Knowledge Management Systems: A Text Mining Perspective

Knowledge Management Systems

A Text Mining Perspective

Hsinchun Chen, Ph.D.

AI
KNOWLEDGE
COMPUTING CORPORATION

" A concise and insightful text by one of the acknowledged leaders in the field. As evidenced by the case studies included here, Professor Chen's research program contributes to numerous application areas and promises to continue to do so."

Stephen M. Griffin, Ph.D.
Program Director, Digital Libraries Initiative Division of Information and Intelligent Systems (IIS) National Science Foundation

" Hsinchun Chen for many years has lead the world in applying new information technologies to large complex information management problems encountered in all large enterprises. This book provides a practical tutorial with compelling case studies on how to address the information avalanche we all face."

Peter Weill, Ph.D.
Director, Center for Information Systems Research &
Senior Research Scientist MIT Sloan School of Management

" Knowledge management is crucial to the success of enterprises of varying sizes. This book gives an excellent explanation and illustration of how some of these new knowledge management techniques and concepts can be leveraged in modern organizations. Its potential for medical informatics and digital government applications, in particular, is tremendous."

Nan-Hung Kuo, Ph.D.
President, Chang Gung University, Taiwan
Former Chairman, National Science Council, Taiwan
Former Minister of Communication and Transportation, Taiwan
Former President, National Chiao-Tung University, Taiwan

" The projects and case studies described in this book are very insightful and useful. I would strongly urge IT researchers and practitioners to look closely into these new techniques and their potential applications in various private enterprises and governments."

Soushan Wu, Ph.D.
Dean, College of Management, Chang Gung University, Taiwan
Advisor, Management Research Program, National Science Council, Taiwan

" Text mining is an increasingly important topic in knowledge management especially after the September 11, 2001 tragedy in the United States. Hsinchun Chen provides us with a delightful set of introductory materials on this topic and how to apply them in the real world. The book comes with a wealth of references, making it easier for novices to find additional in-depth materials."

Susan Butler
Partner, Accenture

" Dr. Chen's book is an essential read for any government manager with enterprise-wide information systems responsibilities. As perhaps the world's largest business, the Federal government is badly in need of KM capabilities, and far behind the best practices of the private sector. The book focuses appropriately on the bi-modal nature of knowledge management; i.e. it requires fresh and cutting-edge thinking in both the technical underpinnings and in organization design, process and culture. Neither is sufficient without the other."

Larry Brandt
Program Manager, Digital Government program
Experimental and Integrative Activities
National Science Foundation

" As a knowledge management practitioner, I am often deluged with vendor hype regarding the latest "fads" in KM technology. Each vendor promises to provide the "complete solution to all your corporate KM needs", leading me to be highly skeptical of any technology that purports to extract meaning from the junk in our unstructured file systems. Thus Dr. Chen's book gives me welcome hope by providing a balanced, clear, factual analysis of all the various KM technologies available today and in the near future. Armed with my knowledge from this book, I am ready to revisit the vendor hype and am encouraged that the new generation of KM tools will provide practical, real business returns through improved decision making."

Margrethe H. Olson, Ph.D.
CKO, Legend Lease
Former CKO, Lotus

" Now more than ever knowledge is power. In this concise book, Professor Chen provides a practical guide to state-of-the-art techniques being used by scholars as well as governments and businesses to harness the power of knowledge."

Mark Zupan, Ph.D.
Dean, Eller College of Business and Public Administration
The University of Arizona

" Dr. Chen is a highly regarded researcher on information retrieval and digital libraries. This short report summarizes research from his university laboratory for a practical business audience and includes succinct explanations of basic concepts such as search engines, information retrieval, and text mining."

Christine Borgman, Ph.D.
Professor and Presidential Chair in Information Studies
University of California, Los Angeles

Knowledge Management Systems

A Text Mining Perspective

By

Hsinchun Chen, Ph.D.

Knowledge Computing Corporation
Department of Management Information Systems
Eller College of Business and Public Administration
The University of Arizona
Tucson, Arizona

Dedication

I dedicate this book, with love,
to the memory of
my father, Kuang-Shen Chen (1933-1997)
and my brother, I-Chun (Jingo) Chen (1961-2000).

This book also owes much to the support from
my mom, Fu-Lien, my wife, Hsiao-Hui (Sherry),
and my daughter and son, Hillary and Hugh.

Hsinchun Chen, Ph.D.
McClelland Endowed Professor of Management Information Systems
Department of Management Information Systems
Eller College of Business and Public Administration
The University of Arizona
Tucson, Arizona

Contents

About the Author

Hsinchun Chen, Ph.D., is the McClelland Endowed Professor of Management Information Systems at the University of Arizona and was honored as the Andersen Consulting Professor of the Year (1999). He received the Ph.D. degree in Information Systems from New York University in 1989, an MBA in Finance from SUNY-Buffalo in 1985, and a BS in Management Science from the National Chiao-Tung University in Taiwan. He is the author of more than 100 articles on text and data mining, medical informatics, intelligence analysis, semantic retrieval, search algorithms, knowledge discovery, and collaborative computing that have appeared in leading information technology publications. He serves on the editorial board of *Journal of the American Society for Information Science and Technology* and *Decision Support Systems*. Dr. Chen is one of the world's leading experts in medical informatics, digital library, e-government, and knowledge management research and his work has been featured in numerous scientific and information technology publications including *Science*, *The New York Times*, and *Business Week*.

Dr. Chen founded the University of Arizona Artificial Intelligence Lab at the Eller College of Business and Public Administration in 1990. The group is distinguished for its adaptation and development of scalable and practical artificial intelligence, neural networks, genetic algorithms, statistical analysis, automatic indexing, and natural language processing techniques. Research development platforms have ranged from supercomputers (CM-5, HP Convex Exemplar, and SGI Power Challenge), to workstations (HP and DEC Alpha), and Pentium-based Windows/NT. Extensive hands-on C, C++, CGI, and JAVA system engineering projects have been under way for the past decade and have made a significant contribution to the research and educational experience of students at all levels in the MIS Department of the University of Arizona (fourth-ranked in the field of MIS graduate education according to *U.S. World and News Report*, April 9, 2001).

Since 1990, Dr. Chen has received more than $10 million in research funding from various government agencies and major corporations including the National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), National Institute of Justice (NIJ), National Library of Medicine (NLM), National Cancer Institute (NCI), National Center for Supercomputing Applications (NCSA), HP, SAP, 3COM, and AT&T.

Dr. Chen is the founding director of the University of Arizona Mark and Susan Hoffman E-Commerce Lab, which features state-of-the-art eCommerce hardware and software in a cutting-edge research and education environment. The lab has received over $2 million in infrastructure funding from Eller College alumnus, Mark Hoffman (founder of SyBase and Commerce One) and HP, as well as more than $10 million of software donations from major IT software companies including SAP, IFS, Commerce One, Oracle, Microsoft, and IBM. Dr. Chen is also founder of an Internet knowledge management company, Knowledge Computing Corporation (KCC), that specializes in medical informatics, law enforcement, and business intelligence applications. The Tucson-based company, has received more than $2.5 million in venture funding and is growing rapidly in the law enforcement and marketing portal sectors.

Dr. Chen received the NSF Research Initiation Award in 1992 and his work has been recognized by major U.S. corporations; he has received numerous industry awards for his contribution to IT education and research. In 1995 and 1996, he received the AT&T Foundation Award in Science and Engineering. In 1998 he received the SAP Award in Research/Applications and became a Karl Eller Center Honored Entrepreneurial Fellow. In 1999, Dr. Chen was awarded the McClelland Endowed Professorship and the Andersen Consulting Professor of the Year Award at the University of Arizona's Eller College. He received the Eller College's Kalt Prize for Doctoral Placement in 2000.

Having been heavily involved in fostering digital library, digital government, and knowledge management research and education in the U.S. and internationally, Dr. Chen was a principal investigator of the NSF-funded Digital Library Initiative-1 project and he has continued to receive major NSF awards from the ongoing Digital Library Initiative-2 and Digital Government programs. He has guest edited special topic issues on digital libraries, knowledge management, web retrieval and mining, and digital government for *IEEE Computer*, *Journal of the American Society for Information Science and Technology*, and *Decision Support Systems*. He also helped organize and promote the Asia digital library research community and has served as the conference general chair of four Asian Digital Library Conferences (in Hong Kong in 1998, in Taipei, Taiwan in 1999, in Seoul, Korea in, 2000, and in Bangalore, India, 2001).

Dr. Chen has frequently served as a panel member and/or workshop organizer for major NSF and DARPA research programs. He has helped set directions for several major U.S. information technology initiatives including: the Digital Library Initiative (DLI), the Knowledge and Distributed Intelligence Initiative (KDI), the Digital Government program, and the Integrated Graduate Education and Research Training (IGERT) program. A total of more than $200 million in federal information technology research spending has been invested in projects with which he has been associated. Dr. Chen is also a recognized advisor for international IT research programs in Hong Kong, China, Taiwan, Korea, and Ireland. ●

Foreword

This book's purpose is to present a balanced and integrated view of what a Knowledge Management System (KMS) is.

We first define Knowledge Management (KM) from various consulting and IT perspectives and then pay particular attention to new and emerging technologies that help promote this new field. In particular, we present a review of some key KMS sub-fields: search engines, data mining, and text mining. We hope to help readers better understand the emerging technologies behind knowledge management, i.e., Knowledge Management Systems.

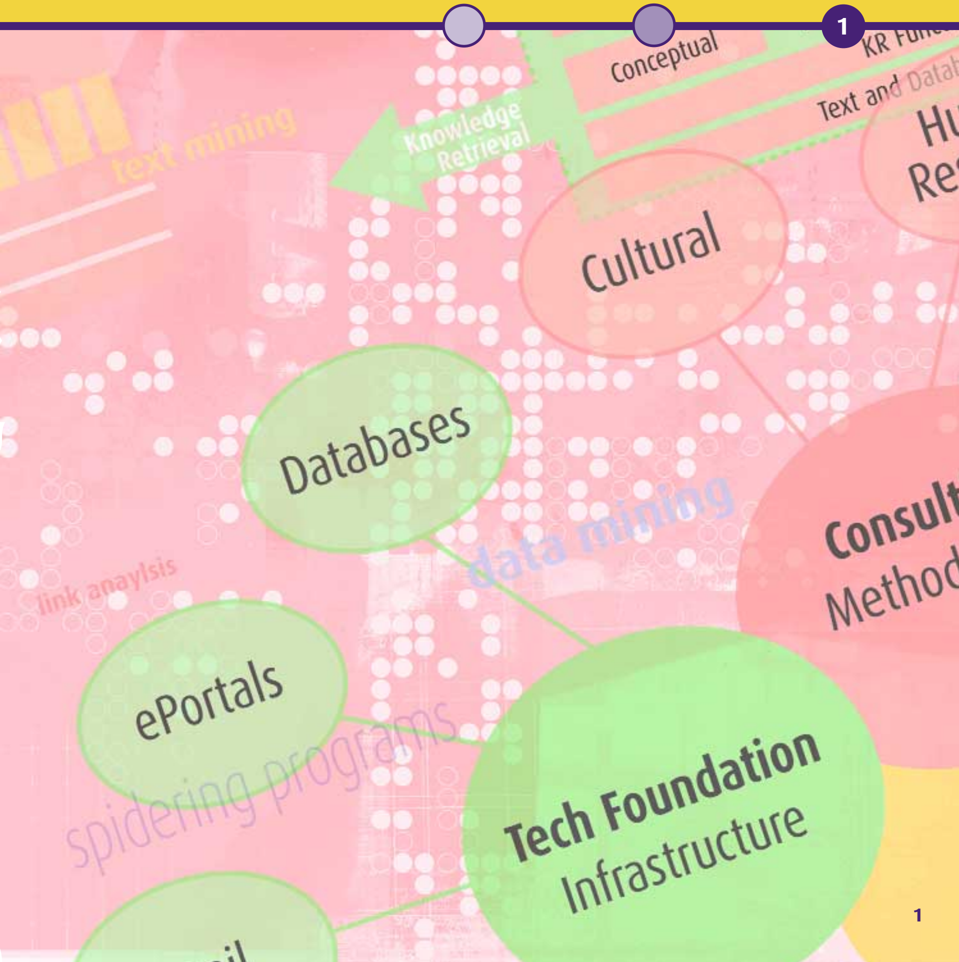
A high-level, although systematic, discussion of text mining is presented. Unlike search engines and data mining that have a longer history and are better understood, text mining is an emerging technical area that is relatively unknown to IT professionals. We therefore present several case studies and conclude with lessons learned and future research and development directions.

This book is intended to provide a gentle introduction to researchers and IT professionals who are new to KMS. We hope it provides a non-technical and practical review of this fascinating field as well as a look at the potential and pitfalls of these emerging technologies.

We would like to acknowledge many organizations and people who have made significant contributions to this field. In particular, we have drawn references from several consulting groups' research reports and industry magazine surveys. We benefited from discussions with many colleagues in the fields of information retrieval, digital libraries, artificial intelligence, and data mining. Several case studies were based on findings originally supported by federal research programs, in particular, National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), National Institute of Justice (NIJ), and National Institutes of Health (NIH).

Last, but not least, we would like to thank faculty members and colleagues at the Eller College of the University of Arizona and the Department of Management Information Systems who have provided many research insights. More importantly, ongoing research could not be accomplished without the dedicated work of my colleagues, co-workers and students in the Artificial Intelligence Lab at the University of Arizona and the university spin-off company, Knowledge Computing Corporation (KCC). ●

Knowledge Management Systems



Knowledge Management Systems

Background

Before discussing Knowledge Management, we need first to understand the unit of analysis, namely, knowledge.

It is generally agreed by IT practitioners that there exists a continuum of data, information, and knowledge (and even wisdom) within any enterprise. The concept of data and the systems to manage them began to be popular in the 1980s. Data are mostly structured, factual, and oftentimes numeric. They often exist as business transactions in database management systems (DBMS) such as Oracle, DB2, and MS SQL. Information, on the other hand, became a hot item for businesses in the 1990s, especially after the Internet web explosion and the successes of many search engines. Information is factual, but unstructured, and in many cases textual. Web pages and email are good examples of “information” that often exists in search engines, groupware, and document management systems. Knowledge is inferential, abstract, and is needed to support business decisions.

In addition to the IT view of the data-information-knowledge continuum, other researchers have taken a more academic view. According to these researchers, data consist of facts, images, or sounds. When data are combined with interpretation and meaning, information emerges. Information is formatted, filtered, and summarized data that, when combined with action and application becomes knowledge. Knowledge exists in forms such as instincts, ideas, rules, and procedures that guide actions and decisions.

