学号： 153541

东 南 大 学

工程硕士学位论文开题报告

院（系、所） 软件学院

工 程 领 域 软件工程

研究生姓名 王海同

校 内 导 师 周德宇

校 外 导 师 彭艳兵

开题报告日期 2016.11.20

入 学 年 月 2015.9

　制

一、学位论文开题报告

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 论文题目 | | 基于语义网络的个性化实体推荐算法的设计与实现 | | | | | | | | | | | |
| 研究方向 | | 软件工程 | | | | | | | | | | | |
| 题目来源 | | 国家 | 部委 | 省 | 市 | | 厂、矿 | | 自选 | 有无合同 | | 经费数 | 备注 |
|  |  |  |  | |  | | √ |  | |  |  |
| 题目类型 | | 工程项目规划管理 | 工程设计 | 工程技术 | 技术攻关改造推广 | | 应用基础性研究 | | 其它 | | |  | |
|  |  | √ |  | |  | |
| 1. **课题来源、选题依据、课题研究目的、工程应用价值**    1. **课题来源**   本课题源自公司系统要求，推荐系统是公司业务的一个重要分支，提高推荐系统的准确性是公司内部一直追求的目标，本课题是在这个背景下提出的。   * 1. **选题依据**   传统的推荐算法大多数都是基于用户和物品两个纬度。而在实际的推荐系统环境中，存在很多用户和物品关系之外的信息。例如在电影推荐中，除去用户对电影的偏爱这一关系外，还存在电影的导演、演员和种类等各种隐含的信息。最近的研究表明，充分利用用户和物品关系以外的信息，能够显著提高推荐系统的准确性。近年来，有许多这方面的研究，但是其中大多数仅利用了单一的用户和物品关系以外的信息，例如很多研究把用户之间的社交关系加入到推荐系统中。本文提出一种基于语义网络的实体推荐模型，本模型充分利用推荐场景中的各种语义信息，将各种潜在的语义信息融合构成一个多源异构图，在该图的基础上利用图论和机器学习的一些算法挖掘隐含的用户和项目之间的联系，进而提高推荐的准确性。  目前在基于图的推荐算法中，大多数都是基于二元图和同构图。这些算法不能应用于有多个实体和多种关系的多源异构图上。本文提出的基于元路径图模型算法，可以解决多源异构图的问题，从而充分利用推荐环境中的隐含语义信息。 | | | | | | | | | | | | | |
| * 1. **课题研究目的**   本文通过扩展传统推荐算法常用的两个纬度，构造出一个多源异构图模型，在多源异构图模型的基础上，通过不同的语义信息构造多种元路径，基于这些元路径形成个性化实体推荐模型，进而解决了传统推荐算法中不能充分利用隐含的语义信息，数据来源单调和冷启动等问题，提高了推荐算法的准确性。  **1.4工程应用价值**  推荐系统中一个重要的推荐技术就是基于网络结构的推荐算法，随着互联网信息的增多，基于语义网络的推荐算法的应用价值越来越高，发展基于基于网络结构的推荐技术不仅能增加用户对网站的忠诚度，还能带来巨大的商业价值。  并且随着Depedia、CN-Depedia等LOD开放数据源的出现，推荐系统中隐含信息扩展变得更加方便，如何能有效的挖掘信息，为用户提供个性化服务，根据用户的兴趣爱好进行广告投放，都已经成为一个重要课题。一个高效的推荐算法不仅能直接带来巨大的利润，并且还能增加用户的黏性。所以也可以这么说，那个企业拥有了更能把握用户兴趣爱好和需求的推荐技术，那么这个企业也就更容易拥有更多用户，也就更容易占领市场，因此对推荐算法的研究具有巨大的实际意义和商业价值。 | | | | | | | | | | | | | |
| 1. **中外文献与资料阅读清单**（要求30篇以上） | | | | | | | | | | | | | |
| 篇数 | 34 | | | | | 其中外文文献与资料篇数 | | | | | 14 | | |
| 序号 | 题 目（及作者） | | | | | | | 出处（书籍及出版社、期刊名及期刊卷期号等） | | | | | |
| 1 | 王茜，段双艳. 一种改进的基于二部图复杂网络的推荐算法 | | | | | | | 计算机应用研究，2013， 30（3）：771-785. | | | | | |
| 2 | 赵新萍．基于图模型的个性化推荐技术研究 | | | | | | | 沈阳航空航天大学，2011． | | | | | |
| 3 | 朱郁筱, 吕琳媛. 推荐系统评价指标综述 | | | | | | | 电子科技大学学报, 2012, 41(2):163-175 | | | | | |
| 4 | 陈克寒, 韩盼盼, 吴健. 基于用户聚类的异构社交网络推荐算法 | | | | | | | 计算机学报. 2013,36(2): 349-359 | | | | | |
| 5 | 彭飞, 邓浩江, 刘磊. 加入用户评分偏置的推荐系统排名模型 | | | | | | | 西安交通大学学报,2012, 40(6):1-5. | | | | | |
| 6 | 王守崑. 透过产品看算法——走进个性化推荐系统 | | | | | | | 程序员, 2009(12):116-118. | | | | | |
