

Classification, Clustering and Information Visualization with Oracle Text

An Oracle White Paper
February 2003

Classification, Clustering and Information Visualization with Oracle Text

Introduction	3
Classification	3
Classification Training.....	4
Clustering.....	4
Information Visualization	5
Development	6
Conclusion.....	6

Classification, Clustering and Information Visualization with Oracle Text

INTRODUCTION

Lately there has been a lot of debate about new features for search systems in general. Search engines technology has moved from a first generation that focuses on refining a hit list to a second generation that focus on relevant information. Classification, clustering, and visualization are at the core of the new generation.

Many small companies are claiming that they have “killer applications” in this area. Oracle sees potential value in classification, clustering, and information visualization as part of a large search platform highly integrated with the database.

In this paper we will present some of the advanced features of Oracle Text that you can take advantage of improving existing search systems or building new applications.

CLASSIFICATION

A document classification application performs some action based on document content. Actions can be assigning category ids to a document for future lookup or sending a document to a user. The result is a set or stream of categorized documents. Documents that are in the same category are more similar to each other than documents in different categories.

For example, assume we have an incoming stream of news articles. We can define a rule to represent the category of Finance. The rule is essentially one or more queries that select document about the subject of Finance. The rule might have the form ‘stock or bonds or earnings’.

When a document arrives about a Wall Street earnings forecast and satisfies the rules for this category, the application takes an action such as tagging the document as Finance or emailing one or more users.

To create a document classification application, you create a table of rules and then create a CTRRULE index. To classify an incoming stream of text, use the MATCHES operator in the WHERE clause of a SELECT statement. Figure 1 shows the general flow of a classification application.

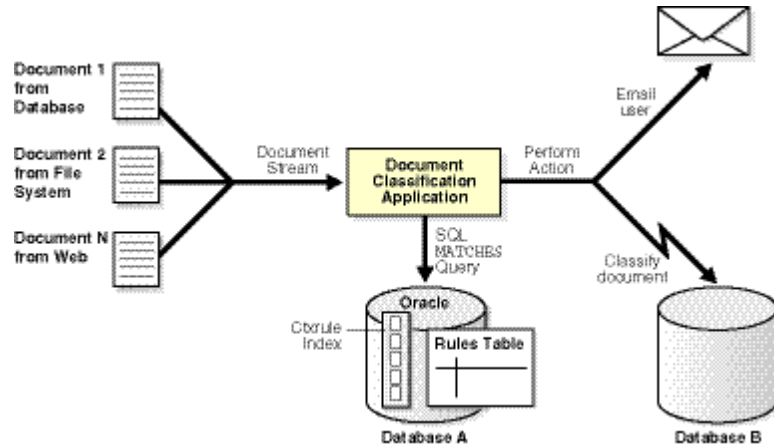


Figure 1 – Overview of a Document Classification Application with Oracle Text.

Classification Training

In the previous section we describe the classification process where for each category there are rules that explain why the documents belong there and using those rules the application classify incoming documents into categories.

The process that defines the rules can be manually which is very accurate but also time consuming and doesn't scale for large collections and categories. Or it can be automatically using Oracle Text's training features that will statistically analyze document groups and automatically generates CTXRULE-compatible rules. The user has to supply a training set consisting of categorized documents and each document must belong to one or more categories.

The benefits are:

- No need to control or supervise large collections.
- Categorize the documents using a statistical approach.
- Leverages existing knowledge for providing the training set

CLUSTERING

As opposed to classification, clustering is the unsupervised division of patterns into groups. Oracle Text offers the CTX_CLUSTER package for building clusters. This package automatically clusters a set of documents according to their semantic meanings. The interface allows users to select the appropriate clustering algorithm. Each cluster contains a subset of documents of the collection. A document within a cluster is believed to be more similar with documents inside the cluster than with outside documents.

The clusters can be used for building features like presenting similar documents in the collection.

The benefits are:

- Automatic discovery of patterns in the collection.
- Useful for identifying categories from the collection.
- Useful for building abstractions (for example: more like this).
- Provides a statistical snapshot of the collection.

INFORMATION VISUALIZATION

Information visualization is defined as “visual representations of abstract data to amplify cognition”. In the context of vast amounts of information, visualization techniques can help users navigate through large sets of documents as well as selecting the appropriate asset. Figure 2 shows a Java implementation of the *stretch viewer* visualization metaphor to browse a subset of the Medical Subject Headings (MeSH) categories.

