Chapter IV
SOM-Based Clustering of Multilingual Documents Using an Ontology

Minh Hai Pham
Swiss Federal Institute of Technology, Switzerland

Delphine Bernhard
Laboratoire TIMC-IMAG, France

Gayo Diallo
Laboratoire TIMC-IMAG, France

Radja Messai
Laboratoire TIMC-IMAG, France

Michel Simonet
Laboratoire TIMC-IMAG, France

ABSTRACT

Clustering similar documents is a difficult task for text data mining. Difficulties stem especially from the way documents are translated into numerical vectors. In this chapter, we will present a method that uses Self Organizing Map (SOM) to cluster medical documents. The originality of the method is that it does not rely on the words shared by documents, but rather on concepts taken from an ontology. Our goal is to cluster various medical documents in thematically consistent groups (e.g., grouping all the documents related to cardiovascular diseases). Before applying the SOM algorithm, documents have to go through several preprocessing steps. First, textual data have to be extracted from the documents, which can be either in the PDF or HTML format. Documents are then indexed, using two kinds of indexing

units: stems and concepts. After indexing, documents can be numerically represented by vectors whose dimensions correspond to indexing units. These vectors store the weight of the indexing unit within the document they represent. They are given as inputs to a SOM, which arranges the corresponding documents on a two-dimensional map. We have compared the results for two indexing schemes: stem-based indexing and conceptual indexing. We will show that using an ontology for document clustering has several advantages. It is possible to cluster documents written in several languages since concepts are language-independent. This is especially helpful in the medical domain where research articles are written in different languages. Another advantage is that the use of concepts helps reduce the size of the vectors, which, in turn, reduces processing time.

INTRODUCTION

In medicine, as in many other domains, text documents are not only numerous, but also written in many different languages. This can produce huge sets of multilingual medical documents that need to be exploited. There are many tools that can facilitate this exploitation. In general, searching for information is a very common task today for anyone who uses the Internet. Searches on the Internet are usually performed using some very popular and powerful search engines. However, results returned by these engines are presented as a very long list and the user still has to spend considerable time to verify if a result corresponds to her or his needs.

Document clustering is thus necessary to help cluster search results into groups of documents. Moreover, the groups must be labeled in order to help users choose the most suitable for their requirements. Among various clustering methods, Self-Organizing Map (Kohonen, 1982) seems to be the one that can best resolve this problem. On the one hand, it clusters documents in groups, on the other hand, it organizes groups on two-dimensional maps in such a way as to conserve the topology of the data structure. Moreover, Kohonen et al. (2000) proves that SOM can organize vast document collections according to textual similarities.

For clustering, documents have to be described by a set of features and values depending on the data representation model chosen. Some data representation models, such as word indexes are very common. However, the ontology-based method would appear to be appropriate for sets of multilingual documents. An ontology provides a mapping from terms to language-independent concepts. Furthermore, an ontology-based method can considerably reduce the dimensionality of input vectors that represent documents, since several terms may denote the same concept. This is extremely valuable because high dimensionality is a major problem in data mining in general.

Suppose that we have a corpus D of N documents and we want to cluster this set of documents into G groups of documents. Each document d_i, *with* $0 < i < N$ ($d_i \in D$) is represented by a list E of M semantic elements. M is the number of semantic elements in N documents. If the index j is the position of a semantic element in the list E then e_i, *with* $0 \leq j < M$ ($e_i \in E$) is the global frequency of this semantic element, i.e., its number of occurrences in N documents. The symbol $d_{i,j}$, *with* $0 \leq i < N$, $0 \leq j \leq M$ corresponds to the importance of the semantic element having index j in the document d_i. The group C_k, *with* $0 \leq k \leq G$ is a set of document indices that have been classified into this group. Note that, we use the concept "semantic element" to replace the concept "stem" or "term" or "concept" or "word category".