

**硕士学位论文开题**

**文献综述报告**

**报告题目** 跨网络数据的表示研究

学 号 **\_\_\_\_\_**2120151008**\_\_\_\_\_\_**

姓 名 **\_\_\_\_\_**李懿**\_\_\_\_\_** **\_**

导 师 **\_\_\_\_\_**礼欣 **\_\_\_\_** **\_\_**

研究方向 **\_\_\_\_\_**网络数据表示 **\_\_\_\_\_**

二级学科 **\_\_\_\_\_**计算机应用技术**\_\_\_\_\_**

一级学科 **\_\_\_\_\_**计算机科学与技术**\_\_\_\_**

学 院 **\_\_\_\_\_**计算机学院 **\_\_\_\_**

2016 **年** 12 **月** 10 **日**

**文献综述报告要求与打印格式说明**

1. 文献综述报告应符合硕士研究生所在学科培养方案的要求。
2. 文献综述报告的内容不再在开题报告中重复。
3. 文献综述报告必须对相关领域已取得之成果进行归纳总结，结合学位论文选题对相关领域未来的发展和研究提出自己的观点。
4. 打印用纸：A4；
5. 页眉为“北京理工大学硕士学位论文开题文献综述报告”； 加黑宋体，小3号，居中。页码居右排版；
6. 页面设计：页眉2.5cm，页脚1.5cm，左边距3cm，右边距2.4cm，正文用宋体，小4号，行间距26磅。

硕士研究生文献综述报告评阅表

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 研究生简况 | 姓名 | 李懿 | | | 性别 | | 男 | 出生年月 | | 1992年8月 | |
| 学号 | 2120151008 | | | 入学时间 | | 2015年9月 | 身份证号 | | 350802199208160511 | |
| 学科、专业 | | | 计算机科学与技术 | | | | | | | |
| 本科毕业时间 | | | 2015年7月 | | | 本科毕业学校 | | 北京理工大学 | | |
| 指导小组 | | | 姓名 | | | 职称 | 工作单位 | | | | 签字 |
| 导师 | | | 礼欣 | | | 副教授 | 北京理工大学计算机学院 | | | |  |
| 小组成员 | | | 李凡 | | | 教授 | 北京理工大学计算机学院 | | | |  |
| 张子剑 | | | 讲师 | 北京理工大学计算机学院 | | | |  |
| 礼欣 | | | 讲师 | 北京理工大学计算机学院 | | | |  |
| 文献综述报告成绩 | | | | | | □ 优 □ 良 □ 通过 □ 未通过 | | | | | |
| 导师评阅意见 | | | | | | | | | | | |
| 作者阅读了大量的参考文献，对与本课题相关的国内外研究现状与发展趋势进行了全面细致的综述，结构研究，逻辑清晰，设立了明确的研究目标和研究内容。本课题的研究内容具有充分的创新性与实用性，本课题的研究方案和计划切实可行。  综上，开题报告达到硕士研究生开题的要求。  签字：  年 月 日 | | | | | | | | | | | |

**1网络表示的研究现状及其影响因素**

**1.1 大数据研究背景**

近年来，随着“云计算”和“物联网”等新鲜词语同时出现的还有“大数据”这个概念。不同于以往在计算机科学中强调算法和数理理论，大数据更加侧重于数据的收集、整理、加工和使用。我们现在身处一个数据爆炸的时代，几乎所有事物都能和数据产生联系，金融、医疗、社交、公共事物等等，我们每天都在产生巨量的数据[1]。这些因为人的使用而产生的数据也仅仅是数据海洋的一小部分，那些在互联网上由机器自动收集的信息才是构成海量大数据的主要组成，同时也是一个未被充分开采的珍贵数据资源。数据成为于自然资源、人力资源同等重要的战略性资源，引起各国政府、学界以及工业界的高度关注。

大数据的数据集大小以难以想象的速度增长，给数据处理带来了极大的挑战。据国际数据公司IDC报道[2]，2011年产生和复制的数据量超过1.8Z字节，是过去5年数据增长的9倍，并将以每两年翻倍的速度增长。预计到2020年，全球将拥有35ZB的数据量，并且85%以上的数据以非结构化或半结构化的形式存在。这些数据背后带来的挑战有一下几个方面：首先，信息技术的发展使得数据的生成和消费变得更容易。例如，每分钟有72小时长度的视频被上传到Youtube服务器[3]。大数据的这种大容量特性使得数据难以可伸缩地从分布式的地点收集并集成。第二，数据采集后，如何以最小的硬件和软件代价存储和管理这些海量异构数据是非常具有挑战性的问题。第三，由于大数据的异构性、规模性、实时性、复杂性和隐私性等特点，大数据分析必须在不同层次(建模、可视化、预测和优化)高效地挖掘数据以提高决策效率。

