



大图搜索：挑战性与相关技术

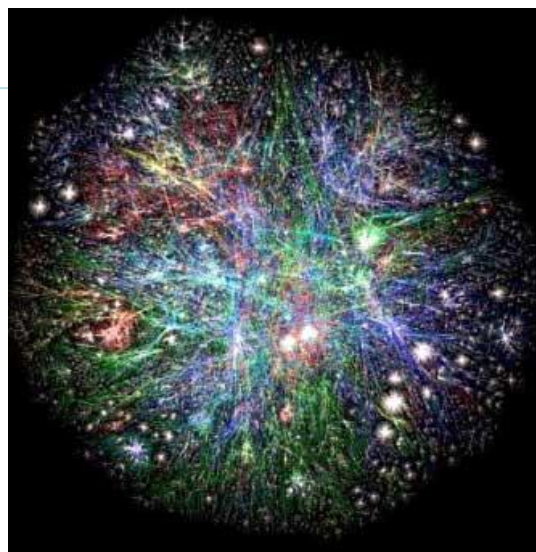
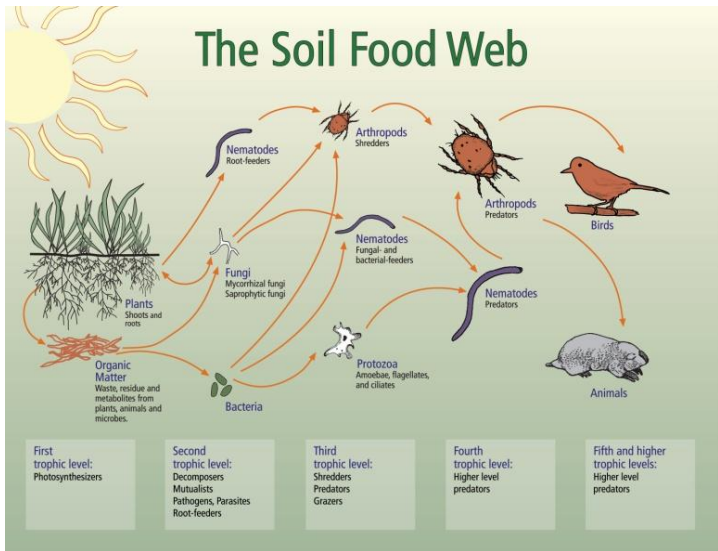
(Big Graph Search: Challenges and Techniques)



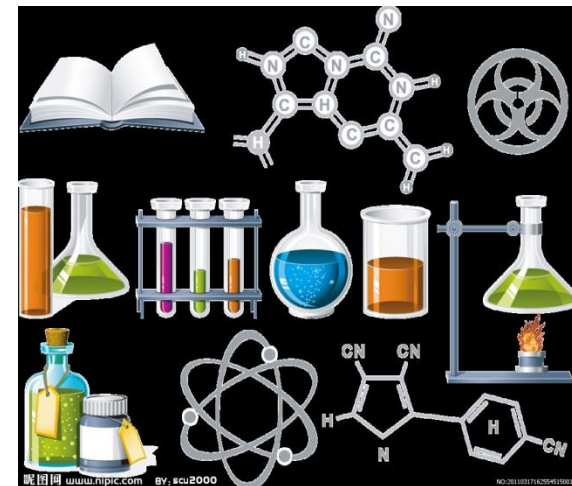
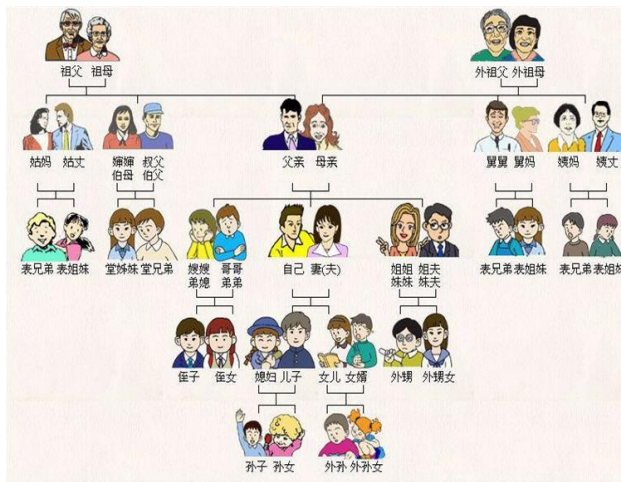
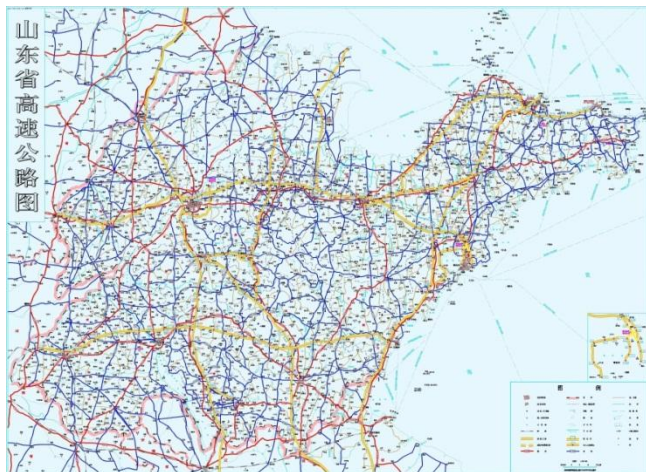
马 帅



