



WHITE PAPER

Why LSI

The Case for Semantic Applications using Latent Semantic Indexing

April 2011



April 2011

The mission of many government agencies is dependent on their ability to transform both structured and unstructured data into meaningful insight and actionable intelligence. Unfortunately, their ability to leverage the large volumes of unstructured and structured data that they collect every day is limited. What's needed is a more scalable and systematic approach for querying, categorizing and analyzing this information.

Increasingly, this void is being filled by applications using latent semantic indexing. While this technique has existed for 20 years, recent advances in computing, such as more scalable systems, are now making it a game changer in many sectors.

LSI-based approaches offer a number of compelling advantages over traditional search technologies, including:

- The ability to search conceptually for broad ideas versus simply looking for verbatim keyword matches;
- Moving beyond the constraints of specific queries, to automatically detect and discover emerging trends in the data; and
- Sophisticated relationship mapping that can uncover otherwise unseen correlations between documents.

These unique capabilities are being applied successfully to address a number of scenarios, including intelligence analysis, medical research, fraud detection and customer service management. And unlike search technologies, they actually become more effective with larger datasets as LSI-based applications now can be applied to collections of 100 million documents or more.

Having led many of the industry's largest LSI implementations, Agilex' Semantic Engineering practice is uniquely qualified to help your organization capitalize on LSI's immense potential. While this White Paper is intended to provide a general introduction to the topic, we would welcome the opportunity to discuss sector-specific ways LSI can be used to improve your agency's performance.

Roger Bradford
Vice President & Semantic Engineering Practice Senior Scientist
Agilex Technologies, Inc.
Roger.Bradford@Agilex.com

Why LSI

The Case for Semantic Applications using Latent Semantic Indexing

| | |
|---|----|
| Introduction | 3 |
| Approaches to Semantic Processing | 4 |
| Real-world Applications | 7 |
| Conclusion | 11 |
| References | 12 |
| The Author | 14 |
| About Agilex Technologies | 15 |

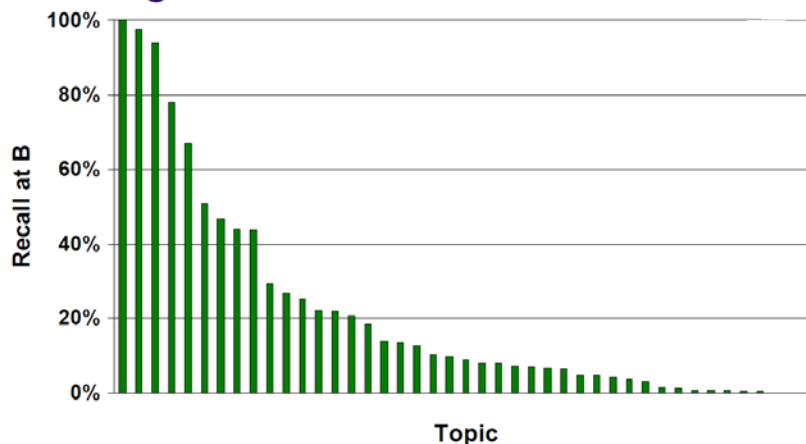
Introduction

Two trends dominate information retrieval and analysis in the federal sector – the volume of information is dramatically increasing, and the value of that information is growing just as quickly. Agencies must deal with terabytes of text, such as email, that often play a significant role in their day-to-day operations. Even small and mid-size agencies are dealing with growing volumes of text that require rapid access and meaningful analysis.

Conventional technologies for retrieving, organizing, and analyzing all of this information have not evolved as rapidly. Most information retrieval systems can only deal with words as strings of characters, or keywords, while advanced users need tools that can find underlying concepts, not just search for keywords.

