

Solving information overload

By Nimish Mehta, President and CEO, PurpleYogi, Inc.



Nimish Mehta
President and CEO

Nimish Mehta, President and CEO of PurpleYogi, joined the company in March, 2001, with more than 20 years of experience in executive management of global software businesses. Mehta most recently served as president and CEO of Impresse Corp., an enterprise software developer and provider of collaborative marketing solutions. Previously Mehta served as senior VP of the Industry & Front Office Applications Division of Oracle Corporation, where he oversaw all product development and marketing for the front office and vertical markets. His proven track record in scaling software businesses and managing operations will be a key factor in PurpleYogi's continued momentum in the marketplace.

Business people now spend half their time looking for information. And finding it will not get any easier—the volume of corporate data doubles each year while the public Web grows by over seven million pages a day. To tame this flood, companies spend billions of dollars on software designed to give their employees better access to the information they need. Such investments in enterprise portals, document management systems, and text retrieval technology give companies better ways to present information.

But such solutions are only as good as the organization of the content within them. Eighty-five percent of corporate information and an even higher percentage of public web content is unstructured, and this information is difficult to organize. Keyword search has real limitations, as anyone who has sorted through a lengthy list of search results can attest. As volume increases, manual tagging of unstructured information quickly breaks down due to its expense and inconsistency.

Without structure around information, workers struggle to locate what they need despite all the money spent to help them. An

EVP at one investment bank described its intranet as “an information landfill.” A managing director at another worried about whether “people will be able to meaningfully access” the 600,000 documents in its document management system.

A Crucial but Underserved Business Need

Only automating the process of organizing corporate information will allow business people to find what they need. Intranets, content management systems, and search engines work much more effectively when they rest on a foundation of structured information. Building such a foundation, based on a consistent and customized organizational framework, helps businesses recoup their investments in these technologies faster.

Automated classification involves three tasks:

- ◆ Building an customized hierarchy for information;
- ◆ Classifying documents quickly and accurately into this information hierarchy; and
- ◆ Presenting documents based on these classifications as users need them.

An ideal solution would automate all three tasks while allowing human judgment to guide the process when appropriate.

First-generation classification solutions have failed to address the first, and most expensive, task: building an information hierarchy. As a result, their customers must depend entirely on human labor to identify the concepts important to their business, organize them into a framework, and flesh out that framework by assembling training documents or building classification rules. These costs are significant: eight employees of a syndication company spent four months building a modest hierarchy of 400 concepts within one common classification solution. Software purchased to save employee time can itself consume thousands of employee hours.

What's worse, these systems either entirely exclude human input from the classification process—as statistical classification systems do—or force every single detail of classification to be painstakingly designed by human beings—as rule-based systems do. In the first case, administrators struggle to understand their expensive “black boxes,” unable to control their behavior. In the latter, administrators suffer under the heavy burden placed on them.

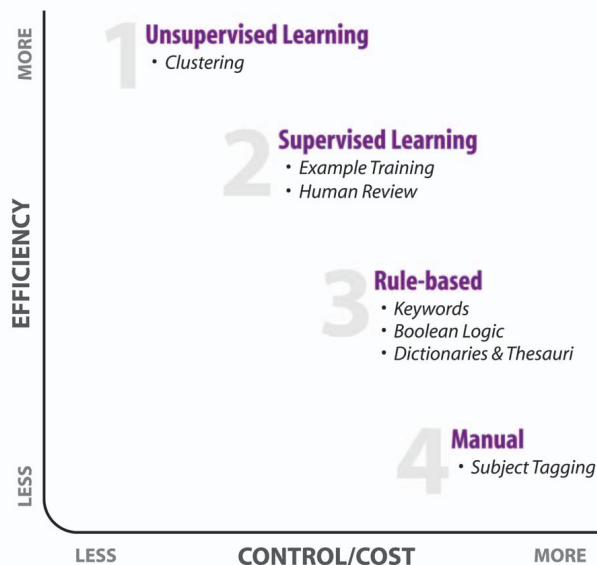
The next generation of classification technology will transcend these limitations to:

- ◆ Automate hierarchy building to reduce set up costs;
- ◆ Allow as much—or as little—human input into the classification process as desired; and
- ◆ Increase classification accuracy in the process.

Automated Hierarchy Building Reduces Set-up Costs

Setting up first-generation classification systems can be costly and time-con-

Classification Methods



Employing multiple classification methods increases accuracy and allows users to select the optimal level of control.

suming, as both the design of the concept hierarchy and the assembly of the training documents or keyword rules are left entirely to human labor. Yet these processes are key to the eventual performance of the solution.

