Extracting Value From Automated Classification Tools

The Role of Manual Involvement and Controlled Vocabularies

By Kat Hagedorn, Argus Associates

March 2001

http://argus-acia.com/

TABLE OF CONTENTS

INTRODUCTION

Automated classification has been trumpeted by vendors as the solution to the problems we face in managing our huge and growing information spaces. Most of us feel like clutching this solution as a drowning person would a life vest because it offers a way of re-gaining control over this information.

We do need a solution, but in fact, as in most things, this type of solution works only if you have a plan and a process for using it. A plan and a process are developed from understanding the environment of your information space—in other words, your business context, content and users. Examples of questions to ask about implementing an automated classification solution in your environment include:

- **Context**. Is there buy-in at the top levels of management for automated classification and for providing the appropriate resources for it? Do you have staff members who fully understand the nature of your business and can translate that into an appropriate classification?
- **Users**. Are your users primarily searchers or browsers? Are they interested in categories that are topical, task-oriented, geographical or audience-based, or do they need customized organizational methods?
- **Content**. Does your information space contain content that is all one type (i.e., homogenous) or different types? Is it in multiple languages?

As part of developing this plan and process, you will also need to:

- **Choose the mix of manual and automated classification**. Do you need humans to create rules used to classify the information? Will you expect the tool to suggest classifications for humans to review? What level of quality control will be necessary?
- **Choosing the controlled vocabularies you use with the automated classification tools**. Will you need an author vocabulary as well as a subject vocabulary? Do you have the resources to build the controlled vocabulary from scratch? Will a thesaurus be more suitable?

This white paper will help answer these questions and others. It discusses the types of automated classification tools that are available, how manual classification fits into the mix, why you need a controlled vocabulary to do the job right, and asks how we can move towards community-wide testing of these automated classification tools.

AUTOMATED CLASSIFICATION 101

Automated classification has been described in a multitude of ways, depending on the point of view of the author (e.g., academician, vendor). A recent excellent discussion is provided in Katherine C. Adams' article "Word Wranglers" (1), which includes a description of the tools available, their features and the techniques used for clustering information.

The following takes a step back and explains what classification is in general, why it is useful and then briefly summarizes the automated classification tool techniques. It is not our intention to discuss details of any of these tools—Adams' article goes into depth on this topic.

DEFINITION OF CLASSIFICATION

Classification is the process by which information, whether in document or data format, is clustered together to make it easier for the user to find it. (For the scope of this white paper, the terms "classification" and "categorization" can be used synonymously.)

For instance, while classifying a collection of documents about food, you might want to cluster documents by subject—e.g., those that discuss "methods of cooking," those that discuss "kitchen tools"—depending on how useful your audience would find that type of classification.

BENEFITS OF CLASSIFICATION

In a nutshell, classification assists people who are:

- **Browsing**. For example, a user may browse through a classified structure (e.g., Men's Clothing: Shirts: T-Shirts) to find relevant information—this structure reflects clusters created by classification.
- **Searching**. For example, a user searching for "cars" retrieves all information that is clustered around this topic. The search engine retrieves these clusters and returns these results to the user.
- **Managing**. For example, an editor looking for the appropriate place to put a piece of information can use a classified structure to help suggest places. The resulting placement of information forms clusters.

The end result of classification is structures that are commonly known as ontologies, taxonomies, hierarchies, controlled vocabularies or thesauri—depending on the discipline involved. A superb article by Dagobert Soergel (2) reflects on the use of these names and notes that one of the main features of any classification structure is to support information retrieval. In this white paper, I will focus on the use of specific types of these structures—i.e., controlled vocabularies and thesauri—to support more efficient and relevant information retrieval.

AUTOMATED CLASSIFICATION TOOLS

Automated classification is the process by which technology is used to create clusters. Some of the most popular vendors that have marketed stand-alone automated classification tools or add-ons to their search or content management tools are Autonomy, Inktomi (formerly Ultraseek), Interwoven (formerly Metacode), Mohomine, Semio and Verity.

These tools cluster information by using one of the following techniques:

- **Statistical clustering**. Employs algorithms to cluster information. Popular methods include term co-occurrence analysis and neural networks. This technique is dependent on the information in the collection—it classifies using the terms in the collection exclusively. Vendors using this technique are Autonomy, Interwoven, Mohomine and Semio.
- **Rules-based clustering**. Necessitates the creation of IF-THEN statements that define the clusters of information. This technique builds a classification structure that can be reviewed and edited manually, while populated automatically. It isn't dependent on the information in the collection—a new collection could use the same statements. Vendors using this technique are Inktomi and Verity.

