软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(10):2611−2624 [doi: 10.13328/j.cnki.jos.005132]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于异构网络面向多标签系统的推荐模型研究*

王 瑜[1,2], 武延军[1], 吴敬征[1], 刘晓燕[2]

[1](中国科学院 软件研究所,北京 100190)

[2](中国科学院大学,北京 100049)

通讯作者: 王瑜, E-mail: wangyu@nfs.iscas.ac.cn

摘 要: 标签成为信息组织的重要方式之一,随着推荐系统的蓬勃发展,标签推荐成为学者们研究的重要问题之一.目前存在各种各样的标签系统,其功能千差万别,标签数据信息越来越复杂.目前研究往往针对特定类型标签数据,缺乏既综合考虑标签数据中不同类型对象的复杂信息又能适用于多种标签系统数据的标签推荐模型.构建了标签推荐模型 HnMTR,该模型首先针对标签数据中不同类型对象构建异构网络模型,其次对异构网络模型中不同类型顶点进行同空间映射,使不同类型的顶点和边可在同一空间进行量化比较;最后基于同空间映射后网络,引入多参数马尔可夫模型进行标签评分和推荐.通过基于豆瓣、Delicious 和 Meetup 这 3 个标签系统数据实验,其结果表明,HnMTR 模型平均准确率比目前主流算法提高 10%以上,取得了较好的推荐结果.

关键词: 异构网络;网络嵌入;标签推荐;标签系统;图模型

中图法分类号: TP181

中文引用格式: 王瑜,武延军,吴敬征,刘晓燕.基于异构网络面向多标签系统的推荐模型研究.软件学报,2017,28(10):2611−2624. http://www.jos.org.cn/1000-9825/5132.htm

英文引用格式: Wang Y, Wu YJ, Wu JZ, Liu XY. Multi-Dimensional tag recommender model via heterogeneous networks. Ruan Jian Xue Bao/Journal of Software, 2017,28(10):2611−2624 (in Chinese). http://www.jos.org.cn/1000-9825/5132.htm

Multi-Dimensional Tag Recommender Model via Heterogeneous Networks

WANG Yu[1,2], WU Yan-Jun[1], WU Jing-Zheng[1], LIU Xiao-Yan[2]

[1](Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Tagging has become one of the most significant methods for information organization. With the proliferation of recommending systems, tag recommendation problem has attracted more and more attention from researchers. Currently, while a variety of tagging systems exist, as the system function becomes more and more complex, the information of tagging data generated by tagging system becomes increasingly complex. In this paper, a tagging system is modeled as a heterogeneous network. To learn the importance of different types of nodes and edges, a general graph-based model, called HnMTR, is proposed. First, HnMTR maps different heterogeneous objects into a unified space so that objects from different dimensions can be directly compared. Then multivariate Markov model is applied to the mapped network to rank tag nodes. Highly ranked tags are recommended for the user. Experiments on three real world datasets with different tagging behavior demonstrate that the proposed method outperforms the state-of-the-art methods significantly.

Key words: heterogeneous network; network embedding; tag recommendation; tagging system; graph model

1 引 言

随着信息技术的不断发展,人类走入信息过载时代,越来越多的系统引入标签作为其信息表示方式,标签推

* 基金项目: 中国科学院先导专项(XDA06010600)
Foundation item: Strategy Priority Research Program of Chinese Academy of Sciences (XDA06010600)
收稿时间: 2015-11-10; 修改时间: 2016-03-17, 2016-06-12; 采用时间: 2016-08-29

© 中国科学院软件研究所 http://www.jos.org.cn

荐技术取得快速发展.《连线》杂志创始人 Kevin Kelly[1]指出,信息组织方式进化经历了 3 个阶段,从文件通过文件夹的组织形式存储在台式机中,到网页以链接的方式形成网络,再到数据流借助标签构成云.以标签来组织信息具有历史必然性,标签既可用于描述和组织信息,又可用于信息检索,此外还具有语义性.随着标签服务的成功,越来越多的系统为用户提供标签服务,标签是用户标记物品和发现物品最自然的方式,用户可用标签标记图片、视频、音乐、文档等各类数据[2].目前标签广泛应用于网站系统[3],Delicious、CiteULike、Last.fm 以及 Meetup 等系统都允许用户为其内容标注标签.然而,目前系统主要利用标签对信息资源进行分类、组织、检索,标签推荐技术也大多采用热门标签、常用标签、PersonalRank 等[3]简单、成熟的方法,推荐效果有待进一步提高.

标签成为信息组织方式,提供标签服务的系统不断涌现,如微博允许用户用标签标识自身特征,用于描述用户相关信息,构成二维标签信息〈用户,标签〉(注:本文所提到的数据维度是指数据中对象类型的个数);数据管理器用标签标识数据,通过标签直接定位数据,解决文件夹式不断双击查找数据的问题,也构成二维标签信息〈数据,标签〉;标签网站系统(如 Delicious,Last.fm,MovieLens 等)允许用户对网站内容进行标注,获得特定用户对特定内容的标签信息,构成三维标签信息〈用户,物品,标签〉;基于活动的应用系统(如 Meetup,Plancast)中,包含活动、群体、用户、活动内容、活动地点等多维信息,其中,用户、群体、活动内容等都可被标签标记,构成多维标签信息〈用户,事件,地点,群体,标签,等〉.标签数据维度从二维到多维不断扩展,随着数据信息量的增多,对用户进行精准标签推荐的难度也不断增大,现有标签推荐系统大多采用成熟、简单的推荐模型[3],难以体现不同类型对象的特征信息差异.目前研究或者针对单一网络结构建模,不区分数据不同维度信息,将数据中不同类型对象看作同一对象,或者针对某一特定维度网络综合考虑多个维度的信息进行标签推荐,缺乏同时考虑多维度信息且适用于不同维度数据的标签推荐模型.

