

# Enhancing Clustering Methods by Inserting Relational Information

**Tijn Witsenburg**  
LIACS, Universiteit Leiden  
Leiden, The Netherlands  
tijn@liacs.nl

**Hendrik Blockeel**  
LIACS, Universiteit Leiden  
Leiden, The Netherlands  
blockeel@liacs.nl

## Abstract

We consider the problem of clustering elements that have both content and relational information (e.g. Web-pages, scientific papers, etc.). Standard clustering methods use content information only, while graph clustering methods are usually based on the graph structure. Relatively recently, researchers have proposed to combine both types of information. In this paper we propose a simple, yet hitherto unexplored, method to insert the relational information into standard clustering methods.

## 1 Introduction

Clustering is an important task in machine learning. Two different types of clustering are distinguished. The first is **standard clustering** where items are grouped in the same cluster when their content has a high similarity. The second is **graph clustering** where a graph is divided into subgraphs so that the nodes in a subgraph are highly connected but there are as little as possible connections between different subgraphs. The question raises how these two types of information can be combined to improve clustering methods for appropriate data sets, i.e. data sets that have both content and relational information.

Neville et al. (2003) discuss this problem and a number of solutions. In the combined method they propose they take the structure of the graph describing the relational information and the edges in this graph are given weights that correspond to the similarity between the nodes they connect and then a graph clustering method is applied to them. Neville et al. compare different graph clustering methods.

In this work an opposite direction is used. Here the relational information is inserted into the similarity function, after which a standard clustering

method is used. It could be said that Neville et al. map the hybrid clustering task onto graph clustering, whereas here it is mapped onto standard clustering.

## 2 The Method

The elements to be clustered are the set  $V$  with  $|V| = n$  and  $v_i \in V$  where  $1 \leq i \leq n$ . For this set two  $n \times n$  matrices can be created:  $S$  and  $W$ .  $S$  is the matrix where each element  $s_{ij}$  is the similarity between  $v_i$  and  $v_j$ .  $W$  is the matrix where each element  $w_{ij}$  is the weight of the edge between  $v_i$  and  $v_j$  in the graph describing the relational information. When there is no edge, the weight is 0. A special case is when the edges are unweighted, then every element  $w_{ij}$  is either 0 or 1. An element  $v_i$  is considered a neighbour of  $v_j$  when  $w_{ij} > 0$ . Standard clustering would use the values in  $S$ . Instead of  $S$  a new matrix  $M$  is created by:

$$m_{ij} = c_1 \cdot s_{ij} + c_2 \cdot asn_{ij} \quad (1)$$

where  $c_1$  and  $c_2$  are constants defining the ratio between the two factors  $S_{ij}$  and  $asn_{ij}$  which are respectively the similarity and the **average similarity of the neighbours** between  $v_i$  and  $v_j$ . This  $asn$  can be seen as the average of the similarities between  $v_i$  and all neighbours of  $v_j$  and the similarities between  $v_j$  and all neighbours of  $v_i$ . Since this can be calculated in two ways, two types of  $asn$  are distinguished:  $asnI$  and  $asnII$ :

$$asnI_{ij} = \frac{\sum_{k=1}^n s_{ik} \cdot w_{kj} + \sum_{k=1}^n s_{jk} \cdot w_{ki}}{\sum_{k=1}^n (w_{kj} + w_{ki})} \quad (2)$$

$$asnII_{ij} = \frac{1}{2} \cdot \left( \frac{\sum_{k=1}^n s_{ik} \cdot w_{kj}}{\sum_{k=1}^n w_{kj}} + \frac{\sum_{k=1}^n s_{jk} \cdot w_{ki}}{\sum_{k=1}^n w_{ki}} \right) \quad (3)$$

Both equations consist of two parts. One part is the average similarity between  $v_i$  and the neighbours of  $v_j$  and the other part is the average similarity between  $v_j$  and all neighbours of  $v_i$ . The difference between the two equations is in the way how they

calculate the final value as an average of the two parts. *asnI* calculates this final score while using the ratio of the total weight of the edges between one element and its neighbours while *asnII* calculates this final score without regarding this ratio. After the new matrix  $M$  is created, these altered similarities can be used in a standard clustering method. In this way existing standard clustering methods can be enhanced as to include also relational information.

### 3 Expectations

#### 3.1 First Results

For this method already some first experiments have been performed. For this, subsets of the Cora data set (McCallum et al., 2000) were used. This is a big data set with scientific papers, including their abstracts and the citation graph, which are divided in several classifications based on their subject. The abstracts were used to compare the content of the papers. The citation graph, where an edge between two papers is inserted when one cites the other, is used for the relational information. First experiments on this data set show a small but significant improvement when the relational information is inserted in the clustering method.

#### 3.2 Advantages and Disadvantages

The big advantage of course is the fact that it allows standard clustering methods to include previously unusable information. This extra information can improve the performance of the clustering method. Another advantage is the fact that it can be used on a variety of already existing standard clustering methods.

Unfortunately this method also has its disadvantages. The first is the fact that the elements lose their location. For instance, in K-means clustering (Steinhaus, 1956) all elements are given a location in a highly dimensional vector-space, which is very important for a proper working. With the proposed procedure the location is lost, hence makes it impossible to use K-means. This problem may be solved by using multi-dimensional scaling. Given the new distances, it relocates the elements of the data set. These new locations could be used by K-means. Whether this solution actually enables a accurate use of the K-means algorithm still needs to be researched.

The second disadvantage is the fact that this procedure has the same complexity as matrix multipli-

cation. This means that it is not suited for very large data sets since then it will take too much time to calculate  $M$ . On the brighter side will this procedure in practice only be used on very sparse graphs, which limits the effect of this problem.

Another problem of using this procedure on a very large data set is the fact that the size of the matrices that need to be in the computers memory will become fairly large. This will at some point limit the size of the data set.

### 4 Future Work

While the first results are promising, still a lot of research needs to be done. Concerning the accuracy of the performance of the procedure, it should be tested on other data sets and for other clustering methods. Also the results need to be compared with the solution of Neville et al. and there are several aspects of this procedure that can be analysed better. These are, for example: ‘What ratio between  $c_1$  and  $c_2$  is best?’, ‘Which is better: *asnI* or *asnII*?’, ‘Would it be wise to also include the neighbours of the neighbours?’, etc.

Besides the accuracy, also subjects as efficiency and scalability should be tested more extensively. Finally, it can be very interesting to research if this procedure can also be used to enhance data mining methods for classification. In principle any distance-based method should be able to exploit the proposed procedure.

### Acknowledgements

Hendrik Blockeel is a postdoctoral fellow of the Fund for Scientific Research of Flanders (FWO). This research is funded by the Dutch Science Foundation (NWO) through a VIDI grant.

### References

- Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163.
- J. Neville, M. Adler, and D. Jensen. 2003. Clustering relational data using attribute and link information. In *Proceedings of the Text Mining and Link Analysis Workshop, Eighteenth International Joint Conference on Artificial Intelligence*.
- H. Steinhaus. 1956. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci., C1. III*, IV:801–804.