Short Text Clustering Based on Keyphrase

Popova Svetlana
Saint-Petersburg State University
Saint-Petersburg, Russia
svp@list.ru

DanilovaVera
Autonomous University of Barcelona
Barcelona, Spain
maolve@gmail.com

Abstract—**In presented paper we deal with narrow-domain short text clustering applied to annotations of scientific articles. Solution of this task is important for development of academic search systems to make them represent information in a structured way, which should reflect the domain of search query. Annotations usually contain short essential information about general contents of the article and they could be a good base for thematic clustering. However sometimes annotations could belong to the same general theme, but different sub-themes. In this case the clustering task becomes harder, because from one hand annotations have significant intersection in many theme related words, and from another hand number of words in annotations is quite small. Such problem could emerge in case a search system tries to present results of narrow-domain query as clusters or thematic groups. This paper investigates solution to the described problem based on keyphrases, which are extracted automatically for each text. In our paper keyphrases were extracted as sequences of nouns and adjectives. Additionally we used stop-words list, which was formed for scientific articles using additional collection of annotations. Two methods of clustering were used: k-means and hierarchic clustering (average link) in Weka implementation. Each document was represented as a vector of values in an attribute space. Two cases of document representation were investigated: in first keyphrases were used and in second single words from keyphrases were used. Obtained results were compared with results, obtained after clustering where each document was represented in a space of collection's full vocabulary ('standard method'). Weight evaluation for each phrase/word in text took place. For phrases weight was calculated like this: 1, if a phrase is contained in text and 0 otherwise. For words: weight was calculated with tf-idf. Cosine distance was used as a distance measure between documents. Three collections with narrow-domain short texts were taken for experiments: CICling 2002, SEPLN-CICling, EasyAbstracts. Experiments have shown, that for k-means quality increase comparing with 'standard method' is achieved when documents are represented with single words from keyphrases and especially with usage of stop-words list. For hierarchic clusterng 'standard method' shows best results.**

Keywords—clustering, narrow-domain short texts, key phrase.