PurpleYogi addresses this problem by partially automating the hierarchy-building process. By analyzing a sample set of documents, a PurpleYogi Discovery System™ builds a customized concept hierarchy without any human intervention in a few hours (depending on the size of the sample). Its proprietary algorithms group documents into natural categories and then organize these categories into a hierarchy. In this way, the underlying structure of the information manifests itself, saving much human effort. Using PurpleYogi software tools, non-technical people edit this initial hierarchy, combining, splitting, adding, and deleting concepts and documents as desired. The resulting hierarchy, with representative documents already resident in each node, gets a classification system up and running immediately. Alternately, customers can launch their system with one of several reference hierarchies, containing up to 12,000 concepts, that PurpleYogi has customized for individual industries.

Human Input Tailors the System to a Company's Needs

Maintaining a classification system is challenging. Concepts change meaning over time as technology and business practices evolve—documents about “microprocessor design,” for example, are very different today from what they were five years ago. And new concepts arise all the time: the newly-inaugurated “George W. Bush administration” suddenly needed to track news about the “California power crisis.” Only human judgment can decide whether and how the classification system should deal with such changes.

For this reason, we at PurpleYogi believe that classification systems should allow as much—or as little—human input into classification as is desired. A set of GUI software tools allows Discovery System administrators to monitor system performance, tweak the definition of a given concept, add and delete concepts, and even alter the classification of an individual document. With tools to adjust system performance, human administrators can balance efficiency and control, selecting the optimal mix of automation and human intervention. (See the accompanying diagram for an illustration.) In this way, human judgment supplements machine efficiency when desired and vice versa.

“Automated classification technology brings intranets to life, turning what was a passive display medium and filing system into a dynamic tool that responds to the changing needs of a corporation and its employees.”

Multiple Classification Methods Increase Accuracy

Any given classification technology has its strengths and weaknesses. Statistical classifiers, for example, are good at detecting the general subject matter of documents (“software industry,” “operating systems”), while rule-based technology is better at discriminating between concepts at finer levels of detail (“Microsoft Windows XP”). Conversely, any solution that relies on a single classification technique will fail to classify all documents with consistent accuracy.

But systems that combine multiple classification techniques can transcend the limitations of any single technique. By classifying a document in multiple ways and then comparing the results, these systems can achieve greater accuracy. Discovery Systems employ both statistical and rule-based classifiers, as well as other proprietary techniques. And they use the structure of the underlying concept hierarchy to present information grouped by topic instead of in a flat list. According to researchers from the University of California-Berkeley and Microsoft, people find information 50% faster when it is presented in this way.

The Benefits

The next generation of classification systems will help enterprises address the problem of information overload. Automated classification technology brings intranets to life, turning what was a passive display medium and filing system into a dynamic tool that responds to the changing needs of a corporation and its employees. When information moves from where it is created and travels around networks freely and effortlessly, employees can know more about what is happening in their business, how their customers

and partners are being served, and what they should do to increase revenue and profits. And they will save time and minimize rework by finding what they need quickly within the vast quantity of information available both inside and outside the company.

A variety of IT systems can benefit from integration with a classification system, including:

- ◆ Department-level document management systems (e.g., Microsoft SharePoint);
- ◆ Enterprise-scale document management systems (e.g., Documentum);
- ◆ Enterprise portals (e.g., Plumtree);
- ◆ Content management systems (e.g., Interwoven);
- ◆ Search engines (e.g., Inktomi);
- ◆ CRM applications (e.g., Siebel); and
- ◆ ERP applications (e.g., SAP).

The ability to better organize and display unstructured content either directly improves or complements all these applications.

The next generation of automated classification technology will include automated set up, the ability to apply human judgment when desired, and the integration of multiple classification techniques. While enterprises have spent billions to address the problem of information overload, the promise of these IT solutions will go unfulfilled without a powerful way to organize, classify, and present information. Only automated classification will turn the abundance of information from a curse to a blessing. ■

About PurpleYogi, Inc.

PurpleYogi creates software that automatically organizes, classifies and manages unstructured information. PurpleYogi allows enterprises to create an information hierarchy of the concepts important to their business, classify internal and external information into this hierarchy and proactively deliver information to the people who need it. PurpleYogi solutions are applicable across a range of industries including professional services, financial services, and high-technology firms. For more information, visit <http://www.purpleyogi.com>.