

## Discover More

A Technical White Paper  
on the Stratify  
Discovery System™



Stratify, Inc.

August, 2001

**S**tratify is the emerging software leader in unstructured data management. The Stratify Discovery System automatically organizes information from internal and external documents into a logical hierarchy with an easy-to-use interface. By structuring previously difficult-to-organize information, the Discovery System dramatically increases the power of existing corporate applications, including document and content management systems, enterprise search solutions, Customer Relationship Management (CRM) and Sales Force Automation (SFA) tools, lead generation software, and corporate portals. As a result, businesspeople easily find the information they need to work more productively and to make better decisions.

## Section I. The Challenge: Managing Information Overload

---

Today's companies need to harness information to succeed. Using information effectively increases employee productivity, and companies that capture, manage and reuse their intellectual assets gain advantages over their competitors. But companies need to manage staggering amounts of information to stay competitive:

- In early 2001, the public Internet, the source of half of the information workers use, exceeded four billion pages and was growing at a rate of 7.3 million new pages every day.<sup>1</sup>
- An estimated eight billion email messages will flood US corporations each day in 2002, up from 3.5 billion in 1999.<sup>2</sup>
- IT departments will need to spend in excess of \$15 billion to manage information in 2001 or face productivity declines of up to 20 percent.<sup>3</sup>
- White collar workers will spend 30 to 40 percent of their time managing documents by 2003, up from twenty percent in 1997.<sup>4</sup>

Hal Varian, the dean of the School of Information Management and Systems at the University of California at Berkeley, neatly summarizes the problem: "The world's total production of unique information per year amounts to about 250 megabytes for each man, woman and child on earth. It is clear we are all drowning in a sea of information."<sup>5</sup>

Businesspeople must wade through this growing sea of information to do their jobs, but the tools they need to navigate it effectively have not kept pace with the expanding quantity of information. As a result, the information-to-noise ratio continues to deteriorate, sapping productivity and causing companies to miss opportunities to innovate and compete. To capture the value now lost to information overload, companies must help their employees manage and harness information effectively.

### Structured vs. Unstructured Information

Business-critical information comes in many forms. Unstructured content, found in documents, presentations and web pages, is often more useful in making decisions than structured content, found in relational databases. While

---

<sup>1</sup> "Sizing the Internet." Cyveillance, Inc., July 10, 2000. Mark Gilbert. "KM and Content: Push! Pull! Publish! Profit?" Gartner Group (Presentation at Symposium and ITXpo), April, 2000.

<sup>2</sup> Joan Hamilton. "Like It or Not, You've Got Mail." Business Week, October 4, 1999, quoting IDC.

<sup>3</sup> Kathy Harris. "KM Scenario: Is there Life after Hype?" Gartner Group (Presentation at Symposium and ITXpo), April, 2000.

<sup>4</sup> "Implementing an Integrated Document Management Strategy." Gartner Group, February 20, 2001.

<sup>5</sup> Alan T. Saracevic. "Quantifying the Internet." San Francisco Examiner, November 5, 2000.

structured information – e.g., revenue trends by department, production costs by location – is critical to managing existing operations, it generally provides few insights into why something is happening or how it might change in the future. Structured information can inform you that sales have been declining in United Kingdom, but learning why this is so – and how the trend can be reversed – requires text-based, unstructured information. Has a competitor introduced a new product there? Has economic turmoil affected overall demand? Has the new regional management changed a formerly effective sales practice? Structured information tells managers what has happened; unstructured information explains why and points the way to a solution. In the end, business leaders need all relevant information to make the best decisions, not just relevant structured information.

Yet most companies spend lavishly on technologies which organize and manage structured information while neglecting the 85 percent of their information found in unstructured documents. As a result, they forfeit value: insights are missed, ideas get shipwrecked, work is duplicated, decisions are made with a fraction of the information needed, and fewer new products come to market. A company's most valuable asset, the time and attention of its most skilled and highly-paid employees, is embodied in the unstructured documents these people create. Most companies still put less effort into managing this intellectual capital than tracking the PCs used to create it.

Without a way to easily find and use unstructured information, workers struggle to locate what they need despite all the money spent on technologies like as enterprise portals, document management systems, and text retrieval technology. As a result, massive IT investments produce dubious results: an EVP at one investment bank described its intranet as "an information landfill." A managing director at another worried whether people would be able to meaningfully access the 600,000 documents in its document management system.

### Today's Tools

Existing technologies which attempt to organize unstructured information are problematic. Keyword search has real limitations, as anyone who has sorted through a lengthy list of search results can attest. Other applications require users to classify and submit their work, set up and maintain agents, and query and read discussion groups when they need advice or assistance, forcing users into awkward and time-consuming behaviors. The unnatural amount of process and activity required to organize and share information bogs these solutions down and reduces their effectiveness.

While businesspeople can now access vast quantities of data, facts, and statistics, they cannot easily create practical knowledge from it because they cannot easily find the unstructured information they need. Despite a flood of information, they still thirst for knowledge. "Unstructured" should not imply "unknown," "unrecognized," or "unused." Corporations clearly need a scalable way to organize and manage unstructured information from a variety of sources, one which can deliver it to the users who need it the most.

## Section II. The Solution: Automatic, Customized Organization and Delivery of Unstructured Information

---

Companies need a way to turn unstructured information into competitive advantage, to capture the value now lost to missed insights and poor decisions. For maximum benefit, such information should appear in the right context, based upon the current activities or established habits of users. Ideally, software should anticipate users' needs for unstructured information and present it to them as they view other structured and unstructured information.

For this reason, incorporating unstructured information into applications like CRM and ERP systems, which now primarily handle structured information, would greatly increase their power and effectiveness. Such applications can manipulate structured information only because they rest on relational databases that organize and manage it. What these applications – and the enterprises that use them – need is software that would fulfill the same function for unstructured information.

With such software as a foundation, a variety of applications could use unstructured information more effectively, including:

- Search applications (e.g., Inktomi);
- Enterprise portals (e.g., Plumtree);
- Document management systems (e.g., Documentum);
- Content management systems (e.g., Interwoven);
- CRM applications (e.g., Siebel); and
- ERP applications (e.g., SAP).

Database design suggests some guiding principles for any software that would organize and manage unstructured information. Just as a schema lends structure to the contents of a database, an information hierarchy – a manipulable directory of topics – lends structure to unstructured information. Standard interfaces allow the programs which rest upon a relational database to remain independent of its logical structure, and such interfaces will allow programs which access unstructured information to remain independent of the underlying information hierarchy as well.

But unstructured information differs from structured information in many ways, and any software that attempts to manage it will have to accommodate these differences. In particular, unstructured information is found in a wide variety of formats – MS Office documents, PDF files, HTML – and in a variety of locations – the Internet, file servers, document management systems, groupware. Moving or copying documents themselves would cause massive confusion and waste storage space, so any organizational system will need to build metadata – information about information – and manage access to the underlying documents through it. Finally, the information hierarchy underlying the system will have to adapt over time to changes in the content within it in ways that minimize human labor but allow human oversight.