An additional technique, designed for improving the above techniques, is:

- **Training**. Compares to-be-classified information with previously well-classified information, and collects these together in clusters. Training can happen at the beginning of the classification process or iteratively throughout. Vendors using this technique are Autonomy, Inktomi and Mohomine.

These automated classification tools work with varying degrees of success. This success depends on the involvement of humans and the incorporation of controlled vocabularies, each discussed in the next two sections.

ADDING HUMANS INTO THE MIX

A number of automated classification tool vendors, pundits and those working in this field have noted that fully automated classification of information is not the complete answer. (1,3,4,5) They state that a semi-automated solution is more appropriate. Peter Morville puts it most succinctly in his ACIA "Little Blue Folders" article: "The key to success in designing information architecture solutions for really large web sites and intranets is to intelligently combine manual AND automated approaches." (6)

Recently, there have been detractors concerned that the cost-benefit gap may be too large, suggesting that involving humans is an unnecessary time expenditure when machines are capable of doing the entire process. (7) I would respectfully disagree. Although there haven't been formal evaluations of this, in my experience manual intervention is necessary, especially for the conceptual parts of the process.

Good indexers and catalogers know the difficulties of developing appropriate clusters of information. Often it isn't enough to just pull out terms that reside in documents; concepts that aren't stated verbatim within a document also need to be reflected in the classification. Computers are not able to understand the meaning in documents or the context surrounding them. As stated by Nancy Mulvany, "computers are capable of automatically manipulating the text of a book in a variety of ways. Yet the computer is incapable of exercising the type of judgment and interpretation applied by experienced indexers." (8)

RANGE OF HUMAN INVOLVEMENT

The use of humans in the classification process runs the gamut from minimal involvement to full-fledged participation.

Humans can be involved with automated classification tools by:

- **Creating the classification**. Developing all or part of the structure to be used during the automated classification process. For instance, manually creating the top-level clusters of the classification structure and having the tool automatically create the bottom-level clusters. This structure can be newly created or tweaked from out-of-the-box versions, and should be predicated on a controlled vocabulary or thesaurus (see next section).

- **Developing the classification rules**. Manually creating and editing the rules that govern the clustering of information. These rules should also be predicated on a controlled vocabulary or thesaurus.
- **Training the collection process**. Moving information from one cluster to another, when necessary, so that the tool learns what information should exist in specific clusters.
- **Implementing suggestions**. Reviewing options provided by the tool for classifying information, analyzing these and using them or choosing other options to best cluster the information.

Throughout the automated classification process, humans should be:

- **Performing quality control**. Reviewing clusters and the information in the clusters after the tool has done its work, and making any changes. Hopefully, the tool also allows humans to perform some of the above listed functions (e.g., creating the classification, training the tool), in order to reduce the need for extensive manual quality control.
- **Testing the classification**. Designing methods for testing the usability (e.g., appropriateness, ease of use) of the resulting classification structure, running tests and analyzing the results. Analysis should provide you with recommendations for improvement to the classification structure and possibly improvements to the classification process.

BENEFITS AND LIMITATIONS OF MANUAL AND AUTOMATED APPROACHES

To fully understand the benefits and limitations of choosing manual and automated classification methods, you will need to develop a table similar to the following:

APPROACH	BENEFITS	LIMITATIONS
Manual	▪ Better understanding of conceptual nuances ▪ Understanding of company environment ▪ High quality ▪ High consistency ▪ Ability to modify structure ▪ Ability to classify images and applications	▪ Slower ▪ Perceived as resource-hungry ▪ Perceived as high cost

APPROACH	BENEFITS	LIMITATIONS
Automated	▪ Perceived cost savings ▪ Perceived as fewer resources required ▪ Frees humans from routine tasks ▪ Minimizes time to complete classification process	▪ Uncertain quality ▪ Uncertain consistency ▪ Overhead for technology and its expertise ▪ Inability to deal with exceptions to rules ▪ Often needs fairly homogeneous collections ▪ Often needs large collections ▪ Often needs full-text collections (not records)

Your environment aside, the most pressing issue of your classification efforts is the quality and consistency of the classification process. The following section addresses how implementing controlled vocabularies assists in limiting poor quality and inconsistency in your classification.

