

Inktomi® White Paper

BEST PRACTICES FOR SEARCH AND CATEGORIZATION



ant categorization search flexible scalable trusted information retrieval relevant categorization search flexible scalable trusted information retrieval relevant categorization
ble scalable trusted information retrieval relevant categorization search flexible scalable trusted information retrieval relevant categorization search flexible scalable trus
on retrieval relevant categorization search flexible scalable trusted information retrieval relevant categorization search flexible scalable trusted information retrieval relev



Inktomi®

BEST PRACTICES FOR SEARCH AND CATEGORIZATION

SAVE, INNOVATE AND PROFIT THROUGH ACTIVE INFORMATION

A well-tuned information retrieval solution can help you save money, increase productivity, and drive innovation. How can your company create results and bottom line impact through information retrieval? For example, by supporting your call center experts with a best-of-breed search application, you can reduce call time, improve service, and lower overhead. By setting up your Research and Development team with a low-cost resource for finding relevant studies or patents across multiple repositories, you help them develop better new products faster and at less expense. And by deploying a retrieval solution to your salespeople, you give them the ability to close more sales by finding the latest product information and online market intelligence.

To ensure that the solution you deploy addresses your full range of needs, you should carefully evaluate your organization's requirements for retrieving, organizing and presenting mission critical information. Whether you already have a solution, or are seeking to acquire a new one, start by reviewing these best practices for search and categorization solutions.

PERFORMANCE BEST PRACTICES

To lower costs and empower employees to drive results, you need to get the right information to the right people, quickly. Your solution has to be fast, accurate, and capable of reaching and indexing all of your information—across portals, applications, databases, and servers. To return value to your users and your bottom line, it needs to honor your information security policies, deliver accurate categorized and ranked answers, and include multi-lingual support and other localization features. An information retrieval solution with these capabilities boosts user adoption for all of your company's applications, improving your ROI and saving your employees valuable time.

SORT FOR RELEVANCE.

The way your information retrieval solution sorts results is critical to your users. While most search solutions offer basic features for ranking and analyzing results, it is vital to select a solution that uses advanced relevance algorithms and linguistic analysis technology to deliver focused, accurate results. Is your organization getting full value from its search solution? If not, you may need to tune for relevance by taking the following steps:

- Incorporate a thesaurus of terms specific to your business, such as product names, technical or domain-specific terminology, and common misspellings. Remember to revisit the thesaurus periodically to keep it up-to-date.
- Adjust the relative weight your solution assigns to the meta-information particularly important to your company—for example, editorial-assigned “description” tags.
- Program “quick link” short cut answers to common search terms—a simple way to tune results without changing source documents or relevance algorithms.
- Configure higher document rankings for wizard-style “hub pages” that lead your users to the information they seek.

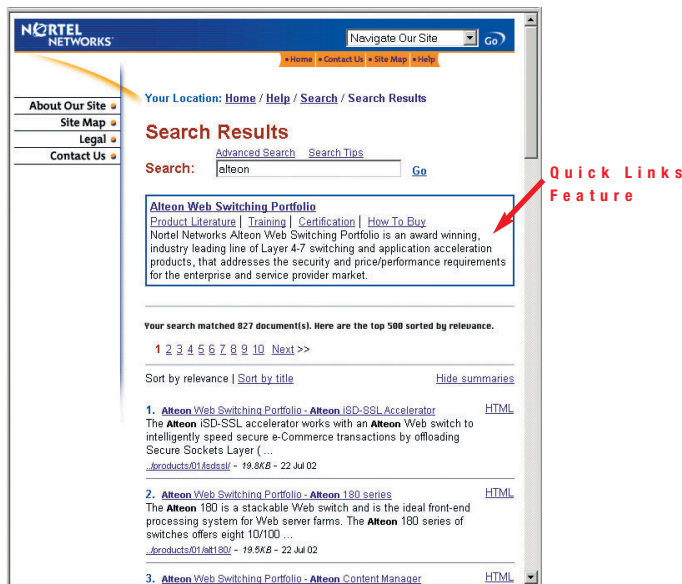
Select a solution that will allow you to tune for many different variables and rules. By setting exactly how your solution determines relevance, you can dramatically improve search performance and save time for your users.

