计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400028

融合语义模型的二分网络推荐算法

周　波
中国科学院科技战略咨询研究院　北京 100190
(806446828@qq.com)

摘　要　当前基于二分网络的推荐算法未考虑推荐对象之间的语义关系,因此文中提出一种融合语义模型的二分网络推荐算法。该算法利用作者主题模型将推荐对象的语义信息降维至二维向量空间;然后计算推荐对象之间的语义相似度,把该语义相似度融合到基于物质扩散的二分网络推荐算法中。以新能源汽车专利权人推荐为实例进行实验验证,结果表明,该算法相比于单一的二分网络推荐算法具有更高的准确率和召回率,准确率提高比率为2.29%,召回率提高比率为4.15%。

关键词:语义模型;作者主题模型;二分网络;推荐算法
中图法分类号　TP391

Bipartite Network Recommendation Algorithm Based on Semantic Model

ZHOU Bo
Institute of Science and Development,Chinese Academy of Sciences,Beijing 100190,China

Abstract　The current research of bipartite network recommendation algorithm does not consider the semantic relationship,so this paper proposes an improved bipartite network recommendation algorithm. Author topic model (AT model) is used to embed the semantic information into a two dimensions semantic space. Then the semantic similarity between the recommended objects is calculated and integrated into the similarity calculation of bipartite network recommendation algorithm. The algorithm is verified by the recommendation of the new energy vehicle patentee. Experimental results show that the new algorithm has higher accuracy and recall rate than the bipartite network recommendation algorithm,the accuracy rate is increased by 2.29%,the recall rate is increased by 4.15%.

Keywords　Semantic model,Author topic model,Bipartite network,Recommendation algorithm

1　引言

在大数据时代,推荐系统尤为重要。在大数据时代,面对过载的海量信息资源,推荐算法至关重要。推荐算法是为人们提供信息的重要途径。协同过滤算法是商业推荐中应用最成熟和最广泛的推荐技术,该算法于1992年由施乐公司Palo Alto研究中心首先提出并实现具有协同过滤特性的邮件推荐系统Taestry;1994年GroupLens的研究团队将协同过滤推荐算法应用在Usenet新闻推荐服务中;2003亚马逊将协同过滤算法应用到商品销售系统中。

当前诸多的学者对基于语义模型的协同过滤推荐算法进行了研究,如Xiao等[1]提出基于项目语义相似度的协同过滤算法;Wu等[2]提出基于本体语义相似度的协同过滤算法;Wang等[3]提出基于知识图谱语义相似度的协同过滤算法[1-3]。

基于二分网络的推荐算法由Zhou等[4]基于复杂网络理论而提出[4],该算法提出后受到广泛关注。基于二分网络的推荐算法较传统的协同推荐算法有着更高的推荐精度。然而,当前对于二分网络推荐算法主要集中在基于网络结构的改进优化[5-7],鲜有学者对基于语义模型的二分网络推荐算法进行研究。Zhang等[8]提出融合IF-IDF的二分网络推荐算法,然而该算法未考虑文本之间的语义关系。因此,本文将对基于语义模型的二分网络推荐算法展开研究,以弥补当前研究的不足。

2　相关理论

2.1　基于物质扩散的二分网络推荐算法的基本原理

该算法假设每个对象均有一定的初始资源,通过对象的度将资源平均地分配给相邻的用户,然后每个用户将自己分到的所有资源再次平均地分配给所选择的对象,通过汇总对象的所有相邻用户分配的资源,得到该对象获得的资源[7]。现已经证明通过初始资源在网络上的扩散原理进行链路预测,比协同过滤算法具有更高的精度[4]。基于物质扩散的算法不考虑用户和对象的内容特征,只是把它们抽象成二分网络的节点。其推荐过程如下。

用X,Y分别表示二分网络两类节点用户和物品的集合,其中X类节点有m个,Y类节点有n个,$a(X_i,Y_k)$表示节点i和节点k之间的关系,如i,k之间存在连接关系,则$a(X_i,Y_k)=1$,否则$a(X_i,Y_k)=0$。节点i获得的资源为$f(i)$,用户i将资源平均分配给相连的物品k,物品k又将所得资源平均分配给其选择过的用户j,用户j将资源平均分配给与之相连的物品a,通过计算用户i最终分配给物品a的资源量,就可以计算用户i和物品a之间的相似度,用户i对所有产品的相似度进行排序,取前top-N作为推荐结果。

通信作者:周波(806446828@qq.com)