北京航空航天大学
BEIHANG UNIVERSITY



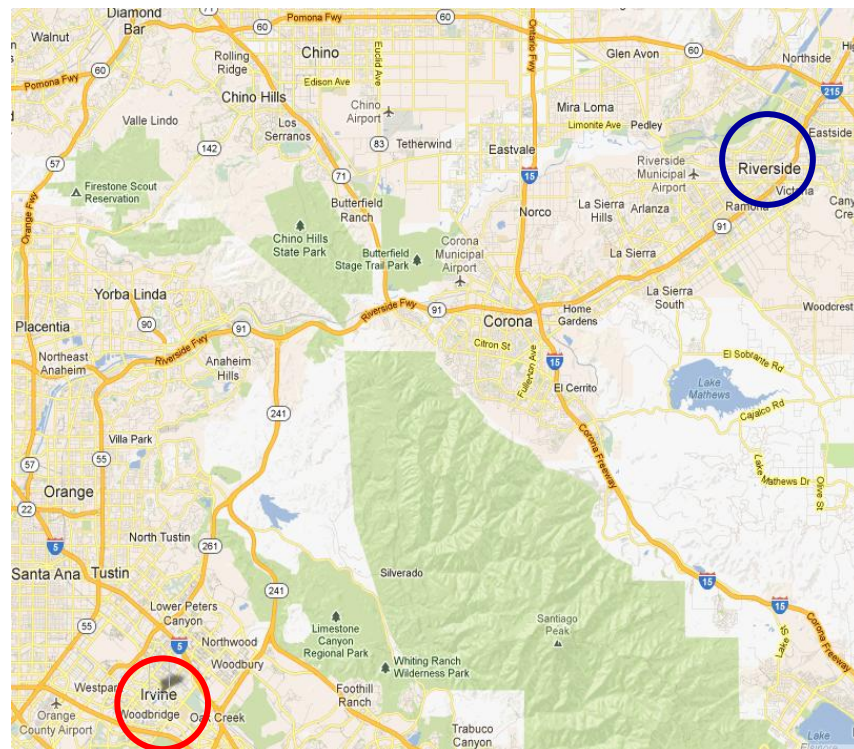
无处不在, 日常接触很多超大规模图!



应用案例1

路线规划^[1]

- 由于“**基于位置的服务(LBS)**”的广泛应用，**图搜索**大量应用到交通网络。
- **Example:** 司机Mark想从美国加州的**Irvine** 到**Riverside**.
 - 如果Mark想驾驶**car**最短的时间到达Riverside,那么这个问题可以看做为**图的最短路径**问题，然后找到的方案是**State Route 261**.
 - 如果Mark想驾驶**truck**运输**危险物品**，则有的**路和桥**是不允许通过的，路线的选择是受约束的. 这样可以通过**正则表达式**等方法来表达约束条件来搜索最佳的交通路线.
 - 如果考虑的**实时交通情况**。。。

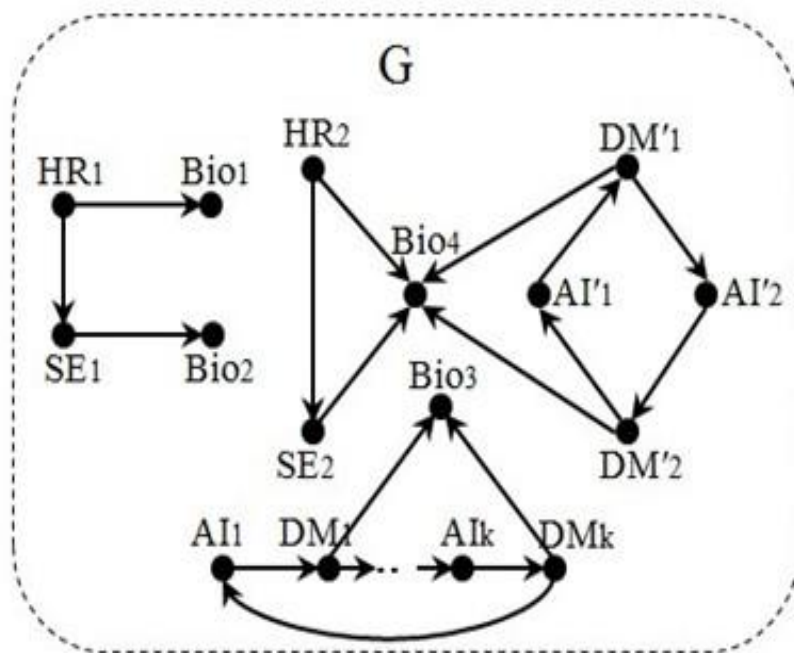


应用案例2

推荐系统 [2]

- 推荐系统有着广泛的应用，如social matching systems.
- 图搜索是一种非常有用的推荐工具.

- 猎头想找一位生物学家(Bio)来帮助一组软件开发人员 (SEs)来分析基因数据.
- 猎头通过专家推荐网(如LinkedIn)搜索
 - ✓ 图中顶点表示人，标签为专长
 - ✓ 图中边表示推荐，如 HR_1 推荐 Bio_1 , AI_1 推荐 DM_1





提纲

- 什么是图搜索？
- 大图搜索挑战性
- 大图搜索相关技术
- 总结



什么是图搜索？

(What is Graph Search?)



图搜索

提出统一的定义 [3]:

- 给定模式图 G_p 和数据图 G :
 - 检测是否 G_p “match” G ;
 - 查找 G 中所有 “match” G_p 的子图

标注:

- 两类查询:
 - 布尔查询(Yes/No)
 - 函数查询, 可以调用布尔查询
- 图中顶点或者边常常带有标签
- 模式图通常比较小(如10个顶点), 但数据图很大(如上亿个顶点)



图搜索

不同的“**match**”语义表示不同类型的图搜索, 包括:

- 最短路径/距离 [1]
- 子图同构 [9]
- 图同态及其扩展 [7]
- 图模拟及其扩展 [6,6]
- 图关键字搜索[4]
- 紧邻查询[8]
- ...

图搜索是一个非常 “general” 概念!



为什么需要图搜索？

(Graph Search, Why Bother?)

