Automated text categorization

September 2006
Warning

This document was created by Digimind.

The content of this document is protected by copyright. It may be partially reproduced only on condition that Digimind is quoted as the author and that the following address is indicated: http://www.digimind.com. For further information please contact Digimind by email at contact@digimind.com or by phone on +33 (0)4 76 70 13 78.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>WARNING</td>
<td>1</td>
</tr>
<tr>
<td>FOREWORD</td>
<td>3</td>
</tr>
<tr>
<td>AUTOMATED CATEGORIZATION: STATE-OF-THE-ART TECHNOLOGY</td>
<td>4</td>
</tr>
<tr>
<td>What is automated categorization?</td>
<td>4</td>
</tr>
<tr>
<td>Existing categorization technologies</td>
<td>4</td>
</tr>
<tr>
<td>Bayesian networks</td>
<td>5</td>
</tr>
<tr>
<td>Decision trees</td>
<td>5</td>
</tr>
<tr>
<td>The k nearest neighbors</td>
<td>6</td>
</tr>
<tr>
<td>DIGIMIND CATEGORIZER</td>
<td>7</td>
</tr>
<tr>
<td>Description of the Digimind Categorizer module</td>
<td>7</td>
</tr>
<tr>
<td>Digimind Categorizer joins the ranks of the world's best technologies</td>
<td>8</td>
</tr>
<tr>
<td>AUTOMATED CATEGORIZATION FOR COMPETITIVE INTELLIGENCE</td>
<td>9</td>
</tr>
<tr>
<td>Categorization &amp; its contribution to Competitive Intelligence</td>
<td>9</td>
</tr>
<tr>
<td>Analyze competitor strategy using automated categorization</td>
<td>10</td>
</tr>
<tr>
<td>How Digimind Evolution applies the technology</td>
<td>11</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>14</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>15</td>
</tr>
</tbody>
</table>
Foreword

Thanks to the soaring increase of information available online, automated text categorization is fast becoming a key technology for intelligence management for companies, both internally and externally.

Yet it remains a complex scientific and technical domain demanding expert knowledge of text and language processing technologies.

This white paper aims to explain the objectives, operating methods and advantages of automated categorization within competitive intelligence activities. It presents Digimind's approach in creating the technology behind Digimind Categorizer.

It is important to emphasize that this white paper focuses on the categorization of information for the purposes of competitive intelligence. Indeed an automated categorization algorithm is of little interest in itself but is useful for handling a certain type of information. For example, an algorithm may be excellent at processing spam, yet completely inapt at classifying competitive intelligence information.
Automated categorization: state-of-the-art technology

What is Automated Categorization?

Automated text categorization is the act of classifying documents automatically according to certain criteria (the text’s subject matter, its style.). Over the last 10 years it has become the subject of renewed interest. This is essentially due to the boom in the availability of digital documents and the need to organize them rapidly. It is a relatively old science, dating back to the 60s, which made considerable advances with the appearance in the 90s of a new generation of algorithms, much more powerful than their predecessors. The technology has been successfully applied in the field of automated classification of documents and email, or anti-spam filters.

Until the early 80s, major resources in human terms were required to create a classifier. Several experts were needed to write the rules manually and then refine them throughout the testing period. The advent of machine learning therefore brought with it a dramatic cut in the time needed. There is no longer any need, for instance, to reconfigure the whole system if the classification system should change.

Categorization is now a reliable tool due to such technological developments and advanced algorithms.

Existing Categorization Technologies

Over the years, automated categorization technologies have used a succession of different algorithms, with each new generation improving on the one that went before. Some of the most longstanding technologies we could mention are semantics-based approaches, which had the disadvantage of high costs in human and financial terms if the classification system required updating. To counter this problem, several companies then introduced new automated categorization
technologies which were purely statistics-based, supported by Bayesian networks. Several generations of statistics-based algorithms followed, which were able to produce ever more relevant results. Below is a presentation of some statistics-based processing methods.

**Bayesian networks**

Naïve Bayes text classification is a method based on learning from a set of training documents. The classification is set up according to probabilities of series of words as they appear in the training documents. It is therefore crucially important that the training documents are representative of the data that will be handled later.

The algorithm uses Bayes' rule to estimate the probability $P(c_j|d_i)$, based on the probability of the category, and the probability that the words comprising $d_i$ belong to $c_j$. It is presumed that the probabilities of these words appearing are independent of each other, although this may not be the case. This is not necessarily the most efficient strategy, as there is a risk that the probabilities might be incorrectly estimated if there are too few examples to work from.

