一种基于领域知识的链路预测方法

李奋华 赵润林
(运城学院数学与信息技术学院 山西 运城 044000)

摘 要 个性化推荐在现代的电子商务中获得了广泛的应用,但是传统的推荐算法存在冷启动、稀疏性和推荐精准度差等问题,并且没有考虑个体间存在的关联信息。针对上述问题,在电子商务中采用链路预测和领域知识相融合的方法进行推荐,并在相关数据集上进行实验研究。结果表明,该方法在电子商务推荐中能够有效地提高推荐的精准度,更贴近用户的需求,能取得较好的推荐效果。

关键词 数据挖掘 社会计算 相似度 链路预测 个性化推荐

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.034

A LINK PREDICTION METHOD BASED ON DOMAIN KNOWLEDGE

Li Fenhua Zhao Runlin
(*Maths & Information Technology School*, *Yuncheng University*, *Yuncheng 044000*, *Shanxi*, *China*)

Abstract Personalized recommendation has been widely used in modern e-commerce, but the traditional recommendation algorithms have problems such as cold boot, sparsity and poor recommendation accuracy, and do not consider the related information among individuals in the process of recommendation. In order to solve the above problems, this paper uses the method of link prediction and domain knowledge to recommend in e-commerce, and carries out experimental research on related datasets. The results show that our method can effectively improve the accuracy of recommendation in e-commerce recommendation, more close to the needs of users, and achieve better results.

Keywords Data mining Social computing Similarity Link prediction Personalized recommendation

0 引 言

个性化推荐技术已经在电子商务领域获得了较广泛的应用[1]。在推荐系统中,推荐算法是推荐系统的核心部分,协同过滤算法作为经典的推荐算法在推荐系统中获得了较成功的应用[2-3]。但是协同过滤算法自身存在一些问题,例如:在数据稀疏的情况下,其推荐效果较差,且在推荐过程中并没有考虑实体间的一些附加有用信息[2]。有些学者提出了基于链路预测的个性化推荐算法,这些算法主要是利用顾客商品二分网络中顾客与商品之间的关联信息来弥补数据稀疏的问题,通过使用二分网络的拓扑结构信息来改善算法的推荐质量。然而,这些算法在推荐过程中仅仅利用了二分网络的拓扑结构信息,并没有考虑商品的领域知识。文献[3]提出采用顾客商品二分网络中商品的领域信息进行推荐,能够进一步改善推荐的精度和效果。

本文结合顾客商品二分网络的拓扑结构和网络中节点的相关属性,提出一种基于领域知识的链路预测方法。在该方法中,推荐给顾客的不同商品有不同的权重,权重的大小由与该商品相关的领域知识来决定,权重越大的商品被认为是越符合顾客需求的,越值得推荐给顾客。

1 链路预测算法

1.1 典型的链路预测算法

链接预测(Link prediction)是指如何通过已知的边(即网络拓扑结构)或者节点的特征等信息来预测

收稿日期:2019-05-14。国家自然科学基金项目(11501498)。**李奋华**,讲师,主研领域:数据挖掘,社会计算,电子健康。**赵润林**,副教授。

评估社会网络中节点之间未知链接(包括已存在而丢失的链接、未来的链接)存在的可能性[4-5]。几种典型的链路预测算法介绍如下:

1) 共同邻居方法(Common Neighbors,CN):该方法是基于待预测节点对的共同邻居节点的数量来对社会网络进行链接预测[6-9]。在这里,假设 $\Gamma(x)$ 代表节点 x 的共同邻居的集合[10-11]。一般来说,如果待预测节点对 (x,y) 具有的共同邻居数量越多,那么就认为节点 x 和节点 y 之间存在连边的可能性越大。因此,共同邻居方法的节点相似性度量指标定义如下:

$$S_{xy}^{\mathrm{CN}} = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

2) Adamic Adar 方法(简称 AA):该方法是基于待预测节点对的共同邻居集合来进行链接预测的。针对待预测节点对的每个共同邻居节点在链接预测中的作用不同,它赋予每个共同邻居一个权值,该权值是对应共同邻居节点度值对数的倒数[12-13]。如果一个节点的度值比其他节点小,那么该节点在链接预测中的作用比其他节点更重要[4-5,14-15]。该方法的节点相似性度量指标定义如下:

$$S_{xy}^{\mathrm{AA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \tag{2}$$