For example, the Text REtrieval Conference (TREC) is an annual conference and competition sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense to advance information retrieval. The results of the TREC Legal Track evaluation in recent years have shown the gross inadequacy of Boolean retrieval as text collections reach levels of millions of documents, as shown in the following graphic:¹

“Boolean” Searches May Miss A Large Percentage of Relevant Documents



78% of relevant documents were only found by some other technique

Source: TREC 2007 Legal Track

Today, it is widely acknowledged that the ability to work with text on a semantic basis is essential to modern information retrieval systems. It is also the case that Latent Semantic Indexing (LSI) has been established as the most broadly applicable and effective capability currently deployed in the field of semantic processing.

Approaches to Semantic Processing

Efforts to incorporate semantic information into text processing systems date back nearly half a century. Over the years, designers have followed various approaches to integrating some degree of semantic processing into their information retrieval systems, including:

- Auxiliary Structures
- Local Co-Occurrence Statistics
- Latent Semantic Indexing

Auxiliary Structures

Controlled vocabularies, or auxiliary structures, such as dictionaries and thesauri, allow broader terms, narrower terms, and related terms to be incorporated into queries.² Controlled vocabularies are one way to overcome some of the most severe constraints of Boolean free-text keyword queries, which include multiple words that have similar meanings (synonymy) as well as words that have more than one meaning (polysemy). Synonymy and polysemy are often the cause of mismatches between the vocabulary used by a document's author and that employed by users of a text retrieval system.³

Over the years, additional auxiliary structures of general interest, such as the large synonym sets of Wordnet, were constructed.⁴ Later approaches implemented grammars to expand the range of semantic constructs. The most recent trend has been to create data models that represent sets of concepts within a domain (ontologies) that can incorporate relationships among terms.

Use of auxiliary structures can improve the efficiency and comprehensiveness of information retrieval and related text analysis operations.* But this approach to semantic processing works best when topics are narrowly defined and the terminology is standardized. It is not a well-suited method for meeting the information retrieval needs of most modern enterprises, with their growing volumes of unstructured data containing vast numbers of unique terms covering an unlimited number of topics.

** In evaluating information retrieval systems, efficiency and comprehensiveness are usually measured in terms of precision and recall. Precision is a standard measure of the efficiency of text retrieval. It equals the number of relevant documents retrieved by a query divided by the total number of documents retrieved. Recall is a measure of the comprehensiveness of text retrieval. It equals the number of relevant documents retrieved divided by the total number of relevant documents in the collection.*

Some other drawbacks of using auxiliary structures include:

- Establishing useful auxiliary structures requires a lot of human input and oversight.
- Language rapidly evolves, requiring the constant updating of auxiliary structures.
- Auxiliary structures can often represent the world view of their creators, introducing a potential source for conceptual mismatches.
- Auxiliary structures capture a world view at a particular point in time. They can be difficult to modify as concepts change in a specific topic area.

Local Co-Occurrence Statistics

Statistical co-occurrence was explored as a means of enlarging and sharpening literature searches by several researchers as early as the late 1950s.⁵ Co-occurrence statistics have also been widely used since the 1990s in synonym mining and word-to-word translation.⁶ Information retrieval systems using this method count the number of times pairs of terms appear together (co-occur) within a sliding window of terms or sentences within a document (for example, ± 5 sentences or ± 50 words).

This approach is simple, but it captures only a small portion of the semantic information contained in a collection of text. At the most basic level, numerous experiments have shown that only approximately a quarter of the information contained in text is local in nature.⁷ In addition, to be most effective, this method requires prior knowledge about the content of the text, which can be difficult with large, unstructured document collections.

As a result, approaches based on counting the local co-occurrence of terms are of limited value in most applications.

Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a *statistical* information retrieval method that can identify text based on the concept(s) it embodies as opposed to simply matching specific keywords. First applied to text at Bell Labs in the late 1980s⁸, it was named *latent* semantic indexing due to its ability to correlate semantically related terms that are “latent” or hidden in a collection of text.

