ELSEVIER

World Patent Information 30 (2008) 280–293

WORLD
PATENT
INFORMATION

www.elsevier.com/locate/worpatin

Text mining and visualization tools – Impressions of emerging capabilities ☆

YunYun Yang *, Lucy Akers, Thomas Klose, Cynthia Barcelon Yang

Bristol-Myers Squibb, P.O. Box 4000, Princeton, NJ 08543-4000, USA

Abstract

Innovation is the underlying foundation of today's competitive economy and technological advancement. There is a plethora of text mining and visualization tools available on the market to facilitate the innovative process in uncovering "hidden nuggets" of information about emerging technologies.

A high-level overview of some key text mining and visualization tools is presented in this paper to provide a comparison of text mining capabilities, perceived strengths, potential limitations, applicable data sources, and output of results, as applied to chemical, biological and patent information. Examples of tools to be discussed include sophisticated text mining software packages, some simpler full-text searching tools, and a few data visualization tools that could be integrated with the more sophisticated software packages and full-text searching tools. Included are comments on our impressions of applicability of these tools to different types of data sources, perceived strengths, potential limitations, and suggestions as to which user groups may benefit from these tools.

Keywords: Text mining; Data visualization tools; Patent information; Competitive intelligence; Intellectual property analysis; Review

1. Introduction

Various text mining and data visualization tools have been previously described in the literature for application to patent information [1–6]. No major breakthroughs in text mining technology have been published since. However, step change improvements have been made over the last few years. For example, there appears to have been a shift in focus from statistical analysis techniques to semantic analysis algorithms. These various tools are based on standard analysis techniques and mainly differentiate in their capabilities to use different data sources and visualize in different ways (tables, maps, graphs, and matrices).

The aim of this paper is to share some of our experiences at Bristol-Myers Squibb (BMS) in learning about and evaluating select tools. We will provide only our impressions of the capabilities of these tools. It is not our intention to cover every available tool on the market nor is it to cover every aspect of the tools described in this paper. We do hope to provide some new insights into the capabilities, data sources, results, our perceived strengths and potential limitations for each of the tools discussed in this paper.

First we will very briefly describe the need for text mining and data visualization and our initial project approach. Then each of the text mining and data visualization tools will be discussed in some detail. Finally, we will provide a summary of the results of our initial evaluation of the selected tools.

The views expressed in this paper are based primarily on vendor demonstrations and discussions within the Patent Analysis Group at BMS. With over 43,000 employees, the mission of Bristol-Myers Squibb Company is to extend and enhance human life by providing the highest-quality pharmaceutical and related health care products.

☆ This article has been developed from a presentation by the authors at the PIUG 2007 Annual Conference, Costa Mesa, CA, USA, May 5–10, 2007 and the PIUG 2007 Northeast Conference, October 8–10, 2007.

* Corresponding author. Tel.: +1 609 818 4726.
E-mail address: yunyun.yang@bms.com (Y. Yang).

0172-2190/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2008.01.007

2. Why text mining and data visualization?

Most of the patent information we have today is in electronic form yet it can be just as disorganized as paper. There is simply too much information to read. In addition, information is buried and difficult to find with conventional searching tools and manual analysis. As patent analysts, we are used to manually poring over large numbers of documents to extract relevant nuggets of information. By exploring other ways to uncover, interpret and digest information, we may begin to provide answers and value, including answers to typical intellectual property management questions [7] such as

- How does our patent portfolio compare with "Company ABC"?
- How many patents are there concerning technology "X"?
- How do our invention disclosures compare with current granted patents?
- Who is citing our portfolio?
- Which patents should we consider divesting as a result of selling "Division XYZ"?
- What would a merged portfolio look like relative to the rest of industry with "Company DEF"?
- How do we improve our patent operations?

3. What are some of the common tasks?

Text mining and data visualization tools have been developed to perform some or all of the following common tasks [1]. Each tool is unique in terms of the tasks and the analytical methods it has been designed to perform.

- List generation and display (histograms).
- List cleanup and grouping of concepts.
- Producing co-occurrence matrices and other graphing.
- Clustering, categorization, grouping and extraction of text.
- Mapping document clusters or concepts.
- Adding temporal components to maps.
- Citation analysis.
- Subject/action/object (SAO), also known as natural language processing.
- Federated searching, e.g. on the intranet or intranet.

4. Project background

In early 2006, the Patent Analysis Group at BMS established a project to evaluate various text mining and data visualization tools. Phase I included gathering background information and brainstorming, identifying the potential tools to evaluate, and conducting on-site demonstrations. Phase II of the project includes piloting a few of the select tools and identifying potential clients groups and real-life case studies for using the tools. The primary focus of this paper is on the results of Phase I of the project since Phase II is still underway.

5. Investigation and process approach

At the onset of the project we first did some research to scout what had been done previously by others. Table 1 lists all of the potential vendors we initially collected. Most of them were invited for in-house demonstrations. This article covers the first 14 vendors listed in Section A of Table 1.

For each tool in Section A, we considered the basic functionality or "type" of tool, the capabilities, the data sources permitted, the results generated, our perceived strengths and potential limitations. Definitions are provided below.

5.1. Type of tool

The "type" includes whether the tool is software designed for text mining/visualization or a patent database content provider, or both.

5.2. Capabilities

The capabilities evaluated include tools performing keyword, statistical and/or linguistic analysis. Keyword analysis refers to extracting nouns or noun phases in text without understanding their meaning or relationships (i.e. out of context). Statistical analysis refers to word frequency-based analysis or counting the number of times a word appears in the text. Linguistic analysis refers to using a trained agent to do natural language processing or semantic analysis.

5.3. Data sources

The specific text mining data sources include: (1) unstructured text (such as full-text documents and emails), (2) structured text (such as database records from STN® or PubMed), and (3) hybrid content (such as patents where the front page information is structured but the remaining text is not).

5.4. Results

The output from each tool is typically generated in lists of documents, tables, charts, graphs, or maps.

5.5. Perceived strengths

Based on vendor demonstrations, we highlight key features that we perceived as strong points for each of the tools.

5.6. Potential limitations

Based on vendor demonstrations, we list certain features that we perceived as lacking or inadequate. These potential limitations are our own personal views.