本文研究一种基于异构网络(heterogeneous network)维度可扩展的标签推荐模型 HnMTR(heterogeneous networks tag recommendation in multi-dimension),该模型可同时适用于不同维度标签数据对用户进行标签推荐,实验结果表明,该模型不仅可以适用于不同维度的数据,而且推荐结果也优于现有流行算法.本文的主要贡献包括以下几个方面.

(1) 在标签推荐应用中,首先提出先进行异构网络同空间映射,再进行推荐的思想.综合考虑各个类型顶点特征后,将顶点进行同空间映射,使该模型适用于具有不同顶点类型数量的标签系统,具有良好的可扩展性.

(2) 在标签推荐应用中,构建面向多标签系统维度可扩展的标签推荐模型.以往模型或者仅考虑部分标签数据信息构建不区分顶点和边类型的同构网络(homogeneous network)模型进行推荐,或者仅基于某一特定类型的标签系统数据构建异构网络模型,其模型只适用于特定类型的标签系统.HnMTR 模型可适用于多种标签系统,且综合考虑数据多维度信息.

(3) 在标签推荐应用中,构建多维度标签数据对象的同空间映射模型.该模型在充分考虑不同顶点特征的前提下,将标签系统中不同类型顶点和边映射到同一空间,使网络中任意两个顶点可以比较,任意两条边的权重可以比较.

(4) 对标签推荐模型和顶点同空间映射模型参数进行最优化学习.针对文中提出的异构网络同空间映射模型和标签推荐模型,分别提出参数优化算法,对模型中的关键参数进行学习,设定最优模型参数值.

2 相关工作

本文研究标签推荐模型,首先对已有标签推荐研究进行分析总结;其次,本文充分借鉴网络嵌入(network embedding)技术已有研究成果,并基于该技术实现异构网络模型顶点和边同空间映射.此外,基于元路径建模是数据挖掘领域进行异构网络分析的另一类方法,很多研究人员将该方法用于相似性计算、链接预测等问题,本文也将简要介绍其相关研究成果.

2.1 面向互联网系统的标签推荐技术

随着互联网标签系统的不断涌现,标签推荐技术引起研究人员的广泛关注.德国卡塞尔大学 Jäschke 等

© 中国科学院软件研究所 http://www.jos.org.cn

人[4]最先通过量化研究方法对标签推荐结果进行评估,并提出基于协同过滤和基于图模型两种标签推荐算法,该研究对标签推荐算法进行初步尝试.美国明尼苏达大学 Vig 等人[5]基于 MoiveLens 数据集,通过分析用户的标签标注行为对电影和标签的评分进行预测,并向用户推荐标签,该研究采用协同过滤相关推荐算法,并未构建图模型,精确度有待提高.美国宾夕法尼亚州立大学 Song 等人[6]将带标签文档通过三元组(Words,Documents,Tags)表示成二部图,并提出高效率算法进行实时推荐,该研究主要专注算法实时特性,难以扩展到多维度标签数据.上海复旦大学 Yang 等人[7]基于微博用户标签数据,针对微博标签数据稀疏性、个人喜好优先等特点,提出先聚类、再通过社会化信息扩展、最后去除语义重复的标签推荐算法,该算法仅适用于微博数据,可扩展性有待增强.新加坡南阳理工大学 Tuan-Anh 等人[8]以基于活动的社会网络(event-based social network)为研究对象,提出用于进行用户个性化推荐的图模型,构建有多种顶点类型的异构网络,并利用该模型为群体推荐标签.该模型引入太多不同顶点类型,未进行顶点同空间映射,直接基于多参数马尔可夫链(multivariate Markov chain)建模,需学习过多参数,给模型计算带来很大负担.清华大学 Wei 等研究人员[9]针对社会化标签系统,构建异构图模型,针对(用户,物品,标签)三维数据进行标签推荐.他们还引入社会网络、标签语义、物品内容等附加信息以提高推荐准确度,是目前针对三维标签数据的最优算法,但该算法难以满足其他维度标签数据的推荐需求.针对以往标签推荐算法可扩展性和准确度无法同时兼顾的问题,本文首先对异构网络进行同空间映射,减少异构顶点类型,降低需学习参数数量.在充分考虑异构顶点信息的前提下,增强算法可扩展性并取得较好的推荐结果.

