第一章　绪论

1.1　研究背景与意义

随着科学技术的进一步发展，我们正逐步从信息时代走入数据时代[11]，全球的数据量正在以一种前所未有的方式增长着。统计表明，全球数据几乎每两年翻一番[10]。数据的迅速增长，在给人们带来便捷信息的同时，也带来了一个巨大的挑战——面对日益复杂的数据，传统的查询方法不再有效，无法快速检索出相关联数据。面对大量有意义的数据，无奈于查询手段的限制，只能将其简化再进行处理。现在的大数据现状就好似守着一座金山，却不知如何开采。为了进一步挖掘有效信息，加速查询速度，提高信息价值，各种数据查询技术便应运而生。

其中最为热门的就是图数据库的查询。图作为计算机科学中的一个数据结构，其数据表达能力较强，可以很好得表示各种关系特征，拓扑结构等，在许多领域都有应用。

举例而言，图数据在以下领域都发挥着显著的作用:

1. 在分子化学[4] 与生物[1] 领域，图可以很好得表示分子结构，用节点代表分子，节点属性为分子属性，边代表分子之间的化学键，边长等可以代表化学键键值等。利用图模型建立分子结构数据库，就可以利用图查询技术快速寻找相似的分子结构，来查询具有特定功能的分子集，如 $DAYLIGHT$ 系统[5] 作为一个商用的分子化合物数据库已经在业界被广泛使用。

2. 在地理信息系统领域[2,3,6]，利用图模型可以完整的表示出各实体之间的关系特征，然后利用这些特征实现进行许多功能，像拓扑关系建模，污染网络绘制，最短路查询[2] 等。

3. 在软件工程中，也可以利用图对程序代码进行建模，获得程序调用图，类关系图等，然后利用图相似性搜索可以进行易发故障判定或者代码剽窃检测[7] 等。

4. 在社会生活中，我们可以利用图结构对人际关系进行建模，进行社群侦查，行为预测等，甚至可以利用其进行犯罪团伙检测[8]。

5. 在 Web 领域，利用图结构进行 XML 文件分析，新闻的聚类分析等早被广泛使用。

除此之外，图数据还在像文献查重[9]，专家推荐系统等众多领域发挥着重要作用。

由此可见，图数据有着广泛的应用前景。因此如何对图数据进行有效管理与使用，愈加成为数据管理领域一个重大挑战。而如何快速检索图结构则是图数据管理中最为重要的一个问题，已备受研究人员的关注。

图查询种类很多，包括子图查询，超图查询，精确查询，近似查询等[12]。子图查询是给定查询图 q 在图数据集中找到所有包括 q 的数据图。子图查询是目前图查询中运用最为广泛的，也是研究最久的一种查询类型。如在生物信息学中[12]，当我们知道某一蛋白质分子结构 CA 是 HIV 病毒中的活跃成分，那就可以通过子图查询，将 CA 作为查询图，找出所有包含该分子结构的大型分子。而超图查询则与子图查询正好相反，是在图数据库中查找所有 q 的子图。精确查询与近似查询相对，可以是超图也可以是子图，主要是强调精确的图匹配过程，即查询图中每条边，每个点，每个关系要在数据图中存在。常用于分子结构匹配，关系网络匹配等领域。而近似查询则不需要完全匹配，只需要大致相似即可，而相似尺度则由各算法决定。近似查询在实际使用中有更大的应用空间，像地理信息系统的位置查询，机器视觉的人脸识别等。

1.2 本文的研究目的与内容

参考文献

[1] Holder L B and Cook D J. Graph-based data mining. *Encyclopedia of Data Warehousing and Mining*, 2005.

[2] Hutchinson D, Maheshwari A, and Zeh N. An external memory data structure for shortest path queries. *Tokyo:Springer Berlin Heidelberg*, 1999.

[3] Chan E P F and Zhang N. Finding shortest paths in large network systems. *Proceedings of the 9th ACM international symposium on Advances in geographic information systems*, 2001.

[4] Cook D J and Holder L B. Substructure discovery using minimum description length and background knowledge. *Proceedings of the National Conference on Artificial Intelligence*, 1994.

[5] C. A. James, D. Weininger, and J. Delany. Daylight theory manual daylight version 4.82. daylight chemical information systems, 2003.

[6] Jing N, Huang Y W, and Rundensteiner E A. Hierarchical encoded path views for path query processing: An optimal model and its performance evaluation. *Knowledge and Data Engineering,IEEE Transactions on*, 1998.

[7] OgataH, FujibuchiW, and GotoS. Aheuristicgraphcomparisonalgorithmanditsapplicationtodetect functionally related enzyme clusters. *Nucleic acids research*, 2000.

[8] Chua P. Catching bad guys with graph mining. *ACM*, 2011.

[9] 朱戈. 基于图的科技文献相似性搜索关键技术研究 [学位论文] 硕士. 黑龙江大学, 2011.

[10] 王海勋. 图数据管理与挖掘. 中国计算机学会通讯, 2012.

[11] (英) 维克托 · 迈尔 -舍恩伯格，肯尼思 · 库克耶. 大数据时代: 生活、工作与思维的大变革. 浙江人民出版社, 2012.

[12] 谭伟，杨书新. 图数据精确查询与近似查询的研究. 2013.