In short, software that organizes and manages unstructured information must do the following:

- Allow applications and users full or permission-based access to all important information, whether internal and external, in whatever location and of whatever format;
- Build an organizing framework – an information hierarchy – tailored to its contents;
- Create rich and consistent metadata by accurately classifying content and applying business rules;
- Expose rich APIs and employ industry-standard interfaces so applications can easily access and manipulate this metadata;
- Learn adaptively and automatically, but under the supervision of people; and
- Present the right information to people when they need it, based upon their current context or habitual interests.

The cornerstone of such software must be an automatic hierarchy-building and classification system that assumes the labor-intensive aspects of categorization but supports human supervision and review.

## Section III. Today's Technology for Organizing a Business' Unstructured Information

---

Vendors have applied a number of technologies to solve the problem of organizing unstructured information. This section surveys these approaches and discusses the merits of each.

### The Importance of Hierarchy Building

Building an information hierarchy is the first essential step in building a system which organizes and manages unstructured information. An information hierarchy, like a database schema, provides an organized framework that allows easy access to information. For optimal value, that framework must reflect the needs of a given enterprise. Defining an information hierarchy that does this, however, is a complicated and time-consuming process.

Unfortunately, it is also a process that must be continually repeated. Changing strategies, evolving products, and advancing technology drive changes in the type of content businesspeople need to do their jobs – and in the underlying information hierarchy. Sooner or later, the topics that are now relevant will go the way of the Y2K bug, and any system that handles unstructured content must address the evolving needs of its users for information.

Some companies have attempted to address the difficulties inherent in hierarchy building by providing prefabricated hierarchies. While these templates help jumpstart the process of organizing unstructured content, they often cannot be customized. Prefabricated hierarchies are a half measure in any case: an ideal hierarchy is tailored specifically for a given enterprise.

A more promising approach – and one which allows for the maintenance as well as the initial construction of a customized hierarchy – is to rely on the structure inherent in the documents themselves. The documents of a given enterprise inherently reflect the subjects and ideas that matter to that enterprise. Software that detects those subjects and reveals the relationships among them can build a hierarchy and maintain it over time.

### Approaches to Classification

A well-defined hierarchy is only the first step toward managing unstructured content. Classifying information into one or more topics within a hierarchy enables end users to find the information they need easily. But classifying information is a difficult problem because unstructured content has implicit meaning that people interpret in different ways, depending on the current context and their individual interests. Today, companies use many different classification methods that vary in the degree of automation provided and in the ability of people to control the outcome of the process. As the following diagram shows, companies classify documents using either manual or automatic means. Automatic technology can be further divided into rule-based systems, supervised learning systems and unsupervised learning systems.

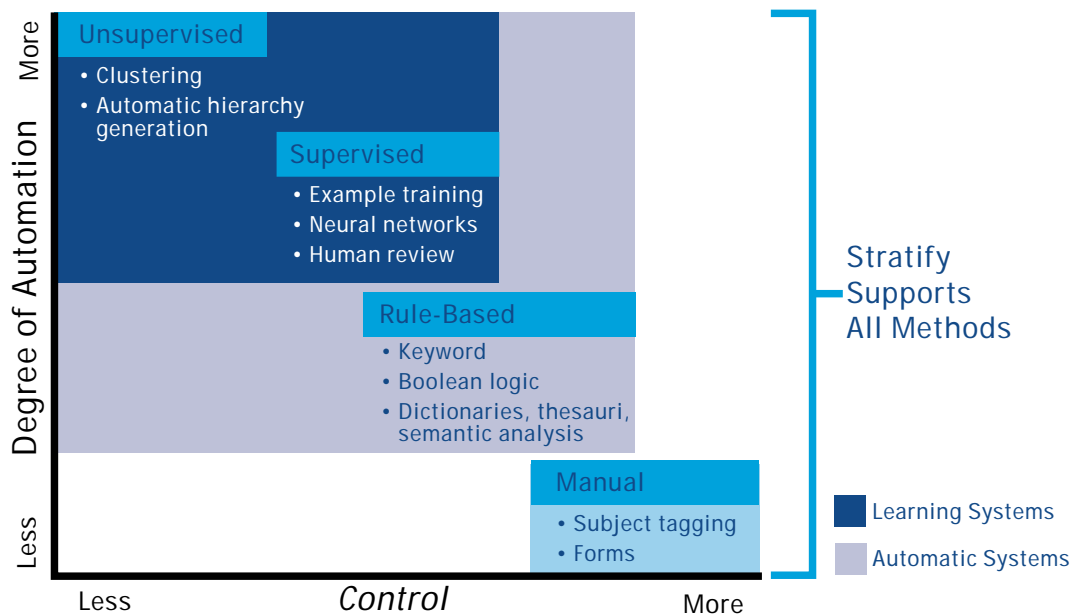


Figure 1: The Spectrum of Classification Options

All methods fall along a diagonal, illustrating the trade-off between automation and control. At the bottom right, manual classification offers the greatest amount of control but offers little automation; at the top left, unsupervised learning offers the greatest automation with little control. Each method works well sometimes, but none work well all the time. The ideal classification system would use each method when it is appropriate.

### Manual Classification

Manual classification requires information managers to review each document, classify it, and assign it to a particular node in the corporate information hierarchy. At small volumes, it is highly accurate because people are better than machines at understanding the meaning in documents. Yahoo! and the Open Directory Project have shown that people can manually classify large amounts of information, but most corporations cannot afford this approach. In practice, manual tagging does not scale well: it is labor-intensive and therefore expensive to set up and maintain. Worse, it is subject to inconsistency because no one person can understand large numbers of subjects well and no group of people can classify documents in a uniform way. Only automated methods can solve the classification problems of a large enterprise.

### Automatic Classification

Automatic classification technology can be divided into two main types of systems: rule-based systems that use human-provided words, operators, criteria or patterns to evaluate whether or not a document belongs to a topic and example-training systems that use example documents to help the system learn how to organize information. An example-training system automatically constructs models of topics from the example documents and uses those models to classify other documents. Depending on the design of the application and tools provided, an example-training system can provide a lot or a little of human supervision over the assembly of training sets.

## Rule-Based Systems

Rule-based systems are popular because users of these systems can precisely define the criteria by which a document is classified. Rule-based systems can support complex operations and decision trees and produce very accurate results. The main drawback of this approach is that, for optimal results, well-trained people with domain expertise must write the rules.

Keyword-based systems, usually based on a search engine, are the most common rule-based technology that companies use to automate information organization and retrieval. At its simplest, the search “rule” is a single word the search engine uses to query an index and find all documents in which the word appears. Keyword-based rules work well when there is little ambiguity in the terms used in the rules and when the volume of documents is relatively small. They are of little use when the user is unsure of the exact spelling of the search term or unfamiliar with any relevant jargon.

When used alone, simple keyword systems lack context and fail to differentiate between words that can have multiple or ambiguous meanings. Because most search engines favor recall over precision, they bring back many results and judge relevance by word frequency, not by what’s important to the user.<sup>6</sup> A query on “penguin” can bring back articles about Antarctic birds, the Pittsburgh hockey team, Batman’s nemesis, and the logo of an open source operating system, leaving the user to pick through clutter to find what they need.

When used with Boolean algebra, refined keyword rules can match documents to categories precisely. But most users cannot form queries that describe the type of information they want. For example, users who enter “Dell Computer” may want specific information on a wide variety of topics related to Dell – e.g., corporate financial information, product reviews, a case study on manufacturing to order. These topics are difficult to define with words and Boolean operators and require multiple iterations to produce the desired results.