2.2 异构网络信息挖掘技术

网络嵌入研究是由网络中协同过滤和连接预测等应用需求驱动而产生的特征嵌入技术分支研究[10].该研究通常将实际应用问题转化为采用代数方法的实体嵌入(entity embedding)问题.美国 NEC 实验室 Zhu 等人[11]提出对网络连接矩阵和文档词频同时进行因数分解对网页进行分类;美国哥伦比亚大学 Shaw 等人[12]提出一个保持网络结构特征的嵌入框架,该框架将网络映射到一个低维度的欧几里德空间.此外,美国纽约州立大学石溪分校的研究人员提出 DeepWalk[13],通过有界随机游走模型学习网络空间的隐含信息.以上方法都仅考虑了同构网络信息,无法对异构网络的多种顶点和边的信息进行有效挖掘.引入异构网络信息最直接的方法是构建异构网络不同顶点和边类型的多个邻接矩阵,再分别对每个邻接矩阵用已有方法进行张量分解[14],这样做需维护大量矩阵信息,无法将算法扩展到大规模网络数据.中国科学院自动化研究所 Yuan 等人[15]提出一个非线性嵌入模型,该模型采用受限玻尔兹曼机(restricted Boltzman machines)对多种边的类型进行分析,但其并未充分利用原数据所有异构信息,且受限玻尔兹曼机在需大量学习参数设定时效率低下.美国伊利诺伊大学香槟分校 Chang 等人[16]提出基于深度学习模型的异构网络嵌入方法,该方法将多种顶点和边的信息映射到同一信息空间,使异构网络中任意两顶点可比较.本文主要借鉴该模型的异构网络信息嵌入思想,对异构网络信息进行同空间映射.

基于元路径的方法是解决异构网络相关问题的另一有效途径,元路径是指异构网络中链接对象关系的组合[17],不同组合表达不同语义.北京邮电大学 Shi 等人[18]针对传统异构网络建模方法难以精确度量顶点间语义信息的问题,提出引入异构网络边和元路径权重的方法,从而能够精确地描述路径语义,最终构建基于语义路径的推荐模型 SemRec.该研究团队已发表多篇基于元路径进行异构网络信息挖掘的科研论文[17,19],将该方法在文献检索、移动通话等领域的应用进行探索.美国印第安纳大学 Liu 等人[20]引入 PRF(pseudo relevance feedback)算法进行科研人员文献推荐.美国伊利诺大学芝加哥分校 Zhang 等人[21]基于多在线社交网络场景,提出内部社交元路径和外部社交元路径概念,构建异构网络模型,对社交网络进行连接预测.此外,美国乔治亚理工学院 Zhou 等人[22]和香港大学 Wan 等人[23]对元路径在异构网络聚类和分类中的应用进行深入探索.由于基于元路径的方法复杂度往往较高,研究者们大多借助分布式架构进行并行计算,本文主要借鉴网络嵌入的相关方法.

总之,已有研究通常针对某一特定标签系统数据建模,难以直接应用到其他系统,可扩展性受到制约.其次,基于异构网络进行信息挖掘技术不断受到研究人员关注,主要采用基于元路径和信息嵌入两种挖掘方法.根据调研结果,还未有学者将信息嵌入挖掘方法应用到标签推荐中,本文将充分借鉴该方法构建具有良好可扩展性的标签推荐模型.

© 中国科学院软件研究所 http://www.jos.org.cn

3 标签推荐模型

本节主要介绍维度可扩展标签推荐模型,可为多种标签系统提供标签推荐服务.其基本思想是:首先构建异构网络模型描述标签数据多维信息;其次对异构网络中除标签以外的异构顶点和边基于同空间映射模型进行计算,使不同类型的顶点和边可进行量化比较,将原异构网络多维度信息映射到由标签顶点和普通顶点构成的二维网络中;最终基于该二维网络引入多参数马尔可夫链构建标签推荐模型,并对模型参数进行最优化学习.本节首先介绍模型基本表示方法,其次介绍异构网络同空间映射模型,最后介绍基于映射后网络的标签推荐模型.

3.1 模型基本表示

异构网络[8]是指一个网络的顶点和边至少有一项存在两种或以上的类型,即顶点有多种类型或边有多种类型或二者都有多种类型.由于本文针对标签系统进行建模,对标签进行特殊处理,将标签系统数据抽象成无向图 $G=(T,ET,V,EV)$,其中,$T=\{t_1,t_2,\ldots,t_{|T|}\}$表示标签集合,$V=\{v_1,v_2,\ldots,v_{|T|}\}$表示图中其他类型顶点集合,$ET$ 表示图中边的两个顶点至少有一个为标签顶点的边集合,EV 表示图中除 ET 以外边的集合.边 $e_{ij}(\forall i,j\in\{1,\ldots,|V|+|T|\})\in EV\cup ET$ 当且仅当顶点 v_i 和 v_j 之间存在无向边,边的权重根据实际数据语义进行定义(在本文中,向量是列向量且用小写加粗字母表示,如 $\boldsymbol{x}$ 和 $\boldsymbol{y}$;矩阵用大写加粗字母表示,如 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$.用大写字母表示集合,如 V 和 E;用$|\cdot|$表示集合的大小,空集用 $\varnothing$ 表示).

此外,无向图 G 通过两个映射函数与对象类型集合 O 和关系类型集合 R 相关联,分别表示为 $f_v:V\to O$ 和 $f_e:E\to R$.每个顶点 $v_i\in V\cup T$ 都对应某个特定的对象类型 $f_v(v_i)\in O$.同理,每条边 $e_{ij}\in EV\cup ET$ 都对应某种特定的关系类型 $f_e(e_{ij})\in R$,每条边的关系类型通过两个顶点的类型确定.

