Decision Support Systems 62 (2014) 131–143

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Orthogonal rotations in latent semantic analysis: An empirical study

Lucian L. Visinescu [a,*], Nicholas Evangelopoulos [b,1]

[a] Department of Computer Information Systems & Quantitative Methods, McCoy College of Business, Texas State University, 601 University Dr., San Marcos, TX 78666, USA
[b] Department of Information Systems and Decision Sciences, College of Business, University of North Texas, 1155 Union Circle #311277, Denton, TX 76203, USA

ARTICLE INFO

Article history:
Received 10 March 2013
Received in revised form 6 March 2014
Accepted 26 March 2014
Available online 3 April 2014

Keywords:
Latent semantic analysis
Factor rotations
Varimax
Quartimax
Equamax
Big data
COALS

ABSTRACT

The Latent Semantic Analysis (LSA) literature has recently started to address the issue of interpretability of the extracted dimensions. On the software implementation front, recent versions of SAS Text Miner ® started incorporating Varimax rotations. Considering open source software such as R, when it comes to rotation procedures the user has many more options. However, there is a little work in providing guidance for selecting an appropriate rotation procedure. In this paper we further previous research on LSA rotations by introducing two well-known orthogonal rotations, namely Quartimax and Equamax, and comparing them to Varimax. We present a study that empirically tests the influence of the chosen orthogonal rotations on the extraction and interpretation of LSA factors. Our results indicate that, in most cases, Varimax and Equamax produce factors with similar interpretation, while Quartimax tends to produce a single factor. We conclude with recommendations on how these rotation procedures should be used and suggestions for future research. We note that orthogonal rotations can be used to improve the interpretability of other SVD-based models, such as COALS.

1. Introduction

In the last two decades, text mining [14] became an ubiquitous research area aiming at facilitating the comprehension of large quantities of unstructured text in order to support different types of business decisions with consequences for both academia and practitioners [45,35]. Originally conceptualized by H.P. Luhn, text mining evolved from simple abstracted principles, definitions and postulates [43], to non-fully automated systems [40] and semi-fully automated systems [20].

Multiple approaches were developed and proposed in order to respond to different needs of utilization of text mining. Latent Semantic Analysis (LSA) is a frequently used text mining method for concept extraction [6,8]. Other well-known concept extraction methods include Probabilistic Topic Modeling [3], Probabilistic Latent Semantic Analysis [18] and Non-negative Matrix Factorization [27]. The concept-based dimensions produced by LSA were found, in their raw format, hard to interpret [25,2]. In more recent years, research started to better address the interpretability issue [41,12] using Varimax rotations on the extracted factors. On the practitioner's side, recent implementations of SAS® Text Miner, beginning with the version 4.2 available in 2010, offer Varimax rotations that significantly improve the analyst's ability to interpret the LSA dimensions [37]. In this paper we further previous research on LSA rotations by introducing two other well-known orthogonal rotations, namely Quartimax and Equamax (also known as Equimax), and presenting an empirical study that compares the resulting rotated factors to those obtained through Varimax.

The paper is structured as follows. We start with a short introduction to LSA. We then present the principle of simplicity, which is the main idea underlying rotation in factor analysis. Subsequently we conduct an empirical comparative study on three orthogonal rotations and report the results. We conclude with a number of observations and recommendations related to the use of orthogonal rotations as part of LSA.

2. Latent semantic analysis

Latent semantic analysis [24], sometimes also referred to as latent semantic indexing [46], has recently gained increased attention in the academic community. A search using "Latent Semantic Analysis" or "Latent Semantic Indexing" terms in eight electronic databases, including Business Source Complete, Academic Search Complete, Library & Information Science Source, MEDLINE, Psychology & Behavioral Sciences Collection, Science & Technology Collection, Humanities Full Text, and SOCIndex Full Text, reveals the increasing trend in research interest shown in Fig. 1, where, for each year since 1989, we have counted all the articles that use LSA/LSI in their titles, abstracts, or keywords. LSA operates on textual data, which uses *words* as the most fundamental measurement unit. Building upon words, with the help of syntactic and grammatical rules, *sentences* are created. Combined sentences result in *paragraphs* that incorporate related ideas, concepts or topics. Multiple paragraphs create sections or chapters that eventually become *documents*. These documents are the most widely used forms in which

* Corresponding author. Tel.: +1 512 245 3801.
E-mail addresses: llv19@txstate.edu (L.L. Visinescu), evangeln@unt.edu (N. Evangelopoulos).
[1] Tel.: +1 940 565 3056.

http://dx.doi.org/10.1016/j.dss.2014.03.010
0167-9236/

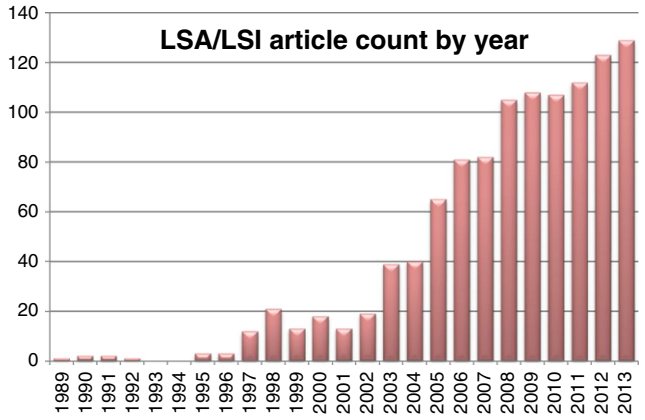

Fig. 1. Counts of LSA/LSI articles from 1989 to 2013 in eight electronic databases.

ideas, or concepts, or latent topics, are encountered. Finally, a collection of documents is defined in the information retrieval/text mining language as a *corpus*. This represents the space in which text mining methods are applied.