| 7 | 项亮. 推荐系统实践 | | | | | | | 北京:人民邮电出版社, 2012 | | | | | |
| 9 | 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究 | | | | | | | 软件学报， 2009， 20(2):1-10. | | | | | |
| 10 | 曾燕 , 麦永浩 . 基于内容预测和项目评分的协同过滤算法 | | | | | | | 重庆三峡学院学报，2007, 3(4): 44-47 | | | | | |
| 11 | 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展. | | | | | | | 自然科学进展, 2009, 19(1):1-12. | | | | | |
| 12 | 李聪.基于领域最近邻的协同过滤推荐算法 | | | | | | | .计算机研究与发展，2008，45（9）：1532 - 1538. | | | | | |
| 13 | 赵培，陆平，罗圣美. 云计算技术及其应用 | | | | | | | 中兴通讯技术, 2010, 4: 9-11. | | | | | |
| 14 | 胡军，崔毅东. Open Stack 环境下计费系统的研究 | | | | | | | 中国科技论文在线, 2013. | | | | | |
| 20 | Xiao Yu†, Xiang Ren, Yizhou Sun .Personalized Entity Recommendation: A HeterogeneousInformation Network Approach | | | | | | | 2014，7th ACM Conference on Web Search and Data Mining, 283–292. | | | | | |
| 21 | Sun, Y., Han, J., Yan, X., Yu, P. S., Wu, T. PathSim: Meta PathBased Top-K Similarity Search in Heterogeneous Information Networks | | | | | | | 2011 Conference on Very Large Database Endowment, 992-1003. | | | | | |
| 22 | Rendle, S. Factorization Machines | | | | | | | 2010，10th IEEE International Conference on Data Mining, 995-1000 | | | | | |
| 23 | Singh, A. P., Gordon, Relational Learning Via Collective Matrix Factorization | | | | | | | 2008，14th ACM Conference on Knowledge Discovery and Data Mining, 650-658 | | | | | |
| 24 | Adomavicius, G., Kwon, Y.. Improving aggregate recommendation diversity using ranking-based techniques. | | | | | | | 2012，IEEE Transactions on Knowledge and Data Engineering, 24(5), 896–911. | | | | | |
| 25 | Jiang G, Chen H, Yoshihira K. Modeling and tracking of transaction flow dynamics for fault detection in complex systems | | | | | | | IEEE Transactions on Dependable and Secure Computing, 2006, 3(4): 312~326 | | | | | |
| 26 | Sotomayor B, Montero R S, Llorente I Metal. Virtual infrastructure management in private and hybrid clouds | | | | | | | Internet Computing, IEEE, 2009, 13(5): 14~22 | | | | | |
| 27 | Michael Cusumano. Cloud computing and SaaS as new computing platforms | | | | | | | Communications of the ACM, 2010, 53(4): 27～29 | | | | | |
| 28 | Saeed Araban, Leon Sterling. Quality of Service for Web Service | | | | | | | In: Proceedings of the 1st Australian Workshop on Engineering Service-Oriented System, Australia, 2004.54～56 | | | | | |
