Automatic Document Categorization
A Hummingbird White Paper
## Table of Contents

The Challenge .................................................. 5
The Solution .................................................... 5
Technology Overview ........................................... 5
Categorization with Fulcrum Knowledge Server ................. 6
Why Neural Networks ........................................... 6
Changing Subject Matter ....................................... 7
Classifying New Documents .................................. 7
Classification and Re-indexing ................................ 7
Scalability ......................................................... 7
Classification According to Pre-defined Categories .......... 8
Administrator Notes: Knowledge Map to Cluster Map ....... 8
The Challenge
The tremendous amounts of information knowledge workers receive and create present two difficult problems. The first is how to organize or classify the information to create an information taxonomy. The second is how to ensure that new information is filed correctly in this taxonomy.

Commonly, organizations try to solve these problems by setting up systems of folders and subfolders. But sometimes new documents are not filed where they should be. And, over time, individuals will tend to create new folders of their own if they can’t find the right folder quickly enough. Before long, the number of folders grows and the categorization system collapses.

To compound the problem, many people have difficulty with traditional information retrieval. Sometimes it is because they don’t know precisely what they’re looking for. Other times it is because of their difficulty translating their need for information into a query a computer can understand.

The Solution
Hummingbird offers a solution to these problems through Automatic Document Categorization, an advanced document concept mapping technology. This technology automatically generates taxonomy based on document contents. At the same time, it provides a procedure for assigning new documents to the taxonomy.

Categorization technology looks inside documents to learn what they are about. It organizes them into an Enterprise Table of Contents™: a browse-able library that can be used by anyone who has ever read a book or used Yahoo!™. With no additional training, users can browse through information to identify the richest sources and discover work that has been done. The result of this more intuitive process is a much higher return on the investment an organization has made in the information it possesses.

An organization may choose one concept map, or many. It depends on how, and by whom, the information is to be used. One taxonomy may be created for sales and another for engineering. Since Automatic Document Categorization works with documents’ virtual, not physical, locations, administrators can offer departments their own views of their documents without affecting the running of the document repository.

Technology Overview
Categorization relies on a neural network technology that organizes documents by content. It is closely related to data mining, in which computational processes and techniques seek patterns in large quantities of structured data. In this case, the data mined is the unstructured content of documents.

Hummingbird’s Automatic Document Categorization involves three stages:

1. Occurrences of terms (words and phrases) in documents are analyzed to find those that are not useful. These include terms that are too specific to some documents, those whose distribution suggests they are typos or OCR errors, and
Automatic Document Categorization

those that are too common to provide insight into the documents’ subjects.

2. Concepts—patterns of terms—are then extracted from the documents using a
neural network algorithm. A concept may contain one or more terms that appear
in another concept and each document may contain more than one concept. By
comparing the set of concepts in one document with the sets of concepts in other
documents, it is possible to show how similar that document is to the others.

3. Documents are then organized in a tree structure according to subject. A first
pass groups documents with similar concepts. Second and subsequent passes
create subgroups. The product is a Cluster Map. The system can create this map or
it can be created to match a previously organized taxonomy of documents.

Categorization with Fulcrum KnowledgeServer

Categorization is integrated with Fulcrum KnowledgeServer, Hummingbird’s flagship
knowledge management product. It is also compatible with Hummingbird EIP, which
includes Fulcrum KnowledgeServer.

Administrators use the Knowledge Manager Workstation to build and maintain
Cluster Maps. Features include the following:

• extract a sample set of documents from repositories that may include file systems,
  Microsoft Exchange, Lotus notes, databases, document indexes and other informa-
tion sources available to Fulcrum KnowledgeServer
• develop a taxonomy from these documents
• monitor, understand and edit the resulting structures
• publish a Cluster Map — the generated taxonomy for the documents and a clas-
sification algorithm — to the Knowledge Server.

Publishing a Cluster Map to a Knowledge Server ensures the following:

• Documents are indexed and automatically classified into one or more subject
groups as determined by the classification algorithm. The resulting classification
(subject groups) is included in Fulcrum KnowledgeServer’s internal catalog.
• Through a Web interface end-users can browse the concept map, select subject
groups and view associated documents. Subject groups can also be used to
refine searches.