The need for a Social Search Engine



- **文件系统** – 1960年代：非常简单的搜索功能
- **数据库** – 1960中期：SQL查询语言
- **互联网** – 1990年代：关键字搜索引擎
- **社会网络** - 1990后期：

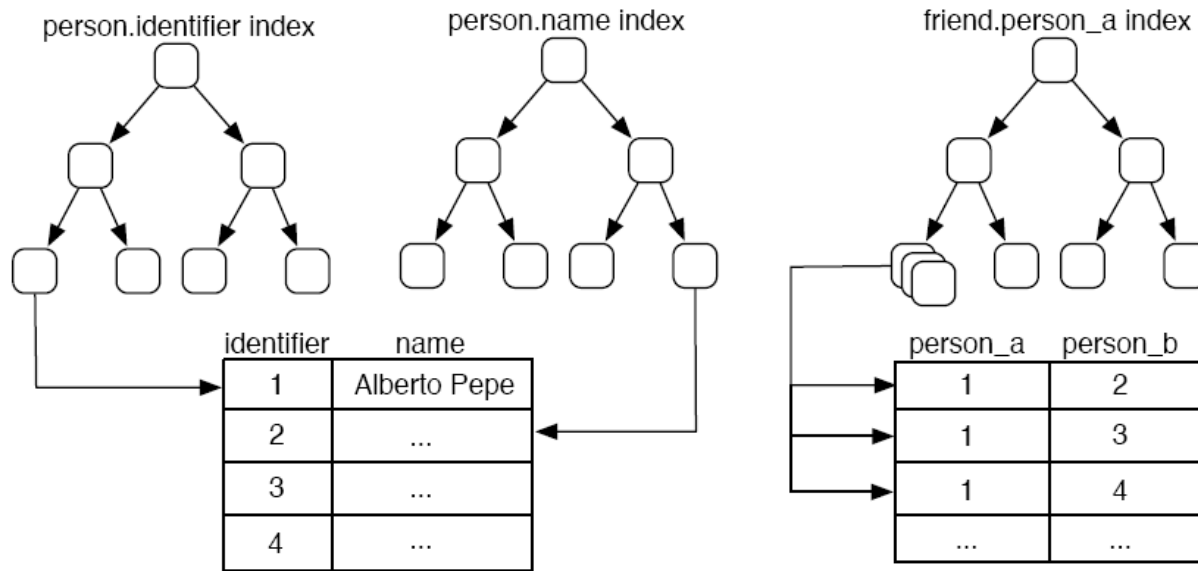


Facebook 与2013年1月16日推出“graph search”

影响到了Google、Yelp和LinkedIn; Yelp股价当天下降7%

图搜索是一种新型社会搜索模式！

图搜索 vs. 关系数据库 [10]



Query:

查找Alberto Pepe所有朋友名字

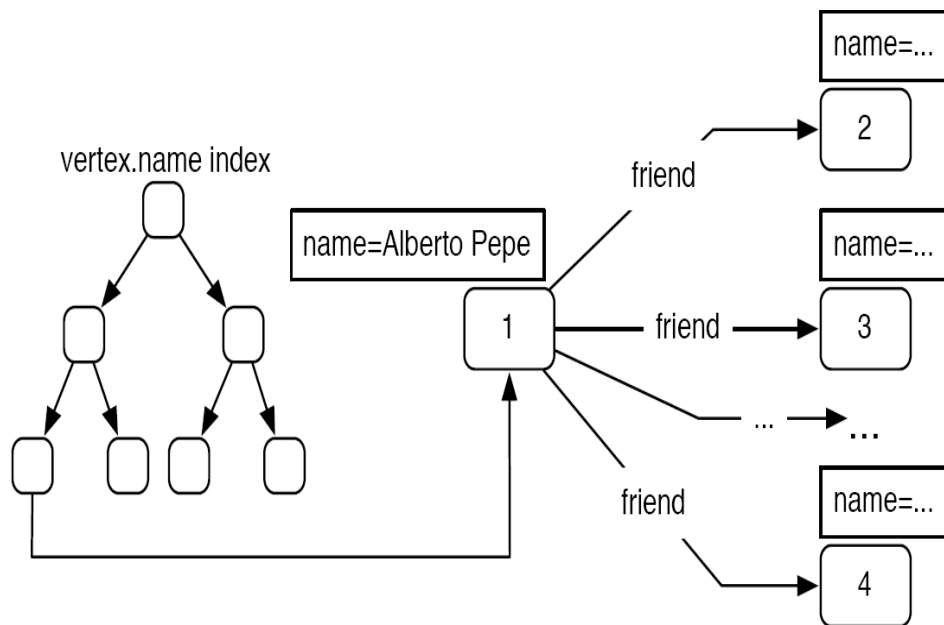
Step 1: The person.name index \rightarrow the identifier of Alberto Pepe. $[\log_2 n]$

Step 2: The friend. person index \rightarrow k friend identifiers. $[\log_2 x : x \ll m]$

Step 3: The k friend identifiers \rightarrow k friend names. $[k \log_2 n]$



图搜索 vs. 关系数据库 [10]



Query:

查找Alberto Pepe所有朋友名字






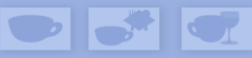






Step 1: The vertex.name index \rightarrow the vertex with the name Alberto Pepe. $[\log_2 n]$