| 29 | Sushil Bhardwaj, Leena Jain, Sandeep Jain. Cloud Computing: A Study of Infrastructure as a Service | | | | | | | International Journal of Engineering and Information Technology, 2010,2(1): 60～63 | | | | | |
| 30 | M. Rosenblum, T. Garfinkel. Virtual Machine Monitors: Current Technology and Future Trends | | | | | | | IEEE Computer Society, May 2005,Volume: 30, issue: 7 :39～47 | | | | | |
| 31 | Paul Barham, Boris Dragovic, Keir Fraser et al. Xen and the art of virtualization | | | | | | | In: Proceedings of the nineteenth ACM symposium on Operating systems principles. USA: ACM Press, 2003. 164～177 | | | | | |
| 32 | Sepolinski, Thain. A Comparison and Critique of Eucalyptus, Open Nebula and Nimbus | | | | | | | In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science. USA: IEEE Press, 2010. 417～426 | | | | | |
| 33 | Sotomayor. Virtual Infrastructure Management in Private and Hybrid Clouds | | | | | | | Internet Computing, 2009, 13(5): 14～22 | | | | | |
| 34 | Dejan. Opennebula: A Cloud Management Tool. Internet Computing | | | | | | | Internet Computing, 2011, 15(2): 11～14 | | | | | |
| 1. **文献综述：国内外研究现状、发展动态**   推荐技术在上个世纪九十年代中期首次被提出， 它是一个典型的交叉研究领域，涉及信息科学、物理学、管理科学、运筹学等多门学科。被提出后就受到了很多专家学者的关注和研究， 在他们的努力下，推荐技术获得了飞速的发展。 在研究初期，对推荐系统的研究相对较为粗浅， 往往对用户行为的预测不准确，但是随着其在互联网各个领域中的广泛应用，使获得用户行为数据变得简单可行，加上研究者的创新算法， 推荐系统的功能迅速得到了改善， 但是目前仍然存在很多问题需要解决。  基于图(网络结构)的推荐算法在 KDD’ 99 中被 Aggarwa 首次提出[18]，结果表明基于图的推荐算法在准确度、可扩展性和效率等方面都优于传统的协同过滤算法。近年来，有学者将复杂网络的一些概念与方法应用到推荐算法的设计和分析中，提出了基于网络结构的推荐方法。由于推荐系统所处理的数据集，可以很好地表示为一个用户-项目二部图，因此可以将复杂网络中资源动态分配、热传导等方法应用于推荐技术中。  2007 年周涛、张翼成等人系统地对基于网络结构的推荐算法进行了分析和研究，认为个性化推荐在一定程度上等价于二部图向一部图的含权投影问题，并依此提出了基于二部图网络资源分配的推荐算法[11]；利用复杂网络动力学中物质扩散[19]和热传导[12]的原理提出了基于网络结构的推荐算法。同样有学者利用复杂网络物质扩散、 链路预测等方法来计算用户间的相似性[20-22]，并应用于协同过滤推荐算法中。 研究表明基于网络结构的推荐算法不但在算法推荐准确性上优于经典的协同过滤算法，而且在算法复杂性上也明显低于经典的协同过滤算法。  图模型在个性化推荐系统的应用，国外早在2000年前后就有研究者对其进行研究。1999年，Aggrarwal etaI．等人建立了用户之间的有向图，并把这个有向图引入到推荐系统中，其中的用户之间的边表明可预测性。SarwarB etal．在电子商务网站中引入了一种智能信息服务，进而进行推荐[121。2009年，You．JinPark et a1．把物品的信息和用户的历史行为信息作为用户模型的一种体现，然后构建了基于用户的模型，进而进行个性化推荐。  尽管推荐系统的应用已经取得了巨大的成功，但是还存在着一些问题需要解决， 还需要不同领域的研究人员对其进行进一步完善和发展。可以说，推荐算法的研究任重而道远。 | | | | | | | | | | | | | |