As the volume of content grows, keyword search returns more matches than users can reasonably sort through. Unable to use Boolean operations to build a fruitful search rule, many users become frustrated with the “needle in a haystack” results search engines produce. This frustration has led to an increase in popularity of “learning” technologies that seek to eliminate ambiguity through the use of statistical or pattern-matching techniques.

## Learning Systems

Learning systems require the user to assemble a set of reference documents pertaining to a particular topic and then use these documents to build a reference model describing the unique properties of the topic at hand. New documents are analyzed and compared to a set of these models, using a statistical or pattern-matching algorithm, to find the closest match.

The training sets used to assemble these models can be collected in a supervised or unsupervised manner. When creating the training set in a “supervised” manner, users select example documents they feel are good representations of the subject matter. In “unsupervised” learning, the system takes a set of documents and clusters it into groups based on word patterns and other statistical similarities, thus assembling multiple training sets from one operation. Both techniques rely on software to examine the content of the documents in the system and make educated guesses about how to organize them.

---

<sup>6</sup> Search engines that favor recall bring back every possible matching document, trying to ensure that potential matches are not missed. Search engines that favor precision will return only those documents that are highly relevant or accurate matches to the query, striving for accuracy and quality instead of quantity.

**Supervised Learning Systems:** A supervised system allows the user to control how the training sets are built and to tune the classification process. Some commercial implementations allow humans to supervise the process of assembling training sets or of the actual classification of documents, but not both. This “black box” approach limits the control users have over the ultimate classification. Stratify technology allows supervision at many different points in the modeling and classification process to produce accurate results tuned to a company’s business needs.

The main drawback of supervised learning systems is that they require people to define training sets for the reference model. Revising models once they are put into production can be difficult because it is not always intuitive how to alter the training sets to improve the model. Stratify provides robust tools (described below) that diagnose problems and allow users to tune the system in multiple ways. Lacking tools like these, many supervised systems become black boxes, behaving arbitrarily and denying control to the user.

**Unsupervised Learning Systems:** Unsupervised systems can bootstrap themselves into operation without human intervention. The major drawback to unsupervised methods is they have difficulty determining whether their results are useful because they lack predetermined benchmarks of accuracy. As a result, unsupervised systems can sometimes focus on features insignificant to humans while neglecting the main ideas within documents.

For these reasons, unsupervised methods like clustering are best used to reduce initial labor, calling on human help to refine their results. Stratify uses them in exactly this way. Starting with a sample of documents from a particular enterprise, Stratify’s patent-pending clustering algorithms analyze the occurrence and density of words within documents to group them into natural clusters and to build a hierarchy linking these clusters together. In this way, it exposes the natural structure hidden within a seeming random collection of documents. A human administrator can then rearrange, expand, or collapse nodes in the resulting structure, using an intuitive visual interface, and label them as desired. The documents in the clusters serve as initial training sets, with model-building software identifying the statistical patterns hidden within the documents. Human administrators can adjust these training sets as desired. The end result is an information hierarchy and a set of topic models fully customized for the needs of a given enterprise.

## Statistical and Pattern-Matching Approaches to Classification

Once the hierarchy and reference models are established, a number of classification methods can determine whether or not a document belongs to a particular category. A summary of the benefits and drawbacks of the most commonly used ones follows:

### Bayesian Probability

The most popular and flexible method to classify documents is Bayesian probabilistic modeling. Bayesian algorithms build statistical models from the words within training sets. When classifying, a Bayesian system compares an individual document to an individual topic model and assigns a probability that the document belongs to that topic. As this process is repeated for all models, the system can assign documents into multiple topics when justified. The statistical nature of classification allows Bayesian approaches to see beyond individual words to the underlying patterns within the documents: a document about hiring policy, for example, could be classified under “human resources” although neither the word “human” nor the word “resources” appears in it. The principles behind the algorithm are easy to understand, and it is easy to teach administrators how to train a Bayesian system to produce good results. As an additional advantage, Bayesian methods are computationally efficient, as they use only the topic models, not all documents in the training set, in the classification process.

In developing the Discovery System, Stratify researchers found that Bayesian methods effectively classify large numbers of documents into multiple topics. Discovery Systems use Bayesian algorithms to derive topic models from sample documents and match content to these models.

### Support Vector Machines

Support Vector Machines (SVMs) calculate the maximum “separation,” in multiple dimensions, between similar groups of documents. While Bayesian systems weigh all training set documents equally, SVMs pay more attention to outlying training documents. Given high quality training sets, SVMs will focus on the crucial documents which help define the borders of the group. With poor quality training sets, however, they tend to focus on erroneous outliers, and their performance suffers markedly. SVMs can effectively decide whether a document belongs to a single topic, but tend to be slow and expensive when used to classify a document into a large number of topics. As John Platt, previously the lead researcher on SVMs at Microsoft, states, “We believe constructing N-class SVMs [which classify a document into more than one topic] is still an unsolved research problem.”<sup>7</sup>

### Neural Networks

Neural networks are computational frameworks composed of nodes which receive an input and compute an output or response. When classifying text, the nodes take training sets as inputs and calculate the topics inferred from these words as the output. Most often, neural networks are used to determine the relationship of similar topics to each other and to allow users to navigate between them. The main drawback to neural networks is that they are “black boxes” – each new training document can affect classifications in unintended ways which cannot easily be explained even in retrospect.

### K-Nearest Neighbors

Another popular statistical method is K-Nearest Neighbors. This algorithm classifies a document by first finding the K “nearest” documents among all the training documents for all categories, based on commonality of words between the document to be classified and each training document. It then returns the most common category among the neighboring documents as the classification. This method produces good results when there are enough training documents to span the breadth of each category. But it does not scale well as the number of training documents and categories increases because the classifier must access each and every training document, making classification slow and memory-intensive. Some commercial systems attempt to solve this problem by summarizing all the training set documents, but this summarization inevitably introduces other errors.

## Combining Multiple Approaches Produces the Best Results

Each of these technologies can produce impressive results. However, machine learning is still more of an art than a science, and users need to understand the circumstances under which each of these algorithms will work well. After reviewing a number of recent research reports on supervised learning algorithms, Andrew McCallum, a leading researcher in text classification, remarked, “Among [Naive Bayes, K-Nearest Neighbors, SVMs, rule-learning algorithms, and others], no single technique has proven to consistently outperform the others across many domains.”<sup>8</sup> Given this, the most promising approach to the problem of classifying unstructured information is to employ multiple classification techniques, using each when most appropriate.

---

<sup>7</sup> J. Platt, N. Cristianini and J. Shawe-Taylor. “Large Margin DAGS for Multiclass Classification.” Advances in Neural Information Processing Systems, 12 ed. S.A. Solla, T.K. Leen and K.-R. Muller, MIT Press, 2000.

## A Note on Semantics and Linguistics

Semantic or linguistic technologies aid document classification by analyzing the meaning of words using custom dictionaries, thesauri, word stemmers, part-of-speech analyzers and/or noun-phrase identifiers. These tools are difficult to develop and maintain, as each language (and dialect) requires its own set of rules. Linguistic technology operates at the level of words and sentences rather than paragraphs and documents. These methods are best used to support other automatic classification methods by helping to make the meaning of words clear.

