Journal of Informetrics 5 (2011) 469–475

Contents lists available at ScienceDirect

Journal of Informetrics

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

Short Communication

The semantic mapping of words and co-words in contexts

Loet Leydesdorff*, Kasper Welbers

Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 23 November 2010
Received in revised form 3 January 2011
Accepted 18 January 2011

Keywords:
Semantic
Map
Document
Text
Word
Latent
Meaning

ABSTRACT

Meaning can be generated when information is related at a systemic level. Such a system can be an observer, but also a discourse, for example, operationalized as a set of documents. The measurement of semantics as similarity in patterns (correlations) and latent variables (factor analysis) has been enhanced by computer techniques and the use of statistics; for example, in "latent semantic analysis". This communication provides an introduction, an example, pointers to relevant software, and summarizes the choices that can be made by the analyst. Visualization ("semantic mapping") is thus made more accessible.

1. Introduction

In response to the development of co-citation maps during the 1970s by Small (1973; Small & Griffith, 1974), Callon, Courtial, Turner, and Bauin (1983) proposed developing co-word maps as an alternative to the study of semantic relations in scientific and technology literatures (Callon, Law, & Rip, 1986; Leydesdorff, 1989). Ever since, these techniques for "co-word mapping" have been further developed, for example, into "latent semantic analysis" (e.g., Landauer, Foltz, & Laham, 1998; Leydesdorff, 1997). These methods operate on a word–document matrix in which the documents can be considered as providing the cases (e.g., rows) to which the words are attributed as variables (columns).

Factor-analytic techniques allow for clustering the words in terms of the documents, or similarly, the documents in terms of the semantic structures of the words (*Q*-factor analysis). Singular value decomposition combines these two options, but is not so easily available in standard software packages such as SPSS. In this brief communication, we provide an overview and summary for scholars and students who wish to use these techniques as an instrument, for example, in content analysis (Danowski, 2009). A more extensive manual can be found at http://www.leydesdorff.net/indicators where the corresponding software is also made available. In this communication, we provide arguments for choices that were made when developing the software. Our aim is to keep the free software up-to-date, and to keep the applications as versatile and universally applicable as possible.

* Corresponding author. Tel.: +31 20 525 6598; fax: +31 84 223 9111.
E-mail addresses: loet@leydesdorff.net, l.a.leydesdorff@uva.nl (L. Leydesdorff).

1751-1577/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2011.01.008

2. The word–document matrix

The basic matrix for the analysis represents the occurrences of words in documents. Documents are considered as the units of analysis. These documents can vary in size from large documents to single sentences, such as, for example, document titles. The documents contain words which can be organized into sentences, paragraphs, and sections. The semantic structures in the relations among the words can be very different at these various levels of aggregation (Leydesdorff, 1991, 1995). Thus, the researcher has first to decide what will be considered as relevant units of analysis.

Secondly, which words should be included in the analysis? An obvious candidate for the selection is frequency of word occurrences (after correction for stopwords). Salton and McGill (1983), however, suggested that the most frequently and least frequently occurring words can be less significant than words with a moderate frequency. For this purpose, these authors proposed a measure: the so-called "term frequency-inverse document frequency," that is, a weight which increases with the frequency of the term i, but decreases as the term occurs in more documents (k) in the set (of n documents). The tf-idf can be formalized as follows:

$$\text{Tf-Idf}_{ik} = \text{FREQ}_{ik} \times \left[\log_2\left(\frac{n}{\text{DOCFREQ}_k}\right)\right] \tag{1}$$

The function assigns a high degree of importance to terms occurring more frequently in only a few documents of a collection, and is commonly used in information retrieval (Spark Jones, 1972). Given its background in practice, however, the measure has not been further developed into a statistics for distinguishing the relative significance of terms.

The proper statistics to compare the rows or columns of a matrix is provided by χ^2 or – using the Latin alphabet – "chi-square" (e.g., Mogoutov et al., 2008; Van Atteveldt, 2005). Chi-square is defined as follows:

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \tag{2}$$

The chi-square is summed over the cells of the matrix by comparing for each cell the observed value with the expectation—calculated in terms of the margin totals of the matrix. The resulting sum values can then be tested against a standard table. Both the relevant routines and the chi-square table are now widely available on the Internet; for example, at http://people.ku.edu/~preacher/chisq/chisq.htm. (If the observed values are smaller than five, one should apply the so-called Yates correction; the corresponding statistics is available, for example, at http://www.fon.hum.uva.nl/Service/Statistics/EqualDistribX2.html.)

Our programs – to be discussed below in more detail and available from http://www.leydesdorff.net/indicators – provide the user with the chi-square values for each word as a variable (by summing the chi-square values for cells over a column of the matrix) and additionally a file "expected.dbf" which contains the expected values in the same format as the observed values in the file "matrix.dbf." The user can thus easily compute the chi-square values using Excel.[1] Furthermore, the comparison between observed and expected values allows for a third measure which is easy to understand, albeit not based on a statistics, namely, the value of observed over expected (*obs*/*exp*). As in the case of chi-square, *observed*/*expected* values can be computed for each cell. However, these values can also be summed, for example, over the columns in order to enable the analyst to assess to which extent a word (as a column variable) occurs above or below expectation.

In summary, one can use four criteria for selecting the list of words to be included in the analysis: (i) word frequency, (ii) the value of tf-idf, (iii) the contribution of the column to the chi-square of the matrix, and (iv) the margin totals of observed/expected for each word. In case studies, we found this last measure most convenient. However, all four measures are made available.

3. The analysis

The asymmetrical word–document matrix – in social network analysis also called a 2-mode matrix – can be transformed into a symmetrical co-occurrence matrix (1-mode) using matrix algebra. This can be done in both (orthogonal) directions, that is, in terms of co-words or co-occurring documents.[2] The resulting matrix is called an affiliations matrix in social network analysis, and the transformation is standardly available in software for social network analysis (such as Pajek and UCINet). This new network is relational and allows, among other things, for the analysis of pathways. Pathways have been identified as indicators of inventions and innovations (e.g., Bailón-Moreno, Jurado-Almeda, Ruiz-Baños, & Courtial, 2005; Bailón-Moreno, Jurado-Almeda, Ruiz-Baños, Courtial, & Jiménez-Contreras, 2007; Stegmann & Grohmann, 2003).

The word–document matrix can also be analyzed in terms of its latent dimensions using factor analysis, multi-dimensional scaling (MDS) or singular value decomposition (SVD), etc. Note that factor analysis and SVD operate in the vector-space that is generated by first transforming the matrix using the Pearson correlation coefficients between the variables. In the vector space, however, similarity is no longer defined in terms of relations, but correlations among the distributions (vectors).

[1] Chi-square is not available for matrices in SPSS because SPSS presumes that the two variables have first to be cross-tabled.

[2] Technically, one multiplies the matrix (A) with its transposed as either AA^T or A^TA.