Why Neural Networks

Many of the computational techniques used in neural networks are closely related
both to conventional statistics and to other computational grouping algorithms. So, it
is reasonable to consider Fulcrum KnowledgeServer’s Categorization technologies as
non-linear developments of well-understood principles.

That said, neural network techniques are generally better at automatically generating
taxonomies than other approaches. In addition, although this approach is computation-
ally intensive when taxonomies are developed, it is efficient at processing queries and
classifying documents online.
Changing Subject Matter
As new documents are filed, a Cluster Map will need to change. Hummingbird offers an approach characterized as static with additional periodic reorganization as its solution.

A static approach allows users to learn how the system has organized the information in it. Familiarity breeds productivity. This approach also cuts the need to reclassify documents often — saving system resources. In addition, it offers a longer lifetime for the use of sub-taxonomies created for departments.

Over time, an organization’s knowledge base may change, using different terminology and introducing new topics into the system. This is usually characterized by a growing number of documents in the “Unclassified Documents” category in Fulcrum KnowledgeServer. When this happens, the Knowledge Manager can refresh the Cluster Map by re-training it with new sample documents. This is done as an offline process, so the production server is not affected. The new Cluster Map is then uploaded to the system as a simple file-copy operation, after which it will be available for production use. The indexing engine will reclassify documents at the next scheduled indexing time to reflect the new Cluster Map.

Classifying New Documents
The steps in classifying new documents are as follows:

- The system reads new documents to describe their contents or it reads descriptions that have already been generated for previously read documents.
- The initial document description is processed through the categorization procedure to point to one or more categories in the taxonomy to which to assign the document.
- The assignment is recorded in the document database and indexed to ensure the categories are available to users during searches.

Classification and Re-indexing
The Categorization procedure is integrated with the indexing process. New documents are indexed and classified at the same time.

Hummingbird has adapted Fulcrum KnowledgeServer’s underlying indexing engine, Fulcrum SearchServer, to ensure that the re-indexing required for the classification of old documents is as fast as possible. To provide for this, the initial description used in classification is stored in every document’s catalogue. A read of this field is used to generate the appropriate category. The field is present in all documents indexed by Fulcrum KnowledgeServer 3.0 (known as DOCFulcrum) or later.

Scalability
To develop a taxonomy, Fulcrum KnowledgeServer can sample all the documents it has indexed, but the recommended sample is 10% for collections of up to 1 million documents. The Categorization technology can deal with a sample of up to 100,000
documents; larger samples are impractical due to the length of time it would take to analyze larger collections. Usually there is little added value to be gained from sample sets larger than this, anyway.

For larger collections, it makes sense to develop independent taxonomies along subject lines. The next step would be to create and analyze samples from these taxonomies. One or more concept maps may then be applied to either the entire collection of, if physically separate, to the individual parts.

**Classification According to Pre-defined Categories**

Fulcrum KnowledgeServer’s Categorization technology allows for the classification of documents into predefined subject groupings (called “Supervised” Clustering). The first step is to place documents into physical folders by subject. Second, organize these physical folders into the desired taxonomy. The software will then build a classification procedure to duplicate this structure.

**Administrator Notes: Knowledge Map to Cluster Map**

**Introduction:** Fulcrum KnowledgeServer provides the environment in which the concept map structure is created. It provides a powerful means of achieving single access to multiple repositories plus a single means of browsing and searching these repositories. Concept maps are provided within Fulcrum KnowledgeServer to enable browsing and searching using subject-based hierarchies.

**Sample documents:** To develop a concept map a collection of documents to analyze is needed. This can be any group of file system documents or a sample of a Knowledge Map extracted using the sampling application. The administrator specifies the parts of the Knowledge Map to sample and the proportion of documents to be included in the sample. The sampling application extracts a random selection of the documents and places them in a file system folder. The documents are converted to html format.

**Indexing the sample set:** The collection of documents to be analyzed is indexed using Fulcrum KnowledgeServer facilities through Knowledge Manager Workstation.