## Section IV. Improving Business Decisions with Unstructured Information

---

Because machine learning is not yet sophisticated enough to support truly automatic classification, the best classification systems use a combination of manual, rule-based, supervised and unsupervised techniques. Humans provide insight, review, correction and tuning – what people do best – and machines handle the repetitive, labor-intensive activities. The best systems also use a variety of methods to analyze content to make intelligent connections on behalf of users. Using these capabilities as a foundation, a variety of applications can incorporate unstructured information and extend their own capabilities from simple data management to producing valuable insights.

### The Stratify Discovery System

The Stratify Discovery System, uses all four of the major classification methods. The software uses unsupervised automatic classification techniques to build an initial information hierarchy and applies a combination of supervised and unsupervised techniques to understand the main ideas in text and organize them into this hierarchy. User-defined rules enhance classification, and people can supervise classification and adjust the system manually when topics change or new categories are needed. By using each classification technology where appropriate – and constantly allowing for human oversight – the Discovery System has an unparalleled ability to put structure around unstructured information and help businesspeople make the right decisions.

Three elements make a Stratify Discovery System unique:

#### 1) Hierarchy Building and Classification Tools Ease Set Up and Maintenance

The Discovery System is the first software to automate the difficult process of hierarchy building and to simplify hierarchy maintenance. As discussed above, the software can automatically create a customized hierarchy from a sample of documents, and allow users to edit the hierarchy as desired. A set of easy-to-use tools enable administrators to monitor the classification system, to tune it for optimum performance, and to alter it to fit new business needs. The tools include diagnostic features that point out potential problems and help identify changes that will resolve them.

#### 2) Modular Classification Yields Maximum Accuracy and Control

The Discovery System uses multiple classification methodologies – manual, rule-based, supervised and unsupervised learning – that operate in concert. At the core of the solution, a flexible classification engine uses only the

---

<sup>8</sup> Kamal Nigam, John Lafferty, and Andrew McCallum. "Using maximum entropy for text classification." In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61 - 67, 1999. Emphasis added.

methods and rules most appropriate to each company's business-critical information. The Classifier, a key component of the Discovery System, harnesses human feedback to improve its accuracy and to adapt to changing business needs.

### 3) Customized Presentation Delivers the Right Information to Users

The Discovery System presents information to users based on their current context and on their habitual interests. The precision of the classification technology allows the system to understand what a user is viewing now and suggest related documents as they become available. Over time, a profile of the documents users access becomes an accurate proxy for their interests, allowing the system to anticipate their current needs by drawing upon their past behavior. These capabilities become even more powerful when embedded within the familiar interfaces of other enterprise applications that use structured information.

## How Does it Work?

The basic building blocks of a Discovery System include modules that collect information, classify it, aggregate metadata about it, and allow a variety of applications to access it. Supervised learning, monitoring and administration tools help users build and maintain the information hierarchy and topic models that power the system.

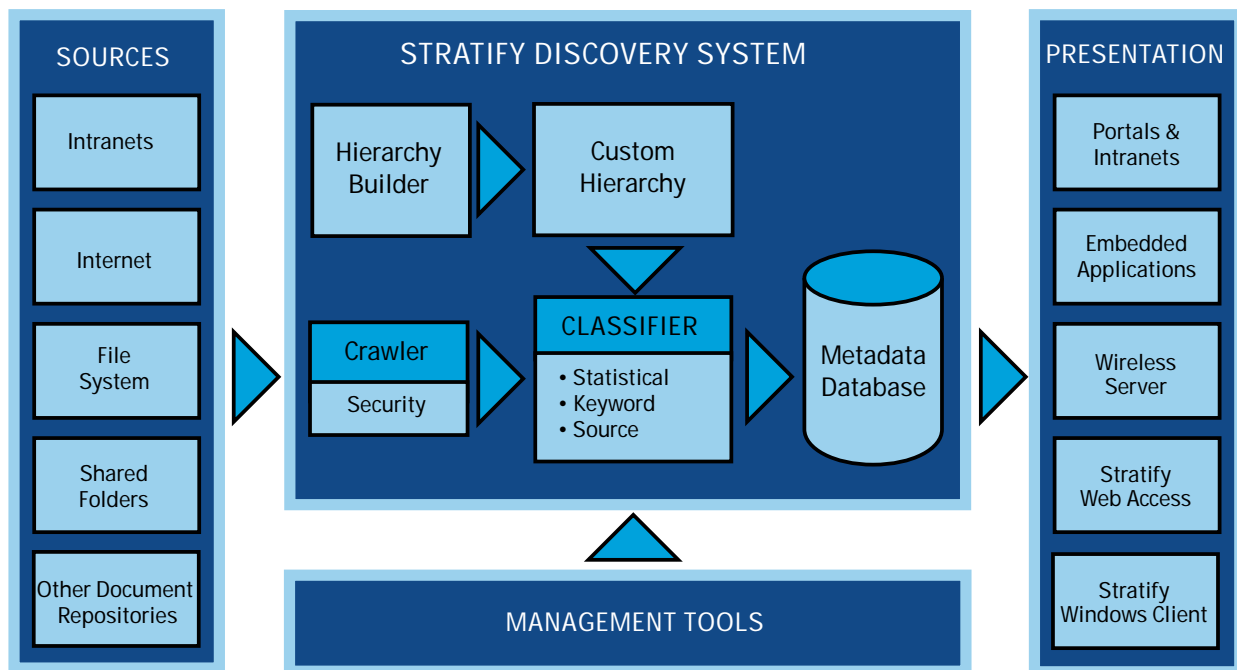


Figure 2. Stratify Discovery System Architecture

### Jumpstarting the System: Creating the Hierarchy and Training Sets

Starting with a sample of documents, the Discovery System's Hierarchy Builder module can create an initial information hierarchy automatically. Depending on the size of the document corpus, either all of the documents or a statistically significant sample thereof are used in the clustering process. The clustering technology works rapidly - a single server can process up to 100,000 pages in less than a day.

Once the documents are collected, the Discovery System uses patent-pending advances in machine learning and information theory to recognize patterns and organize the documents into a hierarchy.

The system first scans all the documents and divides them into groups of similar documents. It clusters documents together only if they are broadly similar, not just if they share a few words in common. Outlying documents that don't fit well with any others are set aside instead of being forced into a group. On the other hand, a document that is similar to more than one group can be included in multiple groups.

This bottom-up approach to hierarchy building, which aggregates individual documents together into clusters, provides better results than a top-down approach, which divides the entire document corpus more and more finely. The latter approach must first assign each document to one of its top-level divisions, which are necessarily large and somewhat arbitrary. Top-down approaches are vulnerable to initial errors in sorting documents, and every additional stage of subdivision adds another chance to misplace a document. Stratify's bottom-up approach, on the other hand, compares documents directly to each other instead of to a top-level category, and continuously adjusts the composition of clusters, correcting errors as it proceeds.

Once documents are aggregated into clusters, the Hierarchy Builder organizes the clusters into a hierarchy. Parent nodes in the hierarchy represent general topics; children represent more specific ones. The system labels each cluster with the most significant words within it, allowing administrators to easily identify the main subject of each cluster. Administrators can manipulate the hierarchy, adding, deleting, merging, and splitting nodes as desired. This naturally leads to a hierarchy which accurately captures the underlying structure in the document corpus.