USING CONTROLLED VOCABULARIES

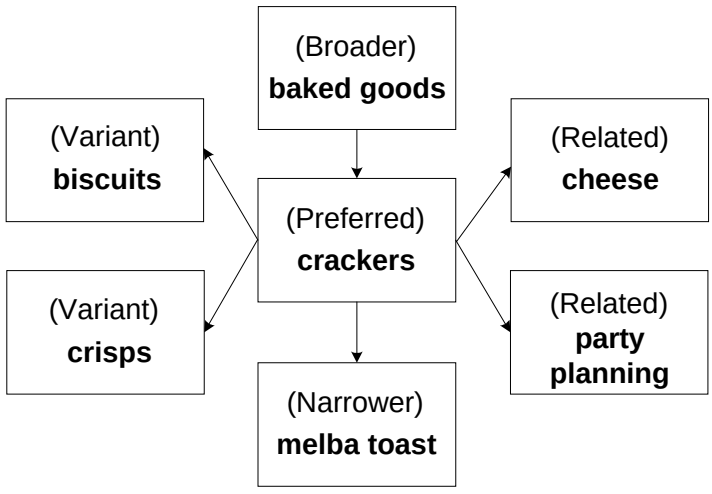
Here's where the users matter the most. You've developed a system that adequately classifies the information you have in your information space, based on the resources and capabilities of your company. You've determined the mix of automated and manual classification and have chosen an appropriate tool. The next step is to identify what the users want.

Most users are interested in directly relevant information (as opposed to all potentially interesting information). Each of the automated classification techniques noted above claims to give users the most relevant information. But if you really want to ensure that users are retrieving what they need, you should be using a controlled vocabulary or thesaurus. Reasons for using these types of classifications are explored in this section.

DEFINITIONS OF CONTROLLED VOCABULARY AND THESAURUS

A controlled vocabulary (CV) is a list of descriptive terms that indicates which terms are preferred and which are variants of the preferred terms. A thesaurus is a special kind of controlled vocabulary that also indicates which terms are broader, narrower and related to the preferred terms. For instance, a controlled vocabulary on food might indicate that "crackers" is the preferred term and "crisps" is a variant term. A thesaurus might also indicate that "cheese" is a related term and that "melba toast" is a narrower term. The following diagram illustrates these relationships.

Figure 1. Thesaurus: relationships between terms.



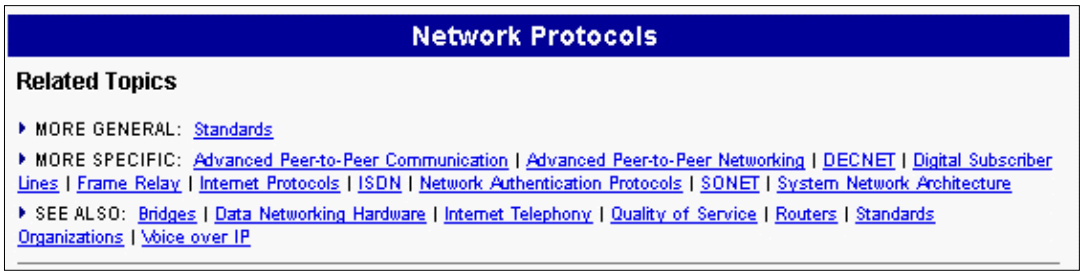
Without a CV in place, automated classification tools can produce something like the following:

Figure 2. Semio: classification of "e-commerce."

In this example, if a CV is in place, a preferred term for "application" or "software" can be chosen, with the other one listed as a variant term. Then, at least three clusters might be merged into one (e.g., "electronic commerce application," "internet commerce application," "electronic commerce software"). And, using a thesaurus, narrower terms might be used for sub-clustering (e.g., "shopping cart tools," "product finders").

Consequently, information is re-distributed into appropriate groupings based on the CV or thesaurus. The end result is more relevant clusters of information for users—they can find what they need gathered in one place. Using this method, well-classified information can look like the following:

Figure 3. Bitpipe: thesaurus term "network protocols."



In this example, broader, narrower and related terms are provided for "network protocols." This gives the user context for this term and avenues for further exploration—it enhances their information retrieval experience and increases the likelihood that they will find relevant results.

For more examples and further resources about CVs and thesauri, see the online materials from the 2001 ACIA "Synonyms and Taxonomies" seminar (argus-acia.com/seminars).

Incorporating Controlled Vocabularies

So, how do you use CVs with automated classification tools? In essence, anywhere a classification is used, a CV can be used. Options include:

- **Creating** the top levels of the classification with a CV and letting the tool automatically create the bottom levels.
- **Informing** the creation and editing of IF-THEN statements with a CV.
- **Training** by selecting preferred and variant terms from a CV and populating clusters with information that includes those terms.
- **Comparing** suggestions for clustering with a CV.
- **Reviewing** statistically derived clusters, matching them against a CV and changing the clusters appropriately.