Further experimentation on Bayesian networks did not produce conclusive results. Bayesian networks are effective for binary classification (determining between spam and not spam for example) using a learning sample consisting of several hundred elements. But once the number of possible categories is multiplied and the set of learning documents consequently smaller, these types of networks are of relatively little relevance: attaining efficiency rates between 20 and 40%.

**Decision trees**

These are tree structures wherein leaves represent classifications and branches represent conjunctions of features, or nodes, that lead to those classifications. A decision tree classifier is learned recursively: at each level a function looks for the most distinguishing feature (a term in this context) and this is then used as an attribute value test for the node. Next, the examples as a whole are separated into several subsets according to the attribute value. This process is then repeated at each level. The leaves can therefore indicate a document's category with a reasonable degree of probability.
The k nearest neighbors

This algorithm does not build a classifier as such. When a test example is presented, it finds the nearest k documents using a measure of similarity. It assigns a majority category to it, or a weighted category according to the distance separating it from the examples.
Digimind Categorizer

DESCRIPTION OF THE DIGIMIND CATEGORIZER MODULE

Digimind Categorizer is one of the constituent blocks of the Digimind Prism development framework, a veritable tool box containing all the technologies needed to build a made-to-measure competitive intelligence solution. Each component is available either individually or as part of the comprehensive package.

Digimind Categorizer uses an automated categorization technology that can detect major themes and file the alerts.

Digimind Categorizer's first step is to constitute a learning database from a set of training documents pre-classified in taxonomy formation. This learning database can for instance use Digimind Evolution documents as its training documents.

As each new piece of information is presented, Digimind Categorizer automatically identifies the language used and then applies a series of preliminary linguistic processes (lemmatization, stemming, deletion of blank words, calculation of repetition, etc.). It then uses algorithms derived from Support Vector Machines to categorize the information.

If the documents' representation space is linear separable, the SVM-based method looks for the separating hyperplane furthest (in distance) from the training documents. Categorization then occurs by positioning the new document in the space and identifying which "side" it is on compared to the hyperplane.

SVMs are very effective for categorizing texts with many attributes, as they are not overly sensitive to excessive learning phenomena.

As we will see in the next section, studies on the subject by various research teams have demonstrated that this approach can attain optimal relevance rates.
DIGIMIND CATEGORIZER JOINS THE RANKS OF THE WORLD’S BEST TECHNOLOGIES

It is difficult to assess and compare the different automated categorization methods. The majority of research articles use the Reuters training corpus for the learning and testing phases. This document base consists of more than 10,000 news items, forming a solid basis for conducting serious studies. Since there are several versions of this document base, comparison is not an easy task. It is also important to have reliable performance indicators with which to measure classifier performance.

The teams at Digimind have therefore restructured the training corpus and the categories within it, in order to conduct tests using topics appropriate to the needs of competitive intelligence.

Different algorithms have been tested and their success rates measured on this new set of training documents. A rate of 100% would indicate that all the information was filed correctly, i.e. in the same folder as in the pre-classified training documents.

![Graph showing success rates of different algorithms](image)

▲ Relevance feedback from the main categorization algorithms.

The algorithm used by Digimind is on a level with the best world players and exceeds Bayesian network technology performance.
Automated categorization for competitive intelligence

Categorization & its contribution to Competitive Intelligence

Ten years ago strategic information was a rare commodity, hard to come by. Intelligence professionals concentrated their efforts and resources on looking for information. Internet was just a minor source with major database suppliers providing the main pool of data, and professionals’ prime focus was on information from the field.

The exponential growth of relevant information available on the Internet is having a radical effect on competitive intelligence practices. Today’s main challenges are the identification of relevant sources and the capacity to handle large volumes of information once sourced, in order to gain a better understanding of the changing environment and conduct analyses.

Not all aspects of intelligence work are affected by these developments. There are still many subjects for which electronic sources are few and far between and the volume of data limited. There is little or no point in applying automated categorization in such cases as these.

For areas that generate a major flow of information, however, it would be difficult to imagine a competitive intelligence technology that is not equipped with automated categorization. By automating text and data classification, categorization eliminates the need for people to validate low added-value information, freeing up their time to focus on analysis instead.