The concept of knowledge has become prevalent in many disciplines and business practices. For example, information scientists consider taxonomies, subject headings, and classification schemes as representations of knowledge. Artificial intelligence researchers have long been seeking such ways to represent human knowledge as semantic nets, logic, production systems, and frames. Consulting firms have also been actively promoting practices and methodologies to capture corporate knowledge assets and organizational memory. Since the 1990s, knowledge management has become a popular term that appears in many applications, from digital library to search engine, and from data mining to text mining. Despite its apparent popularity, we believe the field is rather disjointed and new knowledge management technologies are relatively foreign to practitioners.

Definition

We adopt a layman’s definition of knowledge management in this book. Knowledge management is the system and managerial approach to collecting, processing, and organizing enterprise-specific knowledge assets for business functions and decisions. Notice that we equally stress both the managerial (consulting) and also the system (technology) components.

It is our belief that where a managerial approach lacks a sound technical basis, we will see KM become another casualty of consulting faddism, much as did Business Process Reengineering (BPR) or Total Quality Management (TQM), which, in many cases, did not deliver sustainable values to customers. On the other hand, new KM technologies will fall into misuse or produce unintended consequences if they are not properly understood and administered in the proper organizational and business context.

In light of corporate turnover, information overload, and the difficulty of codifying knowledge, knowledge management faces daunting challenges to making high-value corporate information and knowledge assets easily available to support decision making at the lowest, broadest possible levels.

The Landscape

Knowledge management has academic roots in several research communities that have been developing new technologies and researching their organizational impacts.

The NSF-lead, multi-agency, multi-million-dollar Digital Library Initiative (DLI) has attracted researchers from computer science, information systems and sciences, and social sciences to research issues related to content creation, information management, knowledge extraction, and organizational adoption of different technologies and practices. Similarly, the NSF Digital Government Initiative, the Knowledge Networking Initiative, and the Information Technology Research programs aim to make information and knowledge assets more easily available to scientists, policy makers, practitioners, and the general public.

Such federally funded research programs have also fostered involvement of several key related research disciplines in knowledge management, including data mining, search engines, and information visualization.

While academic communities and federal governments often focus on mid-to long-term research, IT businesses and consulting firms have eagerly embraced methodologies and frameworks that are suitable for immediate corporate knowledge transformation, although they tend to focus on the managerial dimension, paying less attention to emerging technologies.

Several communities have collectively helped contribute to the development of knowledge management.

The consulting community, which is grounded on information system analysis and design methodology, often takes a process perspective. Its members stress best practices, process modeling, learning/education paradigms, human resources, culture and rewards, and systematic methodologies.

Consultants often adopt knowledge management methodologies based on existing and proven technical foundations. Data warehousing, email, e-portals, document management systems, and search engines are examples of popular KM implementation platforms. Despite recognition of the importance of methodologies and technical foundations, their systems suffer from an inability to effectively extract and codify corporate information and knowledge assets.

A content management perspective is exemplified by researchers and practitioners trained in information or library sciences. Stressing content management and system usability, knowledge represented as taxonomies, knowledge map, or ontology are created and maintained by information specialists. However, the process of manual knowledge codification is painstaking and error-prone. A system-aided approach is called for.

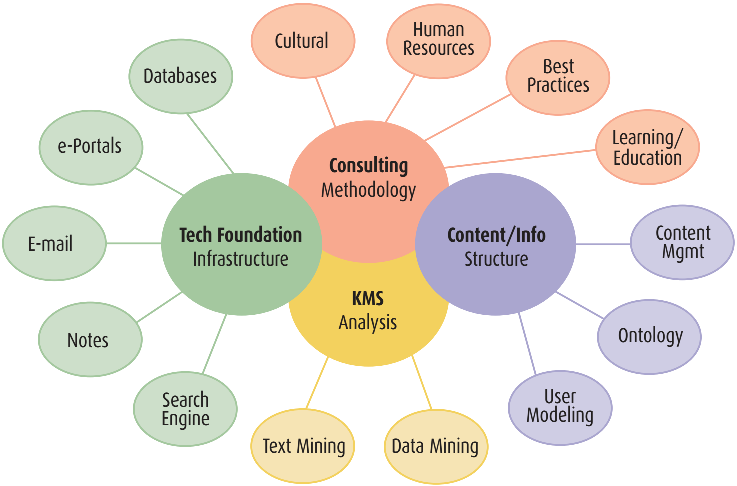


FIGURE 1 KM consists of four main perspectives: Consulting, Content/Information, Technology Foundation, and Knowledge Management System (KMS). KMS includes data mining and text mining.

A significant (emerging) approach to knowledge management is represented by researchers and practitioners who attempt to codify and extract knowledge using automated, algorithmic, and data-driven techniques. We define systems that adopt such techniques as Knowledge Management

Systems (KMS), a class of new software systems that have begun to contribute to KM practices. A KMS focuses on analysis and is the subject of our discussion in this book.

Two of the most relevant sub-fields within knowledge management are data mining and text mining. Data mining, which is better known within the IT community, performs various statistical and artificial intelligence analyses on structured and numeric data sets. Text mining, a newer field, performs various searching functions, linguistic analysis, and categorizations. KMS complements existing IT infrastructure and often requires being superimposed on such foundational systems as e-portals or search engines. Methodologies for practicing these new techniques must be developed if they are to be successful.

The focus of our discussion in this book will be on text mining. However, we will briefly introduce search engines (that focus on text search) and data mining (that focuses on data analysis) as two related and well-known "siblings" of text mining.

Consulting Perspective

Many consulting firms and IT vendors have developed methodologies for their practice of KM. Most, however, have limited experiences in adopting automated knowledge codification and analysis techniques.

Dataware Technology has suggested a 7-step KM methodology that includes: (1) identify the business problem, (2) prepare for change, (3) create a KM team, (4) perform a knowledge audit and analysis, (5) define the key features of the solution, (6) implement the building blocks for KM, and (7) link knowledge to people. Steps (4), (6), and (7) are problematic unless they are provided automated, system-aided support.

The Delphi Group's KM practices look at: (1) key concepts and frameworks for KM, (2) how to use KM as a competitive tool, (3) the culture and organization aspects of KM,(4) best practices in KM, (5) the technology of KM, (6) market analysis, (7) justifying KM, and, (8) implementing KM. Items (5), (6), and (8) require new KM technologies.

Accenture (formerly Andersen Consulting) suggests a simple plan with 6-steps: (1) acquire, (2) create, (3) synthesize, (4) share, (5) use to achieve organizational goals, and (6) establish an environment conducive to knowledge sharing. The "create" and "synthesize" phases are often difficult and problematic.

PriceWaterhouseCoopers has adopted a 5-step approach: (1) find, (2) filter (for relevance), (3) format (to problem), (4) forward (to right people), and (5) feedback (from users). Steps 1-4 require system-aided supports.

Ernst & Young, one of the most savvy KM consulting firms, promotes a 4-phase KM approach consisting of: (1) knowledge generation, (2) knowledge representation, (3) knowledge codification, and (4) knowledge application. However, like most other major consulting firms, they have only begun to include new data mining and text mining techniques within their KM practices.

KM Survey

A recent survey conducted jointly by IDC and *Knowledge Management Magazine* in May 2001 reported the status of KM practices in U.S. companies (Dyer & McDonough, 2001).

Among the top three reasons for a company's adopting KM are: (1) retaining expertise of personnel, (2) increasing customer satisfaction, (3) improving profits or increasing revenues. KM is clearly suited to capturing both internal (employees') and external (customers') knowledge.

The majority of the KM projects (29.6%) are cross functional. In second place, 19.4% of KM projects are initiated by a CEO, rather than by other functional executives. More than 50% of KM projects are completed within 2 years.

The most significant KM implementation challenge is not due to lack of skill in KM techniques. The top four implementation challenges are non-technical in nature: (1) employees have no time for KM, (2) the current culture does not encourage sharing, (3) lack of understanding of KM and its benefits, and (4) inability to measure financial benefits of KM. It seems clear that significant KM education, training, and cultural issues will have to be addressed in most organizations.

Because KM practices are still new to many organizations, it is not surprising that most of the techniques and systems adopted have been basic IT systems, rather than the newer data mining or text mining systems. The most widely used KM software, in ranked order of budget allocations, are: enterprise information portal (e-portal), document management, groupware, workflow, data warehousing, search engine, web-based training, and messaging email.

Knowledge Management Functionalities

The Gartner Group appears to have the best appreciation of Knowledge Management Systems. In one of its reports, a multi-tier Knowledge Management Architecture is presented (Gartner Group, 1999).

At the lowest level, an Intranet and an Extranet that consist of platform services, network services, and distributed object models, are often used

(Dyer & McDonough, 2001) G. Dyer and B. McDonough, "The State of KM," *Knowledge Management*, May 2001, Pages 31-36.

(Gartner Group, 1999) Gartner Group, Knowledge Management Report, Summer, 1999.

as a foundation for delivery of KM applications. Databases and workgroup applications (the first deals with data and the other with assisting people in workgroups) constitute the system components at the next level. In the Gartner Group KM architecture this next level component is called "Knowledge Retrieval" (KR), which consists of text and database drivers (to handle various corporate data and information assets), KR functions, and concept and physical knowledge maps. Above the KR level, a web user interface is often used in business applications.

Two things are especially notable. First, the Gartner Group's KM architecture consists of applications and services that are layered and have complementary roles. No single infrastructure or system is capable of serving an organization's complete KM needs. Second, Knowledge Retrieval is considered the newest addition to the existing IT infrastructure and is the core of the entire KM architecture.

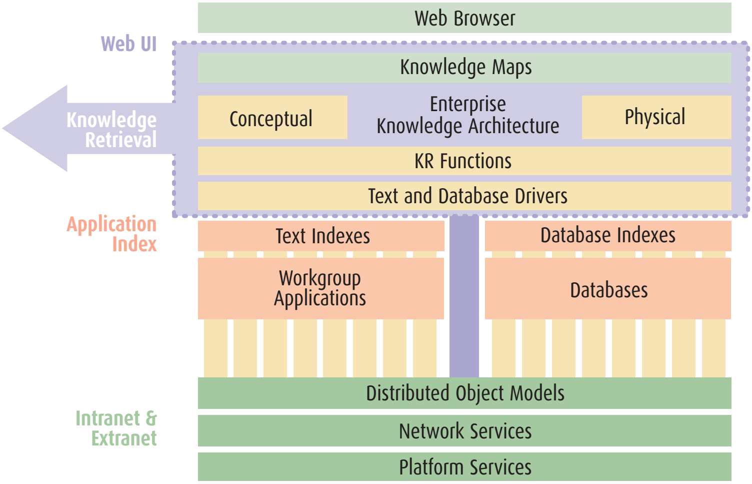


FIGURE 2 The Gartner Group's Knowledge Management Architecture consists of multiple layers and technical components. Knowledge Retrieval is considered the core in this architecture.

The Gartner Group presents KR functions in two dimensions. In the "semantic" dimension, bottom-up system-aided techniques that include data extraction, linguistic analysis (to process text), thesauri, dictionaries, semantic networks, and clustering (categorization/table of contents) are used to create an organization's Concept Yellow Pages. Such Concept Yellow Pages are used as organizational knowledge maps (both conceptual and physical). The proposed techniques consist of both algorithmic processes and ontology generation and usage.

In the second "collaboration" dimension, the goal is to achieve "value recommendations" identified by experts and trust advisors, community building activities, and collaborative filters. Domain experts who hold valuable tacit knowledge in an organization can then be explicitly identified and can be consulted for critical decisions.

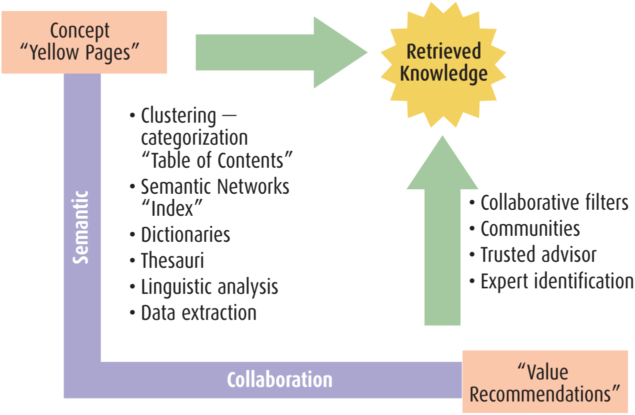


FIGURE 3 Gartner Group's Knowledge Retrieval functions consist of two dimensions: Semantic and Collaboration, each with a set of techniques and components.

The Gartner Group report also presents a KM landscape that consists of several types of industry players. Information retrieval (IR) vendors such as Verity, Excalibur, and Dataware are refining their product functionalities from text retrieval to knowledge management. They have significant experience in text processing. There are also niche document management companies such as PCDOC and Documentum that have developed successful products in managing document content and workflows. Large IT platform companies such as Oracle, Lotus, and Microsoft are aiming to improve the KR functionalities of their popular database or workgroup products. Despite their historical successes, these companies lack significant linguistic and analytical (mining) abilities to create ultimate knowledge maps for organizations. They should be referred to as Information Management vendors instead of Knowledge Management vendors—there is little knowledge analysis or support in their systems.

The last type of vendor in the KM landscape consists of smaller, newer companies and start-ups such as Autonomy, Perspecta, InXight, Semio, Knowledge Computing Corporation (KCC), etc. This set of companies has new linguistic and analytical technologies, but lacks execution experience and integration ability.

KM start-ups differ widely from many Internet start-ups that often rely on fancy business models or just a fast idea. The KM start-ups often result from multi-year hard-core academic or company research making use of significant algorithmic components. For example, InXight is a technology spin-off from Xerox PARC, Autonomy is a technical company that originated at Cambridge University, and Knowledge Computing Corporation is a spin-off company from the University of Arizona Artificial Intelligence Lab.

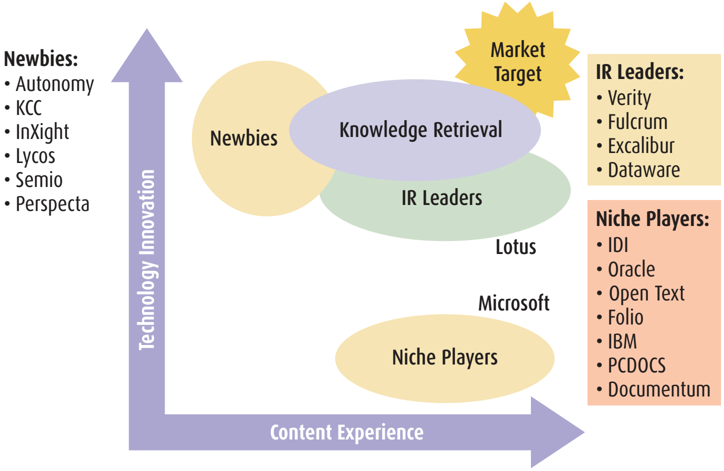


FIGURE 4 The Gartner Group's KM Industry analysis consists of three sets of players: Information Retrieval (IR) Players, Niche Players, and NewBies. Autonomy and Knowledge Computing Corporation (KCC) are among the NewBies.

