Semi-Automatic Knowledge Extraction to Enrich Open Linked Data

Elena Baralis
Politecnico di Torino, Italy

Silvia Chiusano
Politecnico di Torino, Italy

Giulia Bruno
Politecnico di Torino, Italy

Alessandro Fiori
Politecnico di Torino, Italy

Tania Cerquitelli
Politecnico di Torino, Italy

Alberto Grand
Politecnico di Torino, Italy

EXECUTIVE SUMMARY

In this chapter we present the analysis of the Wikipedia collection by means of the ELiDa framework with the aim of enriching linked data. ELiDa is based on association rule mining, an exploratory technique to discover relevant correlations hidden in the analyzed data. To compactly store the large volume of extracted knowledge and efficiently retrieve it for further analysis, a persistent structure has been exploited. The domain expert is in charge of selecting the relevant knowledge by setting filtering parameters, assessing the quality of the extracted knowledge, and enriching the knowledge with the semantic expressiveness which cannot be automatically inferred. We consider, as representative document collections, seven datasets extracted from the Wikipedia collection. Each dataset has been analyzed from two point of views (i.e., transactions by documents, transactions by sentences) to highlight relevant knowledge at different levels of abstraction.

ORGANIZATION BACKGROUND

The Politecnico di Torino offers excellence in technology and acknowledges its historical context. It promotes the ability to carry out theoretical or applied research. Engineers and Architects are the main professional figures at the Politecnico di Torino. Both have strategic planning and a common interdisciplinary approach. The range of studies is broad and ever-widening: it spans space, environment and land, telecommunications, information, energy, mechanics, electronics, chemistry, automation, electrical engineering, industrial design, architecture and building, and many others. The Politecnico has 30.000 students studying on 120 courses (28 Bachelor's degree courses; 32 Master of Science courses; 23 Doctorates and 37 specialization courses). 12 of them are held in English. In the academic year 2011/2012 the Politecnico had around 5,600 students in the first year; in 2010 around 4,500 students graduated with a Master of Science or a Bachelor's Degree. Each year, between lectures, laboratories and practical exercises there are 170,000 hours of teaching. There is a staff of over 890 lecturers and researchers, and around 800 administration staff. There are 5 Schools, 1 Graduate School, 1 Post-graduate specialization, and 11 Departments.

SETTING THE STAGE

The case study exploits data mining techniques to support the semi-automatic inference of interesting knowledge to enrich open linked data. As representative example we consider the Wikipedia dataset, the world's largest online encyclopedia. Authors, affiliated to the Dipartimento di Automatica e Informatica (DAUIN) or to the Dipartimento di Ingegneria Gestionale e della Produzione of Politecnico di Torino, have been active for a long time in the research area of data mining and database systems management. The research activity is mainly devoted to the design and development of innovative efficient techniques for data analysis applied to different domains, including algorithms and data structures to mine large databases, classification algorithms for structured and unstructured (textual) data, algorithms for sequential patterns analysis, algorithms for extracting high level abstraction of the mined knowledge, sensor data analysis, network data analysis, and context-aware applications. Authors have published several papers in both international journals and conference proceedings. They have also been for a long time either teachers or teaching assistants in different databases and data mining courses at the Politecnico di Torino.