Chapter 10

Conclusions

This chapter presents all the conclusions from this thesis.

10.1 Problem Setting

Data mining techniques are designed for a specific type of data. However nowadays, many datasets consist of a combination of different types of data. A data mining technique that is created for one such type can not use all the information in this dataset. This leaves the question of whether it is possible to create a technique that enables data mining techniques designed for one type of data, to also use the information from the other data type, albeit indirectly.

We proposed a method to create hybrid similarity measures. These are measures that alter a similarity measure based on the content of the elements, so that it also includes relational information from this dataset. Theoretical analyses suggest that these hybrid similarities use the homophily in the dataset to compensate for the content variability in it. The experiments described in Part II *Implementations* examine whether it is possible to implement the hybrid similarities succesfully. The experiments described in Part III *Analysis* examine what the reasons could be for the hybrid similarities to outperform the original, content-based similarity. In the following, we summarize the main conclusions from this research.

10.2 Conclusions from Implementations

The hybrid similarities easily can be implemented on data mining techniques that only use a similarity measure between elements (e.g. agglomerative hierarchical clustering, k-medoids, and KNN-classification). Here, the hybrid similarities indeed enhance the performance of the data mining technique when they are used on subsets of the well-known Cora dataset.

Implementing the hybrid similarities with k-means is not straightforward since a similarity between an element and a prototype for a cluster, which is outside the dataset, is not defined. However, it can be approximated by two newly proposed methods. These two methods boil down to using approximate similarity measures so that k-means becomes applicable. For these approximate versions of k-means, the following holds:

- Despite the fact that the hybrid similarities can only be applied to approximate versions of k-means, this setting still will lead to better results.
- Of these approximate versions of k-means, there is no significant difference in the quality of the found clusterings, but k-means-NAMA is much faster than k-means-NAM.
- The more sophisticated versions k-means-NAM(A) were worth developing, as they work better than the more straight-forward approach of applying k-medoids.

10.3 Conclusions from Analysis

The reason that the hybrid similarities improve the performance of these data mining techniques is indeed related to the fact that the homophily in the network compensates for the content varibility. The theoretical arguments for this, presented in Chapter 4, were confirmed experimentally in Chapter 8.

Experiments on a synthetic dataset show that the hybrid similarities consistently outperform the content-based similarity when the dataset has much homophily and much content variability.

Theoretical analyses show that there are five conceptual possibilities to combine the similarities from one element to the neighbors of the other in order to create a hybrid similarity. Experiments on subsets of Cora show for agglomerative hierarchical clustering, k-medoids and KNN-classification the following:

- The five hybrid similarities outperform the original, content-based similarity in every setting.
- There is no significant difference between the performance of any of the five hybrid similarities.

All in all, it can be concluded that, when there is sufficient homophily and content variability, replacing the content-based similarity with a hybrid similarity can lead to improved performance for data mining techniques that use a content-based similarity measure.