Knowledge extraction and processing approaches for structured and unstructured data.

Habilitation Thesis

Mihaela Dînșoreanu, PhD

Abstract

The present habilitation thesis presents the professional activity of the candidate and the scientific activity that was conducted after the defence of her PhD thesis at the Technical University of Cluj-Napoca on 9.07.2004 followed by the title confirmation by the Ministry of Education and Research Order nr. 445/2.08.2004. The research activity related to the PhD thesis was concerned with the design of agent models and multi-agent systems involving cooperating agents. The models were applied in an e-learning system. After finishing the PhD thesis, the candidate remained interested and active in the artificial intelligence field being involved in several national projects. Areas that were addressed in the early projects are: data mining solutions applied on medical data, knowledge extraction and representation solutions applied on archival documents written in natural language. More recently, the research interests of the candidate was focused on various aspects related to semantic data representation and integration, knowledge extraction by employing various machine learning techniques and knowledge processing.

a. *Data representation and semantic integration*. The problem that was addressed by this work is the semantic alignment of data imported from different structured data sources (i.e. relational databases, XML files, CSV files), the automatic design of a unified integration data structure and the automatic design of the corresponding ETL processes. In this respect we developed a solution based on the data sources metadata, synonymy relations from WordNet and the Jaro-Winkler lexical similarity algorithm. We developed a semantically annotated repository that is used in the semantic merging algorithm to produce information as accurate as possible for heterogeneous data integration and dynamic generation of ETL processes. We performed experiments on different sizes of data source structures in order to evaluate the scalability and performance of the solution.

b. *Knowledge extraction*. In the problem of knowledge extraction our latest interests were focused on unstructured data sources, namely natural language texts. In this respect we aimed for topic identification in order to define context for context-sensitive recommendations. Another objective we addressed was opinion mining. For these objectives we explored several dimensions: supervised vs. unsupervised solutions, domain dependent vs. domain independent solutions, language specific solutions for English and Romanian. We proposed a unified topic model based on the Latent Dirichlet Allocation approach for defining the context and also a set of contextual and behavioural distance metrics in order to match content (recommendations) to context. For the opinion mining problem we proposed a processing flow and a learning approach based on an original set of meta-features of text elements (i.e. POS combinations as bigrams). The proposed supervised solution performed very well on English documents and we adapted the solution by employing the Romanian grammar rules to Romanian documents. Another approach that we proposed for the same opinion mining problem was an unsupervised one based on the double-propagation algorithm. We proposed an enhanced set of rules for the double-propagation and also a minimal set of very general seed-words (i.e. two) that proved to perform comparable with the

state of the art results. We also addressed the problem of opinion mining from social networks, namely tweets. The challenges related to tweets are manifolds: tweets do not use a correct language in many cases; they have a limited size (140 characters) and usually contain slang, abbreviations, special symbols, emoticons etc.

Normally an opinion includes the target (i.e. the identity that is referred), the opinion holder (who expresses the opinion) and the opinion itself. In the case of social network content or postings the holder is by default the current user but, although the posting involves a sentiment polarity, there is not always a target. The sentiment might be just a feeling. Supervised approaches are not that useful to classify tweets since there is no relevant annotated corpus and it is very expensive in terms of effort to create such corpus. Moreover, given the diversity of content existing in tweets, supervised approaches are not able to perform acceptable to classify them. We investigated unsupervised approaches for polarity classification of tweets. We also analysed existing lexical, semantic and benchmark data resources that can help to increase the classification performance. We proposed a processing flow and applied our solution to four benchmark datasets in order to compare our results with other state-of-the art solutions. The obtained performance was comparable with the performances reported on supervised solutions.

c. *Knowledge processing*

We investigated methods to derive more information out of the extracted knowledge. In this respect we proposed methods for contradiction detection in opinions, for opinion driven community detection and for financial data streams mining. The first two methods are relying on the results of the opinion mining methods. The contradiction detection method identifies opinions on the same target and evaluates the polarity distance between them. It also identifies basic forms of negation in order to identify contradictory opinions. The community detection solution aims the identification of opinion holders that share similar opinions. The solution also integrates social network data and employs two methods for integrating multidimensional data: Network Integration and Partition Integration. We also experimented 9 multiple-target aggregated similarity functions that lead to the following conclusions: linear functions perform poorly for data sets with multiple targets and functions that calculate the average similarity have greater resilience to noise.

The proposed approaches and obtained results are detailed in the Scientific achievements chapter.

The future research interests of the candidate are shifting towards BigData methods applied in Internet of Things solutions. The envisioned challenges include:

- Semantic alignment of meta-data that describes data received from heterogeneous, dynamic data sources such as devices, sensors, etc.
- Develop efficient data representation and processing methods, with a special focus on data streams
- Integrating machine learning techniques and semantic resources in data (streams) analysis in order to perform context-aware knowledge extraction

More details can be found in the Scientific future development plans chapter.