華中科技大學

本科生毕业设计（论文）开题报告

题　　目：大规模图计算中的最短路径优化问题

院　　系　计算机科学与技术学院

专业班级　信息安全 1301 班

姓　　名　王梦鸽

学　　号　U201315120

指导教师　陈汉华

2017 年 3 月

开题报告填写要求

一、 开题报告主要内容：

1.课题来源、目的、意义。

2.国内外研究现况及发展趋势。

3.预计达到的目标、关键理论和技术、主要研究内容、完成课题的方案及主要措施。

4.课题研究进度安排。

5.主要参考文献。

二、 报告内容用小四号宋体字编辑，采用 A4 号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。

三、 指导教师评语、教研室（系、所）或开题报告答辩小组审核意见用蓝、黑钢笔手写或小四号宋体字编辑，签名必须手写。

四、 理、工、医类要求字数在 3000 字左右，文、管类要求字数在 2000 字左右。

五、 开题报告应在第八学期第二周之前完成。

1 绪论

1.1 研究背景

在图论中两点之间的最短路径距离查询是一个非常基本的操作。目前，图论在解决实际问题中得到了广泛的应用，而最短路径距离查询在这些应用中占有重要的地位。例如，在语义网本体，实体间的距离可以判断他们的关系是多么的密切；在社会网络分析中，两个用户之间的距离可能涉及他们的亲密关系；在网页搜索，短距离的两个网页之间通过 Web 链接意味着他们可能更相似的内容。而其在信息安全方面的重要性更是不言而喻，很多基本的安全算法都需要构建图模型以及计算图中两点的距离。例如，保护用户隐私信息，加权社会网络是一种边上带有权重值的社会网络，其中权重值代表个体关系的强弱。如果加权网络数据未经处理发布会造成用户隐私泄露，而最短路径和节点中通常包含用户隐私；在入侵检测体系中，构建的图模型中的最短路径往往表示表示系统最有可能受到攻击的途径等等。因此，最短路径查询与信息安全密不可分，最短路径查询算法具有广泛而深刻的意义。

然而，在这些应用程序中使用的图的大小通常是在数以百万计的顶点和边的规模。使用传统的算法，如 Dijkstra 算法和广度优先搜索（BFS），在计算 m 条边的图中两个节点之间距离，需要 O（m）的复杂度。在许多实时应用中，长时间运行是不可接受的，它需要查询若干个距离，并用它们来计算或排序，并在微秒级的时间内返回结果。

建立索引的方法被认为是解决这个问题的有效方法。如果预先计算图中每对顶点之间的最短路径距离，并将它们存储在索引中，则可以立即返回查询结果。但用来存储这个索引的空间一般非常大，且可扩展性差。因此，一个有效的索引只存储部分的距离，它可以被看作是标签，查询可以在短时间内通过对相关标签进行运算得到。

在一些简单结构图中的最短路径距离查询，如道路网络图，许多有效的索引算法已经被提出，他们已经非常成功。但在复杂的网络中，大多数算法缺乏良好

的可扩展性，如果他们想要回答确切的距离，他们无法处理具有数以百万计的边的图，因为巨大的索引大小和索引时间。几种近似方法也被提出，但他们只回答近似的最短距离，从而应用范围受限。

1.2 国内外研究现状

最短路径距离查询是一项基本操作，已经研究多年。 Dijstra 算法是用于单源最短路径查询的经典算法。然而，在具有数百万个边的实时复杂网络图中的应用中，该算法对于其高查询时间是无效的。所以在最近的搜索中考虑索引方法。这个想法是在图中预先计算部分距离并存储在索引中，然后可以通过在更短的时间内简单计算来回答查询。

Cohen 等人首先提出了 2 跳标记法建立索引。对于图中的每个顶点，他们计算两个集合，Lin 和 Lout，它们存储距离的部分到这个顶点的最短路径。对于图中的每对顶点 u 和顶点 v，如果存在 w 在 w 到 u 的最短路径上的顶点 w，并且在集合 Lout（u）中已经计算了从 u 到 w 的距离，从 w 到 v 已经在集合 Lin（v）中计算，则可以精确地计算从 u 到 v 的距离。构建这样的索引，所有的最短路径距离可以被精确计算称为 2 跳盖问题。然而，要找到一个有效的算法来获得更小的索引大小与 2 跳覆盖是一个仍然是一个挑战的问题。

另一种具有相似想法的索引算法是分层集线器标签，其用于道路网络图查询。然而，与道路网络图的大多数算法一样，它依赖于特殊属性，例如道路图中的低公路尺寸。对于许多复杂的网络图形，例如社交网络，网络图形，RDF 图形，这些图形与道路网络完全不同。