这些数据有四个主要特征：数据体量巨大（Volume）、数据种类繁多（Variety）、流动速度快（Velocity）、价值密度低（Value）。这些特性决定了传统数据挖掘和分析算法不适用于这样高复杂度的场景，数据被分散在不同的网站，分别使用不同的数据结构进行保存，造成了数据特征抽取的困难[4]。例如，社交网络中的用户关系可以被表示为一个巨大的网络结构，其中的点代表用户，边代表用户之间的关系。同时，用户发表过的言论也被存储在数据库中，独自形成一个巨大的数据网络[5]。显而易见，用户关系网络和用户数据网络的结构是完全不同的，但它们之间的信息却可以互相补充，提供给算法以便充分学习用户的特征表示。

随着大数据的发展，异构网络的跨网络数据分析和表示学习逐渐成为研究热点。异构网络的研究重点主要集中在网络结构的抽象表示、网络数据的联合学习、跨网络数据的融合应用，几个方面。美国早在1989年就首次提出了“数据库中的知识发现（KDD）”的概念，并于1995年召开第一届知识发现与数据挖掘国际学术会议，该会议现在已经成为全球的学术热点[6]。不仅是学术界，国外工业界的巨头IBM、ORACLE、微软、谷歌也都积极投身于新一代知识学习模型的探索。仅2009年，谷歌公司通过对大数据的使用，对美国经济贡献了540亿美元[7]；IBM在2012年以来，依靠大数据分析相关的研究，实现了股价的翻倍[8]。奥巴马政府也在2012年对外公布了“大数据研发计划”，旨在提高和改进人们从大数据中进行收集、存储、保留、管理和分析的核心技术。国内方面虽然起步较晚，但是近几年也呈现出对大数据和跨网络知识学习的高度关注。国家自然科学基金于1993年提出对数据挖掘领域进行相关研究，1999年在北京召开亚太地区知识发现和数据挖掘国际会议（PAKDD）并延续至今[15]。国内互联网企业也积极对各自掌握的海量数据进行学习和研究，百度、阿里、腾讯、网易等公司，都成立了独立的研发团队在数据挖掘和智能学习领域进行研究[16]。

在这种背景之下，对跨网络数据的表示研究可以帮助我们使用不同信息源、不同数据结构、不同时间跨度的数据，构建一个丰富的表达模型。这样的一种跨网络数据表示模型与传统的单数据源模型相比，使用更加丰富的数据进行实体的表示研究，同时使用互补信息实现了协同学习的目标。跨网络数据的表示研究在当前大数据背景下有非常突出的应用价值，也为基础理论研究提供了丰富的算法支持。

**1.2 国内外研究现状:**

网络信息的嵌入式表示(Network Embedding)，基于一个网络，将节点和边的信息投影到低维空间中，使得后续的机器学习模型可以使用这些低维表示进行预测任务。这种方法解决了大规模复杂网络的高信息量问题，取得了很好的效果[17]。目前基于跨网络信息表示的嵌入式表示方法主要有一下几种：

Perozzi 等人[8]在2014年提出了DeepWalk模型，该模型主要借鉴了信息检索领域的随机游走思想，实现了网络信息的重采样。在一个复杂网络上，算法首先从某个节点出发，随机选择自己路径上的下一个节点，重复n次后得到一条长度为n的节点序列。这样的一个节点序列类似于文本处理中的一条语句，在自然语言处理中，近年来最常使用的是基于skip gram[19]算法求解模型。通过将节点作为一个词，来度量词和词之间同时出现频率。可以认为，频率高的词更可能在新的未知语句中出现，借此可以得到对应的节点向量化表示，用于计算节点间的相似度。

Grover 等人[9]在DeepWalk模型的基础上，通过重新定义bias random walk策略的方式，提出了改进的node2vec模型。网络搜索中，有两种普遍采用的策略，深度优先搜索DFS和广度优先搜索BFS[20]。在node2vec中，引入了权重参数α调整深度优先搜索策略和广度优先搜索策略的比重，可以通过对不同类型的网络结构采用不同的策略，得到更加全面的网络路径采样。这种基于随机参数的采样结果，让算法可以适应多样化的需求。随后使用skip gram算法训练，可以得到节点的向量化表示。

国内的杨成和刘知远等人[10]在2015年提出了基于富信息的网络表示学习方法(Network Representation Learning with Rich Text Information)。与之前的DeepWalk模型不同，该模型使用了在实际应用中广泛存在的文本信息，辅助矩阵分解模型的求解。由于文本中包含了更加丰富的信息，使得节点表示的分解结果更加全面。