定义 1(异构网络维度). 对于一个异构网络,其对象类型集合为 O,其顶点有$|O|$种类型,则称该异构网络为$|O|$维异构网络.另外,可以构建$|O|$维异构网络的标签数据集,称为$|O|$维标签数据.

网络的异构性通过集合 O 和 R 的大小体现,如果$|O|=|R|=1$,此时为同构网络;否则为异构网络.每个顶点都属于特定类型 $V_m(m\in[1,|O|])$,不同类型的顶点满足 $V_1\cup V_2\cup\ldots\cup V_{|O|}=V$,$V_m\cap V_n=\varnothing(m,n\in[1,|O|])$,$E_{mn}$ 则表示类型 m 的顶点与类型 n 的顶点间边的集合;每种类型的顶点由一个 d_m 维特征向量进行描述,即对于 $v\in V_m$,其特征向量 $\boldsymbol{x}\in R^{d_m}$.顶点间关系用对称矩阵 $A\in R^{|V|\times|V|}$表示,如果 $e_{ij}\in E$,则 $\boldsymbol{A}_{ij}$ 为 1,否则为–1.

3.2 基于网络嵌入技术的异构网络同空间映射模型

异构网络的顶点和边有多种类型,每种类型顶点具有不同特征,不同类型顶点间相似性难以度量,本节的工作是学习映射函数,将不同空间的数据映射到同一维空间,这样可以方便计算顶点间相似性[16].首先需选取顶点特征构建其特征向量,所选特征应反映顶点独特性质,如用户特征可选用其购买商品数量、消费总金额、标注标签数量等等.特征向量包含信息越多,越能清晰地描述顶点属性,映射结果信息量越大,计算复杂度越高.第 m 种类型的转换矩阵用 $\boldsymbol{U}_m\in R^{d_m\times r}$ 表示.对于第 m 种顶点类型某个顶点特征向量 $\boldsymbol{x}_m$,则有:

$$\tilde{\boldsymbol{x}}_m=\boldsymbol{U}_m^T\boldsymbol{x}_m \tag{1}$$

这样,将所有顶点的特征向量映射到 r 维空间.第 m 种类型顶点 i 表示为 v_{mi},其特征向量为 $\boldsymbol{x}_{mi}$.顶点间余弦相似度可以如式(2)所示,相似度越大,两顶点越相似.

$$s(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})=\tilde{\boldsymbol{x}}_{mi}{}^T\tilde{\boldsymbol{x}}_{nj}=(\boldsymbol{U}_m^T\boldsymbol{x}_{mi})^T\boldsymbol{U}_n^T\boldsymbol{x}_{nj}=\boldsymbol{x}_{mi}{}^T\boldsymbol{U}_m\boldsymbol{U}_n^T\boldsymbol{x}_{nj} \tag{2}$$

其中,$\boldsymbol{U}_m\boldsymbol{U}_n^T\in R^{d_m\times d_n}$.

异构网络中顶点与其他顶点通过显式或隐式方式关联,这些关联信息在网络中通过异构边体现.映射模型基本假设为:若两个顶点相连,则二者相似度应高于不相连的两个顶点.对于两个顶点特征向量 $\boldsymbol{x}_{mi}$ 和 $\boldsymbol{x}_{nj}$,为对边的信息进行形式化表示,设计一个成对判定函数,如式(3)所示.

$$d(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})\begin{cases}>0(\boldsymbol{A}_{ij}=1)\\<0(\boldsymbol{A}_{ij}=-1)\end{cases} \tag{3}$$

所有 $v_{mi},v_{nj}\in V$,t_{mn} 是与顶点类型相关的偏好值[16].令:

© 中国科学院软件研究所 http://www.jos.org.cn

$$d(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})=s(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})-t_{mn} \tag{4}$$

这样,本文采用由网络连接关系引导的二元逻辑回归函数进行参数 $\boldsymbol{U}$ 的学习[16],损失函数表示为

$$L(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})=\log(1+\mathrm{e}^{-\boldsymbol{A}_{ij}d(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})}) \tag{5}$$

以上损失函数基于 Sigmoid 函数,目标为最大化连通顶点与不连通顶点间相似度差异,但上式对 Sigmoid 函数先取倒数再取对数,故目标函数为求其最小值.目标函数形式化为

$$f(\boldsymbol{U})=\min_{\boldsymbol{U}_1,\ldots,\boldsymbol{U}_{|O|}}\left(\sum_{m\in[1,|O|]}\sum_{n\in[1,|O|]}\frac{\beta_{mn}}{N_{mn}}\sum_{v_{mi}\in V_m,v_{nj}\in V_n}L(\boldsymbol{x}_{mi},\boldsymbol{x}_{nj})+\gamma\sum_{m\in[1,|O|]}\|\boldsymbol{U}_m\|_F^2\right) \tag{6}$$