**Building a concept map:** Cluster maps are built using Knowledge Manager Workstation. It leads the administrator through the steps of building a dictionary, extracting the term themes and building the subject tree. The administrator can examine the structures the system has produced, and browse and read documents in the subject categories. The administrator may then edit the automatically generated labels for each subject area and reorganize the subject hierarchy.

**Importing a Cluster Map into Fulcrum KnowledgeServer:** Once the Cluster Map is complete, it can be imported into Fulcrum KnowledgeServer. The Cluster Map is then available for classifying parts or the whole of the Knowledge Map. To do this, the Fulcrum KnowledgeServer administrator assigns one or more Cluster Maps to each index—or this can be performed in reverse, by assigning each index to one or more Cluster Maps.
Classifying documents using a Cluster Map: If a Cluster Map has been associated with an index, every new document added to the index will be automatically classified when it is indexed. When a Cluster Map is first associated with an index, all previous documents will be classified using an efficient mechanism that does not require re-indexing the original documents.

Searching and browsing using a Cluster Map: Client programs used to search document repositories through Fulcrum KnowledgeServer present the Cluster Map subject hierarchies to the user in addition to the Knowledge Map (the physical locations of the documents and the manually organized logical view of these physical locations). Users can search using any combination of physical, logical, subject-based and search query restrictions they desire. For example, they may wish to retrieve all the Exchange documents in one particular subject area or they may wish to view all file systems documents containing the word ‘gold’ in several subject areas. They are then able to combine browsing and searching.
About Hummingbird
Hummingbird is a leader in the development of enterprise software solutions, which provide access to all business-critical information and resources. Hummingbird EIP™ (Enterprise Information Portal) leverages the company’s core strengths in network connectivity, data integration and reporting, and document and knowledge management to connect users to all the business information they need, aggregated and categorized through a single user interface. With a diverse product portfolio, Hummingbird offers complete global enterprise solutions from advanced host connectivity, through sophisticated data exchange, business intelligence and analytic applications, to powerful information management at the desktop or on the Web. Headquartered in Toronto, Canada, the company offers its products, along with related consulting, education, and support services, in more than 50 countries around the world.

Analytic Applications
Analytic Frameworks is a series of industry specific analytic applications that includes components for risk management, profitability reporting, customer analysis, sales forecasting, and quality standards compliance. Initial applications will be directed towards the financial services, telecommunications, and health services industries.

Data Integration and Reporting Solutions
Hummingbird Data Integration and Reporting solutions enable organizations to maximize the value of structured data, allowing them to make better-informed business decisions.

Document and Knowledge Management Solutions
Our solutions enable the enterprise-wide access, analysis and control that organizations need to gain maximum value and competitive advantage from their unstructured information sources.

Enterprise Information Portal Solutions
Hummingbird EIP is a Web-based solution that bridges the disparate worlds of structured and unstructured information across the enterprise. Leveraging Hummingbird’s entire suite of products, the Hummingbird EIP provides knowledge workers with the tools needed to access, extract, and analyze information across the entire enterprise, then act on what they have found.

Network Connectivity Solutions
Hummingbird’s Network Connectivity solutions provide the core technologies to deliver business-critical information and resources to all Windows and e-based desktops to meet organizations’ unique e-business needs.

Corporate Headquarters
Hummingbird Ltd.
1 Sparks Avenue
Toronto, Ontario M2H 2W1
Canada

Toll Free Canada/USA:
+1-877-FLY-HUMM (359-4866)
Tel: +1-416-496-2200
Fax: +1-416-496-2207
E-mail: getinfo@hummingbird.com

North American Sales Offices
Atlanta • Boston • Chicago • Dallas • Los Angeles • New York
Ottawa • Phoenix • Raleigh • San Francisco • Toronto • Washington DC

International Sales Offices
Australia • Belgium • France • Germany • Italy • Japan • Korea • Mexico
Netherlands • Singapore • Sweden • Switzerland • United Kingdom

For more information, visit www.hummingbird.com

Copyright © 2000 Hummingbird Ltd. All rights reserved. Analytic Frameworks, Enterprise Table of Contents, Hummingbird, and Hummingbird EIP, trade dress, manuals, and logos are the intellectual property of Hummingbird Ltd. All other company and product names are trademarks of their respective owners.