Thesis Summary:
Named Entity Disambiguation in Digital Libraries

Le Dieu Thu
Supervisor: Raffaella Bernardi, Massimo Poesio, Patrick Blackburn

In Digital Libraries, ambiguous author names may occur due to the existence of identical names, name misspellings, pseudonyms. Disambiguating these author names is a major problem during data integration and document retrieval.

The main goal of our study is to obtain a disambiguation method that could be applied in any catalogue even if no accurate information about the topic of the records is provided, as it happens in federated portals or digital libraries in which the manual annotations by librarians cannot be exploited. To this end, we look at the author name disambiguation as a clustering problem and propose a framework for disambiguating author name by taking advantage of a large scale external data set, Wikipedia. Using a hidden topic analysis model from this external data set, we automatically enrich topics for each record in the library. By expanding them with hidden topics, we have decreased their vocabularies' difference and improved the clustering quality by taking into account their latent semantic relations. In cluster analysis, we follow the Hierarchical Agglomerative Clustering (HAC) approach since it does not require the estimation of the number of clusters and thus it fits our needs. We experiment with different cutting points to optimize the $F1$ score, a measure of a test accuracy.

Furthermore, to deal with the two problems of feature representation in high dimensional space that happen in our framework, i.e., sparsity problem and visualization for better quantitive analyses of the results, we use the dimensionality reduction technique *PCA* to represent the data in a more compact way.

We evaluate our framework through extensive experiments in the Bolzano Library catalogue: first of all, we exploit as much as possible the features available in all catalogues (i.e., co-author names, titles, publishers) and set it as our baseline (1); then, compare the clustering results based on the manual information added by the librarians (Subject Headings and Classification Numbers) (2) against the clustering results based on feature information extracted automatically via hidden topic analyses (3). It is shown that exploiting the information of Subject Headings and Classification Numbers (2) improves the result of the baseline significantly, and that similar performance is achieved by exploiting instead the features extracted automatically in (3) ($F1$ score increases from 57% to 62% and reduces 12% errors). Therefore, clustering algorithms can be used also in scenarios as the federated portals or the digital libraries in which manual annotations of the librarians are not of support. Finally, it is also shown that after reducing the number of dimensions using *PCA*, *HAC* still achieves an adequate accuracy while reducing the complexity of the clustering process. This compact representation can also be further exploited to visualize the data in a more intuitive way for better quantitive analyses of the results in the future.

Key words:
Named Entity Disambiguation, Digital Libraries, Topic Modeling, Cluster Analysis, Author Name Disambiguation, Dimensionality Reduction