

基于 Skyline 查询的社会网络链接预测

许烁娜 曾碧卿

(华南师范大学软件学院 佛山 528225)

摘 要 在社会媒体中,用户的状态信息实时地更新,用户之间的链接结构也不断改变,这给网络的链接预测提出了严峻的挑战。传统的链接预测方法针对某一特定情景,在非预定情景中效果往往表现不佳。针对单一网路连接预测算法的不足,提出一种基于 Skyline 查询的社会网络链接预测方法。该算法综合运用多种网络链接预测算法,将其预测值作为被预测链接的属性向量,并将 Skyline 点作为链接预测的结果返回给用户。实验表明,基于 Skyline 查询的链接预测方法其准确性明显高于相关链接预测研究的准确性,可应用于实际的社会媒体链接预测和推荐。

关键词 社会网络,链接预测,Skyline,算法

中图法分类号 TP393 **文献标识码** J

Skyline Based Link Prediction on Social Networks

XU Shuo-na ZENG Bi-qing

(School of Software, South China Normal University, Foshan 528225, China)

Abstract In social media, users update their statues in real time, and the links between them change quickly, which poses a huge challenge for link prediction in such social networks. Traditional link prediction algorithms are usually efficient in specific situations, but for other situations, they are not efficient. In order to deal with the deficiency of single link prediction algorithm, a Skyline query based link prediction approach in social networks was proposed. The algorithm combines multiple link prediction algorithms, assumes the values of these algorithms as a vector of the predicted link, and returns users with the Skyline points based on the calculated vector. Experiments show that the Skyline query based link prediction approach is more efficient than related researches, and can be used in real applications of link prediction and recommendation in social media.

Keywords Social networks, Link prediction, Skyline, Algorithm

1 引言

社会网络是人与人之间的关系在计算机世界中的映射。从图论的角度看,社会网络可以看作是一种图结构。图中的节点表示个人或其它实体,图中的边表示两个节点之间的相互关系,如好友关系、引用关系等。由于好友关系的不断增加和解除以及用户状态的不断更新,社会网络是高度动态的,网络中实体之间的关系也在不断发展演化,这种动态性给社会网络的研究带来巨大的挑战。利用图论的知识对社会网络进行分析,即社会网络分析(Social Network Analysis),已成为社会网络中一个热门的研究^[1]。

随着 Web 数据的爆炸式增长,数据挖掘技术不论在工业界还是学术界都越来越得到人们的重视。随着数据挖掘技术的不断发展,将其应用在社会网络分析领域已成为数据挖掘的一个重要的研究方向。在社会网络分析中,应用数据挖掘知识对网络的潜在链接进行预测,即链接预测(Link predic-

tion)^[2],是一项重要的研究内容。在社会网络中,链接预测可以用于网络结构的动力学分析^[3]、好友推荐^[4]、节点的个性化排名^[5]等应用领域。链接预测是社会网络中的一个子问题,它研究的是预测社会网络中未来可能存在的链接。链接预测关心的是节点之间链接未来存在的可能性,而并不关心节点自身的变化。链接预测根据节点自身的属性和网络的整体结构来预测两节点之间存在链接的可能性。

在链接预测过程中,人们针对不同的应用场景定义了不同的链接预测算法,这些算法有基于相似性的链接预测,有基于极大似然概率的链接预测,也有基于概率模型的链接预测方法^[6]。如果对这些算法进行细分,则基于相似性的链接预测又包括局部相似性、全局相似性和半局部相似性。其中局部相似性的定义包含公共邻居^[7]、Salton 指标^[8]、Sørensen 指标^[9]、优先链接指标^[10]、HPI & HDI^[11]等等。

在众多的链接预测算法中,每一种算法都针对特定的应用,没有一种算法适用于任何场合。然而在链接预测中,有时

到稿日期:2013-03-12 返修日期:2013-06-05 本文受国家自然科学基金(71272144),广州市科技计划项目(2013KP084),广东省自然科学基金项目(8151063101000040)资助。

许烁娜(1974—),女,硕士,讲师,主要研究方向为大数据、社会网络分析,E-mail: xsn_nhxy@163.com;曾碧卿(1969—),男,博士,教授,主要研究方向为网络计算、分布式处理、并行 I/O、云计算。

