White Paper

Using LSI for Implementing Document Management Systems

DataCube

Turning unstructured data from a liability to an asset.

Using LSI for Implementing Document Management Systems

By Mike Harrison, Director, Apperception Services

Overview

Latent Semantic Indexing (LSI) is an automated technique for the processing of textual material. It provides state-of-the-art capabilities for:

- automatic document categorization;
- conceptual information retrieval, and;
- cross-lingual information retrieval.

In this document, we intend to show how this technique (as embedded into Apperception's DataCube) can be used as a highly effective method for data migration from unstructured data into a structured Electronic Document and Record Management System (EDRMS)

Role of the Document Management System

The desired benefits of implementing a Document Management solution can include:

- the secure and systematic management of unstructured or semi-structured data such as emails and documents
- a reduction in redundancy and duplication of information
- a reduced risk of not being able to retrieve information when required
- improved security, thereby reducing the risk of unauthorised access
- greater ability to discover and re-use corporate information
- better control of document versions, and
- a reduction in the response time for information requests

However, most implementation plans run into several major issues which are sometimes insurmountable.

1) What taxonomy (file plan) can best represent the requirements of the organisation when setting up the Document Management Syste, ?
2) How is it possible to transfer all of the unstructured legacy data into the structured Document Management System ?
3) How can users be made to use the required Document Management structure ?

Some projects fail because the working party cannot agree what the best (optimum) taxonomy should be whilst others fail because users object to the implementations of the structure imposed part way through the implementation. Many more EDRMS Implementations fail because even though they have been forced through, users soon fall into bad habits and stop using the structure and disciplines that the EDRMS system imposes.

To avoid these pitfalls, Apperception's DataCube can offer help in a number of areas. At the core of the DataCube is a Latent Semantic Indexing engine that indexes and categorises all of the unstructured data, which in turn can assist the EDRMS implementation plan.

How LSI Works

Latent semantic indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings.

Called Latent Semantic Indexing because of its ability to correlate semantically related terms that are latent in a collection of text, it was first applied to text at Bell Laboratories in the late 1980s. Only recently has the computer power and capacity been affordable to make the solution commercially viable. The method, also called latent semantic analysis (LSA), uncovers the underlying latent semantic structure in the usage of words in a body of text and how it can be used to extract the meaning of the text in response to user queries, commonly referred to as concept searches. Queries, or concept searches, against a set of documents that have undergone LSI 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 with the search criteria.

An example of how this is used is to compare a keyword search. If we use the word "bank", it has several meanings – a financial institution, to tilt an aircraft, a hill or slope and as verb to depend on someone or something. Keyword searching would list them all and even with Boolean logic ("AND", "NOT", "OR" etc.) it would not be able to sufficiently distinguish between the uses of the word. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. So when the words around "bank" are examined, the concept or detail how the word is used (the "latent semantics") is then determined.

LSI overcomes two of the most problematic constraints of Boolean keyword queries: multiple words that have similar meanings (synonymy) and words that have more than one meaning (polysemy). Synonymy is often the cause of mismatches in the vocabulary used by the authors of documents and the users of information retrieval systems. As a result, Boolean or keyword queries often return irrelevant results and miss information that is relevant.

Another key feature of LSI is that it is capable of automatically extracting the conceptual content of text items. With knowledge of their content, these items then can be treated in an intelligent manner. For example, documents can be routed to individuals based on their job responsibilities. Similarly, emails can be filtered accurately. Information retrieval operations can be carried out based on

the conceptual content of documents, not on the specific words that they contain. This is very useful when dealing with technical documents, particularly cross-disciplinary material.

LSI is not restricted to working only with words. It can also process arbitrary character strings. Any object that can be expressed as text can be represented in an LSI vector space. For example, tests with MEDLINE abstracts have shown that LSI is able to effectively classify genes based on conceptual modelling of the biological information contained in the titles and abstracts of the MEDLINE citations.

LSI automatically adapts to new and changing terminology, and has been shown to be very 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.

