Web Content Preprocessing for Clustering

Tomáš KUZÁR*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
kuzar@fiit.stuba.sk

There is a huge amount of unstructured content (e.g. blogs, comments) available on the internet. Unstructured data needs to be put into structured representation in order to apply the machine learning techniques. Transformation of unstructured information into structured representation is called preprocessing.

In our paper we focus on preprocessing phase of web content clustering. We evaluate the impact of different data preprocessing methods on success of blog clustering. We found out that applying various text data manipulation techniques in preprocessing can significantly improve the quality of clusters. The quality of clusters is measured by traditional clustering metrics called F-measure.

Data preprocessing consists mainly of term extraction and term selection. Basic term extraction can be provided by tokenization where the terms are delimited by whitespaces. We created list of terms which were filtered before the term extraction process started. The list includes mainly conjunctions, prepositions or pronouns which were filtered during the preprocessing phase.

Most of the languages use suffixes and infixes of the terms. Normalization extracts the bases of the terms. Normalization process is language dependent. Widely used stemmer for English language was developed by Martin Porter. There are various mutations of stemmers developed and used. Stemming removes affixes from words algorithmically and converts the term to its stem. Slovak language uses high number of suffixes. We created a method, which can be considered as basic stemmer for Slovak language, for grouping lexically similar terms into one term. We calculated lexical similarity on terms longer as three characters. If two terms are equal on more on 75% of term length, they are mapped on some lexical term.

Lemmatization for each inflected word form in a document or request, its basic form, the lemma, is identified. Disadvantage of lemmatization is that the terms the dictionary does not contain cannot be lemmatized.

We used Slovak dictionary for lemmatization with more then 100 thousand lemmas. But the problem of lemmatization dictionary is the ambiguity of the word

* Supervisor: Pavol Návrat, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 79–80.

forms. E.g. lemma for word form ‘je’ can be ‘byť” and ‘jesť‘ as well. We fix the ambiguity using just one-one replacement.

Usage of normalization technique is not only language dependent but also application domain and machine learning method dependent according to [2], [3]. Different techniques can be applied which can be applied alone or in combination. In our research we use Eurovoc[1] taxonomy. Taxonomy based term extraction searches for Eurovoc terms in articles and after exact matching replaces the term from the article by more general term in Eurovoc hierarchy.

Usually the number of extracted terms is too high and just most important terms need to be selected. Several term selection methods could be applied. In our experiments we use LDA probabilistic topic model [1], where we represent document as LDA topics. One of the outputs of the LDA model is an article-topic matrix. This article-topic matrix was the input for KNN algorithm in clustering task where we set the number of cluster we want to build.

While information gathering process we downloaded not only articles but also information about count of the discussion posts. We suppose that information about count of the discussion posts can improve the quality of clusters. Some topics have higher average discussion count then some others. We added information about discussion count into article-topic matrix.

We applied the steps mentioned above – term extraction, LDA model building, KNN clustering and clustering metrics calculation (F-measure) – several times using of different term extraction methods or combination of these term extraction methods.

In this paper we focused on term extraction methods applied in preprocessing phase of text mining. Our experiment consisted of several steps: term extraction, LDA model based term selection, clustering and F-measure based evaluation. We used some combinations of term extraction methods and we found out that only lemmatization has always enhanced the quality of clusters. As future work we want to focus on some other term extraction methods – term extraction based on named entity recognition and term extraction method enriched by extensive knowledge of text source.

References

[1] Blei, M., Ng, A., Jordan, M.: Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, 2003.

[2] Korenius, T.: Stemming and lemmatization in the clustering of finish text documents. In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM Press, (2004), pp. 625-633.

[3] Sun, L.: User-driven development of text mining resources for cancer risk assessment. In: *Proceedings of the Workshop on BioNLP*, ACM Press, (2009), pp. 108-116.

[1] http://europa.eu/eurovoc