Search Engine

Knowledge management foundational technologies are many. Among them, the closest siblings to text mining are search engines (that deal with text-based web content) and data mining (that adopts many analytical techniques that are similar to text mining). We briefly review these two better-known subfields of knowledge management and summarize their key system features and business implications.

Search engines rely on text indexing and retrieval technologies much as do those adopted in information retrieval and document management systems. However, search engines have evolved significantly in spidering, ranking, and database capabilities. Most e-portals also contain a significant Internet or Intranet search engine component.

Search engines can be classified by two basic architectures. In *Search Architecture*, spidering programs connect to the web to collect fresh pages that are then indexed via inverted, word-based indexing techniques and are stored in relational databases. Spidering programs need to be spam-resistant and fresh. A search log that stores users' search keywords and other transaction information is a significant component. Browsers are the common vehicles for accessing search engines. Most of the search engines such as Google, Lycos, and InfoSeek belong to this category.

The second architecture is of the *Directory* nature exemplified by Yahoo. A high-precision, low-coverage directory (subject hierarchy) represents subject categories for web sites (instead of web pages) on similar topics. While *Search Architecture* search engines allow users to locate individual web pages, *Directory*-based search engines support user browsing of web sites. Both architectures collect web pages or web sites submitted by users and from automatic spiders (programs) crawling on the web.

The web is, in essence, a graph of web pages connected via hyperlinks. It is estimated that there are currently over two billion web pages and growing fast. A spider is a small Java or C/C++ based program that crawls the web along hyperlinks to fetch remote (live) web pages to the search engine's underlying databases.

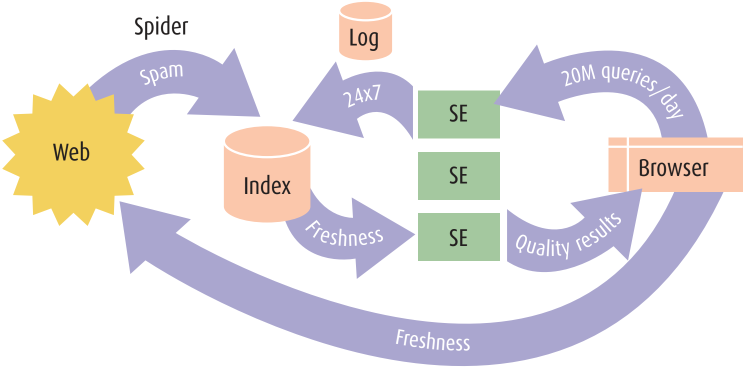


FIGURE 5 The Search Engine (SE) *Search Architecture* consists of spider, index, search log, and browser. Spam and freshness are important for search engines.

Once collected, a web page is indexed by words in its content. Most indexing techniques are based on the traditional inverted index or the vector space model (Salton, 1989) developed by Gerald Salton (generally considered the father of IR). Each keyword is weighted based on the tf x idf function that consists of term frequency (tf) and inverse document frequency (idf). Most search engines maintain 200-500 million popular URLs and have a 50% to 100% overhead in indexing. Some search engines index meta tags and content differently. Some support stemming (removing suffixes), proximity searches (search keywords that appear in certain proximity), and simple natural language processing (NLP) that includes multiple words or removes certain linguistic types (e.g., prepositions such as of, on, in).

Once web pages have been indexed in the databases, they can be accessed by users through a common browser interface. The vector space model creates a weighted keyword score representing the relationship between a user query and any web pages. In addition, most search engines use “inlink” information to provide a quality ranking for each web page retrieved. Based on a form of “popularity” voting, a web page that is pointed to by many quality (and well-known) web sites is considered more relevant. The Page Rank algorithm adopted in Google is a good example of a link analysis technique that is based on a random-walk algorithm with inlink information to selectively collect high-quality web pages (Brin & Page, 1998).

Search engine companies spend an enormous amount of resources to detect “spam” pages (web pages and web sites that are deliberately designed to trick a spidering program to give them high ranking scores). Common spamming techniques include keyword stuffing (placing extraneous keywords multiple times to increase keyword ranking), page-jacking (disguising one’s own web site as another popular and legitimate site by re-routing web traffic), etc. The search engine community is in constant warfare between the vendors and the spammers.

Directory-based search engines require manual categorization and ranking effort. This is labor intensive, painstaking work, but results in excellent precision. Most such search engines contain 500,000 to 800,000 web sites in a manually maintained directory.

None of the search engines completely covers the web, as was reported by Lawrence and Giles (Lawrence & Giles, 1999). In fact, each search engine may

(Salton, 1989) G. Salton, *Automatic Text Processing*. Reading, MA, Addison-Wesley, 1989.

(Brin & Page, 1998) S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine. The 7th WWW Conference, 1998.

(Lawrence & Giles, 1999) S. Lawrence and C. L. Giles, “Accessibility of Information on the Web,” *Nature*, Number 400, 1999, Pages 107-109.

cover only 15% of the web at any given time. The search engine community is also full of hype and excessive claiming, e.g., the largest web repository, the most advanced NLP functionalities, etc. Search Engine Watch (www.searchenginewatch.com) is one of the few resources for unbiased search engine comparison. WWW Conferences also are a source of many interesting and timely reports of new and emerging developments in this community.

Many search engines have attempted to cover other types of content, e.g., newswire, newsgroups, shopping catalogs, Intranet databases (behind firewalls), dynamic content, etc. Vertical search engines are becoming important for both target marketing and web site precision. Search engines for localized markets (e.g., products and services for a given city), shopping softbots, email, newsgroups, chat rooms, and instant messaging are also hot new search engine topics.

Search engine companies have mostly evolved from their technical past to the current media focus that stresses content, service, advertising, and marketing. The majority have abandoned their historically technical and free-sharing roots. New technologies rarely emerge from these companies. Many have formed "keiretsu" with venture capitalists, ad agencies, old media companies, verticals, and even banks, e.g., Kleiner Perkins, AT&T, At Home, Excite, etc.

Some newer search-engine based companies have evolved into e-portals that aim to serve the information needs of a corporate Intranet, e.g., Autonomy and Northern Light. New functionalities such as chat rooms, bulletin boards, calendaring, content pushing, user personalization, publishing, and workflows are added to such a one-stop shopping site for enterprise users. Both internal content (e.g., best practices and communications) and external resources (e.g., industry reports, marketing intelligence) of various formats (e.g., email, power point files, Notes databases, etc.) are captured in such systems.

Despite significant technological advancement of search engines and e-portals, such systems are grounded on basic text processing technologies (inverted index and vector space model) developed in the 1970s. They lack advanced linguistic processing abilities (e.g., noun phrasing, entity extraction—knowing who, what, where, when, etc. in text) and automatic categorization and clustering techniques.

Data Mining

Data mining has emerged over the past decade as an effective business tool for decision support. Many successful applications have been reported including market and sales analysis, fraud detection, manufacturing process analysis, scientific data analysis, etc. Such applications often target structured, numeric transaction data residing in relational data warehouses (Witten & Frank, 2000).

Unstructured, textual document analysis is often considered text mining, a field cross between search engines and data mining. Web mining, the adaptation of data and text mining techniques for web content, transactions, and user behavior is another emerging sub-field of knowledge management (Chen, 2002).

Many disciplines have contributed to data mining including database management systems, statistical analysis, machine learning, neural networks, and visualization.

Data mining is often considered a step within the Knowledge Discovery in Databases (KDD) process, through which an organization's data assets are processed and analyzed in order to gain insights for business decisions (Fayyad, et al., 1996). KDD originates with data residing in a company's relational, transactional database management system or data warehouse. These data are then selected, processed, and transformed to be ready for the data mining step after which the resulting "knowledge" is interpreted and evaluated.

Despite its importance, the data mining phase is not the most dominant step in the KDD process. It is estimated that about 22% of a KDD project effort

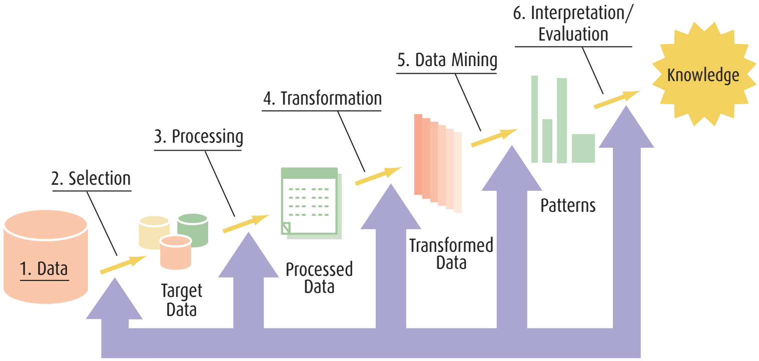


FIGURE 6 A KDD Process consists of many components: Data, Selection, Processing, Transformation, Data Mining, and Interpretation/Evaluation.

(Witten & Frank, 2000) I. H. Witten and E. Frank, Data Mining, Morgan Kaufmann, San Francisco, CA, 2000.

(Chen, 2002) H. Chen. "Web Mining and Retrieval," *International Journal of Decision Support Systems*, 2002.

(Fayyad, et al., 1996) U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, MIT Press, Boston, MA, 1996.

is spent on determination of business objectives, which requires significant joint efforts of IT consultants and business managers. Data preparation, processing, and transformation, which involve significant database work, constitute more than 55% of a KDD effort. Data mining constitutes about 12% of a total KDD effort, while the remaining 10% focuses on analysis, interpretation, and assimilation.

Often considered the "brain" of the entire KDD process, data mining models perform a variety of functions including predictive modeling (classification and prediction), link analysis (associations discovery and sequential pattern discovery), database segmentation (demographic clustering and neural network clustering), and deviation detection (visualization and statistics).

Several popular computing paradigms have been adopted in data mining models.

Statistical analysis relies on stringent mathematical models and assumptions (e.g., normal distribution, independence assumption). Statistical processing is often fast, and its outputs are mathematical and parametric in nature. Well-trained statisticians are often needed for such analyses. Regression analysis, discriminant analysis, factor analysis, principal component analysis, and time series analysis are among the popular statistical analysis techniques.

Neural networks, a computational paradigm in development since the 1960s, became robust and popular in the 1990s (Lippmann, 1987). Mostly grounded on specialized neural network topologies and statistics-based learning or error correction functions, neural networks have been shown to be highly accurate and flexible, but at times slow during training. The multi-layered Feedforward neural network (with the popular Backpropagation algorithm) and Kohonen's Self-Organizing Map (SOM) for clustering are perhaps the best-known and most powerful neural network methods for business data mining.

Symbolic machine learning algorithms that are based on traditional artificial intelligence research have also been adopted in major data mining tools. Techniques such as ID3, C4.5, and CART regression trees are elegant, yet powerful (Quinlan, 1993). Their performances are good, and their results are easy to interpret for business decisions, unlike those of neural networks, which are often treated as a "black box."

In addition to the three popular computational approaches to data mining described, researchers have also adopted other "soft computing" techniques

(Lippmann, 1987) R. P. Lippmann, "An Introduction to Computing with Neural Networks," *IEEE Acoustics Speech and Signal Processing Magazine*, Volume 4, Number 2, April 1987, Pages 4-22.

(Quinlan, 1993) J. R. Quinlan, C4.5: *Programs for Machine Learning*, Morgan Kaufmann, San Francisco, CA, 1993.

such as genetic algorithms, evolutionary programming, fuzzy set, and fuzzy logic. However, their scalability and performances are not as evident as statistical analysis, neural networks, and symbolic learning methods. Most of the popular data mining methods have been incorporated in mature commercial software systems such as: IBM's Intelligent Miner, SPSS, SAS, and SGI's MineSet. In addition, Java and C/C++ freeware are often available in various sources, e.g., the C4.5 book of Quinlan, the data mining textbook of Witten and Frank, and UC Irvine's data mining resource web site.

Unlike search engines, data mining projects are domain-specific, application-dependent, and often require significant business analysis, customization, and refinement. There are some basic requirements for successful data mining. First, there must be a clear sponsor for the application; the business case and value for the application must be clearly understood and measurable; and there must be sufficient business domain knowledge and high-quality, relevant data available.

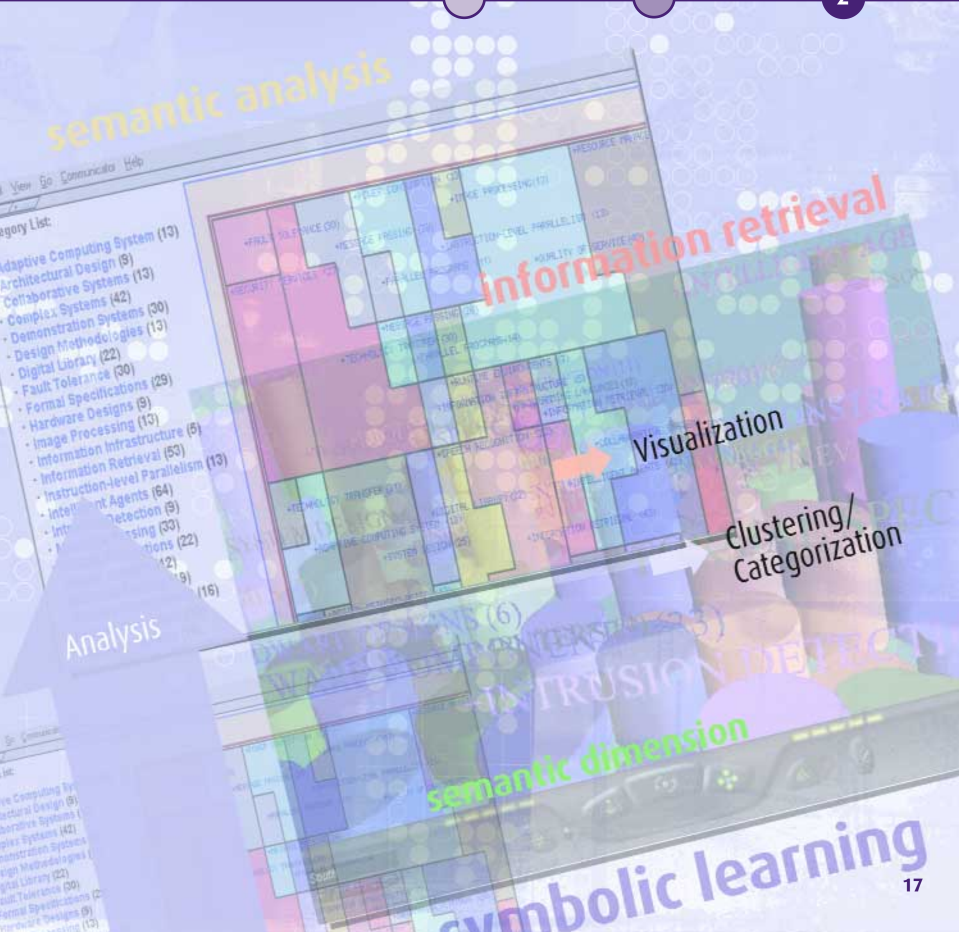
Several areas distinguish text mining from data mining. First, unlike data mining, data characteristics of text mining require significant linguistic processing or natural language processing abilities. Second, data mining often attempts to identify causal relationships through classification (or supervised learning) techniques (e.g., What employee demographic variables affect spending patterns). Text mining, on the other hand, aims to create organizational knowledge maps or concept yellowpages, as described in the Gartner Group Knowledge Management report. Third, text mining applications deal with much more diverse and eclectic collections of systems and formats (email, web pages, Notes databases, newsgroups). Sometimes organization-specific ontologies (e.g., product and service directories) need to be incorporated in order to generate a comprehensive knowledge map.