Operationally, LSI creates a term-document matrix to identify the occurrence of terms within a set of documents; applies term weighting based on term frequencies to reflect the fact that some terms are more important than others in a body of text; and then performs a Singular Value Decomposition (SVD) on the matrix to determine patterns in the relationships between

the terms and concepts used in the documents. LSI also reduces the number of dimensions in the term space of the matrix, making it more useable and efficient.

One consequence of LSI processing is the establishment of associations between terms that occur in similar contexts. This means that queries against a set of indexed documents will return results that are conceptually similar in meaning even if they don't share a specific word or words with the query. The technique has also been shown to capture key relationship information, including causal, goal-oriented, and taxonomic information.⁹

As a result, LSI has proven to be an optimal solution for a wide range of conceptual matching problems.^{10, 11} Several experiments have demonstrated that there are a surprising number of correlations between the way LSI and humans process and categorize text.¹² And in a recent review of 16 different information processing applications, the average performance of LSI was statistically indistinguishable from that of humans.¹³

The theoretical advantages of LSI have been scientifically tested and are supported by the results of multiple studies. For example, the task of categorizing documents based on their conceptual similarities has demonstrated LSI's superiority over other approaches for extracting semantic information from documents.

In terms of automated document categorization, the Reuters 21578 test is used globally as the standard test for benchmarking and comparing various methods. Consisting of 21,578 newswire stories that have been separately categorized by Reuters personnel, it includes detailed specifications for executing the test to maintain consistency.¹⁴ The best results ever reported for the Reuters 21578 test used LSI to categorize the document set.*¹⁵

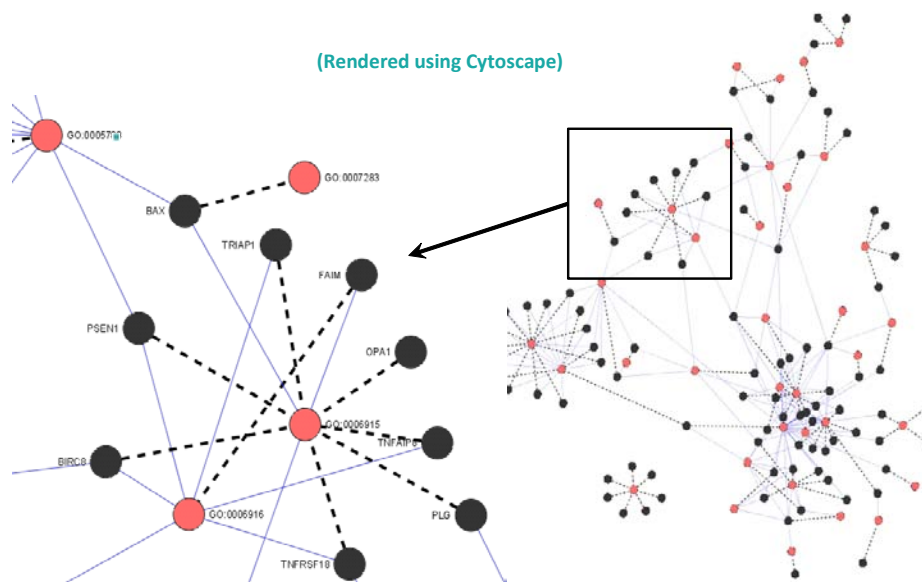
* Categories that did not contain documents were ignored.

Real-world Applications

Conceptual text retrieval and automated document categorization are primary capabilities of LSI.¹⁶ LSI is being used in a variety of contemporary text processing applications, including:

- Text summarization¹⁷
- Information discovery¹⁸
- Relationship discovery¹⁹
- Automatic generation of link charts of individuals and organizations²⁰
- Matching technical papers and grants with reviewers²¹
- Online customer support²²
- Determining document authorship²³
- Automatic keyword annotation of images²⁴
- Understanding software source code²⁵
- Filtering SPAM²⁶
- Information visualization²⁷

Some of today's most important applications emphasize information discovery. The following graphic shows results from applying LSI to discover relationships between genes and diseases:



[Latent gene and function relationships from the June 2006 Gene Ontology; later documented in the June 2007 Gene Ontology.]