N_{mn} 表示连接顶点类型 m 和顶点类型 n 的边的数量.β是平衡参数,用来权衡不同类型边的重要性,γ是用来进行偏置方差[24]折中的偏好参数,这些偏好参数都可以通过学习来设定或者设置为固定值.在本文中,为简化计算,都设置为固定值.在上式中,$\|\cdot\|_F$ 是弗罗贝尼乌斯范数(Frobenius norm).该目标函数可采用坐标下降法进行计算,在计算某个 $\boldsymbol{U}_m$ 时,设定其他变量为固定值,计算过程如下:

$$\min_{\boldsymbol{U}_m}\left(\sum_{n\in[1,|O|]}\frac{\beta_{mn}}{N_{mn}}\sum_{v_{mi}\in V_m,v_{nj}\in V_n}\log(1+\mathrm{e}^{-\boldsymbol{A}_{ij}\boldsymbol{x}_{mi}{}^T\boldsymbol{U}_m\boldsymbol{U}_n{}^T\boldsymbol{x}_{nj}})+\gamma\|\boldsymbol{U}_m\|_F^2\right) \tag{7}$$

下降幅度则可以通过(8)计算,同维映射算法具体过程如算法 1 所示.由于已有研究[16]证明参数趋于收敛,假设该算法经过 n 次迭代趋于收敛,则其时间复杂度为 $O(nN)$.

$$\frac{\partial(\cdot)}{\partial\boldsymbol{U}_m}=\sum_{n\in[1,|O|]}\frac{\beta_{mn}}{N_{mn}}\sum_{v_{mi}\in V_m,v_{nj}\in V_n}\frac{-\boldsymbol{A}_{ij}\boldsymbol{x}_{mi}\boldsymbol{x}_{nj}{}^T\boldsymbol{U}_n}{1+\mathrm{e}^{\boldsymbol{A}_{ij}\boldsymbol{x}_{mi}{}^T\boldsymbol{U}_m\boldsymbol{U}_n^T\boldsymbol{x}_{nj}}}+2\gamma\boldsymbol{U}_m \tag{8}$$

通过以上同维映射,集合 V 中任意顶点 v_i 的特征都可用 r 维向量 $\boldsymbol{r}_i^T$ 表示.此时,顶点 v_i 和 v_j 间边的权重可定义为

$$w_{ij}=s(\widetilde{\boldsymbol{x}_i},\widetilde{\boldsymbol{x}_j}) \tag{9}$$

经过以上处理,对除标签顶点外的顶点及其边进行了同一维度映射,边的权重也进行了统一度量.

算法 1. 同空间映射模型参数学习算法.

输入:顶点初始特征矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$,坐标下降参数$\gamma$,平衡参数$\beta$,映射到维度 r;

输出:学习后的参数 $\boldsymbol{U}$.

1. t=0;
2. 初始化 $\boldsymbol{U}^{(0)}$;
3. **while** $f(\boldsymbol{U})$ 未收敛 **do**
4. **For** 维度 m **from** 1 to N **do**
5. 固定其他维度的参数;
6. 按式(8)所示计算$\partial(\cdot)/(\partial\boldsymbol{U}_{\mathrm{m}})$=0 时的 $\boldsymbol{U}_m$;
7. 更新 $\boldsymbol{U}_m{}^{(t)}$;
8. t=t+1;

3.3 基于多参数马尔可夫链的标签推荐模型

标签推荐的目的是为被标记对象推荐最能描述其功能和所属范畴的标签,如在微博数据中,被标记对象为用户;在文件管理系统数据中,被标记对象为数据;在社交音乐网站数据中,被标记对象为音乐;在基于活动的社会网络数据中,被标记对象为群体或用户.推荐的标签应该综合考虑数据中各个与其相关的维度信息,标签与标记对象在越多的维度相关度高,则其评分越高.

3.3.1 模型描述

通过上一节计算,异构网络 G=(T,ET,V,EV)的顶点 V 和边 EV 映射到同一空间,此时,顶点集合用 V'表示,边集合用 EV'表示,由 V'和 EV'构成的网络为同构网络.此时,网络表示为 G'=(T,ET,V',EV').由于模型的目的是进行标签推荐,在同维映射过程中并未对标签顶点及其相关边进行处理.此时,网络中顶点有两种类型:标签顶点和普

© 中国科学院软件研究所 http://www.jos.org.cn

通顶点;网络中边也有两种类型.带标签边 *et* 和普通边 *ev*.*ET* 中边的权重为标签标记次数,而 *EV*′中边的权重为通过同空间映射后顶点间关系的权重,网络邻接矩阵 $\boldsymbol{B}$ 中的元素值为对应边的权重.

定义 2(状态转移矩阵). 状态转移矩阵 $\boldsymbol{P}$ 是将引入标签顶点后网络 *G*′的邻接矩阵 $\boldsymbol{B}$ 的列进行归一化处理后的矩阵.网络的邻接矩阵表示如下:

$$\boldsymbol{B}=\begin{bmatrix}0 & \boldsymbol{B}_{TV}\\ \boldsymbol{B}_{VT} & \boldsymbol{B}_{VV}\end{bmatrix} \tag{10}$$

经过归一化的矩阵 $\boldsymbol{P}$ 可表示为

$$\boldsymbol{P}=\begin{bmatrix}0 & \boldsymbol{B}_{TV}\boldsymbol{D}_V^{-1}\\ \boldsymbol{B}_{VT}\boldsymbol{D}_T^{-1} & \boldsymbol{B}_{VV}\boldsymbol{D}_V^{-1}\end{bmatrix}=\begin{bmatrix}0 & \boldsymbol{P}_{TV}\\ \boldsymbol{P}_{VT} & \boldsymbol{P}_{VV}\end{bmatrix} \tag{11}$$

