TEXT CLUSTERING ON WEB WITH FREQUENT ITEMSETS - FICWAN

Tomáš KUČEČKA, Daniela CHUDÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
kucecka@fiit.stuba.sk

Abstract. Text data on the web contain additional information when compared to the standard texts. To this information belong references to other resources on the web and structural/style information in the HTML tags and CSS. In this paper we introduce a novel approach that utilizes this additional information for clustering textual data on the web – FICWAN. FICWAN is based on the well-known FIHC algorithm but in the clustering process it also considers neighbourhood of the clustered articles and style information. The experiments we performed clearly show that FICWAN performs better than FIHC.

1. Introduction

Text clustering has a special position in the field of clustering data because of its characteristics. It is highly affected by problems like high dimensionality and labelling of the created clusters. Therefore, standard approaches like hierarchical clustering or partitioning do not suite this problem very well [4]. More suitable approaches represent those based on frequent itemsets. In [2, 3] authors introduced FTC, HFTC and FIHC algorithms for text clustering. All of these work with frequent itemsets. In [3] authors showed, that FIHC outperforms FTC and HFTC. In recent years FIHC has become a very popular algorithm and its different modifications exist.

In this paper we introduce our own solution to text clustering that is based on FIHC - FICWAN. We modified the existing FIHC algorithm in order to utilize information that is specific for web articles. This is the styling information used in HTML language and the neighbourhood of a web article. We assume that the terms in the neighbourhood of a base document together with styles that visually emphasize words in base document are important for the term weighting process.

2. Web article clustering with FICWAN

We successfully implemented and carried experiments with the proposed clustering algorithm FICWAN. The whole principle of our solution consists of the following three steps:

1. Acquisition of documents found in the neighbourhood of the base documents. This acquisition is based on the hyperlinks contained in the base documents. When the documents are acquired we calculate *tf* weights of all of the extracted words and create a term-document matrix filled with these *tf* values.
2. Every word from a document neighbourhood is represented as a triplet *(term, HTML weight, CSS weight)*. The HTML weight and the CSS weight is calculated based on the HTML elements and CSS styling that is relevant for the word in this triplet. The words in the triplets are then preprocessed using the following methods: *stop-words removal*, *lemmatization*, *stemming* and *synonyms replacement*.
3. Clustering with FICWAN. Input to this step are term weights calculated in previous steps. The FICWAN algorithm first used FIHC algorithm to create initial clusters and then performs check on their quality. If needed, reclustering is performed.

In the experiments we compared the two clustering approaches, FICWAN and FIHC, on several datasets. The quality of returned clusters was measured by the F-measure. Overall FICWAN outperformed FIHC algorithm in all of the carried experiments. The most notable results were gained on the *Planetsave* dataset which contained 100 web articles about animals. Each article was about 580 words long. The F-measure value using FICWAN was 0.66 whereas the clusters returned by the original FIHC algorithm had only 0.09. The number of output clusters using FICWAN changed from 45 to 62.

Acknowledgement: This work was partially supported by the Scientific Grant Agency of the Slovak Republic, grant No. VG1/0971/11 and is the partial result of the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

to other papers publishing the results that are summarized here

[1] Kučečka, T., Chudá, D., Sládeček, P.: FICWAN Frequent Itemset Clustering of Web Articles by Analysing the Article Neighborhood. In: *Proc. of the 14th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, IEEE, (2013), [to be published].

Other references

[2] Beil, L., Ester, M., Xu, X.: Frequent term-based text clustering. In: *Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02)*, ACM, New York, USA, (2002), pp. 436-442.

[3] Fung, B. C. M., Wang, K., Ester, M.: Hierarchical Document Clustering Using Frequent Itemsets. In: *Proc. of the SIAM International Conference on Data Mining (SDM13)*, (2003), pp. 59-70.

[4] Yoo, I., Hu, X.: A comprehensive comparison study of document clustering for a biomedical digital library MEDLINE. In: Pr*oc. of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06)*, ACM, New York, USA, (2006), pp. 220-229.