# Defensibility of Content Analyst Analytical Technology (CAAT) for Use in Legal Proceedings

Content Analyst Company, LLC.

## Overview

XERA powered by CAAT provides advanced data analytics technology for searching, analyzing, and reviewing Electronically Stored Information (ESI) for early case assessment and throughout the discovery phase of civil lawsuits (eDiscovery).

CAAT functionality is based on the use of Latent Semantic Indexing (LSI), an indexing and retrieval method that uses a mathematical transform technique called a rank-reduced Singular Value Decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. A rank-reduced SVD was first applied to text at Bell Labs in the late 1980s. This technique was named LSI because of its ability to find the semantic meaning that is latent in a collection of text. Today, LSI is being used in a variety of information retrieval and text processing applications, although its primary application has been for concept searching and automated document clustering and categorization. See http://en.wikipedia.org/wiki/Concept_search for more information about concept searching.

In general, the LSI process involves constructing a weighted term-document matrix, performing a rank-reduced SVD on the matrix, and then using the matrix to identify the concepts contained in the text of the document collection. Matrix decomposition techniques like SVD are data-driven, which avoids many of the drawbacks associated with other approaches to concept search that use auxiliary structures such as controlled vocabularies and ontologies. They are also global in nature, which means they are capable of much more robust information extraction and representation of semantic information than concept search techniques based on local co-occurrence statistics, which capture only a small portion of the semantic information contained in a collection of text. See http://en.wikipedia.org/wiki/Latent_semantic_indexing for a detailed description of LSI and SVD.

## The Mathematics of Latent Semantic Indexing (LSI)

LSI begins by constructing a term-document matrix, A, to identify the occurrences of the m unique terms within a collection of n documents. In a term-document matrix, each term is represented by a row, and each document is represented by a column, with each matrix cell, aij, initially representing the number of times the associated term appears in the indicated document, tfij. This matrix is usually very large and very sparse.

Once a term-document matrix is constructed, local and global weighting functions can be applied to it to condition the data. The weighting functions transform each cell, aij of A, to be the product of a local term weight, lij, which describes the relative frequency of a term in a document, and a global weight, gi, which describes the relative frequency of the term within the entire collection of documents.

A rank-reduced, Singular Value Decomposition (SVD) is performed on the matrix to determine patterns in the relationships between the terms and concepts contained in the text. The rank-reduced SVD forms the foundation for LSI. It computes the term and document vector spaces by transforming the single term-frequency matrix, A, into three other matrices— a term-concept vector matrix, T, a singular values matrix, S, and a concept-document vector matrix, D, which satisfy the following relations:

$$A = TSD^T$$
$$T^TT = D^TD = I_r \quad TT^T = I_m \quad DD^T = I_n$$
$$S_{1,1} \geq S_{2,2} \geq ... \geq S_{r,r} > 0 \quad S_{i,j} = 0 \text{ where } i \neq j$$

In the formula, A, is the supplied m by n weighted matrix of term frequencies in a collection of text where m is the number of unique terms, and n is the number of documents. T is a computed m by r matrix of term vectors where r is the rank of A—a measure of its unique dimensions <= min(m,n). S is a computed r by r diagonal matrix of decreasing singular values, and D is a computed n by r matrix of document vectors.

The LSI modification to a standard SVD is to reduce the rank or truncate the singular value matrix S to size k « r, typically on the order of a k in the range of 100 to 300 dimensions, effectively reducing the term and document vector matrix sizes to m by k and n by k respectively. The SVD operation, along with this reduction, has the effect of preserving the most important semantic information in the text while reducing noise and other undesirable artifacts of the original space of A. This reduced set of matrices is often denoted with a modified formula such as:

$$A \approx A_k = T_k S_k D_k^T$$

The computed Tk and Dk matrices define the term and document vector spaces, which with the computed singular values, Sk, embody the conceptual information derived from the document collection. The similarity of terms or documents within these spaces is a factor of how close they are to each other in these spaces, computed as a function of the angle between the corresponding vectors.

CAAT is able to extract the conceptual content of text based on the associations between words that occur in a similar context. Queries against a set of documents that have been indexed by CAAT will return results that are conceptually similar in meaning to the search criteria even if the results don't share a specific word, or words, within those search criteria.