The resulting clusters of documents, arranged in a hierarchy, can serve as training sets for the classification system. Because of a tight integration with the classification technology, the documents of each cluster tend to make good training sets, allowing accurate and reliable classification of new content.

### Powerful Tools to Manage Classification

The Discovery System supports a powerful and flexible set of tools to help users monitor, modify and manage the classification system. The System Tools use Stratify classification technology to help locate and organize content into training sets. They classify every page that the tool user browses in real time and calculate a relevance score which describes the document's degree of fit with a given training set. With this information, the tool user can easily decide whether a document belongs in a training set.

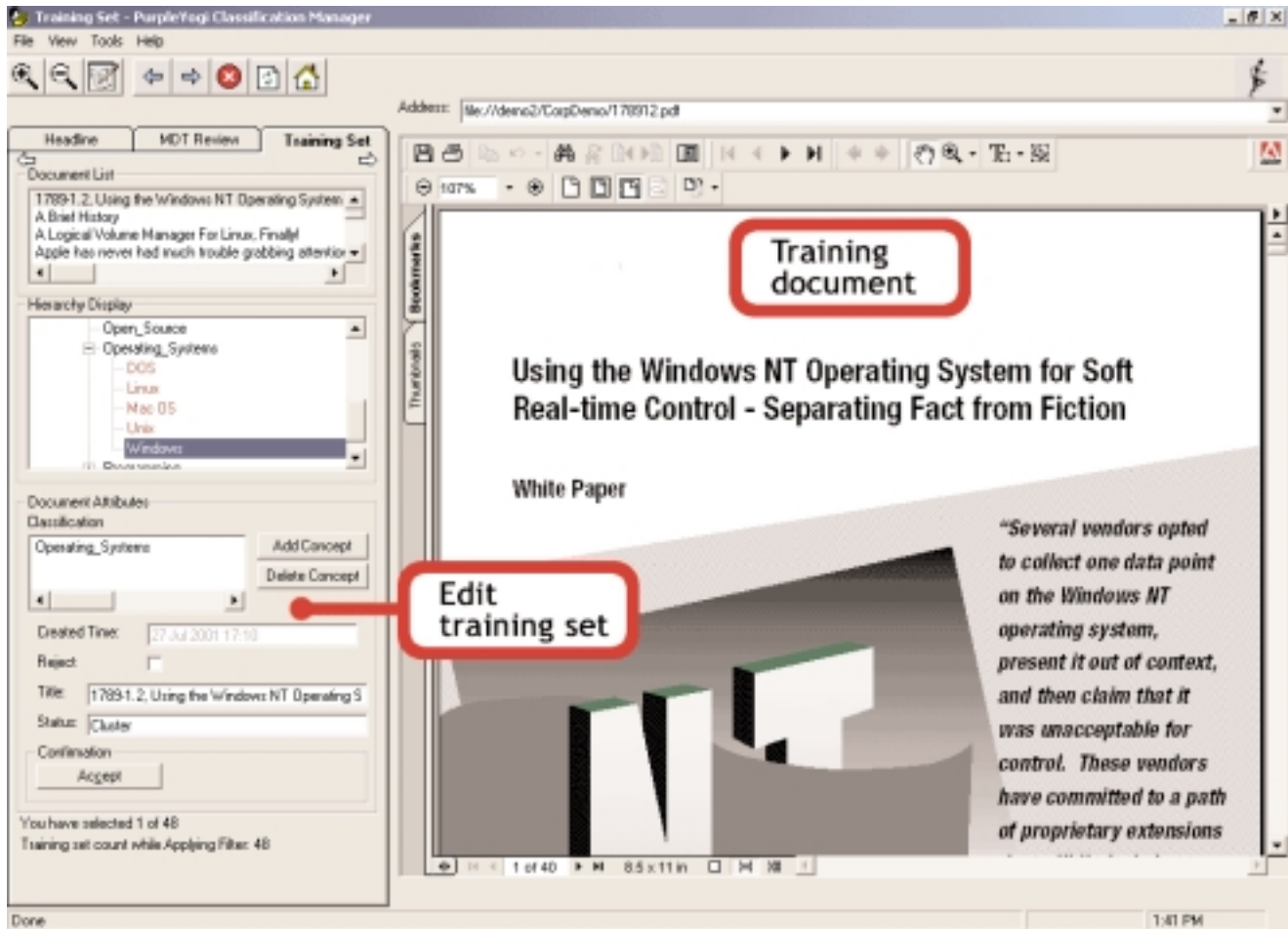


Figure 3: System Tools allow administrators to edit training sets and configure models.

As illustrated in Figure 3, the System Tools allow users to edit classification models in a variety of ways. Users can easily add documents to (or remove documents from) training sets and test the results of these changes in real time. The tools notify the user if a change made to a model or the hierarchy is inconsistent or duplicated in another branch of the hierarchy. The system will show how other sections of the hierarchy are affected by inconsistent manual changes and then ask the tool user if they want to proceed with the change.

