Scalable Similarity Computation in Social Networks

We currently dispose of a large Twitter dataset (2.1 million accounts, 150 million follower/followee links and around 16 million tweets). To validate our theoretical results, we intend to test our architecture and algorithms with large real datasets. We will also study the scalability of graph-based recommendation algorithms for large social networks, and propose methods for on-line recommendation. One important part of the work will be dedicated to optimizing the graph-based similarity scores, which involve the computation of paths between graph nodes. Some existing solutions split the graph into clusters of nodes and propose methods such that the random walk rarely cross the cluster boundaries and cause page faults. In order to reduce computation time, a simple solution would be to precompute and store on disk all pairwise similarity distances offline. Therefore, answering a similarity query online requires only a disk lookup. However, the space requirement is quadratic in the number of nodes in the graph, which is infeasible for current large social graphs. Existing works on distance labeling in graphs were shown to be efficient in rapidly estimating the shortest path distances in large graphs. Some other solutions, based on the usage of landmarks and sketches, which can be viewed as simplified distance labeling algorithms, were also proposed to efficiently estimate shortest paths.

The general idea of these solutions is to store at each node information, like typically a set of vertices and the distance from this node to every one on these vertices, such that the distance queries are answered by using this information. Efficient indexing schemes will allow to optimize the computation cost and to efficiently access the huge amount of data.

Résumen du projet de recherche (Langue 2)

* Study {{state of the art}} of user-based collaborative filtering methods, of graph-based link prediction algorithms and of the performance of their computation on large social networks. * Propose {{methods to compute and store partial path scores}} between the graph nodes based on the usage of well-chosen landmarks or user clusters. At query time, the partial scores will be combined to estimate the actual similarity between the query nodes. * Propose {{storage methods and indexes}} to efficiently manage the partial score and path information. The indexes should be built such that the cost of the access to these information during the computation of the recommendations is optimized. * Tackle the {{dynamicity problem}} and design similarity re-computation strategies possibly by using statistics on graph evolution, that describe the information change frequency and publishing peaks and the user behavior. Propose on-line recommendation algorithms and scalable indexing structures. * Build a {{validation prototype}}. To validate our theoretical results, we intend to test our architecture and algorithms with large real datasets. We currently dispose of a large Twitter dataset (2.1 million accounts, 150 million follower/followee links and around 16 million tweets).
Related work:

- Jin Ruoming, Ruan Ning, Xiang Yang and Lee Victor. A Highway-centric labeling approach for answering distance queries on large sparse graphs. SIGMOD'12
- Ryadh Dahimene, Cédric du Mouza, Michel Scholl: Efficient Filtering in Micro-blogging Systems: We Won't Get Flooded Again. SSDBM'12
- Das Sarma Atish, Gollapudi Sreenivas, Najork Marc and Panigrahy Rina. A sketch-based distance oracle for web-scale graphs. WSDM '10
- Tretyakov Konstantin, Armas-Cervantes Abel, Garcia-Banuelos Luciano, Vilo Jaak and Dumas, Marlon. Fast fully dynamic landmark-based estimation of shortest path distances in very large graphs. CIKM'11
- Song Han Hee, Cho Tae Won, Dave Vacha, Zhang Yin and Qiu Lili. Scalable proximity estimation and link prediction in online social networks. IMC'09
- Jin Ruoming, Ruan Ning, Xiang Yang and Lee Victor. A Highway-centric labeling approach for answering distance queries on large sparse graphs. SIGMOD'12
- Tretyakov Konstantin, Armas-Cervantes Abel, Garcia-Banuelos Luciano, Vilo Jaak and Dumas, Marlon. Fast fully dynamic landmark-based estimation of shortest path distances in very large graphs. CIKM'11