Table 1
Text mining and visualization tools and their vendors

Tool name	Vendor	Vendor site URL
Section A		
Group 1 tools		
ClearForest Text Analytics	ClearForest Ltd.	http://www.clearforest.com/Technology/TechnologyOverview.asp (Note: ClearForest was acquired by Reuters in April 2007)
Goldfire Innovator™	Invention Machine Corp.	http://www.invention-machine.com/GoldfireInnovator.htm
Inxight Smart Discovery® Awareness Server Inxight Smart Discovery® Extraction Server Inxight Categorizer™ Inxight Summarizer™	Inxight Software Inc.	http://www.inxight.com/products/smartdiscovery_as/ http://www.inxight.com/products/sd_es/ http://www.inxight.com/products/sd_es/categorizer/ http://www.inxight.com/products/sdks/sum/
Omniviz	Biowisdom Ltd.	http://www.biowisdom.com/solutions/ (Note: Biowisdom merged with Omniviz Inc. in January 2007)
TEMIS Insight Discoverer™ Extractor Insight Discoverer™ Clusterer Insight Discoverer™ Categorizer XeLDA^{R}eXtraction Terminology Suite™ Skill Cartridge™ Library	TEMIS	http://www.temis.com/index.php?id=59&selt=1
Group 2 tools		
Quosa™ Information Manager Quosa™ Virtual Library™	Quosa	http://www.quosa.com/solutions.html
RefViz™	Thomson ResearchSoft (Thomson Scientific division)	http://www.refviz.com/rvinfo.asp
STN® AnaVist™	CAS	http://www.stn-international.de/stninterfaces/stnanavist/stn_anavist.html
VantagePoint	Search Technology Inc.	http://www.thevantagepoint.com
Thomson Data Analyzer	Thomson Scientific	http://www.thomson.com
Group 3 tools		
Aureka®	MicroPatent/ Thomson Scientific	http://scientific.thomson.com/products/aureka/ http://aureka.micropat.com/7w/html/7w_default.asp
M-CAM Doors™	M-Cam	http://www.m-cam.com/doors
Wisdomain Focust Wisdomain Patent Family Tree	Wisdomain Inc.	http://www.wisdomain.com/Overview.htm
PatAnalyst™	Empolis & Jouve	http://www.patanalyst.com/
Section B		
Anacubis	i2 ChoicePoint	http://www.i2.co.uk/anacubis/
AKS^{2} (alma Knowledge Server)	Bioalma	http://www.bioalma.com/aks2/
BizInt Smart Charts for Patents	BizInt Solutions Inc.	http://www.bizcharts.com/patents/index.html
Delphion – PatLabII Delphion – Clustering	Thomson Scientific	http://www.delphion.com/products/research/products-patlab http://www.delphion.com/products-research/products-clustering
Matheo Patent	Matheo Software	http://www.matheo-patent.com/
PatentExaminer	Questel-Orbit	http://www.patexaminer.com/DisplayLogin
Technology Watch from IBM	IBM-Synthema	http://www.synthema.it/english/index.html
Vivisimo® Velocity – Search Engine, Content Integrator & Clustering Engine	Vivisimo	http://vivisimo.com/html/velocity
Wistract®	Bayer Business Services	http://www.wistract.com/

To summarize, the tool evaluation template looked like this:

- Type of tool
  - Text mining software tool
  - Database content provider
  - Both
- Capabilities
  - Keyword analysis
  - Statistical analysis
  - Linguistic analysis
- Data sources
  - Structured bibliographic data sources
  - Unstructured sources
  - Hybrid sources
- Results
  - Lists of documents
  - Tables
  - Charts/Graphs
  - Maps
- Perceived strengths – our impressions only

– Potential limitations – our impressions only
– Summary

6. Phase I tool evaluation

The 14 text mining and visualization tools in Table 1, Section A, were evaluated based on the "type", the capabilities, the data sources permitted, the results generated, our perceived strengths and potential limitations. It is very important to remember that our evaluations are based solely on the vendor presentations (given in 2006) and our best knowledge at the time. Any improvements and the revisions thereafter are not considered.

We categorized the tools into three groups based on their primary data sources.

Group 1 includes the most flexible tools capable of handling unstructured text:

- ClearForest
- Goldfire Innovator™
- Inxight®
- OmniViz
- TEMIS

Group 2 works best with structured text:

- Quosa™
- RefViz™
- STN®AnaVist™
- VantagePoint
- Thomson Data Analyzer

Group 3 is patent-focused or handles partially structured ("hybrid") data:

- Aureka®
- M-CAM Doors™
- Wisdomain
- PatAnalyst™

These tools are described below in group order. Each of the tools is discussed below in more detail. Some of these tools (e.g., Aureka®, OmniViz, RefViz™, and VantagePoint) were discussed in Eldridge's article [2]. More information is also available from each vendor's website.

6.1. Group 1 tools

The first group of text mining tools we evaluated works best with the unstructured data, such as, full-text patent documents, emails, internal reports, news, journal articles, and web content. This group includes ClearForest, Invention Machine GoldfireInnovator™, OmniViz, and TEMIS. It appears that none of these tools are primary data source providers.

6.1.1. Clearforest

ClearForest offers a Text Analytics solution. It comprises an advanced tagging and extraction platform, an analytics platform, and a development environment. The text mining tool performs list/co-occurrence matrices generation, data clustering/mapping, term extraction and tagging, where the application selects relevant terms from unstructured text such as news articles, web surveys and HTML documents. Once structured, this information can be used in stand-alone analytics applications or combined with structured data to provide more comprehensive business intelligence. Terms are extracted for subsequent analysis, and categorized automatically into pre-defined categories or taxonomies.

The tool permits the visualization of the relationships between collections of taxonomies to extract information that is relevant, actionable and adds value to their Business Intelligence applications. The addition of this situational context to enterprise data systems empowers organizations to: uncover hidden relationships, evaluate events, discover unforeseen patterns and facilitate problem identification for rapid resolution. Thus companies could make better business decisions by turning large volumes of contextually-based information into proactive business intelligence.

One of the unique features and strengths of ClearForest is converting the unstructured to structured data using their Packaged Extraction Module as shown in Fig. 1, within the Text Analytics tool. Text from patent documents obtained from sources like MicroPatent, USPTO or commercial databases can be extracted into structured data entities, such as, claim elements, problems solved, and process technology terms.

Since ClearForest appears to be an enterprise-wide application for text mining and analysis, one of the major issues is defining its practical implementation within an organization. Setting up a cohesive and consistent framework across different business groups may require a significant upfront investment of time, effort and cost. Similar considerations may apply to Goldfire Innovator™, Inxight® and TEMIS.

6.1.2. Invention Machine Goldfire Innovator™

Goldfire Innovator™ is a text mining tool developed by Invention Machine Corporation headquartered in Boston, Massachusetts. The product includes three components: Optimizer Workbench, Researcher, and Innovation Trend Analysis. As shown in Fig. 2, Goldfire Innovator™ uses sophisticated semantic analysis technology that converts unstructured text into a searchable semantic index. Based on the vendor presentation, one of the major capabilities of Goldfire Innovator™ is the creation of the Knowledge Base. It's well known that the SAOs (subject, action, and object) are parts of the language used to describe the meanings and teachings in writing. Goldfire Innovator™ identifies the SAOs so that the teachings or concepts from documents can be extracted from the rest of the documents to create a knowledge base. When a query is sub-

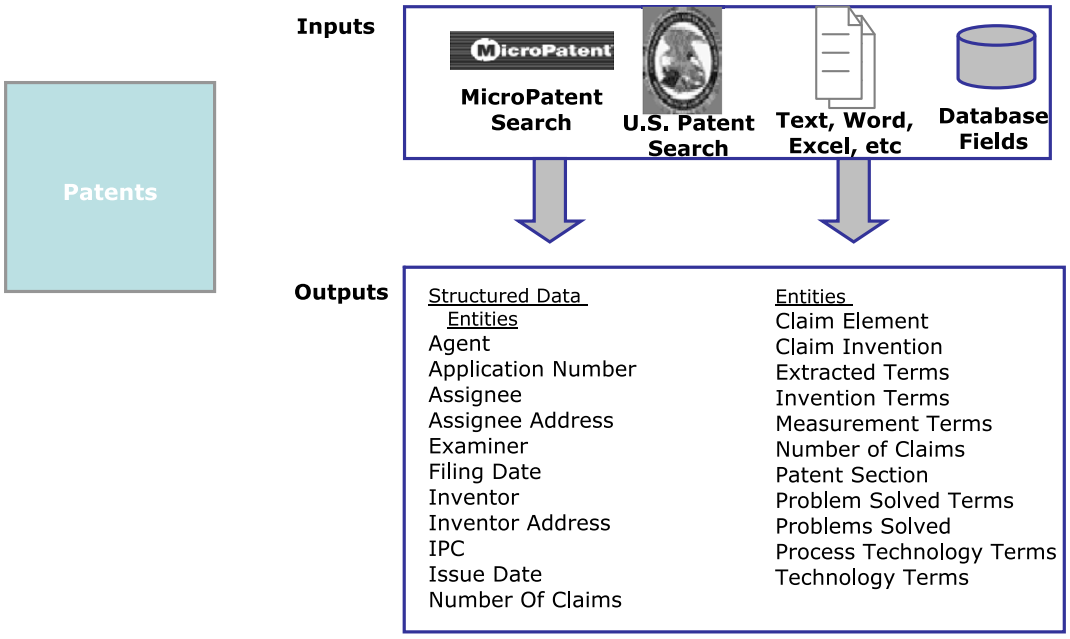

Fig. 1. ClearForest packaged extraction module.