Another key trend is the application of LSI for marshalling data for analysis. In the following graphic, LSI has been used to automatically identify variants of an individual's name within a collection; to highlight snippets of text where those variants appear; and to provide lists of the people, locations and organizations most closely associated with the individual. Such organization of data can relieve a researcher or analyst of much of the burden of manual data organization, allowing them to focus on activities that humans do best: inference, higher-level pattern recognition, extrapolation and related tasks.

The screenshot shows an 'Entity Profile' for 'Amari Saifi'. It includes an 'Export Profile' button and several data sections:

- Name Variants:** A list of names including Amari Saifi, Ammari Saifi, Amara Saifi, Saifi Amari, Ammar Saifi, Leader Amari Saifi, Ammari Sayifi, and Amar Saifi.
- Snippets:** Three text snippets with highlighted names and dates. The first snippet is dated 05/11/2004, the second 11/26/2002, and the third 05/12/2004. Each snippet has a 'Fulltext' link.
- Associates:** A list of names including Amari Saifi, Mokhtar Belmokhtar, Belmokhtar, Ammari Saifi, Abderrezak, Amara Saifi, El Para, Saifi, and El-para.
- Affiliated Organizations:** A list including Gspc, Salafist, Salafi Group, Salafist Group, Algerian Salafist Group, Call, Combat, L'expression, and Al-para.
- Locations:** A list including Algerian Sahara, African Sahel, Merouana, Tessalit, Illizi, Tebessa, Illizi City, El-oued, and Algerian.
- Linked Terms:** A list including Amari Saifi, Abderrezak, El-para, Mokhtar Belmokhtar, Belmokhtar, Ammari Saifi, Gspc, Abderrezak, and Salafist.

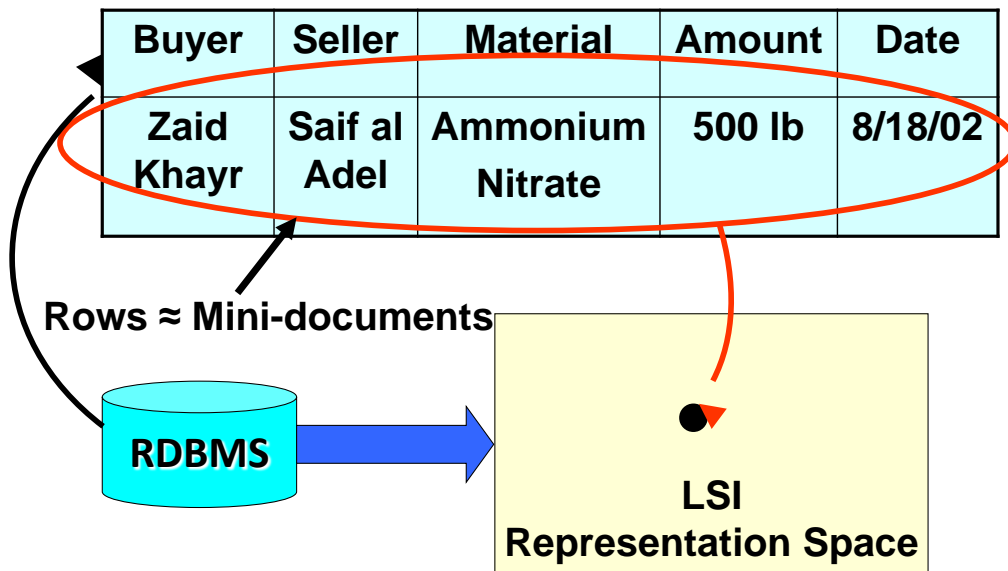
Another advantage of LSI is that it does not require text to be in sentence form. Instead, LSI can deal with lists of names, free-form notes and even tweets. As long as a term-document matrix can be generated, LSI can work with data in any format.

