Use of Text Mining Methods in a Digital Library

Jiří HYNEK[1], Karel JEŽEK[2]

[1]*inSITE, s.r.o., Knowledge Management Integrator*
Rubešova 29, 326 00 Pilsen, Czech Republic
jiri.hynek@insite.cz

[2]*Dept. of Computer Science & Engineering, University of West Bohemia*
Univerzitní 22, Pilsen, Czech Republic
jezek_ka@kiv.zcu.cz

Abstract

The article deals with use of Itemsets classifier based on inductive machine learning in the context of digital library environment. We provide a brief description of a real-world digital library implemented at a power utility. Its implementation and operating experience have motivated our research in inductive machine learning methods for text mining described in the paper.

Being inspired by mining of association rules, we have developed a new categorization method named "Itemsets classifier". By performing various experiments we have proved its ability to surpass some well-known categorization methods, both in terms of precision/recall and efficiency. As the task of classification is closely related to clustering, we have integrated the principles of Itemsets method into a new document-clustering algorithm as well. We are also presenting other Itemsets classifier applications in unsolicited mail filtering and enhancement of the Naïve Bayes classifier. Main ideas and experimental results are presented in the paper.

Partly supported by grant No. MSM235200005