EXPLORING DATA HIERARCHIES TO DISCOVER KNOWLEDGE IN DIFFERENT DOMAINS

April 30, 2019

Giuseppe Ricupero
Politecnico di Torino

Abstract

Data mining techniques are powerful instruments that can be effectively used to analyze data collections and extract hidden and useful knowledge otherwise unavailable. They allow extracting previously unknown interesting patterns such as dependencies among data objects (*association rule mining*), or a model describing data classes (*classification*). However, these kind of data collections are often characterized by a continuously increasing dimension and heterogeneousness, which limit the feasibility of analysis by means of data mining techniques currently available. Therefore, an important question is how these collections can be transformed into exploitable knowledge.

This PhD thesis addresses the study and development of *novel data analysis frameworks* and *patterns* to extract useful insights from the targeted data collections. To this end, the exploration of data taxonomies built on top of the considered data is proposed in this PhD thesis. A *data taxonomy* is a set of is-a hierarchies each one referring to a specific data attribute. Each hierarchy aggregates all the values assumed by the corresponding attributes into higher level concepts in a tree-based structure.

In Chapter 2, the fundamental issue of monitoring the air-pollution in the urban environment is addressed. Two data mining frameworks, *GECKO* (GEneralized Correlation analyzer of pOllution data) and *ARQUATA* (AiR QUAlity patTern Analyzer), are described. The data mining system GECKO leverages the power and expressiveness of the generalized association rules to extract, interpretable correlations among air pollution related data at different granularity levels. The data analyzed by GECKO system covered various aspects of air quality such as meteorological conditions, acquisition times, and vehicular traffic measurements. In its turn, the data analysis performed in *ARQUATA* is targeted at discovering combinations of pollutant concentrations that averagely are in a critical condition. The weighted frequent itemset pattern, is exploited to extract the novel type of pattern designed for this purpose, the *air quality pattern.* To provide different insights to domain experts and municipality actors, these patterns are extracted from several aggregations of the raw data following specific temporal and spatial granularities. GECKO and ARQUATA engines were validated on real open data collected in a major Italian city (i.e., Milan).

In Chapter 3, a novel exploratory data-driven methodology, named ***B***ike Station Ov***E***r***L***oad Ana***L***yzer (*BELL*) is presented with the aim of improving the user perception and ease of maintenance of bike-sharing systems. *BELL* analyzes the occupancy level data acquired from real systems to determine situations of dock overload in mul-

tiple stations which could lead to service disruption. The proposed methodology relies on a new pattern type called Occupancy Monitoring Pattern which has been designed to detect situations of dock overload in multiple stations. Since stations are geo-referenced and their occupancy levels are periodically monitored, occupancy patterns can be filtered and evaluated by taking into consideration both the spatial and temporal correlation of the acquired measurements. The results achieved on real data highlight the potential of the proposed methodology in supporting domain experts in their maintenance activities, such as periodic re-balancing of the occupancy levels of the stations, as well as in improving user experience by suggesting alternative stations in the nearby area. During the empirical study, *BELL* has been thoroughly evaluated using a real dataset acquired from the bicycle sharing systems of two important smart cities, i.e., Barcelona and New York.

Chapter 4 presents a study goal of which is to discover recurrent combinations of items characterized by high profit from transactional datasets. A novel type of pattern, namely the *Generalized High-utility Itemset* (*GHUI*), is defined and developed to combine the expressiveness of generalized and High-Utility itemsets. *GHUI* represents a combinations of items at different granularity levels characterized by high profit (utility). According to a user-defined taxonomy, items are first aggregated into semantically related categories. While profitable combinations of item categories provide interesting high-level information, GHUIs at lower abstraction levels represent more specific correlations among profitable items. A single-phase algorithm is presented to efficiently discover utility itemsets at multiple abstraction levels. The experiments, which were performed on both real and synthetic data, demonstrate the effectiveness and usefulness of the proposed approach.

Chapter 5 introduces the *TACOMA* system. Starting from a real industry case it aims at supporting the integration of data regarding business activities between different web directories (e.g., Google Maps, Pagine Gialle, Apple Maps) characterized by taxonomies of different granularity levels. In particular, *TACOMA* addresses the problem of a source taxonomy, which has categories at a coarser conceptual level of granularity of the target taxonomy. For instance, a source taxonomy with a concept of furniture whose instances should be mapped to the concepts of chair, table, or bookcase in the target taxonomy. The issue is addressed using an classification approach, where the textual data of each case is leveraged to correctly predict the appropriate category of the target taxonomy. The experiments performed on real data coming from a prominent Italian web directory (i.e., Pagine Gialle) have proven the efficacy of the proposed methodology, which has been already integrated in the production system of the company to export its data towards international partner systems (i.e., Apple Maps, Amazon Alexa).