Next Generation Text Analysis Software from InfoTame™

A White Paper by Amy Wohl

Executive Summary

This White Paper is an exploration of new technologies that offer help to information users drowning in the vast seas of raw data we are producing daily.

While these technologies themselves are as sophisticated as their inventors – and the elegant results they can provide – this White Paper is not intended for mathematicians, computer scientists, or theoreticians. Rather, it is for senior business executives who need to equip themselves and their employees with the best tools.

This White Paper contains:

- Information on text searching and text mining
- InfoTame's Value Proposition
- Case Studies describing the experiences of current InfoTame users
- Scenarios describing what future users might do with InfoTame software
- Conclusions and recommendations

At the InfoTame web site, you will be able to find more detailed technical information and a technical white paper, suitable for IT specialists.

The Information Explosion

Researchers at Berkeley have estimated that there is enough information produced in the world each year to equal 250 megabytes for every man, woman, and child. They compare that to the text of 250 books or a total of 1.5 exabytes (10^{18}). (A very detailed explanation of this study appears at http://www.sims.berkeley.edu/how-much-info.) This is unique, new information, in addition to the historic information we already have, stored in files, books, databases, and elsewhere.

IBM estimates that in the life sciences, information doubles every six months, growing by petabytes (10^{15}).

The Internet has provided a perfect environment for hosting this information explosion. From 1990 through 2002, the Internet has grown from a few hosts to over 150,000,000 computers.
(http://www.zakon.org/robert/internet/timeline/#Growth)

Almost anything you need – for work or personal information – is somewhere, but finding that information costs too much in time, skills, and money. At times, when we can't find invisible information we just redo the work and create it all over again, wasting valuable time and money that could be invested in further analysis or new research.

Often, what is needed is not the information itself, but the patterns underlying the information and the insights those patterns might provide. We are looking for the trends that an analysis of the patterns (and changes in the patterns) could offer:

- Directions and pace of trends
- Significant changes in trends, including trend beginnings and endings as well as trend inflection points – the infamous start of the "ramp up curve," indicating that a new activity or a new market has begun in earnest.

Beyond Text Search and Text Mining

The tools we have, mainly search engines, just aren't effective enough.

Many users rely on text search engines for finding information, either within their organization or across the Internet. But text search has its limits. With the millions of documents likely to be available on many subjects, there is simply too much information. Looking for the one or two things that are actually useful is like looking for the proverbial needle in a haystack. Specialized skills in searching are required – and detailed knowledge of the eccentricities of each of the web search engines. Each search engine catalogs, searches, and retrieves and presents information differently, with considerable variation in results.

In any case, search engines can only find what you're looking for – text-based articles which contain the words you've specified. A few can be fine tuned to look for pairs of words, for example, depending on their proximity. None can look through vast collections of documents and discover patterns or trends.

Text mining products come a bit closer. They may use a variety of AI (artificial intelligence) and NLP (natural language processing) techniques to index, search, categorize, and summarize documents.

But InfoTame, which is a next generation technology, combining search, text mining, and analytics into a single product, can do more:

- Offers a new approach to extracting information from all kinds of data
- Discovers hidden relationships; searcher can broaden or redirect a search as new relationships are surfaced

- Extract objective, statistically significant links by creating Information Portraits™ of an entire (and very large) set of documents
- InfoTame is language independent, very fast, highly scalable, and easy to use.

An Information Portrait is an ordered list of the most significant associations extracted from unstructured text data.

A patent for InfoTame's unique technologies is in progress. InfoTame is more of a complement than a competitor to many traditional search, text mining, analytic, and knowledge management tools. It enhances existing tools, offers additional areas of functionality, and opens new areas of information for existing applications. InfoTame does this by analyzing information in ways which provide new dimensions and insights.

- InfoTame can convert intangible or non-quantifiable data into tangible and quantifiable information which can form the basis for key business decisions. For example, references based on emotions can be extracted from emails and notes and used to analyze customer satisfaction, employee or customer perceptions of an event or communication, or reactions to a change in policy.

  This means companies don't have to wait to perform expensive (and often statistically invalid) surveys, but can seek continuous feedback by letting InfoTame explore existing dynamic data collections.

