

HILBERT[®] QUANTUMTEXT OVERVIEW



The “Next Generation” Search Experience – Extracting Intelligence From Document Content

The wealth of information that can be harvested from unstructured data is often critical to effective search or decision-making, yet the varying nature of unstructured data, coupled with the sheer explosion in volume, represents a monumental challenge to any information research endeavor.

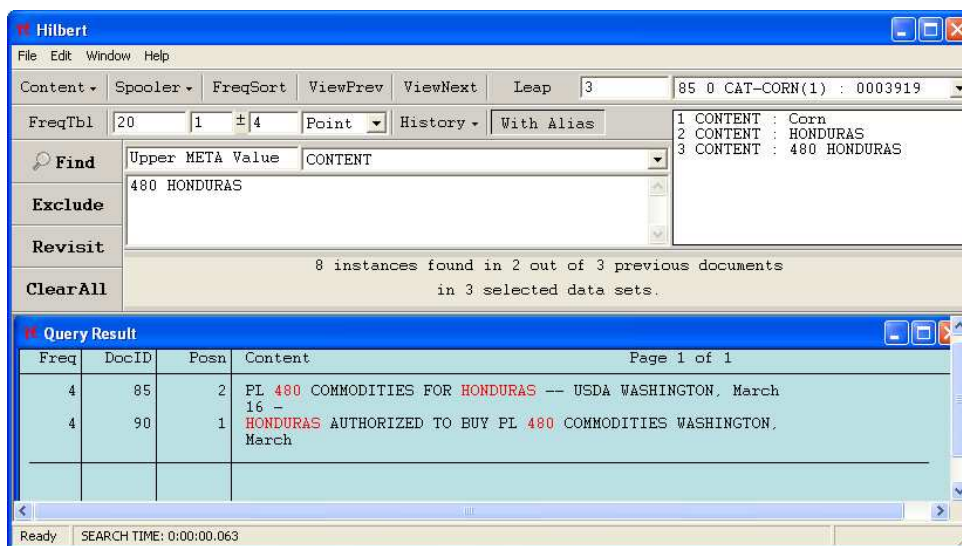
The Challenge

“Searching” implies the analysts know where they are heading in their investigation and where to begin. In many investigations, however, the path to the answer is not clear. What analysts need is a system that can guide them through an overwhelming amount of data to discover the information they seek – a system that merges high performance search with text mining functionality and categorizing capabilities.

The Solution – Searching + Categorization

Searching

We have come to accept the search process as one of entering a few search terms and then navigating through 10,000+ responses. For analysts and researchers who are trying to find “needles in haystacks”, this is not a productive process, and the path to the right answer is not always clear. What analysts need is a system that can guide them through an overwhelming amount of data to discover the information they seek – a system that merges high performance search with text analytics that reveals the most productive paths to follow. Hilbert QuantumText is uniquely designed to do just that.



Why Hilbert?

We are the only solution provider for ultra high-speed access, manipulation, storage and analysis of large volumes of structured and unstructured data.

Speed, Flexibility and Scalability have been the design criteria of the Hilbert solution.

The Hilbert applications are offered as industry or process specific solutions.

Bjorn Gruenwald, CEO and Founder of Hilbert Technology Inc.

Support for different text file types (MS Word, Power Point and Excel as well as Adobe PDF, XML, HTML, MS Outlook emails, ...)

Increased end-user efficiency in exploring new and unused information

Narrow and Drill down searching

Higher efficiency through more accurate specification and limitation criteria

Ultra High-Speed access

Higher user buy-in
Higher end-user productivity

Built upon the patented, numerical processing and storage technologies of the Hilbert Engine, it provides the facilities to guide investigators through the complex maze of data that defines their environment. Some of the highlights include:

- **Drill Down Search**

For many researchers, the answer to one query will LEAD to the structure of the next; certainly in FORENSIC-type analysis this is the norm. Unlike conventional indexed search engines that offer a one-shot query (SQL-like) against the corpus, Hilbert QuantumText provides an unlimited drill down search process that delivers a highly interactive, investigative method of searching. The results of each query are immediately available for the next – you do not have to rethink and redesign your first query. This gives the analyst a hands-on feeling of “driving” through the data with the ability to “steer” through the content, each query result guiding the next step of the investigation.