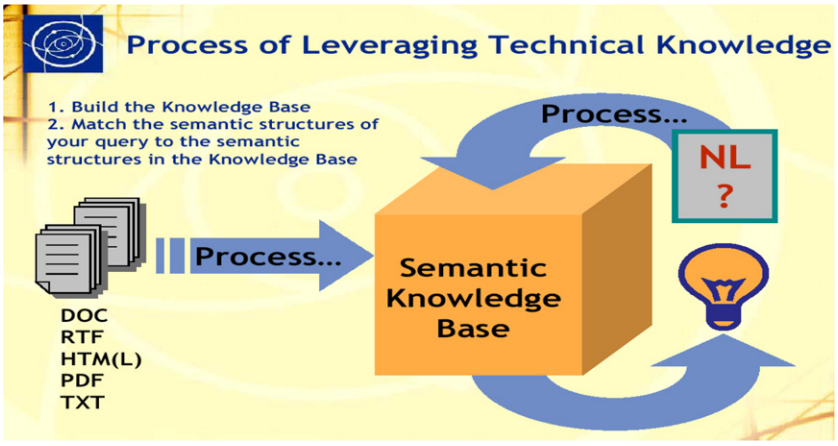

Fig. 2. Goldfire Innovator™ process for leveraging technical knowledge.

mitted, Goldfire matches the semantic structures in the Knowledge Base with the semantic structures from the query.

Through Goldfire Innovator™, 15 million patent documents and over 3000 "deep web" websites as well as more than 8000 scientific journals can be searched. Knowledge Bases can also be created on internal content, including databases, content management systems, filing systems, etc. The combination of the Knowledge Base and the access to searchable content makes Goldfire Innovator™ a very powerful tool to deliver the static categorization of the key concepts, retrieve the relevant answers to questions and provide dynamic document summaries.

Rather than traditional keyword ranking and extraction, Goldfire Innovator™ provides powerful solutions from context. It greatly improves the precision of the information retrieved. One component that we found to be a strong feature is the Innovation Trend Analysis that enables competitive, technology, and citation analysis.

Due to the complexity of the tool, in-house training for end-users as well as database administrator(s) would be required for implementing Goldfire Innovator™. The cost of Goldfire Innovator™ may also set an additional limitation for this excellent tool.

6.1.3. Inxight®

Inxight® was a spin-off of Xerox Palo Alto Research Center (PARC). Its text mining software solutions are empowered by NLP (natural language processing) to enable the contextual extractions. According to Inxight®, the software can read the documents and understand them on a deep linguistically based level to index, categorize, and extract all relevant concepts, entities, and the relationships. The software identifies more than 35 types of information

within a single document. The data sources are unstructured text such as news, websites, internal documents, and full-text patents. The meta-data and the entities can be extracted from the pre-processed documents. Output from Inxight® is hierarchical categorization. The documents are analyzed based on pre-determined categories within hierarchies.

The outstanding feature of the Inxight software is its capability to provide the simultaneous searching in multiple online databases. This capability is also known as federated search. The software works in 32 languages and it understands 27 entity types. Inxight® claims the software has more accuracy than a human reader and it has the most powerful linguistic algorithms in the field.

At the time of the vendor presentation, we felt this tool may require significant investment of time and effort to define more specific categories for implementation in the life science area. The categories demonstrated were too high-level and generic (e.g. medical, science, and news) to be of practical use.

6.1.4. OmniViz

OmniViz is an advanced visual informatics software package designed to provide visualization of numeric data, categorical data, genome sequences, chemical structures, and text documents – literature and patents – all in the same visual framework. OmniViz can deal with very large data sets and can perform case sensitive text analysis to differentiate gene names from proteins, or compound names from common English words. OmniViz combines sophisticated statistical and text analysis algorithms with a range of powerful visualizations to understand data in new ways. Fig. 3 shows the multiple visualization methods offered by OmniViz.

Its key strength is that all data types can be visualized and analyzed together, all in the same tool, providing interactive visualization maps such as Galaxy and CoMet, as shown in Fig. 3. Data selected in one visualization are highlighted in all the other visualizations. OmniViz allows use of different visualizations and tools in connection with one another to explore the full range of connections within data.

OmniViz was initially focused on the life and chemical sciences market, but also delivers its technologies to broader markets including consumer product markets, government, and education. Application areas include research and development, clinical trials, field trials, marketing, finance, legal, and enterprise content management (ECM). OmniViz Inc. is now part of BioWisdom Ltd. as a result of a recent merger.

OmniViz is primarily focused on visualization methods for viewing analysis results. It could be combined with other text mining tools, such as TEMIS, to take advantage of their synergistic capabilities.

6.1.5. TEMIS

TEMIS is a text mining tool as indicated by its name, abbreviated from text mining solutions. TEMIS is empowered by a semantic approach. Its capabilities include extraction, categorization and clustering, which we found to be strong features. At the time of the vendor demonstration, the software was packaged with Insight Discover™ Extractor, Insight Discover™ Categorizer, Insight Discover™ Cluster, a multilingual engine (using natural language processing) XeLDA™, and Skill Cartridges™. Among the many good features this software has to offer, we were most impressed with its Skill Cartridges™. A Skill Cartridge™ is a set of the customized knowledge compo-

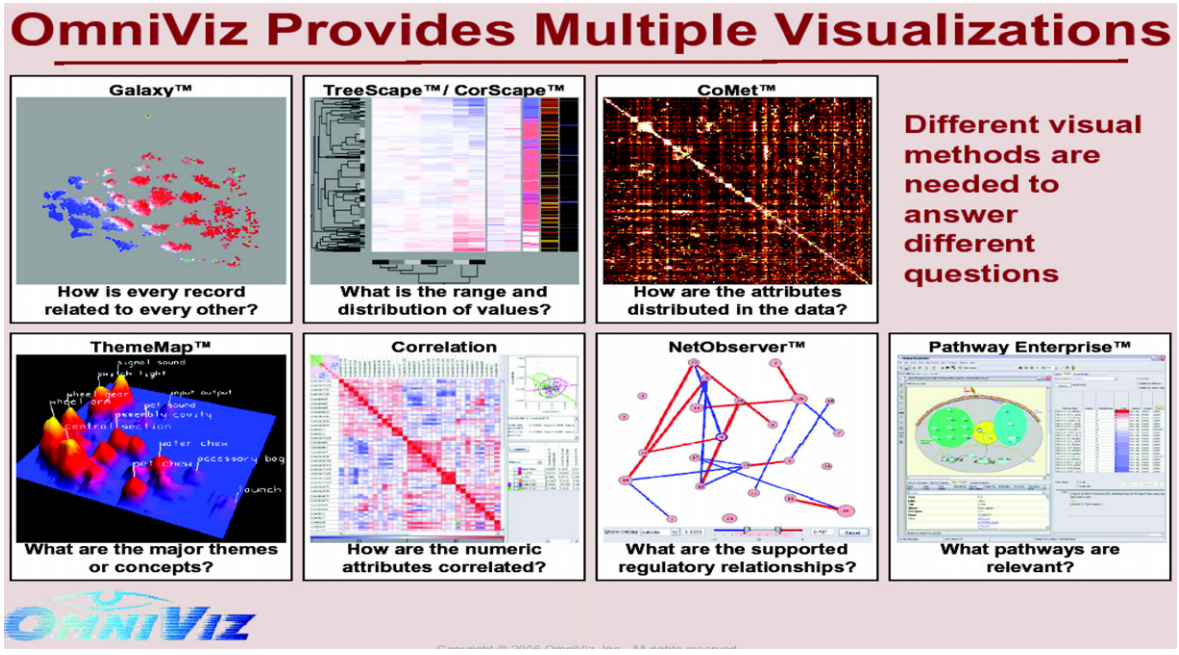

Fig. 3. OmniViz multiple visualization products.