为了更大的可扩展性，几个研究研究了图中的近似最短路径距离，使得它们需要更少的时间和空间来预先计算。然而，这些方法总是导致在复杂网络图中的低精度或更长的查询时间。

Takuya 等人提出了一个高效的算法称为 Pruned Landmark Labeling (PLL)，此算法通过建立两跳覆盖的索引来进行准确的最短路径查询。两跳覆盖的索引意味着两个顶点之间的距离都可以通过储存在索引中的两顶点中的标签集合来计算。每个集合存储的距离从一个顶点到其他顶点的一部分，称为一跳。找到最佳的两跳覆盖的最小尺寸的索引确定为 NP 难问题。PLL 算法从每个顶点执行 BFS，在

BFS 期间构建索引和剪枝，目的是减少索引空间和时间以获得更多的可扩展性。然而，PLL 算法认为具有高的度数的顶点可能是图中的良好的地标，如果他们首先做 BFS，则更多的最短路径可以通过它们，使得更多的距离可以被削减。事实上并不如此，因为一个高度数的顶点，不意味着它具有高介中性。此外，PLL 算法应用到大规模图中表现依旧不能令人满意，仍旧存在着索引空间开销大，索引时间长等诸多可待优化的问题。

1.3 研究目的及意义

我们希望能够针对大数据环境下图规模巨大，给传统最短路径计算算法带来的计算复杂性挑战，设计基于索引的最短路径算法，并在真实大图基础上验证算法性能。即，设计和实现基于索引的最短路算法，在较低的时间和空间复杂性下解决大图的最短路径计算的性能瓶颈问题。

2 课题研究方案

2.1 主要研究内容

考虑现有的最短路径距离查询算法所存在的问题，即索引空间开销大，查询时间仍可优化等，通过算法优化从而降低索引大小以及加快查询时间

2.2 解决思路

PLL 算法简单地通过 BFS 算法建立索引，在查询过程中通过求两点标签交集并相加取最小值的方法得到两点之间的最短距离。然而，值得注意的是，在 BFS 过程中我们可以获得很多路径信息，在同一条最短路径上的两个点本身在该路径上的位置差即为两点之间的最短距离。因此，如果记录下 BFS 过程中的部分最为有价值的路径，路径上所有点之间的两两之间的最短距离都可以通过为其加上路径标签，利用同一个路径标签相减得到。在这种情况下，两点的标签不仅可以相加，也可以通过新增一类标签的方法进行相减。更为重要的是，不用再记录路

径中任何两点之间的标签，因为通过新增的这一类路径标签相减即可得到。从而能够降低索引空间的开销。同时，当两点具有同样的路径标签时，可以不用进行求交集、比较、取最小值等一系列操作，只需要进行一次相减即可得到对应的准确最短距离，从而能后缩短部分查询时间。

2.3 有效性分析

对于一条具有 v 个顶点的路径，如果将其记录下来，在优化后的算法中只需要记录 v 个标签，而在 PLL 算法中，存在两种极端情况。

极端情况 1：该路径中的点按照 BFS 先后顺序排列，则任何一个点都会与其之后 BFS 的点产生一个标签，这时候 PLL 对于此路径会记录$\frac{v^2+v}{2}$个标签。

极端情况 2：该路径中均分点都是按照 BFS 选后顺序排列的，这时候由于进行过 BFS 的点两边的点不在需要记录标签，此时 PLL 对于此路径会记录 2v-1 个标签。

因此，对于优化算法的有效性分析如下所示。

优化后的算法	PLL	优化效率
v	极端情况 1：$\frac{v^2+v}{2}$	$\frac{2}{v+1}$
	极端情况 2：$2v-1$	$\frac{1}{2-\frac{1}{v}}$

2.3 研究方案

该解决思路下主要待攻克的难题就是怎样选取价值最大的路径从而能最大程度降低索引的大小，而该难题的最优解等同于最小覆盖子集问题。因而日后需要着重进行理论研究和实验测试，尽量能够接近最优效果。

路径的最大收益问题定义如下，

定义（路径的最大收益问题）：在一个图 G（V,E）中，有 M 条最短路径，如何在此之中选择一个路径子集，使得该子集中所有路径的收益是所有路径子集中收益最大的。

每条路径 Pj 的收益 Q（Pj）等于路径中每个点的收益 Q（v）的总和，即

$$Q(Pj) = |\sum_{v \in Pj} Q(v)| - |Pj|$$

其中,$Q(v) = Pj \cap V(v)$, V(v)表示点 v 的所有标签 label(v)中的顶点的集合，Pj 表示点 v 所在路径中所有点的集合。每选取一条路径后，V(v) 更新为 V(v)－Q(v)。

