TU Clausthal

Bachelor-, Master- und Doktorandenseminar
des Instituts für Informatik

A Link-Density-based Algorithm for Community Finding in Graph Databases

Vladivy Poaka, TU Clausthal

More and more data is represented through graph databases, due not only to exibility and scalability of their data models, but also because of shorter execution time of queries. Furthermore, graph databases management systems are particularly appropriate to represent social and information networks in the most natural sense. This allows to take advantage of one of the most relevant features of graphs representing real systems, which is community structure, or clustering, i.e. the organization of vertices in subsets called clusters, with more edges joining vertices of the same cluster than those connecting vertices of different clusters. Such clusters, or communities, can be considered as independent compartments of a network, playing a similar role like, for example groups of proteins having the same specific function within a cell in biology, clusters of customers with similar tastes in a social network, or groups of paper/documents about same or similar topics. This may also be helpful to optimize complex networks by finding shortest paths, or rearranging elements to minimize links between them. Detecting such characteristics of networks is very useful in sociology, biology, engineering, economics, computer science and politics. The clustering problem is very hard and not yet completely solved, despite big efforts of an interdisciplinary community of scientists working on it over the past few years. There are numerous community detection techniques in the literature, but they suffer in some way from limitations such as the accuracy of results, the complexity time and space, and/or the strong dependence on the application domain.

We started with an exposition of the state-of-the-art, from definitions of main concepts of the problem, to the most popular techniques, their various improvements, and how to compare them against each other. Then, we designed and proposed a solution with some improvements with the use of efficient data structures and parallelizing the execution of the program that allows to detect overlaps between communities. More, we design an approach, called Link-Density-based Preferential Attachment, only based on the intrinsic structure of the graph at study, to stabilize and to increase the accuracy of the well-known Label Propagation Algorithm. This approach is also able to distinguish core from boundary and noise (or isolated) points.

Identifying groups and their boundaries permits a classification of vertices, according to their structural position in the group. Also, vertices with a central position in their clusters may have an important function of control and stabilization within this group, and those lying at the boundaries between modules play an important role of mediation, and lead the relationships and exchanges between different communities. Our program is implemented in this vein to detect such vertices with three different approaches. Another aspect we discussed is prediction in networks. It is a process that consists of learning from previous and current states of a graph, and propose (new) links based on its inherent characteristics. The application of this feature is recommendations of products through targeted marketing in the business world. It may also help to model the evolution of a (social) network, so that some preventive measures could be taken soon enough. We implemented this function with an evaluation of it. After many tests and benchmarking we made, we illustrated the results we obtained and discussed potential lines of future works on complex networks.

Donnerstag, den 06.11.2014
10 Uhr c.t. in Raum 106, IfI, Julius-Albert-Straße 4