QQML 2009
International Conference on
Qualitative and Quantitative Methods in Libraries
26 - 29 of May 2009 Chania Crete Greece www.isast.org

Document-Term Clustering Technique: Mining the Bibliographic Data for Finding the Hide Relationships in Disorder-Therapeutic Literature of Psychology

Mohammad Tavakolizadeh-Ravari[1] and Maryam Nejabatian[2]
[1] Yazd University, Yazd, IRAN, [2]National Youth Organization, Tehran, IRAN
mravari@yahoo.com

Document-Term Clustering is a classification method that is based on the term co-occurrence of documents within a corpus. The result of co-occurrence analysis is a table called proximity matrix. In this table, every row is labeled by a distinct term and follows by the count of its co-occurrence frequency with other terms, so that each case is considered a vector. The base of clustering is measuring the distance between these vectors.

To give an instance of how this technique works, we examined it to find the existent relationships between the psychological disorders and the therapeutic methods in biomedicine literature. To this end, Medical Subject Headings (MeSH) was searched for finding the therapeutic methods. The culled terms were queried by "OR" operator in PubMed/MEDLINE. It retrieved over 65.000 relevant bibliographic records. Saving them in a text-format made the data mining process by programming possible. The focus was on the descriptors (MeSH Headings) that those documents received by human indexers. The corpus entailed about 4.000 distinct descriptors. We ignored those with less than ten times frequency. About 100 of the most relevant descriptors up to the rest 2.000 were selected to determine the relationships among them.

Two fields of records were automatically transferred into a SQL Server Table: 1. the sequence number of records within the corpus, 2. their MeSH Headings. The co-occurrence count of each term with other recent culled terms was determined latter by Query Search Techniques. The process automatically followed by inserting the counts into a proximity matrix.

The proximity of descriptors was measured by Euclidean Distance Method (EDC).They were also clustered hierarchically based on the method suggested by Ward (1963).

The outcome revealed that we can classify the disorders into eight or nine segments. In addition, we could find out which disorders have not consistently neighbored with certain therapeutic methods. On the other hand, we discovered which disorder(s) was/were often taken into consideration with certain therapeutic method(s). Finally, we could determine how the segments correspond with each others.

Key words: Cluster Analysis; Data Mining; Subject Classification; Psychology, Disorders; Psychology, Therapeutics Methods