nents that define the information to be extracted. As depicted in Fig. 4, the text is first processed by the multilingual engine XeLDA to convert it into "out-of-context" words. The meanings of the text are then extracted by its Discover™ Extractor which uses a specialized Skill Cartridge™. The text is then converted into meanings. TEMIS works with any text from any document in any format. Output of TEMIS includes clusters, rankings, and lists of references that can be used to uncover trends and relationships. Searching by concepts is one of the strong features we like. Its combination with specialized Skill Cartridges™ offers powerful text discovery solutions to help users drill down the full text to discover the most relevant answers.

TEMIS also provides a Chemical Document Browser, currently marketed as Desktop Miner® for Chemistry (DM4Chem). This browser is a specialized extraction module that translates chemical names into chemical structures.

At the time of the product demonstration, TEMIS seemed to lack a variety of visualization options. We felt the tool could be integrated with other, more robust visualization tools like OmniViz to improve its visualization capabilities. The high cost may be another factor potentially restricting its use.

6.2. Group 2 tools

The second group of text mining tools that we evaluated works best with the structured data, also known as fielded data, such as patent front pages and bibliographic information from databases. This group of tools includes Quosa™, RefViz™, STN®AnaVist™, VantagePoint, and Thomson Data Analyzer.

6.2.1. Quosa™

Quosa™ is a text mining tool based on concept extraction and clustering. It does statistical analysis including term extraction and frequency ranking. Its concept extraction uses a dynamic extraction algorithm developed by mathematicians from MIT and Harvard.

Although we grouped Quosa™ into the group of tools that are most applicable to structured data, the vendor claims that the tool also works well with unstructured text such as text from OVID, Google Scholar, patents, and full text journal articles. Quosa™ users can mine the full text to retrieve the information that is difficult to discover by other means.

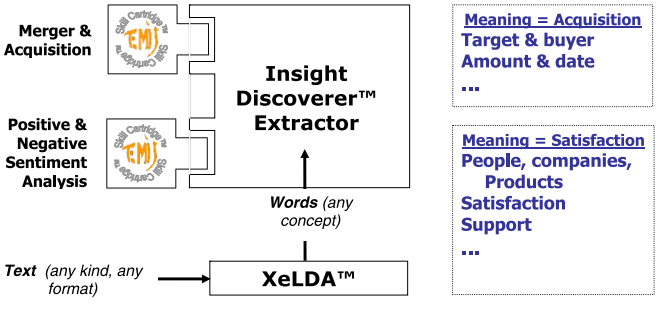

Fig. 4. TEMIS text mining process showing how words are converted to meanings.

We were impressed with Quosa™'s full text searching capability and its ability to pull out or download a large number of full text journal articles and patents. Another good feature of Quosa™ is that the articles can be downloaded to Endnote for further study.

The results of Quosa™ are highly organized collections of documents on a shared server or local computer. The information can be shared by team members. Annotating is also available.

Quosa™ seemed to be a very powerful tool that applies primarily to literature sources, such as PubMed, Ovid, and Google Scholar. We did not see as strong an application of this tool in patent information sources as in the non-patent information sources.

6.2.2. RefViz™

RefViz™ is a data visualization and analysis tool, specifically designed for bibliographic references. It was created by OmniViz Inc. (now part of BioWisdom Ltd.) for Thomson ResearchSoft, who markets and sells this software. It has capabilities to perform statistical and linguistic analysis. RefViz™ works only with structured data, such as titles and abstracts from databases or bibliographic information from EndNote, ProCite, and Reference Manager. The results from RefViz™ are Galaxy review and matrix review.

The main feature of RefViz™ is the Reference Retriever™ which can be used to search multiple online sources simultaneously. The large set of references can be analyzed by thematic content.

RefViz™'s key strength is post-processing of bibliographic information. It utilizes a small subset of OmniViz's text analysis and visualization capabilities in a user friendly interface that is designed for use by anyone who conducts literature searches; has a personal database of literature that they need to comprehend; currently works with citation management software packages and wants to visualize the literature contained within; and needs a product to help them analyze their document information. In RefViz™, documents are organized by thematic content and presented in two interactive visualizations that facilitate rapid identification of major themes and areas of interest. RefViz™ can be used to learn more from expanded literature searches, and thereby minimize the loss of important information that may occur by narrowing a search too soon.

Reference Retriever™ works best with the following literature sources: Web of Science, PubMed, Ovid, OCLC, Library of Congress, and Purdue University. Additional configuration is required for other sources.

6.2.3. STN®AnaVist™

STN®AnaVist™ is an interactive, frequency-based analysis, and visualization tool that offers a variety of ways to analyze search results from both scientific literature and patents. Patterns and trends can be visualized. Answer sets are created and exported with STN Express from multiple

databases – including CAplus, PCT, and US full-text patent databases, and, more recently, from Derwent DWPI. STN®AnaVist™ has the capability to analyze the patent landscape, track competitive intelligence, discover new applications for existing technology, determine research trends, and support strategic business planning.

A key feature is the application of CAS vocabulary to standardize technology terms across databases to reduce data scatter. Interactive relationships among data and charts are visualized in the visualization workspace. Data can be highlighted during the analysis to easily see relationships. Key capabilities include the use of a company name thesaurus to group and cleanup assignee data. Visualization results can be shared with others within an organization using STN Login IDs for Shared Projects.

The current STN AnaVist™ (version 2.0) works only with references downloaded from CAPLUS^{SM}, PCTFULL full-text database, USPATFULL full-text database and, more recently, Derwent World Patent Index DWPI^{SM}. Although it works well with bibliographic data analysis, it does not appear to have significant text mining capabilities for full-text information.

6.2.4. VantagePoint

VantagePoint, developed by Search Technology Inc., is a desk-top text mining and visualization tool empowered by natural language processing (NLP). It provides rapid navigation through structured text, such as, the bibliographic information obtained from the online hosts to discover the hidden patterns, trends, and relationships. The lists generated from various fields (including the NLP phrases) can be cleaned so that the concepts can be grouped, clustered, or categorized. The outcome of the VantagePoint analysis is a series of matrices, factor maps, correlation maps, and Excel charts. Fig. 5 shows a sample VantagePoint template applied to fuel cell automotive technology.

VantagePoint comes with a Reader module that enables end-users throughout a company to interact with, but not edit, the results output and data analyzed by a professional information analyst.

One of the strengths of VantagePoint is the analytical tool box feature. It helps to answer who, where, when, and what types of questions for anyone who would like to plot such information. We think this is a very useful tool for business intelligence and competitive intelligence in terms of monitoring competitors' research trends, or identifying the top inventors and/or acquisition targets. It is most useful for company profiling and technology assessment.