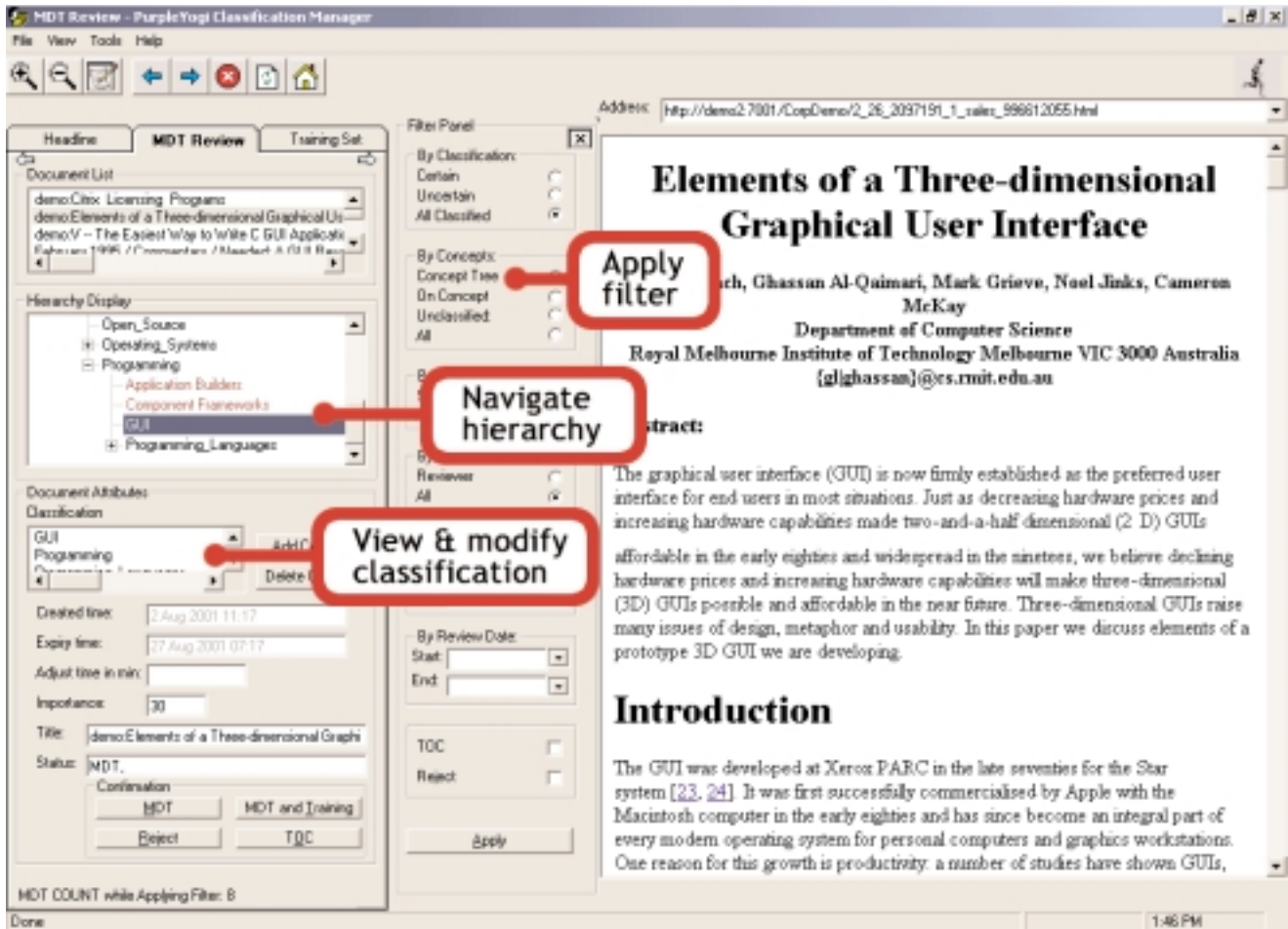


Figure 4: Users can easily browse and edit the entire hierarchy, set a topic's relevance, importance and timeliness, and set expiration policies on documents.

From the interface shown in Figure 4, users can inspect documents in the system and modify their classifications. They can find the documents they want by browsing the information hierarchy or by applying filters on a variety of topics. Users can also set business rules which tell the system that some documents or sources are more important than others and specify when to remove particular documents or sources from the system. The System Tools are designed to allow multiple users to interact with the system at the same time without sacrificing reliability or security.

## Refining the Hierarchy and Statistical Models

As discussed earlier, statistical classification methods do not work equally well in all cases. Automatic methods can produce unintended results and concept meanings can become fuzzy or indistinct over time. To overcome these challenges, the System Tools allow users to manually annotate the individual metadata descriptions of the documents, to change classification models, and to modify the hierarchy.

To refine classification accuracy, the System Tools allow the user to add keyword terms to the topic models. Keywords allow new content to be classified into very narrowly-defined categories with maximum precision. When working in concert with the statistical classification, keywords are very effective. Because the system bounds

the keyword-rule evaluation to the higher-level topics automatically matched by the classifier, ambiguous meaning in the keywords is eliminated.

The clustering technology can monitor the document corpus on an ongoing basis and use new clusters to suggest new topics to add to the hierarchy. To do this, the user would compare the outcome of a new clustering run to the existing hierarchy and find any new clusters. By clustering unclassified documents, the system can help users discover new topics that could be of value to their business.

## Collecting the Right Content

To begin the classification process, the user specifies the file servers, Web sites, table-of-contents pages and other sources of information that the system should crawl. Discovery System classification is unobtrusive; it doesn't change or move any of the original content. Instead, it produces a streamlined description of the important features and topics embedded in the documents. This metadata description uniquely identifies the document and makes it easy to match it with other important content and topic areas.

The Discovery System includes sophisticated programs controlled by flexible administrative tools, called Crawlers, which continuously add documents to the system. Crawlers parse and filter pages to identify which the text should be presented to the Classifier, eliminating links and mark-up that could obscure the meaning in the document. They also extract and store the document's title, location, source, and date. Using patent-pending text mining and pattern matching algorithms, the Crawlers detect and eliminate duplicate documents.

The Crawlers ensure that all of the content in the system is fresh, accessible, and relevant. They support a variety of metadata expiration policies so administrators can control how long the system retains document data. Crawlers also detect when a document has been removed from its original source and expire the metadata automatically, making sure users see only the freshest, most relevant information when using a Discovery System.

Through the Crawl Manager interface, administrators can test repositories and Web sites they want to crawl before adding the documents within them to the system. This keeps users from experiencing "page not found" error messages or similar crawl failures. Instead, they see accurate and relevant results.

## The Classification Process: Parallel Deployment of Multiple Technologies

Using patent-pending advances in information theory, artificial intelligence, and machine learning, the Classifier analyzes the text the Crawler extracts for it. As shown in Figure 5, the Classifier architecture is modular, applying multiple classification methods to each document. The Combiner module within the Classifier aggregates the results of each method and chooses the best classifications, based on its knowledge of the strengths of each classifier and the degree of certainty each classifier has about a particular document. By using multiple complementary techniques, the Classifier achieves much greater accuracy than single-method classifiers. It has achieved precision and recall rates of more than ninety percent on certain document corpuses.

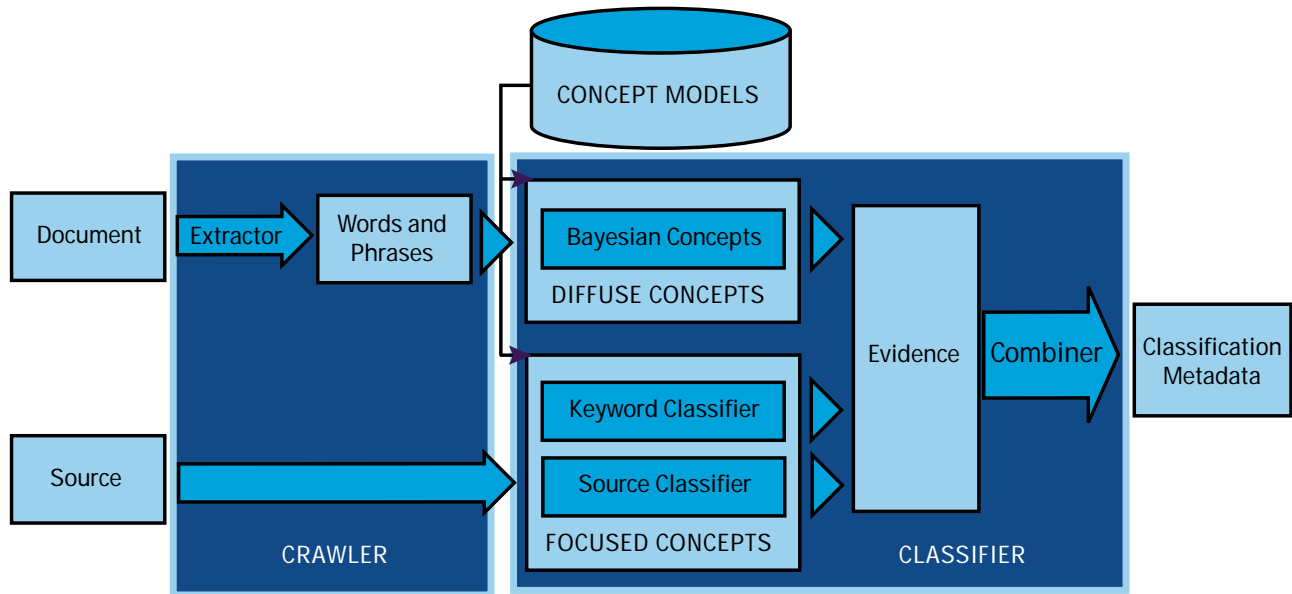


Figure 5: The Classification Process

Depending on format, length, and style, different content will be classified more accurately by one method than another. For example, the Bayesian statistical classifier works well when classifying documents into broad topics, say, different industries like “Airlines,” “Biotechnology,” and “Insurance.” However, documents about a particular industry can contain more than one company name and make it difficult to determine which company is the main subject of the document. To easily determine whether a document is about one company or another, the Discovery System uses rules that capture other specific information about the company, such as variations in the company name, unique product names, and the names of key management team members. Because the Bayesian classifier has already identified a broad context for the document, any ambiguities in the meaning of keywords are resolved.