Both data mining and text mining adopt significant analytical methods and their results are often highly visual and graphical. Data visualization and information visualization techniques attempt to create an interface that is well suited for human decision making.

Even in light of the highly technical nature of the data mining and text mining systems, Knowledge Management requires a balanced managerial and technical approach. Quality content management and data assurance, careful business and system requirement solicitation, effective implementation methodologies, and supportive organizational culture and reward systems are critical for its success. ●

Text Mining

2



Text Mining

IR and AI

As the core of the knowledge management systems, text mining is a cross between information retrieval (IR) and artificial intelligence (AI).

Gerald Salton, a pioneer in IR since the 1970s, is generally considered the father of IR. His vector space model has become the foundation for representing documents in modern IR systems and web search engines.

IR is a field that has gone through several major generations of development. In the 1970s, computational techniques based on inverted indexes and vector spaces were developed and tested in computer systems. In addition, Boolean retrieval methods and simple probabilistic retrieval models based on Bayesian statistics were created. Although more than 30 years old, this set of techniques still forms the basis of modern IR systems.

In the 1980s, coinciding with the developments of new AI techniques, knowledge-based and expert systems that aim to emulate expert searchers and domain specialists were developed. User modeling and natural language processing (NLP) techniques were developed to assist in representing users and documents. Research prototypes were created to represent information specialist (e.g., reference librarian) heuristics for effective online searching.

Realizing the difficulties of creating domain-specific knowledge bases and heuristics, researchers in the 1990s attempted to adopt new machine learning techniques for information analysis. AI techniques, including neural networks, genetic algorithms, and symbolic learning, were tested in IR (Chen, 1995).

Since the mid 1990s, the popularity of search engines and advances in web spidering, indexing, and link analysis have transformed IR systems into newer and more powerful search tools for content on the Internet. The diverse multimedia content and the ubiquitous presence of the web make both commercial users and the general public see the potential for utilizing unstructured information assets in their everyday activities and business decisions.

It is estimated that 80% of the world's online content is based on text. We have developed an effective means to deal with structured, numeric content via database management systems (DBMS), but text processing and analysis is significantly more difficult. The status of knowledge management systems is much like that of DBMS twenty years ago. The real challenges, and the potential payoffs for an effective, universal text solution, are equally

(Chen, 1995) H. Chen, "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," *Journal of the American Society for Information Science*, Volume 46, Number 3, Pages 194-216, April 1995.

appealing. It is inevitable that whoever dominates this space will become the next Oracle (in text).

Herbert Simon, a professor at Carnegie Mellon University, is considered one of the founding fathers of artificial intelligence (AI), which has long been striving to model and represent human intelligence in computational models and systems.

Simon and his colleagues pioneered the early research in AI, most notably by creating General Problem Solvers (GPS) that emulated general human problem solving. By the 1970s, computer programs were developed to emulate rudimentary but human-like activities such as cryptarithmetic, chess, games, puzzles, etc (Newell & Simon, 1972) .

In the 1980s, there was an explosion of AI research activities, most notably in expert systems. Many research prototypes were created to emulate expert knowledge and problem solving in domains such as medical and car diagnosis, oil drilling, computer configuration, etc. However, the failure of many such systems in commercial arenas led many venture capitalists to back away from any ventures associated with AI.

Nevertheless, commercial expert systems have made both researchers and practitioners become realistic about the strengths and weaknesses of such systems. Expert systems may not be silver bullets but they have been shown to be suited for well-defined domains with willing experts.

In the 1990s, AI-based symbolic learning, neural-network, and genetic-programming technologies have generated many significant and useful techniques for both scientific and business applications. The field of data mining is the result of significant research developed in this era. Many companies have since applied such techniques in successful fraud-detection, financial-prediction, web-mining, and customer-behavioral analysis applications.

Both IR and AI research have contributed to a foundation for knowledge representation. For example, indexing, subject heading, dictionaries, thesauri, taxonomies, and classification schemes are some of the IR knowledge representations still widely used in various knowledge management practices. AI researchers, on the other hand, have developed knowledge representation schemes such as semantic nets, production systems, logic, frames, and scripts.

With the continued expansion and popularity of web-based scientific, governmental, and e-commerce applications in the 2000s, we foresee active research leading to autonomous web agents with learning and data mining abilities. The field of web mining promises to continue to provide a challenging test bed for advancing new IR and AI research.

(Newell & Simon, 1972) A. Newell and H. Simon, *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

The Industry

As the Gartner Group report suggested, the new KMSs require a new set of knowledge retrieval (KR) functionalities, broadly defined in two areas: the semantic dimension (with the goal of creating concept yellowpages) and the collaboration dimension (with the goal of providing value recommendations from experts). Such functionalities can be used to create knowledge maps for an organization.

As evident in the *Knowledge Management Magazine* survey, most of the software systems currently in use for knowledge management purposes are of the basic infrastructure types (e.g., e-portals, groupware, search engines, document management). The IR vendors (e.g., Verity, OpenText), document management companies (e.g., PCDOC, Documentum), and groupware/ email firms (e.g., Microsoft, Lotus, Netscape) are essential to knowledge management deployment but generally lack critical analytical and linguistic processing abilities.

New knowledge management companies excel in text mining ability but often lack execution and delivery abilities (e.g., Autonomy, Knowledge Computing Corporation, inXight, and Semio). Autonomy is probably the most

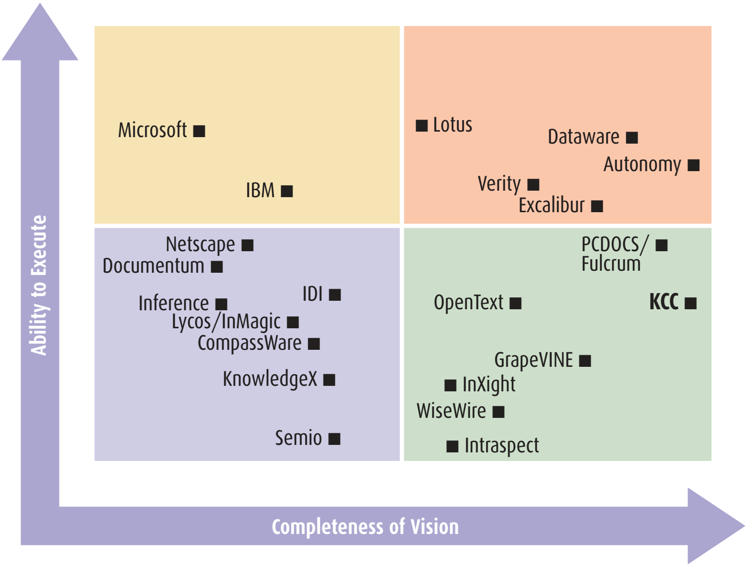


FIGURE 7 Autonomy is performing well in the emerging KM product market. KCC has significant technologies and visions but lacks execution experience at this point.

successful company in this category so far. Founded by Michael Lynch in 1997, Autonomy has successfully integrated Bayesian statistics technologies into its e-portal product, The company experienced significant growth stage during the Internet explosion phase (IPO) and has suffered stock devaluation similar to many IT firms in early 2001. Autonomy products are often used to create e-portals for corporate Intranets or for news media applications.

Technology comparisons among these IR, document management, and new-style knowledge management companies can be classified in three areas: text analysis ability, collection creation and processing ability, and retrieval and display ability.

Text analysis includes such features as natural or statistical language processing, indexer or phrase creator, entity extraction (for entity types), conceptual associations (automatic thesauri), domain-specific knowledge filter (using vocabularies or ontologies), automatic taxonomy creation (clustering), multi-document format support (e.g., Word document, Power Point file, web pages), and multi-language support. Most of the new KM vendors, especially Knowledge Computing Corporation (KCC), are more complete in offering such features, with the exception of Autonomy, which is primarily statistics-based

	Vendors									
Text Processing & Analysis Techniques	KCC	Autonomy	Hummingbird (DOCS/Fulcrum)	Open Text (Leading Side/ Dataware/Sovereign Hill)	Verity	Excalibur	Documentum	Semio	InXight	e-Gain (Inference)
Natural/Statistical Language Processing	x	Bayesian statistics		probablistic model		x		x	algorithms	x
Indexer/Phrase Creator	x				x	x			x	
Entity Extractor	x									
Conceptual Associations/ Thesaurus	x		x	x	x	x		x	x	x
Domain-Specific Filter using Manually Developed Vocabularies/Ontologies	x			flexible filtering				x		
Automatic Taxonomy/ Clustering	x	x	x	x	x			x	x	
Multi-document Format Support		x	x	x	x	x	x	x		
Multi-language Support	x	x	x		x		x	x		

FIGURE 8 KMS Industry Competitive Analysis: Text Analysis Abilities

Collection Creation/ Processing Techniques	Vendors									
	KCC	Autonomy	Hummingbird (DOCS/Fulcrum)	Open Text (Leading Side/ Dataware/Sovereign Hill)	Verity	Excalibur	Documentum	Semio	InXight	e-Gain (Inference)
Spider (HTTP Document Collection)	X	X	X			X				
Data Warehousing	X			X				X	X	
Content Categorization	X	X	X		X	X		X	X	
Hyperlink Creation		X		X						
Community Content Development/Sharing		X								
Automatic Document Summarization	X	X	X	X				X	X	

FIGURE 9 KMS Industry Competitive Analysis: Collection Creation and Processing Abilities

(i.e., without a linguistic engine). However, most vendors do not provide supports for phrasal extraction, entity extraction, or conceptual associations.

Autonomy is strongest in its collection creation and processing ability, which consists of spidering (http document collection), data warehousing, content categorization, hyperlink creation, community content development and sharing, and automatic document summarization. Most IR document management software and new KM vendors have deficiencies in some aspects of collection creation and processing.

In retrieval and display, Autonomy, along with most major IR and document software firms, are doing an excellent job. They typically support such func-

Retrieval/Display/ Delivery Techniques	Vendors									
	KCC	Autonomy	Hummingbird (DOCS/Fulcrum)	Open Text (Leading Side/ Dataware/Sovereign Hill)	Verity	Excalibur	Documentum	Semio	InXight	e-Gain (Inference)
Search Engines	X	X	X	X	X	X		X	X	X
Visualizer(s)	X							X	X	
Security/Authentication				X			X			
Wireless Access	X	X					X			
Metadata XML Tagger		X		X	X		X		X	
Personalized Delivery		X			X	X	X			

FIGURE 10 KMS Industry Competitive Analysis: Retrieval and Display Abilities

tionalities as search engine, visualizer, security and authenticator, wireless access, metadata or XML tagger, and personalized delivery.

Collection creation and processing ability and the retrieval and display ability both require significant system integration with an existing IT infrastructure. Text analysis and processing, however, are algorithmic in nature and are considered unique additions to the knowledge management functionalities. Such techniques are new, but they nevertheless exhibit overwhelming potential for business text mining.

Text Analysis Techniques

Core KMS text mining analysis techniques can be classified into four main layers: linguistic analysis/NLP, statistical/co-occurrence analysis, statistical and neural networks clustering/categorization, and visualization/HCI.

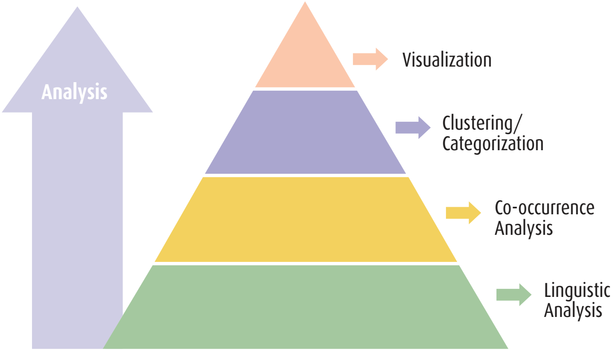
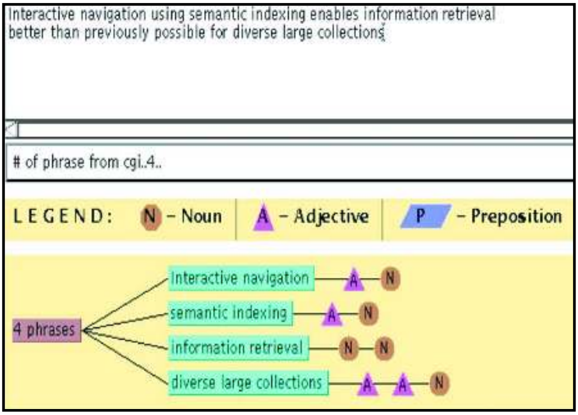


FIGURE 11 The Text Mining Techniques Pyramid

At the lowest level, linguistic analysis and natural language processing (NLP) techniques aim to identify key concept descriptors (who/what/when/where) embedded in textual documents. Different types of linguistic analysis techniques have been developed. Word and inverted index can be combined with stemming, morphological analysis, Boolean, proximity, range, and fuzzy search. The unit of analysis is word. Phrasal analysis, on the other hand, aims to extract meaningful noun phrase units or entities (e.g., people names, organization names, location names). Both linguistic and statistical (such as mutual information) techniques are plausible. Sentence-level analysis including context-free grammar and transformational grammar can be performed to represent grammatically correct sentences. In addition, semantic analysis based on techniques such as semantic grammar and case grammar can be used to represent semantics (meaning) in sentences and stories. Semantic analysis is often domain-specific and lacks scalability.

Based on significant research in the IR and computational linguistics research communities (e.g., TREC and MUC Conferences sponsored by DARPA), it is generally agreed that phrasal-level analysis is most suited for coarse but scalable text mining applications. Word-level analysis is noisy and lacks precision. Sentence-level is too structured and lacks practical applications. Semantic analysis often requires a significant knowledge base or a domain lexicon creation effort and therefore is not suited for general-purpose text mining across a wide spectrum of domains. It is not coincidental that most of the subject headings and concept descriptors adopted in library classification schemes are noun phrases.