Another strength of VantagePoint is its powerful title reviewing window feature which gives users easy access to

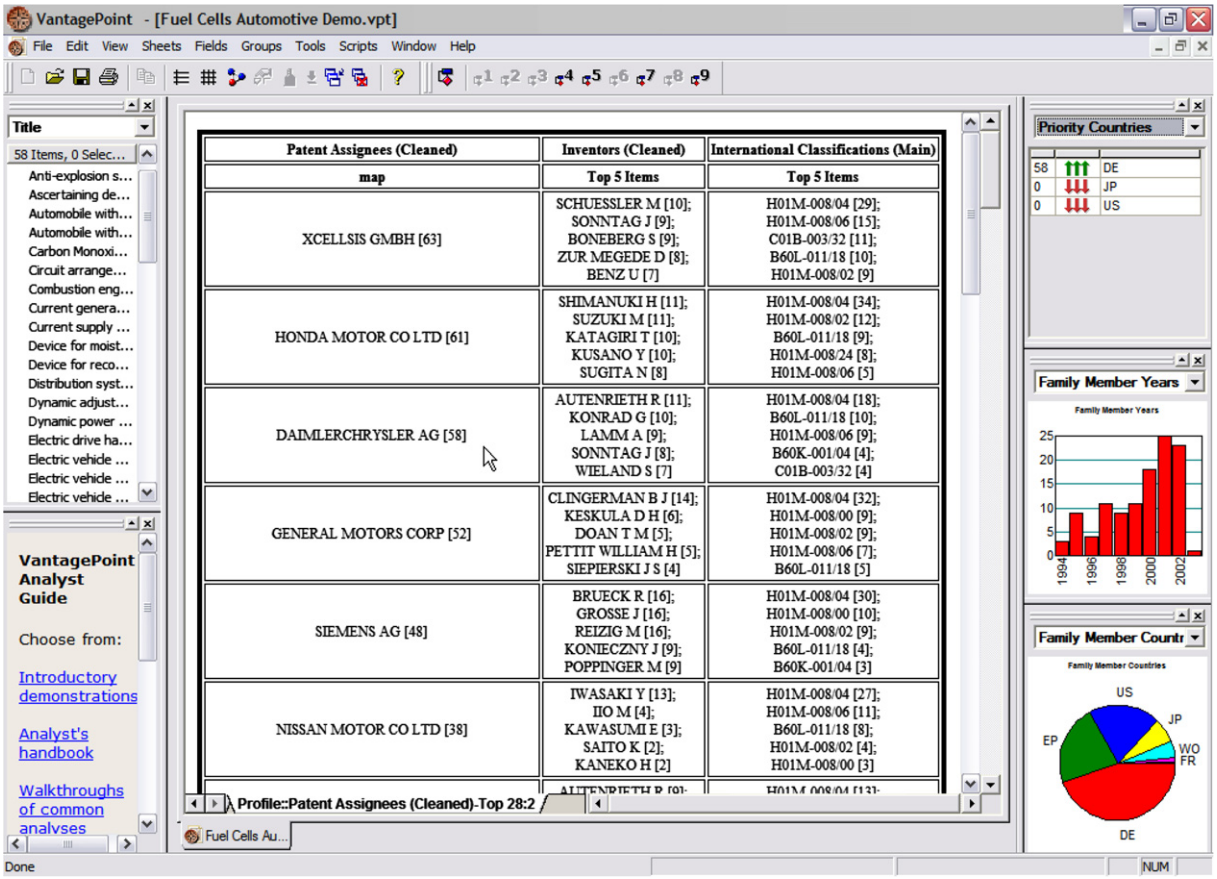

Fig. 5. VantagePoint – sample template applied to fuel cell automotive technology.

the original text. Scientists may use this function to drill down to the text to quickly locate required information.

List clean-up is the most important task to perform in order to ensure data integrity. Although VantagePoint provides pre-defined thesauri (author, affiliation, and British American spelling) to automate the list clean up process, it still involves considerable manual work to do the data cleaning, particularly when the size of the document set is large. Since the list cleanup is a critical step to generate meaningful analysis results, any improvement in list cleanup would be highly desirable.

6.2.5. Thomson data analyzer

The successor to the Derwent Analytics[SM], Thomson Data Analyzer (TDA) is fundamentally a VantagePoint tool configured to optimally analyze value-added data from Derwent patent sources, such as World Patent Index (DWPI). TDA provides the VantagePoint user-defined analytical tools that allow for creation of cleaned lists, co-occurrence matrices to see trends and relationships, and maps that visually represent relationships within large data sets. By combining the power of VantagePoint data mining software with Derwent value-added patent data, users can analyze trends, profile competitor's patenting activity, track the progress of a specific technology area, identify areas of strategic R&D opportunities, and identify prolific inventors/collaborators. The main strength of TDA is the set of pre-defined analytical tools that perform automated analyses (macros) optimized for Derwent data, and generation of reports with a single click. Its toolkit includes user-defined thesaurus creation, and removal of duplicates.

Like VantagePoint, Thomson Data Analyzer comes with a Reader module that enables users to interact with, but not edit, the data and results output from TDA.

Since Thomson Data Analyzer is based on VantagePoint technology, the limitations of TDA are basically similar to VantagePoint. Whereas VantagePoint can be applied to a variety of structured data sources, TDA is optimized for Thomson data sources.

6.3. Group 3 tools

The third group of text mining tools we evaluated works well with hybrid data, or partially structured text, such as patent documents. This group includes Aureka®, MCAM Doors™, Wisdomain and PatAnalyst™. These tools provide access to their own patent databases.

6.3.1. Aureka®

Aureka®, a pioneer in the patent text mining and visualization arena, allows organization and management of intellectual property (not just patents, but corporate documents as well). It offers both ThemeScape thematic text mining capability and a search engine for patent content collection. The text mining capability is based on keyword and statistical analysis.

The search engine accesses MicroPatent's full-text patent collection and includes a search engine for identifying relevant references. These references can be saved, creating sets for further analysis and sharing with colleagues. Another nice feature of the Aureka® platform is the ability to annotate documents. One of the key strengths is the ability for individuals within an organization to create sets of patents, analyze them, annotate them, generally create intelligence from them, and save all of this knowledge in a single place.

One of the unique analytical tools built into the Aureka® system is the ThemeScape thematic text-mining tool that allows the creation of customized stopword lists. ThemeScape employs a concept mapping method of creating technology landscapes. As shown in Fig. 6, Aureka's ThemeScape is the contour map display which gives users a birds-eye view of the common concepts from a given document set. The program reads full-text documents, identifies themes that occur throughout the references, and employs clustering algorithms to organize documents by co-occurrence of the identified themes.

Another strong feature of Aureka® is the citation analysis tool which incorporates an interactive, hyperbolic tree viewer for US patents. Citation trees can support a rapid visual review of the citation history for a single US patent by color-coding and labeling in a number of different ways, including by assignee, publication date, or inventor.

New users of Aureka's ThemeScape may find it challenging to understand the relationships between clusters as the peak labels may not provide a meaningful definition. At the time of our evaluation, citation analysis worked only for US patents.

6.3.2. M-CAM DOORS™

M-CAM Inc. is a global, full-service firm providing intellectual property and rights (IP&R) and intangible asset (IA) financial services. It offers text analysis and risk management solutions. It is also a database provider, with access to patents from over 88 patenting authorities, and more than 50 million patent documents. Strengths are in patent uniqueness and enforceability analysis.