Obtaining Controlled Vocabularies

Procuring pre-built CVs can be difficult for a variety of reasons. They can be:

- **Inexistent**. Is there a CV that has been created in your subject domain and that fits how your users look for information?
- **Incomplete**. Does the CV cover all the topics you're trying to cover?
- **Inconsistent**. Does the CV consistently organize items at appropriate levels of detail?
- **Incomprehensible**. Is the CV too technical for your needs?

The bottom line is: How much will the CV need to be tweaked? Sometimes the best approach is to build one from scratch and refine it during the classification process. This means involving more resources (time, money, people), but the end result will be more satisfactory than a pre-built CV.

Another option is to use the text to generate the controlled vocabulary (as opposed to statistically generating clusters that aren't controlled). There has been research done on the use of tools that automatically extract terms from information and then build a thesaurus from those terms. (9) This method is best used in conjunction with manual involvement to ensure the relevancy of the CV. The benefit of this method is that the CV is directly connected to the information being clustered.

See Lou Rosenfeld's recent article "Looking for Metadata in All the Wrong Places" for an accessible discussion of CVs and how to obtain them. (10)

Issues to be Aware of During Development

Developing CVs involves issues that may not be obvious initially. Some of these are:

- **Creating cross-references or double-posting**. It's often useful to provide "see" references (e.g., "cars *see* automobiles") and "see also" references (e.g., "cars *see also* travel") among clusters, or to selectively place information in more than one cluster (i.e., double-post it). This makes the relationships among terms obvious and makes the resulting classification more usable.
- **Building more than one CV**. To reach your primary users and to reflect all your content, you may need to develop multiple types of CVs. Examples include those organized by user role (e.g., engineer, administrator), topic (e.g., crackers, cookies), action or task (e.g., selling, buying) and geography (e.g., France, England).
- **Building CVs in more than one language**. If your collection contains considerable amounts of information in multiple languages, you should have a CV for each language. This often involves more than simple translation of terms, since concepts, jargon and relevancy of terms differ among languages.
- **Building a thesaurus**. If you have the resources, a thesaurus can greatly help your users and your business. You will be offering users navigational assistance (e.g., broader, narrower, related terms), which can help them generate more relevant results. Managing a classification structure that includes more relationships among terms can be a very powerful avenue for translating your business strategy—e.g., promoting products using broader clusters ("you're interested in 'construction tools,' you may want to see everything related to 'home improvement'").

A Call for Testing

I've discussed the need for manual involvement in your classification process and I've discussed the need for controlled vocabularies in that process. What I think is missing in these discussions is testing of these variable, and sharing of the results with the wider community.

Situations exist in which testing can occur, but inevitably there are limitations to doing this kind of testing. These situations include:

- **Partnering with an automated classification tool vendor** so you can run pilot tests with actual data using their tool. The limitation is the partnership agreement and how that affects your company's strategy.
- **Obtaining a trial version of a tool** to run against actual or contrived data. The limitation is that the trial version may limit you to a certain number of documents, not allow the integration of controlled vocabularies or not allow you to customize the classification process.
- **Contracting with a vendor to do a full test of their tool**, which usually involves the assistance of consulting engineers from the vendor. The limitation is the added expense of paying for the consultants just to see whether you want to use the tool.

I feel that there is a greater opportunity out there—an opportunity to provide the community with real data that has been tested in a variety of ways using a variety of automated classification tools, with results that can be shared and compared without sharing proprietary vendor information with the community. In other words, I feel there is an opportunity to form an international collaboratory testbed for these types of tools.

The Text Retrieval Conferences (TREC) project provides an excellent example of a collaboratory testbed (trec.nist.gov). Each year, the project provides data sets that interested parties can use to run against new technology to see where that technology succeeds and fails. It's not a giant leap to assume that something like this could be done to test automated classification tools. Since TREC's environment is set up for new technology and not established technology, it may not be the best venue for the proposed testing.

Testbed Features

In order for use of the testbed to be most effective, the community needs to enforce:

- **Description of the information** in the testbed. What type of information are you testing (e.g., engineering, religious studies)? Is it mostly text or records-based? Is it highly conceptual content or fact-based data?
- **Information in multiple languages**. This will assist in reaching the largest community of testers.
- **Selection of controlled vocabularies** to run against information. Multiple types of CVs are a necessity. Multiple language CVs will be necessary for information in multiple languages.
- **Ability to perform the test in a multitude of ways**. For instance, running text-based English language content with multiple topic domain CVs using three selected tools, or running records-based French and English language data without a CV through five selected tools.