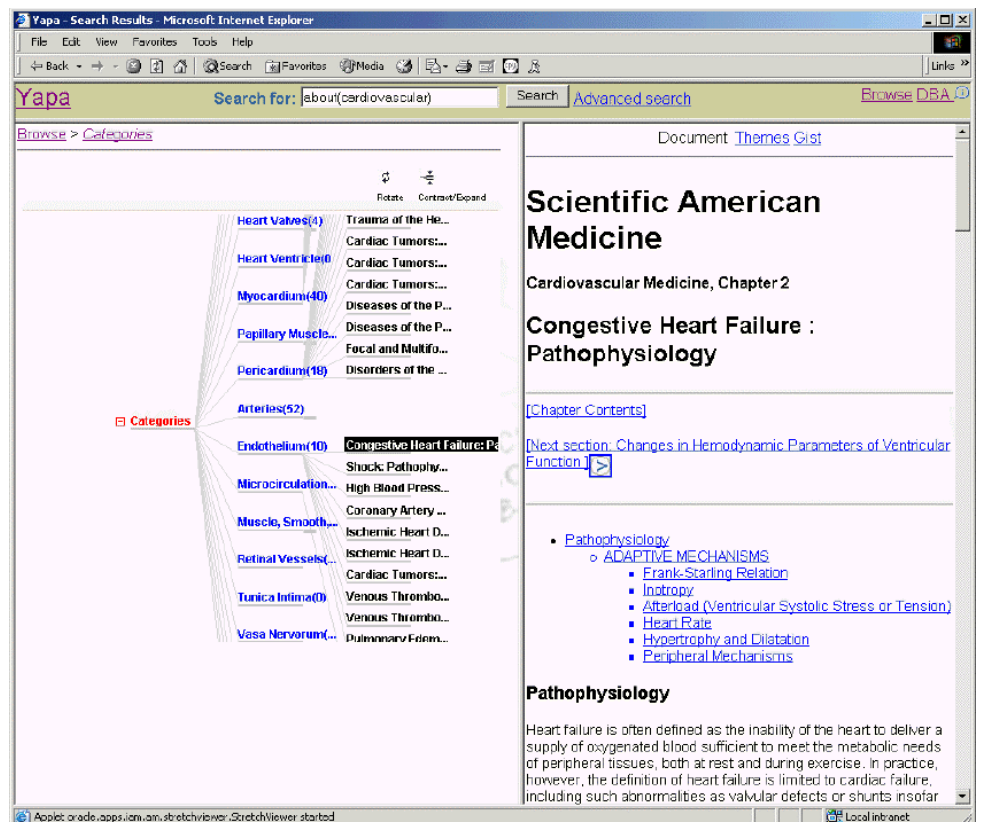


Figure 2. The stretch viewer visualization for browsing categories.

DEVELOPMENT

Oracle Text has an API that application developers can use to construct new systems or embed particular features in other systems.

Oracle Text also provides code generator wizards as part of the JDeveloper9i IDE environment. Users who are not entirely familiar with the API or who want to reuse components can take advantage of these mechanisms.

CONCLUSION

Oracle Text offers a complete technology stack for content classification, clustering, and information visualization. These advanced features can be combined with other Oracle Text characteristics like XML searching, document services, Boolean operators, etc.

Oracle Text has all the normal benefits of an industrial-strength database available, without the cost of learning and supporting extra APIs and duplicated data. The ability to find, classify, cluster, and visualize documents based on their textual, content metadata, or attributes, makes the Oracle database the single point of integration for all data management.

FURTHER READINGS

1. Oracle Text Reference Guide. Oracle Corp., Redwood Shores, CA (2002).
2. Oracle Text Application Developer's Guide. Oracle Corp., Redwood Shores, CA (2002).
3. S. Alpha *et al.* "TREC-10 Statistical classification (adaptive and batch) and question-answering"
(<http://trec.nist.gov/pubs/trec10/papers/orcltrec10.pdf>)
4. Oracle Text Home Page (<http://technet.oracle.com/products/text/>)



White Paper Title
February 2003
Author: Oracle Text Team
Contributing Authors:

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com

Oracle is a registered trademark of Oracle Corporation. Various product and service names referenced herein may be trademarks of Oracle Corporation. All other product and service names mentioned may be trademarks of their respective owners.

Copyright © 2001 Oracle Corporation
All rights reserved.