| 1. **研究基础** 2. 熟悉推荐算法的基础理论知识，理解经典的推荐算法，如协同过滤推荐算法，基于内容的推荐算法。 3. 熟悉基本的机器学习算法，如矩阵分解，K-means聚类等。了解比较深入的机器学习算法，如谱聚类。 4. 了解图论的基本知识。 5. 熟悉Python，能够使用Python进行数据分析和编码工作。 | | | | | | | | | | | | | |
| 1. **主要内容与关键技术**    1. **主要内容**   本文提出了一种基于多源异构网络的个性化实体推荐模型。本模型融合了用户-物品之间的关系以及其他实体之间的关系。该模型基于多种语义关系，通过对应的多种元路径将用户对物品的已有的喜爱关系传播出去，根据扩展的PathSim算法得出相应的传播偏爱分数，填充原有评分矩阵，从而得到与多种语义关系对应的多个传播评分矩阵。之后采用矩阵分解技术将所有传播矩阵分解，得到多组用户矩阵和物品矩阵。采用贝叶斯最优排序算法计算得出每组用户矩阵和物品矩阵在权重参数，基于这些矩阵和对应的权重参数计算得出用户对所有物品的综合评分，对这些评分进行排序，取Top-K作为推荐结果。 为了进一步区分用户的兴趣爱好，本文在全局推荐模型的基础上，通过谱聚类算法将用户分类，对每类用户建立相应的全局推荐模型，根据用户所涉及的用户种类以及这些用户种类与中心分类之间的关联性，计算得出用户对每个物品的评分，取top-k作为推荐结果。  本文的主要内容总结如下：   1. 本文提出了一种基于多源异构网络的推荐模型。 2. 本文将用户对物品的偏好程度通过多源异构网络中的各种语义元路径传播出去，进而得到用户和其他无直接联系物品之间潜在的偏爱程度，解决了冷启动问题。 3. 本文基于多种语义传播评分和对应的权重得出了全局推荐模型。 4. 本文在全局推荐模型的基础上，通过谱聚类将用户分组，进而提出了个性化推荐模型。 5. 本文在两个真实数据集IMDbMovieLens-100K 和Yelp做实验。    1. **关键技术和难点**   **（1）推荐模型中权重参数的确定**  　为了计算推荐模型中的各语义矩阵的权重参数，本文基于已有的用户-物品矩阵采用贝叶斯最优化排序模型估计参数  贝叶斯公式的最优化准则是使得后验概率最大化。  公式1  其中θ=｛θ1,…θL｝表示全局推荐模型中的权重参数，p(R|θ)表示物品被正确排序的可能性。  假设所有的用户评分和被排序的物品之间是相互独立的 ，我们可以扩展概率函数  p(R|θ)如下所示:    公式2  其中（ea> eb）代表了所有基于用户ui已有评分中具有正确顺序的物品集。  我们定义p(ea> eb;ui|θ) 如下：  公式3  其中σ是sigmoid函数 。  假设p(θ)服从均值为0协方差矩阵为的高斯分布，根据上述概率定义，我们可以得到目标函数如下：  公式4      通过最小化目标函数，我们能够计算出推荐模型中的参数θ。  目标函数O是可微的，可以使用很多最优化算法（例如SGD,BFGS-B）来估计参数θ，目标函数O的梯度计算如下：  公式5  其中ri,ab=r(ui,ea)-r(ui,eb)  **（2）谱普聚类在多源异构网络中的应用**  个性化推荐模型需要现将用户分类，本文采用谱聚类的算法对用户进行分类。谱聚类(Spectral Clustering, SC)是一种基于图论的聚类方法——将带权无向图划分为两个或两个以上的最优子图，使子图内部尽量相似，而子图间距离尽量距离较远，以达到常见的聚类的目的。本文是的推荐模型是基于多源异构图模型的，这个特点使得本文更适合采用谱聚类进行兴趣社团分割。  谱聚类的步骤：  第一步：数据准备，生成图的邻接矩阵；  第二步：归一化普拉斯矩阵；  第三步：生成最小的k个特征值和对应的特征向量；  第四步：将特征向量kmeans聚类(少量的特征向量)； | | | | | | | | | | | | | |
| 1. **研究方向、实施方案及可行性分析**    1. **研究方向**   本课题的研究方向基于语义网络的个性化推荐,是结合图论和机器学习算法改进推荐的一种方式。 | | | | | | | | | | | | | |
| * 1. **实施方案**   **（1）构造多源异构网络**  一个信息网络可以用一个有向图G(V,E)、一个实体种类映射函数φ:V→A和一个关系种类映射函数ψ:E→R表示，对于任意实体v∈V,都有φ(v)∈A,同样，对于任意l∈E,都有ψ (l )∈R。当︱A︱>1或者︱R︱>1时，我们把这种信息网络称为多源异构网络。类似于数据库中的E-R图，我们采用一种抽象的图GT(A,R)来表示多源异构网络中实体种类和关系种类。具体例子如下图所示。    **（2）分析网络中的元路径**  在多源异构网络中，两个实体能够通过多种不同的路径连接。由于实体和关系的多样性，这些路径上可能包含多种实体和关系，以及不同的顺序和长度。为了描述路径的种类，本文定义了元路径。元路径P=A0 A1 … Ak 是网络GT(A,R)中的一条路径，它定义了A0和Ak之间的一种新的合成的关系R1R2…Rk,其中Ai∈A，Ri∈R，(i=0,…,k)，A0=dom(R1)=dom(P)，Ak=range(Rk)=range(P)，Ai=range(Ri)=dom(Ri+1)，(i=1,…,k)。其中dom(.)表示某个映射关系函数Ri的定义域，range(.)表示值域。具体如下图所示。  