Text does not need to be in sentence form for LSI to be effective. It can work with lists, free-form notes, email, Web-based content, etc. As long as a collection of text contains multiple terms, LSI can be used to identify patterns in the relationships between the important terms and concepts contained in the text. LSI has proven to be a useful solution to a number of conceptual matching problems. The technique has been shown to capture key relationship information, including causal, goal-oriented, and taxonomic information.

Document Classification

LSI is also used to perform automated document categorization. In fact, several experiments have demonstrated that there are a number of correlations between the way LSI and humans process and categorize text. Document categorization is the assignment of documents to one or more predefined categories based on their similarity to the conceptual content of the categories.

LSI helps to classify documents in two main ways :

1) In the **UNSUPERVISED** mode, the DataCube scans all of the organisation's data and groups the document into a specified number of clusters of documents that are conceptually related to each other. This process (known as Dynamic 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. This is very useful when dealing with an unknown collection of unstructured text. The DataCube provides the ability to group the documents into as many categories as the user specifies and even suggests a naming convention. In this way, the taxonomy will reflect the actual data rather than a proposed

structure. Dynamic clustering based on the conceptual content of documents can also be accomplished using LSI.

2) In S**UPERVISED** mode, the data is categorised according to an existing Taxonomy using example documents (Exemplars) for each category. For example, the in DataCube, we use the Local Government Classification Schema when working with data from Local Government Authorities. The LGCS is fairly common across most UK Local Government Authorities for structuring data and consists of around 750 categories which define all areas of their business. Using Apperception has defined each category in some detail and attached around 20 to 25 Exemplars to each category. Exemplars are best (or worst) examples of documents that match the concepts of documents that belong to that category. The DataCube scans every document in an organisation's data network and compares it using LSI to the exemplars in the schema and defines the category according to the strongest correlation to the concepts in the exemplars. When the optimum correlation is found, the document is categorised according to which category the exemplar belongs to. This method has been proven during in-depth trials with a leading London Borough to be over 85% accurate on large volumes (over 2 million documents). With further refinement of the process (e.g. by more in-depth analysis of the exemplars by Subject Matter Experts) and by complimenting this process with a rules-based system with exceptions, the level of success is expected to be over 90%.

Creating Order out of Chaos

A successful document classification methodology depends on a number of factors :

1) **The organisational or categorical structure, often referred to as taxonomy, needs to be well-defined so that the user or automated technology knows the characteristics a document must have to be classified accordingly.**

   The DataCube provides a Schema Management module as the basis of the automated system. Each category is defined and the attributes and properties described, including default properties for the category e.g. Retention Period. Exemplars are then associated with this category, which provide best and worst examples of documents that fit into this category. Schemas can be created to best fit the organisation's need and can be bespoke. However, sometimes Industry standard schemas used.

   For example, Apperception has used the Local Government Classification Schema for UK Local Authorities which defines 750 categories which we have defined in detail and populated with over 15,000 exemplars from

which all of the personal data has been removed. An index and categorisation system is therefore ready to deploy into any UK Local Authority, although it does allow modification and enhancements to reflect differences for each authority.

2) **This structure needs to reflect the needs of the end consumers of the document as well as have a clear labelling or tagging scheme for all organisational categories**

As above, the schema management module provides this capability. In addition to the standard default attributes for each category, the system also provides rules and exceptions for each category that reflect the requirements and feedback from users in each organisation. Also, trigger points have been identified as a requirement, so that some documents are tagged with details such as contract termination date, which in turns determines the retention policy. In addition, additional categories are enables to reflect authorised users' needs to additional categories.

3) **Either a human or software technology must read and assess all document content in order to make the proper categorical assessments**

The DataCube scans and identifies all of the files in the network and creates a data inventory of every relevant data file, which contains details about the file including its file path. It then uses this information to collect text from every file and build an index. Finally, it compares the text from every document against all of the exemplars to find the best correlation, when it marks the document against the category of the exemplar that it has matched against. This process has been extensively tested on Local Authority data over millions of documents and found to be over 85% accurate and with further retraining of the system, it is expected to be achieve over 90%. This compares with the ability for manual labelling which is believed to be less accurate and, given the volume of legacy data, impossible to do.