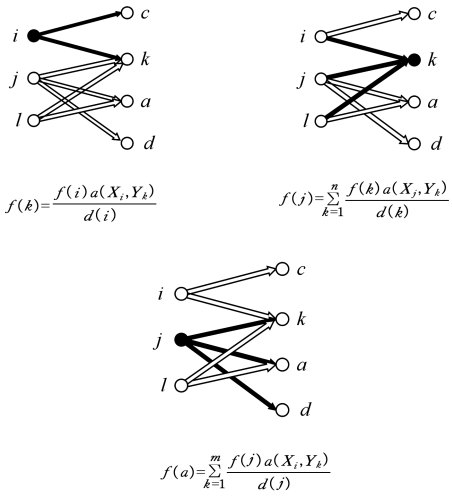

图 1　物质扩散算法的三次资源扩散过程

Fig. 1　Three times resource diffusion process of material diffusion algorithm

2.2　作者主题模型的基本原理

Blei 等在前人的基础上提出了概率生成模型 LDA[9]，LDA 模型只能用来生成文档的主题，然而在现实中一个作者可能有多篇文档，因此 Steyvers 等在 LDA 模型中引入作者隐变量，用作者主题分布取代 LDA 模型中文档主题分布，从而提出了 AT 模型(Author-Topic)，即作者主题模型[10]。该模型的优点在于将高维的词空间映射到二维的主题空间上，提取主题语义结构，将作者和主题之间的相似度通过计算语义空间的距离进行量化[11]，从而有效计算作者与作者之间的相似度。

具体来说，AT 模型的基本原理如下：

(1)每个作者 $a=1,2,\cdots,A$ choose $\theta_a\sim Dirichlet(\alpha)$；每个主题 $t=1,2,\cdots,t$ choose $\phi_t\sim Dirichlet(\beta)$。

(2)每篇文档 $d=1,2,\cdots,D$，给定作者 ad 的向量和每个单词 $w_i(i=1,2,\cdots,N_d)$。其中，X_i 条件依赖于 ad，$X_i\sim Uniform(ad)$，$Uniform$ 表示均匀分布，即从第 d 篇文档的作者中按均匀分布概率抽取一个作者；z_i 条件依赖于 X_i，$z_i\sim Multinomial(\theta(X_i))$；$Multinomial$ 为多项分布，表示从作者-主题的多项分布中抽取一个主题[11]；w_i 条件依赖于 z_i，$w_i\sim Multinomial(\phi z_i)$，表示从主题-词项多项分布中抽取一个单词。整个过程如图 2 所示。

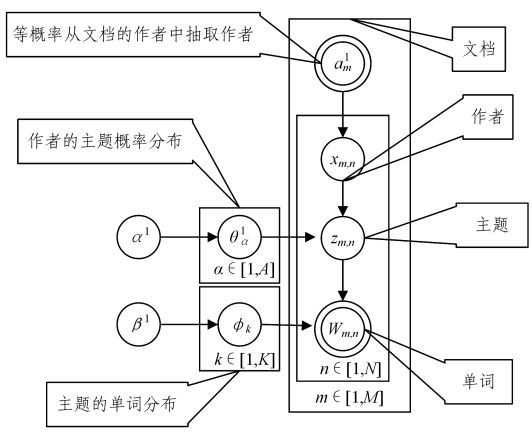

图 2　AT 模型的贝叶斯网络图

Fig. 2　Bayesian network diagram of AT model

Wang 等证明当主体之间的平均距离最大时，主题模型达到最优[12]。EsTAArooks 等提出使用 KL 距离计算主题之间的距离，计算方法如下[13]：

第一步　采用 KL 距离计算主题之间的两两相似度如下：

$$KL(Z_i\parallel Z_j)=KL(P(Z_i,w_k)\parallel P(Z_j,w_k))=\sum_{k=1}^{n}P(Z_i,w_k)\log_2 P(Z_i,w_k)/P(Z_j,w_k) \tag{1}$$

第二步　主题 Z_i 和 Z_j 之间的距离可以定义为：

$$dis(Z_i,Z_j)=1/2*(KL(Z_i\parallel Z_j)+KL(Z_j\parallel Z_i)) \tag{2}$$

第三步　计算主题之间的平均相似度，选择最大的 $avearge\text{-}dis$ 对应的主题数量作为最优的主题数。

$$average\text{-}dis=\frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}dis(Z_i,Z_j)}{K*(k-1)/2} \tag{3}$$