3 课题研究进度安排

表 1 课题研究进度安排表

学期	周次	工作任务
2017 第二学期	3 周——7 周	查阅相关文献材料
	7 周——9 周	提出初步可行算法
	9 周——11 周	编写实验代码
	11 周——14 周	算法优化并进行实验验证
	14 周——15 周	完成论文撰写

4 主要参考文献

[1] Takuya Akiba, Yoichi Iwata, Yuichi Yoshida. Fast exact shortest-path distance queries on large networks by pruned landmark labeling. Proceedings of SIGMOD 2013.

[2] I. Abraham, D. Delling, A. V. Goldberg, and R. F. Werneck. Hierarchical hub labelings for shortest paths. In ESA, pages 24–35. 2012.

[3] V. Agarwal, F. Petrini, D. Pasetto, and D. A. Bader. Scalable graph exploration on multicore processors. In SC, pages 1–11, 2010.

[4] T. Akiba, C. Sommer, and K. Kawarabayashi. Shortest-path queries for complex networks: exploiting low tree-width outside the core. In EDBT, pages 144–155, 2012.

[5] R. Albert, H. Jeong, and A. L. Barabasi. The diameter of the world wide web. Nature, 401:130–131, 1999.

[6] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In KDD, pages 44–54, 2006.

[7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. Physics reports, 424(4-5):175–308, 2006.

[8] P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: a multiresolution coordinate-free ordering for compressing social networks. In WWW, pages 587–596, 2011.

[9] P. Boldi and S. Vigna. The webgraph framework I: compression techniques. In WWW, pages 595–602, 2004.

[10] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. Physical Review Letters, 85:5468–5471, 2000.

[11] W. Chen, C. Sommer, S.-H. Teng, and Y. Wang. A compact routing scheme and approximate distance oracle for power-law graphs. TALG, 9(1):4:1–26, 2012.

[12] J. Cheng and J. X. Yu. On-line exact shortest distance query processing. In EDBT, pages 481–492, 2009.

[13] E. Cohen, E. Halperin, H. Kaplan, and U. Zwick.Reachability and distance queries via 2-hop labels. In SODA, pages 937–946, 2002.

[14] W. Fan, J. Li, X. Wang, and Y. Wu. Query preserving graph compression. In SIGMOD, pages 157–168, 2012.

[15] A. Gubichev, S. Bedathur, S. Seufert, and G. Weikum. Fast and accurate estimation of shortest paths in large graphs. In CIKM, pages 499–508, 2010.

[16] H. He, H. Wang, J. Yang, and P. S. Yu. Blinks: ranked keyword searches on graphs. In SIGMOD, pages 305–316, 2007.

[17] R. Jin, N. Ruan, Y. Xiang, and V. Lee. A highway-centric labeling approach for answering distance queries on large sparse graphs. In SIGMOD, pages 445–456, 2012.

[18] C. Jordan. Sur les assemblages de lignes. J. Reine Angew Math, 70:185–190, 1869.

[19] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In KDD, pages 137–146, 2003.

[20] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In WWW, pages 641–650, 2010.

[21] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In CHI, pages 1361–1370, 2010.

[22] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In KDD, pages 177–187, 2005.

[23] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics, 6(1):29–123, 2009.

[24] C. Magnien, M. Latapy, and M. Habib. Fast computation of empirically tight bounds for the diameter of massive graphs. J. Exp. Algorithmics, 13:10:1.10–10:1.9, Feb. 2009.

[25] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: a system for large-scale graph processing. In SIGMOD, pages 135–146, 2010.

华中科技大学本科生毕业设计（论文）开题报告评审表

<table>
<tr><td>姓名</td><td></td><td>学号</td><td></td><td>指导教师</td><td></td></tr>
<tr><td colspan="2">院（系）专业</td><td colspan="4"></td></tr>
<tr><td colspan="6">指导教师评语
1. 学生前期表现情况。
2. 是否具备开始设计（论文）条件？是否同意开始设计（论文）？
3. 不足及建议。</td></tr>
<tr><td colspan="6">

指导教师（签名）：

年　　月　　日</td></tr>
<tr><td colspan="6">教研室（系、所）或开题报告答辩小组审核意见</td></tr>
<tr><td colspan="6">

教研室（系、所）或开题报告答辩小组负责人（签名）：

年　　月　　日</td></tr>
</table>