国内的唐杰等人[11]在2015年提出了适用于大规模信息网络的嵌入式表示模型(Large scale information network embedding, LINE)。LINE模型通过分析网络中的一阶邻居关系和二阶邻居关系，对节点之间的相似度进行了重新定义。这里一阶邻居指两个节点直接相连，且边上的权重越大两个点就更相似；二阶邻居指两个节点之间共享了很多邻居，说明相似度很高[21]。整个模型通过相似度和KL散度作为度量，计算得到每个节点最优的向量化表示。

**2 网络数据表示算法**

网络数据的表示（embedding）通常从网络的结构入手进行考虑。在一个网络中，节点之间依靠边互相连接，这些边代表的信息可能是社交网络中的互相关注、可能是文献中的共同作者、也可能是同时出现在用户购物车里的两件物品。总之，这里的边体现的是一种亲密关系，有边相连的节点于不相连的节点相比，有着更高的相似度。

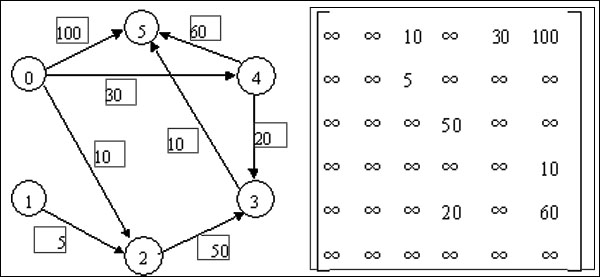
这里我们可以定义两种邻居信息：一阶相似度和二阶相似度。一阶相似度指的是有直接的边把两个节点连接到一起，它是一种局部的相关性。此时边的权重可以定义为边的强度，也就是归一化之后的出度值。这样的信息可以使用一个邻接矩阵很好地保存下来，矩阵的形式如图1所示。在矩阵中，行列中的元素代表的是节点，两个节点对应的值是连接的权值。二阶相似度保存的是节点周围邻居的信息，是一种全局的相关性。数学上，我们可以将一个节点的一阶邻居信息用一维数组进行表示 。那么节点 和 之间的二阶相似度就可以被看作是 和 之间的相似度。

图1 邻接矩阵

假设节点 和 之间存在一条边，一阶相似度可以用它们的联合概率计算，此时 是节点的低维向量化表示：

计算节点i和j 的二阶相似度，可以看作是在比较节点j的邻居信息和节点i之间的相似度，如果节点i的低维表示于节点j的邻居相似，则说明i和j存在二阶相似度，计算公式如下：

在此基础上，还有很多学者提出了基于异构网络的表示学习算法。

唐杰等人[12]在2015年，还提出了大规模异构网络的文本预测模型(Predictive Text Embedding through Large-scale Heterogeneous Text Networks, PTE)。这个模型主要需要将要预测的信息使用embedding的方法在最后一步提取出来，但是同时避免嵌套类似CNN和RNN这样的复杂预测模型。通过使用类似二部图的网络结构，将每个部分的损失函数累加到一起，得到结果。

黄志鹏等人[13]对多元异构网络和知识图谱进行嵌入式的表示学习，提出基于概率元路径的异构信息网络嵌入式表示学习模型（Heterogeneous Information Network Embedding for Meta Path based Proximity）。通常计算相似度时，都是使用两个节点直接相连作为标准，但是当引入元路径的概念后，可以将不同的节点之间的连边组装起来，把网络的嵌入式学习推广到其他场景中。

**3 深度学习框架**

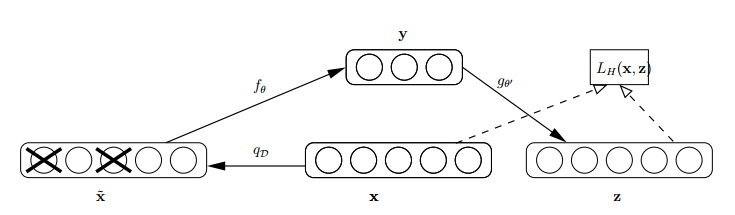
随着计算机硬件的不断革新，人工神经网络的计算速度得到了巨大的提升，与之相关的一系列算法也重新受到大家的重视。其中，对网络结构进行抽象化表示的深度学习框架就包括了降噪自动编码器（Denoising Autoencoder）。它的计算流程如图2所示。

