An Augmentation Hybrid System for Document Classification and Rating.

Richard Dazeley and Byeong-Ho Kang

School of Computing, University of Tasmania, Hobart, Tasmania 7001, Australia.[1]
Smart Internet Technology Cooperative Research Centre, Bay 8, Suite 9/G12
Australian Technology Park Eveleigh NSW 1430[1]
{rdazeley, bhkang}@utas.edu.au

Abstract. This paper introduces an augmentation hybrid system, referred to as Rated MCRDR. It uses Multiple Classification Ripple Down Rules (MCRDR), a simple and effective knowledge acquisition technique, combined with a neural network.

Introduction

As we move from the Information Age to the Age of Information Overload, Information Filtering (IF) has gained significant attention in the research community. This paper briefly introduces a new method based on a variant to the Multiple Classification Ripple Down Rules (MCRDR) methodology, called Rated MCRDR (RM) [1]. Rated MCRDR is an augmentation hybrid intelligent system developed to provide both classifications and a relevance ranking of cases and can be applied in many domains [1]. One of the key areas that the algorithm was designed for was information filtering and in fact draws heavily on ideas found in the information filtering research. The main idea behind the system is to significantly reduce the feature space, so that it is of a size that a neural network is capable of handling, in such a way that we don't effectively loose any relevant information.

Rated MCRDR (RM)

To achieve this, RM adopted the basic premise that while the majority of features may be statistically relevant [2] it is safe to assume that an individual user is not interested in all the possible features. Therefore, RM attempts to identify keywords, groups of words, phrases or even compressed features, outputted from some other feature reduction method, by using simple user interrogation, by using the Multiple Classification Ripple Down Rules (MCRDR) [3]. This incremental Knowledge Acquisition (KA) methodology allows a user to perform both the KA process and the maintenance of a Knowledge Based System (KBS) over time [3]. The basic concept

[1] Collaborative research project between both institutions

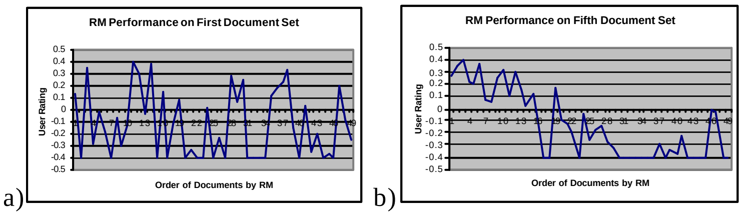

Fig. 1. Ability of RM to order cases according to the simulated-user's preference. a) Shows RM's performance prior to any training. b) Shows RM's performance after 5 document sets.

behind MCRDR is to use the user's knowledge within the context it is provided [1, 3] to produce multiple classifications for an individual document. Therefore, if the expert disagrees with one or more of the conclusions found by the system, knowledge can be easily added to improve future results.

It then learns further information, through observing user behaviour, about the relationships between groups of identified features to capture a deeper sociological meaning behind the selected features as well as to associate a set of relevance rankings. When a new feature or set of features are identified by the user, the specifically designed neural network steps to a rating that accurately identifies its relevance to the user immediately. After the initial learning step, any further documents receiving the same classification allow the network to learn more intricate non-linear relationships. Thus, RM has the ability to learn both classifications for documents if required, as well as being able to learn both linear and non-linear ratings effectively. The remainder of this paper will discuss RM in detail.

Results and Discussion

The system has undergone preliminary testing with a simulated expert using a randomly generated data set. Figure 1, illustrates how RM was able to place the documents with a higher relevance to the user first after only seeing 5 groups of 50 documents. These tests were done primarily to show that the system was able to learn quickly and to be used for parameter tuning purposes. Clearly a more rigorous testing regime needs to be used in order to fully justify the algorithm's ability to learn within the information domain.

References

1. R. Dazeley and B. H. Kang. Rated MCRDR: Finding non-Linear Relationships Between Classifications in MCRDR. in *3rd International Conference on Hybrid Intelligent Systems*. 2003. Melbourne, Australia: IOS Press
2. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. in *European Conference on Machine Learning (ECML)*. 1998: Springer
3. B. H. Kang, Validating Knowledge Acquisition: Multiple Classification Ripple Down Rules. 1996, University of New South Wales: Sydney.