式中:$\Gamma(x)$、$\Gamma(y)$ 分别代表节点 x、y 的共同邻居的集合;K_z 表示节点 z 的度。

3) Jaccard Index 方法(简称 JA):该方法由 Jaccard 提出[10],是信息检索领域被广泛应用的一种相似度度量方法[16-17]。它的主要思想是:给定一个节点对 (x,y),在节点 x、y 的邻居并集中,将随机选择一个邻居节点是该节点对共同邻居的概率作为节点相似度的度量指标。该方法定义如下:

$$S_{xy}^{\mathrm{JA}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{3}$$

一般来说,评估预测算法精确度的主要有两个指标:AUC 和 Precision。AUC 是从整体上来考虑算法的预测精度;Precision 是从局部层面来考虑算法的预测精度[18]。

1.2　顾客商品二分网络

复杂系统在现实生活中普遍存在,复杂网络是表示和研究复杂系统的有效方法之一。在复杂网络中,节点表示复杂系统中的个体,边表示个体之间存在的关系[5,16]。二分网络是复杂网络的一种特殊形式,在该网络中节点被分成不同的两类节点,同类节点之间不存在关系,只有在不同类节点之间才存在关系[19]。在电子商务领域,顾客购买商品构成的网络就是二分网络,假设 $P = \{p_1, p_2, \cdots, p_n\}$ 表示顾客商品网络中的商品集合,$C = \{c_1, c_2, \cdots, c_m\}$ 表示顾客商品网络中的顾客集合,因此,能够获得一个隶属关系矩阵 $\boldsymbol{A} = (a_{ij})_{n \times m}$,其中 a_{ij} 表示网络中节点 i 代表的特定对象(即商品 i)与节点 j 代表的特定对象(即顾客 j)的隶属(即购买)关系值,也就是说,在该矩阵中如果顾客 c_i 购买了商品 p_j,那么 a_{ij} 赋值为 1,否则 a_{ij} 赋值为 0。在顾客商品二分网络中,如何选择顾客没有购买过的合适商品推荐给每个顾客是个性化推荐中很关键的问题,传统的协同过滤算法虽然有时能够取得较好的推荐效果,但是该算法仅仅考虑网络节点(顾客/商品)的直接邻居,具有一定的局限性[20]。图 1 为一个顾客商品二分网络,在该网络中平行四边形代表顾客,椭圆代表商品,实线表示顾客已购买商品,虚线表示将来顾客可能购买的商品(即个性化推荐)。

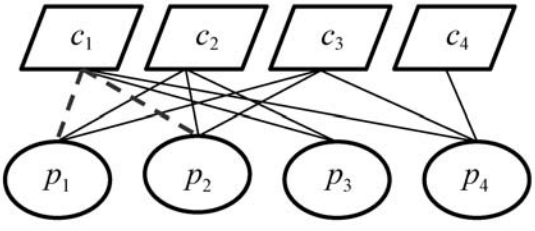

图 1　顾客商品二分网络

2　方法设计与实现

2.1　基于领域知识的链路预测方法

链路预测实际上是根据网络的拓扑结构的特点来推断未来有可能出现的关系。在生物研究领域,研究者根据共有邻居的数量来计算蛋白质对之间的拓扑相似度,以此来进行预测和推荐。文献[3]利用二分网络的拓扑特点,在电子商务中的商品推荐方面做了一定的研究。在此启发下,根据顾客商品二分网络的拓扑特点,把链路预测和领域知识相融合,描述了一种顾客商品推荐方法。假设顾客商品二分网络 G,通过相似度分值来计算未来顾客 c 购买商品 P 的连边 $<p,c>$ 的可能性大小,以此作为推荐的依据。对于节点 x,$\Gamma(x)$ 代表节点 x 的共同邻居的集合,那么节点 x 的邻居的邻居集合定义如下:

$$\Gamma'(x) = \Gamma_{c'' \in \Gamma(x)}(c'') \tag{4}$$

对式(1)作修改后,连边 $<p,c>$ 可能性的度量标准如下:

$$CP\_CN(p,c) = |\Gamma(p) \cap \Gamma'(c)| \tag{5}$$

同理,对式(3)作修改后,可获得连边 $<p,c>$ 可能

性的度量标准如下：

$$CP\_JA(p,c) = \left| \frac{\Gamma(p) \cap \Gamma'(c)}{\Gamma(p) \cup \Gamma'(c)} \right| \tag{6}$$

在文献[3]的启发下，根据实际顾客商品网络中商品的分类，构建商品的分类层次结构树，在此分类结构树的基础上，构建商品之间的语义相似度的度量标准，如下：

$$SPP(p_i,p_j) = \frac{2 \times \sum_{dis(p' \in set(p))} dis(p')}{n \times (dis(p_i) + dis(p_j)} \tag{7}$$

式中：$dis(p_i)$表示商品分类树中商品 p_i 与根节点的路径长度；$set(P)$表示商品 p_i 和商品 p_j 共同的祖先节点的集合；n 表示 $set(P)$中祖先节点的个数。

