



WHITE PAPER

Why LSI? Latent Semantic Indexing and Information Retrieval

By Roger Bradford

Agilex Technologies, Inc., Chantilly, Virginia

Two trends dominate information retrieval and analysis in today's enterprise: the volume of information is dramatically increasing, and the value of that information is growing just as fast. Modern enterprises must deal with terabytes of text, such as email, that often play a significant role in their day-to-day operations. Even small and medium-sized enterprises are dealing with growing volumes of text that require rapid access and meaningful analysis.

Conventional technologies for retrieving, organizing, and analyzing all of that information have not evolved as quickly. Most information retrieval systems can only deal with words as strings of characters, or keywords, although advanced users need tools that can find underlying concepts, not just search for keywords. 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 apparent that Latent Semantic Indexing (LSI) is one of the most effective approaches 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:

- » 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. Controlled vocabularies are one way to overcome some of the most severe constraints of Boolean free-text keyword queries—multiple words that have similar meanings (synonymy), and words that have more than one meaning (polysemy). Synonymy and polysemy are often the cause of mismatches in the vocabulary used by the authors of documents and the users of text retrieval systems.

Over the years, additional auxiliary structures of general interest, such as the large synonym sets of Wordnet,

were constructed. 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), which can incorporate relationships among terms.

Controlled vocabularies can contribute to 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 isn't well-suited to the information retrieval needs of most modern enterprises and the growing volumes of unstructured data which contain thousands of unique terms covering an unlimited number of topics.

Some other drawbacks of using auxiliary structures:

- » Establishing useful controlled vocabularies requires lots of human input and oversight.
- » Language rapidly evolves, requiring the constant updating of controlled vocabularies.
- » Controlled vocabularies can often represent the world view of their creators, introducing a potential source for conceptual mismatches.
- » Controlled vocabularies 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. Co-occurrence statistics have also been widely used since the 1990s in synonym mining and word-to-word translation.

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 (for example, ± 5 sentences or ± 50 words) within a document. 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 $\frac{1}{4}$ of the information contained in text is local in nature. 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 is a statistical information retrieval method that is capable of retrieving text based on the concepts it contains, not just by matching specific keywords. First applied to text at Bell Labs in the late 1980s, it was called LSI because of its ability to correlate semantically related terms that are “latent” in a collection of text.

LSI uses 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 uses a mathematical transform technique to reduce the number of dimensions in the term space of the matrix to make it more useable

and efficient. One consequence of LSI processing is the establishment of associations between terms that occur in similar contexts. As a result, queries against a set of documents that have undergone LSI will return results that are conceptually similar in meaning to the query even if they don't share a specific word or words with the query.

LSI has proven to be an optimal solution for a wide range of conceptual matching problems. The technique has been shown to capture key relationship information, including causal, goal-oriented, and taxonomic information. Several experiments have also demonstrated that there are a surprising number of correlations between the way LSI and humans process and categorize text.

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

The Reuters 21578 test for automated document categorization has been used worldwide for at least ten years as the standard test set for benchmarking various approaches to automated document categorization. Consisting of 21,578 newswire stories that have been categorized by Reuters personnel, the test set includes a specification for the test procedure including the partitioning of the test set into training and test data, the metric to be used for measuring the test results, and the technique for calculating the metric values. The best results ever reported for the Reuters 21578 test used LSI to categorize the document set.



Real World Application

LSI is being used in a variety of information retrieval and text processing applications, although its primary application has been for conceptual text retrieval and automated document categorization. Below are some other ways in which LSI is being used:

- » Text summarization
- » Information discovery
- » Relationship discovery
- » Automatic generation of link charts of individuals and organizations
- » Matching technical papers and grants with reviewers
- » Online customer support
- » Determining document authorship
- » Automatic keyword annotation of images
- » Understanding software source code
- » Filtering SPAM
- » Information visualization



The use of LSI does not require that text be in sentence form. LSI can deal with lists of names, free-form notes, etc., as well as e-mails and blogs. As long as a term-document matrix can be generated, LSI can work with data in any format or language. LSI also uses no auxiliary structures such as controlled vocabularies—it is completely data-driven.

