A VALIDATION STUDY OF "LEXIMAPPE"

L. LEYDESDORFF

Department of Science Dynamics, Nieuwe Achtergracht 166, 1018 WV Amsterdam (The Netherlands)

(Received October 7, 1991)

Clusters of normalized title-words in two sets of patent data in the food-sector (from 1985 and 1989, respectively) are analyzed in terms of their underlying document and word structures. The clusters were generated by using the system LEXIMAPPE of the Paris School of Mines. Both input and output data were kindly made available for validation purposes. Analysis of the data shows that the "centrality" and the "density" of the clusters produced by LEXIMAPPE are primarily dependent on the number of word occurrences in the corresponding parts of the input matrix. While the clusters are kept approximately equal in terms of the number of words (with a maximum of 10), they vary widely in terms of the number of word occurrences in the underlying document sets. "Centrality" and "density" vary correspondingly. The contribution of the smallest cluster to the reduction of uncertainty in the prediction of the document structure is even smaller than that of 77 (other) single words. In the dynamic analysis, I found significant stability where LEXIMAPPE indicated major changes. However, like every clustering algorithm LEXIMAPPE is based on specific assumptions which may lead to specific results that cannot be simulated by using other methods. Researchers who base their results on LEXIMAPPE should be aware of the peculiarities specific to this system.

Introduction

At the Paris School of Mines during the 1980s, an automated system for the scientometric analysis of co-occurrences of words was developed and extensively tested. The system is also known as LEXIMAPPE.\(^1\) Increasingly, results from LEXIMAPPE are used for the understanding of dynamic developments in science ("actor networks"),\(^2\) political processes related to technological innovation ("translation"),\(^3\) and as input to cognitively oriented science and technology studies.\(^4\)

In a previous publication,\(^5\) I raised the issue of whether one should use co-occurrences of words only, or rather analyze the distributions of words over documents (including, for example, tri-occurrences and single occurrences). Additionally, I argued in favour of the use of title words over keywords, because of the "indexer effect." The latter I specified both in terms of a (dynamic) delay and as a first taxonomy, i.e. clustering step. After subsequent debate on these issues,\(^6\) I was
invited by the Parisian group to study two datasets used as input to LEXIMAPPE, and kindly provided with full details about the data, processing and results.

The two datasets are observations of patent data in 1985 and 1989, respectively, which are given in terms of (normalized) title-words. The data was downloaded from the DERWENT-database by the Parisian researchers in the context of another project. The processing of the data in my study is strictly similar to that by the Parisian group.

In order to be able to relate and assess the input and output data of the system in one comprehensive framework, and additionally for the purpose of a dynamic comparison of the results for different years, I will use a set of information theoretical measures which I have elaborated in more detail elsewhere. I discuss the quality of the clusters produced by LEXIMAPPE in the two years separately, the relation of the results with the underlying data structures, and the comparison between the datasets for the two years.

**LEXIMAPPE**

LEXIMAPPE is based on single linkage clustering using as a similarity criterion the product of the relative co-occurrences for each pair of words. For each word \( A \), the probability of co-occurrence with word \( B \) is: \( f_{AB}/f_A \). Therefore, the similarity between \( A \) and \( B \) is given by what the makers of LEXIMAPPE call the equivalence index \( E \):

\[
E_{AB} = \left( \frac{f_{AB}}{f_A} \right) \cdot \left( \frac{f_{AB}}{f_B} \right)
\]

After computation of the value of \( E \) for all combinations of two words which are contained in the document set, the co-occurrence relations are sorted in decreasing order of \( E \). The clusters are essentially created by extracting one word as the centre of a cluster, after it occurred ten times in this listing. This word functions as the centre of a star-formed network, which by definition cannot contain more than ten words. The cluster is designated by the central word.

After extraction, words do no longer figure in the extraction procedure. Therefore, it obviously may happen that a core word for another cluster cannot function in that role, because it was already extracted in a previous round. In that case, in the system of LEXIMAPPE the later cluster will probably be taken as an extension of the former one, because after extraction the clusters are assessed, and if