图数据研究的一点体会

NDBC 2016研究生辅导报告

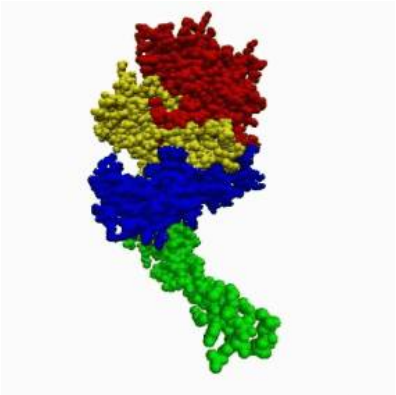
李荣华
深圳大学计算与软件学院

报告大纲

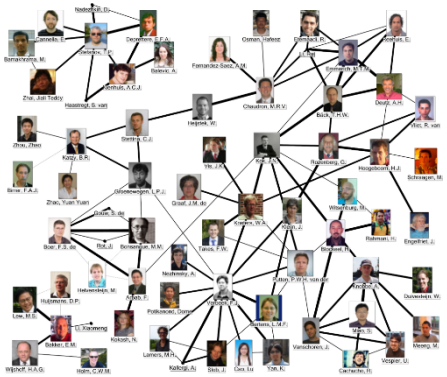
- 图数据简介
- 图数据的算法基础
- 图数据研究：如何寻找问题？
- 研究方法：结合我所做过的研究经验

图数据无处不在！

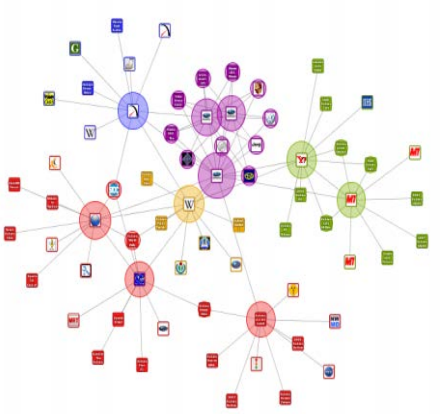
社交网络



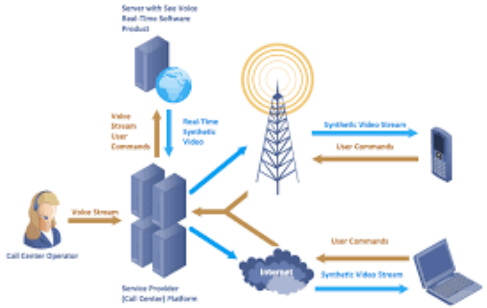
蛋白质交互网络



科学家合作网络



万维网



电信网

路网

常用的图数据下载

- 斯坦福大学公开数据集
  - http://snap.stanford.edu/data/index.html
- 亚利桑那州立大学公开数据集
  - http://socialcomputing.asu.edu/pages/datasets
- WebGraph数据
  - http://law.di.unimi.it/datasets.php
- Konect数据集
  - http://konect.uni-koblenz.de/

图数据库的算法基础

- 磨刀不误砍柴工！

- **“刀”＝“图算法”**

基本图算法

- **把显示器搬走都能写出来的算法！**
  - 广度优先搜索
  - 深度优先搜索
  - Prim/Kruskal算法
  - Bellman-Ford算法
  - Dijkstra算法
  - 拓扑排序算法
  - 强连通分支算法

高级一些的图算法

- 最大团、三角形枚举算法
- 网络流算法
  - 应用：求稠密子图的Goldberg算法[1]
- K-core、k-truss、k-edge connected subgraph分解算法[2-4]
- 图同构、图模式匹配算法[5]
- 求top-k最短路径的Lawler-Yen算法[6-7]
- 动态规划算法，例如求TSP、Steiner Tree等问题[8-9]
- A星算法[8-10]
- 图聚类算法（e.g. SCAN算法[11]）
- PageRank、SimRank等算法
- Distance Oracle算法（参考Eidth Cohen的论文）
- 频繁子图模式挖掘算法等
- 。。。

图数据的研究问题

- 查询处理问题
  - 可达性查询、路径查询、相似性查询、连接查询、关键词搜索、子图匹配等等
- 挖掘问题
  - 图聚类、社区探测、子图模式挖掘等等

如何寻找新的研究问题？

- 方法一：问题突破

新问题=老问题+（新）数据类型+（新）代价模型

- 方法二：应用驱动的新定义、新问题
- 方法三：算法突破
  - 奥林匹克精神：更高、更快、更强！
- 一个有价值的研究问题的求解算法不应太简单！

图数据类型

- 静态图数据
- 边加权图数据
- 动态图数据
- 节点带标签属性图数据
- 不确定性图数据
- 时态图数据（点或者边带有时间信息）
- 时空图数据（点或者边带有时间和空间信息）
- 。 。 。

代价模型

- 流图数据模型
  - 图数据只能读取有限次数，且内存无法保存整个图
- （半）外存模型
  - 半外存：内存可以存储所有的节点，但不能存储所有的边
  - 外存：内存无法存储所有的节点
- 并行图计算模型
  - 例如，多核的共享内存
- 分布式图计算模型（MapReduce、Pregel等）
  - 图分布式地存储在一个集群中，通常需要优化通讯代价和迭代计算的次数
- 外包图计算模型
  - 图数据存储在云端，并可能已经做了隐私保护处理。
- 。 。 。