LSI automatically adapts to new and changing terminology, and it has been shown to be remarkably tolerant of noise (i.e., misspelled words, typographical errors, unreadable characters, etc.). This is especially important for applications using text derived from Optical Character Recognition (OCR) and speech-to-text conversion. LSI also deals effectively with sparse, ambiguous, and contradictory data.

Because LSI uses a strictly mathematical approach, it is inherently independent of language. It can be used to process information in any language—generally any language that can be represented in Unicode.

Conclusions

There is little argument in the industry that basic Boolean and keyword-driven search techniques are being rapidly outpaced by both user demands and software requirements across a variety of text analytics and processing solutions. Of the various semantic analysis technologies in use today, LSI has proven itself to be a stable platform offering the greatest application benefit.

Earlier challenges for LSI in terms of scalability and required computing horsepower have been addressed both by software refinements and the overall computer hardware industry. Today's servers are well-positioned to take advantage of LSI, and the web client architectures adopted by companies like Content Analyst lend themselves to traditional environments and emerging Software as a Service (SaaS) markets. 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 (word lists, thesauri, etc.) makes it ideally suited for complex analysis of unstructured document collections. Finally, LSI supports the creation of very flexible products that can cross industries and even languages without costly and lengthy development.

###

Roger Bradford is the Technical Director of the Semantic Processing Practice at Agilex Technologies Inc. Previously he was a Senior Scientist and Technical Fellow at SAIC. In 2003 he received the SAIC Excellence in Science and Technology award for his work in LSI. He has ten years experience in developing information systems that employ LSI technology. He currently holds five patents and has an additional five pending that deal with extensions and applications of LSI techniques. He has presented papers on LSI at ACM, IEEE, and SIAM conferences. His published papers in this area cover applications of LSI in information discovery, social network analysis, and secure data sharing. His current work focuses on application of LSI for data mining in large text collections (tens of millions of documents).

References

- 1Dubois, C., The Use of Thesauri in Online Retrieval, *Journal of Information Science*, 8(2), 1984 March, pp. 63-66.
- 2Furnas, G., et al, The Vocabulary Problem in Humansystem Communication, *Communications of the ACM*, 1987, 30(11), pp. 964-971.
- 3Miller, G., Special Issue, WordNet: An On-line Lexical Database, *Intl. Journal of Lexicography*, 3(4), 1990.
- 4Maron ME, Kuhns JL, On relevance, probabilistic indexing and information retrieval, *Journal of the ACM*, 1960;7, pp. 216-244.
- 5Turney PD, Mining the web for synonyms: PMR-IR versus LSA on TOEFL, ECML, 2001, pp. 491-502.
- 6Landauer, 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.
- 7Deerwester, 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.
- 8Ding, 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.
- 9Bartell, 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.
- 10Graesser, A., and Karnavat, A., Latent Semantic Analysis Captures Causal, Goal-oriented, and Taxonomic Structures, *Proceedings of CogSci 2000*, pp. 184-189.
- 11Landauer, 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.
- 12<http://www.daviddlewis.com/resources/testcollections/reuters21578/>
- 13Zukas, A., and Price, R., Document Categorization Using Latent Semantic Indexing, in: *Proceedings, Symposium on Document Image Understanding Technology*, 2003, pp. 87-91.
- 14Dumais, S., Latent Semantic Analysis, in *ARIST Review of Information Science and Technology*, vol. 38, 2004, Chapter 4.
- 15Gong, Y., and Liu, X., Creating Generic Text Summaries, in: *Proceedings, Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 903-907.
- 16Bradford, 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.
- 17Bradford, 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.
- 18Bradford, 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.
- 19Yarowsky, 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.
- 20Caron, J., Applying LSA to Online Customer Support: A Trial Study, Unpublished Master's Thesis, May 2000.
- 21Soboroff, 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.
- 22Monay, 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.
- 23Maletic, 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.
- 24Gee, K., Using Latent Semantic Indexing to Filter Spam, in: *Proceedings, 2003 ACM Symposium on Applied Computing*, Melbourne, Florida, pp. 460-464.
- 25Landauer, 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.
- 26Price, 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.