LSI is also completely data driven as it does not rely on auxiliary structures such as controlled vocabularies. And because LSI uses a strictly mathematical approach, it is inherently independent of language and can be used to process information in any language that can be represented in Unicode.

Likewise, LSI automatically adapts to new and changing terminology, and it has been shown to be remarkably tolerant of noise (e.g., misspelled words, typographical errors and transliteration variants).²⁸ This is especially important for applications using text derived from Optical

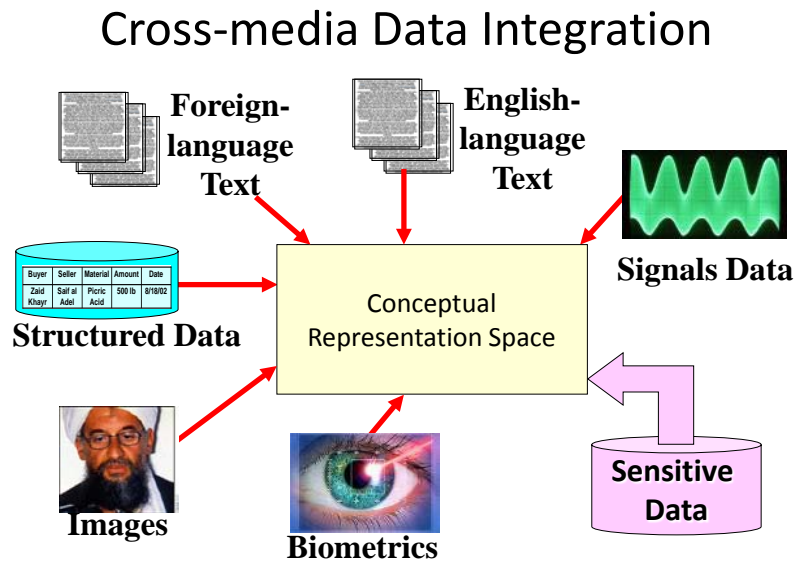
Character Recognition (OCR) and speech-to-text conversion. It also deals effectively with sparse, ambiguous, and contradictory data.

LSI is not limited to textual data. It can, in fact, be used to analyze any type of data linking events and observables. For example, information from structured databases readily can be incorporated into LSI semantic representation spaces. As shown in the following figure, rows of data can be drawn from a structured database and treated as “little documents”.



Mapping structured data into an LSI space presents an entirely new perspective on that data. In particular, subtle relationships among entities in the data can be uncovered.

LSI has also been used with great success in working with multimedia information, including audio, image, and video data.^{29 30 31 32} Of great significance is the fact that multiple types of data can be represented and analyzed simultaneously in a single LSI semantic representation space, as indicated in the following figure. This ability to fuse information from a wide variety of disparate sources is of great practical value in situational awareness and data mining applications.



Finally, LSI has been shown to scale well. Today, repositories of up to 100 million documents routinely are being addressed using high-end servers. Initial forays into the application of cloud computing environments are showing the way towards even larger repositories.

Conclusion

There is little argument in the industry that basic Boolean and keyword-driven search techniques are being rapidly outpaced across a variety of text analytics and processing scenarios by both user demands and application requirements. Of the various semantic analysis technologies in use today, LSI has proven itself to be a robust platform offering the most precise and comprehensive performance. And it is being used today in mission-critical, 24X7 applications.

Today's multicore, multiprocessor server architectures support routine creation of LSI semantic representation spaces for collections of up to 100 million documents. Larger applications are being implemented in cloud environments. The wealth of applications developed using LSI underscores the technology's tremendous power and flexibility. The ability of LSI to operate without the auxiliary structures needed by other semantic techniques (i.e., word lists and thesauri) makes it ideally suited for complex analysis of unstructured document collections. Finally, LSI supports the creation of very flexible systems that can cross applications and even languages without costly and lengthy development.