3　融合作者主题模型的物质扩散算法

鉴于当前推荐算法常用的 movielens 公开数据集缺乏足够的语义信息，本文采用合作专利数据进行专利权人合作推荐。因此与用户向用户推荐产品不同，本文是向专利权人推荐可能合作的专利权人，算法的核心是计算专利权之间的相似度。

3.1　基于二分网络模型的专利权人相似度

用 X,Y 分别表示二分网络两类节点专利权人和专利的集合，其中 X 类节点有 m 个，Y 类节点有 n 个，$a(X_i,Y_k)$ 表示节点 i 和节点 k 之间的关系，如 i,k 之间存在连接关系则 $a(X_i,Y_k)=1$，否则 $a(X_i,Y_k)=0$。设初始状态专利权人 i 的获得资源为 $f(i)$，专利权人 i 将资源平均分配给相连的专利 k，专利 k 又将所得资源平均分配给合作专利权人 j，专利权人 j 将资源平均分配给与之相连的专利 a，专利 a 又将资源平均分配给专利权人 l，如此进行 4 次资源分配，将专利权人 l 分配的资源量 $f(l)$ 作为其与专利权人 i 的相似度，最终推荐结果为 $R_1(i,l)=f(l)$，专利权人 i 对所有专利权人的相似度进行排序，取前 top-N 作为推荐结果。

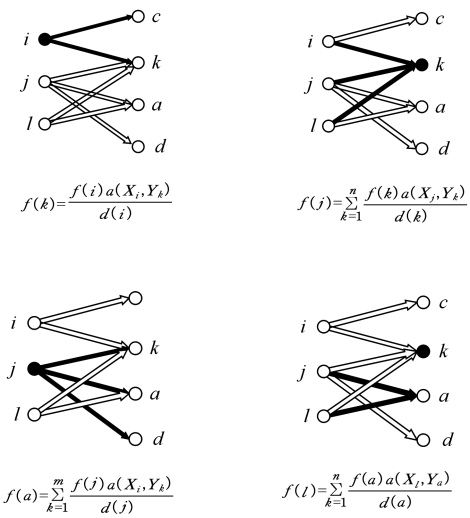

图 3　基于物质扩散算法的专利权人推荐的 4 次资源扩散过程

Fig. 3　Four times resource diffusion process of patentee recommendation based on material diffusion algorithm

3.2　基于主题模型的专利权人相似度

本文采用 EsTAArooks 等所采用的 JS 距离计算专利权

人对应的主题概率分布相似度，JS 距离是可以测量一对随机变量概率分布的相似性的方法[13]。假设专利权人 i 在合作主题 k 上的概率分布为 $p(i,k)$，专利权人 l 在合作主题 k 上的概率分布为 $p(l,k)$，则专利权人 i,l 之间的相似度可以表示为：

$$R_2(i,l)=0.5*\sum_{k=1}^{n}(p(i,k)*\ln(\frac{2p(i,k)}{p(i,k)+p(j,k)})+ p(l,k)*\ln(\frac{2p(l,k)}{p(i,k)+p(l,k)})) \quad (4)$$

3.3 融合算法专利权人相似度

$R_1(i,l)$ 表示基于二分网络的推荐结果，$R_2(i,l)$ 表示基于主题模型的推荐结果。本文借鉴 Hong 等[14]中提出的融合框架，构建融合语义模型的二分网络推荐算法，公式如下：

$$R(i,l)=\rho\alpha e^{R_1(i,l)}+(1-\rho)e^{R_2(i,l)},0\leqslant\rho\leqslant1 \quad (5)$$

当 $\rho=1$ 时 $R(i,l)$ 是二分网络的推荐结果；当 $\rho=0$ 时 $R(i,l)$ 是主题模型的推荐结果；当 ρ_1 在 0 到 1 之间时，$R(i,l)$ 是融合算法的推荐结果。

4 实验结果及分析

4.1 实验数据

鉴于当前推荐算法实验常用的 movielens 公开数据集缺乏足够的语义信息，本文自建数据集。本文采用文献[15]提供的德温特专利数据进行专利权人推荐。该数据共有 1 558 件专利，9 692 对合作关系，3507 个专利权人。随机抽取 20% 的数据作为测试数据，剩下的 80%作为训练数据，如此重复 5 次，生成 5 组测试集和训练集。