候用户并不熟悉潜在用户的特点,因此无法选取适合的链接预测模型。本文按照多个标准计算各个节点与目标节点之间的链接预测值,然后将这些值作为节点的属性值,采用数据库中 Skyline 查询的思想,将 Skyline 点作为候选节点返回给目标用户。

2 Skyline 查询

Skyline 查询又称为 Pareto,是数据库领域的热点研究内容,它在不损害其他方利益的条件下,自身达到最优值。Skyline 查询属典型的多目标优化问题,广泛应用在多目标决策中。具体地讲,Skyline 查询处理是指从一个给定的 n 维空间对象集合 S 中选取一个子集,该子集中的任何一点都不能被 S 中任意其它点所支配。Skyline 查询中的支配关系是指:给定一个 n 维空间集合 S 以及 S 中的两个点 p 和 q ,如果 $p_i \leq q_i (1 \leq i \leq n)$,我们说 p 支配 q ;如果存在 $p_i < q_i, p_j > q_j (1 \leq i, j \leq n)$,那么 p 和 q 相互不支配。

下面看一个经典的例子——假日酒店问题^[12]。假设某旅客在假日去 Nassau 旅游,他想找一家价格便宜且离海边近的酒店入住。图 1 中横坐标表示酒店的价格,纵坐标表示该酒店离海边的距离。从图中可以看出,该旅客要找的酒店一定在图中的折线上,而不在连线上的酒店根本不用考虑。这条折线上的数据的集合就是 Skyline 查询的结果集。

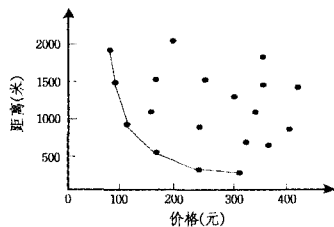


图 1 假日酒店的 Skyline 问题

3 相关工作

本节介绍社会网络中的链接预测和数据库中 Skyline 查询的相关研究。

3.1 链接预测

社会网络中,节点的链接预测方法主要有 3 种:基于相似性的方法、极大似然估计法和概率模型方法。

基于相似性的方法是最简单的链接预测方法。在基于相似性的节点链接预测方法中,任意两个节点之间都有一个相似性值。根据这些相似性值对链接预测进行排名,相似性值越大,说明这两个节点在将来越有可能建立新的链接。在该类方法中,如何定义节点间的相似性是最大的挑战。典型的方法有基于局部相似性的链接预测方法,如公共邻居^[7]、优先链接^[10]等;基于全局相似性的链接预测方法,如矩阵森林指数^[13]、余弦相似性^[14]等;基于半局部相似性的链接预测方法,如局部路径指数^[15]、局部随机游走^[16]等。

极大似然估计法的假设前提是网络的组织结构符合某种规律,然后依照此规律对网络进行建模,并且模型中含有参数。在建立了带参数的网络模型后,按照已有的网络数据对

网络进行评估。在依照极大似然估计法计算出社会网络的参数之后,根据该网络模型对未来可能建立的链接进行预测。典型的基于极大似然估计法的网络链接预测方法有层次结构模型^[17]和随机块模型^[18]。极大似然估计法虽然准确度高,但是处理能力有限。当网络中含有几千个节点时,极大似然估计法能在合理的时间内给出满意的结果,但是当网络中含有几百万节点时,该类方法却无能为力。此外,极大似然估计法的准确性明显低于基于相似性的算法。然而,极大似然估计法可以很好地分析网络的组织结构,这是其它链接预测算法达不到的^[19]。

概率模型方法是对社会网络的结构进行抽象,然后利用机器学习的方法预测该网络中尚未存在的边。给定一个社会网络 $G=(V,E)$,概率模型方法首先建立一个包含多个参数 Θ 的目标函数,此目标函数可以很好地描述该网络的结构。假设当前时刻节点 i 和 j 尚未建立链接,那么节点 i 和 j 之间建立链接的概率为 $P(A_{ij}=1|\Theta)$ 。通过对目标函数进行优化,可以得到 $P(A_{ij}=1|\Theta)$ 的实际取值。典型的概率模型方法有概率关系模型(或关系贝叶斯网络)^[20]、概率实体关系模型^[21]和随机关系模型^[22]。

