

The ABCs of Content Organization

Verity White Paper



The ABCs of Content Organization

July 2002

Table of Contents:

The ABCs of Content Organization	3
Automatic Classification	3
Business Rules	3
Concept Extraction	4
Combining Domain Expertise and Machine Efficiency	4
The Most Flexible, Accurate Content Organization Available	4

Publisher's Note: Information contained in this document is intended for guideline purposes only. Verity product documentation supercedes information contained in this document. The situations described in this document are offered as examples; actual configurations and results will vary from system to system.

The ABCs of Content Organization

In the past corporations had to make a choice. They could either classify information accurately, but expensively, or inaccurately for a much lower cost. Manual methods required extensive human effort to classify every single document in an enterprise, and suffered from time, scalability and consistency problems. To decrease the time and cost of manual classification, many software vendors have developed automatic classification solutions over the last couple years. These solutions were a step forward, because they eliminated the cost of hiring human subject experts to tag documents one-by-one. Unfortunately, they often lacked the accuracy necessary in a business environment.

Another unfortunate side effect of the proliferation of vendors claiming “automatic” classification solutions is the confusion they have created. Before you can determine which classification solution is right for your enterprise’s information environment, you should ask yourself a number of questions.

- Does your enterprise content have associated metadata that could be used for organization purposes?
- Has an organization system (taxonomy) already been developed?
- Is suitable representative data available for building a taxonomy?
- Is domain expertise available to specify characteristics for certain categories?
- Is the content you want to organize of relatively high or low value?

These questions will determine how accurate the content classification needs to be, and the investment in terms of time, resources and money that will be required to achieve that level of accuracy. Some businesses will find that completely automatic classification will produce satisfactory results. At the other end of the scale, some companies will decide that the complexity of their corporate information and its high value to their core business require that automatic classification be augmented with significant amounts of domain expertise. Verity’s content organization tools can build automatic classifiers for both of these extremes, as well as classifiers that meet the needs of businesses that fall somewhere in the middle.

The ABCs of Content Organization

There are three methods to automatically organize content. Verity refers to these as the ABCs of Content Organization:

- Automatic classification
- Business rules
- Concept extraction

Automatic Classification

This method assumes that a taxonomy already exists, possibly created by domain experts or from metadata, URLs and/or file paths. Classification rules that define each category are

automatically learned from exemplary documents in the category using Verity’s benchmark-leading, patent-pending Logistic Regression Classification (LRC) technology.

With this method, the two ways that categories can be defined are:

- Using the Verity Intelligent Classifier (VIC), Verity’s content organization graphic user interface (GUI), to extract a hierarchy of categories in a taxonomy based on URL paths or file paths. This method lets you mirror existing Web site or file system structures in the taxonomy. Documents are automatically associated with corresponding categories.
- Using VIC to base categories on a Verity Collection field in document metadata. This method can be used when the categories are explicitly listed in a field in the collection. For example, categories can be specified in HTML documents by metatags that are indexed into a Verity Collection field. Documents are then automatically associated with corresponding categories.

Once a set of documents is associated with each category, LRC can be invoked to automatically learn rules that define each category. This can be done one category at a time, or recursively for all the subcategories under the specified top-level category. Verity’s LRC can learn rules based on positive as well as negative exemplary training documents, which results in a higher level of accuracy. Many automatic classifiers on the market can only work with positive examples, and the accuracy of their classification suffers correspondingly.

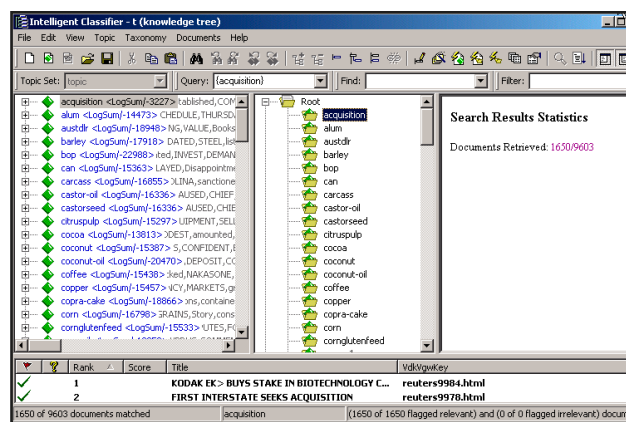


Figure 1. Automatically created rules for all categories.

Business Rules

This method lets you leverage your enterprise’s domain expertise using VIC to construct a hierarchy of categories. The characteristics of each category in the taxonomy can then be specified with easily created and reused rules. These characteristics can make use of structured data like the “author,” “location” and “deal size,” as well as concepts like “good deal,” “difficult customer” or “up-sell opportunity.”

Using domain expertise to specify characteristics of categories in this way is a very efficient and accurate method to organize content.

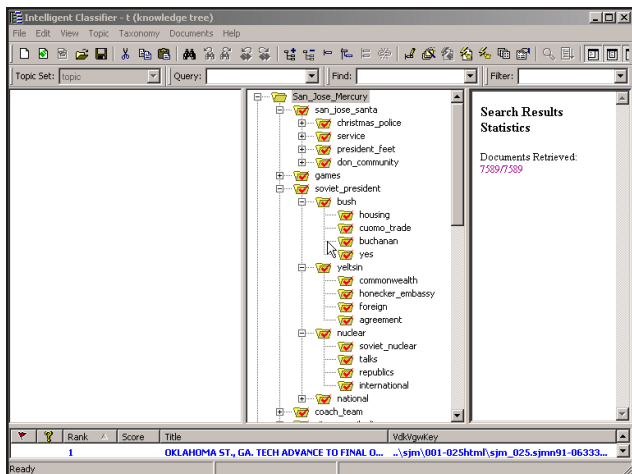


Figure 2. Category hierarchy automatically created by Verity's Thematic Mapping.