其中, $\boldsymbol{D}_V^{-1}$ 和 $\boldsymbol{D}_T^{-1}$ 为两个对角矩阵,二者的第 *i* 个元素计算如下:

$$\boldsymbol{D}_T^{-1}(i,i)=\frac{1}{\sum_{M\in\{T,V\}}\sum_{k=1}^{M}\boldsymbol{B}_{MT}(k,i)},\boldsymbol{D}_V^{-1}(i,i)=\frac{1}{\sum_{M\in\{T,V\}}\sum_{k=1}^{M}\boldsymbol{B}_{MV}(k,i)} \tag{12}$$

定义 3(查询向量). 给定一个标记对象及其与标签推荐相关的顶点集合 *S*,查询向量 $\boldsymbol{q}\in R^{|V|}$ 定义如下:

$$\boldsymbol{q}_i=\begin{Bmatrix}\omega_i(v\in S)\\ 0(v\notin S)\end{Bmatrix} \tag{13}$$

其中,$\omega_i>0$ 且 $\Sigma\omega_i=1$,令 $\boldsymbol{\omega}^T=\{\omega_1,\omega_2,...,\omega_{|S|}\}$ 表示查询向量的非零值向量.开始时,首先将 ω_i 设置为 $\sqrt{\sum_{k=1}^{r}\boldsymbol{r}_{ik}{}^2}$, 再对 $\boldsymbol{q}$ 中所有的非零值进行归一化处理,得到 ω_i 的初始值.

可重启随机游走模型(random walk with restart)被广泛应用于基于图的推荐模型中,其基本思想是将推荐问题转化为图顶点相似性计算问题.在异构网络中,该模型的缺陷是对网络中顶点和边仅仅看作是一种类型,无法区分异构网络中顶点和边的不同类型[8].在标签推荐系统中,也有研究将该模型应用到异构网络中,证明参数最终趋于收敛[9],但其方法并未将异构网络不同维度顶点和边进行同空间映射,其推荐模型仅适用于具有三维标签数据.本文为了对标签顶点和普通顶点进行区分,引入多参数马尔可夫链进行建模,顶点的状态转移函数用公式(14)、公式(15)表示.

$$\boldsymbol{t}^{(t+1)}=\alpha_{VT}\boldsymbol{P}_{VT}\boldsymbol{v}^{(t)} \tag{14}$$

$$\boldsymbol{v}^{(t+1)}=\alpha_{TV}\boldsymbol{P}_{TV}\boldsymbol{t}^{(t)}+\alpha_{VV}\boldsymbol{P}_{VV}\boldsymbol{v}^{(t)}+(1-\alpha_{TV}-\alpha_{VV})\boldsymbol{q} \tag{15}$$

从上两式可以看出,在该模型中,参数 $\alpha_{VT},\alpha_{TV},\alpha_{VV}$ 和查询向量 $\boldsymbol{q}$ 直接决定推荐结果.接下来对模型参数进行优化学习,令 $\boldsymbol{a}^T=\{\alpha_{VT},\alpha_{TV},\alpha_{VV}\}$.初始时,$\boldsymbol{t}^{(0)}$和 $\boldsymbol{v}^{(0)}$均设置为{0}.

3.3.2 模型参数优化

本文采用 BPR(Bayesian personalized ranking)优化框架来构建目标函数[8].在标签推荐应用场景中,给定标记对象,已被标记的标签集合表示为 *PT*,其余标签集合表示为 *NT*,则有 $T=PT\cup NT$.对于上式中 $\alpha_{VT},\alpha_{TV},\alpha_{VV}$ 参数,则应满足 *PT* 中标签排序比 *NT* 中标签排序靠前,即 *PT* 中标签比 *NT* 中标签有更大概率被标记到该对象上.采用 AUC(area under the roc curve)目标函数对该语义进行建模,假设总共有 *m* 个实例,则目标函数可表示为

$$\max_{\boldsymbol{w},\boldsymbol{\alpha}}f(\boldsymbol{w},\boldsymbol{\alpha})=\sum_{k=1}^{m}f_k(\boldsymbol{w},\boldsymbol{\alpha})=\sum_{k=1}^{m}\frac{\sum_{i\in PT_k}\sum_{j\in NT_k}\prod(\boldsymbol{t}(i)-\boldsymbol{t}(j))}{|PT_k||NT_k|} \tag{16}$$

其中,$NT_k=T-PT_k$,$\prod(\cdot)$为指示函数,$\boldsymbol{t}(i)$表示通过式(14)、式(15)得到的标签 *i* 的推荐评分,如果 $\boldsymbol{t}(i)-\boldsymbol{t}(j)>0$,则为 1,否则,为 0.指示函数一般定义为不可微的 Sigmoid 函数 $\sigma(x;\theta)=1/(1+e^{-\theta x})$,其中,参数 θ 控制误差,并通过实验设定.将该函数的对数形式代入式(16),可得:

$$\max_{\boldsymbol{w},\boldsymbol{\alpha}}f(\boldsymbol{w},\boldsymbol{\alpha})=\sum_{k=1}^{m}\frac{\sum_{i\in PT_k}\sum_{j\in NT_k}\ln\sigma(\boldsymbol{t}(i)-\boldsymbol{t}(j))}{|PT_k||NT_k|} \tag{17}$$

本文采用随机梯度下降法(SGD)计算式(17)中使目标函数最大化的参数 $\boldsymbol{\alpha}$,在每一步中,偏导求解算法按照式(18)所示梯度进行下降,其中 *lr* 决定了下降梯度的大小.