3.2 Skyline 查询

Skyline 查询问题一经提出便引起了广泛而深入的研究。然而随着数据量的爆炸式增长、流数据应用的大量涌现,Skyline 查询研究正在向并行与分布式 Skyline 查询、数据流上的 Skyline 查询两方面发展。

随着并行与分布式计算方法的发展,采用廉价的并行或分布式系统来并行处理 Skyline 查询是应对大数据量的首选方法。Wu 等人^[23]研究了如何在内容可寻址的分布式网络中进行有效的 Skyline 查询,并提出了一种基于递归区间划分及动态区间编码的分布式算法。Cosgaya-Lozano 等人^[24]将待查询的数据集水平划分为多个子集,将其存储在不同的磁盘上,提出一种两阶段的并行 Skyline 查询方法。该算法第一阶段在局部数据集上进行 Skyline 查询,第二阶段合并所有的局部 Skyline 查询结果,并返回全局的 Skyline 查询结果。

数据流广泛存在于环境监测、交通监控、网络通信以及金融股票交易等应用中。数据流分析与挖掘是当前数据库与数据挖掘领域的研究热点,然而目前关于数据流上的 Skyline 查询的研究却很少。Lee 等人^[25]提出了一种数据流上的处理多个连续 Skyline 查询的方法——FAST。该方法应用过滤技术快速丢弃那些不可能成为 Skyline 点的数据,并快速判别内容中的数据可能成为哪个 Skyline 查询的结果。Zhang 等人^[26]通过滑动窗口技术研究了不确定数据流上的 Skyline 查询。

4 基于 Skyline 查询的节点链接预测

在社会媒体中,随着节点的不断加入和新的链接的不断形成,网络的拓扑结构不断地发生变化,同时由于网络中的个体实时地更新自己的状态,导致节点的内容也不断地发生变化。在这样一个网络结构和节点属性不断变化的动态社会网

络中,用户的需求很难预先定义好,因此传统的网络链接算法不能很好地满足用户的需求。

4.1 系统设计

本文将 Skyline 查询引入节点间的链接预测,即在网络中设置一个链接预测模型库,该库中预先定义多个链接预测的方法,其中每个链接预测方法的结果作为总的链接预测的一个属性。系统根据链接预测模型库上的模型将网络的节点链接预测值计算完毕并存储在链接预测结果库中。在查询处理过程中,用户提交链接预测请求,查询处理引擎分析该用户的资料,然后选择出可能用到的模型,并将这些模型的预测结果的 Skyline 查询结果返回给查询用户。系统的工作步骤如下:

1. 系统将社会网络的结构与节点的属性按照链接预测模型库里的模型进行节点间链接预测的计算;
2. 将计算结果存于链接预测结果库中;
3. 用户提交链接预测请求;
4. 链接预测提出用户的属性信息,然后根据该属性信息在链接预测结果库中选取适合的属性;
5. 将提取的属性信息的 Skyline 集合返回给用户。

系统结构及工作流程图如图 2 所示。

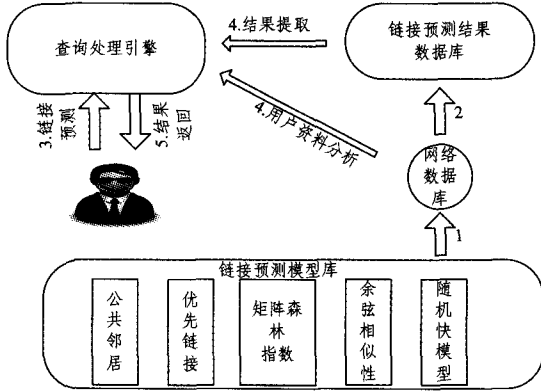


图 2 基于 Skyline 的链接预测系统结构

4.2 基于 Skyline 的链接预测算法

给定一个社会网络 $G=(V, E)$, 其中 V 为节点的集合, 节点的数目用 $|V|$ 表示, E 为边的集合, 边的数目用 $|E|$ 表示。当指定了链接预测的节点 u 后, 根据链接预测模型库计算其余各节点 $v(v \in V$ 且 $v \neq u)$ 在指定预测模型 $i(1 \leq i \leq m)$ 下的预测值, 结果是一个 $(n-1) \times m$ 的矩阵 M 。其中 $M_{i,j}$ 表示节点 i 在模型 j 下的预测值。在指定了节点 u 后, 其链接节点的返回是一个 Skyline 查询过程。