M-CAM DOORS™ was the product demonstrated to us. It is a patent risk management and analysis tool for determining the unique commercial opportunities and threats of patents issued, and applied for, in the United States and around the world. It is marketed as a tool for companies to help identify prior art and licensing opportunities for their portfolios. The system works by combining advanced semantic analysis with co-citation analysis. Combining the two techniques allows easier identification of highly related patent references.

M-CAM DOORS™ uses visual displays, such as the "Compass" citation view, and "Magellan" telescope and hourglass views, as shown in Fig. 7, to help analysts keep track of a collection of patents concurrently. We were very

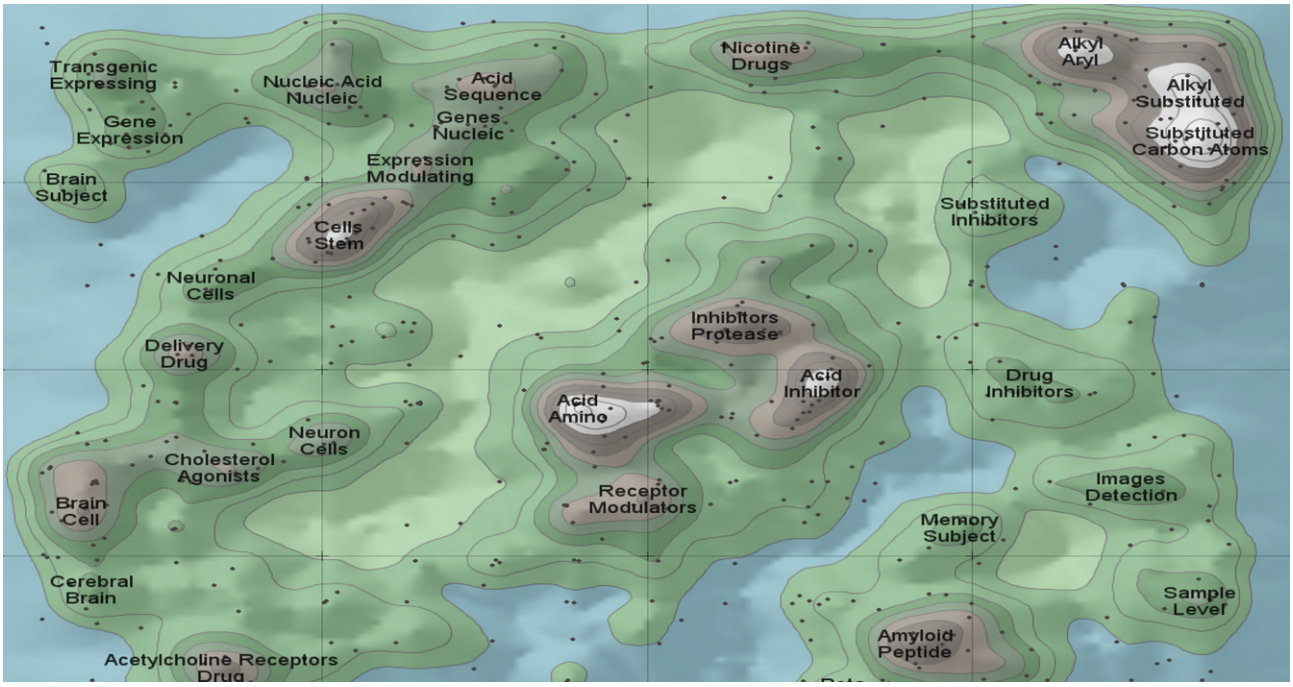

Fig. 6. A ThemeScape map of a large set of documents provides an initial view of the content. Additional probing and analysis of the map will help to reveal more insight.

impressed by these visual displays which can be used to review patents from statutory, potential infringement, and commercial applications perspectives by patent searchers/analysts and attorneys.

Through multi-lingual linguistic and semantic-based analysis, M-CAM DOORS™ also enables competitive intelligence analysis and financial risk analysis for mergers and acquisitions, and stock trading activities.

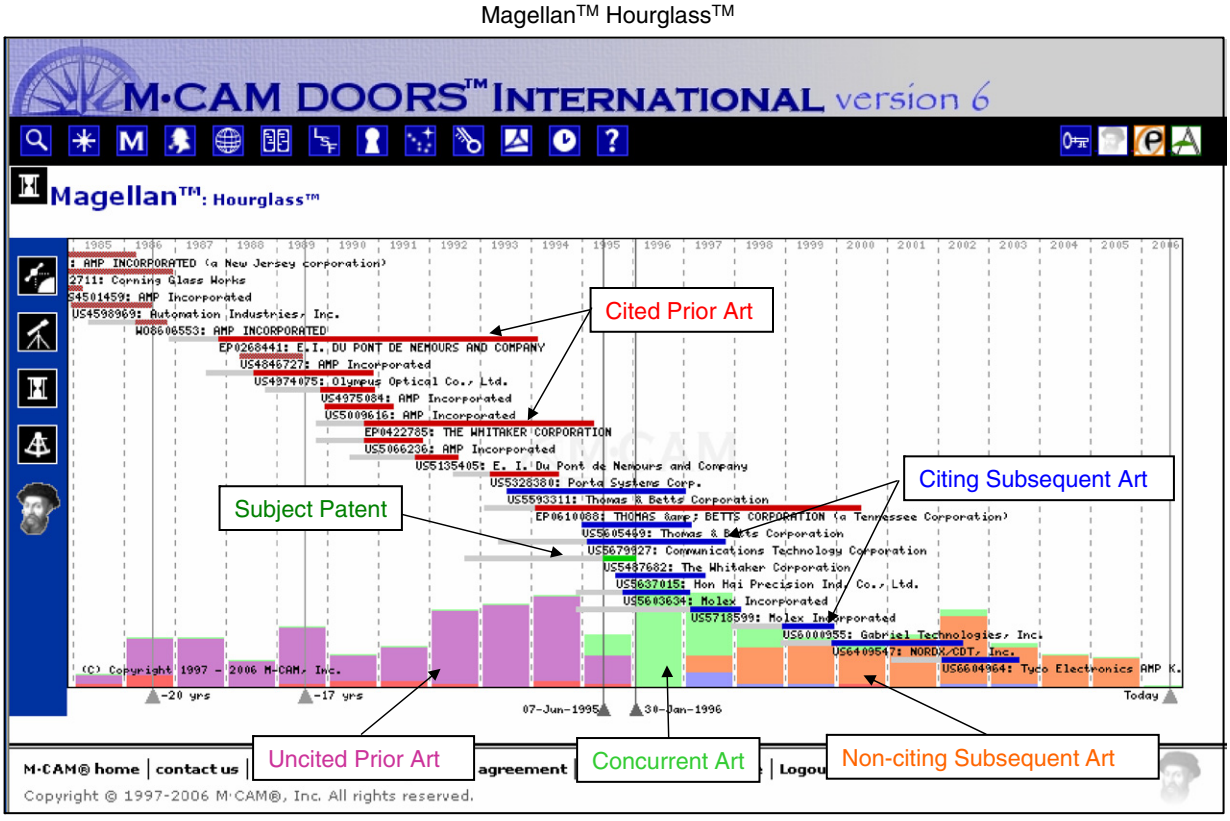

Fig. 7. M-CAM DOORS™ Magellan™ Hourglass™ view of patents.