Dynamic clustering based on the conceptual content of documents can also be accomplished using CAAT. Clustering is a way to group documents based on their conceptual similarity to each other without using example documents to establish the conceptual basis for each cluster. The titles for each cluster are derived from the concepts contained in the clustered documents. Clustering is very useful for identifying the conceptual content of a collection of unknown documents.

The ability to extract the conceptual contents of documents also enables CAAT to perform automated categorization (classification), which assigns documents to predefined categories based on their similarity to the concepts that each category represents. Example

documents (exemplars) are used to establish the conceptual basis for each category. During categorization processing, CAAT compares the concepts contained in the documents being categorized to the concepts contained in the example items, and documents are assigned to a category (or categories) based on the similarities between the concepts contained in the documents and the concepts that are contained in the example documents. Several experiments have demonstrated that there are a number of correlations between the way LSI and humans process and categorize text.

CAAT also provides the ability to perform metadata searches and includes a keyword search component for executing Boolean and keyword queries. Metadata searches and keyword searches can be combined with concept searches to limit the results of concept searches and provide powerful search capabilities that are particularly useful for early case assessment and for the culling of non-responsive documents. CAAT also utilizes concept searches to help clarify the learned context of specific words, which can be used to enlarge keyword lists for finding conceptually related content.

## The Benefits of CAAT Technology in Litigation Support

According to International Data Corporation (IDC), in 2006, companies in the U.S. created, captured, and stored 161 billion gigabytes of ESI. Richard Stout of Lawdable.com stated in March 2009 that a "medium-size matter" now averages 30 gigabytes of data. This tremendous growth in the volume of ESI, all of which is potentially discoverable in litigation, has required the development of new forms of search and retrieval technologies that can overcome some of the inherent weaknesses of keyword or Boolean search techniques in the processing and review of relevant documents.

The CAAT LSI-based functionality overcomes two of the most severe constraints of Boolean and 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 members of litigation support teams responsible for document analysis and for identifying non-responsive documents. For the 200 most-polysemous terms in English, the typical verb has more than twelve common meanings, or senses. The typical noun from this set has more than eight common senses. For the 2000 most-polysemous terms in English, the typical verb has more than eight common senses and the typical noun has more than five. Boolean and keyword queries often return irrelevant results and miss information that is relevant.

Concept searches using CAAT are also very tolerant of noise (i.e., misspelled words, typographical errors, unreadable characters, etc.), which can render data inaccessible to simple keyword searches. This is especially important during the review phase in regard to documents derived from Optical Character Recognition (OCR) and speech-to-text conversion. The CAAT concept search technology also deals effectively with sparse, ambiguous, and contradictory data.

The CAAT automated categorization capability is a cost-effective way for litigation support teams involved in performing early case assessment. Implementing the same LSI-based functionality used to conduct concept searches, automated categorization is an efficient way to classify documents based on their conceptual content because it relies on example documents to establish the concepts that define the categories of interest. Benchmark tests have demonstrated that CAAT technology is capable of categorizing millions of documents per day with an accuracy that is consistent with, or better than, human-based categorization.

## Defending the Results of CAAT Technology

The primary uses of CAAT in litigation support are for finding, identifying, analyzing, and grouping conceptually related documents to determine if they are responsive to a lawsuit, whether or not they contain privileged information, and if they include issue-based information about a case. CAAT supports focused reviews for which the documents for review are organized by topic and/or concept, and which are four to five times more efficient than traditional linear reviews. In eDiscovery, the ability to search, cluster, and categorize large collections of unstructured text on a conceptual basis is much more efficient than traditional linear review techniques, which require manual review of each document usually following a first received/first reviewed process.

The CAAT LSI-based functionality ensures that the most conceptually similar documents are always identified and included in the results of concept searches as well as in clustering and categorization results. Although CAAT technology has never actually been challenged in a courtroom, it satisfies three key tenets of software defensibility:

### 1. Is there solid technical ground for the technique and can the technique be explained?

» The LSI-based technology used by CAAT is patented, well-tested, and profusely documented. The linear algebra used to compute the rank-reduced SVD converts each document into a mathematical representation called a vector, and the original document, which must be preserved for litigation, is not altered in any way. The most widely used variant of the algorithm used for computing the SVD was first published in 1970. See, http://en.wikipedia.org/wiki/Singular_value_decomposition for a detailed explanation of Singular Value Decomposition.