Step 2: The vertex returned \rightarrow the k friend names. $[k + x)]$

搜索效率由 $(k + 1)\log_2 n$ 提高到 $\log_2 n$

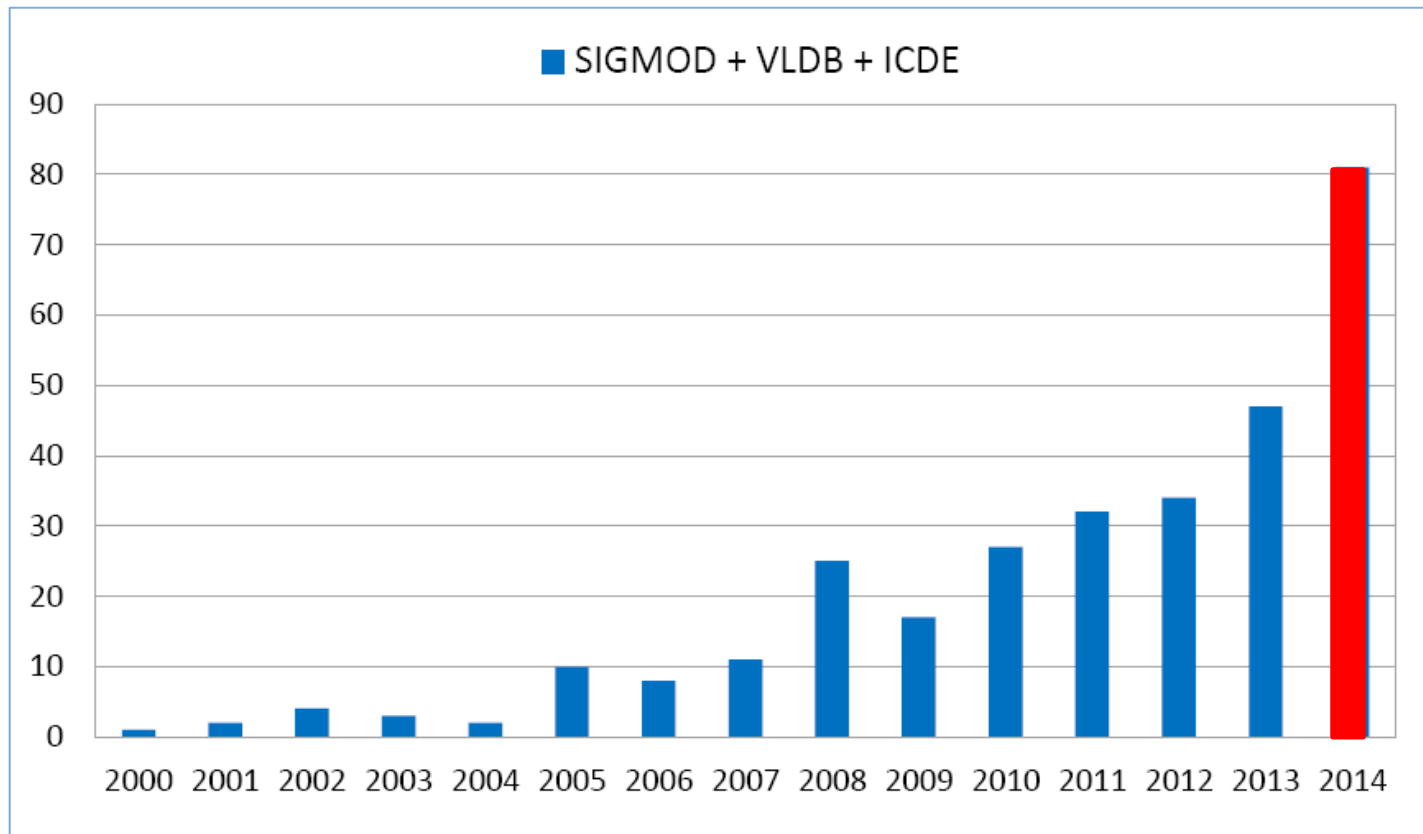
图搜索 vs. Web搜索



Google Search	VS	Graph Search
Keyword Based		Natural Language
 Friends near me FriendsNearMe.com A company in Guantanamo Bay providing entertainment for friends		 Friends near me 
 Best coffee shops Did you mean... <i>Dr. Best's coffee shop?</i>		 Best coffee shops 
 Interesting music Intersting's youtube channel Intersting is a popular music band that never produced a nice song		 Music my friends like 
 John Doe John Doe on Wikipedia Read about John Doe's history and legacy because everybody is a professor.		 John Doe  John Doe Email Phone Activities Photos Friends Family

- 关键字 vs. 短语、短句子
- 网页 vs. 实体(人, 社群等)
- 无生命特征 vs. 人的行为特征
- 历史 vs. 未来

学术界关注



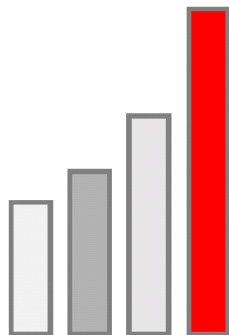
**Social computing
&
Web 2.0**



大图搜索的挑战性 (Challenges)



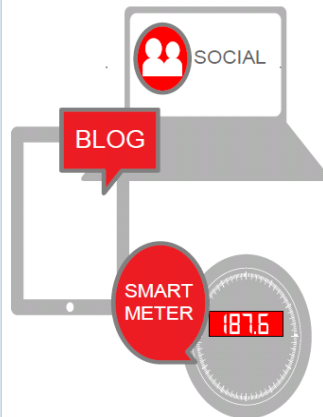
大图数据，如社会网络等



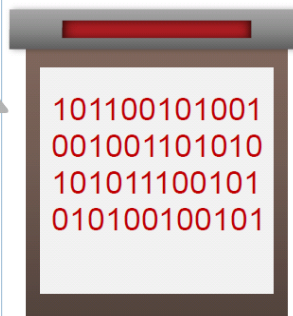
VOLUME



VELOCITY



VARIETY



VALUE

数据量大： 高效的图搜索需要在均衡查询性能与准确性

数据变化频繁： 融合数据的动态性和时间特征

数据丢失和不确定性： 提高数据的质量，减少负面影响。

FAE法则

- 在大量、动态和不确定图数据中：

- **F**: 如何提供**友好的**图搜索界面？
- **A**: 如何搜索“信息”**更准**？
- **E**: 如何搜索“信息”**更快**？



友好性(Friendliness)