把式(7)和式(5)、式(6)相融合就得到在顾客商品二分网络中基于链路预测和领域知识的推荐评估指标，如下：

$$CPS\_CN(p,c) = |\Gamma(p) \cap \Gamma'(c)| \times \frac{\sum_{p_i \in CHB(c)} SPP(p,p_i)}{m} \tag{8}$$

$$CPS\_JA(p,c) = \left| \frac{\Gamma(p) \cap \Gamma'(c)}{\Gamma(p) \cup \Gamma'(c)} \right| \times \frac{\sum_{p_i \in CHB(c)} SPP(p,p_i)}{m} \tag{9}$$

式中：p_i 表示顾客 c 已经购买的商品；$CHB(c)$表示顾客 c 已经购买商品的集合；m 表示顾客已经购买商品的数量(即 $CHB(c)$集合的大小)。

2.2 实验分析

根据 2.1 节中描述的顾客商品推荐方法，在实际的超市购买数据集上进行实验，该数据集包含了近 1 年 3 542 名顾客的购买信息，其中涉及到的商品有 487 种，交易次数达到 36 873 次。在本实验中，用链路预测中的 AUC 作为推荐结果的度量标准，把该方法中的分值最高的前 20 种商品推荐给顾客，分别把数据集的 70%、80%作为训练集，剩余的作为测试集，实验结果如表 1 和图 2 所示。

表 1 顾客商品二分网络中不同推荐方法的推荐精度比较

推荐方法类型	评估指标	
	AUC(70%)	AUC(80%)
$CP\_CN(p,c)$	0.624 1	0.654 1
$CP\_JA(p,c)$	0.685 3	0.715 3
$CPS\_CN(p,c)$	0.725 6	0.765 6
$CPS\_JA(p,c)$	0.777 8	0.797 8

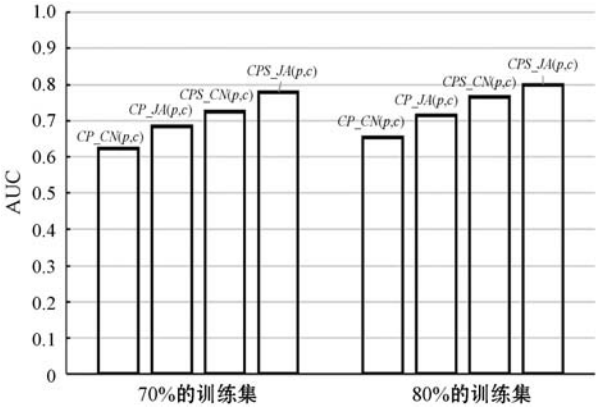

图 2 顾客商品网络中不同推荐方法的精度比较

从表 1 和图 2 的实验结果来看，在顾客商品二分网络中采用链路预测和领域知识相融合的推荐方法能够取得较理想的推荐效果。

3 结 语

基于电子商务二分网络个性化推荐精度低的现状，本文描述一种基于领域知识的链路预测方法，并在真实的超市顾客商品数据集上进行实验。结果表明，该方法推荐效果较好，能够在一定程度上提高个性化推荐的精度，具有一定的实用价值。

参 考 文 献

[1] 刘建国，周涛. 个性化推荐系统的研究进展[J]. 自然科学进展，2009，19(1)：1－15.

[2] 夏培勇. 个性化推荐技术中的协同过滤算法研究[D]. 青岛：中国海洋大学，2011.

[3] Chen P Y S, Wu S, Yoon J. The impact of online recommendations and consumer feedback on sales[C]//25th International Conference on Information Systems, 2004.

[4] 汪小帆，李翔，陈关荣. 网络科学导论[M]. 北京：高等教育出版社，2012.

[5] 郭世泽，陆哲明. 复杂网络基础理论[M]. 北京：科学出版社，2012.

[6] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报，2010，39(5)：651－661.

[7] 刘宏鲲，吕琳媛，周涛. 利用链路预测推断网络演化机制[J]. 中国科学物理学力学天文学，2011，41(7)：816－823.

[8] Sekara V, Stopczynski A, Lehmann S. Fundamental structures of dynamic social networks[J]. Proceedings of the National Academy of Sciences, 2016, 113(36): 9977－9982.

[9] Adamic L A, Lento T M, Adar E, et al. Information evolution in social networks[C]//Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016.

