Redescription Mining from Boolean data with a hierarchy

Viet Ta

February 27, 2020

In the contemporary world of unprecedented amount of data we produce every second, data mining is no longer a trendy jargon that people use to decorate their profile, but rather a essential practice for the new wave of technological innovations to happen. Consequently, the modern idea of data mining is not only to discover information from crude data, but also to condense that information down to concise descriptions and insights. One solid representation of such ”insights mining” is the redescription mining.

As in its name, redescription mining is the mining of redescriptions. In a nutshell, a redescription is a pair of descriptions, i.e. queries, those have comparable supports. Although, to fully understand the concept, a framework of theories is prerequisite. For example, we need to be familiar with how the data is defined, structured, and constrained; how to formalize the descriptions and redescription; how to measure the similarity between two queries' results, etc.

One classic example of redescription mining is about bioclimatic niche, where we want to find out the relation between temperature and the distribution of the lynxes on the map. The first query is to find the geographical area where the maximum March temperature is in a specific range, e.g. $-24.4°C \leq T \leq 3.4°C$. The second query is to find the geographical area on the map where the lynxes inhabit. If we can find a pair of queries, one from each set of queries, such that the outputs of them, in this case spatial data, overlap at a certain predefined threshold; then we have successfully found a redescription.

Among various types of data those can be mined, Boolean data is a common one, which requires it own techniques and tricks to deal with.