Interactive navigation using semantic indexing enables information retrieval better than previously possible for diverse large collections

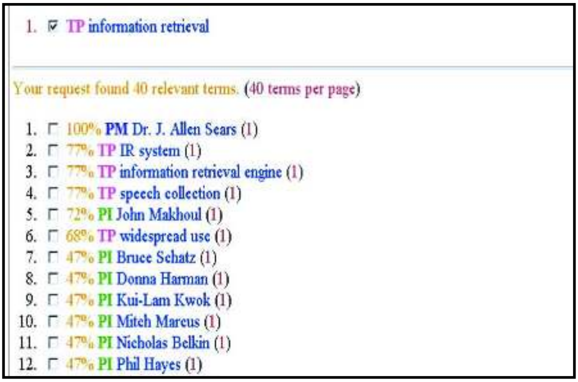
# of phrase from cgi.4..

LEGEND:
N - Noun
A - Adjective
P - Preposition
4 phrases
Interactive navigation A N
semantic indexing A N
information retrieval N N
diverse large collections A A N

FIGURE 12 Arizona Noun Phraser parses English sentences with part-of-speech-tagger and lexicons to produce meaningful noun phrases, e.g., "interactive navigation," "information retrieval."

Based on statistical and co-occurrence techniques, link analysis is performed to create automatic thesauri or conceptual associations of extracted concepts. Similarity functions, such as Jaccard or Cosine, are often used to compute co-occurrence probabilities between pairs of concepts (noun phrases). Some systems use bi-gram, tri-gram, N-gram, or Finite State Automata (FSA) to further capture frequent concept association patterns. Existing human-created dictionaries or thesauri can also be integrated with the system-generated concept thesauri.

Statistical and neural network based clustering and categorization techniques are often used to group similar documents, queries, or communities



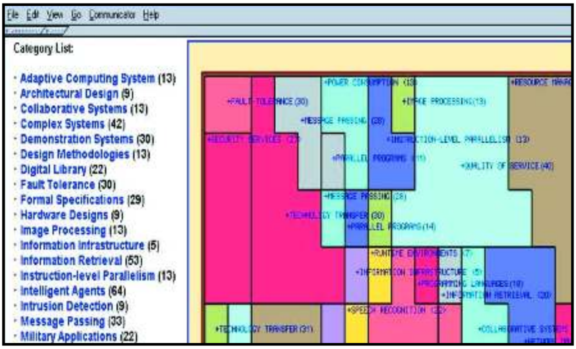
1. ☑ TP information retrieval

Your request found 40 relevant terms. (40 terms per page)

1. □ 100% PM Dr. J. Allen Sears (1)
2. □ 77% TP IR system (1)
3. □ 77% TP information retrieval engine (1)
4. □ 77% TP speech collection (1)
5. □ 72% PI John Makhoul (1)
6. □ 68% TP widespread use (1)
7. □ 47% PI Bruce Schatz (1)
8. □ 47% PI Donna Harman (1)
9. □ 47% PI Kui-Lam Kwok (1)
10. □ 47% PI Mitch Marcus (1)
11. □ 47% PI Nicholas Belkin (1)
12. □ 47% PI Phil Hayes (1)

FIGURE 13 Highly related concepts (noun phrases/terms) are identified based on co-occurrence analysis, e.g. "Dr. J. Allen Sears," "IR System."

in subject hierarchies, which could serve as corporate knowledge maps. Hierarchical clustering (single-link or multi-link) and statistical clustering (e.g., multi-dimensional scaling, factor analysis) techniques are precise but often computationally expensive. Neural network clustering, as exemplified by Kohonen's self-organizing map (SOM) technique, performs well and fast. It is our experience that such a technique is most suited for large-scale text mining tasks. In addition, the SOM technique lends itself to intuitive, graphical visualization based on such visual parameters as: size (a large region represents a more important topic) and proximity (related topics are grouped in adjacent regions).



Category List:

- Adaptive Computing System (13)
- Architectural Design (9)
- Collaborative Systems (13)
- Complex Systems (42)
- Demonstration Systems (30)
- Design Methodologies (13)
- Digital Library (22)
- Fault Tolerance (30)
- Formal Specifications (29)
- Hardware Designs (9)
- Image Processing (13)
- Information Infrastructure (5)
- Information Retrieval (53)
- Instruction-level Parallelism (13)
- Intelligent Agents (64)
- Intrusion Detection (9)
- Message Passing (33)
- Military Applications (22)

FIGURE 14 SOM technique produces 2D graphical knowledge map (cyber-map) of important topics, e.g., "Quality of Service," "Technology Transfer," "Digital Library." The same 2D knowledge map is also displayed in alphabetical order on the left-hand side.

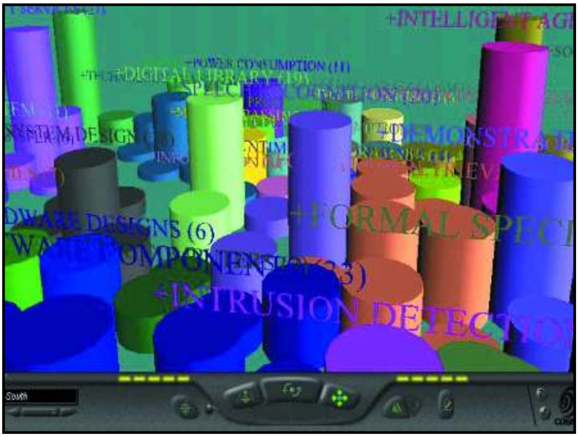


FIGURE 15 The same 2D knowledge map can be displayed in 3D, where the height of cylinders represents number of documents. Users can navigate with a VRML interface.

Visualization and human-computer interactions (HCI) help to reveal conceptual associations and visualize knowledge maps. Different representation structures (e.g., tree, network), display dimensions (1D, 2D, 3D), and interaction techniques (e.g., zooming, spotlight) can be adopted to reveal knowledge more completely. For example, the same 2D SOM representation can be visualized in 3D, by using a helicopter style navigation based on VRML. However, such advanced information visualization may not be practical for business applications until more HCI field research has been performed.

OOHAY and Benefits

It is our belief that the old way of creating subject hierarchies or knowledge maps based on human efforts (such as the Yahoo's directory structure) is not practical or scalable. The existing amount of business information, the speed of future information acquisition activities, and the amount of human effort involved make the manual approach obsolete. Only by leveraging various system-based computational techniques can we effectively organize and structure the ever-increasing textual (and maybe multimedia) content into a useful Object Oriented Hierarchical Automatic Yellowpage (OOHAY)—a computational approach that we believe in and is the reverse of Yahoo!

Implementing an OOHAY approach to knowledge management offers multiple benefits. First, a system-aided approach could help alleviate the "information overload" problem faced by practitioners and managers.

Oftentimes, conventional search engines return too much instead of too little information. A dynamic, customizable analysis component based on text mining techniques can help alleviate such a search bottleneck. Secondly, the system-generated thesaurus can help resolve the "vocabulary differences" faced by searchers—user search terms are often different from a database's index terms, and thus return low relevant results. The system-aided approach also helps to reduce the amount of time and effort required for creating enterprise-specific knowledge maps. The system-generated knowledge maps are often more fine-grained and precise than the human-generated knowledge maps. Organizations could benefit from such an efficient, high-quality, and cost-effective method for retaining internal and external knowledge. In the next chapter, we summarize several case studies and their implications to businesses.

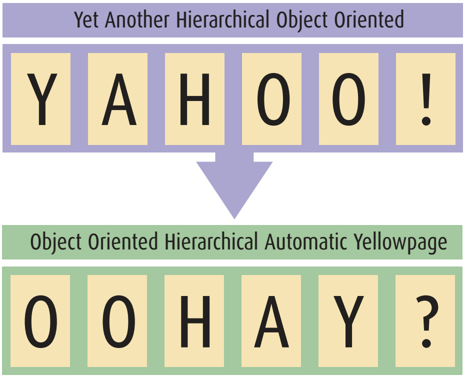
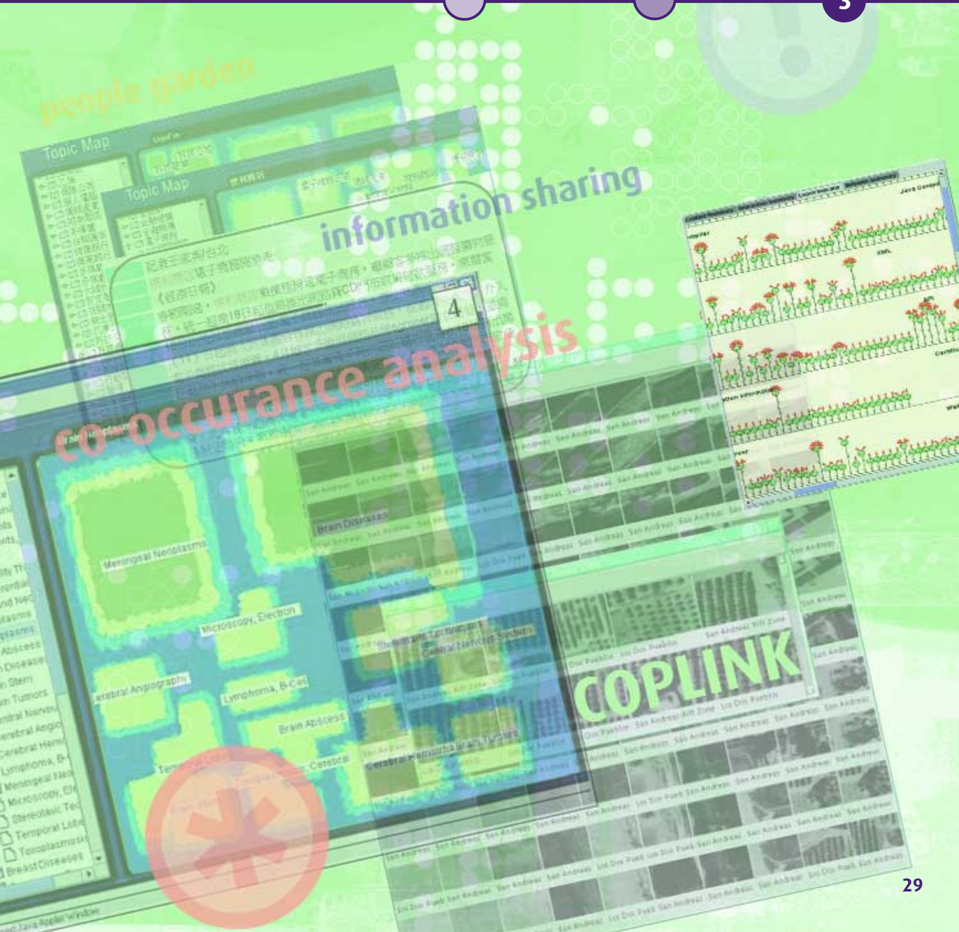

Yet Another Hierarchical Object Oriented
Y A H O O !
Object Oriented Hierarchical Automatic Yellowpage
O O H A Y ?

FIGURE 16 We believe an automated OOHAY approach to knowledge management is necessary to complement the manual YAHOO classification approach.

Case Studies

3



Case Studies

Based on significant academic research conducted at the Artificial Intelligence Lab of the University of Arizona and commercial R&D performed at the Knowledge Computing Corporation (KCC), we summarize several case studies that utilized selected text mining technologies in different applications and organizational contexts. We hope to highlight key application characteristics, system functionalities, and lessons learned in these actual case studies.

We would like to acknowledge the support and assistance of the following

Web Analysis

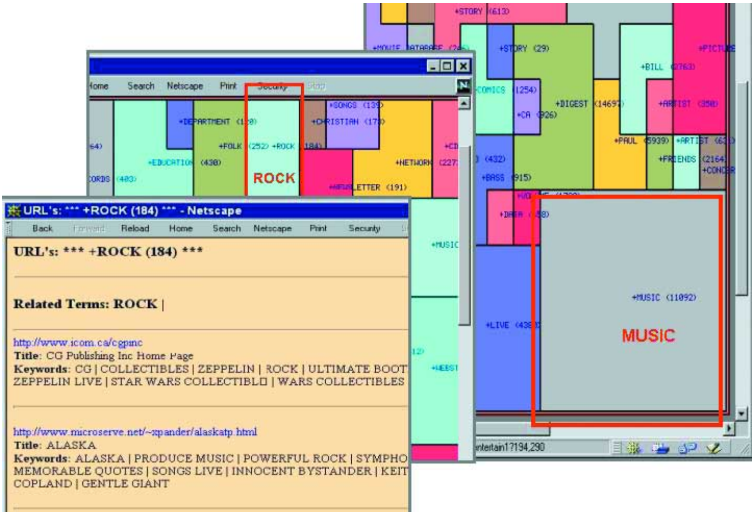
PROBLEM: Web content has become the fastest growing information asset in both Internet and Intranet. Directory-based search engines such as Yahoo organize web content (sites) based on their subject categories. Given the painstaking process of manual categorization, can web content be automatically analyzed, indexed, and categorized?

SOLUTION: Based on selected text mining techniques described earlier, a test bed of about 100,000 entertainment-related web pages was created by multi-threaded Internet spiders. After the web pages had been collected, high-precision noun phrase indexing was performed on each web page. A vector space model of noun phrases and their associated weights was used to represent each web page. All web pages were then categorized by a self-organizing map (SOM) clustering program. A multi-layered, graphical, 2D jigsaw-puzzle-like "knowledge map" was created by the program, with each jigsaw piece representing a given topic. The pieces varied in size with a large jigsaw piece representing a more important concept. Similar topics appeared adjacent to each other on the knowledge map. Users were able to browse this 2D knowledge map of the web to locate topics of interest, much as they can browse in Yahoo (Chen, et al., 1998).

