A word-based distance for supervised learning over e-commerce shop items

Zoltán Ács, Zoltán Vincellér

Eötvös Loránd University, Department of Information Systems

acszolta@inf.elte.hu, vzoli@inf.elte.hu

Nowadays, product suggestion systems and solutions became more and more important in e-commere solutions. There are many reasons for this popularity. On the one hand, traders want to personize their offers for their customers to increase their income. On the other hand, customers want to find the best offers as fast as it possible. There exist many different solutions, e.g. hierarchical classification [1] or Support Vector Machines [2]. The former is used by ebay.com. A common property of these solution is that we can handle them as a classification problem over a large amount of text contents.

Generally, text classification problems are based upon a string compersion problem, which suppose that we have a proper metric for meassuring the distance. In these fields of science, there are a lot of suggested distances ready to use, e.g. Hamming distance, Damerau-Levenshtein distance or edit distance [2]. In this paper, we introduce a $d_{cond}(s_1, s_2, c)$ distance which combines the edit distance between s_1 and s_2 and the available statistical informations over c cathegory. We will show the benefits and drawbacks of this distance on an e-commerce product database, which is a collection from three public database, namely e-Bay.com, Amazon.com and theFind.com.

Here, we will focus on the properties and the "goodness" of the introduced distance, but we will also discuss some special matters and their soultions, e.g. the occurence of popular words, effect of recurrences and abbreviations or handling mismatch. We will show that this extended edit distance can also help and improve the correctness of a classification over an arbitrary item set where the items are described by text labels.

References

[1] Dan Shen,Jean-David Ruvini, Badrul Sarwar; Large-scale Item Categorization for e-Commerce

[2] Thorsten Joachims; Text Categorization with Support Vector Machines: Learning with Many Relevant Features

[3] William J. Masek, Michael S. Paterson; A faster algorithm computing string edit distances