4.2 评价指标

常用的推荐算法的评价指标有两个：准确率和召回率，本文使用这两个指标来评价推荐结果。准确率反映的是推荐结果中正确推荐的比例，召回率反映的是推荐结果在测试集中出现过的比例。准确率和召回率的定义如下，表 1 给出了公式中变量的含义。

$$precision=TP/(TP+FP)$$

$$recall=TP/(TP+FN)$$

表 1 实验结果的评价标准

Table 1 Evaluation criteria for experiment

	测试集中	非测试集中
推荐结果	TP	FP
未推荐	FN	TN

4.3 实验结果及分析

实验具体步骤如下：

Step1 将专利权人之间的合作关系转化为矩阵专利权人-专利矩阵 $a(X_i,Y_k)$，并计算各个节点的度。

Step2 基于物质扩散算法的 4 次资源扩散，计算各个专利权人之间的相似度 $R_1(i,j)$。具体过程见图 3。

Step3 使用作者主题模型计算专利权人的主题分布，基于式(3)，选择最优主题个数。

Step4 基于专利权人-主题分布矩阵，采用式(4)计算专利权人之间的相似度 $R_2(i,j)$。

Step5 采用式(5)计算的融合相似度，计算专利权人之间的融合相似度 $R(i,j)$。

Step6 根据 Step5 的预测评分从高到低进行排序，优先选择前 N 个专利权人进行推荐，本实验 N 取 50。

Step7 改变训练集和测试集，重复 Step3－Step6 的实验过程 5 次，计算 5 次实验的准确率、召回率的平均值，作为最终的推荐结果

实验结果表明，基于二分网络模型和作者主题模型融合的推荐算法，对推荐结果有提升作用，但提升效果有限。当 $\rho_1=0.8,\rho_2=0.9$ 时，融合算法的结果取得最优结果，此时融合算法推荐结果比单一的二分网络推荐结果的准确率提高了 0.2%，召回率提高了 1.2%，准确率提高比率为 2.29%，召回率提高比率为4.15%。

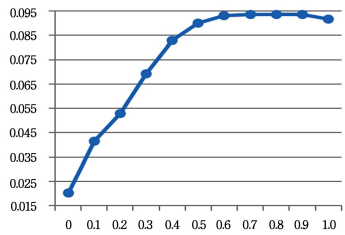

图 4 融合算法的准确率随融合系数的变化

Fig. 4 Accuracy rate of combination algorithm varies with fusion coefficient

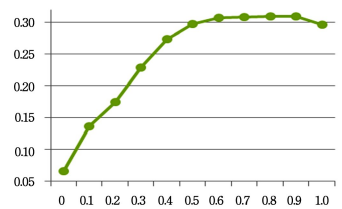

图 5 融合算法的召回率随融合系数的变化

Fig. 5 Recall rate of combination algorithm varies with fusion coefficient

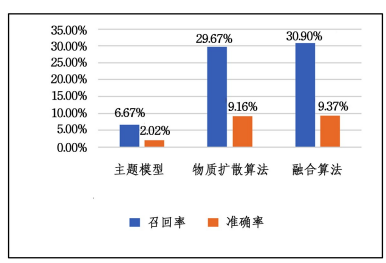

图 6 3 种算法的准确率和召回率对比

Fig. 6 Comparison of the recall rate and accuracy rate between three kind algorithms

结束语 本文针对协同二分网络推荐算法仅利用节点之间网络结构信息而没有考虑推荐对象间语义关系的问题，提出了一种融合语义模型的二分网络推荐算法。该算法既利用了节点之间的网络结构信息，又使用了节点的语义信息。通过实验证明，该算法提高了二分网络推荐算法的准确率、召回率。

但本文的研究存在一定的不足：1)由于缺乏公开的语义推荐算法数据集，本文使用自建数据进行测试，虽然使用随机抽样，取 5 次实验的平均值来提升研究的严谨性和科学性，但相比其他研究采用公开数据集展开实验的研究，本研究对同行研究进行对比分析带来不便，期待未来有相关的公开数据集以进行进一步的实验；2)本文采用准确率和召回率进行评价，未来的研究中可采用更全的指标如覆盖率、新颖性等进行测试。

参 考 文 献

[1] XIAO M,XIONG Q X. Collaborative Filtering Recommendation Algorithm Based on Semantic Similarity Between Items[J]. Journal of Wuhan University of Technology,2009,31(3):21-23,32.

[2] WU Z Y,TANG Y,FANG J X,et al. Collaborative Filtering Recommendation Algorithm Based on Ontology Semantic Similarity [J]. Computer Science,2015,42(9):204-207,225.

