Chapter 15
Latent Semantic Analysis for Text Mining and Beyond

Anne Kao
Boeing Research & Technology, USA

Steve Poteet
Boeing Research & Technology, USA

Jason Wu
Boeing Research & Technology, USA

William Ferng
Boeing Research & Technology, USA

Rod Tjoelker
Boeing Research & Technology, USA

Lesley Quach
Boeing Research & Technology, USA

ABSTRACT

Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI), when applied to information retrieval, has been a major analysis approach in text mining. It is an extension of the vector space method in information retrieval, representing documents as numerical vectors but using a more sophisticated mathematical approach to characterize the essential features of the documents and reduce the number of features in the search space. This chapter summarizes several major approaches to this dimensionality reduction, each of which has strengths and weaknesses, and it describes recent breakthroughs and advances. It shows how the constructs and products of LSA applications can be made user-interpretable and reviews applications of LSA beyond information retrieval, in particular, to text information visualization. While the major application of LSA is for text mining, it is also highly applicable to cross-language information retrieval, Web mining, and analysis of text transcribed from speech and textual information in video.

DOI: 10.4018/978-1-61350-126-9.ch015

INTRODUCTION

A vast amount of information exists in text form, such as free (unstructured) or semi-structured text, including many database fields, reports, memos, email, web sites, blogs, and news articles. Various web mining and text mining methods have been developed to analyze textual resources. *Latent Semantic Analysis (LSA)* (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), or Latent Semantic Indexing (LSI) when it is applied to document retrieval, has been a major approach in text mining. It is an extension of the *vector space method* in Information Retrieval (Salton, Wong, & Yang, 1975), using a mathematical approach to represent documents as numerical vectors but with a more sophisticated means of characterizing the essential features of documents and reducing the number of dimensions needed to describe documents to a manageable size. There have been several major approaches to address this *dimensionality reduction*, each of which has strengths and weaknesses. A major challenge in using LSA is that it is typically considered a black box approach that makes it difficult to understand or interpret the results. However, more recent research has not only overcome this challenge, but also demonstrates that the use of LSA extends beyond information retrieval and text document clustering to become a major player in the area of text information visualization. This chapter will summarize the major approaches to LSA, their strengths and weakness, as well as recent breakthroughs and advances and applications beyond information retrieval.

Text mining has adopted certain techniques from the more general field of data analysis, including sophisticated methods for analyzing relationships among highly formatted data, such as numerical data or data with a relatively small fixed number of possible values. Such techniques can expose patterns and trends in this type of data. Text mining can identify relationships between individual unstructured or semi-structured text documents, as well as more general semantic patterns across large collections of such documents. Latent Semantic Analysis, like many other methods of text mining, depends on the twin concepts of "document" and "term." As used in this chapter, a "document" refers to any body of unstructured or semi-structured text. The text may include the entire content of a document in the general sense, such as a book, an article, a paper, or the like -- or only a portion of a document, such as an abstract, a paragraph, a sentence, or a title. Ideally, a "document" describes a coherent topic. In addition, a "document" can be the text field of a database, or encompass text generated from an image or graphic, or it may be text recovered from audio or video formats. We will use the term "document" in this general sense.

A document can be represented as a collection of "terms," each of which can appear in multiple documents. Typically, a "term" consists of an individual word used in the text. However, a "term" can also include multiple words that are commonly used together, for example, "landing gear", or even consist of a string that need not appear explicitly in the text but rather result from token normalization or standardization. Token normalization will be discussed further below.

In vector-based methods of text data analysis, after a suitable set of terms has been defined for a document collection, the collection can be represented as a set of vectors. With traditional vector space methods, individual documents are treated as vectors in a high-dimensional vector space in which each dimension corresponds to some feature of a document, typically a term. A collection of documents can thus be represented by a two-dimensional matrix $A_{(t,d)}$ of features (terms) and documents. In the typical case, the value of each matrix entry is the number of occurrences of that term in the specified document, or some weighting or principled transformation of that number. LSA, as an extension of the vector space method, involves methods of transforming A by various means, e.g. *singular value decomposition*