- 如何以**友好的方式**提供“图搜索”的查询界面？
 - **关键字**的搜索模式非常友好
 - 直接让用户输入模式图非常不友好
 - 提供方便的方式**隐式**表达查询图
 - 如，Facebook采用简单化的自然语言



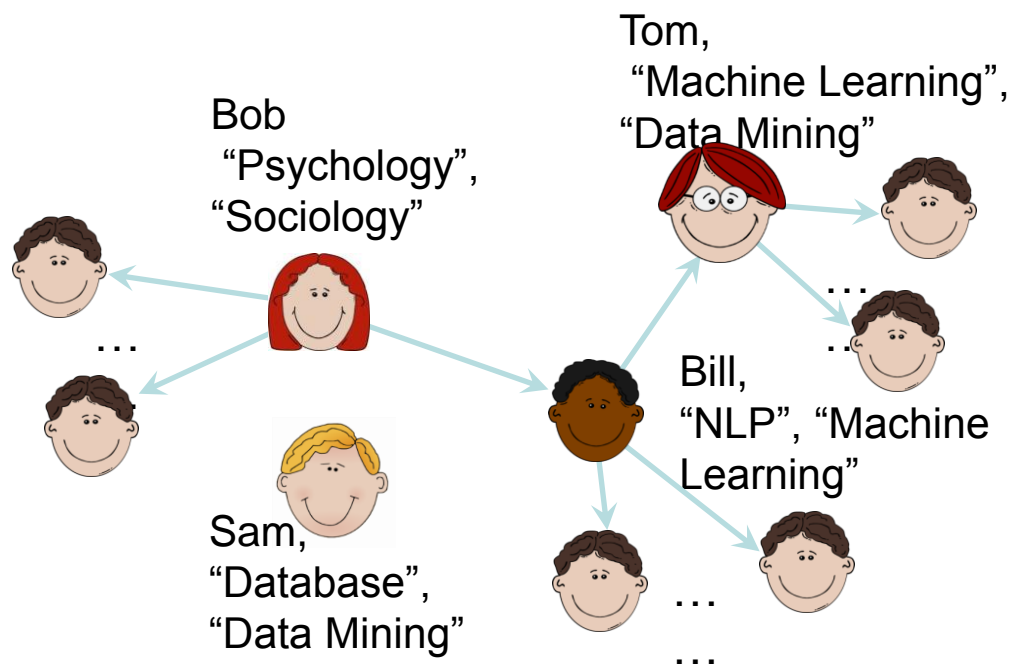
如，影响力事件组织者搜索

- 以关键字的方式搜索社会网络图中k个事件组织者
 - 融合了图上的关键词搜索
 - 融合了事件的影响力传播
 - 提出了具有性能保障的近似算法 - 近似比 $(1/2 - \xi)$

查询Q示例：

$K = 2$

$Q = \{\text{Psychology, Sociology, Data mining}\}$



准确性(Accuracy)

- 如何搜索“信息”更准？
 - 用户意图理解（融合用户的行为特征）
 - 融合知识图谱- Knowledge Graph
 - 基于知识/用户意图的查询转换

搜索北航周围的饭店

- 人在美国 vs. 人在北航
- 人在北航: 中午12点 vs. 半夜12点

搜索北航的信息：

- 北航、北京航空航天大学、北京航空航天大学、Beihang、Beihang University、Beijing University of Aeronautics and Astronautics

移动互联网

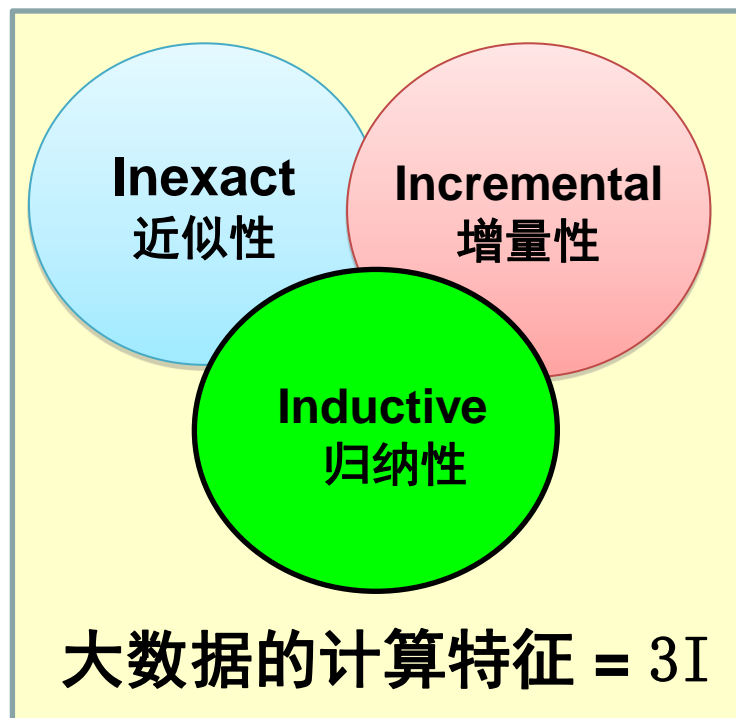


知识图谱



高效性(Efficiency)

- 如何搜索“信息” 更快?
 - 查询近似技术
 - 数据近似技术



$$R = Q(D)$$



天下武功 唯快不破



大图搜索的查询技术

(Query Techniques for Big Graph Search)

$$R = Q(D)$$

查询近似技术

主要思想： 对一类查询复杂性高的查询语言 Q ，变换为一类查询复杂性低的查询语言 Q' ，并且尽量不影响查询结果的准确性。



挑战： 平衡查询的复杂性和查询的准确性!