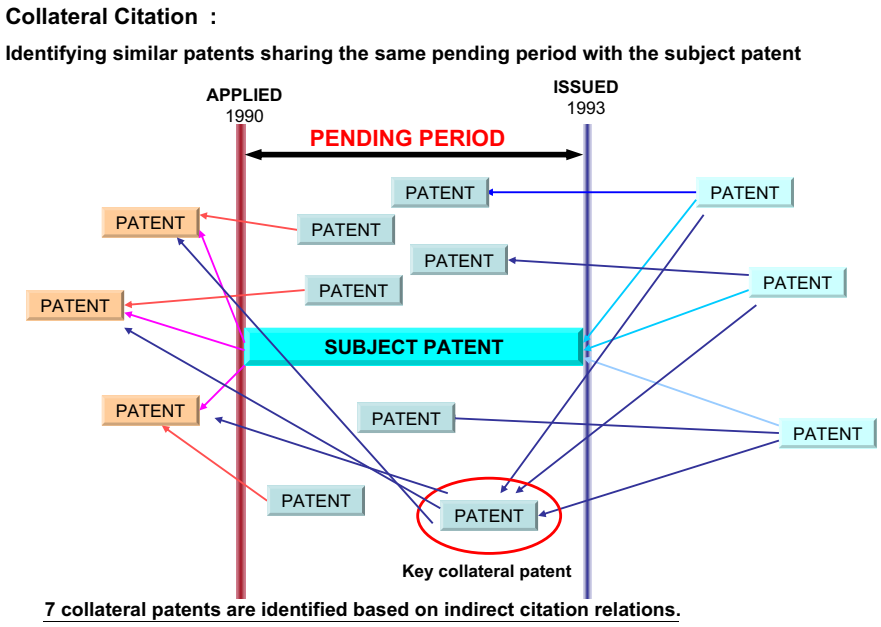

Fig. 8. Wisdomain: sample collateral citation analysis – identifying similar patents sharing the same pending period with the subject patent.

Although M-CAM DOORS™ appeared to be one of the better tools for patent analysis, the cost for conducting a trial limited further evaluation of this tool.

6.3.3. Wisdomain

Wisdomain Inc. offers a suite of integrated tools with patent searching and text mining capabilities. Their major product is FOCUST which provides a searchable patent database coupled with advanced analysis software. FOCUST provides access to US, EP, PCT, JP, CN (abstracts), and KR patents as well as Inpadoc legal status and patent family data.

Wisdomain's FOCUST consists of a Search Module, a Citation Analysis Module, and an Analysis and Visualization Module. Some of FOCUST's capabilities apply only to US patent data. We were quite impressed with the Citation Analysis Module, in particular their Collateral Citation Analysis, which helps to identify similar co-pending patents to a target patent. Fig. 8 shows a view of seven collateral patents identified based on indirect citation relationships to the target patent. Patents in the pending period may not be retrieved by traditional citation analysis.

Other Wisdomain products are the Patent Family Tree and PatentMagnet. The Patent Family Tree combines two pieces of family information: the patent Genealogy tree for US patents, reflecting US patent system specific Continuation, Continuation-in-Part, and Divisional patent information, plus Inpadoc patent family data. This information is provided in a user-friendly tree mapping format. PatentMagnet is a patent search facility that guarantees finding similar patents to a target patent. With just a single patent number, the tool allows users to find potential prior art in patent infringement or validity cases.

Wisdomain uses its own patent collection for analysis and does not seem to allow for importation of datasets from other sources. Like Aureka, Wisdomain citation analysis is limited to US patents only.

6.3.4. PatAnalyst™

Although the name PatAnalyst™ implied text mining capability to us, we found that it is actually a tool for searching patents like MicroPatent or PatBase, with collections of full-text patents from at least eight patenting authorities, and bibliographic data from more than 70 countries. It does not have a text mining algorithm capability. However, it is quite impressive in terms of its clean search interface, highlighting features, viewing capabilities, and organization of patents into folders/drawers.

7. Conclusions

The results of our evaluation are summarized in Table 2. Key findings include the type of tool, its capabilities, the data sources permitted, results generation, and potential user groups that were identified for each tool. Our perceived strengths and potential limitations are listed in Table 3.

We have looked at three groups of tools in this Phase I evaluation and identified their strengths and potential limitations from a patent analysis perspective. The first group that can process unstructured data includes ClearForest, Goldfire Innovator™, Inxight®, OmniViz, and TEMIS. Amongst these tools, semantic analysis appeared to be the most powerful technique for text mining large sets of documents. The second group of tools we evaluated works well for structured text. Within this group, VantagePoint, Thomson Data Analyzer, and Quosa™ appeared to have

Table 2
Evaluation summary

Vendor tool	Type of tool	Capabilities	Data sources	Result output	Potential users
Group 1 tools					
ClearForest TextAnalytics	Text mining	Semantic analysis/ natural language processing	Structured and unstructured text from web, internal documents patents, etc.	Structured data entities, lists, visualization tools – trend graphs, category maps	Business Intelligence
Goldfire Innovator™	Text mining	Semantic analysis	Unstructured text from personal data, corporate data, deep web, patents, etc.	Summarization, categorization	R&D scientists
Inxight Smart Discovery®	Text mining	Natural language processing, contextual extractions	Unstructured text from web, internal repositories; pre-process documents to extract meta-data and identify entity types	Hierarchical categorization	R&D Informatics
Omniviz	Visual-based text/data mining	Statistical analysis	Structured and unstructured text, numeric, chemical structures	Interactive visualization maps	R&D scientists
TEMIS	Text mining	Natural language processing	Structured and unstructured text from web, internal docs, patents, clinical trials, email, bioinformatics, etc.	Clusters, lists, rankings	R&D scientists, Business Intelligence
Group 2 tools					
Quosa™	Text mining based on concept extraction/ clustering	Statistical analysis	Structured and unstructured text from patents, internal documents, PubMed, Google Scholar, Ovid	Organized document collection, team sharing, annotation	R&D scientists
RefViz™	Text analysis and data visualization	Statistical and linguistic analysis	Structured text from ISI Web of Science, PubMed, OCLC	"Galaxy" and matrix visualization	R&D scientists, information professionals
STN® AnaVist™	Text analysis and database provider	Statistical analysis	Structured text – CAPLUS, USPatful, PCTFull, DWPI	Charts, research landscape map	Information professionals, business intelligence, R&D scientists
VantagePoint	Text mining	Pattern matching, natural language processing	Structured text from bibliographic fields	Lists, summaries, charts, maps, matrices, graphs	Information professionals, Business Intelligence
Thomson Data Analyzer	Text mining	Pattern matching, natural language processing	Structured text from bibliographic fields	Lists, summaries, charts, maps, matrices, graphs	Information professionals, Business Intelligence
Group 3 tools					
Aureka®	Text mining and database provider	Keyword and statistical analysis	Patents from MicroPatent database	ThemeScape maps, hyperbolic citation trees, text clusters	Legal/Patent Dept., R&D scientists, Information professionals, Strategic Planning, Business Intelligence
M-Cam Doors™	Database provider, with text analysis/ risk mgmt solution	Linguistic and semantic analysis, multi-lingual	Patents from 88 patenting authorities and journals (start summer of 2006)	"Compass" citation view, "Magellan" telescope view, competitive analysis, financial risk analysis for merger/acquisition stock trading, patent uniqueness and enforceability analysis	Business Intelligence, Legal/Patent Dept, Information professionals
Wisdomain	Database provider and text mining	Keyword analysis, citation map visualized searching	Patents from US, EP, PCT, PAJ, Chinese, and Korean abstracts, Inpadoc legal status	Genealogy tree, citation maps, tables, charts	R&D scientists, Information professionals
PatAnalyst™	Patent database provider	No text mining capability	Patents from US, EPO, PCT, PAJ, UK, Germany, France, Switzerland	Viewer with organized set of patents, highlighted keywords	Information professionals, R&D scientists, Information professionals

broader application than RefViz™ and STN®AnaVist™, which are limited only to bibliographic data and abstracts of a few data sources.The third group works well with hybrid data and includes Aureka®, M-CAM Doors™, Wisdomain, and PatAnalyst™. Each of these tools has very specific functionalities and strengths that can be applied