老问题+新数据类型

- K-core分解+动态图数据
  - Rong-Hua Li etc. Efficient Core Maintenance in Large Dynamic Graphs，TKDE 2014
- 社区搜索问题+节点带权值的图数据
  - Rong-Hua Li etc. Influential Community Search in Large Networks， PVLDB 2015
- 图聚类问题+属性图数据
  - Yang Zhou etc. Graph Clustering Based on Structural/Attribute Similarities，PVLDB 2009
- 可达性查询问题+时态图数据
  - Huanhuan Wu etc. Reachability and time-based path queries in temporal graphs, ICDE 2016

老问题+新代价模型

- 图的深度优先搜索问题+外存模型
  - Zhiwei Zhang etc. Divide & Conquer: I/O Efficient Depth-First Search, SIGMOD 2015
- K-core分解+外存模型
  - Dong Wen etc. I/O efficient Core Graph Decomposition at web scale, ICDE 2016
- 基本图算法+分布式图计算模型
  - Lu Qin etc. Scalable big graph processing in MapReduce, SIGMOD 2014
- 最短路径问题+外包图数据模型
  - Jun Gao etc. Neighborhood-privacy protected shortest distance computing in cloud, SIGMOD 2011

新定义、新问题

- 以社交网络中的影响传播为背景，如何找到k个人使得别人通过随机游走的方式容易找到他们。
  - Rong-Hua Li etc. Random-walk domination in large graphs, ICDE 2014
- 如何找到图中所有的稠密子图?
  - Lu Qin etc. Locally Densest Subgraph Discovery, KDD 2015

算法突破

- 通常需要New ideas!

- 图结构聚类的快速算法
  - Lijun Chang etc. pSCAN: Fast and exact structural graph clustering, ICDE 2016
- 关系数据库关键词搜索的快速算法
  - Rong-Hua Li etc. Efficient and Progressive Group Steiner Tree Search, SIGMOD 2016
- 最短路查询的快速算法
  - Takuya Akiba etc. Fast exact shortest-path distance queries on large networks by pruned landmark labeling，SIGMOD 2013

研究方法

- 推广法
  - Ruoming Jin etc. Distance-Constraint Reachability Computation in Uncertain Graphs, PVDLB 2011
  - 我们将上述算法推广至更一般的情形，采用类似于Lawler-Yen算法的空间划分思想，得到了一种递归分层抽样框架
  - Rong-Hua Li etc. Efficient and Accurate Query Evaluation on Uncertain Graphs via Recursive Stratified Sampling, ICDE 2014

研究方法

- 统一法
  - 将以往的主要算法建立联系，得到统一的算法框架
  - Rong-Hua Li etc. On Random Walk Based Graph Sampling, ICDE 2015
  - 通过引入一个参数，得到参数化的随机游走算法，从而使得以往的算法都是我们算法的特殊情况，因此统一了所有的随机游走抽样算法。

研究方法

- 类比法
  - 类比已有的算法思想，设计新的算法
  - Rong-Hua Li etc. Efficient and Progressive Group Steiner Tree Search, SIGMOD 2016
  - 最短路径问题的双向Dijkstra算法。路径问题能双向，树问题是否也可以？我们为此设计了一个在树上类似于双向Dijkstra的算法。

研究方法

- 通过研究问题的固有属性和不变量设计算法
  - Rong-Hua Li etc. Influential Community Search in Large Networks, PVLDB 2015
  - 通过研究问题中节点影响力的排序不变，我们发现了在k-core的k值不同的索引树中share了这个排序，从而设计了一个更为巧妙而且高效的索引构造算法。

研究方法

- 它山之石！
  - 利用其它学科的方法来解决本领域中的问题
  - Rong-hua Li etc. Link prediction: the power of maximal entropy random walk，CIKM 2011
  - 最大熵随机游走是理论物理中的一个概念，它每次都是有偏向地往特征向量中心性高的节点游走，因此符号链路预测的直观感觉。我们采用这一方法设计了很多链路预测的算法，得到了非常好的效果。

谢谢！

参考文献

- [1] A. V. Goldberg. Finding a maximum density subgraph
- [2] V. Batagelj and M. Zaversnik. An O(m) algorithm for cores decomposition of networks
- [3] J. Wang and J. Cheng. Truss decomposition in massive networks
- [4] L. Chang. Etc. Efficiently computing k-edge connected components via graph decomposition
- [5] W. Fan. Etc. Graph pattern matching: From intractable to polynomial time
- [6] J. Y. Yen. Finding the k shortest loopless paths in a network
- [7] E. L. Lawler. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem
- [8] R-H. Li. Etc. Optimal Multi-Meeting-Point Route Search
- [9] R-H. Li. Etc. Efficient and Progressive Group Steiner Tree Search
- [10] A. V. Goldberg and C. Harrelson, “Computing the shortest path: A* search meets graph theory
- [11] SCAN: A Structural Clustering Algorithm for Networks