Document Indexing Techniques for Text Mining

José Ignacio Serrano
Instituto de Automática Industrial (CSIC), Spain

Mª Dolore del Castillo
Instituto de Automática Industrial (CSIC), Spain

INTRODUCTION

Owing to the growing amount of digital information stored in natural language, systems that automatically process text are of crucial importance and extremely useful. There is currently a considerable amount of research work (Sebastiani, 2002; Crammer et al., 2003) using a large variety of machine learning algorithms and other Knowledge Discovery in Databases (KDD) methods that are applied to Text Categorization (automatically labeling of texts according to category), Information Retrieval (retrieval of texts similar to a given cue), Information Extraction (identification of pieces of text that contains certain meanings), and Question/Answering (automatic answering of user questions about a certain topic). The texts or documents used can be stored either in ad hoc databases or in the World Wide Web. Data mining in texts, the well-known Text Mining, is a case of KDD with some particular issues: on one hand, the features are obtained from the words contained in texts or are the words themselves. Therefore, text mining systems faces with a huge amount of attributes. On the other hand, the features are highly correlated to form meanings, so it is necessary to take the relationships among words into account, what implies the consideration of syntax and semantics as human beings do. KDD techniques require input texts to be represented as a set of attributes in order to deal with them. The text-to-representation process is called text or document indexing, and the attributes and called indexes. Accordingly, indexing is a crucial process in text mining because indexed representations must collect, only with a set of indexes, most of the information expressed in natural language in the texts with the minimum loss of semantics, in order to perform as well as possible.

BACKGROUND

The traditional "bag-of-words" representation (Sebastiani, 2002) has shown that a statistical distribution of word frequencies, in many text classification problems, is sufficient to achieve high performance results. However, in situations where the available training data is limited by size or by quality, as is frequently true in real-life applications, the mining performance decreases. Moreover, this traditional representation does not take into account the relationships among the words in the texts so that if the data mining task required abstract information, the traditional representation would not afford it. This is the case of the textual informal information in web pages and emails, which demands a higher level of abstraction and semantic depth to perform successfully.

In the end-nineties, word hyperspaces appeared on the scene and they are still updating and improving nowadays. These kind of systems build a representation, a matrix, of the linguistic knowledge contained in a given text collection. They are called word hyperspaces because words are represented in a space of a high number of dimensions. The representation, or hyperspace, takes into account the relationship between words and the syntactic and semantic context where they occur and store this information within the knowledge matrix. This is the main difference with the common "bag of words" representation. However, once the hyperspace has been built, word hyperspace systems represent the text as a vector with a size equal to the size of the hyperspace by using the information hidden in it, and by doing operations with the rows and the columns of the matrix corresponding to the words in the texts.

LSA (Latent Semantic Analysis) (Landauer, Foltz & Laham, 1998; Lemaire & Denhière, 2003) was the first

one to appear. Given a text collection, LSA constructs a term-by-document matrix. The A_{ij} matrix component is a value that represents the relative occurrence level of term *i* in document *j*. Then, a dimension reduction process is applied to the matrix, concretely the SVD (Singular Value Decomposition) (Landauer, Foltz & Laham, 1998). This dimension-reduced matrix is the final linguistic knowledge representation and each word is represented by its corresponding matrix row of values (vector). After the dimension reduction, the matrix values contain the latent semantic of all the other words contained in all each document. A text is then represented as a weighted average of all the vectors corresponding to the words it contains and the similarity between two texts is given by the cosine distance between the vectors that represent them.

Hyperspace Analogue to Language (HAL) (Burgess, 2000) followed LSA. In this method, a matrix that represents the linguistic knowledge of a text collection is also built but, in this case, is a word-by-word matrix. The A_{ij} component of the matrix is a value related to the number of times the word *i* and the word *j* co-occur within the same context. The context is defined by a window of words, of a fixed size. The matrix is built by sliding the window over all the text in the collection, and by updating the values depending on the distance, in terms of position, between each pair of words in the window. A word is represented by the values corresponding to its row concatenated with the values corresponding to its column. This way, not only the information about how is the word related to each other is considered, but also about how the other words are related to it. The meaning of a word can be derived from the degrees of the relations of the word with each other. Texts are also represented by the average of the vectors of the words it contains and compared by using the cosine distance.

Random Indexing (Kanerva, Kristofersson & Holst, 2000) also constructs a knowledge matrix but in a distributed fashion and with a strong random factor. A fixed number of contexts (mainly documents) in which words can occur, is defined. Each context is represented by a different random vector of a certain size. The vector size is defined by hand and corresponds to the number of columns, or dimensions, of the matrix. Each row of the matrix makes reference to one of the words contained in the text collection from which the linguistic knowledge was obtained. This way, each time a word occurs in one of the predefined contexts, the context vector is summed up to the row referent to the word. At the end of the construction of the knowledge matrix, each word is represented as a vector resulting from the sum of all the vectors of the contexts where it appears. Other advantage relies on the flexibility of the model, because the incorporation of a new context or word only implies a new random vector and a sum operation. Once again, texts are represented as the average (or any other statistical or mathematical function) of the vector of the words that appear in it.

Unlike the previous systems, in WAS (Word Association Space) (Steyvers, Shiffrin & Nelson, 2004) the source of the linguistic knowledge is not a text collection but data coming from human subjects. Although the associations among words are represented as a word-by-word matrix, they are not extracted from the co-occurrences within texts. The association norms are directly queried humans. A set of human subjects were asked to write the first word that come out in their mind when each of the words in a list were presented to them, one by one, so that the given words correspond to the rows of the matrix and the answered words correspond to the columns of the matrix. Finally, the SVD dimension reduction is applied to the matrix. The word and text representations are obtained the same way as the systems above.

The FUSS (Featural and Unitary Semantic Space) system (Vigliocco, Vinson, Lewis & Garrett, 2004), the knowledge source also comes from human subjects. It is based on the state that words are not only associated to their semantic meaning but also to the way humans learn them when perceive them. Then, human subjects are asked to choose which conceptual features are useful to describe each entry of a word list. So a word-by-conceptual feature matrix is constructed, keeping the *k* most considered features and discarding the others, and keeping the *n* most described words by the selected features. Once the matrix is bounded, a SOM (Self Organizing Map) algorithm is applied. A word is represented by the most activated unit of the maps, when the feature vector which corresponds to the word is taken as the input of the map. The texts are then represented by the most activated units which correspond to the words that appear inside it.

In Sense Clusters system (Pedersen & Kulkarni, 2005), the authors propose two different representations for the linguistic knowledge, both of them matrix-like, called representation of first order and second order, respectively. In the first representations, matrix values