The system can also pre-process content and direct it to the sections of the hierarchy where it is most likely to classify with a high probability of relevance. For example, when the crawlers return a URL that contains a common stock ticker symbol, the system immediately routes these Web pages to the corresponding industry or company topic, bypassing any statistical or keyword processing.

A modular architecture and rich set of APIs make it easy to integrate the Classifier into a variety of enterprise applications. Stratify itself supports a variety of deployments of the Classifier, from the compact footprint of a client-side plug-in to a distributed configuration hosted on network servers.

## Standard Interfaces Make Metadata Accessible

As content is classified, the Discovery System creates and maintains a compact summary of every classification. This summary, which is called a Metadata Database entry, includes the original location of each document (file system directory path, web page URL etc.), the document title, topic(s) matched, key words and other information such as the time the document entered the system, its original source and expiration date. This metadata allows access to any document by a variety of criteria.

Unlike other classification systems, the Discovery System uses standard interfaces so other applications can easily access the metadata it generates. A SQL database, provided by any common vendor, serves as the foundation of the Metadata Database, and other applications can interface with it using standard queries. Customers can add customized fields to the database and specify when and how metadata updates should be transmitted to applications. Since most enterprise applications already rest on SQL-capable relational databases, integration with a Discovery System is painless. As a final aid to integration, the Discovery System formats metadata in XML.

## Related Article Matching and User Interest Profiling

Applications that draw on a Discovery System use metadata entries to match the main topics in pages to other pages that contain similar or related topics. If business needs require, the matching process can take factors like date, source, title, location and degree of similarity into account. The power and flexibility of the classification system give administrators the ability to specify what is relevant and accurately deliver the right information to users.

The end result is a set of links to documents organized by subject area. This consistent, global view of relevant information from a variety of sources allows users to find useful information quickly. By browsing through a structured tree of topics, users can access not only the most similar documents, but also those in related subject areas that the user might not have connected to the original document. The advantage over traditional keyword search is clear: users can easily discover useful information without sifting through pages of search results or endlessly revising their search queries.

The Stratify Windows Client can provide users with information related to their current work within a number of commonly used desktop applications, such as Microsoft Office applications, HTML pages, and Adobe PDF. As a user reads or writes within the active window on the desktop, the Client classifies its contents and serves up links to related documents. In this way, useful information reaches users automatically as they work, without them having to search for it.

In addition, the Windows Client can learn about user interests and automatically build and refine user profiles over time. By observing the content and pages users access, the Windows Client technology builds a unique profile for every registered user, reflecting the primary topic areas they find personally relevant. The profile changes with a user's interests change, automatically adding new topics and removing unused ones. Users can explicitly add or remove topics from their profile as well. In a window personalized for the user, the Windows Client recommends and publishes new or related content as matching links. As new information is published on the network, the Discovery System automatically delivers it to users with matching profiles.

The Discovery System integrates unstructured information into the workflow of their users. Whether they are searching for information, reading documents composed by other people, or writing text of their own, businesspeople can take advantage of the system's suggestions without leaving their current context. And, as other enterprise software providers integrate the Discovery System into their own offerings, users will have even more ways to access unstructured information within familiar interfaces.

# Section V. Case Study

An example will help illustrate how a Discovery System differs from other automatic or rule-based systems. In Figures 6 and 7, we compare the related articles recommended on a local entertainment and news destination site in Northern California to the recommendations given by a Stratify application. Figure 6 features an article about declining sales in the semiconductor industry. A related article panel powered by another classification solution suggests a flat list of articles, one about how the California electricity crisis is affecting chip manufacturers, another about small display screens for Web content, and a third about nanotechnology. In this case, the first article is only tangentially related to the story, and the latter two are entirely unrelated.