图2 降噪自动编码器

在模型的复杂度和数据量都已经确定的前提下，防止过拟合的一种办法是减少数据中的噪音数量，即对训练集数据做清洗操作。然而，如果我们无法检测并删除掉数据中的噪音。另一种防止过拟合的办法就是给数据中增加噪音，这看似与之前的结论矛盾，但却是增强模型鲁棒性的一种有效方式。假如现在我们输入的是一副含有一定噪音的图片，例如图片中有污点，图片中的数字倾斜等，并且我们仍然希望解码后的图片是一副干净正确的图片，这就需要编码器不仅有编码功能，还得有去噪音的作用，通过这种方式训练出的模型具有更强的鲁棒性。降噪自动编码器就是在Autoencoder的基础之上，为了防止过拟合问题而对输入的数据（网络的输入层）加入噪音，使学习得到的编码器W具有较强的鲁棒性，从而增强模型的泛化能力[22]。

基于深度学习模型的在网络表示学习中也有广泛的使用，Wang D等人[14]在2016年提出了结构化的深度网络嵌入式学习模型（Structural Deep Network Embedding）。模型对网络节点描述为特征向量，具体可以写成邻接矩阵的形式，然后使用自动编码器进行编码，同时对于结果不为0的点进行惩罚。通过此种方法，可以得到两个节点之间的相似度表示。最后，将多个自编码器模型同时训练，得到网络节点的向量化表示。

**4 结论**

跨网络数据的表示研究可以使用使用深度学习方法和传统网络嵌入式表示学习方法进行构造，可以在社交、安防等领域帮助自动抽取大规模网络的内在表示，自动学习低维向量化的表示，以便提供给后续的机器学习模型使用。

**参考文献**

[1] Y. Wu, Y. Zhu, and B. Li, “Trajectory improves data delivery in vehicular networks,” in INFOCOM, 2011 Proceedings IEEE. IEEE, 2011, pp. 2183–2191.

[2] M.Uddin,M.Amin,and et al,“On diversifying source selection in social sensing ,”in Ninth International Conference on Networked Sensing Systems(INSS),2012,pp.1-8

[3]Luis G.Jaimes,Idalides Vergara-Laurens and Miguel A.Labrador, “A Location-based incentive mechanism for participatory sensing systems with budget constraints”, in IEEE. Percom,2012.

[4]Z.Feng,Y,Zhu，and et al, “Trac: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing”, in INFOCOM, 2014, pp. 1231 - 1239.

[5]L.Gao, F. Hou, J Huang, “Providing Long-Term Participation Incentive in Participatory Sensing”, In Proceedings of IEEE INFOCOM, 2014.

[6]S Reddy D Estrin and M Sruvastava, “Recruitment framework for participatory sensing data collections” in Pervasive Computing, ser. LNCS, 2010, vol. 6030, pp. 138–155.

[7]D Deng ,C Shahabi, and U Demiryurek,“Maximizing the number of worker’s self-selected tasks in spatial crowdsourcing ” In Proceedings of the 21st SIGSPATIAL GIS 2013 pp. 314–323.

[8] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 701-710.

[9] Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016: 855-864.

[10] Yang C, Liu Z, Zhao D, et al. Network Representation Learning with Rich Text Information[C]//IJCAI. 2015: 2111-2117.

[11] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[12] Tang J, Qu M, Mei Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1165-1174.

[13] Huang Z, Mamoulis N. Heterogeneous Information Network Embedding for Meta Path based Proximity[J]. arXiv preprint arXiv:1701.05291, 2017.

[14] Wang D, Cui P, Zhu W. Structural deep network embedding[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016: 1225-1234.

[15] C. Cheng, H. Yang, M. R. Lyu, and I. King, “Where you like to go next: Successive point-of-interest recommendation,” in IJCAI, 2013.

[16] H., W. "The Hungarian method for the assignment problem." Naval Research Logistics Quarterly2(1955):83-97.

[17] Patriksson, Michael. "The Traffic Assignment Problem---Models and Methods." Traffic Assignment(1994).

[18]O. Choi, S.Kim, J.Jeong, H.W.Lee, and S.Chong, “Delay-Optimal Data Forwarding in Vehicular Sensor Networks”, Arxiv preprint arXiv:1209.4527,2012

[19] L. Chisalita and N. Shahmehri, "A peer-to-peer approach to vehicular communication for the support of traffic safety applications", in Proceedings of 5th IEEE Conference on Intelligent Transportation Systems, Singapore, 2002, pp. 336-341.

[20] H. Hartenstein, B. Bochow, A. Ebner, M. Lott, M. Radimirsch, and D. Vollmer, "Position-aware ad hoc wireless networks for inter-vehicle communications: the Fleetnet project", in Proceedings of ACM Mobihoc, 2001, p.262.

[21] P. Gibbons, B. Karp, Y. Ke, S. Nath, and S. Seshan, "Irisnet: An architecture for a worldwide sensor web", IEEE Pervasive Computing, 2003, pp. 22-33.

[22] 彭曲,丁治明,郭黎敏.基于马尔可夫链的轨迹预测,计算机科学,37(8),2010.