Rk  R2  R1    图2 不同的元路径  观看  观看-1  观看  观看-1  观看  P1：用户 电影 用户 电影 用户 电影  主演  主演-1  主演  主演-1  观看  P2：用户 电影 演员 电影 演员 电影  本文给出了图2中两条元路径的例子，其中图2中的蓝色实体直线代表P1，红色实体直线代表P2。这两条元路径基于两种不同的语义假设把用户和电影连接起来。P1通过找到和目标用户看过同一部电影的用户扩展出去。P2则利用电影和主演电影的演员这一关系建立起用户和他没有看过的电影的关系。通过计算基于不同元路径的用户和电影之间的接近度，我们可以从不同的语义角度向用户做推荐。  **（3）计算用户偏爱度传播分数**  在众多形式的元路径中，本文提出的推荐模型将集中在形式为user – item – \* – item的元路径上。 这种形式的元路径可以把用户对有直接联系的物品的偏爱程度通过各种元路径传播到没有直接联系的物品上，从而使用户与其他物品产生联系。为了衡量这种潜在语义联系的可能性大小，本文定义了偏爱传播分数。  给定一条元路径P=R1R2…Rk,其中dom(P)=user,range(P)=item,令P’= R2…Rk，其中dom(P’)=item,range(P’)=range(P)=item。本文定义user i和item j 之间基于元路径P的偏爱传播分数如下：  公式6  其中代表e和ej之间的一条路径，代表e和e之间的一条路径，代表ej和ej之间的一条路径。  用户ui和物品ej之间的偏爱传播分数包含两部分：(1)所有显示存在的用户ui和物品之间的关系。（2）所有用户已知感兴趣的物品和潜在感兴趣的之间的路径。  **（4）全局推荐模型**  根据元路径P，计算用户和所有物品之间的偏爱传播分数，我们可以得到一个用户偏爱矩阵S∈Rm\*n，其Si代表用户ui基于元路径P，扩展了新的语义（例如根据用户喜爱的电影类型，或者用户喜欢的导演）后的所有可能的偏爱评分。基于L种元路径重复上述过程，可以计算出L个相应的评分矩阵，我们定义这些矩阵为S(1)，S(2)….. S(L)。使用S(q)代表基于第q条元路径的评分矩阵。通过矩阵分解技术，对矩阵S(q)分解如下:  公式7  其中U(q) ∈Rm\*d，代表了用户矩阵，V(q) ∈Rn\*d表示物品矩阵。  对L个评分矩阵进行如上的矩阵分解，可以得到L组用户和物品矩阵（U(1),V(1)），….（U(L),V(L)）。每组矩阵代表了不同元路径对应的语义特征。在定义推荐模型的时，不同的语义有着不同的重要性。例如用户在选择电影是会更加注重演员是谁，而不是这部电影是那个公司出版的。所以我们定义全局推荐模型如下：  公式8  其中θ(q)是第q组用户和物品矩阵的权重。  根据上面公式给出的推荐模型，给定一个用户，能够计算出他和所有物品之间的推荐分数，之后对这些分数进行排序，取top-k作为推荐结果。  **(5)个性化推荐模型**  全局推荐模型是将一个推荐模型应用到所有用户。但是这样不能把用户区分开来。例如，全局推荐模型会把最流行的电影，以及由大牌明星出演的电影推荐给用户，但这并不适用于所有用户。为了进一步提高推荐的准确性，本文将全局推荐模型进行扩展，根据每个用户不同的偏好和兴趣，为他们建立相应的个性化推荐模型。  尽管不同的用户有着不同的兴趣和偏好，但是从某个角度出发，一小群人可能具有相似的兴趣。例如喜爱漫画的用户喜欢超级英雄，冒险类的电影。喜欢Steven Spielberg的用户，喜欢他导演的电影。基于这一发现，本文首先根据用户的兴趣将用户分类，然后对每一类用户应用一个推荐模型。由于用户可能有多种偏好（一个人既是漫画爱好者，又喜欢Steven Spielberg），所以在推荐时，本文会融合所有与用户相关的分类。具体如下所示：  公式9  其中，C表示与用户ui相关的分类的，sim(.,.)定义中心集群Ck和ui之间的余弦相似性。  在对每个用户分类使用全局推荐模型估计完参数后，当我们为ui推荐物品时，我们首找到与用户ui的相关用户分类（与中心集群有较高的相似性），然后根据上面公式中的参数把这些用户融合在一起。对计算所得的所有物品的推荐分数排序，取Top-K最为推荐结果。具体算法流程如下表所示。    **个性化推荐模型学习算法**    **//** Input：评分矩阵和信息网络  **//** Output:每个用户分类的推荐模型  Input R,G  Output:θ{.}  提取L条形式为user - item- ∗ - item的元路径。  **//** 用户偏爱程度沿各元路径传播  For q ← 1 to L do  Forearch ui and ej do  (公式6)  End  矩阵分解得到U(q),V(q)  End  **//** 对用户进行分类  C= Spectral Clustering(R)  **//** 学习推荐模型  For each Ck in C do  使用用户子集Ck中的评分数据使得参数θ{k}最优（公式4）  End | | | | | | | | | | | | | |
| **6.3可行性分析**  目前推荐算法相关技术已经比较成熟，经过实习期间对推荐算法的了解以及对相关理论知识的学习，实现去基于多源异构网络的个性化推荐模型具有可行性。 | | | | | | | | | | | | | |