(下转第 238 页)

(2):222 - 235.
[8] Malawski M, Figiela K, Bubak M, et al. Scheduling multilevel deadline-constrained scientific workflows on clouds based on cost optimization[J]. Scientific Programming, 2015, 2015:5.
[9] 郑敏,曹健,姚艳. 面向价格动态变化的云工作流调度算法[J]. 计算机集成制造系统,2013,19(8):1849 - 1858.
[10] Rodrigo C, Rajiv R, Anton B, et al. CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms[J]. Software: Practice and Experience, 2011, 41(1):23 - 50.
[11] Juve G, Chervenak A, Deelman E, et al. Characterizing and profiling scientific workflows[J]. Future Generation Computer Systems, 2013, 29(3):682 - 692.
[12] Ostermann S, Iosup A, Yigitbasi N, et al. A performance analysis of EC2 cloud computing services for scientific computing [C]//International Conference on Cloud Computing, 2009.
[13] Schad J, Dittrich J, Quiané-Ruiz J A. Runtime measurements in the cloud: Observing, analyzing, and reducing variance [J]. Proceedings of the VLDB Endowment, 2010, 3(1): 460 - 471.
[14] Pan G, Li K, Xu Y, et al. A novel task scheduling scheme in heterogeneous computing systems using chemical reaction optimization[J]. Communications in Computer & Information Science, 2014, 472:328 - 335.
[15] Tawfeek M A, El-Sisi A, Keshk A E, et al. Cloud task scheduling based on ant colony optimization [C]//International Conference on Computer Engineering & Systems. IEEE, 2014.
[16] Huang J. The workflow task scheduling algorithm based on the GA model in the cloud computing environment[J]. Journal of Software. 2014, 9(4):873 - 880.
[17] Pandey S, Wu L, Guru S M, et al. A particle swarm optimizationbased heuristic for scheduling workflow applications in cloud computing environments[C]//2010 24th IEEE International Conference on Advanced Information Networking and Applications, 2010.
[18] Topcuoglu H, Hariri S, Wu M Y. Performance-effective and low-complexity task scheduling for heterogeneous computing [J]. IEEE Transactions on Parallel and Distributed Systems, 2002, 13(3):260 - 274.

(上接第 158 页)

[12] Watanabe Y, Kato T, Ishikawa M. Extended dot cluster marker for high-speed 3D tracking in dynamic projection mapping [C]//2017 IEEE International Symposium on Mixed and Augmented Reality(ISMAR), 2017.
[13] Pjanic P, Willi S, Grundhofer A. Geometric and photometric consistency in a mixed video and galvanoscopic scanning laser projection mapping system[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(11):2430 - 2439.
[14] Bermano A H, Billeter M, Iwai D, et al. Makeup lamps: live augmentation of human faces via projection[J]. Computer Graphics Forum, 2017, 36(2):311 - 323.
[15] Kurth P, Lange V, Siegl C, et al. Auto-calibration for dynamic multi-Projection mapping on arbitrary surfaces[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(11):2886 - 2894.
[16] Narita G, Watanabe Y, Ishikawa M. Dynamic projection mapping onto deforming non-rigid surface using deformable dot cluster marker[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(3):1235 - 1248.

(上接第 208 页)

[10] Martínez V, Berzal F, Cubero J C. A survey of link prediction in complex networks[J]. ACM Computing Surveys (CSUR), 2017, 49(4): 69.
[11] Chuan P M, Ali M, Khang T D, et al. Link prediction in co-authorship networks based on hybrid content similarity metric[J]. Applied Intelligence, 2018, 48(8):2470 - 2486.
[12] Li J, Cheng K, Wu L, et al. Streaming link prediction on dynamic attributed networks[C]//Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018.
[13] Brochier R, Guille A, Velcin J. Link prediction with mutual attention for text-attributed networks[EB]. arXiv: 1902.11054, 2019.
[14] Kim J, Hastak M. Social network analysis: characteristics of online social networks after a disaster[J]. International Journal of Information Management, 2018, 38(1): 86 - 96.
[15] Serrat O. Social network analysis[M]//Knowledge solutions. Springer, 2017: 39 - 43.
[16] Soares P R, Prudêncio R B. Proximity measures for link prediction based on temporal events[J]. Expert Systems with Applications, 2013, 40(16):6652 - 6660.
[17] Benson A R, Abebe R, et al. Simplicial closure and higher-order link prediction [J]. Proceedings of the National Academy of Sciences, 2018, 115(48): 11221 - 11230.
[18] 张宗宇. 社会化网络的链接预测[D]. 北京:北京邮电大学,2009.
[19] 朱陈平,张永梅,刘小廷,等. 复杂网络稀疏性的统计物理研究综述[J]. 上海理工大学学报,2011,33(5):425 - 432.
[20] 周涛,柏文洁,汪秉宏,等. 复杂网络研究概述[J]. 物理,2005,34(1):31 - 35.