Enrichment of Taxonomy

Mili Biswas

Master thesis in Computer Science

Building a taxonomy from textual data is an essential task in various applications and systems. A taxonomy created from textual data helps to structure information into categories that further enables search and reuse of those information effectively. This is the case of Vogue retailer, where taxonomies are used to improve the online search and recommending products. Taxonomies allow also to define rules and relationships among different information categories in an abstract way that facilitates further development and refinement of a knowledge-base system. Many applications such as Information Retrieval, Text Clustering and Classification or Text Mining heavily relies on taxonomies built on textual data.

Similarly to the automatic creation of taxonomy, enriching an existing taxonomy based on new text data is also gaining popularity now-a-days. For example, in fashion technology, new product concepts or categories are needed to be identified from new textual data and then these can be added in an existing product category taxonomy. This helps to create new product and better product recommendations. However, both creating a taxonomy and enriching an existing one are challenging tasks because these require domain knowledge of underlying data which is not available most of the time. Moreover, enriching an existing taxonomy based on little new data is almost impossible as most of the algorithms require large amount of data to train the associated model.

In this thesis, we study different state-of-the art algorithms used for building taxonomy automatically from textual data. We also propose a novel technique TaxoTL for enriching existing taxonomy from new data. Our empirical results on several real-world datasets show that TaxoTL is capable of enriching taxonomy correctly. Our measured scores of TaxoTL has comparable accuracy with the other state-of-the-art algorithms. We also show that TaxoTL is on average more than 20 x faster compared to the other state-of-the-art algorithms (2min vs.40mins). In addition, TaxoTL is more memory efficient and consumes on average 10 x less memory than other algorithms.

Prof. Philippe Cudré-Mauroux