### 2. Are the eDiscovery processes provided by CAAT technology easily repeatable as required by the Federal Rules of Civil Procedure (FRCP)?

» The mathematical foundation of the CAAT LSI-based technology ensures that its processes and findings are completely repeatable and systematic. When given the same data and settings, CAAT will always produce the same results for concept searches, dynamic clustering, and document classification.

### 3. Are the results of the CAAT eDiscovery processes inclusive? Are all relevant documents included in the results?

» CAAT does not exclude results as keyword searching does. The CAAT user-specified relevancy thresholds allow the most conceptually similar documents to be consistently identified and included in its results. Setting the relevancy threshold higher means that only the documents that are the most conceptually relevant to a query will be included in concept search results. Setting the threshold lower allows less conceptually similar documents to be included in the results. Setting the relevancy threshold to zero means that all of the documents in a dataset will be included in the results of a concept search no matter how relevant they are to query. Result documents are also ranked, and can be sorted, by their conceptual relevance values.

### 4. Is there any legal precedent?

» There are numerous examples in which advanced analytics feature in judicial recommendations and even legislation. Advanced analytics are being increasingly referenced as an alternative to keyword-only searches. In the 2007 Disability Rights Council v Washington Metro Area Transit Authority, the parties' inability to come to agreement over keyword search terms prompted Honorable John Facciola to issue the following recommendation: "I bring to the parties' attention recent scholarship that argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results."

» On the legislative front, advanced analytics such as CAAT are also gaining acceptance. FRE-502, which was signed into law in September 2008, was intended to address attorney-client privilege and to bring clarity to how inadvertent disclosures could be handled. When interpreting FRE-502, attorneys will typically look at Congressional working committee notes; for 154 CONG. REC. S1318 (Feb. 27, 2008), these specifically state, "...a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken 'reasonable steps' to prevent inadvertent disclosure."

Well positioned to be defensible in legal proceedings, much more efficient and timely than traditional linear review techniques, and technologically superior to simple keyword or Boolean search tools, CAAT technology provides an ideal solution for controlling the total cost of review and meeting the diverse challenges of eDiscovery.

## References

Bradford, R. B., Why LSI? Latent Semantic Indexing and Information Retrieval, White Paper, Content Analyst Company, LLC, 2008.

Landauer, T., et al., Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report, M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10, Cambridge: MIT Press, 1998, pp. 45–51.

Bradford, R. B., Word Sense Disambiguation, Content Analyst Company, LLC, U.S. Patent 7415462, 2008.

United States District Court, District of Columbia, Disability Rights Council Of Greater Washington, et al., Plaintiffs, v. Washington Metropolitan Transit Authority, et al., Defendants, Civil Action No. 04-498 (HHK/JMF), June 1, 2007, Memorandum Opinion

Public Law 110–322, 110th Congress, Rule 502. Attorney-Client Privilege and Work Product; Limitations on Waiver, Advisory Committee Note Federal Rule Of Evidence 502 With Congressional Statement Of Intent

## About Content Analyst Company

We provide powerful and proven Advanced Analytics that exponentially reduce the time needed to discern relevant information from unstructured data. CAAT, our dynamic suite of text analytics technologies, delivers significant value wherever knowledge workers need to extract insights from large amounts of unstructured data. Our capabilities are easily integrated into any software solution, and our support strategy for our partners is second to none. For more information visit www.contentanalyst.com.



## About iCONECT Development, LLC

iCONECT Development, LLC is a global market leader in developing innovative legal review software, services and programs that empower people to complete complex legal review projects more cost-effectively.  As early innovators of eDiscovery review software since 1999, and with the release of its revolutionary web-based XERA eDiscovery Review Platform, iCONECT continues to raise the bar for advancements in developing intelligent, easy-to-use tools that help hosting providers, law firms, corporations and government agencies worldwide optimize workflows and manage big data more efficiently.

iCONECT's worldwide corporate headquarters are based in Reston, Virginia (Washington, D.C.), with satellite offices in Los Angeles, California, and London, Canada.