如，强模拟图查询



子图同构^[11]:

- 给定模式图 Q , 数据图 G 的子图 G_s :
 - Q 图同构 G_s 如果存在一一映射函数 $f: V_Q \rightarrow V_{G_s}$ 满足:
 - ✓ 对 Q 中的任何顶点 u , u 和 $f(u)$ 有相同的标签
 - ✓ 边 (u, u') 在 Q 当且仅当 $(f(u), f(u'))$ 在 G_s
 - Q 子图同构 G , 如果 G 中存在如上子图 G_s

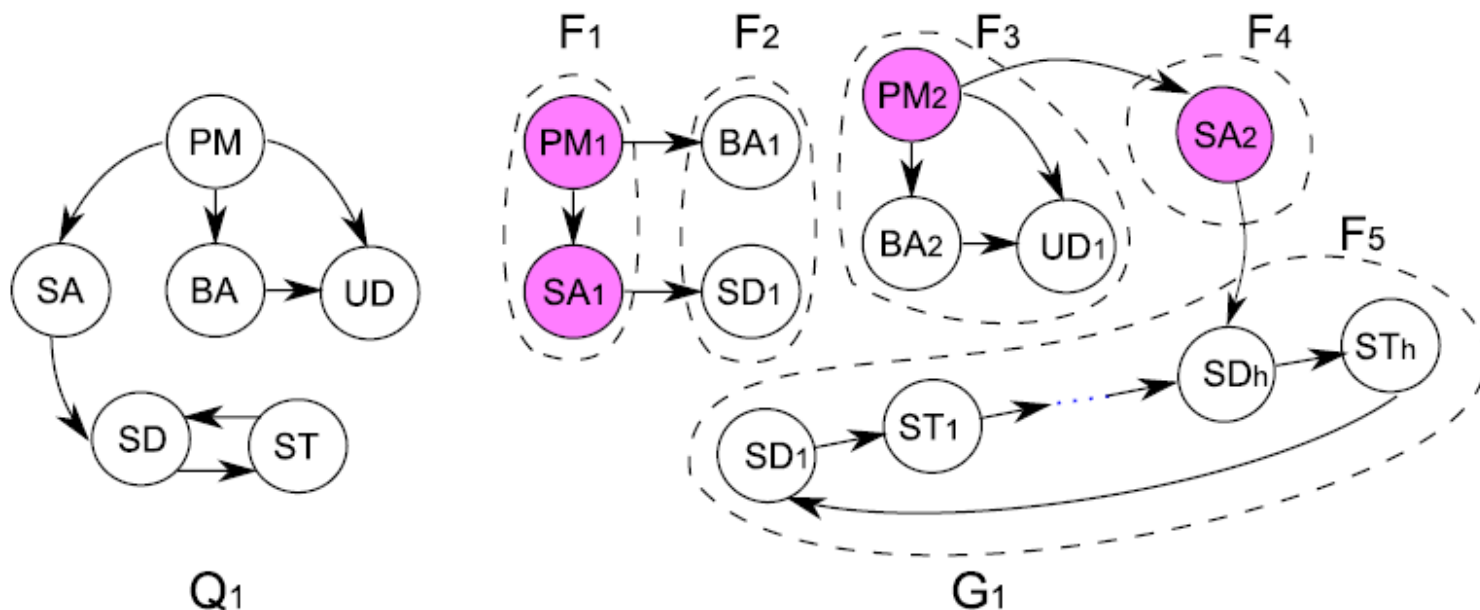
优点 : Q 和 G_s 一模一样

缺点 : NP完全问题 ; 最坏情况下指数个匹配子图 ; 约束过于严格

[12] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. Strong Simulation: Capturing Topology in Graph Pattern Matching. **TODS 2014**.

[13] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching. **VLDB 2012**.

子图同构图查询



组成一个软件开发团队

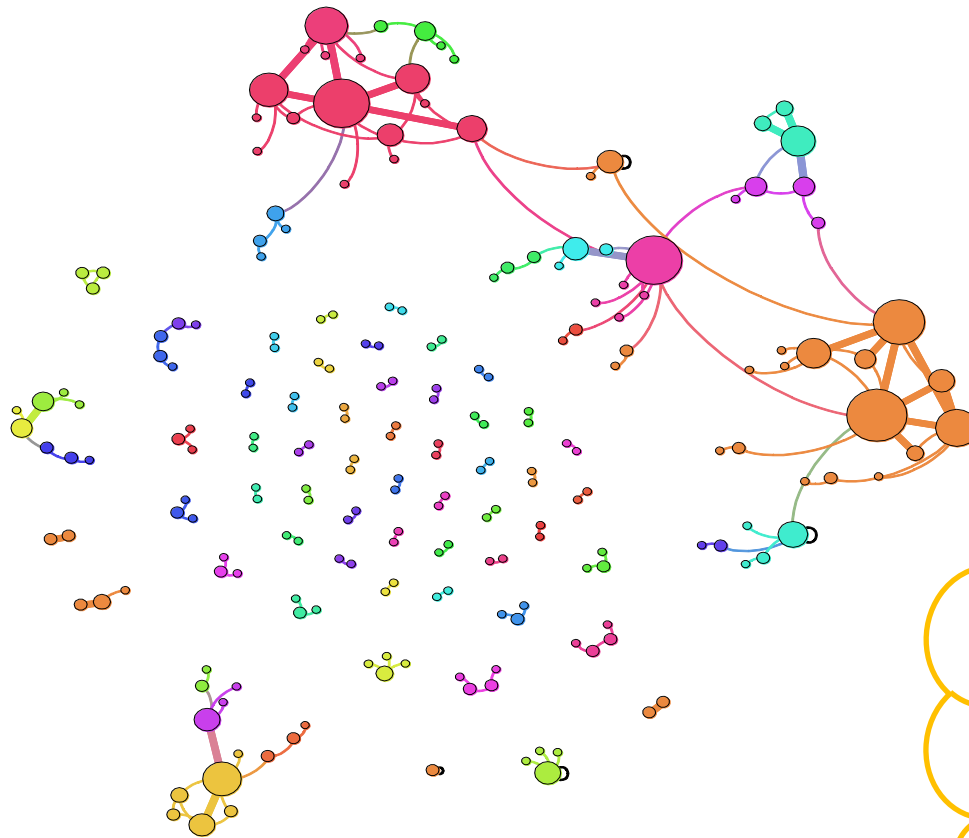
强模拟：返回 $F3 + F4 + F5$;