- **Alias-Lists**

The user defined Alias-Lists specify which words are used as an alias for others equivalently, i.e. board = director = CEO. Each of the equivalent words are then used the same way as the alias (includes all synonyms).

- **Ignore-Lists**

The user defined Ignore-Lists are used to suppress commonly used words (the, a, and, ...). Used case specific these Ignore-Lists can be used to find the specific differences in documents with a high degree of affinity.

- **Frequency Sort**

This feature organizes documents based on the quantity of occurrences of the query term(s), providing one form of relevance to help guide the researcher.

- **Word Histograms**

Often, analysts are just not sure what to look for next. Why not let the data itself guide that decision? At every level of investigation, Hilbert QuantumText can generate a word frequency list, i.e. a list of every word in alphabetical order that occurs in the current result set, showing how many times each word occurs and in how many different documents each occurs. This view of the data can give insight and guide the next step of the query process. A person's name that occurs many times, a common city mentioned, an email sent to a strange address, at a frequency that seems unusual – this is the kind of information that can drive the investigation forward in the right direction and save hours of hit or miss querying.

Categorization using Reference Documents instead of keywords

Higher accuracy of categorization

User friendly set-up of categorization environment and parameters

Scoring and affinity calculations

Higher degree of objectivity in classification and categorization

Elimination of ‘false-positive’ scores often introduced by buzzword overload

Scalable to millions of documents

Available in desktop, net-worked or server configurations

Categorization

Researchers often need to find documents that are similar to one(s) they have identified. Key word searching assumes a particular word or phrase must be present, which is too often not the case. Hilbert QuantumText provides an accurate, easy-to-use document categorizer that automatically identifies a relevant set of documents by comparing a selected document or set of documents to the entire corpus or any subset. Hilbert QuantumText automatically ranks each document according to its relevancy to the subject document(s) for up to 10 categories at once. This high accuracy categorization process saves the analyst from having to read through hundreds of documents to determine correlations and patterns in text.

The screenshot shows the Hilbert software interface. The top window, 'Hilbert Categorizer in Categorization Mode', has a menu bar (File, Edit, Window, Help) and a toolbar with buttons for 'From Browsed', 'From Queried', 'Load Category', 'Load Corpus', 'Run', and 'Report Setup'. A 'Contrast: 3' dropdown is also visible. Below the toolbar is a 'Browse' field containing the path 'C:\SifTextPlus-New\CAT-corn\'. A table below lists categories:

Category Name	Category Data Folder	Category Source Folder	No. of File	Category ID
CAT-Corn	C:\Program Files\Hilbert\SifText Plus\Categorize...	FROM SIFTEXT DATABASE	181	0
CAT-Crude	C:\Program Files\Hilbert\SifText Plus\Categorize...	FROM SIFTEXT DATABASE	389	1
CORPUS	C:\Program Files\Hilbert\SifText Plus\Categorize...	FROM SIFTEXT DATABASE	117	10000

The bottom window, 'Hilbert Categorization Report', displays a table with the following data:

Filename	CAT-Corn %1	CAT-Corn %2	CAT-Corn S1	CAT-Corn S2	CAT-Corn Relv
trade-0012396.txt	96.5	94.4	277	106	134410
trade-0012346.txt	96.2	91.5	503	106	112547
trade-0009622.txt	92.3	93.1	303	167	86445
trade-0012249.txt	96.7	93.6	296	87	63166
trade-0010917.txt	90.7	86.7	162	59	26558
trade-0010898.txt	90.7	86.7	162	59	26558
trade-0012318.txt	93.9	81.5	232	43	25967
trade-0012325.txt	91.7	91.7	93	67	21740

The status bar at the bottom indicates 'Ready' and 'ELAPSED TIME: 0:00:28.844'.

Text categorization as performed by the Hilbert QuantumText provides two main functions:

- **Information Retrieval**
in which documents will be scored based on their relevance to a specific topic
- **Text Classification**
where text documents will be grouped according to their relevance to a pre-defined set of topics.