Table 3
Perceived strengths and potential limitations

Vendor tool	Perceived strengths	Potential limitations
Group 1 tools		
ClearForest TextAnalytics	Extraction modules	Significant upfront investment
Goldfire Innovator™	Sophisticated semantic analysis	Requires in-house training; high cost
Inxight Smart Discovery®	Extraction and federated search	Requires input of more specific categories for practical use in life science area
Omniviz	Interactive visualization	Focused primarily on visualization
TEMIS	Extraction using Specialized Skill Cartridges™	Limited visualization options; high cost
Group 2 tools		
Quosa™	Full-text retrieval and management	Primarily used for non-patent literature sources; additional configuration for other sources
RefViz™	Bibliographic data post-processing	Primarily used for non-patent literature sources; additional configuration for other sources
STN®AnaVist™	Standardized vocabulary; company thesaurus	Applicable to very few STN data sources
VantagePoint	Analytical tool box for technology and company assessment	List cleanup of large datasets can be challenging
Thomson Data Analyzer	Pre-defined macros using Derwent value-added data for analytical tool box for technology and company assessment	Optimized for Thomson data sources
Group 3 tools		
Aureka®	Patent mapping, clustering and citation analysis	Difficult to understand Themescape labels; citation analysis for US patents only
M-Cam Doors™	Patent uniqueness and enforceability analysis	Focused on patent information; high cost
Wisdomain	Strong collateral citation analysis	Citation analysis for US patents only; appears to have no data import capability
PatAnalyst™	Powerful full-text patents user interface with display features	Not a text mining tool

to various methods of analyzing patents. These methods may include manual, automated citation, and claims analyses.

Based on our Phase I evaluation results, all of these tools have powerful features. It appears that there is no single tool that can perform all text mining and visualization functions. Each tool may have been designed to target different user groups, providing different aspects of text mining and visualization capabilities. It would be ideal to combine two or more of these tools to take advantage of their optimal features, such as semantic-based analysis with data visualization. The complexity and capabilities of some tools may exceed our current requirements and, therefore, may limit the usage of these tools.

For the pharmaceutical and chemical-related industries, chemical structure information is of paramount importance. A major limitation we found in these tools is the lack of chemical structure mining capability, with the possible exception of Temis, and, perhaps Omniviz, which appeared to have some chemical structure mining features.

Based on what we have learned from the product demonstrations, certain factors such as the complexity of the tools, scope of data sources and types, visualization options, as well as the upfront cost, ongoing maintenance and support issues may set some limitations. However, in key business-critical decision-making processes, the potential value derived from these tools may provide justification for the initially-perceived high cost and other limiting factors.

We have observed that the tools described in Group 1 are most flexible due to their capabilities to process a broad range of data sources and use of advanced semantic technologies. The tools in this group may require in-house training or a significant upfront investment of time, effort and cost. The tools in Group 2 work primarily with structured data sources and therefore, their use may be optimal for bibliographic information. Limitations in Group 3 are tool-dependent. Some tools in this group provide citation analysis but they work only with US patents.

In conclusion, we have attempted to summarize many of the text mining and visualization tools currently available in the market. We have discussed their unique features, perceived strengths, potential limitations, applicable data sources, and output of results. We hope this review will help readers to understand the current landscape of text mining/visualization tools. This field continues to evolve as users and vendors work together to improve the capabilities and usefulness of these tools. As machines revolutionized the industrial age in the 19th century, text mining/visualization tools have the potential to revolutionize the information age and drive innovative solutions in the 21st century by providing facile mechanisms for tackling the ever-growing deluge of patent information. As patent analysts, our challenge is to keep up with these emerging technologies and strive to transform our manual analysis methodologies by incorporating these new tools into our daily workflow.

Acknowledgments

We would like to thank our colleagues in Information & Knowledge Integration group at Bristol-Myers Squibb for their active involvement, support, and dedicated assistance in this project: Peter Mattei, Shelley Pavlek, Joanne Freeman, Marlene Khouri, Heahyun Yoo, Tony Medina, Michael Rogers, Tisha Zawisky, and Claudia Powers. We extend our special thanks to Ramesh Durvasula from BMS Informatics, and Ronald Stoner from Mead Johnson for sharing their past experiences and learnings as well as providing their insightful suggestions and support for this project. Special thanks to the vendors who demonstrated their products and allowed us to use their product information for this article.

References

[1] Trippe A. Patinformatics: tasks to tools. World Patent Inform 2003;25 (3):211–21.
[2] Eldridge J. Data visualisation tools – a perspective from the pharmaceutical industry. World Patent Inform 2006;28 (1):43–9.
[3] Stembridge B, Corish B. Patent data mining and effective portfolio management. Intellect Asset Manage 2004 (October/November):30–5.
[4] Fischer G, Lalyre N. Analysis and visualisation with host-based software – the features of STN® AnaVist™. World Patent Inform 2006;28 (4):312–8.
[5] Fattori M, Pedrazzi G, Turra R. Text mining applied to patent mapping: a practical business case. World Patent Inform 2003;25 (4):335–42.
[6] Porter A, Cunningham S. Tech mining, exploiting new technologies for competitive advantage. Wiley & Sons; 2005.
[7] Markley M. Aureka® overview; May 18, 2006 presentation to BMS.

YunYun Yang is currently a Senior Patent Analyst in the Patent Analysis group at BMS providing patent information support to Legal/Patent Department and the cardiovascular drug discovery group. she also leads the initiative to evaluate text mining and visualization tools in the Information & Knowledge Integration (IKI) group at BMS. Prior to joining BMS, she worked as chemistry editor at the ISI, and as information scientist at DuPont. She has been actively involved at PIUG and was Exhibits Committee Chair for the 2007 PIUG Annual Conference.

Lucy Akers is a US patent agent with a degree in Chemistry. She worked for five years as an independent patent information consultant and has held corporate positions with Shell, ExxonMobil and Nerac, in management, advisory or individual contributor roles. She is the Past Chair of the Patent Information Users Group (PIUG) Inc. and is currently a Senior Patent Analyst with Bristol-Myers Squibb.

Thomas Klose is currently a Senior Patent Analyst in Bristol-Myers Squibb's Patent Analysis Group. He first worked with Eastman Kodak Co. in Rochester, NY as a Research Chemist. Later on, he transferred to Kodak's legal division and became a Patent Searcher. After a short interval as a Senior Information Scientist with American Cyanamid in Princeton, NJ, he moved to BMS in 2001. Tom graduated from the University of Adelaide with a PhD degree in Organic Chemistry. He is a member of Patent Information Users Group (PIUG) and the American Chemical Society (ACS).

Cynthia S. Barcelon Yang is currently Director of Patent Analysis Group at Bristol-Myers Squibb. Prior to joining BMS in 2001, She had worked at DuPont, initially as a bench chemist and later as information scientist, group leader, project manager, patent searcher, and team leader in the Patent Information Services group at DuPont Pharmaceutical subsidiary. She obtained her M.S. degree in Chemistry from Marquette University and a M.S. degree in Information and Library Science from Drexel University. She is a member of the Patent Information Users Group (PIUG), the American Chemical Society (ACS), and the Patent Documentation Group (PDG).