Benefits of Testing

The testbed would benefit both the automated classification vendors and those interested in using the tools. It can assist in:

- **Selling the need for the tool to the leadership**. Quantifiable effects are easier for upper management to buy into.
- **Providing a way for content managers to compare tools** and see what types of tools work best for their types of data and resources.
- **Comparing manual and automated methods of classification**. The level of human involvement could be tested under a variety of conditions.
- **Getting feedback from users**. Testing with users can provide subjective and objective data on which tools work in which environments.
- **Selling the tool to you**. Vendors with quantifiable data on their tool's performance can use this to market the tool.

As a consultant, I also find the opportunity to share results from prior and present engagements to be a strong selling point for future clients. There is a dearth of tangible outcomes in the field of information architecture. It would be useful to have concrete results showing that controlled vocabularies are effective using certain tools and certain types of information—this would benefit the client and the entire information architecture community.

REFERENCES

1. Adams, K.C. “Word Wranglers: Automatic Classification Tools Transform Enterprise Documents from ‘Bags of Words’ into Knowledge Resources.” (www.intelligentkm.com/feature/010101/feat1.shtml)
2. Soergel, D. “The Rise of Ontologies or the Reinvention of Classification.” JASIS, 50(12), 1999. pp. 1119–1120.
3. Koch, T. “Specification for Resource Description Methods. Part 3. The Role of Classification Schemes in Internet Resource Description and Discovery.” A DESIRE deliverable. Section 3.7. (www.lub.lu.se/desire/radar/reports/D3.2.3/class_v10.html)
4. Shafer, K.E. “Automatic Subject Assignment via the Scorpion System.” (www.oclc.org/oclc/research/publications/review96/scorpion.htm)
5. Larson, R.R. “Experiments in Automatic Library of Congress Classification.” JASIS, 43(2), 1992. p. 147.
6. Morville, P. “Little Blue Folders.” Strange Connections, July 10, 2000. (argus-acia.com/strange_connections/strange003.html)
7. SIG-IA discussion list. Weeks of 2/4/01 and 2/11/01. (www.asis.org/Conferences/Summit2000/Information_Architecture/listserv.html#archives)
8. Mulvany, N.C. Indexing Books. Chicago, University of Chicago Press, 1994. p. 46.
9. Chen, H., et al. “Automatic Thesaurus Generation for an Electronic Community System.” JASIS, 46(3), 1995. pp. 175–193.
10. Rosenfeld, L. “Looking for Metadata in All the Wrong Places: Why a Controlled Vocabulary or Thesaurus is in Your Future.” WebReference, February 2, 2001. (www.webreference.com/authoring/design/information/cv/index.html)

About the Author

Kat Hagedorn (kat@argus-inc.com) joined Argus Associates as an Information Architect in 1998. Her background in biological sciences and information science (a B.S. from Cornell University and an M.S. from the University of Michigan School of Information) brings an added dimension to her expertise in designing organizational systems for clients.

Kat's previous publications include "Bottom-up Information Architecture: Leveraging Metadata" in SLA's Content Management book (1998) and the "Information Architecture Glossary" for the ACIA (2000). She has spoken in several venues, including at the American Society of Indexers Annual Meeting in May 2000.

At Argus, Kat has provided information architecture consulting services to a variety of clients including 3Com, Applied Materials, AT&T, Corning, Procter & Gamble, Square D and The Weather Channel.

About the Argus Center for Information Architecture

Mission

The Argus Center for Information Architecture provides leadership in defining and advancing the evolving discipline of information architecture.

What We Do

The Argus Center serves as a focal point for learning about the theory and practice of information architecture. Towards this goal, we:

- Manage a selective collection of links to the most remarkable content, events, and people in our field.
- Produce original articles, white papers, conferences, and seminars that draw from the experience and expertise of the Argus team.
- Conduct research, independently and through partnerships, focused on improving our collective understanding of information architecture.

Who We Are

The Argus Center for Information Architecture was created by information architects for information architects.

It is sponsored by Argus Associates, a consulting firm that specializes in information architecture design. The entire Argus team contributes to its development.

The Argus Center also draws from the broader community of information architects, through partnerships with individuals, corporations, and universities.

Learn More

To learn more about the publications and events of the Argus Center, please visit our web site at argus-acia.com.