在结果返回时, 往往根据系统的需要返回 top- k 个结果 (即按照某种标准返回排名靠前的 k 个结果), 然而在算法进行 Skyline 查询时, 返回的结果往往超过 k 个, 这时需要对结果进行取舍。由于每一种预测算法都在其特定的情形下表现得很好, 因此在进行结果的取舍时, 不能对任何一种算法赋予过高的权值。为了提高预测结果的多样性, 即不能让某一种或几种算法统治整个结果集, 令 top- k 个结果中每个预测算法含有的结果数量与 Skyline 中该预测算法的结果数目成正比。

给定节点 u , Skyline 查询的结果为 S_u , S_u 中属于模型 $i(1 \leq i \leq m)$ 的结果为 S_u^i , 并且 $S_u = \bigcup_{i=1}^m S_u^i, S_u^i \cap S_u^j = \emptyset (1 \leq i, j \leq$

$m)$ 。用 $|S_u^i|$ 表示 S_u^i 中含有的元素个数, 最终返回的 top- k 结果为 $S_{u,k}$ 。

$$S_{u,k} = \bigcup_{i=1}^m S_u^{i'}, S_u^{i'} = \text{Truncate}(S_u^i, k_i), k_i = \frac{|S_u^i| \cdot k}{\sum_{j=1}^m |S_u^j|} \quad (1)$$

式中, $\text{Truncate}(S_u^i, k_i)$ 表示从 S_u^i 中截断选取 k_i 个元素, 并且 $k = \sum_{i=1}^m k_i$ 。

在对模型 i 的结果进行截断的过程中, 为了充分应用该模型的优势, 对每一个结果向量的元素进行平方, 然后基于平方和对结果进行排序, 并取其前 k_i 项。即,

$$S_u^{i'} = \{r_j \mid \forall 1 \leq j \leq p \leq k_i, r_j \leq r_p\} \quad (2)$$

$$r_j = \sum_{p=1}^m (r_{j,p})^2 \quad (3)$$

式中, $r_{j,p}$ 为向量 r_j 的第 p 个元素值。

5 实验及分析

本实验对链接预测所采用的评价标准是查准率 (precision)、召回率 (recall) 和 F-measure, 其定义为如下公式:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

其中, $|TP|, |FP|, |FN|$ 分别表示 True Positives, False Positives 和 False Negatives。

为了对本文提出的基于 Skyline 查询网络链接算法进行评价, 我们采集了部分的新浪微博的数据和腾讯微博。采集的数据有网络的结构关系, 即好友的相互关注关系。本实验在新浪微博上采集了 204075 个用户以及他们的相互关系 1298219 个, 在腾讯微博上采集了 196592 个用户及他们的相互关系 2197438 个。

在本实验中, 对采用 Skyline 查询的链接预测与典型的链接预测方法 (公共邻居、优先链接、矩阵森林指数、余弦相似性和随机快模型) 进行了对比, 结果见图 3 和 4。从图中所示的结果可以看出, 采用 Skyline 查询的链接预测方法明显优于采用单个指标的链接预测方法。

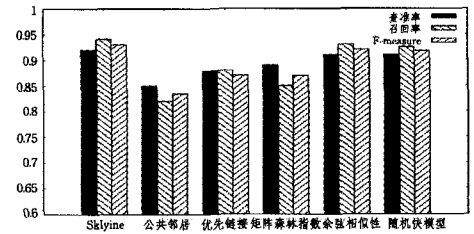


图 3 链接预测结果对比图 (新浪数据集)

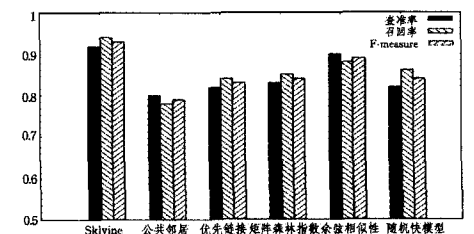


图 4 链接预测结果对比图 (腾讯数据集)