LESSONS LEARNED: Web pages, especially personal web pages dealing with entertainment topics, are very noisy. Typos, grammatical errors, and html irregularities occur frequently. In order to increase indexing precision, different filters and stopwords must be used to filter out noise. An SOM algorithm was found to be a computationally efficient tool (2-3 hours to analyze a 100,000-page collection on a mid-size workstation). However, performance (recall/precision) dropped to only about 50% due to content noise. The resulting SOM knowledge map was very interesting for serendipity browsing, but its effectiveness for searching (a

organizations: National Science Foundation (NSF), Defense Advanced Research Project Agency (DARPA), National Institutes of Health (NIH), National Institute of Justice, National Center for Supercomputing Applications (NCSA), Arizona Cancer Center, Tucson Police Department, Phoenix Police Department, United Daily News, Knowledge Computing Corporation, and the Artificial Intelligence Lab at the University of Arizona's Eller College of Business and Public Administration.

specific topic) was not as apparent. Instead of hiring 50 ontologists for multiple years to organize web sites/pages much like what has been done in a Yahoo directory, the proposed OOHAY approach could create a fine-grained topic-based directory of a similar size in less than one day on a multi-processor computer. Periodic update of such a directory can also be performed in less than a day. We believe that with the higher-precision content within a corporate Intranet, the proposed spidering, noun phrasing, and SOM techniques would show promising performance. ●



URL's: *** +ROCK (184) *** - Netscape

URL's: * +ROCK (184) *****

Related Terms: ROCK |

http://www.icom.ca/cgpinc
Title: CG Publishing Inc Home Page
Keywords: CG | COLLECTIBLES | ZEPPELIN | ROCK | ULTIMATE BOOT
ZEPPELIN LIVE | STAR WARS COLLECTIBL□ | WARS COLLECTIBLES

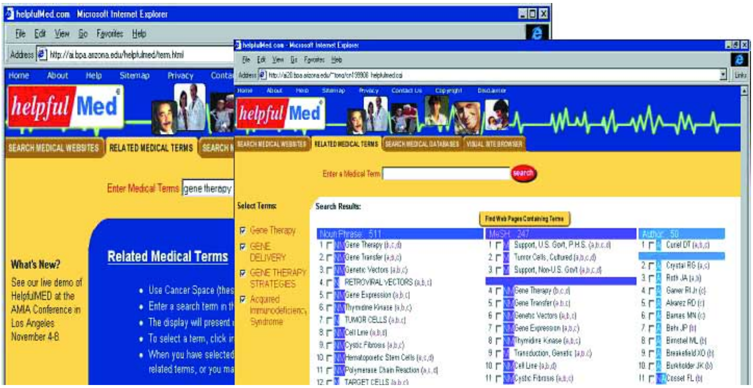
http://www.microserve.net/~xpander/alaskatp.html
Title: ALASKA
Keywords: ALASKA | PRODUCE MUSIC | POWERFUL ROCK | SYMPHO
MEMORABLE QUOTES | SONGS LIVE | INNOCENT BYSTANDER | KEIT
COPLAND | GENTLE GIANT

FIGURE 17 A top-level map reveals many entertainment-related topics, e.g., "MUSIC." Clicking on a topic brings about a second-level map that contains more fine-grained topics, e.g., "ROCK." At the lowest level, clicking on a region reveals web pages that belong to the topic.

(Chen, et al., 1998) H. Chen, A. L. Houston, R. R. Sewell, and B. R. Schatz, "Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques," *Journal of the American Society for Information Science*, Volume 49, Number 7, Pages 582-603, 1998.

PROBLEM: Health-related content has long been known to be among the most sought-after information on the web (probably right behind sex-related topics). A great deal of free health-related content has been made available on the web, but the quality varies. The National Library of Medicine (NLM) has developed many excellent, high-quality medical databases such as MEDLINE and CANCERLIT that are widely used by medical professionals. Unfortunately, because of the domain and terminology knowledge required, even physicians would have trouble stating queries unless they have been trained in using the NLM's MeSH subject headings or Unified Medical Language System (UMLS). How can we identify high-quality web content and provide support for health-related search needs?

SOLUTION: A medical search spider was created and equipped with a medical vocabulary knowledge base from UMLS. An inlink analysis algorithm similar to that of Google was developed. A medical web page/site is considered high-quality if it contains UMLS phrases and is pointed to by other known quality medical web sites, e.g., John Hopkins University Medical School web site, Mayo Clinic web site, etc. The spider collected about 1 million English language medical web pages, which were then indexed by a medical noun phraser. An automatic medical thesaurus of about 40 million terms was generated based on co-occurrence analysis of medical noun phrases. In addition, a multi-layered, graphical medical knowledge map was created by the SOM algorithm. Medical


helpfulMed.com Microsoft Internet Explorer
File Edit View Go Favorites Help
Home About Help Sitemap Privacy Conta
helpful Med
SEARCH MEDICAL WEBSITES
RELATED MEDICAL TERMS
SEARCH M
helpfulMed.com - Microsoft Internet Explorer
File Edit View Go Favorites Help
Links
Home About Help Sitemap Privacy Contact Us Copyright Disclaimer
helpful Med
SEARCH MEDICAL WEBSITES
RELATED MEDICAL TERMS
SEARCH MEDICAL DATABASES
VISUAL SITE BROWSER
search
Find Web Pages Containing Terms

Address http://ai.bpa.arizona.edu/helpfulmed/term.html

Enter Medical Terms gene therapy

What's New?

See our live demo of HelpfulMED at the AMIA Conference in Los Angeles November 4-8

Related Medical Terms

- Use Cancer Space (thes
- Enter a search term in th
- The display will present
- To select a term, click in
- When you have selected related terms, or you ma

Address http://ai20.bpa.arizona.edu/~tona/cn199908 helpfulmed.cgi

Enter a Medical Term

Select Terms:

☑ Gene Therapy
☑ GENE DELIVERY
☑ GENE THERAPY STRATEGIES
☑ Acquired Immunodeficiency Syndrome

Search Results:

Noun Phrase: 511	MeSH: 247	Author: 50
1 □ Gene Therapy (b,c,d)	1 □ Support, U.S. Gov't, P.H.S. (a,b,c,d)	1 □ Curiel DT (a,b,c)
2 □ Gene Transfer (a,b,c)	2 □ Tumor Cells, Cultured (a,b,c,d)	2 □ Crystal RG (a,c)
3 □ Genetic Vectors (a,b,c)	3 □ Support, Non-U.S. Gov't (a,b,c,d)	3 □ Roth JA (a,b)
4 □ RETROVIRAL VECTORS (a,b,c)	4 □ Gene Therapy (b,c,d)	4 □ Garver RI Jr (c)
5 □ Gene Expression (a,b,c)	5 □ Gene Transfer (a,b,c)	5 □ Alvarez RD (c)
6 □ Thymidine Kinase (a,b,c)	6 □ Genetic Vectors (a,b,c)	6 □ Barnes MN (c)
7 □ TUMOR CELLS (a,b,c)	7 □ Gene Expression (a,b,c)	7 □ Behr JP (b)
8 □ Cell Line (a,b,d)	8 □ Thymidine Kinase (a,b,c)	8 □ Bimstiel ML (b)
9 □ Cystic Fibrosis (a,b,c)	9 □ Transduction, Genetic (a,b,c)	9 □ Breakefield XO (b)
10 □ Hematopoietic Stem Cells (a,c,d)	10 □ Cell Line (a,b,d)	10 □ Burkholder JK (b)
11 □ Polymerase Chain Reaction (a,c,d)	11 □ Cystic Fibrosis (a,b,c)	11 □ Cosset FL (b)
12 □ TARGET CELLS (a,b,c)		

FIGURE 18 A 40 million term medical thesaurus was created using medical noun phrasing and co-occurrence analysis techniques. The system suggests medical terms relevant to users' queries, e.g., "Genetic Vectors" and "Gene Transfer" are suggested for "Gene Therapy." The system also maps user terms to relevant MeSH terms.

users were able to use the automatic thesaurus to identify appropriate query terms during searches and the SOM-based medical knowledge map could be used to assist in the browsing of medical knowledge and content.

A large medical ontology such as the UMLS was created by merging many different human-generated thesauri and term definitions. Many person-years efforts have been invested to create such a community asset. However, once optimized, the complementary system-aided approach to creating automatic thesaurus and medical knowledge maps would take only a few days—term definitions and relationships are extracted from the documents directly. Knowledge update would become efficient and seamless.

LESSONS LEARNED: Unlike entertainment-related web content, medical abstracts stored in the NIH-sponsored MEDLINE are very precise, but the terminologies employed are fluid and cryptic (e.g., acronyms for genes, diseases, etc.). Medical professionals and the general public would benefit greatly from online search aides such as the automatic thesaurus and SOM-based medical knowledge map. With the high-precision NIH medical content, both the system-generated medical thesaurus and knowledge map showed excellent results (approaching 80% in recall/precision) and are able to complement the human-created MeSH and UMLS ontologies. We believe such techniques would be useful for text mining of medical diagnosis records as well. However, web pages collected from the Internet sources remain suspect because of their lack of uniform quality. ●

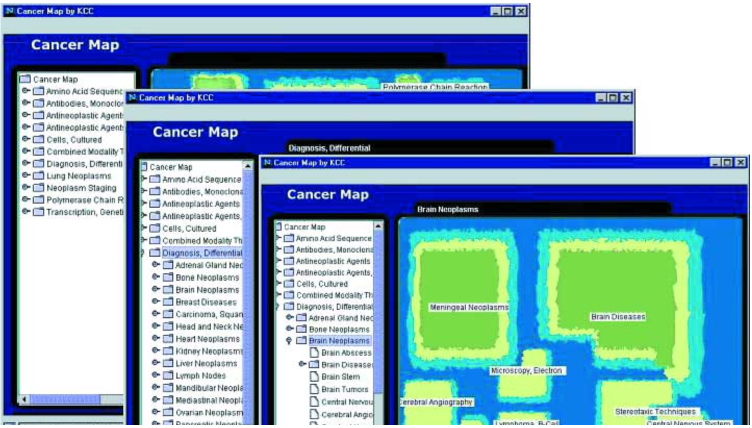
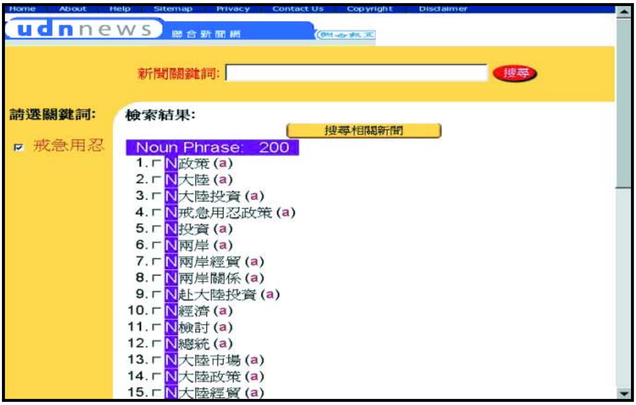


FIGURE 19 A multi-layered cancer map reveals critical cancer research topics, e.g., "Diagnosis, Differential." Clicking on a region brings about another finer-grained topic map that contains topics such as: "Brain Neoplasms."

PROBLEM: News media have been among the most widely used content resources on the web. Many corporations also rely on news sources for their industry and competitive intelligence analysis activities. In addition, we are interested in knowing whether the new text mining techniques could be applied in languages other than English.

SOLUTION: United Daily News (UDN) in Taiwan probably has the largest Chinese news content in the world. Dating back 50 years, UDN recently has made its collection (4-5 years, 600,000 articles) available online and is in the process of completely digitizing its entire collection of about 20 million articles. A statistic-based mutual information program was used to index Chinese content. Unlike the English-based noun phraser that is based on English grammar and unsuited for other languages, the statistics-based mutual information technique looks for long, frequently-occurring, multi-word patterns. After mutual information, a similar co-occurrence and SOM approach was adopted to create an automatic thesaurus for news events and a knowledge map for news browsing.



新聞關鍵詞:

請選關鍵詞:

☑ 戒急用忍

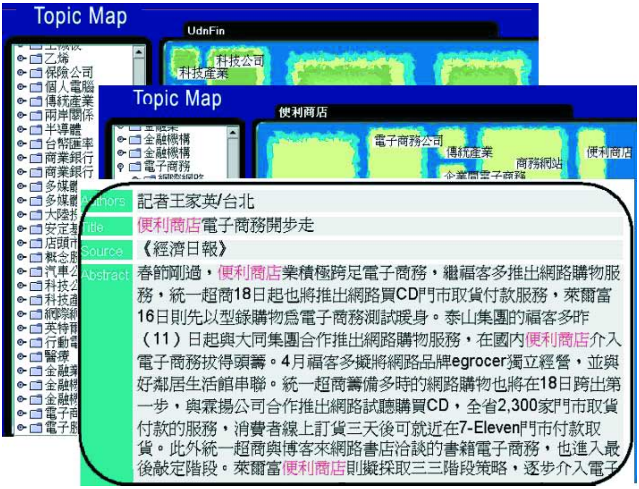
檢索結果:

Noun Phrase: 200

1. ☐ N政策 (a)
2. ☐ N大陸 (a)
3. ☐ N大陸投資 (a)
4. ☐ N戒急用忍政策 (a)
5. ☐ N投資 (a)
6. ☐ N兩岸 (a)
7. ☐ N兩岸經貿 (a)
8. ☐ N兩岸關係 (a)
9. ☐ N赴大陸投資 (a)
10. ☐ N經濟 (a)
11. ☐ N檢討 (a)
12. ☐ N總統 (a)
13. ☐ N大陸市場 (a)
14. ☐ N大陸政策 (a)
15. ☐ N大陸經貿 (a)

FIGURE 20 Related Chinese news topics are suggested based on statistical parsing and co-occurrence analysis techniques.

LESSONS LEARNED: News articles are extremely structured (people, places, events, etc.) and precise. The statistical indexing technique worked superbly for Chinese content and would work well for many other languages. The resulting news event automatic thesaurus and knowledge map achieved a performance level at 90%. We believe text mining would work well for industry-specific marketing or competitive intelligence analysis. By analyzing multiple external and internal intelligence sources (e.g., Gartner Group, IDC, WSJ, etc.) automatically, market analysts could more effectively and efficiently survey the industry landscape and identify threats and opportunities. Significant improvements in productivity and quality of marketing research could be expected by adopting the new text mining techniques on a traditional e-portal platform. ●

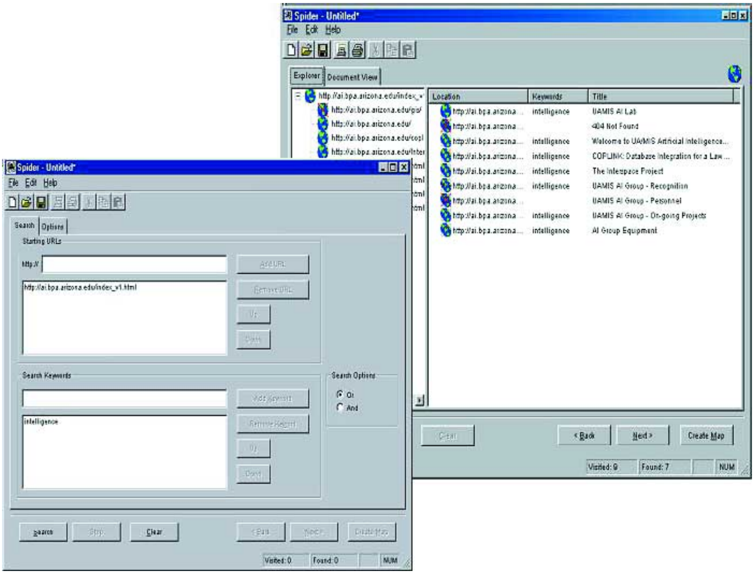


Authors	記者王家英/台北
Title	便利商店電子商務開步走
Source	《經濟日報》
Abstract	春節剛過，便利商店業積極跨足電子商務，繼福客多推出網路購物服務，統一超商18日起也將推出網路買CD門市取貨付款服務，萊爾富16日則先以型錄購物爲電子商務測試暖身。泰山集團的福客多昨（11）日起與大同集團合作推出網路購物服務，在國內便利商店介入電子商務拔得頭籌。4月福客多擬將網路品牌egrocer獨立經營，並與好鄰居生活館串聯。統一超商籌備多時的網路購物也將在18日跨出第一步，與霖揚公司合作推出網路試聽購買CD，全省2,300家門市取貨付款的服務，消費者線上訂貨三天後可就近在7-Eleven門市付款取貨。此外統一超商與博客來網路書店洽談的書籍電子商務，也進入最後敲定階段。萊爾富便利商店則擬採取三三階段策略，逐步介入電子

FIGURE 21 A Chinese business intelligence topic map reveals critical e-commerce topics discussed in Chinese news media. Clicking on a region reveals more fine-grained topics and the related news articles.