© 中国科学院软件研究所 http://www.jos.org.cn

$$\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} + lr\frac{\partial f_k(\boldsymbol{\omega},\boldsymbol{\alpha})}{\partial \boldsymbol{\omega}}, \boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + lr\frac{\partial f_k(\boldsymbol{\omega},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \tag{18}$$

接下来计算函数偏导数,方法如下:

$$\frac{\partial f_k(\boldsymbol{\omega},\boldsymbol{\alpha})}{\partial \boldsymbol{\omega}} = \frac{\sum_{i\in PT_k}\sum_{j\in NT_k}\frac{\partial \ln\sigma(\delta_{ij})}{\partial \delta_{ij}}\left(\frac{\partial \boldsymbol{t}(i)}{\partial \boldsymbol{\omega}} - \frac{\partial \boldsymbol{t}(j)}{\partial \boldsymbol{\omega}}\right)}{|PT_k||NT_k|}, \frac{\partial f_k(\boldsymbol{\omega},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \frac{\sum_{i\in PT_k}\sum_{j\in NT_k}\frac{\partial \ln\sigma(\delta_{ij})}{\partial \delta_{ij}}\left(\frac{\partial \boldsymbol{t}(i)}{\partial \boldsymbol{\alpha}} - \frac{\partial \boldsymbol{t}(j)}{\partial \boldsymbol{\alpha}}\right)}{|PT_k||NT_k|} \tag{19}$$

在式(19)中,$\delta_{ij}=\boldsymbol{t}(i)-\boldsymbol{t}(j)$,由于其为 Sigmoid 函数,则有$\partial\ln\sigma(\delta_{ij})/(\partial\delta_{ij})=\theta(1-\sigma(\delta_{ij}))$,而$\theta$又可以与参数 lr 合并,故可表示为$\partial\ln\sigma(\delta_{ij})/(\partial\delta_{ij})=1-\sigma(\delta_{ij})$.最后需计算式(14)和式(15)的偏导数,由$\boldsymbol{q}$定义可知,$\partial\boldsymbol{q}/(\partial\omega_i)\in R^{|V|}$,且$\omega_i$对应的值为 1,其余值为 0,计算可得:

$$\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\omega}} = \alpha_{VT}\boldsymbol{P}_{VT}\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{\omega}} \tag{20}$$

$$\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{\omega}} = \alpha_{TV}\boldsymbol{P}_{TV}\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\omega}} + \alpha_{VV}\boldsymbol{P}_{VV}\frac{\partial \boldsymbol{v}}{\partial \boldsymbol{\omega}} + (1-\alpha_{TV}-\alpha_{VV})\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{\omega}} \tag{21}$$

$$\begin{cases} \dfrac{\partial \boldsymbol{t}}{\partial \alpha_{VT}} = \boldsymbol{P}_{VT}\boldsymbol{v} + \alpha_{VT}\boldsymbol{P}_{VT}\dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VT}} \\ \dfrac{\partial \boldsymbol{t}}{\partial \alpha_{TV}} = \alpha_{VT}\boldsymbol{P}_{VT}\dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VT}} \\ \dfrac{\partial \boldsymbol{t}}{\partial \alpha_{VV}} = \alpha_{VT}\boldsymbol{P}_{VT}\dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VT}} \end{cases} \tag{22}$$

$$\begin{cases} \dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VT}} = \alpha_{VT}\boldsymbol{P}_{TV}\dfrac{\partial \boldsymbol{t}}{\partial \alpha_{VT}} + \alpha_{VV}\boldsymbol{P}_{VV}\dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VT}} \\ \dfrac{\partial \boldsymbol{v}}{\partial \alpha_{TV}} = \boldsymbol{P}_{TV}\boldsymbol{t} + \alpha_{TV}\boldsymbol{P}_{TV}\dfrac{\partial \boldsymbol{t}}{\partial \alpha_{TV}} + \alpha_{VV}\boldsymbol{P}_{VT}\dfrac{\partial \boldsymbol{v}}{\partial \alpha_{TV}} - \boldsymbol{q} \\ \dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VV}} = \alpha_{TV}\boldsymbol{P}_{TV}\dfrac{\partial \boldsymbol{t}}{\partial \alpha_{VV}} + \boldsymbol{P}_{VV}\boldsymbol{v} + \alpha_{VV}\boldsymbol{P}_{VV}\dfrac{\partial \boldsymbol{v}}{\partial \alpha_{VV}} - \boldsymbol{q} \end{cases} \tag{23}$$

通过式(20)和式(21)可以计算参数$\omega_i(i\in[1,|S|])$的微分,通过式(22)和式(23)可以计算参数$\alpha_{VT},\alpha_{TV},\alpha_{VV}$的微分.通过以上过程可以得到优化后的参数$\boldsymbol{\omega}$和$\boldsymbol{\alpha}$.模型的参数优化过程可表示为算法 2.由于已有研究[8]证明模型参数趋于收敛,假设该算法经过 n 次迭代趋于收敛,则其时间复杂度为 $O(nm)$,推荐模型时间复杂度与参数学习算法 2 时间复杂度一致.接下来,将对算法的结果与目前流行算法进行实验对比.