Traditionally, categories are specified by keywords and documents are 'keyed out,' one by one into hierarchical niches using rules of taxonomy. However, keywords are not always 'categorically' unique across documents, and sometimes not present in documents at all. The Hilbert QuantumText specifies categories by means of sample documents called Reference Documents, not keywords. A corpus of Foreign Documents is a set of documents to be categorized, according to the provided categories. The relevance of any foreign document to any category is determined in relation to its relevance of all reference documents in that category. This introduces the synergy of 'context,' and so makes Hilbert QuantumText a dynamic process that uses all of the documents at hand, not just a series of isolated keywords. Internal parameters are available to sharpen and broaden this categorization process.

Conclusion

Hilbert QuantumText combines high-speed search with categorization technologies to give analysts and researchers the advantage they need to quickly discover vital information. Hilbert QuantumText searches text, MS Word, Power Point and Excel as well as Adobe PDF, XML, HTML, MS Outlook emails and other text-based documents. Users quickly discover correlations in data, reducing the number of documents to only those that are relevant. The analyst using Hilbert QuantumText becomes much more efficient and effective.

Solution Areas

Hilbert QuantumText is a high performance search and discovery application essential to any industry that deals with large sets of documents. Hilbert QuantumText can effectively be applied to a variety of solution areas. These include:

Bio-Informatics	BI / Reporting	CC–Feedback/Complaints
Contract Management	University Research	Cross Database Analysis
Text and Data Mining	Data Profiling	Data Quality Improvement
E-mail/Doc search/Analysis	Forensic Research	Underwriting (Risk)
Homeland Security	Knowledge Management	Law and Litigation Support
Law Enforcement	Matter Management	Medical Records
Medical Research	News Services	Personal Desktop Search
Pharmaceutical Research	Regulatory Compliance	Sarbanes – Oxley
Streaming Data Analysis	Maintenance/Warranty	

Supported File Types

- Adobe Acrobat: PDF
- Microsoft Office: Word, Excel, PowerPoint (DOC, XLS, PPT, DOT, ...)
- Microsoft Office 2007: DOCX, XLSX, PPTX
- Standard-Text: RTF, TXT (ASC, LOG, PRN, ...)
- Outlook and Outlook-Express: PST, DBX, with Attachments
- ZIP-Compressed Folders (Archived Data)
- Web Pages: HTM, HTML, XML, JAVA
- Microsoft Works, Publisher: WPS, WDB, WKS, PUB
- Database Exports: CSV, TSV, TXT
- Some historical formats such as dBase and WordPerfect

Your Advantages

Narrow and Drill down searching

- Higher efficiency through more accurate specification and limitation criteria

Ultra High-Speed access

- Higher user buy-in
- Higher end-user productivity

Categorization using Reference Documents instead of keywords

- Higher accuracy of categorization
- User friendly set-up of categorization environment and parameters

Scoring and affinity calculations

- Higher degree of objectivity in classification and categorization
- Elimination of 'false-positive' scores often introduced by buzzword overload

Support of different text file types

- Increased end-user efficiency in exploring new and unused information

Scalable to millions of documents

- Available in desktop, networked or server configurations

About Hilbert

Pennsylvania based Hilbert Technology Inc. is an international provider of data management and analysis solutions for large and medium size enterprises worldwide. The offering is based on the revolutionary, patented Hilbert Engine technology for the ultra high-speed access, manipulation, storage and analysis of large volumes of structured and unstructured data. The Hilbert solutions are offered as industry or process specific solutions. Organizations in public services, law, government, finance, communications, wholesale and retail sales, transportation & tourism and chemical & pharmaceutical can employ Hilbert solutions to gain unparalleled speed in access and analysis over large data volumes.

Notes

This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor is it subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Hilbert Technology Inc.
P.O. Box 330
Newtown, PA 18940
USA
Web: www.hilbertcompany.com
E-Mail: info@hilbertcompany.com
Tel: +1 (212) 252 1600
Fax: +1 (212) 252 1615