PROBLEM: Search engine content and indexes are often out of date. On average, most search engines collect and refresh their collection once a month, thus retaining many obsolete pages and dead links. For marketing analysts or business managers who are monitoring an industry or many competitors, a "fresher" and more dynamic approach to monitoring web content is needed. In addition, the amount of content collected from Internet sources could be overwhelming. How can a system help users collect fresh and comprehensive Internet content and support dynamic, personalized analysis of content?

SOLUTION: Instead of using a server-based approach that requires monthly pre-spidering and indexing, a client-based approach to meta-spidering and web analysis could be adopted. A personal search spider could connect to multiple search engines (e.g., Lycos, Google), databases (e.g., MEDLINE, CancerLit, NBC news, CBS news), or competitors' public web sites (e.g., www.ibm.com) and collect comprehensive and timely information. Such information is then loaded into a user's own computer to allow him/her to perform a more detailed analysis. The



http://ai.bpa.arizona.edu/index_v
http://ai.bpa.arizona.edu/gis/
http://ai.bpa.arizona.edu/
http://ai.bpa.arizona.edu/copl
http://ai.bpa.arizona.edu/Inter

Location	Keywords	Title
http://ai.bpa.arizona...	intelligence	UAMIS AI Lab
http://ai.bpa.arizona...		404 Not Found
http://ai.bpa.arizona...	intelligence	Welcome to UA/MIS Artificial Intelligence...
http://ai.bpa.arizona...	intelligence	COPLINK: Database Integration for a Law ...
http://ai.bpa.arizona...	intelligence	The Interspace Project
http://ai.bpa.arizona...	intelligence	UAMIS AI Group - Recognition
http://ai.bpa.arizona...		UAMIS AI Group - Personnel
http://ai.bpa.arizona...	intelligence	UAMIS AI Group - On-going Projects
http://ai.bpa.arizona...	intelligence	AI Group Equipment

Visited: 9 Found: 7

Starting URLs
http://
http://ai.bpa.arizona.edu/index_v1.html

Search Keywords
intelligence

Search Options
● Or
○ And

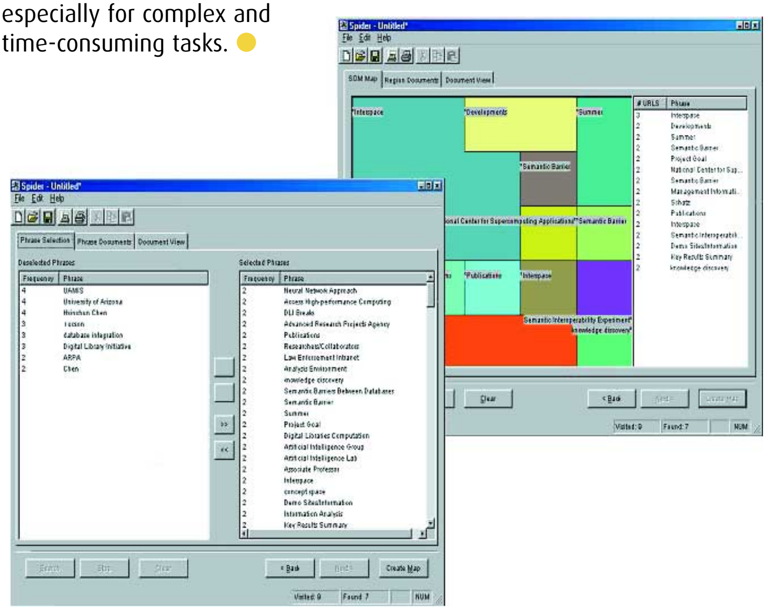
Visited: 0 Found: 0

FIGURE 22 Personal Spiders collect web pages and perform meta-searching on different business intelligence databases. Internet information is downloaded to a user's computer.

web content is summarized dynamically based on noun phrasing and an SOM knowledge map. In addition, the search sessions can be saved and users can set their own search parameters.

LESSONS LEARNED: The concept of one-stop meta searching was shown to be very powerful. Users were able to quickly get all relevant information from one interface. Dynamic noun phrase summarization was very useful for helping users get a quick feel of the thematic topics in the data collected. SOM output was interesting and appealing, especially for graphics-oriented users. Users found client-based searching and personalization (through filters and dictionaries) to be particularly suited for time-critical marketing and business intelligence analysis. For corporate marketing e-portals, both server and client-based searching could be invaluable.

The dynamic term summarization approach is useful for alleviating the "information overload" problem faced by traditional search engines. A several-fold improvement in search and analysis efficiency could be obtained especially for complex and time-consuming tasks.



# URLS	Phrase
3	Interspace
2	Developments
2	Summer
2	Semantic Barrier
2	Project Goal
2	National Center for Sup...
2	Semantic Barrier
2	Management Informati...
2	Schatz
2	Publications
2	Interspace
2	Semantic Interoperabil...
2	Demo Sites/Information
2	Key Results Summary
2	knowledge discovery

Deselected Phrases

Frequency	Phrase
4	UAMIS
4	University of Arizona
4	Hsinchun Chen
3	Tucson
3	database integration
3	Digital Library Initiative
2	ARPA
2	Chen

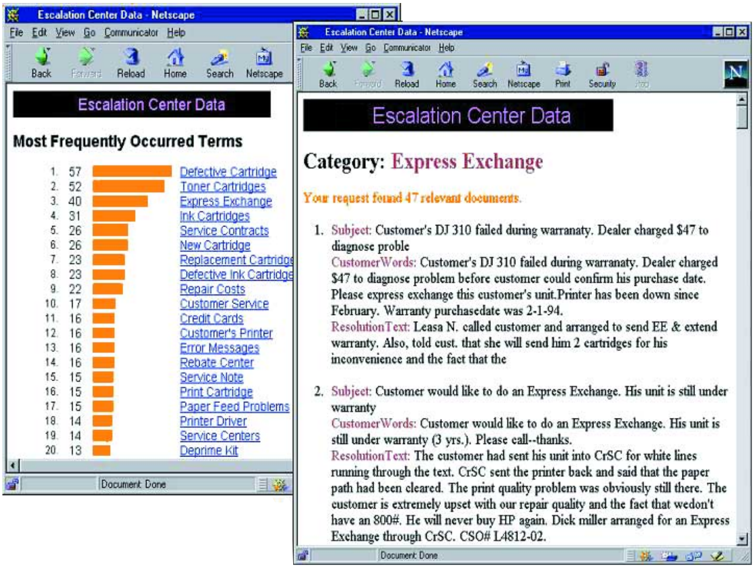
Selected Phrases

Frequency	Phrase
2	Neural Network Approach
2	Access High-performance Computing
2	DLI Breaks
2	Advanced Research Projects Agency
2	Publications
2	Researchers/Collaborators
2	Law Enforcement Intranet
2	Analysis Environment
2	knowledge discovery
2	Semantic Barriers Between Databases
2	Semantic Barrier
2	Summer
2	Project Goal
2	Digital Libraries Computation
2	Artificial Intelligence Group
2	Artificial Intelligence Lab
2	Associate Professor
2	Interspace
2	concept space
2	Demo Sites/Information
2	Information Analysis
2	Key Results Summary

FIGURE 23 Extracted business intelligence from the web can then be automatically analyzed to identify emerging themes and topics (using noun phrases and topic maps).

PROBLEM: Customer Relation Management (CRM) data collected through a call center are valuable assets of corporations, but are rarely utilized. With the steadily maturing virtual call center technologies and growing acceptance, corporations are beginning to record and analyze customer engagement data (questions and answers) collected through email, online chats, and VoIP (voice over IP) recording. Can text mining techniques help retain agent knowledge through analysis of past engagements? Can such techniques help reveal the problems that customers are facing and suggest proper solutions?

SOLUTION: Most CRM software supports only analysis of numeric data stored in customer engagement databases, e.g., volumes, agent productivity, call distribution, etc. They lack support for textual analysis of frequently occurring problems. Using noun phrasing and SOM categorization, we were able to categorize and visualize significant problem areas faced by call centers.


Escalation Center Data - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Document: Done

Escalation Center Data

Most Frequently Occurred Terms

1.	57	Defective Cartridge
2.	52	Toner Cartridges
3.	40	Express Exchange
4.	31	Ink Cartridges
5.	26	Service Contracts
6.	26	New Cartridge
7.	23	Replacement Cartridg
8.	23	Defective Ink Cartridg
9.	22	Repair Costs
10.	17	Customer Service
11.	16	Credit Cards
12.	16	Customer's Printer
13.	16	Error Messages
14.	16	Rebate Center
15.	15	Service Note
16.	15	Print Cartridge
17.	15	Paper Feed Problems
18.	14	Printer Driver
19.	14	Service Centers
20.	13	Deprime Kit

Escalation Center Data

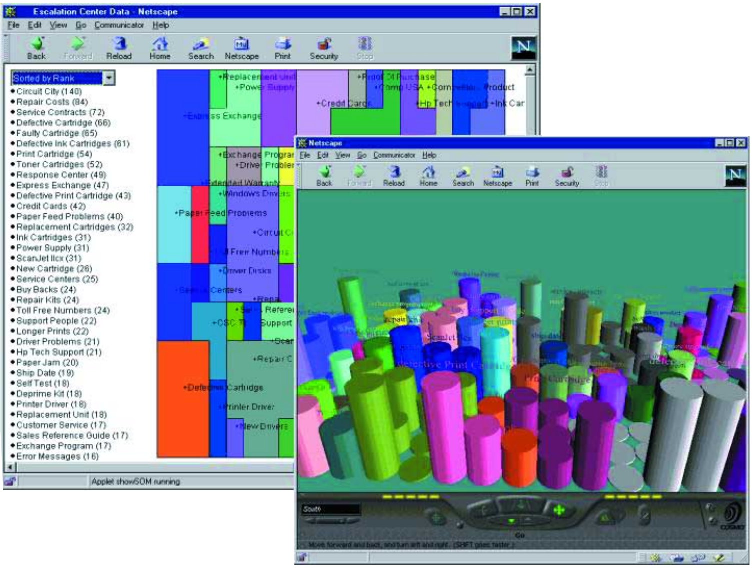
Category: Express Exchange

Your request found 47 relevant documents.

1. Subject: Customer's DJ 310 failed during warranaty. Dealer charged $47 to diagnose proble
   CustomerWords: Customer's DJ 310 failed during warranaty. Dealer charged $47 to diagnose problem before customer could confirm his purchase date. Please express exchange this customer's unit.Printer has been down since February. Warranty purchasedate was 2-1-94.
   ResolutionText: Leasa N. called customer and arranged to send EE & extend warranty. Also, told cust. that she will send him 2 cartridges for his inconvenience and the fact that the
2. Subject: Customer would like to do an Express Exchange. His unit is still under warranty
   CustomerWords: Customer would like to do an Express Exchange. His unit is still under warranty (3 yrs.). Please call--thanks.
   ResolutionText: The customer had sent his unit into CrSC for white lines running through the text. CrSC sent the printer back and said that the paper path had been cleared. The print quality problem was obviously still there. The customer is extremely upset with our repair quality and the fact that wedon't have an 800#. He will never buy HP again. Dick miller arranged for an Express Exchange through CrSC. CSO# L4812-02.

FIGURE 24 CRM engagement data can be analyzed and tabulated for easy browsing.

LESSONS LEARNED: It was observed that call center engagement data are quite noisy, with many typos, acronyms, and grammatical errors. However, noun phrasing analysis and frequency count enabled us to summarize problem areas that were buried deep inside the agents' textual descriptions. By storing the resulting question-answer pairs and providing online suggestions (of similar engagements) during an agent's customer engagement process, an agent could become more productive. In addition, inexperienced agents could be brought up to speed much faster, thus addressing the critical agent turn-over problem that faces the call center industry. A several-fold improvement in agent productivity is expected with the proposed text mining approach. We believe the KM-enhanced virtual call center will be the future for this billion-dollar industry. ●


Escalation Center Data - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Sorted by Rank
Applet showSOM running
Netscape
South

- Circuit City (140)
- Repair Costs (84)
- Service Contracts (72)
- Defective Cartridge (66)
- Faulty Cartridge (65)
- Defective Ink Cartridges (61)
- Print Cartridge (54)
- Toner Cartridges (52)
- Response Center (49)
- Express Exchange (47)
- Defective Print Cartridge (43)
- Credit Cards (42)
- Paper Feed Problems (40)
- Replacement Cartridges (32)
- Ink Cartridges (31)
- Power Supply (31)
- ScanJet IIcx (31)
- New Cartridge (26)
- Service Centers (25)
- Buy Backs (24)
- Repair Kits (24)
- Toll Free Numbers (24)
- Support People (22)
- Longer Prints (22)
- Driver Problems (21)
- Hp Tech Support (21)
- Paper Jam (20)
- Ship Date (19)
- Self Test (18)
- Deprime Kit (18)
- Printer Driver (18)
- Replacement Unit (18)
- Customer Service (17)
- Sales Reference Guide (17)
- Exchange Program (17)
- Error Messages (16)

FIGURE 25 The same CRM data can be revealed in 2D and 3D formats.

Newsgroup Visualization

PROBLEM: As described in the Gartner Group report, Knowledge Retrieval (KR) consists of two dimensions: semantic and collaboration. Collaboration entails identifying trusted advisors and experts in an organization. In newsgroup, groupware, and corporate email applications, many discussion items are created, but searching for relevant topics and trusted experts in such a community-generated repository is often difficult. In a newsgroup, many people may respond to multiple topics in multiple threads. How can text mining and information visualization techniques help reveal thematic topics in newsgroup collection and help identify experts in the community?