结束语 在社会网络中,链接预测可以应用在个性化节点排名、节点推荐和网络动力学研究中。然而社会媒体中,由于网络结构和用户状态的不断更新,使得单一的链接预测算法并不能满足用户的不同需求。本文提出了一种基于 Skyline 查询的链接预测方法。该算法将现有的各种链接预测方法的取值作为链接预测的节点的属性、综合用户的资料 and 状态信息,将链接预测的结果以 Skyline 查询的方式返回给用户。该方法综合了现有的各种链接预测方法的优点,并充分考虑了用户需求的不断变化,其预测结果明显优于现有的相关研究。

参 考 文 献

- [1] Scott J. Social network analysis[M]. SAGE Publications Limited, 2012
- [2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031
- [3] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J]. Physics reports, 2006, 424(4): 175-308
- [4] Hsu W H, King A, Paradesi M, et al. Collaborative and structural recommendation of friends using weblog-based social network analysis[C]//AAAI Spring Symposium Series. 2006
- [5] Yin D, Hong L, Davison B D. Structural link analysis and prediction in microblogs[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 1163-1168
- [6] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications, 2011, 390(6): 1150-1170
- [7] Newman M E J. Clustering and preferential attachment in growing networks[J]. Phys. Rev. E, 2001, 64
- [8] Salton G, McGill M J. Introduction to Modern Information Retrieval[M]. New York: McGraw-Hill Co., 1983: 30-42
- [9] Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons [J]. Biol. Skr., 1948, 5(4): 1-34
- [10] Zhang Guo-qing, Wang Di, Li Guo-Jie. Enhancing the transmission efficiency by edge deletion in scale-free networks[J]. Phys. Rev. E, 2007, 76
- [11] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(30): 1551-1555
- [12] Borzsonyi S, Kossmann D, Stocker K. The skyline operator [C]//Proc. of the 17th Int'l Conf. on Data Engineering. Heidelberg, IEEE Computer Society Press, 2001: 421-430
- [13] Chebotarev P, Shamis E. The matrix-forest theorem and meas-

- uring relations in small social groups[J]. Automation and Remote Control, 1997, 58(9): 1505-1514
- [14] Fouss F, Pirotte A, Renders J, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. IEEE Trans. Knowl. Data. Eng., 2007, 19(3): 355-369
- [15] Lu L, Jin C, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Phys. Rev. E, 2009, 80(4): 046122
- [16] Liu W, Lu L. Link prediction based on local random walk[J]. EPL, 2010, 89(5): 58007
- [17] Sales-Pardo M, Guimerà R, Amaral L A N. Extracting the hierarchical organization of complex systems[C]//Proc. Natl. Acad. Sci. U. S. A., 2007, 104: 15224
- [18] Airoldi E, Blei D, Fienberg S, et al. Mixed-membership stochastic block models[J]. Machine Learning Research, 2008, 9: 1981-2014
- [19] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453: 98-101
- [20] Friedman N, Getoor L, Koller D, et al. Learning Probabilistic relational models[C]//Proceedings of the 16th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 1999, 1300
- [21] Heckerman D, Meek C, Koller D. Probabilistic Entity-Relationship Models, PRMs, and Plate Models[C]//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada, 2004, 55
- [22] Yu K, Chu W, Yu S, et al. Stochastic Relational Models for Discriminative Link Prediction[C]//Proceedings of Neural Information Processing Systems. Cambridge MA: MIT Press, 2006, 1553
- [23] Wu P, Zhang C, Feng Y, et al. A Parallelizing skyline queries for scalable distribution[C]//Proceedings of the 10th International Conference on Extending Database Technology. Munich, Germany: Springer, 2006, 112-130
- [24] Cosgaya-Lozano A, Rau-Chaplin A, Zeh N. Parallel computation of skyline queries[C]//Proceedings of the 21st International Symposium on High Performance Computing Systems and Applications. Saskatchewan, Canada: IEEE, 2007: 12-18
- [25] Lee Y W, Lee K Y, Kim M H. Efficient processing of multiple continuous skyline queries over a data stream[J]. Information Sciences, 2012, 221(1): 316-337
- [26] Zhang Wen-jin, Lin Xue-min, Zhang Ying, et al. Probabilistic skyline operator over sliding windows[J]. Information Systems, 2013, 38(8): 1212-1233
- [27] 王茜, 张鲲鹏. 隐私保护数据挖掘算法 MASK 的改进[J]. 重庆理工大学学报: 自然科学版, 2012, 26(6): 63-66