> **Pitfall: Not Tuning for Document Relevance**

Often, companies select a solution that does not let them easily set relevance filters. This forces users to burn time clicking on links that do not answer their problem and makes it harder to promote user adoption.

www.nortelnetworks.com

Nortel Networks customized the search results to match their look and feel. Here, the “Quick Links” section speeds visitors to the most common destinations for searches on product names.



LEVERAGE METADATA.

The most effective means of improving your information retrieval accuracy is to make the most of the metadata that already exists in your enterprise. Your solutions should allow you to leverage this high-quality “information about information” available from content management systems, document repositories, or hidden document fields. Ensure that your solution allows you to refine your current metadata and can be tuned to accommodate your unique tagging policies.

EASE OF USE.

You need a solution that even casual or first-time users can easily master. The best search engines offer intuitive, customizable user interfaces, the ability to search in natural language, require no knowledge of complicated Boolean search terminology, provide advanced search features, and allow queries in familiar forms—specific phrases, keywords, even field searches. To help make the solution more useful to your end users, it should be easy to customize, allowing you to configure a look and feel that is familiar to all of your users.

CONTINUOUS INDEXING.

To keep content fresh and relevant for all of your users, your solution should include a continuous, incremental content-gathering “spider.” This means that your content is always being indexed in real-time, not in batches, giving your users much faster access to newly published information. The best crawlers automatically learn how often certain areas of your content is updated and adjust their crawling frequency to match that pattern.

“ We are constantly reshaping and enhancing our Internet content based upon customer feedback and usability testing in order to provide greater value to our online visitors worldwide. Our search engine must keep pace with these shifts by quickly detecting and indexing new content. ”

- JENNIFER MOYER, PROGRAM MANAGER FOR HEWLETT-PACKARD'S COMPANY-WIDE SEARCH APPLICATIONS

> Pitfall: Batch Indexing

Some companies deploy portals that assimilate new content according to a periodic schedule.

This approach, called batch indexing, restricts users from seeing the latest information and documents, potentially forcing, for example, sales for customer service to use obsolete information.

SINGLE SEARCH POINT ACROSS APPLICATIONS.

Your information retrieval solution should index all of your data, no matter what format it is in or where it is located. Companies have heterogeneous information architectures, with data stored within and created by several databases and applications. Many search solutions are only effective searching across either unstructured data, such as Web pages and text documents, or structured data, such as databases. Your solution should be able to handle both and to plug into diverse pools of data, index them in real-time, and make them immediately available to all users from a single search box.

Why does this matter? Imagine a salesperson pulling together a proposal under time pressure and needing information from the following sources: a best practices guide in Documentum, a sales intelligence guide on an intranet, and customer profile information in a database. To meet her deadline, she would need to be able to find everything with a few quick searches from one search location.

> **Pitfall: Multiple Search Boxes**

Many enterprises have their content unconnected and distributed across several applications. In such an environment, users might have to search with two or three different applications to access the right information and content repository. This awkward process discourages users from using the online information resources at their disposal and undermines ROI.

ENSURE CONTENT SECURITY.

It is essential that you be able to control who in your company accesses sensitive information, such as documents containing market forecast information or employee salary information. Your search and retrieval solution should accommodate your organization's protocols for both external and internal

security. The solution also needs to have built-in capabilities for setting and enforcing real-time access privileges to documents and information, and should integrate with security infrastructure platforms, such as Netegrity, as well as application-specific security frameworks, such as those found in enterprise content and document management systems.

> **Pitfall: Revealing Titles of Confidential Documents**

Many search applications allow all users to view the titles of confidential documents in their search results, a practice that can compromise secure information. Select a solution with hit-level security to protect the titles of your confidential documents.

BE GLOBAL.

If your business is international, your search solution needs to be as well. Your search solution should support multiple languages, allowing employees and customers across the world to get value from your information retrieval solution. This localization should be both at the level of the user interface and the actual search functionality.

LEVERAGE CATEGORIZATION.

Categorization is a very powerful way of cutting the "noise" out of searches. According to a recent study¹, search solutions with advanced categorization features help users find results as much as 50% faster than those without categories. Users get to the information they need faster by searching in a much more focused way—for example, searching just within one category or subcategory. A search solution incorporating advanced categorization technologies provides your users with multiple routes to the most relevant content, enabling them to navigate to the content that they are seeking in whichever way they find most natural.

