, only a small leap of imagination is needed to envision comprehensive financial tools for analyzing intangibles. Markets and investors will benefit greatly from them.

August 20, 2004

[22] e.g. Thomson SDC Platinum ™ databases with standard financial reporting.