- News data bases, chat rooms or other sources can be mined to determine the hot current topics. These topics can become the basis for new predictive models of more accurate financial analysis.

  Imagine a consumer company that can tune into the ten most significant topics associated with, for example, "sports" in teenage chat sites and builds product promotions around this "buzz." Think of looking for changes in attitude toward travel as an input to planning inflection points for changes in ticket prices.

  None of this need invade individual privacy. InfoTame is looking for aggregated data to create trend information. Individual consumers' identities are entirely protected.

- Security and law enforcement investigations can increase their quality by employing standard InfoTame probes (pre-built complex queries) and comparing the Information Portraits obtained.

InfoTame creates an impact on a business not only by automating existing processes or by improving speed or efficiency, but by enabling new methods of research and new insights into existing areas. These new techniques allow companies to make significant progress rather than just incremental improvements.

The InfoTame Value Proposition

InfoTame's search, analysis, and pattern identification tool has business benefits for nearly any business.

InfoTame provides benefits to nearly every department of its customers:

- Not just a search tool, but a tool that will return relevant, relational search results
- Results that go beyond Search to Analysis
- In a real time environment, supporting applications that require continuously updated information
- Including any data or combinations of data required (not just data indexed by web search engines)

In particular vertical markets, InfoTame can offer additional benefits which are especially appealing.

CRM (Customer Relationship Management)

Business organizations of any size or type have learned that the best customers are the ones you already have. But this means knowing as much about your customers as possible and leveraging this knowledge into enhanced relationships and increased revenues. That means being able to look at every kind of organizational information describing the customer relationship – from orders and invoices to emails, sales reports, and focus group data – and figuring out just what it means.

All this must occur quickly, so that sellers can react swiftly to customer concerns and needs, lest customers go elsewhere and opportunities be lost. Even more important, as noted in a recent McKinsey two-year study of customer loyalty, small changes in customer spending can be more important than actual defections.

The InfoTame solution allows sellers to include ***any*** type of textual data in their analysis, without the need to perform time consuming and complex transformations. It also permits not only searching for the occurrence of specific words or phrases, but also the identification of patterns and time-related trends, often as they are occurring. This allows the organization, for example, to detect customer mood changes early and react appropriately.

Competitive text mining products typically require extensive work to permit searching across multiple document data bases and data types, so that real time analysis is not an option. Competitive information retrieval products focus on finding the individual instances of particular words – generally offering up the documents in which they're found rather than any underlying patterns these words might represent. Too, the information retrieval engines either offer up enormous volumes of information – too much to allow a researcher to look through much more than the first few per cent – or require very high levels of training and skill to create smaller and more relevant document collections, running a risk of throwing away some relevant information during the narrowing process.

Financial Services

Financial Services firms are typically looking at the marketplace, searching for information on the firms they cover and for overall marketplace trends. They seek advantage in discovering information and forming opinions ahead of their competitors – even if it's just a few minutes or hours earlier.

With its ability to offer real time analysis of very large, unstructured document collections, InfoTame can permit financial services firms to search broadly and still quickly pinpoint trends and inflection (change) points in the markets they follow. Competitive text mining products might provide them with some pattern information, but not likely in real time, especially if it required looking through multiple and heterogeneous document data bases. Competitive information retrieval products simply return too much information to be useful in this application.

Biotechnology

In various biotechnology applications, such as pharmaceutical research, the problems are various and different. Researchers are trying to keep track of research across a vast number of topics and geographies – nearly any new discovery – plant life, animals, changes in knowledge about pathology and treatment – might be important. At the same time, researchers must track what is being done successfully (and not successfully) in their own companies, in their competitors, and in the academic research environment.

InfoTame is particularly useful for biotechnology research because it can search across very large, heterogeneous unstructured text. Much of the material biotechnologists need to be looking through for new information exists in the form of journals and other papers published in various forums (conferences, web sites, etc.) as well as the public filings of governments. Other products would find it difficult to handle the heterogeneity or the scale of such searches without extensive preparation time – and, for these researchers, time is of the essence here. But one unique benefit of InfoTame's product is perhaps the most important for Biotechnology applications – discovering hidden links between a gene and a desease, drugs and side effects, symptoms and syndromes and treatments, and so on. You simply cannot find them with other technologies because you cannot search for a link that you don't know exists.