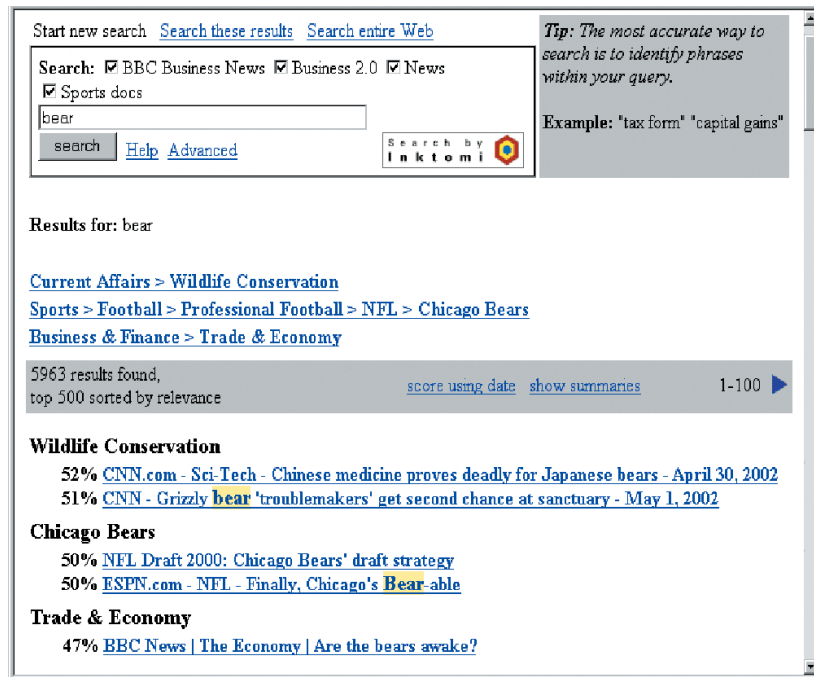
KEEP PACE WITH YOUR TAXONOMY.

There are three methods in which your enterprise can deploy and update a topic hierarchy, or taxonomy. Determine which approach is right for your company.

- **Manually develop a taxonomy.** This approach can be effective for organizations with moderate content complexity, limited sources of content, and stable user needs. Unfortunately, most companies find that their current taxonomy is outdated and requires too much manual attention to add new categories and classify documents.

Sample User Interface

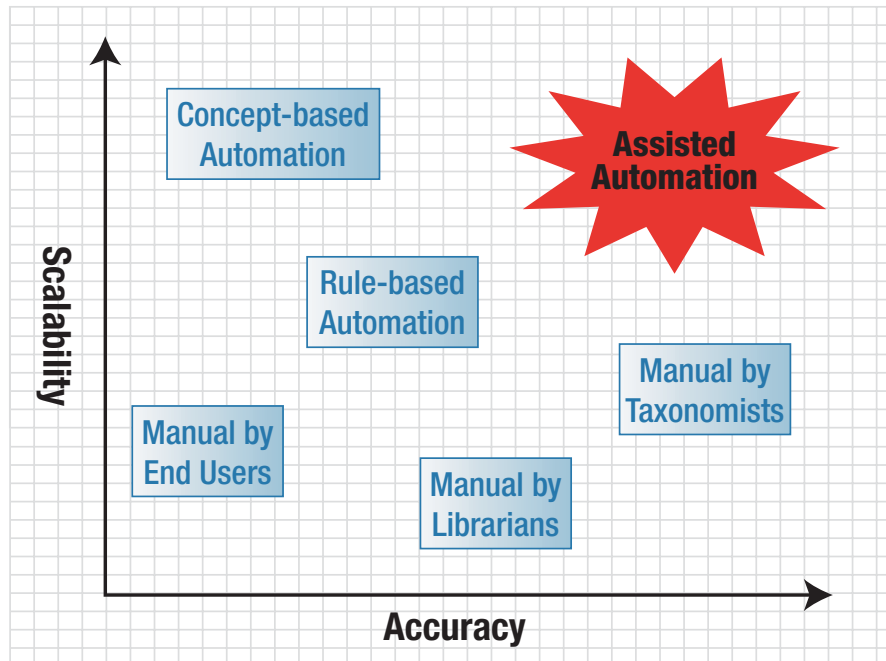
A search for the word "bear" returns results in several distinct categories. This enables fast and intuitive identification of information.



