Comparison of different "window-size" key phrase co-occurrence for knowledge representation

Liu Jianhua[1], Alan Porter[2], Zhang Zhixiong[1], Guo Hongmei[1]

[1]*National Science Library, Chinese Academy of Sciences, Beijing,China*

Address: 33 Beisihuan Xilu, Zhongguancun , Beijing P.R.China ,100190.

Email: liujh@mail.las.ac.cn

[2]*School of Public Policy, Georgia Institute of Technology, Atlanta (USA), and Search Technology, Inc., Norcross, Atlanta, GA 30092, USA*

Abstract:

Word/phrase co-occurrence (if two words/phrases -- p and p' -- are seen in the same window, they are usually related) is a basic method to find the word/phrase associations for various bibliometric or informetric analyses. Such information can aid semantic association, topic identification, theme clustering, and knowledge structure profiling of a target domain record set. To construct an effective word/phrase co-occurrence matrix, there are two important factors -- word/phrase selection and suitable word/phrase co-occurrence window size identification. In this paper, the authors focus on most effective co-occurrence window size identification for knowledge representation through comparing five different window sizes.

To implement the comparison, the authors design data processing.

1. Pre-processing of data. (1) obtain "big data" related publication abstracts set (11684 records) from the Web of Science. (2) For co-occurrence matrix construction, use the same key phrase sets. Here, we use GATE for Automatic Key Phrase Extraction (extract the key phrase from the title and abstract, use some rules for phrase exclusion), and get the final term set. During this process, some Nature Language Processing results are recorded for the next step, such as the offset of each key phrase, sentence number that each key phrase occurs, and Part-of-Speech (POS) tagging.

2. Construction of five different co-occurrence matrices. Based on the same key phrase set, we construct five co-occurrence matrices according to different "window sizes": (1) Full text/paragraph size -- phrase co-occurrence in the whole title and abstract. (2) Sentence-wise -- key phrase co-occurrence in the same sentence. (3) Fixed window size -- it only considers phrases in a fixed window surrounding one phrase. In other words, for two phrases p and p', the co-occurrence of p and p' under this situation is set as the number of times both p and p' appear in the window. According to referenced research, the authors use co-occurrence within 5 meaningful tokens (removing the useless words such as "the\a\an" and so on from the title and abstract based on the POS information and some stopword lists). (4) Syntactic relationship -- it is a special co-occurrence of sentence-wise. It means there should be a syntactic relation between two phrases in one sentence. The authors obtain this information based on a syntactic dependency parser. (5) Semantic relationship -- words/phrases should have pre-defined semantic relationships, such as hierarchical relationship, synonyms, possessive, and so on. The authors used several dictionaries, such as verbNet.

3. Based on different matrices, the authors generate key phrase networks to check the effect for knowledge representation, with the help of experts checking the real situation about the results and comparing the results with a literature review.

This research is still in progress, and we expect to find the most effective window size for co-occurrence matrix construction.

Keywords: key phrase co-occurrence; co-occurrence matrix; window size; knowledge representation;