子图同构：返回空集！



子图同构约束过于严格，并不适合一些新型应用！

Terrorist Collaboration Network



**“Those who were trained to fly didn’t know the others.
One group of people did not know the other group.”
(Osama Bin Laden, 2001)**



强模拟图查询



Matching	children	parents	connectivity	cycles
\prec	✓	×	×	✓ (directed), × (undirected)
\prec_D	✓	✓	✓	✓ (directed & undirected)
\prec_D^L	✓	✓	✓	✓ (directed & undirected)
\triangleleft	✓	✓	✓	✓ (directed & undirected)

locality	matches	Bisimilar&b'ed-cycle
×	✓	×
×	×	×
✓	✓	×
✓	×	✓

查询结果保持70-80%子图同构结构，效率提高100倍！



大图搜索的数据技术

Data Techniques for Big Graph Search

$$R = Q(D)$$

数据近似技术

主要思想：对一类查询复杂性高的查询语言Q，将查询数据D变换机器能够高效处理的较小量D'，并且尽量不影响查询结果的准确性。

$$Q(D) \xrightarrow{\text{approximation}} Q(D')$$

二八定律：在众多现象中，80%的结果取决于20%的原因

$$D = \text{HARD}(D) + \text{SOFT}(D)$$

×

挑战：平衡查询的效率和查询的准确性！



如，链接预测

- 直接采用非负矩阵分解的代价高

- 效率低
- 数据越稀疏，效果越差

- 基于抽样的Ensemble方法

- 采样要保证一定的覆盖率

PROPOSITION 2. *The expected times of each node pair included in μ/f^2 ensemble components is at least μ .*

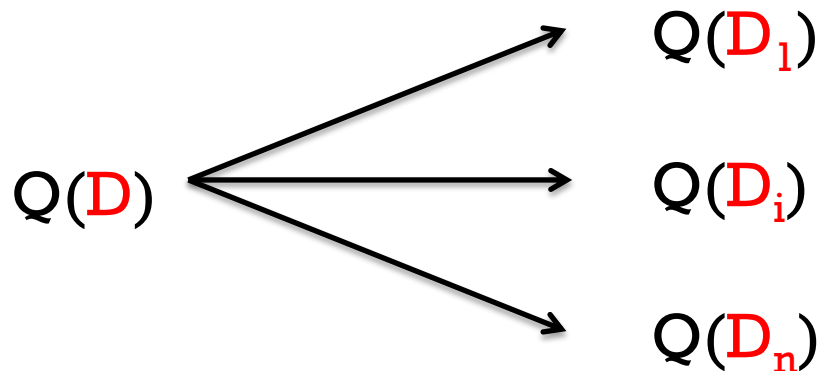
- 基于链接预测特征的抽样
- 结合Ensemble的思想：链接e的预测分值是所有Ensemble中的最大值

- 实验结果

小数据	准确性	大数据	效率
YouTube	高18%	Friendster	快31倍
Flickr	高4.4%	Twitter	快21倍
Wikipedia	高16%		



分布式数据处理技术

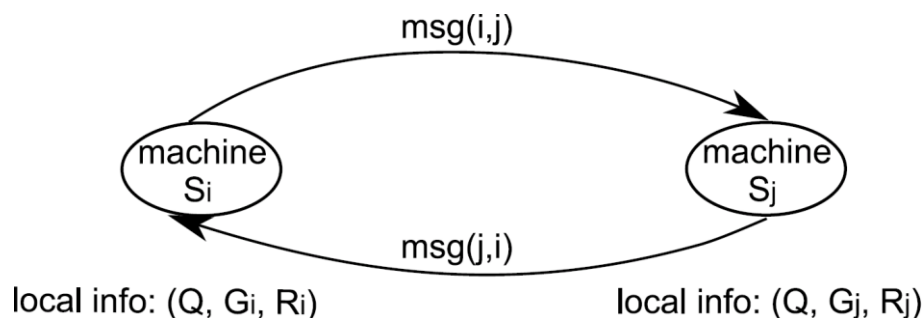


- 现实中的图通常非常大，使用单机来管理和查询图不现实：
 - Yahoo! Web图有140顶点
 - Facebook: 超过10用户
- 现实活中的图通常是分布式的：
 - Google, Yahoo! and Facebook都有大规模的数据中心存储数据ss

如，分布式图模式匹配

提出分布式计算模型 [2]:

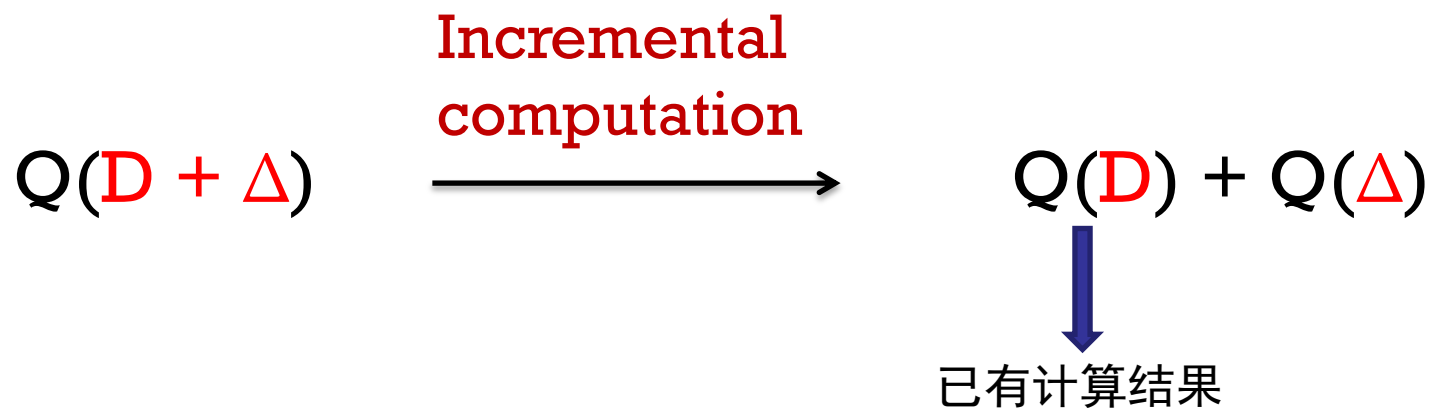
- 机群：具有**等同计算能力**的多台机器(发起查询的指定为**协调者**);
- 任何一台机器能够**直接**向其他机器发送**任意数量**的消息;
- 所有机器通过**本地计算**和**消息传送**协同完成任务.



分布式算法复杂性指标:

1. **机器访问次数**: 访问一台机器的最大次数(**交互复杂性**)
2. **最大完工时间**: 所有机器中最长的完工时间(**效率**)
3. **通讯数据量**: 不同机器之间的通讯消息的量和 (**网络带宽的消耗**)

增量计算技术



增量计算技术



如，增量模式匹配 (VLDB 2010 [6]):

- 提高效率，同时也是应对数据动态性的一种有效方法

Google Percolator [19]:

- 将索引系统改为增量的方法：
 - 将文档的平均处理时间减少为1%
 - 当每天处理的文档数据一样是，将文档的平均老化时间减少50%

从“零”开始是对计算资源的极大浪费!



其它数据技术

- 数据索引：空间代价、构建时间代价、查询效率提高

- 数据压缩：
$$Q(\mathbf{D}) \xrightarrow{\text{compression}} Q(\mathbf{D}')$$

- 数据划分：
$$Q(\mathbf{D}) \xrightarrow{\text{partitioning}} Q(\mathbf{D}_1) + \dots + Q(\mathbf{D}_n)$$

Work in progress !

Ring系统：凝聚理论、算法和技术



Ring典型应用

异常事件
分析与预警

微博
图片识别

微博
心情搜索



360度全面事件预警
one ring to rule them all





小结

图搜索是一种新型社会搜索模式

大图搜索的应用与挑战(FAE法则)

解决大图搜索挑战的相关技术

**Just a start,
there is a long way to go for Big Graph Search!**

Acknowledgements

Collaborators:

Charu Aggarwal, Sourav S Bhowmick, Yang Cao, Gao Cong, Liang Duan, Wenfei Fan, Kaiyu Feng, Renjun Hu, Jinpeng Huai, Jia Li, Jianxin Li, Xudong Liu, Haixun Wang, Tianyu Wo, Weiren Yu ...

They are from:



THE UNIVERSITY of EDINBURGH



NANYANG
TECHNOLOGICAL
UNIVERSITY



THE OHIO STATE UNIVERSITY



Microsoft

Research
微软亚洲研究院



facebook

References



- [1] Rice, M. and Tsotras, V.J., Graph indexing of road networks for shortest path queries with label restrictions, VLDB 2010.
- [2] Shuai Ma, Yang Cao, Jinpeng Huai, and Tianyu Wo, Distributed Graph Pattern Matching, WWW 2012.
- [3] Shuai Ma, Jia Li, Chunming Hu, Xuelian Lin, and Jinpeng Huai, Big Graph Search: Challenges and Techniques, Frontiers of Computer Science, 2015, to appear.
- [4] C. C. Aggarwal and H. Wang. Managing and Mining Graph Data. Springer, 2010.
- [5] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu, Adding Regular Expressions to Graph Reachability and Pattern Queries. ICDE 2011.
- [6] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu, Graph Pattern Matching: From Intractable to Polynomial Time. VLDB 2010.
- [7] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu, Graph Homomorphism Revisited for Graph Matching. VLDB 2010.
- [8] Hossein Maserrat and Jian Pei, Neighbor query friendly compression of social networks. KDD 2010.
- [9] Brian Gallaghe, Matching structure and semantics: A survey on graph-based pattern matching. AAAI FS. 2006.
- [10] Marko A. Rodriguez, Peter Neubauer: The Graph Traversal Pattern. Graph Data Management 2011: 29-46

References

- [11] Kaiyu Feng, Gao Cong, Sourav S. Bhowmick, Shuai Ma: In search of influential event organizers in online social networks. SIGMOD 2014.
- [12] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo. Strong Simulation: Capturing Topology in Graph Pattern Matching. TODS 2014.
- [13] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching. VLDB 2012.
- [14] P. Bogdanov, M. Mongiovì, and A. K. Singh, Mining heavy subgraphs in time-evolving networks, in ICDM, 2011.
- [16] Liang Duan, Charu Aggarwal, Shuai Ma, Renjun Hu, and Jinpeng Huai, Scaling up Link Prediction with Ensembles, WSDM 2016.
- [17] Weiren Yu, Charu C. Aggarwal, Shuai Ma, Haixun Wang: On Anomalous Hotspot Discovery in Graph Streams. ICDM 2013
- [19] Daniel Peng, Frank Dabek: Large-scale Incremental Processing Using Distributed Transactions and Notifications. OSDI 2010.



Homepage: <http://mashuai.buaa.edu.cn>

Email: mashuai@buaa.edu.cn

Address:

Room G1122,
New Main Building,
Beihang University
Beijing, China



Thanks!