¹ H. Chen and S.T. Dumais.

Approaches to Classification

Combining automation with manual review provides the most accurate and scalable classifications.



Source: Philip Russom

- **Purchase a pre-built taxonomy.** This is a fast way to initiate a manual taxonomy. However, because every company is different, most enterprises find themselves continuously making substantial changes to their taxonomies to get value from them.
- **Deploy a taxonomy generation solution.** Solutions of this type can intelligently analyze your data and suggest enhancements to a taxonomy, or generate a complete taxonomy from scratch. This is a scalable approach that allows you to keep up-to-date quickly and easily. The best solutions give your content experts the ability to collaboratively refine the taxonomies for their respective domains.

CLASSIFY YOUR DATA AND DOCUMENTS.

Once your taxonomy is in place, you need to classify data and documents as they are published into the knowledge base. Can you manage a solution that requires significant time and effort to maintain? Or do you want an advanced solution that demands less attention, while still offering full control for review and adjustments? Because these approaches require different levels of resources and offer varying degrees of control over content, it is important to select a system that your organization will be able to sustain over time.

- **Classify with a simple rules-based solution.** In this approach, companies manually specify classification rules for each content “node” on the hierarchy. While this approach can be very accurate, it can require a substantial amount of

solution administration time, depending on the breadth and complexity of the content. This approach is effective for, say, a small manufacturer classifying product information on its external Web site for customers.

- **Classify with an advanced, automated solution.** This approach routes content to the appropriate categories based on a combination of statistical formulas. To ensure that the solution is effective, your topic experts should review its classification recommendations regularly. While pure automation is fast and scalable, it can easily misclassify complex documents, thus making them harder for users to find. Using a blend of automatic and manual refinement provides the most accurate classification and enables the system to “learn” how to better classify content over time.

ROI BEST PRACTICES

When it comes to your information retrieval solution, you need a fast, secure path to value. With information retrieval becoming increasingly important to your enterprise, it makes sense to adopt a solution that will create results for you quickly and remain reliable, easy to integrate, easy to maintain, and scalable through all stages of its lifecycle.

RAPID DEPLOYMENT AND EASE OF INTEGRATION.

Your solution should be easy to deploy without long, costly delays and configuration processes. You want to expend a minimum of time and resources on building your search and categorization system to maximize its utility. Your solution should support a wide variety of information access protocols and standards out of the box, including HTTP, HTTPS, NNTP, Microsoft Exchange, Lotus Notes, and SQL. It should also offer open HTTP, XML, and Java APIs to provide you with the ability to integrate search into your Web-based applications with a minimum of developer effort.

EASE OF MAINTENANCE.

From the user interface, to the taxonomy, to the means it uses to index data, your solution should allow you to control every aspect of its functionality. You need an application that your system administrators can access, tune and maintain without bringing in costly outside professional services assistance. For added flexibility, it should have a browser-based administrative interface to incrementally adjust the search engine's configuration, without requiring programming or systems administrator resources.

“ We've found that less of the administrators time is taken up by the solution since we've made Inktomi the search engine. We're able to connect to more information while spending less time managing the solution. ”

- GAIL LESLIE, MANAGER
OF KNOWLEDGE MANAGEMENT
PROCESSES, TOWERS PERRIN

> *Pitfall: High Cost of Tuning*

With many search solutions, maintenance consumes an unexpectedly large amount of time and resources, with awkward configuration processes forcing system administrators to tune the system repeatedly. This burns expensive internal IT human assets and drives up the total cost of the solution.

SCALABLE.