###

References

- ¹ Baron, J., Beyond Keywords: Emerging Best Practices in the Area of Search and Information Retrieval, New Mexico Digital Preservation Conference, 5 June, 2008.
- ² Dubois, C., The Use of Thesauri in Online Retrieval, *Journal of Information Science*, 8(2), 1984 March, pp. 63-66.
- ³ Furnas, G., et al, The Vocabulary Problem in Human-system Communication, *Communications of the ACM*, 1987, 30(11), pp. 964-971.
- ⁴ Miller, G., Special Issue, WordNet: An On-line Lexical Database, *Intl. Journal of Lexicography*, 3(4), 1990.
- ⁵ Maron, M. and Kuhns, J., On Relevance, Probabilistic Indexing and Information Retrieval, *Journal of the ACM*, 1960;7, pp. 216-244.
- ⁶ Turney P.D, Mining the web for Synonyms: PMR-IR versus LSA on TOEFL, *ECML*, 2001, pp. 491-502.
- ⁷ Landauer, T., and Dumais, S., A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge, *Psychological Review*, 1997, 104(2), pp. 211-240.
- ⁸ Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing, in: *Proceedings of the 51st Annual Meeting of the American Society for Information Science 25*, 1988, pp. 36-40.
- ⁹ Graesser, A., and Karnavat, A., Latent Semantic Analysis Captures Causal, Goal-oriented, and Taxonomic Structures, *Proceedings of CogSci 2000*, pp. 184-189.
- ¹⁰ Ding, C., A Similarity-based Probability Model for Latent Semantic Indexing, in: *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 59-65.
- ¹¹ Bartell, B., Cottrell, G., and Belew, R., Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling, in: *Proceedings, ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992, pp. 161-167.
- ¹² Landauer, T. , et al., Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report, in: M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10*, Cambridge: MIT Press, 1998, pp. 45-51.
- ¹³ Bradford, R., Comparability of LSI and Human Judgment in Text Analysis Tasks, in: *Proceedings, Applied Computing Conference, Athens, Greece, 28-30 September, 2009*, pp. 359-366.
- ¹⁴ <http://www.daviddlewis.com/resources/testcollections/reuters21578/>
- ¹⁵ Zukas, A., and Price, R., Document Categorization Using Latent Semantic Indexing, in: *Proceedings, Symposium on Document Image Understanding Technology, 2003*, pp. 87-91.
- ¹⁶ Dumais, S., Latent Semantic Analysis, in *ARIST Review of Information Science and Technology*, vol. 38, 2004, Chapter 4.
- ¹⁷ Gong, Y., and Liu, X., Creating Generic Text Summaries, in: *Proceedings, Sixth International Conference on Document Analysis and Recognition, 2001*, pp. 903-907.
- ¹⁸ Bradford, R., Efficient Discovery of New Information in Large Text Databases, in: *Proceedings, IEEE International Conference on Intelligence and Security Informatics, Atlanta, Georgia, LNCS Vol. 3495, Springer, 2005*, pp. 374-380.
- ¹⁹ Bradford, R., Relationship Discovery in Large Text Collections Using Latent Semantic Indexing, in: *Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism, and Security, SIAM Data Mining Conference, Bethesda, MD, 20-22 April, 2006*.

