**电 子 科 技 大 学**

**20 16 级本科毕业设计（论文）开题报告表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学号： 2016180201020** | | | **姓名：周良军** | **学院：资源与环境学院** | **专业：空间信息与数字技术** |
| **论文题目** | **基于Graph Embedding的机场航线预测研究** | | | | |
| **题目来源： 1.科研√　 2.生产　 3.教学（含实验） 4.创新创业**  **（在选项上打勾选择）** | | | | | |
| **成果形式： 1.硬件 2.硬件+软件 3.软件 4.理论研究√**  **（在选项上打勾选择）** | | | | | |
| **学位论文　　研究内容** | | **1、研究现状及发展态势**  真实的图（网络）往往是高维、难以处理的，20世纪初，研究人员发明了图形嵌入（Graph Embedding，下面简称图嵌入）算法，作为降维技术的一部分。他们首先根据实际问题构造一个D维空间中的图，然后将图的节点嵌入到d（d<<D）维向量空间中。嵌入的思想是在向量空间中保持连接的节点彼此靠近。  受到自然语言处理中word2vec方法的启发，通过DeepWalk方法来实现图嵌入成为当下热门，该方法首先选择某一特定点为起始点，做随机游走得到点的序列，然后将这个得到的序列视为句子，用word2vec来学习，得到该点的表示向量。DeepWalk通过随机游走去可以获图中点的局部上下文信息，因此学到的表示向量反映的是该点在图中的局部结构，两个点在图中共有的邻近点（或者高阶邻近点）越多，则对应的两个向量之间的距离就越短。  但DeepWalk方法所获得的嵌入质量和信息量偏低，时常无法满足要求，在数据挖掘领域的顶级会议KDD-2016上提出了一种新的node2vec方法，与DeepWalk相似，node2vec通过最大化随机游走得到的序列中的节点出现的概率来保持节点之间的高阶邻近性。与DeepWalk的最大区别在于，node2vec采用有偏随机游走，在广度优先（bfs）和深度优先（dfs）图搜索之间进行权衡，从而产生比DeepWalk更高质量和更多信息量的嵌入。  本文拟从图嵌入领域中的node2vec方法入手，将表示机场航线的数据转为向量，再通过机器学习来进行训练，实现对感兴趣的机场之间建立航线可行性的评估。  **2、选题依据及研究意义**  图，如社交网络、单词共存网络和通信网络，广泛地存在于各种现实应用中。通过对它们的分析，我们可以深入了解社会结构、语言和不同的交流模式，因此图一直是学界研究的热点。航线的建立需要考虑经济水平、文化差异、地理位置等诸多要素，传统的解决方案需要对庞大的参数和数据进行复杂的建模，将耗费巨大的人力物力。  本研究通过机器学习对已有的航线数据进行学习，将这些庞大而复杂的要素看作“不能被直接观察到，但是对系统的状态和能观察到的输出存在影响”的隐变量进行学习，可以大大简化预测的成本，避免了复杂建模过程中可能出现的参数设置不全、数据缺陷而带来的影响，从而实现更可靠的预测。  **3、课题研究内容**  （1）基于Graph Embedding技术的图降维研究  为了更好地利用原始数据，获得更高质量和更多信息量的嵌入，本研究拟采用node2vec方法，node2vec通过最大化随机游走得到的序列中的节点出现的概率来保持节点之间的高阶邻近性，需要通过设置合理的参数来决定下一个到访的节点。在实验过程中，需要不断地调参以获得最佳效果   1. 合适的机器学习模型选取   不同的机器学习模型适合处理不同的数据，基于目前本研究的样本数量和需求，可以选取决策树和深度学习模型进行尝试。  **4. 拟解决的关键问题和最终目标，以及拟采取的主要理论、技术路线和实施方案等**  **4.1拟解决的关键问题**  （1）如何有效地学习机场航线的图特征，将航线转化为含有丰富信息的高维向量  （2）对机场航线预测问题进行机器学习建模，对多模型进行实验与调优。  **4.2最终目标**  对已有航线数据的学习，产出航线特征信息，并且高准确率地预测未来航线或者未知航线。  **4.3主要理论**  node2vec方法：  node2vec依然采用随机游走的方式获取顶点的近邻序列，不同的是node2vec采用的是一种有偏的随机游走。给定当前顶点v，访问下一个顶点x的概率为    πvx是顶点v和顶点x之间的未归一化转移概率，Z是归一化常数。node2vec引入两个超参数p和q来控制随机游走的策略，假设当前随机游走经过边(t,v)到达顶点v，设πvx=αpq(t,x)⋅wvx，wvx是顶点v和x之间的边权，    dtx为顶点t和顶点x之间的最短路径距离。下面讨论超参数p和q对游走策略的影响:   1. Return parameter,p  参数p控制重复访问刚刚访问过的顶点的概率。注意到p仅作用dtx = 0的情况，而dtx = 0表示顶点x就是访问当前顶点v之前刚刚访问过的顶点。 那么若p较高，则访问刚刚访问过的顶点的概率会变低，反之变高。 2. In-out papameter,q   q控制着游走是向外还是向内，若q>1，随机游走倾向于访问和t接近的顶点(偏向BFS)。若q<1，倾向于访问远离t的顶点(偏向DFS)。    图1 从t访问到v时，决定下一个访问顶点时每个顶点对应的α  **4.4技术路线**  用航线建立图，用graphembedding提取路径，用word2vec学习特征表达，数据预处理（分数据），用机器学习训练与预测。  **4.5实施方案**   1. 从OpenFlights网站获取航线相关数据 2. 对数据进行分析、预处理 3. 对数据集进行划分：按照8:1:1的比例，将数据集划分为训练集、测试集和验证集 4. 使用决策树、深度学习模型进行训练、预测，并不断调优 5. 对模型进行评价   **5、论文特色或创新点**  将Graph Embedding技术用于机场的航线预测中，预计将简化建立新航线的流程，大大提高预测的准确性和效率。  参考文献：  [1] Aditya Grover, Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In KDD, 2016.  [2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In ICLR, 2013.  [3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.  [4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. IEEE TPAMI, 35(8):1798–1828, 2013.  [5]J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In EMNLP, 2014.  [6] S. Cao, W. Lu, and Q. Xu. GraRep: Learning Graph Representations with global structural information. In CIKM, 2015.  [7] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale Information Network Embedding. In WWW, 2015.  [8] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online learning of social representations. In KDD, 2014.  [9] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. RolX: structural role extraction & mining in large graphs. In KDD, 2012.  [10]P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, Verspoor, et al. A large-scale evaluation of computational protein function prediction. Nature methods, 10(3):221–227, 2013.  [11] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.  [12] K. Li, J. Gao, S. Guo, N. Du, X. Li, and A. Zhang. LRBM: A restricted boltzmann machine based approach for representation learning on linked data. In ICDM, 2014.  [13] X. Li, N. Du, H. Li, K. Li, J. Gao, and A. Zhang. A deep learning approach to link prediction in dynamic networks. In ICDM, 2014.  [14] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel. Gated graph sequence neural networks. In ICLR, 2016.  [15]S. Zhai and Z. Zhang. Dropout training of matrix factorization and autoencoder for link prediction in sparse graphs. In SDM, 2015. | | | |
| **导师审查意见** | | **签名：**    **日期：　 年　　月　 日** | | | |