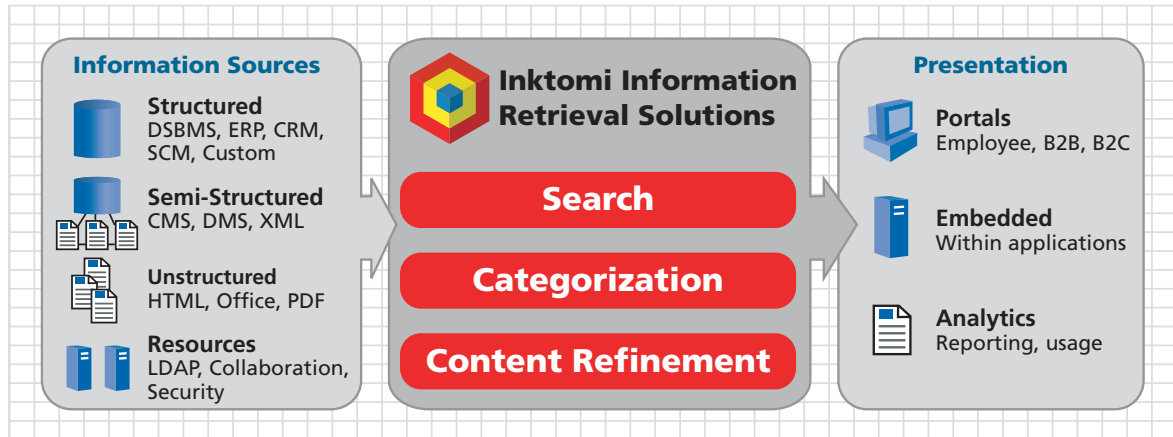
Your company changes rapidly, and your information retrieval solution should reflect that. As your organization and knowledge base grows and changes to reflect the demands of the market, the solution should scale easily and affordably to match your new requirements. This includes the ability to handle more queries, introduce more complex search functionality, and extend the categorization taxonomy to include new topics.

> *Pitfall: Hidden Hardware Costs*

Some search solutions that seem scalable at first glance actually require substantial new investment in hardware. Companies can end-up spending far more to scale the solution than planned. Be sure to consider all costs of scaling your solution before you invest in a new application.

Inktomi Information Retrieval Solutions

Inktomi provides a full suite of solutions to find the right information any time.



INKTOMI INFORMATION RETRIEVAL SOLUTIONS

Inktomi, a leading provider of information retrieval solutions, is the validated, right choice for your company's search and categorization solution. Over 2,500 customers, including 8 of the top 10 Fortune 500 enterprises, trust Inktomi Enterprise Search to deliver unconstrained information access. Inktomi's base of customers and strategic partners includes such leading companies as Cable & Wireless, Dell, Hewlett-Packard, IRS, Merrill Lynch, Microsoft, Nokia, Sun Microsystems, the U.S. Library of Congress, and Yahoo!

Inktomi Enterprise Search helps individuals rapidly find the right information at the right time within enterprise applications, portals, databases and networks. Inktomi's hallmarks are developing software that is simple to deploy and manage, and offers powerful administration tools and an open architecture.

Inktomi Classifier efficiently organizes distributed information into an intuitive topic hierarchy, or taxonomy. Through comprehensive, collaborative tools that provide clear visibility into every stage of the information organization process, customers are able to build and maintain an accurate and flexible information resource tailored for their organizations. Inktomi Classifier utilizes a combination of Bayesian and rules-based classifiers to provide the greatest accuracy with minimal effort.

Inktomi Topic Advisor is the most accurate and efficient way to create and update taxonomies. The system provides information managers complete control with category review and approval of changes prior to publishing. Inktomi Topic Advisor enhances existing taxonomies by identifying new topics or suggesting the consolidation or splitting of existing topics. The product can also help create taxonomies from scratch by analyzing

unstructured data and suggesting topics or nodes around subject matter clusters. Inktomi Topic Advisor is an optional component of Inktomi Classifier.

Inktomi Content Classification Engine (CCE) organizes content into browseable topics for easy searching. This is a cost-effective, basic categorization solution that leverages rules-based classifiers. It is ideal for small to medium-sized organizations or organizations with simple classification requirements.

CONTACT US

If you are currently evaluating the effectiveness of your organization's current approach to information retrieval, we can help. Let us assess your organization's specific needs and help you develop a detailed ROI case. Contact us at search-solutions@inktomi.com or 1-888-INKTOMI (465-8664) to speak with a sales representative and schedule a consultation.



I n k t o m i®

Inktomi Corporate Headquarters
4100 East Third Avenue, Foster City, CA 94404
1-650-653-2800
www.inktomi.com

Copyright © 2002 Inktomi Corporation. All rights reserved. Inktomi and the tri-colored cube logo are trademarks or registered trademarks of Inktomi Corporation in the United States and other countries. All other trademarks mentioned herein are properties of their respective owners.

INK-WP-CAT (08/02)