Concept Extraction

Here the system analyzes the entire set of documents with no human supervision, and extracts the dominant themes (for example, "Petrochemicals," "Human Resources," or "International Equities"). In addition to labeling each theme, it suggests a taxonomy for organizing these themes.

First, the document collection is specified and Verity's patent-pending Thematic Mapping technology organizes documents by themes. The Verity Intelligent Classifier (VIC) then groups these themes into a multi-level taxonomy, with each theme representing a category or subcategory. Finally, the themes are automatically labeled with an easily understood name that can be modified to match your corporation's vocabulary or industry terminology. Figure 2 shows a few of the categories and subcategories created by Verity's Thematic Mapping when applied to a collection of San Jose Mercury News articles.

This process is fully automatic, and requires no human intervention. The automatically generated category hierarchy organizes content into automatically generated subject areas using business rules that are fully compatible with those created using the Automatic classification and Business rules methods described previously.

Combining Domain Expertise and Machine Efficiency

There are a variety of ways to combine the ABCs of Content Organization to efficiently and accurately organize any set of documents, depending on your business needs. For example:

- Construct the top level of a taxonomy by specifying several broad categories, associate a set of exemplary documents with each category, and use Thematic Mapping to extract concepts and automatically create subcategories. Next, employ domain expertise to tailor the automatically created subcategories to fit specific business needs. Finally, use LRC to automatically create rules that define all the categories and subcategories.

- Create the rules that define the categories, and test the rules to see if they retrieve documents that match them. Next, view the result documents, mark them as "relevant" (positive examples) or "irrelevant" (negative examples), and use LRC to update the rules.
- Select a set of relevant and irrelevant documents for a category, and use LRC to create the rules that define the category. Examine the rules and edit them to incorporate domain expertise about the subject represented by the category.

The Most Flexible, Accurate Content Organization Available

Verity K2 infrastructure delivers the broadest and most accurate range of content organization tools currently available. Its flexibility allows businesses like yours to select the tools and methods that make the most sense to your business model, so you can realize the full return on investment that properly organized content delivers. In fact, only Verity provides the full range of classification methods—from totally automatic classification to automatic methods augmented by various degrees of domain expertise—required to maximize the value of your growing information assets.

Verity, Inc.

894 Ross Drive
Sunnyvale, CA 94089
t. 408.541.1500
f. 408.541.1600

Verity New York

230 W. 41st Street, 12th Floor
New York, NY 10036
t. 646.366.9500
f. 646.366.9540

Verity Washington

2941 Fairview Park Drive, 8th Floor
Falls Church, VA 22042
t. 703.289.8800
f. 703.289.8840

Verity Chicago

200 South Wacker Drive, 31st Floor
Chicago, IL 60606
t. 312.674.4525
f. 312.674.4571

Verity Boston

35 Corporate Drive, 4th Floor
Burlington, MA 01803
t. 781.685.4881
f. 781.685.4601

Verity Dallas

12377 Merit Dr., 11th Floor
Dallas, TX 75251
t. 972.455.4640
f. 972.455.0474

Verity Canada

400 – 4th Avenue SW, Suite 2600
Calgary, AB T2P 0J4
Canada
t. 403.750.4000
f. 403.750.4100

Verity Mexico

Camelia 253-6, Colonia Florida
Alvaro Obergon, Mexico D.F.
01030 Mexico
t. 52.5.661.7126
f. 51.5.598.9526

Verity Benelux

Coltbaan 31
3439 NG Nieuwegein
The Netherlands
t. 31.30.669.2120
f. 31.30.662.2094

Verity GB Ltd.

The Pavilions
Kiln Park Business Ctr.
Kiln Lane
Epsom KT17 1JG
t. 44.1372.747076
f. 44.1372.747071

Verity Germany

Babenhäuser Straße 50
D-63762 Großostheim
Germany
t. 49.6026.9710.0
f. 49.6026.9710.20

Verity France

14, Place Marie Jeanne Bassot
92593 Levallois Perret Cedex
France
t. 33.1.41.49.0450
f. 33.1.40.89.0981

Verity South Africa

492 Brown Street
Laudium, 0037
Centurion, Pretoria
Gauteng, South Africa
t. 278.3603.1134

Verity Australia

Level 20, 99 Walker Street
North Sydney NSW 2060
Australia
t. 61.2.9657.1055
f. 61.2.9657.1059

www.verity.com

© 2002 Verity, Inc. All rights reserved. Verity®, TOPIC®, KeyView®, and Knowledge Organizer® are registered trademarks. The Verity logo, Verity Portal One™, and Verity® Profiler are trademarks of Verity, Inc., in the United States and numerous other countries. All other trademarks or symbols are those of their respective owners.

**Sales and Product Information**

info@verity.com

Partnerships

partners@verity.com

Technical Support for Existing Customers

tech-support@verity.com

Verity, Inc.

894 Ross Drive, Sunnyvale, CA 94089 | t. 408.541.1500 | f. 408.541.1600 | www.verity.com