Exact and inexact subgraph matching in large networks

Di Natale R[1], Ferro A[1], Giugno R[1], Mongiovì M[1], Pulvirenti A[1], Sharan R[2], Shasha D[3]

Motivation

The increasing availability of large biological networks has recently stimulate a steeply rise of interest in efficient tools for analyzing them. Comparing biological networks is crucial for understanding the fundamental mechanisms underlying biological processes of living organisms. For example, comparing different species can assist in the identification of conserved complexes and in the annotation of proteins of newly studied species. Network querying plays an important role in this field, allowing to identify, among a database of target networks, sub-networks that are "identical" or "similar" to a given query network. Although a lot of effort has been invested on searching for "identical" sub-networks (exact matching problem), solving this problem is still unfeasible for large networks. The situation is even worse when it comes to cope with the more general problem of querying for "similar" sub-networks. The latter problem, referred as inexact matching, is fundamental for cross species comparison. Recently, many graph indexing systems have been developed for addressing the exact matching problem. They aim to speed up the query processing, by using features, previously extracted from the database of target networks. Basically such features (i.e. small sub-graphs, sub-trees or paths) are used to filter out networks of the database that do not contain all the features of the query. These techniques are effective on databases of small networks but become often infeasible when applied to huge networks (e.g. protein-protein interaction networks). We developed two graph indexing tools based on innovative approaches for the exact matching and inexact matching problem respectively: SING (Sub-graph search In Non-homogeneous Graphs) and SIGMA (A Set-cover-based Inexact Graph Matching Algorithm). The former aims to solve the exact matching problem in large database of small and large networks. The latter provides a novel graph indexing method to cope with the inexact matching problem.

Methods

SING uses a novel approach for exact matching on large graphs that makes use of paths as features. In contrast to systems that use more complex features such as sub-graphs or sub-trees, our index includes all paths of bounded length and

[1] Dipartimento di Matematica ed Informatica, Università di Catania, Catania Italy [2] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv Israel [3] Courant Institute of Mathematical Sciences, New York University, New York USA

considers the position of a feature within the graph. A new pruning rule is defined and shown that it captures the structure of the graphs much better, leading to a strong reduction of candidates. Using this additional information, SING is able to improve the filtering power and to optimize the verification phase. SIGMA builds on approximating a variant of the set-cover problem that concerns overlapping multi-sets. The algorithm is based on associating a feature set with each edge of the query and looking for collections of such sets whose removal will allow exact matching of the query with a given graph. This translates into the problem of covering the missing features of the graph with overlapping multi-sets. We formulate this variant of Set Cover and provide a greedy approximation for it.

Results

Extensive tests on a database of 40,000 chemical compounds show that our tools outperform the most popular systems when both exact and inexact matching is applied on databases of small graphs. To evaluate the performance of SING for exact matching on large networks, three datasets are used. In the first we consider a single randomly generated scale free network of 2000 nodes and 4000 edges. In a second test, we query a set of protein complexes of yeast against the whole human protein-protein interaction network. In these two tests we compare SING against VF2, which is considered the state-of-the-art algorithm for exact matching. The results show that SING outperforms VF2 in all the experiments. The latter dataset was build considering several copies of the transcription regulation network of Escherichia Coli, annotating the nodes of each network with discretized gene expressions profiles. We query the network database with a set of network motifs labeled with gene expression levels. Even on this dataset SING ouperforms the other tools. We evaluated the performance of SIGMA in looking for similar protein complexes. We applied our algorithm for comparing yeast and human protein complexes. Matched protein complexes are likely to share similar biological functions. Each human complex was queried against the yeast complex collection. The best matching complexes, in terms of number of matching edges, were filtered out and separately analyzed. The results are consistent with what is reported in the literature.

Contact e-mail

dinatale@dmi.unict.it