Security and Law Enforcement

Security is a high profile subject today. The focus is on preventing security breeches, rather than identifying the culprits after the harm is done. The latest FBI report on security makes it clear that there is much work to be done in identifying potential problems and acting to eliminate unintentional access.

For law enforcement, prevention and detection in a dangerous world happen in all too real time.

InfoTame helps security and law enforcement professionals by offering them an additional tool to search through very large unstructured text files, seeking to find patterns to help identify potential trouble – and potential troublemakers. Again, the emphasis is on scalability as well as language independence, since crime fighting and security are global issues. Real-time trend analysis can help pin-point dangerous situations so that security resources can be focused to their best advantage. InfoTame's ability to accept native language queries and offer graphical presentation of trends can be particularly helpful in letting trained security or law enforcement professionals work directly with the tools, rather than through intermediaries who might slow the process. And, similar to Biotechnology, the discovery element is very important: with InfoTame, you can find links that are totally unexpected and may lead to breakthroughs in very convoluted and complex cases involving well-developed networks of criminal elements.

Succeeding With InfoTame; Current Users

InfoTame is not a new company, but rather a European company which is now introducing its technology and products to the North American market. It has a 5 year track record and more than 25 customer success stories since the commercial introduction of its first product in April 2000. The three stories below will give you a good idea of the range and depth of InfoTame's appeal.

A Major Media (TV and Radio) Company

Media outlets need to **track the interests** of their public on a daily and weekly basis. They do this by collecting and analyzing press coverage (from many sources, not just their own). In this major media organization, 4,000 to 5,000 individual documents are being analyzed daily – too many for human attention, but not for InfoTame, which can highlight major topics and analyze interest levels.

In a test of the InfoTame product, the Media Company had InfoTame analyze 4.7 million documents from 500 newspapers, magazines, and journals for September 5, 2001 and asked InfoTame's trend analysis to predict what the major news topics would be on September 9. InfoTame correctly selected 3 of the top 4 major topics, 10 of the 18 total major topics. InfoTame and the Media Company concluded that the Information Analysis Engine can be used to select the most significant topics for in-depth studies or reviews and to recommend topics for coverage.

Radio and TV stations need to **provide background** for news stories – an analysis of causes, related events, historic data, and relevant players. InfoTame can provide this background information to the writer and save hours of research time and cost. When stories break or change course unexpectedly, the speed of providing background information may, in fact, be more valuable than the cost savings.
InfoTame creates an Information Portrait of each analyzed object. As changes occur, subsequent searches of the document database will verify elements in the Information Portrait and **identify trends**. Forecasts can be provided, by the system, based on these identified trends.

InfoTame Technology can also help define the main topics for objects to research, creating a set of dynamic categories for the documents retrieved. This avoids the need for a skilled taxonomist or a separate taxonomy product.

A European Political Party

A European political party must collect and analyze a broad range of information about all of the parties and politicians in its region, including the interests of all the players, the electorate, and connections between politics and business organizations and activities. This includes analyzing such regional information as corporations' financial interests, business problems, and comparisons of the region's image to that of its neighbors.

To assist in these tasks, InfoTame provides the Party technology to understand voters' concerns by **discovering** the most significant topics in their letters and identifying voter **trends** that will require political action.

Geographically Distributed Oil Company

A large and geographically distributed oil company with a headquarters operation, regional offices, and branch offices needs to be aware of all of the information being collected across its organization.

- Headquarters needs to monitor the internal document flow and to process large volumes of internal and external corporate correspondence, much of it flowing across disparate local area networks, into a single, logically aggregated, textual collection.

- The corporation needs to be aware of changes in the regions and branch offices. InfoTame can monitor and alert management when events occur by detecting changes in the Information Portrait caused by external events even if the event's connection is hidden from view.

- InfoTame can compare the Information Portraits of various elements of the Oil Company, using documents from different sources, revealing distinctions and abnormalities which would not otherwise come to management's attention.