二、论文工作进度与安排

|  |  |  |  |
| --- | --- | --- | --- |
| 起讫日期 | | 工 作 内 容 和 要 求 | 备 注 |
| 2015.7.13-2015.8.31 | | 系统了解OpenStack、KVM等云计算相关技术，为开题做准备 |  |
| 2015.9.1-2015.10.1 | | 初步确定开题方向，查找相关资料，提出问题并寻找相应解决方案。 |  |
| 2015.10.2-2015.11.30 | | 编写开题报告，并根据导师意见进行修改，完成开题 |  |
| 2015.12.1-2016.3.10 | | 完成论文初稿，对实现的系统进行测试验证。 |  |
|  | |  |  |
|  | |  |  |
| 学校指导教师对开题报告的综合意见 | 云计算是目前IT业界较为流行的技术，成为很多企业采用的流行技术之一。然而，云计算给企业和客户带来IT资源使用便利的同时，也存在着如何有效监控云平台的内部状态，这也成为了很多应用云平台企业面临的挑战。作者课题结合企业实际项目，对云环境下数据采集与分析系统展开研究和实现，具有一定的技术新进性和良好的工程应用价值，同意开题！  指导教师（签字）  年 月 日 | | |
| 校外导  教师对开题报告的综合意见 | 指导教师（签字）  年 月 日 | | |
| 开  题  报  告  审  议  情  况  记  录 | １、审议小组意见（审议小组成员3-5人）  ２、投票表决结果  审议小组出席 人  通 过 人  不 通 过 人  ３、审议小组组长（签名）\_\_\_\_\_\_\_  审议小组成员（签名）\_\_\_\_\_\_\_ \_\_\_\_\_\_\_ \_\_\_\_\_\_\_  \_\_\_\_\_\_\_ \_\_\_\_\_\_\_  年 月 日 | | |
| 院（系、所）意见  负责人（签字或印章）  年 月 日 | | | |