The rationale for using LSA, the mathematical foundations behind the method, and a series of examples in which the method was successfully applied, can be found in an edited volume published in 2007 [25]. The main motivation for using LSA is that the algorithm was found able to model human understanding of word meaning and perform humanlike activities such as: evaluation of multiple-choice tests results, modeling of paragraph to paragraph coherence, and improvement of information retrieval. In addition to being a computational algorithm, LSA is also perceived as a theory of meaning in which words and documents represent vectors in the LSA space. Thus, LSA is able to match similar ideas by mapping them on a system of coordinates of reduced dimensionality.

The foundation of LSA is the vector space model (VSM), which treats every document as a *bag-of-words*, eliminating the syntactic and grammatical structure of the text. In VSM, documents are represented as mathematical vectors in a multi-dimensional space where each unique term that appears in the corpus is a dimension. Comparisons between documents can be performed by computing the cosine of the angle formed by the respective vectors, also known as document similarity. When the cosine between two vectors is 0 it is said that the vectors are orthogonal therefore the documents have nothing in common, whereas if the cosine value is 1 the documents include identical sets of terms. A more detailed presentation on the vector similarities in the VSM can be found in the work of Salton and his colleagues [36]. The VSM was shown to have several limitations in dealing with the length of the documents, the order of the words in a document, the synonymy, and the polysemy of the words [11,26]. To cope with these limitations, various corrective approaches were introduced. These include the reduction of the space of words by eliminating very common terms (stop words), the stemming of terms to their roots, and the establishing of a cosine threshold for retrieving similar documents [17]. When LSA was introduced, its authors proposed it as a solution to address the synonymy and polysemy issues existing in VSM [11]. More recent studies continue this endeavor [49,16].

LSA starts with the creation of the VSM, represented by a typically large term-by-document matrix in which, initially, the frequency of a word appearing in all documents is inserted at the intersection of a row and a column. Most often, the resulting term-by-document matrix is sparse. That is, it has a big number of "zero" values, indicating that the average document has a small fraction of terms appearing in it. To cope with the sparsity and keep the size of the term-by-document matrix manageable, a series of adjusting measures are taken. For example, a stop-word list is used to eliminate very common terms ("a", "of", "but", etc.), which tend to appear in a large number of documents and therefore play a very small part in the formation of language use patterns. Next, a stemming procedure further reduces the number of words that enter the final analysis. After applying the stop-words and stemming, the term frequencies are weighted using local information from each document as well as global information from the entire collection of documents. One of the most commonly used weighting scheme is TF–IDF, where terms frequencies are locally not weighted ("TF"), but globally weighted using inverse document frequency ("IDF"). Another widely used weighting scheme is the Log-Entropy transformation.

In order to describe the mathematical operations of LSA, let us denote the term frequency matrix that results from the aforementioned steps as matrix A. Using singular value decomposition (SVD), matrix A is decomposed into term eigenvectors U, document eigenvectors V, and singular values Σ (see Fig. 2). The main premise of LSA is the option to reconstruct A by multiplying the matrices on the right-hand side of Fig. 2, without using all the available dimensions (factors), but by keeping only a few top dimensions. The number of dimensions retained for further analysis can be established based on threshold values for their corresponding eigenvalues, or manually selected by the researcher. The extracted dimensions can be associated with terms and documents through eigenvector matrices U and V, respectively, but, as previously mentioned in this paper, the dimensions may not be easily interpretable through this association.

To facilitate the labeling of LSA factors, Varimax orthogonal rotation, traditionally used in exploratory factor analysis, was introduced as an additional step following SVD [41]. However, there is a little research on the influence that various types of orthogonal rotation may have on the resulting factors. In addition to Varimax rotation, this study includes Equamax and Quartimax rotations. The main purpose of the study is to clarify the influence of the aforementioned orthogonal rotations on the resulting rotated LSA factors. We begin our investigation by presenting some related principles in the next section.

3. Simplicity and orthogonal factor rotations

3.1. The principle of simplicity

Invented more than 100 years ago to describe the influence of an unobservable factor called *g* (for general intelligence) on examinee's test scores on several domains, the single factor model (SFM) assumes that items form a homogenous set and measure just one common attribute [42]. The single factor model was later improved to cope with the functions of multiple latent variables and became known as common factor model (CFM) [44]. Based on the common factor model assumptions, factor analysis differentiates between exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). While EFA seeks to uncover the number of existing factor as well as the variable–factor relationship, CFA seeks to validate the factor structure which is a priori assumed to exist in a set of data, and evaluate the relationship between each factor. More on the philosophy of EFA can be found in Mulaik [30] and on the philosophy of CFA can be found in Jöreskog [21]. For the purposes of this paper, we observe that LSA is similar to EFA, as it attempts to uncover (explore) latent factors in unstructured textual data.

Several estimation methods generally used in EFA include the principal component method (PCA), unweighted least squares (ULS), generalized least squares (GLS), maximum likelihood (ML), principal axis factoring (PAF), alpha factoring (AF) and image factoring (IF). Nearly most of the time the extracted factors are orthogonal being ordered based on the amount of variance in the original data set explained by each factor [1].

To facilitate the interpretation of extracted factors that describe a subspace of the original data space (i.e., explaining less variance), two main types of rotations are traditionally used: orthogonal rotations and oblique rotations. The rationale for the use of orthogonal rotations is based on the principle of simple structure as proposed by Thurstone