Extracting tag hierarchies

Gergely Palla
Statistical and Biological Physics Research Group of HAS, Eotvos University, Hungary,
pallag@hal.elte.hu

Keywords: *tags; hierarchy; networks; folksonomy;*

Tagging items with descriptive annotations or keywords is a very natural (and human) way to compress and highlight information about the properties of the given entity. In the recent years tags have become very prevalent in various online platforms and databases ranging from e.g., blogs and online stores through scientific publications to protein databases. Furthermore, tagging systems dedicated for voluntary tagging of photos, films, books, webpages, etc. with free words are also becoming popular. The emerging large collections of tags associated to different objects are often refereed to as folksonomies, highlighting their collaborative origin and the “flat”, egalitarian organisation of the tags opposed to hierarchical ontologies. An additional tag hierarchy in case of a folksonomy, (or an online store), can very effectively help narrowing or broadening the scope of search, and accordingly, several tag hierarchy extraction methods have been proposed over the years.

Here we present a complete framework for automated tag hierarchy extraction based on tag occurrence statistics. Along with proposing new algorithms, we are also introducing different quality measures enabling the detailed comparison of competing approaches from different aspects. Furthermore, we set up a computer generated benchmark providing a versatile tool for testing, with a couple of tunable parameters capable of generating a wide range of test beds. Beside the computer generated input we also use real data in our studies, including a biological example with a pre-defined hierarchy between the tags. The encouraging similarity between the pre-defined and reconstructed hierarchy, as well as the seemingly meaningful hierarchies obtained for other real systems indicate that tag hierarchy extraction is a very promising direction for further research with a great potential for practical applications.