Redescription Mining from Boolean data with a hierarchy

Viet Ta

March 5, 2020

With the enormous amount of data we produce nowadays, data mining is becoming more and more prevalence. 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 method of such "insights mining" is the redescription mining.

The intuition of redescription mining is to find two different ways to describe the same thing. Before we can go for the definition of it, several fundamental terminologies need to be addressed. Firstly, the redescription mining data model $\mathcal{D}$ is a tuple of 3 *sets* : (*entities* $\mathcal{E}$, *attributes* $\mathcal{A}$, *views* $\mathcal{V}$). Each entity $e \in \mathcal{E}$ is associated with a set of *attributes* from $\mathcal{A}$. The *attributes* are partitioned into disjoint *views* $V \in \mathcal{V}$. Intuitively, this model is analogous to the concept of data table, where entities are equivalent to rows, attributes correspond to columns, views are just partitions of columns. Secondly, a description is simply a query q evaluates each entity e and output a Boolean value. And thirdly, a support of a query is a set of entities in the data that renders true in the query.

In a nutshell, a redescription is a pair of descriptions with disjoint views and sufficiently similar supports. The similarity of two supports can be measured by a distance function.

One classic example of redescription mining is about finding bioclimatic niche. In the example, 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 results of them, in this case spatial data, overlap at a certain predefined threshold; then we have successfully found a redescription.

If we keep going deeper from here, redescription mining can be branched out by the way we define structure the data, special constrains, etc. 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.