- ²⁰ Bradford, R., Application of Latent Semantic Indexing in Generating Graphs of Terrorist Networks, in: Proceedings, IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24, 2006, Springer, LNCS vol. 3975, pp. 674-675.
- ²¹ Yarowsky, D., and Florian, R., Taking the Load off the Conference Chairs: Towards a Digital Paper-routing Assistant, in: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very-Large Corpora, 1999, pp. 220-230.
- ²² Caron, J., Applying LSA to Online Customer Support: A Trial Study, Unpublished Master's Thesis, May 2000.
- ²³ Soboroff, I., et al, Visualizing Document Authorship Using N-grams and Latent Semantic Indexing, Workshop on New Paradigms in Information Visualization and Manipulation, 1997, pp. 43-48.
- ²⁴ Monay, F., and Gatica-Perez, D., On Image Auto-annotation with Latent Space Models, in: Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA, 2003, pp. 275-278.
- ²⁵ Maletic, J., and Marcus, A., Using Latent Semantic Analysis to Identify Similarities in Source Code to Support Program Understanding, in Proceedings of 12th IEEE International Conference on Tools with Artificial Intelligence, Vancouver, British Columbia, November 13-15, 2000, pp. 46-53.
- ²⁶ Gee, K., Using Latent Semantic Indexing to Filter Spam, in: Proceedings, 2003 ACM Symposium on Applied Computing, Melbourne, Florida, pp. 460-464.
- ²⁷ Landauer, T., Laham, D., and Derr, M., From Paragraph to Graph: Latent Semantic Analysis for Information Visualization, in: Proceedings of the National Academy of Science, 101, 2004, pp. 5214-5219.
- ²⁸ Price, R., and Zukas, A., Application of Latent Semantic Indexing to Processing of Noisy Text, Intelligence and Security Informatics, Lecture Notes in Computer Science, Volume 3495, Springer Publishing, 2005, pp. 602-603.
- ²⁹ Kurimo, M., Indexing Spoken Audio by LSA and SOMS, IDIAP Research Report 00-06, April, 2000.
- ³⁰ Bellegarda, J., Exploiting Latent Semantic Information in Statistical Language Modeling, in: Proceedings of the IEEE, Volume 88, Issue 8, Aug 2000 pp. 1279 – 1296.
- ³¹ Praks, P., Dvorsky, J., Snasel, V., Latent Semantic Indexing for Image Retrieval Systems in: Proceedings of the SIAM Conference on Applied Linear Algebra, 2003.
- ³² Souvannavong, F., Merialdo, B., and Huet, B., Latent Semantic Analysis for an Effective Region-based Video Shot Retrieval System, in: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York, 2004, pp. 243 – 250.

The Author

Roger Bradford is a Vice President and Senior Scientist with Agilex Technologies, Inc. He has led the company's Semantic Engineering Practice since its inception. His current work focuses on the application of LSI for data mining within very large text collections, encompassing hundreds of millions of documents.

Previously, he served with SAIC as a Senior Scientist and Technical Fellow. He was presented the SAIC Excellence in Science and Technology Award in 2003 for his pioneering work with Latent Semantic Indexing.

Mr. Bradford currently holds six patents relating to LSI applications and extensions, with additional patents pending. He has published numerous academic papers on LSI-related topics, including information discovery, social network analysis and secure data exchange. He has also presented LSI-focused research at numerous conferences, including those of the ACM, IEEE and the Society for Industrial and Applied Mathematics (SIAM).

About Agilex Technologies, Inc.

Agilex is an employee-owned provider of mission and technology solutions to the national security, healthcare and public sectors. Realize the Value of Information® is our corporate mantra as we help clients unlock the value of information while reducing the cost to manage it. Our notable professionals include former federal CIOs, multiple patent holders, and architects for some of the U.S. government's most critical IT systems.

Headquartered in Chantilly, Virginia, Agilex has delivered significant results for an impressive list of clients throughout federal, state and local government, and within global 2000 corporations. The *Washington Business Journal* named Agilex the Washington, DC area's Fastest Growing Company in 2010. The company was also recognized by the Northern Virginia Technology Council (NVTC) as the organization's 2010 Hottest Emerging Government Contractor for the Washington, DC metro region.

For more information, go to www.Agilex.com or call 1-888-3AGILEX.

The logo for Agilex, featuring the word "Agilex" in a bold, italicized sans-serif font. The "A" is teal, and the "gilex" is gold. A thin gold arc is positioned above the "gilex" portion of the text.

Agilex