[3] WANG G S,PAN F Z. Collaborative filtering recommendation algorithm based on semantic similarity[J/OL]. [1-9]2020-04-05]. https://kns-cnki-net. e1. buaa. edu. cn/kcms/detail/34. 1054. N. 20200306. 2103. 006. html.

[4] ZHOU T,JIE R,MATĆS M,et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4):70-80.

[5] ZHOU B. Collaborative filtering algorithm—a special case of the bipartite network recommendation algorithm [J]. Computer Science,2019,46(S2):163-166,177.

[6] ZHOU B,YANG C F. The research of Bipartite Network Recommendation Algorithm Based on Transmitter and Acceptor Ability[J]. Technology Intelligence Engineering,2016,2(2):71-80.

[7] ZHOU B,YANG C F. The Research Progress of Recommendation Algorithm Based on Bipartite Network [J]. Technology Intelligence Engineering,2016,2(1):77-90.

[8] ZHANG X M,JIANG S Y,ZHANG Q S,et al. Hybrid recommendation by combining network-based algorithm and user preference [J]. Journal of Shandong University (Natural Science), 2015,9:29-35,41.

[9] BLEI D M,NG A,JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research,2003,3:993-1022.

[10] ESTAAROOKS A,JO T,JAPKOWICZ N. A Multiple Resampling Method for Learning from Imbalanced Data Sets[J]. Computational Intelligence,2004,20(1):18-36.

[11] WANG Y G,ZHANG X,LIU X G. Research on micro-blog user's interest mining based on author-topic model[J]. Computer Engineering and Applications,2015,51(13):126-130.

[12] WANG L D,WEI B G,YUAN J. Document Clustering Based on Probabilistic Topic Model [J]. Acta Electronica Sinica, 2012, 40(11):2346-2350.

[13] STEYVERS M,SMYTH P,ROSEN-ZVI M,et al. Probabilistic author-topic models for information discovery[C]// Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2004:306-315.

[14] HONG Y T. Collaborator Recommendation System Based On Coauthor Network [D]. Hangzhou:Zhejiang University,2015.

[15] ZHOU B,YANG C F. The Recommendation of Patentee Based on Bipartite Network——A Case Study of New Energy Vehicles [J]. Technology Intelligence Engineering,2016,2(4):56-68.

ZHOU Bo, born in 1991, master. His main research interests include data mining, and intelligence analysis.

(上接第 470 页)

[3] CAO W D,FANG X N. Airport Flight Departure Delay Model on Improved BN Structure Learning[J]. Physics Procedia,2012, 33.

[4] ARORA S D,MATHUR S. Effect of airline choice and temporality on flight delays[J]. Elsevier Ltd,2020,86.

[5] YAO Y,ZHU J F. An early warning model of abnormal flight management based on extension correlation function[J]. Journal of Southwest Jiaotong University,2008(1):101-106.

[6] XU T,DING J L,WANG J D,et al. Flight Delay and Spread Analysis Model Based on Bayesian Networks[J]. Journal of System Simulation,2009,21(15):4818-4822.

[7] LV Z P,HU X,DING J L. Flight Delay Early Warning Index System and Early Warning Level Construction[J]. Aviation Computing Technology,2010,40(1):1-4.

[8] WU R B,LI J Y,QU J Y. Flight delay prediction model based on dual-channel convolutional neural network[J]. Computer Applications,2018,38(7):2100-2106,2112.

[9] WU R B,ZHAO T,QU J Y. Flight delay prediction model based on deep SE-DenseNet[J]. Journal of Electronics & Information Technology,2019,41(6):1510-1517.

[10] ZHANG Z N,ZHANG J. Prediction method of large-scale flight delays[J]. System Engineering,2020,38(4):115-121.

[11] ZHU M M. Research on Bayesian Network Structure Learning and Reasoning[D]. Xi'an:Xidian University,2013.

[12] TENG H Z,JIA X S,ZHAO J M,et al. Research on Application of Hierarchical Hidden Markov Model in Equipment State Recognition[J]. China Mechanical Engineering, 2011, 22(18): 2175-2181.

[13] HE Y C,CHEN Z G,WANG H F,et al. Research on Bayesian Network Model of Missile Fault Diagnosis Based on Netica[J]. Aviation Weaponry,2020,27(1):89-95.

ZHANG Cheng-wei, born in 1990, postgraduate, teaching assistant. His main research interests include airline operation management and data mining.