Future Scenarios for InfoTame Users

Corporate Portals

Intranets have changed in the past few years from being interesting IT experiments to being the major way for organizations to communicate with their employees. Increasingly, we do this through Corporate Portals. That is, users are linked to organizational information through their identity and their organizational role, with portal filtering software providing each user with a personal view of the information. This provides a kind of first-order filter, so that some of the completely irrelevant information is removed and important information is more readily noted.

Extranets are simply Intranets which extend beyond the corporate boundaries to include important partners – contractors, suppliers, and even customers.

In a partner extranet, InfoTame might help insure that partners were satisfied with a new program or pick up their concerns about changes in pricing, distribution policies, or allies, allowing the site owner to quickly apply appropriate course corrections and make certain that important programs meet their goals.

Companies also employ portals to support **Customer** relationships. This may include product and support information, on-line order-taking, or simply interaction with customers to support good customer relationships – everything from a place to file complaints and suggestions, to information (perhaps travel, food, or health) related to the company's products and interests. Once customers use the portal to interact with their supplier, InfoTame can peer through their correspondence and other text-based interactions (protecting customer anonymity, if desired) and report back the results of advertising and marketing campaigns (both the supplier's and his competitors), the effect of outside events, and changes in customer attitudes. This permits the supplier to better understand his customers and to quickly and proactively change messages, campaigns, and even products.

Market Research

Market researchers are in the business of teasing needles out of haystacks. They routinely exercise their pretty good search skills on behalf of their clients' needs, but that still leaves them at the mercy of the limits of search engines and their eccentricities. In fact, every market researcher knows that sometimes just not searching on-line may be faster, even though you're sure the best stuff must be there, somewhere.

With InfoTame there are far more choices. Researchers are not restricted to engine-indexed documents, but can access other document collections and include them in the search. Research is not restricted to turning up documents that contain particular words, but can include pattern and trend searches which may be particularly useful in pinpointing critical time periods of intense activity or change in direction.

Market researchers may be more willing to take on prediction assignments, using trend information discovered by the InfoTame engine.

Legal: Litigation, Content Analytics, and IP-Related Research

Today, the legal profession is moving toward supporting litigation with computer-based tools, but the tools we have can be frustrating and slow. We amass huge volumes of documents – emails, depositions, files of every description, but we have only fairly crude search techniques to look through them for interesting and useful information. This requires human eyes, trained in computing and the law, and much time.

InfoTame can do much better, breezing through large volumes of any kind of textual information with ease and speed and retrieving specific items or producing comparisons between different participants in the legal process – comparing their InfoTame Information Portraits to help understand the dynamics of their document trail as well as any underlying patterns or trends hiding in the text.

InfoTame tools can be used both in the pre-trial preparation phase, as well as during the trial when its real time features are particularly valuable.

InfoTame can be used to add content analytics to existing legal document management systems that a law firm may already have in place.

For firms involved in IP (Intellectual Property) Research, such as patents, InfoTame can be very useful for finding relevant information about existing patents or patent applications as well as published information about prior art, looking across a vast array of news, academic, and other textual data bases.

Conclusions

InfoTame provides a better way to find new kinds of information in large volumes of textual data. Through its speed, its ability to create Information Portraits of Document collections, and its ability to analyze documents not just for their word content, but also for their meaning and for the rise and fall of trend patterns, InfoTame can enable unique new ways of exploiting an organization's documents.

InfoTame is particularly well suited for certain kinds of customers and certain types of applications.

Suitable Customers	***Suitable Applications***
Large, Information-dependent Organizations of any Type	Customer Relationship Management
The Real Time Media (Radio, TV, Web)	Background Information Searches; Trend Forecasting
Law Enforcement	Real-Time Searches for Suspect Information; Discovery of Hidden Links
Large, Heterogeneous Organizations	Exception Event Reporting by Pattern Identification

It's easy to get started:

(1) Ask InfoTame for a demonstration based on their sample data bases of documents; or

(2) Ask for a trial pilot based on your own information needs

More detailed technical information is, of course, available from InfoTame at www.infotamecorp.com.

For information on Wohl Associates, please go to http://www.wohl.com; a free subscription to Amy D. Wohl's Opinions is available at http://www.wohl.com/signup.htm.