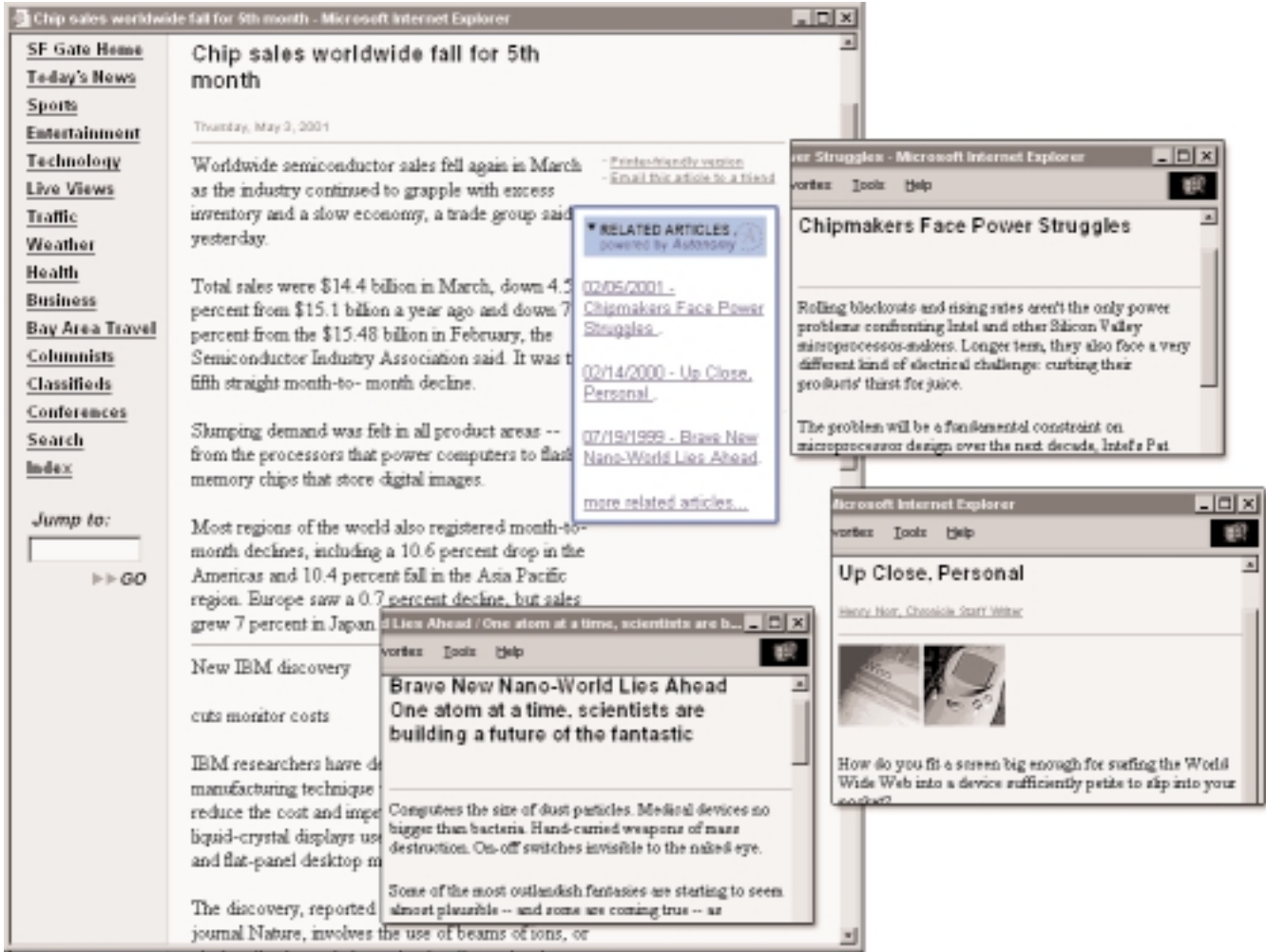


Figure 6: Another classification solution delivers a flat list of tangentially related articles.

By contrast, the Stratify application, shown in Figure 7, showed more closely related articles from the same site. In addition, it classified them under subject headings. News of declining chip sales could spur a user to read more about other bad business news under the "Layoffs & Restructurings" heading or to read more about the "Semiconductor Industry" itself. Additional related subject headings about IBM and the computer hardware industry follow. By using the structure of its information hierarchy, the Stratify application provides a richer experience for users, allowing them to browse other subjects and discover interesting articles they might not have found

otherwise. None of the related articles that the Stratify system found were featured as related articles by the classification system used on the destination site.

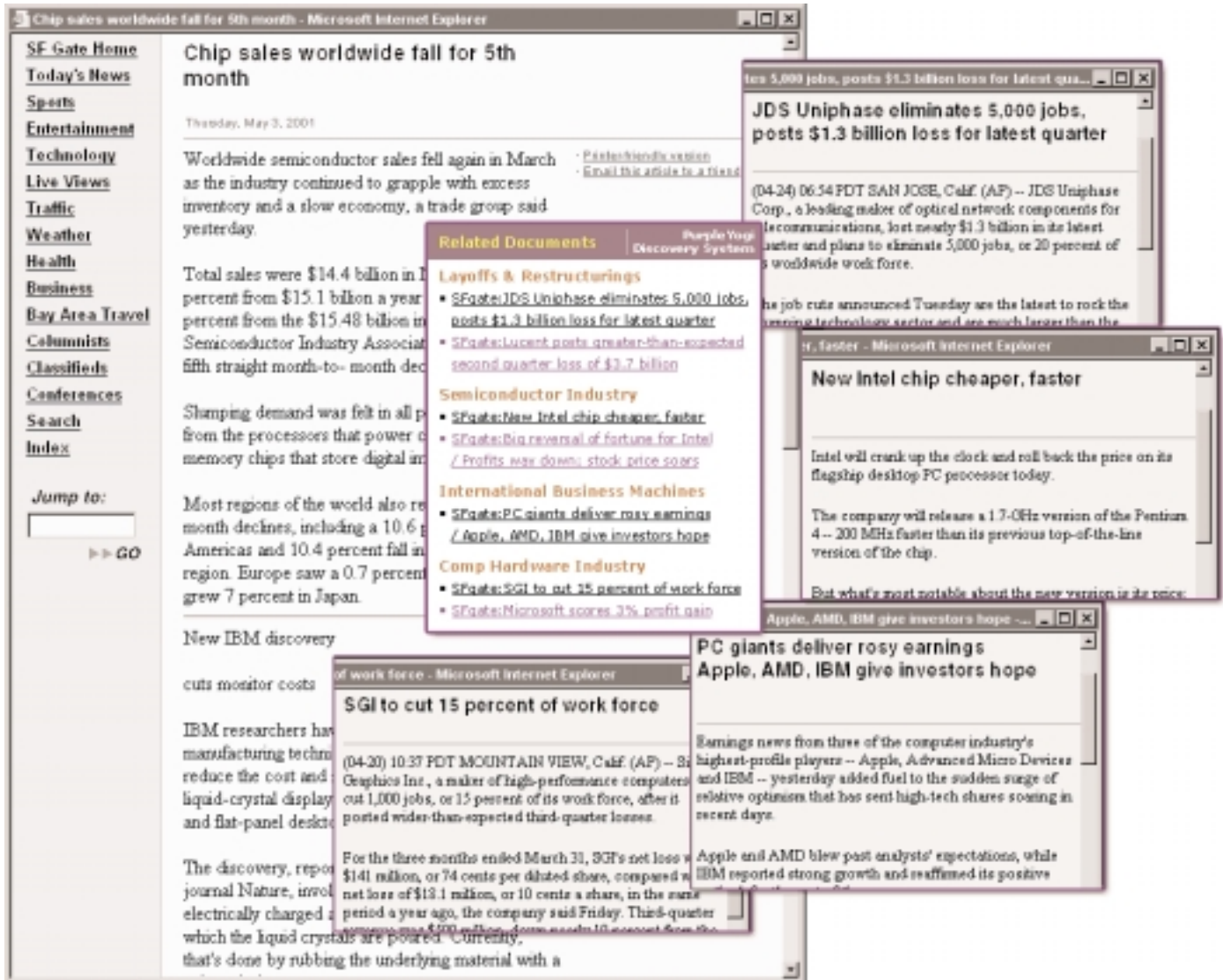


Figure 7: The Stratify application found more, and more closely related articles, from the same site. By organizing these articles under subject headings, the application allows users to explore related topics.

## Section VI. Better Information Leads to More Informed Decisions

Accurate, automatic classification of unstructured information allows businesspeople to work smarter and to make better decisions. Today, a host of applications allow them to access structured information to manage their businesses. These same applications – and the people they serve – would benefit greatly from integrating unstructured information with these applications. The resulting blend of structured and unstructured information would help them explain, instead of just perceive, what is happening to their business and decide what to do next.

**T**he Stratify Discovery System combines the best of human and machine methods to overcome the challenges companies face in organizing and managing information. When the right information automatically finds the people who need it the most, executives can know more about their customers, their products, and their competitors and make decisions that will increase profits and create enduring value. Instead of drowning in a swelling flood of unorganized information, workers can drink deep of the knowledge they need to do their jobs. The Stratify Discovery System gives companies the power to harness unstructured information for competitive advantage, producing better-informed businesspeople and better business.



Stratify, Inc. • 501 Ellis Street • Mountain View, CA 94043 • USA  
Phone: 650-988-2000 • Fax : 650-988-2159 • [www.stratify.com](http://www.stratify.com)

---

Stratify, Stratify Discovery System, and "Discover More" are trademarks of Stratify, Inc.

All other company and product names mentioned are used for identification purposes only and may be trademarks of their respective owners.

© 2001 Stratify, Inc. All Rights Reserved.

---