4) **End users must have a means by which to search for and find the documents they need quickly and efficiently**

LSI provides a highly accurate and consistent search mechanism that exceeds all other commercial options and as the index is held in memory, it is extremely fast, even when searching millions of documents. It was explained earlier how LSI compares against keyword searches, although the DataCube provides this as well. Other products use Bayesian techniques which are based on the probability of documents belonging to a category but LSI has been proven to be more accurate and provide a better performance. The DataCube therefore provides an interrogation

agent that enables text or whole articles to be pasted into the search box and rapidly and accurately matched against the data set. Search results can be further filtered by keywords and metadata searching to return a highly selective set of results that can be viewed individually or exported as a data set into other applications such as Excel, SQL Server, Sharepoint and other third party applications.

5) **The methodology must incorporate the proper checks to ensure accuracy and reduce categorical assignment errors. It must a consistently repeatable process**

   Whilst the document classification can be conducted as a one-off exercise and data tagged or moved into specified locations, the DataCube provides an Enforcement framework to ensure that the data stays in a managed state. Having built the initial index, newly created documents are dynamically added to the index without having to rebuild the system, although the index should be retrained periodically, the frequency of which is determined by the volume of new data created.

In addition to the points above, the DataCube assists the process of cleansing the data, such as identifying duplicate files and redundant data, as well as assisting to migrate the data into the Document Management system, meaning that the legacy data can be transferred and become usable in the new structure.

Enforcement of the Document Management Structure

Using the approaches outlined above, it has been shown how the DataCube can help overcome two of the major problems encountered during the implementation of an EDRMS, which are :

1) How to define a workable taxonomy for the Document Management system
2) How to migrate the data into the new Document Management system when the taxonomy has been defined

So what about the users ? How does an organisation ensure that the Document Management System continues to be used correctly ?

Audit of the EDRMS Data

Whilst the DataCube works to provide a categorisation system for unstructured data, it can also check the contents of document management systems such as Sharepoint and confirm that the data has been correctly placed into the EDRMS. The DataCube can analyse the metadata of the document as well as the content and can provide reports and checks that the two are accurate within the defined taxonomy.

Scan of Unstructured Data

Where data continues to be saved in an unstructured manner, the DataCube can continue to be used to migrate the document into the EDRMS but also to name and shame users who continue to behave in this way, leading to appropriate enforcement of the policy.

In-flight Analysis of the Content

A further aspect of the DataCube is its ability to analyse the content in realtime to analyse the concepts and compare it to the defined categorisation system. As the index is held in memory, this process is extremely quick and has no discernable delay for the user. This approach is used for a Data Loss Prevention module within the DataCube but in this context, it can be used to suggest to the user which category is appropriate for the document that they are using.

Do I need a Document Management System at all ?

That is not a decision we make as the DataCube works according to the requirements of an organisation and in our experience, many organisations want the confidence, structure, version control and security that a Document Management system can bring. As shown above, the DataCube supports the implementation and ongoing management of data in an organisation to support this.

However, there is a growing view that with the volume of unstructured data now being generated, existing Document Management Systems and more widely EDRMS systems will not cope or will be swamped with data that will make them unusable. In that scenario, a tool like to DataCube will be required to help create order out of potential chaos and may be the only way that this can be achieved. By implementing the DataCube, we are future proofing your decision.

Seeing is Believing..

To see a demonstration of the product or to discuss how it can be implemented in your organisation, contact Apperception at info@apperception.co.uk or call us on 0845 644 3479

Apperception Limited,
Unit 2A,
Berol House,
25 Ashley Road
London, N17 9LJ

Tel: 08456443479
Fax: 0871 288 1023
Email: info@apperception.co.uk
Website: www.apperception.co.uk