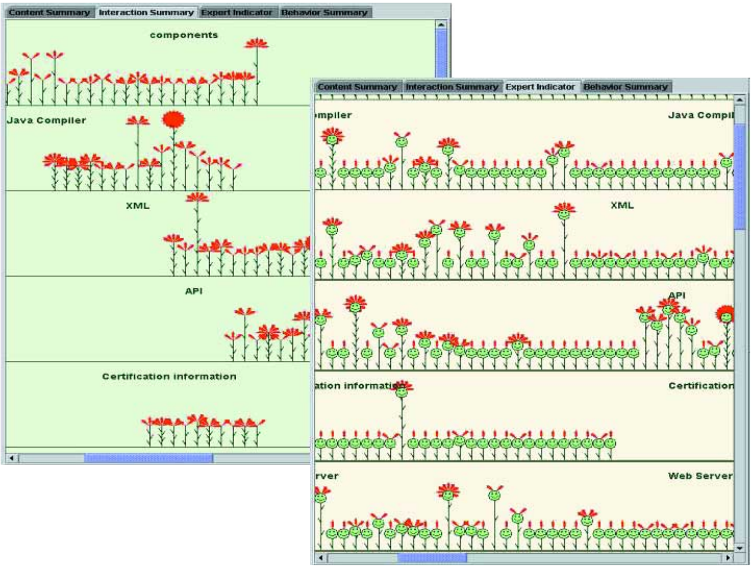
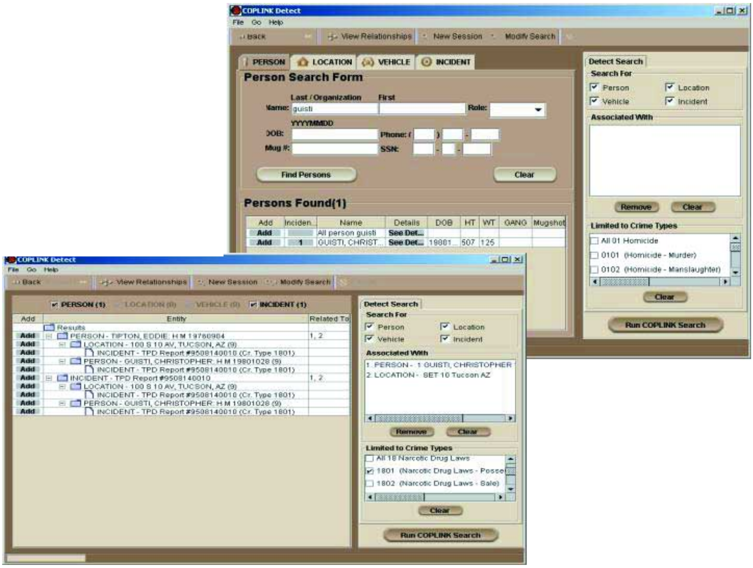


FIGURE 26 Communication Garden reveals important threads (represented as flowers) and experts (represented as people) hidden in newsgroup databases. Instead of reading text, users strode through the garden to visually identify important threads and experts. Clicking on a flower reveals the underlying textual email messages.

SOLUTION: Newsgroup email items could be analyzed based on noun phrasing and SOM categorization techniques. Using the “glyph” representation for multiple features, we created a “communication garden” visualization metaphor that represents people (who contributed their knowledge to the newsgroup) in terms of long, blooming flowers. A healthy garden of many blooming flowers would thus represent a lively discussion with many good experts. Conversely, a newsgroup reader would not wish to consult news items in an unhealthy garden (of email messages).

LESSONS LEARNED: With effective visualization, a picture indeed is worth a thousand words (P1000, as code-named by the CIA visualization research program). Expert identification is critical for knowledge management support. The glyph-based communication garden display was shown to be a very effective tool for expert identification. Instead of relying on cumbersome cognitive processing (by reading large amounts of text), knowledge workers could utilize their visual recognition ability (by glancing at pictures) to engage in more effective decision making activities. Instead of spending multiple hours browsing textual documents, a visual browsing process could take only minutes.

PROBLEM: Many government agencies, such as local law enforcement units, produce electronic reports that are difficult to search and analyze. Often, only experienced and knowledgeable workers are able to use such organizational resources effectively. Just as e-commerce activities (e.g., CRM data, web search engines, intelligence analysis) require knowledge management support, the voluminous textual reports generated from many e-government programs also call for advanced text mining methods. Can local law enforcement agencies use text mining to support criminal analysis?



Person Search Form

Name: Last / Organization: guisti First: Role:
DOB: YYYYMMDD
Phone: ( ) -
Mug #:
SSN: - -

Persons Found(1)

Add	Inciden...	Name	Details	DOB	HT	WT	GANG	Mugshot
Add		All person guisti	See Det...					
Add	1	GUISTI, CHRIST...	See Det...	19801...	507	125		

Detect Search

Search For
☑ Person ☑ Location
☑ Vehicle ☑ Incident

Associated With

Limited to Crime Types
□ All 01 Homicide
□ 0101 (Homicide - Murder)
□ 0102 (Homicide - Manslaughter)

COPLINK Detect

☑ PERSON (1) LOCATION (0) VEHICLE (0) ☑ INCIDENT (1)

Add	Entity	Related To
	Results	
Add	PERSON - TIPTON, EDDIE: H M 19760904	1, 2
Add	LOCATION - 100 S 10 AV, TUCSON, AZ (9)	
Add	INCIDENT - TPD Report #9508140010 (Cr. Type 1801)	
Add	PERSON - GUISTI, CHRISTOPHER: H M 19801028 (9)	
Add	INCIDENT - TPD Report #9508140010 (Cr. Type 1801)	
Add	INCIDENT - TPD Report #9508140010	1, 2
Add	LOCATION - 100 S 10 AV, TUCSON, AZ (9)	
Add	INCIDENT - TPD Report #9508140010 (Cr. Type 1801)	
Add	PERSON - GUISTI, CHRISTOPHER: H M 19801028 (9)	
Add	INCIDENT - TPD Report #9508140010 (Cr. Type 1801)	

Detect Search

Search For
☑ Person ☑ Location
☑ Vehicle ☑ Incident

Associated With
1 PERSON - 1 GUISTI, CHRISTOPHER
2 LOCATION - SET 10 Tucson AZ

Limited to Crime Types
□ All 18 Narcotic Drug Laws
☑ 1801 (Narcotic Drug Laws - Posse
□ 1802 (Narcotic Drug Laws - Sale)

FIGURE 27 COPLINK Detect helps detectives identify crime associations, i.e., people, location, vehicle, and incident, e.g., "Eddie Tipton" was found to be strongly associated with "Christopher Guisti" and "100 S 10 AV TUCSON" in narcotic cases.*

**All names and addresses are changed to protect confidentiality.*

SOLUTION: A federally funded project called COPLINK was designed with such a purpose in mind. COPLINK is a distributed data warehouse that supports inter-jurisdictional information sharing and knowledge management. The system consolidates (textual) police reports collected from different legacy systems and performs co-occurrence analysis of criminal elements, e.g., people's names, organizations, license plates, crime types, etc. The resulting concept space reveals weighted relationships between various law enforcement associations. Investigators were able to use the system to perform more efficient analyses and improve case solvability.

Time is of the essence for criminal analysis. The 24 hours (or less) after a crime occurs are the most crucial during investigation because crime scene evidence may be fast deteriorating and criminals who committed the crime are often fast escaping the scene or even the jurisdiction. Our many case studies involving real investigative cases revealed that the COPLINK concept space allows investigators to reduce their investigative time-often from multiple days to less than one hour. As a result, both investigator productivity and case solvability improve significantly. For time-critical applications (e.g., business and health intelligence) or time-sensitive knowledge workers (e.g., investigators and market analysts), a text mining approach to knowledge management is a necessity, not a luxury.

LESSONS LEARNED: Crime investigation without proper system support is a complex and time-consuming job. Co-occurrence analysis and case narrative analysis are extremely useful for investigative work. In addition, collaboration mechanisms (linking different investigators), push technology (monitoring new data content), and wireless application (through laptop, PDA, and cell phone) would be essential for more effective and efficient law enforcement work. Due to the intertwining nature of government, we believe the COPLINK solution also would be appropriate for other local e-government activities involving corrections, courts, social services, etc.

PROBLEM: With the increasing amount of multimedia content on the Internet and Intranets, is multimedia data mining mature enough for business applications? A progression from data mining, to text mining and then to multimedia data mining seems obvious. But are the multimedia analysis techniques such as audio analysis or image processing robust enough for commercial multimedia content such as corporate training tapes, design drawings, news clips, etc.?

SOLUTION: Image segmentation and processing are techniques that are needed for processing still pictures and videos (of sounds and images). Most content-based image segmentation techniques are based on texture pattern, color, and shape. Similarity-based image content retrieval is feasible (e.g., find all the images that are similar to these sample colored textures), but semantics-based retrieval (e.g., find

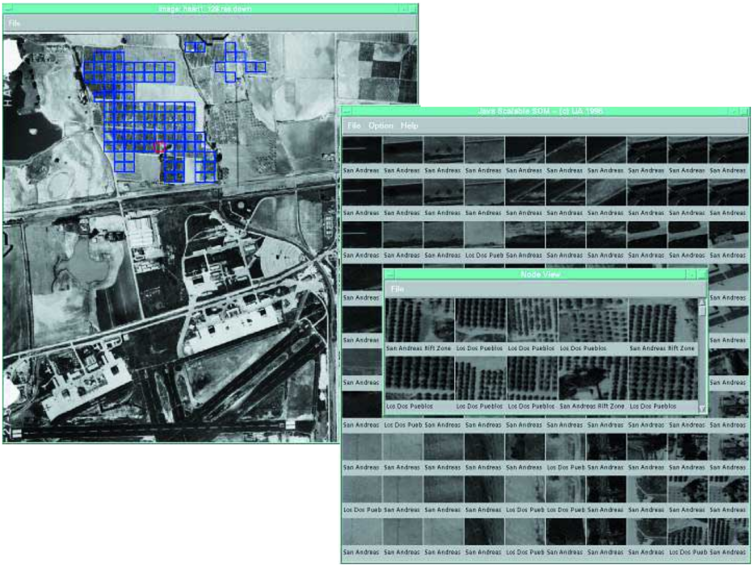

File
File Option Help
San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas
San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas
San Andreas San Andreas San Andreas Los Dos Pueb San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas
San Andreas
San Andreas
San Andreas
San Andreas
San Andreas
San Andreas
Node View
File
San Andreas Rift Zone Los Dos Pueblos Los Dos Pueblos Los Dos Pueblos San Andreas Rift Zone
Los Dos Pueblos Los Dos Pueblos Los Dos Pueblos San Andreas Rift Zone Los Dos Pueblos
San Andreas Los Dos Pueb San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas San Andreas
San Andreas San Andreas San Andreas San Andreas San Andreas Los Dos Pueb San Andreas San Andreas San Andreas San Andreas
Los Dos Pueb San Andreas San Andreas San Andreas Los Dos Pueb Los Dos Pueb San Andreas San Andreas San Andreas San Andreas
San Andreas San Andreas San Andreas San Andreas Los Dos Pueb San Andreas San Andreas San Andreas Los Dos Pueb San Andreas

FIGURE 28 A red square indicates a location with vegetation image pattern of interest to a searcher. The air photo image matching system helps identify areas on the map that have similar vegetation pattern.

all pictures that have yellow birds) is still far from being a reality. Some application areas of image analysis, such as air photo content-based retrieval, seem promising. Users would be able to find similar orchards or housing development patterns based on texture-based image segmentation techniques.

LESSONS LEARNED: Multimedia data mining clearly is an emerging area of great potential. However, most images or videos need to be tagged manually with textual labels (i.e., metadata). Users can then search for images using text. Although content-based retrieval is becoming mature, it is still limited to selected domains. Image analysis techniques are still very application dependent (unlike data or text analysis). In addition, multimedia applications may also require integration of data, text, and image mining techniques, a very challenging research problem.

The Future

4

multilingual web
wireless web
multilingual knowledge portal
semantic web

The Future

What are the future research areas for Knowledge Management Systems and text mining? What are the applications and technologies that may affect future knowledge workers?

Semantic Web: The current web infrastructure is one of hypertext syntax and structure. The html hyperlinks do not suggest any semantic or meaningful relationships between web pages. We only know that two hyperlinked pages have some sort of relationship. How to represent semantics on the web and to create a semantic web of meaningful interconnected web objects and content is a challenging research topic. Some researchers suggest richer web content notations and inference rules such as XML and RDF, others suggest a system-aided machine learning approach to extracting semantic associations between objects.

Multilingual Web: The web has increasingly become more international and multi-cultural. Non-English web content has experienced the strongest growth over the past few years. In addition, the globalization and e-commerce trend has created much multilingual Intranet content for multi-national corporations or companies with international partners. How can we create a multilingual knowledge portal such that users can experience seamless cross-lingual information retrieval (e.g., searching for Chinese government regulations using English queries) and real-time machine translation? The most immediate application of a multilingual web would be in international marketing and intelligence analysis for multi-national corporations.

Multimedia Web: We believe multimedia data mining is a trend that cannot be reversed. Although it may not become a dominant part of corporate information assets (unlike structured data and unstructured text), it does fill an important gap in corporate knowledge management.

Wireless Web: Although we believe the majority of the web content will still be accessed over high-speed wired networks, wireless applications will continue to grow in years to come. They will also emerge and proliferate rapidly in selected high-impact application areas, e.g., email, financial stock quotes, curbside law enforcement alerting (via PDA and cell phone), etc. For knowledge workers who are mobile and time-pressed, wireless knowledge management will not be a luxury but a necessity in a not-so-distant future.

IDC Group predicts the KM service market to be worth $8 billion dollars by 2003 and the KM software market to be worth $5.4 billion dollars by 2004. The unstoppable trend towards the semantic web, multimedia web, multilingual web, and wireless web will only accelerate its growth. Knowledge is king, not information or data!

Despite this real market potential, KM presents a significant challenge to practitioners and researchers alike in choosing the "right" technology in the "right" organizational context with the "right" methodology. The road ahead is uncertain, but we firmly believe that the next dominant player in this space would eventually emerge as the next Yahoo or even Oracle in this new millennium. ●

Professor Hsinchun Chen
Director, Artificial Intelligence Lab
Director, Mark and Susan Hoffman E-Commerce Lab
Founder, Knowledge Computing Corporation

Department of Management Information Systems
Eller College of Business
and Public Administration
The University of Arizona
McClelland Hall 430
P.O. Box 210108
Tucson, Arizona 85721-0108

Tel: (520) 621-2748
Fax: (520) 621-2433
Email: hchen@eller.arizona.edu

Web site: http://ai.eller.arizona.edu

Knowledge Computing Corporation
3915 East Broadway, Suite 301
Tucson, Arizona 85711

Tel: (520) 574-1519
Toll free: (877) 522-9599
Fax: (520) 574-0870
Email: hchen@knowledgecc.com

Web site: www.knowledgecc.com

THE
ELLER
COLLEGE®

THE UNIVERSITY OF ARIZONA®