Categorization is thus able to industrialize part of the intelligence process that has until now been time-consuming and resource-hungry. Analysis comes is given due priority and the intelligence project’s return on investment is optimized.
ANALYZE COMPETITOR STRATEGY USING AUTOMATED CATEGORIZATION

The standard version of Digimind Categorizer 2.0 includes training documents relevant for analyzing organizational strategy. This corpus was based on the IPTC (International Press Telecommunications Council) standard, but the classification was simplified and completed in order to include only categories that are applicable to the analysis of organizational strategy.

It incorporates several areas which correspond to the main functions within companies:

- Human resources
- Research and development
- Marketing
- Finance
- ...

As soon as information is detected by the competitive intelligence software, Digimind Categorizer can determine whether it concerns the launch of a new product, or the nomination of a key person, the opening of a new plant, etc. At the user’s request, it can automatically file the information into the appropriate folder or simply suggest where it should be filed.
The main analysis categories provided as standard in Digimind Categorizer 2.0

The analysis of a competitor strategy is just one example of how Digimind Categorizer can be applied. It can be extended to handle highly specialized categorization issues:

- **Per business sector**: pharmaceutical industry, energy, telecoms, banking, insurance, ministries, etc.
- **Per profession**: marketing, customer service, strategy department, etc.

The analysis of each new domain simply requires a new collection of documents from which to build the training corpus and this is a relatively rapid and low-cost task.

**How Digimind Evolution Applies the Technology**

Digimind Categorizer is available as a Web Service, compatible and able to work alongside all formats, applications and software programs that operate on all types of platforms.
It is also available as an autonomous module that can interface with all types of documents and information flow.

**Types of information**

Digimind Categorizer can be used on all databases and indexes, simplifying the task of information retrieval by directly filing the results in the user-defined categories.

**News feeds**

Digimind Categorizer can interface with all types of RSS and XML feeds, which it automatically enriches in metadata. The feeds are filed in the user's categories as they arrive.

Last but not least, the application improves productivity as highly relevant information can be consulted directly in the folders, pre-classified according to user-defined topics.

For example, a user monitoring an RSS feed from a competitor's site had to track the source continuously in the past, but can now choose to be alerted only when press releases mention new product launches.

**Integrated in Digimind Evolution**

Digimind Categorizer aims to facilitate the task of validating information.

Although it is available as an autonomous module, its full potential is put into play when operating within Digimind Evolution, the leading competitive intelligence platform on the market. Here it enables users to validate alerts automatically as they are received.

Several options are open to the user:

- Alerts are systematically validated in a particular folder
- Alerts are validated according to key words
- Alerts are automatically validated

- **Systematic classification:** during surveillance agent configuration, one or several folders are specified as the target locations for automatically filing the agent's alerts.
This type of automated validation is the simplest but also the most categorical, in the sense that all the alerts will be filed as determined by the user, even if they do not bear any relation to that folder.

**Classification according to key words**

This type of validation has the most potential to benefit users, but requires additional work when creating and configuring folders.

The user must first link key words, or a query, to the folder(s) where he or she would like the alerts to be classified.

As an alert arrives, the automatic validator applies all the folders’ queries to the alert. After comparison, the alert will be validated by the appropriate query and automatically filed in the associated folder.

The outcome is an automatic filing system for alerts which is flexible and relevant thanks to folder-specific queries.

**Classification using the categorization engine**

This uses the documents’ automated categorization engine to identify where the alert should be validated depending on the names of the users’ folders. Only one folder will be indicated for each classification.

This method is the simplest automatic validation of the three. The alerts are automatically validated in the most relevant folder.
Conclusion

Automated categorization technology need not be systematically integrated into each competitive intelligence project. It is only appropriate if the volume of information to be handled exceeds several hundreds or thousands of items per day.

We would always recommend taking great care when selecting sources for monitoring, and filters must be correctly configured to ensure that the information retrieved is of high quality and not too voluminous, that is to say an amount that can be handled by the user.

It is, however, increasingly common to see competitive intelligence projects focusing on topics that generate an information flow that does justify the deployment of an automated categorization solution.

This latest version of Digimind Categorizer heightens the relevance of categorization to join the ranks of the best technologies worldwide. Furthermore it provides a customized corpus of training documents ready for use in the field of competitive intelligence.
References

Study by Jon Udell on the use of Bayesian networks for automated categorization of content:


