SPECIAL SECTION

Search 2009 | Marjorie M.K. Hlava, Guest Editor

7] Introduction to Search 2009
by Marjorie M.K. Hlava and Jay Ven Eman

11] An Evolution of Search
by John D. Holt and David J. Miller

16] Next Generation Search Platforms: How Vendors Are Searching Unstructured Content
by Rich Turner

20] Reconsidering Relevance and Embracing Interaction
by Daniel Tunkelang

24] Searching Real-Time Financial News: PAR for the Course
by Lawrence C. Rafsky

31] Integration of Taxonomy and Keyword Searches: A Comparison of Two Implementations
by Ronald P. Millett

36] Semantic Search
by Darrell W. Gunter

FEATURE

38] Information Science Beyond the Border
by Luanne Freund and Heather O'Brien

DEPARTMENTS

[2] Editor's Desktop

[3] President's Page

[4] Inside ASIS&T

COLUMNS

[40] IA
Hasta Luego
by Stacy Surla

Introduction from Thom
by Thom Haller

OPINION

The NISO Standard for Controlled Vocabulary:
A Blueprint for Revision
by Bella Hass Weinberg
nterest in retrieval of non-structured information, whether full
text or surrogates, has been a common bond for a large part of
our Society’s membership. In this issue we focus on a core
component of this problem – text search. Marge Hlava, president
of Access Innovations, Inc., and chairman of the Bulletin Advisory
Board, has put together six articles, each from a scientist and/or
executive of a commercial firm that develops and markets search
software. A few articles from other vendors were solicited but not
received in time; some may appear in later issues. The emphasis
is on case studies and on classes and capabilities of newer
generation search software. We are very pleased to be able to
publish articles covering such an interesting range of innovations
from entrepreneurs in the field.

On a related topic, in our Opinion section, Bella Hass
Weinberg proposes revisions to the 2005 Guidelines for the
Construction, Format, and Management of Monolingual
Controlled Vocabularies, which should soon be up for review.

After four years as associate editor for information architecture,
Stacy Surla has decided to take a well-deserved break. Over these
years, Stacy has either solicited or written more than a dozen IA
columns and produced four widely read and appreciated special
sections on the topic. We have benefited greatly from the wide
variety of topics she pursued, the distinguished authors she
persuaded to write for us and her practical and pragmatic focus.
One of the authors she has corralled in the past, Thom Haller,
has agreed to replace Stacy for at least the next year. Thom is a
teacher, lecturer and writer in the field of information architecture
and has been associated with ASIS&T’s IA Summits since their
inception. We welcome him to the Bulletin.

As we turn to news of the Society, ASIS&T president Don
Case also provides a review of recent actions by the ASIS&T
Board of Directors, which include quite a few new initiatives of
interest such as raising the amount of money per capita given to
active local chapters.

Finally, this year’s Annual Meeting in Vancouver, Canada, is
being held jointly with the Canadian Association for Information
Science/L’Association Canadienne des Sciences de
L’Information (CAIS-ACSI). As part of our continuing series on
sister information societies and in honor of the event Luanne
Freund and Heather O’Brien have prepared a history of CAIS-
ACSI, including its long association with ASIS&T. As I sit in
hot and drought-stricken central Texas, Vancouver seems a
heavenly vision, and it has been refreshing at least be able to
read about Canada. Given an exciting program as well as a
beautiful location, I look forward to this meeting with even more
anticipation than usual.
The ASIS&T Board recently held its annual summer retreat, this time on the campus of the University of North Carolina, in lovely Chapel Hill. The location had much to recommend it, as four of the members, including the incoming President Gary Marchionini, are Tar Heels; it made the logistics for the meeting – which is more typically held near Washington, D.C. – much simpler.

Now North Carolina may lead the nation in creepy place names – think Great Dismal Swamp, Kill Devil Hills and Cape Fear – but it is unfailling in its welcome to visitors, even if they are Yankees or other foreigners. We were shown real southern hospitality and all the grits we could eat. We were even blessed with mild weather, rather than the usual humidity that hangs like kudzu over the southeast during mid summer.

But, back to the meeting itself. I must say that it is really nice having a Board with so many good ideas. In this, and an earlier, meeting we discussed a number of ways in which we might do more for the membership, including (but not limited to) the following initiatives:

- Increasing the rate for funding for regional chapters on a member-by-member basis. Under this scheme, active chapters will receive at least five times as much per regular member, an average increase of several hundred dollars per chapter. [Approved]
- Doubling the amount of money given to student chapters at the start of their year. [Approved]
- Responding to the recession by defining a “hardship” category so that more low-income members can renew at a special $40 rate. [Approved]
- Improving and publicizing the placement service at the Annual Meeting. [In process]
- Holding a new summit related to scientific data collection and management during 2010. [In planning stages]
- Creating additional scholarships for students to attend the Annual Meeting; seven scholarships would be competitive on a regional basis and provide an additional incentive for students to become involved in local activities. [Under discussion]
- Making more effective use of social media, such as blogs, wikis and Second Life, to keep in touch with members. [Under discussion]
- Hiring a part-time “interaction manager” devoted to accomplishing the previous goal. [Under discussion]
- Finding ways to make Annual Meeting presentations more interactive and involving through changes in scheduling and format and through use of technologies. [Under discussion]
- Devising a mechanism to solicit real-world information problems and have design teams compete for a solution. [Under discussion]

As indicated above, several of these changes are already in effect, while others await a financial analysis or further tweaking. The remainder of the initiatives will take a few months to flesh out, but we expect them to be in place sometime in 2010.

I left the University of North Carolina encouraged by our discussions, and certain that North Carolina deserves a better official motto than “We’re the state north of South Carolina.” With four Board members, not to mention three LIS programs, maybe its motto ought to be “The State of Information Science.”
Muhammad Rafiq of Pakistan is the winner of the 2009 SIG/III International Paper Contest for his paper titled “[The] LIS Community’s Perceptions towards Open Source Software Adoption in Libraries.” Rafiq will be awarded a two-year membership in ASIS&T, as well as funding to travel to this year’s Annual Meeting in Vancouver.

Second place honors go to Muhammad Arif and Saima Kanwal, also of Pakistan, for their paper titled “Acceptance of Digital Library among Female Students and Effects of Limited Access of Digital Library on their Performance in Research Work: A Case of International Islamic University.” As principal author of the paper, Arif will be awarded a two-year membership in ASIS&T.

Rafiq will attend the International Reception at the ASIS&T Annual Meeting on Monday, November 9, 2009, at 8 p.m. He will discuss his paper, which is to be published in the September 2009 issue of the International Information and Library Review, edited by Toni Carbo. Publisher Elsevier has generously donated copies of the IILR issue which will be available as long as the quantity lasts.

Other activities at the International Reception are the annual SIG/III InfoShare Silent Auction and a raffle. If you have items you would like to donate to the Silent Auction, please contact InfoShare officers Abebe Rorissa (arorissa<at>albany.edu) or Sarah Emmerson (saemmerson<at>yahoo.com) and bring your items to the conference with you.

The prize in the raffle is a gift basket put together by the local chapter of ASIS&T with locally produced products and goodies.

All proceeds from the Silent Auction and Raffle Ticket Sale go to the SIG/III InfoShare Fund, which offers ASIS&T memberships to information professionals in developing countries for whom the cost of membership would otherwise be a financial burden.

Thriving on Diversity: Information Opportunities in a Pluralistic World

November 6-11, 2009, Vancouver, BC, Canada

The 2009 ASIS&T Annual Meeting is fast approaching. Make sure you’ve marked your calendar. Check your mailbox for a printed version of the preliminary program; or go online for an electronic version. Registration forms are available in both formats. Make your travel plans now. We all look forward to a large crowd in Vancouver. See you there.

ASIS&T Networks, Blogs and Tweets

The ASIS&T membership sure knows its interactions! The organization now invites interested information folks to stay in touch with the organization and the people through its presence at Facebook, Twitter and in the blogosphere.

The American Society for Information Science and Technology group at Facebook has more than 800 members. The group shares ASIS&T news and information, discussion boards and the ever-popular wall for open messaging.

As for blogging, www.asis.org/wiki/bits is a space where all those interested in ASIS&T-related news, events and issues can come together for discussion. Contributions to the blog are encouraged from any ASIS&T member and can take many forms. If you are interested in blogging (either as a one-time or regular contributor) please email Cassidy Sugimoto at csugimoto<at>unc.edu.

Twitter is another way in which the membership of ASIS&T can be more engaged. You can follow the latest happenings in the information world at asist_org at http://twitter.com. Twitter posts are also linked to the blog. Want to tweet for ASIS&T? Email Cassidy Sugimoto at csugimoto<at>unc.edu.
News about ASIS&T Chapters

NEW JERSEY CHAPTER
The New Jersey Chapter of the American Society for Information Science & Technology (NJ/ASIS&T) has named Nicholas J. Belkin, Rutgers University School of Communication and Information, the recipient of the 2009 Distinguished Lectureship Award. The award, established in 1985, honors individuals who have made significant contributions to the field of information science.

An awards ceremony honoring Belkin was to be held on September 25 where he would deliver a lecture entitled Personalizing Support for Interaction with Information.

In addition to teaching at Rutgers, Belkin is the author of numerous articles on topics such as information retrieval, human information behavior and human-computer interaction. He has presented at nearly 200 conferences throughout the United States, Great Britain and Europe. Belkin is also past president of the American Society for Information Science and Technology, which bestowed him with its prestigious Award of Merit in 2003, its Award for Excellence in Research in Information Science in 1997 and its Outstanding Information Science Teacher Award in 1990.

Belkin also serves on the editorial board of Information Processing and Management and Information Retrieval.

LOS ANGELES CHAPTER
The Los Angeles Chapter of the American Society for Information Science and Technology (LACASIS) awards its 2008 Margaret McKinley Memorial Student Scholarship to Shilpa Rele, a student in the department of information studies at the University of California, Los Angeles. Shilpa is currently completing a master’s degree in library and information science.

The annual scholarship is the result of an essay competition open to all current library and information science graduate students. First established in 1992, and renamed in memory of the late Margaret McKinley in 1993, the competition is intended to encourage students in librarianship and information science programs to consider the benefits of participation in the activities of professional societies. The winner receives reimbursement funding up to $1,000 for registration, airfare and hotel expenses to attend ASIS&T’s Annual Meeting and a one-year membership in ASIS&T. Two runners-up receive one-year memberships in ASIS&T.

Shilpa’s professional interests relate to the description and retrieval of visual and archival information and the management of digital image collections. With a previously awarded master’s degree in art history from the University of Minnesota, Shilpa looks to combining her interests in art history and information studies while working with visual collections.

A runner up in this year’s essay competition is Matthew Mayernik, also of the University of California, Los Angeles. Mayernik is currently a Ph.D. student in the department of information studies. He is interested in how scientific information is created, and how it can be made available through online catalogs and digital libraries.

News about ASIS&T Members

Barbara M. Wildemuth, professor at the School of Information and Library Science at the University of North Carolina at Chapel Hill, is the author of the recently published Applications of Social Research Methods to Questions in Information and Library Science (Libraries Unlimited), a book that is described as a “first of its kind for librarians.” Written in a conversational tone, the book describes several methods available for the conduct of a research study.

Gary Marchionini, Cary C. Boshamer distinguished professor and incoming president of ASIS&T, has been appointed chair of the Biomedical Library and Informatics Review Committee of the National Library of Medicine (NLM). The committee reviews and advises the NLM and the National Institutes of Health on grants in the areas of medical library resources, integrated advanced information management systems, training in medical library and other information sciences, research in medical and biotechnology informatics and biomedical scientific publications. Marchionini is serving the final year of a four-year term.

Microsoft’s Dumais Earns Salton Award

Susan T. Dumais of Microsoft Research is the recipient of the 2009 Gerard Salton Award bestowed by the ACM Special Interest Group on Information Retrieval (SIGIR). Dumais was cited for her innovative contributions to information indexing and retrieval systems that have
widely impacted the quality of search from the desktop to the Web.

Dumais was an early developer of innovative interfaces and algorithms that reflected an understanding of computer users as well as the context of their search and information retrieval efforts. Her focus has been to improve the lives of users by incorporating the context of their work into search applications.

Her current research focuses on personal information management, user modeling and personalization, tightly coupling search and browsing, and implicit measures of user interest and activity. She has contributed to both the theoretical developments and practical implementations of key search issues, many of which incorporate knowledge of users and their context to improve the search process.

The Gerard Salton Award is presented every three years to an individual who has made significant, sustained and continuing contributions to research in information retrieval. It is named in honor of Gerard Salton, developer of SMART (System for the Mechanical Analysis and Retrieval of Text) and a longtime member and former director of the American Society for Information Science and Technology.

Her current research focuses on personal information management, user modeling and personalization, tightly coupling search and browsing, and implicit measures of user interest and activity. She has contributed to both the theoretical developments and practical implementations of key search issues, many of which incorporate knowledge of users and their context to improve the search process.

The Gerard Salton Award is presented every three years to an individual who has made significant, sustained and continuing contributions to research in information retrieval. It is named in honor of Gerard Salton, developer of SMART (System for the Mechanical Analysis and Retrieval of Text) and a longtime member and former director of the American Society for Information Science and Technology.

Her current research focuses on personal information management, user modeling and personalization, tightly coupling search and browsing, and implicit measures of user interest and activity. She has contributed to both the theoretical developments and practical implementations of key search issues, many of which incorporate knowledge of users and their context to improve the search process.

The Gerard Salton Award is presented every three years to an individual who has made significant, sustained and continuing contributions to research in information retrieval. It is named in honor of Gerard Salton, developer of SMART (System for the Mechanical Analysis and Retrieval of Text) and a longtime member and former director of the American Society for Information Science and Technology.
Introduction to Search 2009
by Marjorie M.K. Hlava, Guest Editor, and Jay Ven Eman

Search is like the blind men and the elephant. The story goes that four blind men approached an elephant. By touching the hide, trunk, tail and legs each thought of the elephant as an entirely different animal. From the hide we can conclude that search is a contextual process where searchers must aimlessly feel around until they either accidentally find what they want or give up. From the trunk we can conclude that search should be like a fire hose where we flood the searchers with content, overwhelming them. The tail? From this perspective search is seen as the infamous “long tail” wherein searchers are fed minuscule tidbits of esoterica, leaving them curious, but famished. And those feeling the legs would conclude that we should stomp the content into submission, pulverizing until it is unrecognizable.

The make-up of our blind crew is further complicated because it represents software developers, entrepreneurs, content producers and searchers – different perspectives and different needs. And each has different needs at different points in time. If we focus on just the searchers, their needs change over time. Search behavior is not uniform across a given population of searchers and any individual’s search behavior will change with his needs. Information requirements at the beginning of a project are likely to change through the course of a project. The serendipity created by browsing the stacks – or dynamically generated clusters produced from a search query – will likely be appreciated at the onset of a knowledge intensive project.

Once the project parameters are set and the project approved, precision and recall in retrieval are paramount. For a typical project timeline stage would be about 95% of its duration.

To compound the problem, Google has made search known. Prior to Google’s ascendance, search was not really known outside the information industry, where it was the purview of information professionals serving professional knowledge workers. Now search is everywhere. Google has also biased the landscape of expectations with little differentiation given to the World Wide Web versus large, complex, internal information environments – the world of filing cabinets, billions of C drives and an equal number of desks stuffed with paper documents, and of electronic content in a near infinite number of formats, spread across thousands of applications.

The elephant parable illustrates the complexity that is search and helps explain why search is so frustrating and doesn’t work most of the time. It helps explain why the typical, large organization has, on average, five separate search applications. It also highlights why there are so many different approaches to solving search.

What are the basic approaches to search to ensure that end users will be able to find their data easily, quickly and accurately? The kinds of search algorithms used to build and implement search software systems vary widely. There are Boolean engines, Bayesian engines, inference, vector, ranking, natural language processing (NLP) and its parts (semantic, syntactic, phraseological, morphological, grammatical, common sense), co-occurrence, clustering, sequel rules, neural nets, latent semantic and others, not to mention various combinations thereof.

Most search applications now use more than one approach in an attempt to overcome various weaknesses inherent in each approach as well as trying to complement strengths. Whether this will result in fuller, richer, more
satisfying experiences for searchers or add to their frustration and to the complexity and cost of implementation and maintenance is yet to be determined for many of the offerings in the marketplace today. There is not much research on whether combined solutions are necessarily any better than single design approaches. Does combining latent semantic indexing with NLP compound the weaknesses of each, or is the whole better than its parts? In modern search technologies much is done to surround search to bring a good answer to the requestor. This issue considers search on the practical side, using case studies to illustrate varying points of view. Some of the matters covered include dealing with unstructured information, meaning text that is not field-formatted into relational databases, but rather has been left as a full-text document in formats such as PDF or the various formats included in Microsoft Office. Other approaches suggest semantic and structural enrichment of the content as a critical part of the search solution.

Measuring search results by relevance, precision and recall (or hit, miss and noise statistics) has been discussed in detail in the *Journal of the American Society for Information Science and Technology* (JASIST) from a theoretical perspective. Visualization of the results is also an important topic, but in this issue we focus on what search software developers are doing to produce better results metrics. Pragmatic case studies look at the options and the challenges of implementation at an organization’s site. The authors were chosen to represent the breadth of actual approaches from Boolean to latent semantic and the wide territory in between.

Search challenges have been with us a long time. At the turn of the 1900s, Charles Ami Cutter and Melville Dewey took very different approaches to the way in which items should be stored and retrieved. Cutter [1] leaned toward multiple classifications and heavy cross-referencing to “make the hoard an army,” while Dewey preferred single placement and detailed classifications with see also references [2].

In the 1930s, GE Research under innovators like Homer Hall was able to use early computer logic to organize information resources. [3]

The Cold War brought us the massive research of COSATI (Council on Scientific and Technical Information, 1964) [4, 5], leading to the Recon and later Dialog systems, which dominated our world for two decades. [6, 7, 8] Many of their recommendations are still waiting to be implemented. At the same time the SDC Orbit system and the ELHILL work at the National Library of Medicine took another approach. [9] These Boolean systems were the vanguard of the 1990s.

Recently we have seen a profusion of search software from university researchers: from Autonomy, growing out of research at the University of Cambridge, to the Google cluster of algorithms and computer hardware power, born from the Stanford University research. The discussion continues as to whether it is better a) to add value to data with controlled vocabularies and metadata or b) let the software, properly tuned, do it all with no need for human intervention. The truth, as those blind men found, is the whole is usually greater than the sum of its parts – there will not be “The One Solution.”

We once worried about reaching the “elusive end user.” Now, with the advent of Internet protocols and reaching near ubiquitous accessibility, everyone has begun to search. The challenges faced in the last century are still with us. The papers in this issue exemplify some best-of-breed directions from practical perspectives. All of the authors are involved in commercially viable systems. Their pragmatic approaches are, by necessity, steeped in the results of search research. Meetings like the ASIS&T Information Architecture Summits address the frontiers of user interface design. The underlying architecture is discussed in this issue.

The article by John Holt takes us through the evolution of search to its current level. Similarly, Rich Turner takes us on an exploration of next generation search platforms. The article by Daniel Tunkelang outlines the options for relevance and ways to measure the results. Rafsky takes us one step further demonstrating through a case study the challenges encountered in building a search service for business and financial news including real-time delivery and the need to measure the impact for each returned search and to bridge the gap from system-based to user-based ranking.

The differentiation between the search presentation layer, the user interface and the search software underneath is not always apparent. The article by Ronald Millett makes this distinction clear and provides examples of the data on two different search software applications with the same overlying presentation layer. He explores the relationships between taxonomy,
metadata, fields and facets, and the ways in which taxonomy can work with search. In the end, there is always an inverted file to associate the query with the data. Most systems separate the display files from the search files to speed the response time for the user. The article by Darrell Gunter shows the levels of options available, including personalization through search.

Here’s an article-by-article preview of the contents of this issue:

**John D. Holt and David J. Miller, “An Evolution of Search”**

A new stage in the evolution of search has arrived with the advent of entity-based search. The linking or clustering of documents or records into sets of references that describe an entity can be used for much more than just reporting on an entity. The entity can become the object of the search. Entity search builds upon Boolean and relevance ranking techniques, providing improvements in both precision and recall. Precision is improved because the entities returned are consistent with the attribute values supplied in the search. Recall is improved because the combination of entity values specified in the search expression need not appear in any particular underlying reference document or record.


Search software companies are looking to develop the next generation technologies that will drive past keyword search by understanding the concepts within unstructured information. Three technologies are receiving wide attention: natural language processing, Bayesian inference and latent semantic indexing and analysis. This article examines all three technologies, exploring what they do and the pros and cons behind each technique.

**Daniel Tunkelang, “Reconsidering Relevance and Embracing Interaction”**

The query does not provide search engines with enough information to reliably determine how relevant a document is to users’ information needs. Human-computer information retrieval (HCIR) tools provide users with opportunities to clarify and elaborate their intent. If the engine isn’t sure what users want, it can ask them. The author considers the three goals of transparency, control and guidance.

**Lawrence C. Rafsky, “Searching Real-Time Financial News: PAR for the Course”**

Many challenges are encountered in building a search service that includes real-time delivery for business and financial news. PAR–Presumed Action Response is built to measure the impact for each returned search and bridge the gap from system-based to user-based ranking.

**Ronald Millett, “Integration of Taxonomy and Keyword Searches: A Comparison of Two Implementations”**

Two search implementations, one Bayesian and one Boolean, which combine taxonomy and keyword search using the same database and thesaurus are compared. The advantages of including taxonomy-based, subject descriptor searches and keyword searches in the same system were seen in both test systems. More relevant results were achieved through tapping multiple dimensions of search. The Search Harmony system combines several search methodologies in one, unified system. These methodologies include navigational trees, subject descriptors, auto-completion, search within results and relevance-based full-text searches including fuzzy search features such as stemming. Combining all of these features in a single interface improves search results.

**Darrell W. Gunter, “Semantic Search”**

The author offers two case studies to support the exploration of the differences between traditional Boolean search and semantic search. Semantic search is a process used to improve online searching by using data from semantic networks to disambiguate queries and web text in order to generate more relevant results. Using semantic technology creates proprietary aggregated visualizations of the key concepts of the aggregated dataset – presented in a bar chart of relevant weighted concepts. The contention is that this allows researchers to be more proficient and efficient in their search. ■

**RESOURCES on next page**
Resources Mentioned in the Article


An Evolution of Search
by John D. Holt and David J. Miller

The technology of information retrieval systems continues to evolve, and in particular, the technology of search has continued to evolve. A new stage in the evolution of search has arrived with the advent of entity-based searching. This paper provides a brief review of some of the earlier stages of search evolution in the context of the evolutionary pressures of the concurrent improvement of both precision and recall.

Early Boolean Search
Efficient mechanical Boolean search of records is a 19th century invention. The extension of Boolean search from records (structured data) to text documents was initially accomplished by the simple expedient of creating a summary record of the text using a controlled vocabulary and a taxonomy. The records under search could refer to books, papers, notes, works of art or even public records. The subset of records that satisfied the Boolean predicate was selected, and the records that did not satisfy the predicate were left behind. In the case of needle-sort cards in buckets, this is a literal description of the process. The scope of the Boolean predicate was by necessity the record or fields on the record. The searcher could use the AND, OR and AND NOT Boolean operations in the construction of the search predicate.

The resolving power provided by the mechanical search approach was very low. The number of controlled vocabulary terms or taxonomy leaves was limited to the length of the card edges. In addition to searching through cards, print indices could be searched as well. A much larger number of controlled vocabulary terms could be used in the construction of an index such as the Readers Guide to Periodic Literature, first published in 1901. An industrious researcher could perform the AND, OR and AND NOT operations by keeping lists of the document references and manually performing the operation upon those lists.

Boolean Search and the Digital Computer
The digital computer allowed for a large increase in the number of controlled vocabulary terms by the middle of the 20th century, and the number of terms that could be used expanded from about a hundred to several thousands. The search was still performed upon an extract of the original papers and upon indices. Computer memories were too small to seriously consider large collections of text or structured records.

The primary and secondary memory of the digital computer increased in the last quarter of the 20th century to the point where the complete text of papers, articles or large collections of structured data could be stored in computer systems. Additional operations such as adjacency became required to allow users to specify search predicates consisting of word patterns. These word patterns provided the context necessary to disambiguate the individual words.

A deep understanding of the document collection was required to properly specify the patterns of words. If the patterns were too loose, the number of documents that satisfied the predicate became too large for effective review. The resulting documents could sometimes be sequenced to allow the searcher to stop the review of the answer documents before all of the documents were examined. Unfortunately, most of the time there were no ways to naturally sequence the answer documents to provide an early termination of the review.
The search system stemmed the words to varying degrees to reduce the word variants that a searcher would need to enter. The system generally would allow the word strings to include wild card characters to provide additional flexibility. Some systems introduced word equivalents for common spelling variations. The notion of word equivalents eventually expanded to include alternate names of famous organizations.

The search of structured and semi-structured data collections presented similar problems. To achieve a result set of reasonable size, a highly specific search would need to be formulated. Unfortunately, a highly specific search was unlikely to return all of the records of interest. The negative impact of recording error upon structured and semi-structured data searches is higher than the impact of recording errors in text documents because there is much higher degree of redundancy found in text documents.

**Relevance Ranking the Result**

Gerald Salton is generally credited with the observation that more frequently occurring terms are less distinguishing than less frequently occurring terms. The inverse term frequency of the search terms can be used as weights to rank the documents. Most text search systems today use some variation of the vector space model where each word in the collection is considered an attribute and therefore a dimension. The search expression is treated as a vector of terms, and each document in the collection is also treated as a vector of terms. The similarity of the search vector to each document vector is calculated using the weights of the matching terms to score the importance of the match.

The document ranking provided good results with very loose searches. The searcher no longer needed a deep understanding of the word patterns that were likely to be found in the relevant documents. The searcher was still required to know the vocabulary of the subject.

The presence of phrases and wild-carded string expressions complicates the process of ranking, as it is not obvious how to determine the weights efficiently. Index terms assigned to documents can also be used in the search process but can be difficult to use in the relevance ranking calculation.

Statistical indexing can also be performed upon text documents. The statistical indexing approach uses the number of occurrences of content words and noun-noun phrases to calculate a statistical digest of the document. The search can be restricted to the set of statistical index terms to provide improved precision. The ranking can then be performed using the document terms.

Relevance ranking can be extended to semi-structured and structured information searching. The search again is treated as a vector, and each record in the collection is also a vector. The degree of similarity between the search and each of the records in the collection can be measured, and the most similar records are returned to the researcher. In practice, the usual technique is to calculate the degree of mismatch between the search and each of the records in the collection instead of the similarity.

Structured information has far fewer attributes than text documents where each unique word can be considered a different attribute. Since each attribute is treated as a dimension, the structured information record has a much lower dimensionality than a typical text document. This lower dimensionality of the structured information record enables the use of more sophisticated measures for similarity. In the case of the document, a term was either present or it was not present. For structured information, each attribute can be evaluated for the degree of similarity between the attribute values. For example, two strings can be compared to each other using an edit distance measure to determine the degree of similarity.

**Smarter Data**

The problem of vocabulary mismatch can be ameliorated by adding indexing terms. Indexing terms also remove some of the need for the searcher to anticipate the word patterns. Indexing alone falls short of achieving complete disambiguation.

Additional terms can be added to the search index of the document collection to supplement an original sequence of words. The sequence “Falkland Islands” can be supplemented with “Malvinas” as an alternate term. This is a simple example of a data enhancement. In the instant example, a search looking for “Malvinas” in the same paragraph as “Exocet” would find a story about the Falklands war and the use of an Exocet missile to sink the HMS Sheffield.
Semantic indexing and entity recognition are tools that can be used to make the documents easier to find. The user now only needs to specify the concepts or entities that must be discussed for the document to be returned. Unfortunately, semi-structured data such as public record data does not lend itself to semantic indexing techniques.

Entity recognition is a difficult problem. A natural language parsing process can be used to say that a particular string references an entity or concept, but cannot say anything about the entity referenced. An ontology is required to resolve the reference. An ontology for concepts, historical events, historically important people and organizations and places can be constructed by human efforts from historical accounts, biographies, gazetteers and maps. But what of an ontology of ordinary people and organizations or even of emerging concepts?

**Discovery of Entities in Data**

Text documents, semi-structured data (for example, bibliography entries) and structured data can all be processed to discover entities. Entities can be individual people, transient organizations (for example, the co-authors of a paper), organizations, places or events. The document text or record refers to an entity, but is not itself the entity.

Some entities can have concrete identities. Place names can be tied to particular geographic areas. A geographic area can be definitively identified with coordinates and boundary information. Time entities, like 7 December 1941, can be definitively identified. Some organizations can be definitively identified, though in practice fewer than one would think.

In general, very few non-geographic and non-chronological entity references can be definitively identified as a reference to a particular entity. What can be done is to group or associate the references into internally consistent sets of references. A particular set of references then describes an entity, and an entity is described by a set of references. The preferred error is to have more than one set of references describe a particular entity. In the domain of document collections, the process of the creation of sets of references to the entities is a form of clustering. In the domain of records management, the process of the creation of sets of references to the entities is known as record linkage.

A reference to an entity has attributes. A set of references that describe an entity will in practice have some subset of attributes in common for a subset of the references. It is quite possible and desirable that there not be any attributes that are in common for all of the references. There need only be enough commonality among the disparate references to link all of the references together.

For example, consider the case where there are three sources of information (source A, source B and source C), and each source has a variety of fields. There are three different pairings of the sources, A-B, A-C and B-C. As long as any two pairings have distinguishing fields in common the records can be linked. Each record is a reference to an entity.

Once the references (either records or documents or passages of documents) have been gathered into sets, the fact of set membership can be used for searching and reporting. The use of the set for reporting is simply the aggregation of the references that make up the set into a virtual document or record. There are two methods of leveraging the set of references to improve the search process for finding the documents or records of interest. Given a found record of interest, the information extracted from the other members of the set of references can be used to form a search that is both more precise and more inclusive. Alternatively, instead of using the individual records or documents as the object of the search, the entity is directly used as the object of the search by using the complete virtual document or record.

**Example**

Consider a collection of bibliographies and the abstracts and the keywords for each of the referenced papers. We can discover the ad hoc or transient organizations of co-authors and enhance the collection by adding entity information indicating the group of co-authors responsible for a particular paper. Group membership can be used to disambiguate the names of the individual authors. The journals and conference proceedings referenced can be categorized by clustering the keyword lists from the papers published in the journal or conference proceeding.

In some cases, there will be additional information available, such as
e-mail addresses of one or more of the authors, the academic affiliations of the authors and a credit to the grant and sometimes the grant identification.

Crawling the websites of academic institutions gathering lists of then-current faculty can augment the bibliography collection. Most faculty will maintain a list of their publications as well, and these lists will provide additional statistical information. Note that without human intervention it is not possible to assign a semantic meaning to a list of bibliographic references on a faculty member’s web page. It may be a list of publications authored by the faculty member or it may be a suggested reading list.

The group entities are discovered using statistical record linking or clustering. The individual entities can be disambiguated using group membership as well as other attributes. The statistical linking or clustering will use the list of authors, the journal or conference proceeding, the cluster identity of the journal or conference proceeding, the publication date, idiomatic expressions found in the abstract and any other attributes that are available on a sufficient number of the records in the collection.

Using Entities to Improve Search Results

A search finds the set of records that match (or best match, if a relevance-ranked search) a predicate. The notion of expanding or augmenting a search has been shown to be an effective method of improving recall. The typical method of expansion is to use a statistical thesaurus. The query terms are used to look up thesaurus entries. The associated terms are pooled and a subset of terms is selected based upon criteria such as the degree of commonality.

A collection of entity information can be used as a statistical thesaurus for the purpose of query expansion. The expansion of the query with information from the entity may not be quite as simple or as straightforward as traditional query expansion. The number of terms in the traditional expanded search will vary with the number of matching thesaurus entries. Unfortunately, the number of expansion terms available from an entity-based thesaurus will vary by the number of entities related to the initial set of search terms. A simple expansion will only be practical when the number of initial entities involved is reasonably small, such as a few hundred entities.

Another form of query expansion is a “More Documents like this Document” search. The researcher runs an initial search that is fairly broad, browses the list of answer documents and finds a suitable seed document. The researcher then requests that the search system use the descriptive terms and phrases of the seed document as a second search to create a highly relevant document. A similar process can be used with entities. It is important to recall that our entities are really just references and there may be multiple sets of these references to each actual entity in our collection.

The ability to specify more complex search predicates is another means of improving the search results. Consider the search where you might want to find authors who had published papers on the same disparate set of topics. The search system could simply use the least frequently occurring topic in the set to get the initial list of author entities and use the list of author entities to retrieve the corresponding publication records and then filter the list of publication records by the remaining topics.

The expressive range of this approach for efficient search and retrieval is limited to the information found in the individual records in our collection and upon search criteria that produce intermediate result sets of manageable size. Information that is derived from some aggregate operations upon sets of records will require another approach. Search criteria that create intermediate result sets that are too large will also require a different approach.

Using Entities as the Objects of the Search

As was noted above, we really do not know the entities. What we do know are internally consistent sets of records that reference an entity or entities. The sets are constructed to prefer an error where more than one set refers to a particular entity and to avoid an error where a set refers to more than one particular entity. It is convenient to adopt the fiction that these sets of references are the entities.

A document can be constructed for each of the entities in the form of a report. An author group entity report can contain the reports of any individual author entities that can be unambiguously linked to the author group. The report can show the group core members, those authors that have participated at a very high rate, peripheral members that have participated in a plurality
of the publications and ad hoc members. The report need never actually exist, but can be virtual.

The virtual document is more than just the union of the individual records that comprise the virtual document. A variety of aggregations can also be included in the virtual document. In our present example, aggregations such as the average number of publications per year or the average length of time between publications can be included in the virtual document.

The creation of a set of complex derived attributes, such as aggregates, during the creation of the database supports efficient search and retrieval. If the data collection is small enough, these operations can be performed during the search process. However, for any reasonably sized collection there will be far too many intermediate results.

The search expression is matched against the virtual document or record. The virtual record has a significant number of attributes. It can be both convenient and effective to leverage the availability of these attributes by using a forms-based search interface. The system presents the researcher with a form. The entry boxes on the form solicit the attribute values from the researcher. The researcher now only needs to know the attribute values of interest. The system can match these attribute values against the attribute values for the entities and deliver the closest matching entities. It is important to note that it is not necessary for a single record to exist that contains all or even a plurality of the attribute values supplied.

Semi-structured information and text documents contain large numbers of text strings as attribute values. In the instant case, the abstract is an example of an attribute with a large number of text strings. The search form mechanism discussed above is inadequate when the attribute has a large number of text strings.

It can be helpful to support extensions to the search form interface to solicit the specification that the strings entered on the form should be from different or the same component records. The system can support a search specification of more complex relationships between text strings because the system assembled the virtual document.

**Conclusion**

Entity search is another step in the evolution of information retrieval systems. Entity search builds upon Boolean and relevance ranking techniques. Entity search provides improvements in both precision and recall over traditional Boolean and relevance ranked search techniques.

Boolean search techniques require the researcher to be knowledgeable of the words and expressions used in the document or record collection. Precise results can be obtained, but at the cost of a significant drop in recall. Recall can be achieved, but only at a significant drop in precision.

Relevance ranking via statistical techniques can be used to improve apparent precision in some cases. However, the statistical techniques do not apply well to searching structured and semi-structured data with attribute values.

The linking or clustering of the documents or records into sets of references that describe an entity can be used for much more than just reporting on an entity. The information from the set can be used in some cases to improve recall by broadening the search. Alternatively, and more powerfully, the entity can become the object of the search.

A search expression that specifies a set of attribute values can be used when the entity is the object of the search. Both precision and recall are improved. Precision is improved because the entities returned are all consistent with the attribute values supplied in the search. Recall is improved because the combination of entity values specified in the search expression need not appear in any particular underlying reference document or record.
All of us are very familiar with search – it’s Google, and it’s ubiquitous. Google has done a very good job at monetizing search for the masses. The company incorporates elements of many different technologies, both Boolean and conceptual; it leverages these technologies to quickly mine content; and it presents results in a very familiar, comfortable, non-threatening way. Google has even extended its platform into the enterprise search market, and Google One-Box has a commanding market share.

What is very important is to separate Google the business from Google the software. As a business, Google is an advertising company – plain and simple. It is not unlike the early TV broadcasting companies, when 1950s era soap operas were just that – engrossing melodramas written for housebound homemakers, sponsored by laundry detergent manufacturers, with the sole purpose of promoting more soap.

This business model works very well for the masses. Rail as we might that Google doesn’t provide unbiased answers, that Google searches only present a small snippet of all the information that’s out on the web, or that Google seems to be promoting one company over another, without advertising revenue there would be no possible way a company could – or would – invest the hundreds of millions of dollars Google has spent to develop and deploy its search.

The model works somewhat for business-to-consumer marketers. The Google search paradigm (its uncluttered keyword search window) is copied by most websites, and the Google advertising paradigm (presenting relevant advertisers within or as part of the results) can be seen on most consumer information sites such as maps and news services.

But it doesn’t work well for businesses. Even One-Box adopters struggle with the “fixed parameters” that drive Google. In business, there is no advertising revenue for search: It is a means to an end, a necessity to find critical business information and to act upon it.

Therefore, over 100 companies – a number that continues to grow rather than shrink – provide some type of search software to a variety of business customers. These companies exist because they are pioneering technologies that focus on searching unstructured information, which is the bulk of what we store and capture today. Searching structured data like spreadsheets and databases is fairly mature. The same holds true for keyword or Boolean search – by themselves, these mark-and-index solutions are widely deployed, and some are even available as open source software. The focus for search software companies is the next generation technologies that will drive past keyword search and better address the exponential explosion of unstructured information by understanding the concepts within that information.

Many enterprise implementations combine techniques – much like Google did – but focus on providing businesses with robust platforms that can address the challenge of finding actionable information. There are three technologies that are getting wide press and have found applicability in numerous business solutions. This article will look at all three, what they do and the pros and cons behind each technique. Which is best? The sad answer is that it all depends. Knowing what your options are is the first step to finding that answer.

The three technologies we’ll look at are natural language processing or NLP, Bayesian inference and latent semantic indexing and analysis. One of these is based on sentence structure, and the other two are based on mathematical algorithms, but all three try to solve the challenge of understanding meaning so that search can be based on concepts versus

Rich Turner is with Content Analyst Company. He can be reached by email at rtuener<at>contentanalyst.com
keywords. Why is this important? Simply, because language is complicated and fluid, and there are many ways to express similar things. For cognitive beings, this complexity is a wonderful adaptation, but for computers – whose processing is binary by nature and design – it is a huge drawback. Boolean operands attempt to bridge this gap, but remembering if/and and or/else combinations, performing “fuzzy search” tricks, don’t address the problem head-on. These technologies all do, in one form or another.

Natural Language Processing

Natural language processing or NLP is the oldest and best-known computer-based language technology. The idea behind NLP – which sprang from the same Xerox-sponsored Palo Alto Research Center (PARC) think tank that brought us the mouse and ultimately Windows – is that there is structure behind language, and if you can understand the structure, you can understand the language. This assumption makes perfect sense: We communicate using sentences, and they are comprised of nouns, verbs, adjectives and the like. Noun + action = result.

NLP doesn’t work by itself; what researchers found was that within the language there was no inherent expression of meaning. Simple sentences could be expressed as “noun + action = result” but more complicated thoughts sent early NLP systems into a tailspin. What was needed – and are now available in incredible variety – were word lists. Some lists are dictionary-based – they impart meaning into the words that NLP has deconstructed from language. Others tackle language nuances called synonymy (multiple words with similar meanings) and polysemy (words with multiple meanings often based on context).

From the start NLP needs structure – that is, a taxonomy or word hierarchy because word lists are very specific to subject matter. A medical term list makes no sense to an aerospace engineer, and 20th-century synonyms might not be meaningful to Y-generation conversation and writing. So NLP correlates lists of terms and definitions (and how those terms and meanings are combined) to express meanings, concepts and context. Does NLP work? Yes, very well, and it’s deployed in many variations across a range of solutions. Are there challenges? Definitely. Because NLP depends upon outside structures such as dictionaries and thesauri to derive ultimate meaning, those structures must be maintained – constantly, as language continually evolves. Which outside structure is also important – not all words are created equal, to paraphrase Thomas Jefferson. And finally, a good taxonomy is needed to put all this into context.

Technologies Based on Mathematical Algorithms

The two other technologies are related in that they use mathematics instead of language structure. Why math? Two reasons, actually. The first is that the structure of language – like any other repeatable structure – can be expressed in mathematical terms. The second is that by expressing language in terms of math, you can potentially avoid the need to use the outside structures discussed above. The promise of using mathematics-based search technologies is that you can create and deploy a more robust engine that won’t need to know particular information about what it’s indexing yet can still derive meaning. In theory, size would also not be as big a challenge, for example, to scalability. It shouldn’t matter how complex a sentence or thought was if you could express it mathematically. The two mathematical techniques look at different spectrums of the mathematical nuances of text.

Bayesian inference. The first we’ll discuss is Bayesian inference, which actually harks back to Rev. Thomas Bayes, who in the 17th century proposed his famous Bayesian Theorem. (Note to readers: whether Bayes actually came up with this theorem is one of those hotly contested historical debates). Bayes’ Theorem is a probability theorem, originally proposed to explain gambling outcomes. The idea is that if something occurs along with other things, and a similar thing occurs elsewhere along with other similar things, then those things are all likely related. Bayesian inference, as the practical application of the theorem has been called, has been successfully applied to a number of scientific questions, from thermal transfer to energy generation.

It wasn’t until the late 20th century that companies started looking at how Bayesian inference could apply to language. It turns out that Bayesian inference can perform quite well and is able to accurately determine conceptual similarity among documents and terms. Better still, Bayesian can handle a lot of documents – in fact, it needs a fairly large corpus in order to
establish patterns, probabilities and relationships. Bayesian inference can also be successfully used to organize documents, called categorization or clustering, based on conceptual relevance.

The biggest challenge that Bayesian inference has is dealing with very complex concepts and relationships. Here, the very elegance of Bayes’ Theorem is its downfall: For each additional condition, the theorem must investigate all permutations around that condition. Complex concepts – with a lot of contingencies, conditions and facets – can cause a Bayesian engine to perform a lot of extraneous calculations, bringing throughput to a crawl. The providers of Bayesian-powered search engines compensate for this problem by linking word lists and dictionaries to their engines. While they’re not technically needed, they provide a useful assist when dealing with such concepts. One other caveat for Bayesian is that in order to accurately perform categorization based on example documents, the engine needs both relevant examples and non-relevant examples – again, the provision of both types of documents speeds the engine up, enabling it to also know what it’s not looking for when grouping documents.

**Latent semantic indexing.** The other mathematical search technology is based on word co-occurrence rather than probability. It is also a geometric model, using vector-based math across hundreds of dimensions to identify conceptual relevance. This technology is latent semantic indexing and analysis, which is not quite as old as Bayes’ Theorem, but does go back to Bell Labs in the early 1980s. The scientists at Bell Labs patented LSI, but decided it had no future at AT&T because it dealt with text and needed a lot of computer horsepower.

Recently, things have changed – quite dramatically – for LSI and LSA-based solutions. First, the computer horsepower and, more specifically, the memory that these engines need for efficiency (typically they load the entire index into memory) is inexpensive and readily available. In the 1980s memory was measured in kilobytes; today, it’s measured in gigabytes. Second, 21st-century mathematicians and engineers have figured out how to apply LSI across a wide range of challenges (it turns out LSI isn’t limited to text; it can handle voice and video and even mixed media); as a mathematical solution it is very pure.

LSI works on the idea that words (or entities if the material is non-textual) expressing the same or similar concept will typically be used together – co-located. Like Bayesian, LSI works best with a large corpus, as this technology virtually learns conceptual relationships as it indexes documents. The benefit to this approach is that LSI does not need any outside structures like dictionaries and thesauri to work its magic. The other unique feature of LSI is that its ability to learn conceptual relationships between words and documents can also be focused on learning conceptual relationships between languages. In other words, LSI can train itself on an identical set of documents translated into different languages and it will figure out, all on its own, that “Comment allez-vous?” in French is the same concept as “How are you?” in English. When used in cross-lingual applications, LSI allows searchers to input queries in, say, English, yet find relevant documents in French and German without prior translation. One LSI vendor, Content Analyst Company, even has a training space for their LSI engine. The training space is used for things like cross-lingual indexing but is segregated from the documents being searched so that the engine is easily trained for a multitude of search tasks without polluting the search results.

LSI also comes at a price: the use of advanced math with vectors being drawn across hundreds of dimensions requires robust computing power. Today’s computers – even desktops and laptops – are thankfully up to this task. The other necessity for LSI is RAM memory – lots of it. Most LSI users are running 64-bit operating systems and will load their servers up with memory to get the optimum performance from these engines. Again, the continual drops in memory cost and increasing RAM footprints in modern servers have made LSI a worthy contender in advanced search technologies.

**Making Choices**

So which technology should you use? The answer is that it just depends. What you are doing and what kind of outcome you need very much dictates the kind of environment you should consider. Companies that want to “index our entire two petabytes of corporate information” haven’t really thought that through. What do they want to find? Who needs to find it? What is their budget? How disciplined are they? That’s like saying “I want a big car, and I
want it to go fast” and then agonizing over that vehicle’s fuel consumption and mechanical intricacy because you bought a race car but you’re using it as a daily commuter.

Any search engine depends upon its index – add new information and you need, at least eventually, to rebuild that index. Multiple search engines means multiple indexes. Unlike keywords that are static and have no meaning attached in the search world, conceptual search engines find and identify concepts from the whole text being indexed. Throw in enough new terms, and new concepts will emerge. Most engines can rebuild an index while still allowing users to perform searches, but those new concepts won’t emerge until that re-index is done. Again, it comes back to what you are trying to accomplish.

Enterprises are wise to take a page from Google if they are assembling their own search environments: Use the best appropriate technology for the specific task, and don’t be afraid to have multiple technologies in your shop. One size doesn’t fit all. Vendors are increasingly tuned into this as well. Industry giant Microsoft, who has its own proprietary search technologies, integrates its FAST subsidiary’s advanced search into a variety of products. NLP vendor Cognition is as well known for its industry-specific word lists as it is for its NLP search product. And CAAT, a conceptual search engine from Content Analyst Company, is shipped to customers tightly coupled with dtSearch®, a Boolean and keyword search engine.

Which technology is better? The fact that this debate has raged for nearly a decade ultimately means that they are all worth considering for next-generation search requirements. Each has its strengths and weaknesses and nuances where each outperforms the other. What is far more significant is how the technology is implemented. A fantastic search engine can be quickly hobbled by an inadequate user interface. A simple, elegant UI (user interface) will quickly be dismissed if it lacks the expected horsepower beneath.

There are certainly areas where one technology outshines the others – and depending upon your industry, one or two of these technologies may dominate. That popularity should not be the guiding consideration, however. What is of paramount consideration is the degree to which the implementation satisfies the requirements of your user community—or if you’re creating your own interface, your development specs. Search needs to be a tool, and next-generation search is no different. The deployment of any search solution needs to be in lockstep with your users’ expectations, skill levels and work requirements. Those requirements put the onus squarely back on the vendors. Technology for its own sake rarely satisfies. Make sure your vendors understand your expectations, make sure you and your users are comfortable with the usability of their solution and make sure that deploying it will be seen as a solution, not another issue in your users’ daily work lives.

---

**Resources for Further Reading**

For more information about NLP (natural language processing), visit the following Wikipedia page: [http://en.wikipedia.org/wiki/Natural_language_processing](http://en.wikipedia.org/wiki/Natural_language_processing)

For more information on Bayesian inference, the following Wikipedia page is very helpful: [http://en.wikipedia.org/wiki/Bayesian_inference](http://en.wikipedia.org/wiki/Bayesian_inference)

For details on latent semantic indexing, the appropriate Wikipedia page is [http://en.wikipedia.org/wiki/Latent_semantic_indexing](http://en.wikipedia.org/wiki/Latent_semantic_indexing)
Reconsidering Relevance and Embracing Interaction
by Daniel Tunkelang

We live in an age of ubiquitous search. Children growing up today in the developed world would probably find a world without search engines as inconceivable as the rest of us would find a world without electricity and running water. Moreover, search engine users—that is to say, almost all of us—expect those engines to work like magic, extrapolating our information needs from a pair of words (on average) and responding with the content most relevant to those needs.

Of course, search engines do not always meet these expectations. Thousands of engineers work at—or with—search engine companies to tune the ways those engines model and compute relevance. Relevance, however, is not a one-size-fits-all measure. For example, web search engines typically measure the relevance of web pages using bibliometric considerations such as Google’s PageRank measure and the “anchor text” associated with links connecting into pages. In contrast, enterprises cannot typically rely on a rich network connecting the documents stored in their repositories; hence, their search engines look for relevance signals in document structure (for example, title and keyword fields) and consider factors specific to their domains or applications. Of course, almost every search engine’s filtering and ranking algorithms take into account the relationship between words in the search query and words that occur in documents.

The Problem with Relevance

But relevance as a concept is problematic. In 1964, William Goffman wrote, “[T]he relationship between the document and the [search] query, though necessary, is not sufficient to determine relevance” [1]. In other words, the query does not provide the search engine with enough information to reliably determine how relevant a document is to the user’s information need. Library and information scientists like Tefko Saracevic and Nicholas Belkin have been saying as much for years, emphasizing an approach to information seeking that focuses on the user [2, 3]. Ironically, the success of commercial search engines has marginalized many of their concerns: search engines do well enough without considering the users—at least for web search—that many search engine researchers and developers have been content to ignore Goffman’s warning.

Lately, however, there have been signs that our honeymoon with purely ranked retrieval systems is coming to a close. Bing, Microsoft’s recent relaunch of its web search offering, touts itself as a “decision engine” rather than simply an engine to match search queries to documents. More substantively, Bing’s interface offers users a variety of ways to interact with the search engine. For example, a search for “Barack Obama” offers a set of query refinements that includes news, biography and interviews. Google, though not quite as aggressive on this front, has been slowly promoting similar features, initially relegating them to its advanced search and experimental interfaces but increasingly raising their visibility. Last but not least, tools like Yahoo!’s Search Assist provide real-time query suggestions to users as they enter search queries. On the web, we are seeing the initial signs of search engines engaging users in a more interactive query elaboration process.

Meanwhile, search applications outside of general web search have been embracing interaction for several years. In particular, faceted search has become a staple of online retail sites. Endeca established faceted catalog search as an industry standard, and mega-retailers Amazon and eBay have...
since built out their own implementations. Today, site developers have a variety of commercial and open-source options available to them.

Faceted search combines conventional text search with the ability to refine query results using multiple facets of those results, enabling users to navigate content in multiple ways, rather than in a single, pre-determined, hierarchical order. For example, a shopping site might allow users to refine product results by such facets as product type, brand and price, while a directory might allow users to refine people results by age, gender and location.

Online retailers are some of the most visible users of faceted search, but they are by no means the only ones. Libraries are increasingly using it for their next-generation online public access catalogs (OPACs), almost a century after the Indian librarian S. R. Ranganathan promoted the advantages of a faceted classification scheme as an alternative to hierarchical schemes like the Dewey Decimal system [4]. More broadly, site and enterprise search applications have learned that the best way to help users satisfy their information needs is to provide users with a richer interface and user experience. The “10 blue links” interface is clearly not enough to satisfy a wide variety of information seeking applications.

Human-Computer Information Retrieval

Guided query refinement and faceted search fit into a class of approaches called human-computer information retrieval (HCIR), a term coined by Gary Marchionini [5]. The HCIR approach advocates for tools that bring human intelligence and attention actively into the search process. Rather than guessing what users need, these tools provide users with opportunities to clarify and elaborate their intent. If the engine isn’t sure what users want, it just asks them.

In economic terms, HCIR aims to offer users better return on investment. Instead of slavishly accepting the constraints of the current interaction metaphor (users enter two words as input and see a ranked list of ten results as output) and attempting to optimize the user experience within those constraints, a search engine can allow users to get more if they give more. But what should it ask users to give? And what will users get in return?

We answer those questions in reverse order. A search engine should provide users with an interface that offers transparency about the engine’s internal state, control over the filtering and ranking of results, and guidance to inform an adaptive information seeking strategy. In return, users can and will help the search engine help them, using the available query elaboration options to clarify and adjust their expressed information needs throughout the information seeking process.

Let us now consider the three goals of transparency, control and guidance.

Transparency

The premise of transparency is simple: users should know why a search engine returns a particular response to their query. Note the emphasis on “why” rather than “how” since most users don’t know – or care – what algorithm a search engine uses to in order to assemble a response to a query. What users do care about is whether the engine “understood” their query – that is, whether the engine is at least attempting to address the information need that the user intended to express.

A recent study by Autobytel and Kelyon Research, entitled “The State of Search,” reported that most users expect search engines to read their minds [6]. Of course, such an expectation is unreasonable – even our closest friends cannot read our minds. What frustrates users most, however, is when a search engine not only fails to read the user’s mind, but also gives no indication of where the communication broke down, let alone how to fix it. In short, the search engine fails to provide transparency.

We are familiar with the quip that “to err is human, but to really foul things up requires a computer.” It shouldn’t surprise us that a search engine may really foul things up when it attempts to extrapolate an information need from a couple of words. If, however, the search engine provided the user with a clear picture of how it arrived at its response, then the user could adapt to the engine’s limited cognitive capabilities – in particular, the engine’s inability to read the user’s mind.

Control

Transparency is necessary for users to establish effective communication with search engines, but it is not sufficient. Indeed, transparency might even cause more harm than good if the users have visibility into the system’s
misunderstandings but no ability to resolve them. To overcome this problem, search engines also need to offer users control. The ultimate arbiter of the information seeking process must be the user, not the system.

Interestingly, early search engines offered users more control than most modern ones. In particular, commercial search engines in the 1970s and 1980s, some of which are still used for applications like electronic discovery, offered users Boolean search, enabling users to construct arbitrarily complex queries that the system would follow precisely. Indeed, Boolean search interfaces not only provide control to users, but also offer them a high degree of transparency. Unfortunately, their lack of guidance (which we will discuss in a moment) proved fatal: most users – even professional information seekers – are not adept at query construction, especially when they are querying unfamiliar content repositories [7].

But many people learned the wrong lesson from the weaknesses of Boolean search interfaces. As a result, modern search engines generally offer a ranked retrieval model that not only is opaque to users, but also offers them minimal control over the information-seeking process beyond the ability to enter words into the search box. In particular, the ranking of results, an essential function of the search engine when the number of matching results far exceeds the number shown to the user, is typically outside the user’s control.

There is no reason to prevent users from having full control over the filtering and ranking of results. Offering users control does not preclude (or excuse) a search engine from making a best effort to satisfy users who do not exercise that control. Web search providers have argued that the threat of spam and unscrupulous search engine optimization requires protecting users with secret filtering and ranking algorithms – an argument reminiscent of security through obscurity. To the contrary, giving users more control would actually offer them more robust protection, since the diversity of user behavior would dilute the spammers’ efforts. All search engines should treat their users as the primary actors in the information-seeking process, rather than mostly passive recipients of their paternalism.

Guidance

While transparency helps users understand the system internal state, and control allows users to override the system, these two ingredients, as shown by the early experience with Boolean search systems, are necessary but not sufficient. The third essential ingredient to ensure a successful user experience is guidance: The system must not only respond to users’ queries, but also help users formulate those queries.

In the 1990s, Peter Pirolli and Stuart Card developed a theory of information foraging, establishing an analogy between the way users seek information and the way animals forage for food [8]. The core concept in this theory is “information scent,” the indicators or clues that lead users to choose the next step in the information-seeking process with the expectation that it will help them progress toward their goal. Earlier work by Marcia Bates described a “berrypicking” model of information retrieval, in which a user does not satisfy an information need in a single query, but rather through a series of steps in which the user either finds information directly relevant to the information need or learns how to adapt the search strategy [9].

Like the information architects who organize the content on websites, search engine designers should aspire to provide users with scent at every step of their information-seeking process. Techniques like query suggestions, faceted search and results clustering all offer users the opportunity to make progress on their next step, rather than always having to restart the information-seeking process from scratch. Indeed, faceted search is a popular technique for offering users such guidance. While users are ultimately responsible for expressing their information needs, it is the search engine’s job to act like a reference librarian and help the users in this process.

From Vision to Reality

Hopefully you are now convinced that search engines should offer users transparency, control and guidance. Granted, it is one thing to advocate these goals and another to actually achieve them. A key obstacle is inertia. Both users and search engine developers are comfortably familiar with a standard interface and wary of making radical changes to it. Nonetheless, the successes in highly visible domains like online shopping have demonstrated that users will take advantage of better interfaces, given the opportunity. Meanwhile, researchers and practitioners are continuing to invest in better
methods to support the query elaboration process. The status quo is a significant challenge, but not an insurmountable one.

As users, we must demand to be treated as masters of our information-seeking destiny, embracing the responsibility concomitant with that power.

As search engine developers, we must place users first, giving them control in order to get their active participation in the information-seeking process. Together, we can bring forth a brave new world of human-computer information retrieval.

Resources Mentioned in the Article

Searching Real-Time Financial News: PAR for the Course
by Lawrence C. Rafsky

Real-time news (also known as live news, streaming news or breaking news) – especially news focused on business and financial matters – is widely read and builds up quickly. A typical newspaper/newswire aggregation system will process three stories per second around the clock, adding approximately 250,000 stories every business day to the collection. Large commercial systems substantially surpass this total.

The needs and behavior of end-users searching collections of this type differ from general searching norms in several key aspects:

- The computational burden is closer to the classic alert/routing problem than it is to the ad-hoc search problem.
- Business work is about the opportunities of today and tomorrow.
- All news is structured.
- User queries do not typically consist of a few terms.
- Users have a high degree of topic familiarity and topic focus.

The Computational Burden

The computational burden is closer to the classic alert/routing problem than it is to the ad-hoc search problem (these are standard terms-of-art in IR – as, for example, Robertson [1] explains). In fact it closely matches the batch routing problem, because (1) users save searches and re-execute them frequently (rather than entering searches in an ad-hoc manner) and (2) very little latency is tolerated by users. A news article arriving in the last few moments should be returned as a hit if the article is a search match to a just-executed query.

Combine this low-latency requirement with a saved-search, re-execution cycle that is often a minute or less (programmed as web page auto-refresh), and the similarity with alerting becomes obvious. Thus the old folk-theorem, “In search you have all the time you need to study the archive, and no time to study the query; in alerting you have all the time you need to study the query, and no time to study the archive,” doesn’t hold here.

However, real-time news is still search. A static set of matches (in this case, news headlines and summaries) is returned, and the user demands some ranking of the results by meaningfulness and pertinence. In the alert problem it is sufficient to present the user with a temporal stream of matches; not so here.

Thus real-time news search demands the computational agility of alerting and the semantic processing of search. The job is made somewhat easier by the batch nature of the query collection – the saved searches. In the commercial system NewsEdge™, of the last million searches (looking backward from August 15, 2009), approximately 96.5% of executed searches were repetitive submissions of stored queries. This statistic excludes users who specifically posted alerts and received a content push from NewsEdge – it refers only to users who requested (perhaps automatically) and received a web page of results.

The Opportunities of Today and Tomorrow

For the most part, real business work (as opposed to academic business study) is about the opportunities of today and tomorrow. Searchers care less and less about stories as they come from earlier and earlier time points. Thus “article publication date” is not merely a time facet or a temporal query profile [2]; it is a fundamental part of the search process.

Furthermore, there is a fundamental time flow to news: themes develop,
expand and mature over time. Arriving news stories reflect this rhythm. Thus news temporal order has a fundamental effect on how a retrieval system handles clustering, novelty discovery and redundancy [3].

One common approach is to enforce a time cut-off, to limit retrieval to, say, the last 180 days (unless specifically directed otherwise by the user) and then proceed normally with all calculations such as relevance. Another approach is to use story time (more correctly, age) directly in relevance formulas, down-weighting older stories.

Taking a cue from the similarity between the low-latency news search problem and the alert problem, some systems display results only in reverse chronological, or time, order, and attach a relevance value to each headline. A twist on this, which is perhaps more user-friendly, is to flag or star stories – still displayed in time order – that have a relevance score above a threshold.

All News Is Structured

All news is structured – whether or not the document arrives in html, xml or plain text. News – by definition – has at least a headline and one or more of the following: deckline (sub-headline), byline (author and location) and body (the story itself). And like all documents there is a source and potentially a copyright, but these features are not distinguishing characteristics of news. Moreover, news articles from printed newspapers – and some websites – have typeface weight and page/section placement data that convey valuable information, which has been exploited in a number of IR systems, although we are covering unified collections of news here, including newswires and blogs that lack this structure. There can also be summaries or leads (similar to abstracts in scholarly papers) – or, if not, these items can be synthesized by exploiting the top-down structure of news writing (more on that below) and metadata sections containing such information as company identifiers.

This structure can and must be exploited in relevance and novelty/redundancy calculations. Most work on IR systems for structured documents focuses on segment (as opposed to entire document) retrieval, and techniques such as those for computing relevance for xml documents or for defining term weights within segments are very valuable in real-time news retrieval. An article by Weigel, Schulz, & Meuss [4] is one of many possible references on these topics.

Beyond simple structure, sophisticated relevance calculations should take into account that editorial rules and conventions (writing styles) exist, are tightly enforced for some of these structures, and they should therefore weight terms accordingly. Reading Ellis’ Copy-Editing and Headline Handbook [5] is eye-opening for anyone writing a news retrieval system.

But the world of journalism is not a simple one. New websites are setting new rules, although the story concept remains the same. The top-down nature of news stories also comes into play when computing relevance, except for “Summary,” “Headlines of the Hour” or NIB (news in brief) stories, since they consist of multiple small stories knitted together (the print media equivalent of an RSS file, except often without links to longer stories).
User Queries

User queries in the real-time news environment do not typically consist of a few terms. The “2.4-terms-per-query” rule [6] simply does not hold. Taking a random sample of (roughly) 50,000 saved searches on the NewsEdge system, we find an average of 44 terms per query. This characteristic has profound implications for relevance calculations that use term frequencies (as nearly all do, in some capacity or another). Inspection shows that many of these terms are proper nouns, in particular, company names (or company ticker symbol identifiers), whose presence make perfect sense, since a list of company names is a proxy for an industry or directly represents an investment portfolio or sales territory. In the case of an industry query, it might be argued that a metadata value (an industry category) would be a better (single term) query, but not everyone defines industries the same way, and end users very much prefer the definitive nature of an explicit company list (usually with additional keywords and phrases).

Unlike most OR clauses that a search algorithm would encounter, here the expected number of term matches is often one (of, say, 44). That more terms may match in no way implies greater relevance. In fact too many matches would indicate some type of boring list or table of companies, the kind of story nobody wants to read.

Topic Familiarity

Users in our environment have a high degree of topic familiarity [7] and topic focus. This characteristic dramatically affects the use of relevance criteria. No one in the business world (we are excluding academic business research here) searches business news for fun – only for profit. Give the user high precision but low recall, and he misses opportunities, losing money. Give the user high recall but low precision, and you waste his time. Luckily the topic focus, combined with the structure analysis discussed above, helps with both recall and precision.

Consider a story on a concert to be held at Verizon Center in Washington, D.C., with a box office that accepts Visa, MasterCard and American Express. It is unlikely that a business searcher wants this story returned as a highly ranked hit for a query that included the company symbols NYSE:MA, NYSE:VZ and others. But it has two perfect term matches. Possible saving graces include:

1. It is not a business story.
2. The phrase “Verizon” appears next to “Center,” and that phrase only appears once (so it is most likely not a story on Verizon’s investment in the naming rights, which came along with the acquisition of MCI Worldcom).
3. The term “MasterCard” appears at the bottom of the story.
4. The ticker symbols (NYSE:MA and NYSE:VZ) do not appear in the metadata record that accompany the story.

There are other cautionary notes for search engineers, however. We need to worry about a story that has Verizon offering cell phone deals to MasterCard cardholders – highly relevant – and snippets like “work at Verizon centers on increasing renewal contracts,” which reveal problems with ignoring case and conflating singular and plural versions of terms.

Some IR systems are designed to maximize diversity of results; that is, when a search term is ambiguous they try to return stories covering all possible meanings. The topic focus of business news retrieval systems eliminates this requirement. We can be assured that the user with a query containing “poultry stocks” wants hits on Perdue, not cooking. Recipes would routinely be included in stories retrieved for commodities traders were it not for the recognition of topic focus.

Alternatives for Defining and Calculating Relevance in the Business News Environment

It should be clear from the above discussion that standard, classic, vector-space relevance, which is calculated by the formula “term frequency multiplied by inverse document frequency” (tf*idf), will not be sufficient in and of itself for business news retrieval systems. Even when we take document structure, query length and topic focus into account, more is needed to do a good job for working business news searchers.

First, users this sophisticated understand Boolean logic and will expect all hits to be matches. This expectation is not really at odds with relevance-
based retrieval, since even Salton suggests that Boolean and relevance search can be combined by returning all true Boolean matches, computing the relevance score for each match and then sorting (or flagging so that stories can be kept in reverse chronological order, as discussed above) [8, Chapter 10]. This procedure is routinely carried out in nearly all business news retrieval systems in the commercial marketplace.

But we know now that other measures – not based on closeness to the query – need to be brought to bear. Three main alternatives to classic relevance have surfaced (in addition to using print and web position and placement clues, as discussed above, which we will not pursue further here):

a) Evaluate novelty of results; promote matches (flag or sort to top) that bring new information to light. [9]

b) Derive ranks based on authority, using either author analysis [10] or link analysis of the collection. (For example, see [11] or [12].) Of course this approach includes Google-like inbound link ranking, usually attributed to the founders of Google but priority would appear to belong to Yahong (“Robin”) Li, the founder of Baidu, alone – based on Li’s paper at the 1997 Philadelphia SIGIR workshop [13], the date of Li’s patent filing (1997, 5 February) [14], and the submission date of Li’s IEEE Internet Computing paper [15].

A weaker version of authority is popularity, used on social media sites, measuring “clicks” (in this context, reading the full story after being presented only with the headline and summary – “most read” voting) or simply counting how many readers comment on a story.

c) Attempt to figure out the user “use case” – the context of the search. What is the user going to do with the results? How can the search results help with the “work task”? [16] Elevate those matches that help the most.

We can dismiss (b) almost immediately; it is of no help in this problem since real-time news contains neither the links nor the author attributes necessary to fuel this ranking engine. And if we consider popularity measures, there are two problems. The first is due to the low-latency requirements since ranks need to be computed in real-time. The system simply cannot wait and collect “story read” information from users before ranking the matching stories that just arrived. Second, while such information and more could be collected and used to inform the rankings in subsequent searches, user-behavior privacy issues – vital in a competitive business setting – make any attempt to collect (much less mine) such data ill-advised.

Approach (a) is very likely to fail in this context as well. Novelty analysis depends vitally on thematic isolation. The thinking behind much novelty analysis is that if the theme is new the match provides novel information. This assumption fails badly in financial news, however, since financial events unfold over time in unexpected ways. Here are four (hypothetical) financial news threads – the headlines in each thread are presented in time order, and the key stories (the ones that deserve high rank) are shown in bold.

Company XYZ announces its earnings results will be available on July xx

**Company XYZ announces its earnings results**

Company ABC has begun a clinical trial on Wonder Drug ZZZ

**Company ABC announces successful results from its ZZZ clinical trial**

Company ABC wins FDA approval for ZZZ

Company P and Company Q are rumored to be in merger talks

**Company P and Company Q to merge**

Company P and Company Q announce a merger

Company P and Company Q complete their merger

Company Z will have to file for bankruptcy protection if things don’t improve

**Company Z makes an initial bankruptcy filing**

Company Z bankruptcy hearings proceed

Some interpretation of novelty should indeed be a key component of search-result ranking for real-time business and financial news, but “new theme” is not the answer.

Eliminating the options above leaves us with (c) – rank the stories on how we believe the stories might or will influence the user’s actions – the eventual actions the user takes (or decides not to take) after studying the
search results presented. Van Rijsbergen has put it beautifully – this (abridged by us) quote is worth studying:

The nature of what is wanted by a user is a matter for debate. . . . It is not enough to say that what is wanted is a matching item, matching items may be irrelevant or useless. . . . It is a convenience to conflate relevance with aboutness, but now especially in the context of Web searching it may be necessary to begin separating these. . . . [I]ncreasingly searches are done within a context of performing a task; the nature of the task could have significant effect on what is worth retrieving [6, p. 13].

For example, I need to find the exact address of a hotel in a city foreign to me. Typing the hotel name into a web search engine is usually of little value – it has been assumed (either by the wisdom of crowds or some hidden commercial interest) that I want to make a reservation at the hotel, and often I have to wade through pages of discount hotel booking site URLs to find the hotel’s actual website. This mistake is one of “use.”

And this is our suggestion: Presume a particular action response on the part of our business/finance news searcher. We call this technique PAR searching – Presumed Action Response.

The key observation here is that the world that invests in business is a parallel, shadow world to the one that operates and cares for business. A person in business searching for news on Company Big Widgets might care about the following:
- competing with Big Widgets
- selling to Big Widgets
- buying from Big Widgets
- investigating the executives who work for Big Widgets
- getting a job with Big Widgets
- suing Big Widgets
- determining how Big Widgets is doing financially (since that affects all the above)

or something similar. A person investing in business searching for Company Big Widgets cares about the exact same things, since they expose the levers that determine whether investments in Big Widgets will make or lose money.

For publicly traded companies the stock market provides insight retrospectively as to whether a news story on a particular company “moved the market” for that company’s stocks or bonds – which means the news story in question should have been highly ranked if the searcher had issued a search on that company name just as the story was received by the system.

Therefore, using historical data analysis, we should be able to determine those features in a story (on a public company) that cause it to “move the market.” We are not suggesting this analysis is an easy undertaking – there are a number of issues and problems with such studies, touched on below. But we do claim that it is “Job One” for developing a search-result ranking algorithm for real-time business and financial news.

Having found these market-moving, news story features – words and phrases – in public company stories, we can extend the analysis ipso facto to stories on non-public companies and more generally to non-company-centric business stories, since these stories will contain similar features (using standard IR similarity formulas but paying attention to the structure issues discussed above).

We define the impact of a story as some quantitative scoring that measures the presence or absence of such features. The exact numerical scales used and other details are not important for this paper – what is important is this: In the domain of real-time business and financial news, we believe impact should form the basis of result ranking.

Observe that classic tf*idf calculations still play a key role, but indirectly – they are used in similarity measurements for stories that cannot be related to movement in public markets.

(Note: Government economic-condition news releases move interest rate and foreign exchange markets in ways similar to the movements produced by breaking stories on public companies. Our methods thus extend to them as well, although we do not concern ourselves with this type of story further in this paper.)

The concept of “moving the market” is not trivial. It does not necessarily mean a stock price has changed. Often one finds that, in reaction to a major story, there is wild trading, but the price moves very little on large volume. In these cases, volatility is the right measure (and we look for changes in
To assess impact we carefully measure over time the actual changes in markets when stories break, using time-series statistics, and isolate the features involved. Here are some of the factors we weigh for news on public companies:

- Is the timing of the news event known in advance (earnings reports), or is the fact that there is news itself a surprise, representing “true novelty” (patent infringement case, for instance)?
- How close are we to the next earnings announcement date?
- How close are we to the next options expiration date?
- Has volatility decreased across the time of the news event since uncertainty has been reduced, or the opposite?
- Has volume dried up ahead of the news release since uncertainty has increased the real (unobserved) bid-ask spread to the point where there are no trades to be done?

Having in this way assessed (actual) impact for a large number of news stories on public companies, we then run a standard categorizer (a separating hyper-plane algorithm or a support vector machine), eventually building a linguistic feature set, a set of linear rules and a threshold computation that can be applied in real-time to any story (whether on a public company or not). Distance (on either side) from the threshold can be turned into a numeric measure taking values from 0 to 1 of (predictive) impact and be used to rank stories.

The method appears to work well. Here is a screen shot from our test bench.

Predicted impact is shown as a number from 1 (low) to 9 (high) – just a re-scale and rounding of the actual continuous value. The query was: Biotech Earnings. Note that this is not a query on a public company name.

You can see that new earnings reports got high impact scores (mostly 9s, one 7), a “reiterate” research report got a 6, a research report with little to say got a 2 and (most impressively) an announcement of an earnings conference call – an “announcement about an announcement” – was marked down to an impact ranking of 1.

As long as our assumption on end-user actions is correct, we believe this PAR approach, using our impact measure, produces meaningful rankings of Boolean-match results for working users of real-time business and financial news search systems.

PAR can be viewed another way. Since investors do actually read news stories and trade securities on the basis of what they read, if we have accurately captured their actions (in the sense that we can determine that a market movement resulted from the issuance of a news story), then impact is in fact relevance feedback from an informed community. It is not direct feedback, but a proxy. Hence the other meaning of PAR: Proxy Activated Relevance-feedback. It is usually easy to focus on a graphic rather than a number, so in our commercial NewsEdge products we are introducing impact as a bar graph:

Note the IBM story with no impact. We are going to allow both a sort on impact and a filter (show no stories below an impact threshold).

We believe we have done our job. We have made PAR.

Acknowledgements
The author wishes to thank the members of the NewsEdge team at Acquire Media, and in particular Kristen Carney, Zhi Chen, Boris Guralnik, Dan O’Connor and Nancy Yee (alphabetical order). Thanks are due as well to Irene Travis, editor of the Bulletin, for suggestions on the first draft that substantially improved the exposition.
Resources Mentioned in the Article


Integration of Taxonomy and Keyword Searches: A Comparison of Two Implementations

by Ronald P. Millett

Subject taxonomies have a long history of usefulness in indexing libraries of documents or records. One of the weaknesses of this approach is the difficulty of traversing the taxonomy hierarchy to find the subject term. Full text indexing of keywords enables automatic discovery of records containing terms ordered by relevance including phrase, title text, word proximity, emphasized text, linguistic and other criteria. Major weaknesses of the keyword approach are the lack of the ability to use controlled subject terms and result hits that are false-positive. This paper compares two implementations that combine taxonomy and keyword approaches in an attempt to address these weaknesses for efficiency, ease of use, recall, precision and hit relevance.

NICEM Database Search Implementations

These two search implementations combine taxonomy and keyword searches:

1. A system based on Access Innovation’s Search Harmony product, using their subject taxonomy and user interface with the Perfect Search search engine.

2. A system that is an earlier Access Innovation’s implementation using the Lucene open-source search engine with the same subject taxonomy.

The comparison search data set for both systems uses about 500,000 records from summaries of multimedia videos or presentations in the National Information Center for Educational Media (NICEM) database stored as separate XML files. The subject taxonomy terms were specified with a XML subject category tag. Other important tags include title, abstract and publication date. Table 1 summarizes the sample database and taxonomy information.

<table>
<thead>
<tr>
<th>TABLE 1. NICEM test database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records (in separate XML files)</td>
</tr>
<tr>
<td>Average record size</td>
</tr>
<tr>
<td>Database size</td>
</tr>
<tr>
<td>Number of subject taxonomy terms</td>
</tr>
<tr>
<td>Average subject terms per record</td>
</tr>
</tbody>
</table>

The database was indexed by both systems on similar 64 bit Windows servers, with from 3 to 16 GB of RAM. The indexing process was single threaded even though there were several cores available. Table 2 summarizes indexing and retrieval information for each system. The Lucene implementation used the Java version.

| TABLE 2. Lucene and Perfect Search Indexing and Retrieval Information |
|--------------------------|----------------------|
|                         | Lucene               | Perfect Search       |
| Indexing Time            | 161 minutes          | 13 minutes           |
| Index Size               | 168 MB               | 440 MB (main index) + 192 MB search accelerators |
| Threading                | Single               | Single               |
| Typical Engine Retrieval Time | Not Available   | 1,200 - 7,000 queries per second |

Ronald P. Millett is chief scientist at Perfect Search Corporation, 1145 South 800 East, Suite 325, Orem, Utah. He can be reached by email at ron.millett<at>perfectsearchcorp.com or through the company’s website at www.PerfectSearchCorp.com
Retrieval Speed vs. Indexing Cost

One of the apparent anomalies of the Perfect Search system is that the typical inverse relationship between retrieval speed and indexing cost is missing. Faster retrieval speed usually implies that a slow and expensive indexing process needs to be executed. However, Perfect Search’s approach uses an indexing pipeline that creates temporary files and utilizes indexing pattern speedups at every point of the indexing process to improve the indexing functionality. Note that even though the Perfect Search index and accelerators used in retrieval are much bigger than Lucene’s index, these files were generated much faster than the Lucene index. This indexing and retrieval architecture results in both impressive indexing speeds and retrieval speeds for the same system.

Once indexing speeds are a small multiple of file copying speeds, as Perfect Search’s data for this test illustrates, the whole character of creating and using indexes can change. Formerly expensive re-index operations are no longer a problem, if needed. Creation of incremental index files can become almost instantaneous and merging together many small index files can also be an easy part of making a very responsive indexing process to support the search of dynamic data.

Document caching and acceleration files are also generated and included in the numbers in the table for the Perfect Search-based system. The 192 MB acceleration files were generated for the test system in under a minute. On a large system such as the installation for WorldVitalRecords.com, the 149 GB of indexes require acceleration files of only 13 GB (9% of index). 440 MB is not considered a large index for a Perfect Search-based system and the acceleration files are a much bigger percentage of the total index size (192 MB, 43%) than they would be for a larger search system.

These acceleration files play a key role in enabling very fast retrieval speeds for the system. Even when a single query can actually trigger as many as 20 sub-searches for different relevance rankings, a typical retrieval speed of 1,200 to 7,000 queries per second is truly impressive for each complete search. These speeds are for a system where both accelerators and index files have been cached.

Autocompletion of Subject Taxonomy Terms

Using precategorized subject terms, searches can avoid irrelevant hits and find records or documents that directly pertain to the subject category selected. An important weakness of taxonomies is the difficulty of finding the descriptor that the user has in mind if it has to be traversed from the top of a hierarchy. Both test systems use the Access Innovations autocompletion module that uses a “key word in context” (KWIC) [3] list and term matching to allow the user to find any term that he has in mind that might be part of a subject description. Figure 1 shows a sample screen shot of this powerful taxonomy autocompletion window.

The taxonomy tree structure is still very important. It shows subject descriptors in parent/child/sibling set relationships with other descriptors. A disadvantage of this approach can result when a pure binary tree or single top node results in too many operations to find the needed term like the “20 questions” game. Access Innovations uses a forest of taxonomies approach with typically non-binary child divisions that allows for rapid traversal of
Precision and Recall Comparisons

Except for one characteristic of the Lucene-based system that lowered precision to 99%, the two systems both receive 100% precision and recall values for the taxonomy subject descriptors. When the Perfect Search stemming module and its effects are added, the recall of the Lucene-based system went down. The Perfect Search algorithmic stemmer was too broad in some cases, which reduced precision for a few specific terms.

Table 3 compares the Lucene and Perfect Search hits for the text “communications” in descriptor and keyword modes. This search was selected because it shows some interesting precision and recall differences.

Note that because the Lucene implementation can get a hit for “communications” as a sub-term of subject taxonomy descriptors such as “communications industry” and “satellite communications,” Lucene has a higher retrieval count on hits for this descriptor entry than Perfect Search. Because the actual descriptor “communications” was not in some hits, the precision value for this search for the Lucene implementation is 99%.

Note also that the Perfect Search system uses a stemming subsystem to include terms like “communication,” “communicate” and “communicating” as lower relevance hits for a keyword search for “communications.” The current implementation uses an algorithmic stemming module that also included hits for “community,” “communities,” “communist” and “communism” as possible word forms for “communications.” This algorithmic stemmer created false positive lower relevance hits. A dictionary-based stemming module being implemented for the next version of Perfect Search will avoid this problem and not include the extraneous forms.

When correctly stemmed terms are included for this keyword search, Lucene then has its recall percentage for “communications” go down to 22% while Perfect Search maintains a 100% recall percentage. When the false positive stems are counted, Perfect Search has its precision percentage go down to 70%. Lucene does have a similar algorithmic stemming module available, but it was not used in this implementation [4, pp. 282-284].

| TABLE 3. Hit analysis for “communications” subject descriptor and keyword searches |
|---------------------------------|-----------------|-----------------|
|                                  | Lucene          | Perfect Search  |
| Taxonomy descriptor hits        | 2,760           | 2,738           |
| Hits from “communications” sub-terms “communications industry” “satellite communications” | 22               | 0               |
| Recall % for exact descriptor hits | 100%            | 100%            |
| Precision % for exact descriptor hits | 99%             | 100%            |
| Keyword hits                    | 3,678           | 23,960          |
| Keyword hits with correctly stemmed terms | 0               | 13,018          |
| Keyword hits with stemming errors | 0               | 7,264           |
| Precision for exact keyword hits | 100%            | 100%            |
| Precision including stemmed terms | N/A             | 70%             |
| Recall for keyword hits         | 100%            | 100%            |
| Recall including correctly stemmed terms | 22%             | 100%            |
Relevance Algorithms

The descriptors receive a relevance score of 100 for the Search Harmony Perfect Search-based system and are ordered according to the order they were indexed. Keyword searches can have some false positive low relevance hits due to the stemming algorithm currently in use as previously mentioned. Because of the ambiguity inherent in all languages, the use of a word in a search may not match the query’s intended meaning. We did not evaluate that factor in our search precision numbers. Having exact phrase and near proximity in compact fields like the title helps to have more relevant hits and avoid false positive hits.

Combinations of the title, abstract and other fields plus term proximity (exact phrase, near content word proximity) and term linguistic stemming provide relevance scoring of hits from scores of 100 to 50. These orderings and values can be customized according to the users’ preferences using XML configuration file information.

The Lucene-based system also has calculated relevance for keyword searches. Boosts are available for low frequency terms, special fields and normalized term weights [4, pp.78-81] [5].

Mixing and Matching Taxonomy and Keyword Terms

The Lucene-based search system has two modes of searching: preferred terms (taxonomy subject categories) and keywords. The two modes cannot be mixed in the same query. The preferred terms are matched using the autocompletion typing window.

The Perfect Search-based search system allows the user to mix and match preferred taxonomy descriptors and keywords and use a similar syntax to Lucene. Each subject taxonomy descriptor is indexed as a single complex term that is not divisible (e.g., SC:“children’s literature”). New searches and refinement of existing searches can select from the taxonomy terms or the keyword terms.

Figures 3 and 4 illustrate multiple search terms using taxonomy descriptors and keywords in the same search using the Search Harmony Perfect Search-based system. Tables 4 and 5 show the hit frequency for each search term and each stage of the search intersection process.

If, for example, a single term retrieves 1000 hits, the user cannot and will not in normal practice look through the hits to find the right hit. Typically, unless the hit is in the first few pages of results, it might as well be invisible as far as the user is concerned. Precision and recall numbers for large sets might as well not exist when the set cannot reasonably be examined. The power of having both carefully defined descriptors and full text terms in the same query results in a smaller search result set that can easily be viewed and examined.

In Table 4 and Figure 3 we see a search where, even after two subject-descriptor terms were entered, there were still over 1,000 hits in the results list. The keywords “lion” and “witch” also occurred in many database entries,
but after the descriptors filtered the set down to 1,078, the two terms resulted in a short list that contained two copies of an audio book of the C. S. Lewis’ work *The Lion, The Witch and the Wardrobe*.

Table 5 and Figure 4 illustrate another search where a single category descriptor “business planning” was combined with keywords “computers” and “forecasting.”

More Relevant Results through Tapping Multiple Dimensions of Search

The key to success in a search system is to help users quickly find the data they are looking for. The ambiguity of language, syntactic and UI details and the amount of data that a search system has to winnow out to find key results for the users are formidable obstacles for any system. The advantages of including taxonomy-based subject descriptor searches and keyword searches in the same system were emphasized by searches on both test systems. The lack of being able to mix and match descriptor and keyword searches in the Lucene implementation made it much more difficult to retrieve a result list that can easily be examined by the user.

The Search Harmony system combines several search methodologies in one unified system, including mixing and matching descriptor and keyword search terms. These methodologies include navigational trees, subject descriptor autocompletion, search within results and relevance-based full text searches including fuzzy search features such as stemming. Combining all of these features in a single interface, each as it were in a different search dimension, provides an important laboratory for improving search results and finding that “Voila!” best result.

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Frequency</th>
<th>Cumulative Hit Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject descriptor: business planning</td>
<td>516</td>
<td>516</td>
</tr>
<tr>
<td>Keyword: computers</td>
<td>12,638</td>
<td>37</td>
</tr>
<tr>
<td>Keyword: forecasting</td>
<td>384</td>
<td>2</td>
</tr>
</tbody>
</table>

I T O P O F A R T I C L E

C O N T E N T S

N E X T PA G E >

N E X T A R T I C L E >

M I L L E T T , c o n t i n u e d

Resources Mentioned in the Article


Semantic Search
by Darrell W. Gunter

Over the last four decades steady improvements have been made in search engines, including those that use Boolean search interfaces as well as those that rely on other ways of matching query terms to words in the documents searched. In this paper we consider the further gains that can be achieved with a technique called semantic search.

Wikipedia defines semantics as the study of meaning [1]. The word semantics itself denotes a range of ideas, from the popular to the highly technical. It is often used in ordinary language to denote a problem of understanding that comes down to word selection or connotation. Semantic search is a process used to improve online searching by using data from semantic networks [2] to disambiguate queries and web text in order to generate more relevant results.

For example, Collexis [3] utilizes semantic technology to create their proprietary Fingerprints. A Fingerprint represents a visualization of the aggregation of the key concepts of the aggregated dataset – presented in a bar chart of relevant weighted concepts. For example, if a researcher has found an article about diabetes in the New York Times and wants to know if there is any research in PubMed [4] that matches this article, she can paste the article into the Collexis engine, and it will create a Fingerprint of the article. Further it will return a result set of articles from the PubMed database that can match any portion of the article. In addition to creating Fingerprints and providing matching articles based on a comparison of the relevant concepts, this semantic technology also allows the researcher to profile experts, determine the research trend of key concepts and conduct hypothesis generation. Further, she is able to build on her search by adding or deleting concepts based on the direction of her research.

Semantic search as described above has many benefits over traditional search. It allows searchers the following capabilities:

- To refine their search by adjusting the settings of the Fingerprint
- To combine or remove a concept(s)
- To create dashboards based on the key relevant concepts

Imagine a search in Google about juvenile diabetes. The result set will contain thousands of items. In order to know the content of all retrieved resources the researcher would have to read them. With the Collexis technology the result set is presented in an aggregated Fingerprint that consists of the weighted bar chart of key relevant concepts. The researcher is able to dive into any concept to determine the articles that are associated with the concept. In a nutshell, the semantic technology allows search based on concepts in context.

These features allow researchers to be more proficient and efficient in their searches. You may know that Microsoft has launched “Bing,” its first semantic entry in the global search space. Bing’s technology came from Microsoft’s acquisition of Powerset last year. Bing has begun to get very good reviews, and I would suspect that Google will not be far behind with a semantic search engine.

Let’s look at a couple of case studies that demonstrate the adaptability and power of this approach. The first case study is from Johns Hopkins University, where they faced a fundamental problem. With hundreds of researchers scattered throughout their institution it was very difficult for the research community to determine the best person to collaborate on a very specific research project. To help with this problem Johns Hopkins
Leadership built a centrally located coffee shop – hoping that the research community would get to know their fellow researchers while getting a cup of joe. While everyone enjoyed a great cup of coffee or latte, folks were not making the necessary connections for their research. The university turned to Collexis to help with the expertise identification problem. Collexis was able to aggregate the research profiles of the entire research community and create an institutional dashboard. This institutional dashboard allowed them to immediately find the expert on any biomedical topic or category by searching based on key concepts. Johns Hopkins was so excited about the institutional dashboard that they opened it up to the world via the Internet. Now its research community can easily identify potential collaborators in a matter of seconds. Its pool of collaborators is also visible to everyone. The coffee shop is still very active, but the new collaborators meet there to discuss research ideas over a cup of joe. It is important to note that this case study was replicated at many top research institutions such as the NIH, the University of Michigan, and the University of Miami to name a few.

The second case study that I would like to discuss is how the American Association of Cancer Research utilizes the Collexis Reviewer Finder to enhance the editor’s capability to determine the best pool of peer reviewers. The demands on editors’ time are quite significant, and one of their most important duties is to manage the peer review process for their respective journals. Managing the hundreds, if not thousands, of submitted manuscripts is a daunting task. The editors must understand the specifics of the manuscript and then determine who within their peer review pool is best qualified to review it. Today this task is usually managed in a very ad hoc manner. Editors must use a plethora of tools to achieve their objective of getting the manuscripts peer reviewed in a timely and professional manner. To help with this productivity issue, Collexis has developed and launched the Collexis Reviewer Finder application.

Utilizing the Collexis proprietary Fingerprinting technology, the Reviewer Finder applications allow the editor to first Fingerprint the submitted manuscript and then compare it against the 1.8 million research profiles in the BiomedExperts.com database. The Reviewer Finder instantly provides the editor with a list of potential peer reviewers based on the specific research profile and how the profile dovetails with the submitted manuscript. Further, the Reviewer Finder will also detail if anyone in the peer-reviewer pool has a conflict of interest with the author of the submitted manuscript. A conflict of interest is noted if the potential reviewer has either co-authored a paper/grant or has worked in the same location/institution with the submitting author. The Collexis Reviewer Finder provides editors with an invaluable tool that saves them considerable time and improves their productivity greatly.

Semantic search makes possible the recreation of the two key features in these case studies – aggregating and visualizing a large dataset and conducting a conceptual search. Standard search tools do not support these tasks in a very precise or timely way. While they will continue to be used, it is time for new technology to step forward that can increase the researcher productivity. I am confident that this new technology is semantic search. By this time next year we will see a flood of exciting applications based on semantic technology. For more information about the Semantic Web visit the Project10x website [5] and download the executive summary for Semantic Wave Report: Industry Roadmap to Web 3.0 and Multibillion Dollar Market Opportunities [6].

Resources Mentioned in the Article

Most ASIS&T members probably know that at any Annual Meeting, Canadians make up the largest group of international attendees. With the increasingly international orientation of ASIS&T, the Canadian presence is growing: in the past five years new student chapters have opened in Vancouver and Toronto, and in 2007 ASIS&T had its first “cross-border” president, Edie Rasmussen. In 2009, the Canadian members are very excited to welcome ASIS&T members to the Annual Meeting in Vancouver, on Canada’s beautiful west coast.

What many ASIS&T members may not know is that for almost 40 years, Canadians have had a separate information science association of their own. The Canadian Association for Information Science/L’Association Canadienne des Sciences de L’Information (CAIS-ACSI) was formally founded in 1971 “to promote the advancement of information science in Canada and encourage and facilitate the exchange of information relating to the use, access, retrieval, organization, management and dissemination of information.” [1]. The origins of the association are interesting, as it formed as an offshoot of the much older American Documentation Institute (ADI) in the same period that the ADI underwent the name change to the American Society for Information Science (ASIS). The story, as reconstructed by Kirsti Nilsen [2] from CAIS-ACSI archival material, is that in 1969, a group of Canadian ASIS members met in Ottawa with the intent to form a local ASIS chapter but instead ended up proposing, and eventually founding, a separate Canadian association. Like ASIS&T, early membership in CAIS-ACSI was made up of computer scientists and information professionals from government and industry but gradually shifted to include greater numbers of academics and researchers.

Today, CAIS-ACSI is a single national association, run by a seven-person executive that is largely drawn from the academic faculty of the eight Canadian LIS programs. The current president, Kate Johnson, and vice-president, Nadia Caidi, are active members of both ASIS&T and CAIS-ACSI. This dual membership is true of many CAIS-ACSI members. Given the geographical span of LIS institutions in Canada – from the west coast (University of British Columbia, University of Alberta) to the central region (University of Western Ontario, University of Toronto, Université de Montréal, McGill University, University of Ottawa) to the east coast (Dalhousie University), CAIS-ACSI serves an important role in connecting and representing Canadian LIS researchers and practitioners from all regions of the country.

### CAIS-ACSI Activities

CAIS-ACSI achieves its goal of advancing information science in Canada through its journal, the *Canadian Journal of Information and Library Science (CJILS)* [3], and its annual conference, both of which are open to international participation. Like the *Journal of the American Society for Information Science and Technology (JASIST)*, *CJILS* has a broad scope, publishing articles on a wide range of topics in information science and library science, as well as reviews of books, software and technology. *CJILS* also serves as a venue for French language publishing in information science, as it publishes both English and French language material and includes abstracts in both Canadian official languages. The journal was founded in 1970 and became a peer-reviewed journal in 1986 under the editorship of Charles Meadow.
(another well-known ASIS&T member). CJILS is published by the University of Toronto Press and is widely indexed. A subscription to the journal includes a membership in CAIS-ACSI and can be had for an economical $49 (students) and $75 (individuals).

The CAIS-ACSI conference takes place every year in late May or early June, which makes it an excellent complement to the ASIS&T meeting each fall. In comparison to ASIS&T meetings, CAIS-ACSI meetings are smaller and more intimate affairs, although they have the same focus on the communication of current research and networking. Typically, there will be under 100 attendees in total, from Canada, the US and farther afield, and between 30-50 papers presented in one or two tracks over three days. Given the size of the meeting and its casual atmosphere, it is easy to get to know other participants, to learn about and discuss diverse research areas and to make new professional connections. It offers a warm and welcoming opportunity for graduate students and new researchers to present their work and for established scholars to test out new ideas.

Over the years, the CAIS-ACSI conference has been held in cities across Canada, usually in conjunction with the Congress of the Humanities and Social Sciences [4], a large-scale, multidisciplinary meta-conference, which attracts scholars, researchers and students from around the world. So while the CAIS-ACSI meeting itself is small, it takes place in the context of a much larger “intellectual festival” and offers the best of both worlds. The June 2010 conference will be held at Concordia University in Montreal, Quebec, and we encourage the ASIS&T community to consider taking part. Top conference papers are considered for publication in CJILS and the remaining papers are published in the online proceedings on the CAIS-ACSI website [5]. There are currently over 500 CAIS-ACSI conference papers and abstracts available in this collection, going back to 1993. The conference also features a student paper award, which offers students a great opportunity to travel and attend the conference and see his or her work published in CJILS.

Given the close ties between CAIS-ACSI and ASIS&T through shared origins, goals and overlapping membership, it is worthwhile thinking about how these two organizations can work together in the future. CAIS-ACSI plays an important dual role in the Canadian context, as an arena in which to examine issues of particular interest to Canadians and a means of connecting the Canadian information science community to broader international perspectives and issues. Opportunities to build ties between CAIS-ACSI and ASIS&T through shared projects, events and initiatives would serve the interests of both organizations and strengthen the position of information science as a profession and discipline in North America.

---

**Resources Mentioned in the Article**

Hasta Luego
by Stacy Surla, outgoing associate editor for information architecture

When I took over the job of editing this IA column four years ago, my selfish goal was to learn more about what interests me most, directly from the people whose work I most admire. I’m an inquisitive dilettante (as Heinrich Zimmer defines that notion) and proud of it. As one who takes delight in information architecture’s obscure and thrilling connectivities, the column gave me a wonderful opportunity to commission a diverse collection of articles from authors and practitioners who are exploring the most thought-provoking domains in user experience.

I will no longer be the one petitioning these explorers for their articles, their trip reports from a still largely undiscovered country. Instead, to my deep pleasure, writer, teacher and speaker Thom Haller will carry on this work. Thom is himself a noted explorer, one who also takes delight in the field of IA, and a deeply gifted, Virgil-like guide who excels in introducing others to the venerable, yet future-facing, practice of making the complex clear.

Good night, then, and a thank you to everyone who enabled me to make this adventure – including Andrew Dillon who passed the scroll to me; Irene Travis who was a warm and ever-tolerant editor-in-chief; Dick Hill who is unfailingly supportive of all my ambitions; and the authors who gave so generously. See you later.

Stacy Surla served as the Bulletin’s associate editor for IA. She serves on the IA Institute Board of Directors and is a past chair of the IA Summit. She can be reached at stacy<at>greenfx.net.

Introduction from Thom
by Thom Haller, incoming associate editor for information architecture

I often think of myself as the man who is most likely to get lost in information. I have trouble finding my way across sentences, I expect my paragraphs to be well structured, I want document clarity and visible structure, I expect systems to be consistent and I anticipate good user experience.

Sometimes I’m disappointed. OK, often I’m disappointed. I become frustrated and grumpy. And I sometimes envy those of us who move through turgid prose and improbable structure without batting an eye. I am not among them. I stand with the humans who cannot understand the kiosk interface and cry out for plain language in government documents. Lots of challenges, but with them, great possibility. Obviously, we have a job to do.

How I Landed Here

I stepped into the job of making the complex clear 15 years ago. I recall standing at the top of an escalator at a communications conference in 1995. “I’m interested in information structure and visuals,” I told my colleague. “But I’m also interested in how people process information and what they do with information. I think I’m a data stylist.”

“No,” he responded. “You’re an information architect.”

“Sounds cool,” I thought.

My colleague directed me to books by Richard Saul Wurman who coined the term. “These are ideas I can use for improving my PRINT documents,” I squealed (oblivious to the web world forming at the same time). But by 1999 business had embraced the web, Rosenfeld and Morville had released the Polar Bear Book and I was teaching an information architecture course – one more theoretical than its current day descendent.

It’s as a teacher that I step into this column. As with my classes, I anticipate I will find I’m more frequently a learner than a pontificator. I hope I’ve learned from Andrew and Stacy before me and

Thom Haller, teacher, speaker, writer and user advocate, teaches principles of performance-based information architecture and usability. Since 1998, Thom has taught classes on architecting usable web/Intranet sites. As a teacher, Thom enables students to structure information so people can find it, use it and appreciate the experience. He can be reached at thom<at>thomhaller.com
that I can use this space to bring you current conversations in the IA and UX communities.

**Passion for Foundation**

Once, when chatting with Peter Meholtz, president/founding partner at Adaptive Path, Peter related how he liked that the field provided him with different speaking topics. “Really?” I replied. “I like how the field provides me with the same topics – it’s only the details that keep changing.” I long ago condensed the field into what I call eight Ps of information architecture:

1. **Performance** – What do users need or want to do in this information space? How do we measure successful performance?
2. **Perspective** – What is our focus? Are we focusing on who’s using the communication product, what they want from the product and their context?
3. **Product** – What technological choices best suit the needs of the users of this information?
4. **Presentation** – How can we visually structure information so people can accomplish what they want to accomplish (without becoming thwarted)?
5. **People** – Who is involved in this endeavor? Who will provide content and expertise? How do we build beneficial relationships?
6. **Politics** – How do conceptual barriers and organizational context keep us from meeting the needs of the site’s users?
7. **Process** – What is the framework we follow to develop this communication product?
8. **Possibility** – What are the ways we can respect and help users, while maintaining our focus on a site/product’s mission?

These questions have guided me for 11 years. They welcome differing opinions, and my job is to bring opinions, stories and synthesized research to you.

**I Need Your Help for the Next Issue**

Throughout my career, I have been fascinated by the idea of improving human (and organizational) performance – following the idea, “How can information structure help people do their jobs better?” Years ago, engineer Thomas Gilbert laid out the foundation for human performance thinking in his book *Human Competence*. He proposed “[i]mproving the information used by workers raises their measured performance substantially – never less than a 20% improvement, and sometimes as high as 600%.”

I have yet to see the studies that show me 600% improvement. But I believe the work we do is directed at this size organizational gains. I’d like to know what you think. What studies, summaries, stories and findings show the relationship between good architecture and accomplishment? I need your help for the next issue. Feel free to comment on any of my foundational questions – or let me know which ones I’ve missed.

Write me at thom<at>thomhaller.com.

**P.S.** As I step into this column I want to thank Stacy Surla for all her support and thank my frequent colleague Deborah Aker for supporting me with editorial counsel.
I chaired the committee that developed the NISO Guidelines for the Construction, Format, and Management of Monolingual Thesauri [1], which was published in 1994 and reaffirmed in 1998. I participated in the NISO Workshop on Electronic Thesauri, held in 1999, at which it was decided to update the guidelines, but I was not part of the revision committee that produced the Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies [2], issued (only in electronic form) in 2005. That standard should soon be up for review and public comment; in this article I share my thoughts on the need for revision of the standard.

Precoordination

At the initial meeting of the committee that developed the 1993 thesaurus standard, I recall asking, “How are we going to distinguish thesauri from subject heading lists?” The consensus was that thesauri are designed for postcoordination (combination of terms at the searching stage) and subject heading lists for precoordination (combination of terms at the indexing stage). Late in the comment stage, someone asked why the draft standard provided no guidance on subheadings to be used with descriptors, and I replied, “Because they are out of scope for a standard on thesauri.”

Given that the 2005 standard substituted the term controlled vocabularies for thesauri in the title, I expected to find guidance on subheadings in it. Therefore, I was stunned to read, “The rules for developing precoordinated indexing terms are generally outside of the scope of this standard. These guidelines are available in sources such as the Library of Congress Subject Cataloging Manual: Subject Headings . . .” (p. 37). Nothing was thus accomplished by broadening the scope of the standard if it provides no guidance on subheadings.

My impression is that a global change was done to convert the term thesaurus to controlled vocabulary throughout the standard. This process resulted in the following ludicrous sentence in the definition of “subject heading”: “Precoordination of terms for multiple and related concepts is a characteristic of subject headings that distinguishes them from controlled vocabulary terms” (p. 9). (The term controlled vocabularies includes subject heading lists.) Another example: “Each term included in a controlled vocabulary should represent a single concept . . .” (p. 23). That is a requirement of thesauri, not subject heading lists, which often include conjoined terms, for example, Banks and Banking. As the previous quotation notes, subject headings precoordinate “multiple and related concepts.”

Definitions

The 2005 Guidelines have eight pages of definitions close to the beginning of the standard (section 4) and a more extensive glossary in an (unnumbered) appendix. I much prefer the structure of the 1993 standard, which defined terms at the point of their first use in the text, and collected the definitions in a glossary that was treated as part of the text.

The 2005 Guidelines have inconsistencies between the definitions and the text. For example, entry term is defined as
“The non-preferred term in a cross reference . . .” (p. 5). The guideline on place names, however, says, “The form most familiar to the users of the controlled vocabulary should be designated as the entry term [i.e., the preferred term], and cross-references should be provided from the variants” (p. 33).

The replacement of descriptor (the word for the preferred term in a thesaurus) by term in this standard led to many awkward expressions, such as “entry terms and terms” (p. 63). The definitions section has a separate entry for preferred term, but that expression is not used consistently instead of descriptor. The definition of indexing term states, “Terms, subject headings, and heading-subheading combinations are examples . . .” and concludes, “Also called descriptor” (p. 6). The index says, “descriptors see terms” (p. 169).

In the definition of controlled vocabulary, the note “This rule does not apply to synonym rings” was erroneously placed after the rule for qualifiers of ambiguous terms rather than the rule for selecting a preferred term (p. 5). This observation is confirmed by the correct placement of this note under the rule for preferred terms in the body of the standard (p. 13). The exclusionary note for synonym rings is missing in the section on the purpose of controlled vocabularies, which specifies that they should “Provide consistent and clear hierarchies” (p. 11). Hierarchy is conspicuously absent from the abstract of the standard.

The 2005 Guidelines refer from quasi-synonym to near-synonym (p. 8). The definition previously given for quasi-synonym was assigned to generic posting, with the example “furniture UF beds” (p. 6). I like the expression quasi-synonym for this case, where the terms are treated as equivalent, although they are not even near-synonyms. Longman Advanced American Dictionary explains quasi-as “acting in some way like” [3, p. 1177].

The 2005 Guidelines treat candidate term and provisional term as synonyms (p. 4), in contrast to the 1993 Guidelines, which had separate definitions for them. Yet chart B.1 in the 2005 edition (p. 135) has separate entries for “Candidate terms” and “Provisional terms.”

Not all technical terms have definitions in the standard. An example is “descriptive metadata” (p. 19). Librarians distinguish “descriptive cataloging” from “subject cataloging,” but “descriptive metadata” includes subject headings. Typographical errors in definitions are discussed below in the section entitled “Editorial Flaws.”

Illustrations

The 1993 standard placed illustrations from published thesauri in an appendix because some formatting guidelines of the standard may not have been followed in these thesauri. The 2005 Guidelines embed small excerpts from published thesauri into the text. Many of these illustrations violate the rule that terms – except for proper names – should be lowercased. Even the original illustration provided to illustrate ambiguity (p. 13) capitalizes “Mercury (metal),” although lowercasing serves to distinguish common nouns from proper names. The metal is lowercased on page 21. The parenthetical qualifiers for the four meanings of mercury are not in alphabetical order and are in different sequences on pages 1 and 13.

The illustration for the section of the standard that defines thesaurus (5.4.4) is a single descriptor record from the DTIC [Defense Technical Information Center] Thesaurus, with terms in all uppercase and no RT (related term) reference in the sample record.

Some of the annotations to illustrations are incorrect. For example, the explanation of Example 185, the record for “Ferula” in Medical Subject Headings (MeSH), says, “This history notes (sic) indicates that from 1986-1990 the term was not authorized . . .” (p. 98). That is not true. In MeSH, the note
“was seen under PLANTS, TOXIC 1986-90” refers to a period when there were two categories of subject headings: those authorized for use in the print index as well as online and those authorized only for use in the online database. “Ferula” was in the latter category, and users of the print index were referred to a broader term. The distinction was abolished in 1991.

The illustration of a pick list includes the term “Miscellaneous” (p. 78). This selection is a poor model for the design of a category structure. If it can be justified at all, the heading Miscellaneous can be assigned only after all other headings have been examined, but the screen shot shows only a partial array – from Manuscripts to Vessels.

Another poor illustration is given for Top Term Structure (p. 71): a top term with only a single level of narrower terms, all aligned. The essence of a top term structure is that it shows multiple levels of hierarchy.

Most serious is the misidentification of the illustration for the multilevel display. The Guidelines include an example produced with Synaptica software (p. 69), which can generate a top term structure, but not a multilevel display. The latter gives all levels of broader and narrower terms for each descriptor, often coded BT1, BT2; NT1, NT2. There are small excerpts from multilevel thesauri in the 2005 Guidelines, for example, on p. 19, an “Online thesaurus entry” from AGROVOC, and on p. 85, an excerpt from the UNESCO Thesaurus used to illustrate a multilingual vocabulary. Both of these illustrations feature only two levels of broader terms, and no narrower terms.

Examples

The preceding section focuses on examples excerpted from published controlled vocabularies. In this section, I discuss examples provided by the standard’s authors, without attribution to a published vocabulary.

The standard includes poor examples of precoordinated terms, for example, United States – History – Civil War, 1861-1865 (p. 37). Jessica Milstead’s dissertation pointed out the sorting problem caused by such headings [4, pp. 174-175], which are intended to be filed on the date; without complex programming, however, computers process the words for the historical event.

Another poor example used to illustrate precoordination is “Searching, Bibliographic” as this flouts the guideline to enter terms in natural language order (section 7.8). The justification for the inverted order – “Precoordination is often used to ensure logical sorting of related expressions” (p. 8) – brings to mind Charles Ammi Cutter’s murky exception to the rule for direct order: “inverting the phrase only when some other word is decidedly more significant . . .” [5, (rule) 76, p. 42]. This exception led to numerous classified arrays in Library of Congress Subject Headings. Little by little, these inverted terms are being flipped to direct order.

The first example of node labels has incorrect indentation (p. 61); the subsequent page gets it right.

Legalese

Throughout the standard, the words may, should and must appear in boldface italic type, which I find jarring. This boldfacing recurs in the Summary of Standard Requirements, in Appendix A. (I’m not sure a 31-page section – more than a third of the body of the work – merits the title “Summary,” but this appendix is useful.)

The legalese is inconsistent. In some places, the standard says that semantic relationships of thesauri should be reciprocated, for example, on pp. 9 and 166; in others, for example, in section 8.1.1 and the definition of reciprocity (p. 8 and p. 164), that they must be. I favor the latter requirement.
Thesaural Relationships

The 2005 standard has less information on thesaural relationships than its predecessor. The current *Guidelines* mention the possibility of developing more refined codes for the associative (RT) relationship (section 8.4.4), but omit the illustration of such codes that was appended to the prior edition (Figure A22). Given that the 2005 standard discusses semantic networks (section 10.9.3), which have more refined relationships than thesauri, this omission is hard to understand. Moreover, after the 1993 standard was issued, most thesaurus software vendors developed the capability to accommodate special relationship indicators.

Vocabulary Mapping

After the 1993 *Guidelines* were published, I was invited to participate in a panel about the NISO standard, held at a conference of the American Library Association. Another panelist criticized the standard for not helping him with his problem of integrating two disparate controlled vocabularies. I replied that Z39.19-1993 was a standard for developing a monolingual thesaurus, not for merging incompatible subject heading lists.

The current *Guidelines* have sections on merging vocabularies (10.6-10.8) that essentially enumerate the problems inherent in this process. I’m not convinced that these sections belong in a standard. The appendix on interoperability is useful, except that it is replete with vogue words, for example, *satellite vocabularies* (p. 143), for the well-established *microthesauri*. My *Bulletin* article on vogue words [6] could use a sequel.

Index

The index is not comprehensive. For example, there is one locator for the heading “spelling,” but the important point in section 11.4.2 about non-displayed spelling variations and typographical errors is not indexed. The same unindexed point occurs on page 93.

The summary is not indexed, except for its range of pages under the heading-subheading combination “standard: summary of” (p. 171). (Would you look this up?)

The index lacks continuation headings, even on left-hand pages. For example, p. 170 begins with the subheading “and metadata.” One must flip back to the prior page to find the heading “interoperability.” Having participated in several juries of the American Society of Indexers, I know that the lack of continuation headings in a book index causes it to be placed on the reject pile immediately, even before the structure of the index is examined.

The index has an incorrect locator for “vocabulary switching” – “44-145.” That type of error can be caught by a simple reading of a draft index.

Information Design

The only running head in the document is the standard number. The 1993 edition had running heads for chapter titles and appendix numbers, which made the standard browsable and facilitated following references such as “see Appendix D.” Given the focus of the 2005 edition on web display, in which orienting the reader is so important, more attention should have been given to the information design of the standard.

Although the standard provides a guideline on the use of running heads to identify various sequences in a printed vocabulary (section 9.4.1.3), in the glossary, *running head* is defined incorrectly as “A page heading indicating the first and last entries that appear on that page” (p. 165). The correct term for that feature, which is often found in dictionaries and encyclopedias, is *guideword* [7, p. 498].
Editorial Flaws

The sequence of bulleted items in several sections is not logical and would benefit from rearrangement. The hyperlinking of the document is incomplete; cross references to related sections and definitions are often missing. For example, there is no link between the definitions of *term* and *preferred term* even though these expressions are used interchangeably. Thorough hyperlinking would have compensated, in part, for the incomplete index.

There is an error in the summary’s reference to section 9.2.5 (p. 131) for guidelines on sorting. That section deals with capitalization. Conversely, there is a correct number in the reference to section 11.3 (p. 141), but the reference is inappropriate, as it follows the statement that relationships in graphic “displays are often built dynamically (in real-time)” [by computer], while section 11.3 deals with human maintenance of vocabularies.

In most cases, the summary replicates the rules from the body of the standard and omits the examples. In some cases, the summary provides only rubrics, resulting in misleading information. For example, the summary of section 8.4.1 (p. 115) lists among the uses of Related Term references “Mutually exclusive sibling terms.” The body of the standard, however, explains that RT references should *not* be provided between such terms (section 8.4.1.2).

This American standard uses the British spelling “Acknowledgements” (p. viii). Within a single paragraph (sect. 2.2), the spellings “multi-lingual” and “multilingual” co-occur. The heading of section 6.4.2.3 hyphenates “Cross-references,” while the text does not. A gross spelling error occurs on p. 14: “rather then.” The headword “graphics display” in the glossary (p. 160) should read “graphic.”

The definitions section does not consistently italicize terms within definitions that are themselves headwords, for example, *keywords* in the definition of *natural language* (p. 7).

Quite a few terms are not typographically distinguished from the surrounding text, for example, “if the editor changes passenger cars to automobiles . . .” (p. 101). The font size of Example 131 (p. 56) is much larger than that of other examples on the same page. The sample Candidate Term Form from DTIC (p. 147) is very fuzzy; it seems to have been faxed.

Bibliographic References

Appendix F of the Guidelines contains the references, arranged numerically in two sections: (1) Controlled Vocabularies Used as Examples and (2) Documents Referenced in the Standard. The opposite sequence would have been more logical. In any case, the references in the two sections are numbered consecutively, and these numbers are used in the body of the work.

The first reference from the text is “[29]” (p. 2). It would have been useful to explain the reference apparatus at the beginning of the standard. Given the lack of running heads in the document, it is not easy to locate the reference list. The second reference (ibid.), to the NISO technical report on indexes, follows the statement, “For indexing procedures and practices see. . . .” That report [8] deals only with indexing *principles*, not procedures and practices.

A useful bibliography follows the references. The entries are grouped in ten categories whose headings are not arranged alphabetically. Every classification needs an outline, but no list of the section headings is given at the head of the bibliography or in the table of contents.

Clearly, the bibliography will require updating – at least to cite new editions of books and standards, including RDA [9] instead of Anglo-American Cataloguing Rules, and the new IFLA Guidelines for Multilingual Thesauri [10]. In several places in the body of the standard, the reader is
told to “Check the Internet,” for example, “for current software tools” (p. 99). References to specific websites that list such tools would have been more helpful.

**Conclusion**

The 2005 Guidelines definitely require revision – at the very least to correct errors and inconsistencies and to update the bibliography.

I think it is useful to place thesauri within the spectrum of controlled vocabularies, but nothing was gained by broadening the scope of Z39.19 from its 1993 version to the 2005 edition. The latter provides no guidelines on the development of lists of subheadings, and it provides less information on thesauri than did its predecessor.

In my view, the reason the draft revision of the (now withdrawn) NISO standard Basic Criteria for Indexes failed to be approved (and hence was issued as a technical report) is that it had too broad a scope – book indexes, journal indexes, database indexes and automatic indexes. The Talmudic proverb that comes to mind is “Tafasta merubah, lo tafasta,” which means – If you try to grab too much, you grab nothing at all.

I recommend that the next edition of Z39.19 reinstate the word thesauri in the title, while explaining how thesauri differ from other controlled vocabularies. One might say that the NISO standard for controlled vocabularies should be withdrawn and the new British Standard [11] adopted instead. That standard is very expensive, however. I was told that the (reasonably priced) 1993 NISO standard for thesauri was a best-seller, and I’m sure that the 2005 edition is often downloaded.

I perceive that interest in thesauri is increasing; therefore, we need an accurate set of guidelines for this important tool for indexing and searching.

---

**Resources Cited in the Article**