Email labeling by rough clustering*

László Aszalós, Mária Bakó, Tamás Mihálydeák

University of Debrecen
laszalos@unideb.hu, bakom@unideb.hu
mihalydeak.tamas@inf.unideb.hu

Abstract

Previously one did not have many possibilities to sort mails, and later emails: we could only arrange them into folders. One mail or email could be put into exactly one folder, we could sort it based on sender, subject or priority into different folders. Later in Gmail the labeling of emails was introduced: one email could get several labels, and virtual folders were generated by these labels. This kind of labeling was taken by other mailers, photo and music organizer software, too.

It is a pleasure to use a well organized collection, but usually a big pain to set up the labeling of it. We undertake to simplify these kind of tasks by using our experience in rough set theory and clustering. The clustering is a well-known part of the data mining, where the elements are grouped by their similarity. The similarity is an inexact concept in real life, e. g. we easily mix up two Japanese persons, however a Chinese man easily differentiates them. The methods of data mining usually uses the concept of the distance of objects. But what is the distance of two persons, or two emails? Each parameter of the objects could be quantified, but they could have different dimensions, hence the objects are not easily comparable.

The correlation clustering [1] we used does not depend on the concept of the distance, but on a similarity relation. This relation is partial, so it is possible that it does not state anything about similarity of two objects. Moreover it could state that objects are similar, or dissimilar.

Based on a such similarity relation the correlation clustering produces a near-optimal partition. To find the best partition according to the similarity relation is a NP-complete problem, therefore it cannot be aimed at real life problems. Running the method several times we could get different partitions. The best ones can be combined by concepts of the rough set theory [2]. Two objects are tightly similar, if they are together in all the best partitions. Therefore we can label one of them with the labels of the other. Two objects are weakly similar, if they occur at least once together in some of the best

*The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

partition. This relation is not an equivalence relation – unlike the strong one – hence in some cases we use its transitive closure. Then there is a reason to check manually that a label of an object is suitable for some weakly similar object, or not.

We remark that the relations' strong and weak similarity is based on the initial relation. By changing it, the resultant relations are changed, too. In this article we present the mathematical background and the application of the method for labeling emails.

Keywords: rough set, correlation clustering, data mining

MSC: 03E02, 62H30, 90C27

References

[1] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, July 2004.

[2] Zdzisław Pawlak. Rough set approach to knowledge-based decision support. *European journal of operational research*, 99(1):48–57, 1997.