算法 2. 标签推荐模型参数学习算法.

输入:m 个实例,下降梯度参数 lr;

输出:学习后的参数$\boldsymbol{\omega}$和$\boldsymbol{\alpha}$.

1. t=0;
2. 初始化$\boldsymbol{\omega}^{(0)}$和$\boldsymbol{\alpha}^{(0)}$;
3. **while** $f(\boldsymbol{\omega},\boldsymbol{\alpha})$未收敛 **do**
4. 对 m 个实例进行随机洗牌;
5. **Foreach** 实例 k **do**
6. 按式(14)、式(15)所示迭代计算向量 $\boldsymbol{t}$ 和 $\boldsymbol{v}$;
7. 根据式(20)、式(21)计算$\partial\boldsymbol{t}/\partial\boldsymbol{\omega}$和$\partial\boldsymbol{v}/\partial\boldsymbol{\omega}$;
8. 更新$\boldsymbol{\omega}^{(t+1)}\leftarrow\boldsymbol{\omega}^{(t)}+lr(\partial f_k(\boldsymbol{\omega}^{(t)},\boldsymbol{\alpha}^{(t)}))/\partial\boldsymbol{\omega}^{(t)}$;
9. 根据式(22)、式(23)计算$\partial\boldsymbol{t}/\partial\boldsymbol{\alpha}$和$\partial\boldsymbol{v}/\partial\boldsymbol{\alpha}$;
10. 更新$\boldsymbol{\alpha}^{(t+1)}\leftarrow\boldsymbol{\alpha}^{(t)}+lr(\partial f_k(\boldsymbol{\omega}^{(t)},\boldsymbol{\alpha}^{(t)}))/\partial\boldsymbol{\alpha}^{(t)}$;
11. t=t+1;

© 中国科学院软件研究所 http://www.jos.org.cn

4 实验与评估

本节首先进行评估指标和实验设计的介绍,然后进行实验结果的阐述与评估.

4.1 评估指标

4.1.1 准确率

准确率定义为在推荐的 Top-K 标签集合中,真正标记在数据上的标签数与推荐集合标签总数的百分比,其定义如下所示.其中,D 表示测试数据集合,$R(d)$是推荐给数据 d 的标签集合,$T(d)$是数据 d 真实拥有的标签集合.

$$P@K=\frac{\sum_{d\in D}|R(d)\cap T(d)|}{\sum_{d\in D}R(d)} \tag{24}$$

4.1.2 召回率

召回率定义为在推荐的 Top-K 标签集合中,真正标记在数据上的标签数与数据的真实标签数的百分比,其定义如下所示.

$$R@K=\frac{\sum_{d\in D}|R(d)\cap T(d)|}{\sum_{d\in D}T(d)} \tag{25}$$

4.1.3 平均准确率(mean average precision)

平均准确率(MAP)用来衡量 Top-K 标签集合的平均准确度,它用来评测推荐的平均质量,其定义如下.其中,$rel(k)$是一个指示函数,如果第 k 个标签与真实标签匹配,则其值为 1;否则,为 0.$P(k)$是前 k 个标签的匹配比率.

$$MAP@K=\frac{\sum_{d\in D}\sum_{k=1}^{K}(P(k)\times rel(k))/|R(d)\cap T(d)|}{|D|} \tag{26}$$

4.2 实验设计

本文分别基于二维、三维和五维标签数据对推荐模型进行实验验证,其中,二维数据采用豆瓣读书数据[25],三维数据采用 Delicious 数据[26],五维数据采用 Meetup 数据[8].不同维度数据对比算法设置如下.

豆瓣数据集

(1) TF-IDF.基于信息检索领域的加权技术,采用统计方法评估标签重要程度.在二维数据集中,标签权重随其在同一标记对象上被标记次数的增多而增大,随其被不同标记对象所标记次数的增多而减小.

(2) CF.协同过滤算法,用到该数据集〈图书,标签〉二维信息.在标签推荐应用中,图书间相似度计算的基本思想是:两图书拥有的共同标签数越多,则二者越相近.用两图书共同标签数除以两图书标签集合大小乘积计算相似度.为了获得较好的推荐结果,对热门标签评分进行惩罚.

(3) RWR.可重启随机游走算法,其基于图模型,根据标签和图书的标记关系构建网络.其基本思想是赋予要推荐的图书较高权重,并以此顶点为出发点进行随机游走,以一定概率走到当前顶点的邻居顶点,以一定概率重新回到出发顶点.最后,顶点权重分布趋于收敛,该权重即为标签排序权重.

Delicious 数据集

(1) CF.协同过滤算法,在本数据集中采用〈网页,标签〉二维数据信息.计算方法与豆瓣数据集基本相同.

(2) RWR.可重启随机游走算法,在三维数据集中,该算法不区分顶点和边的不同类型,将整个网络看作同构网络.其基本思想是赋予被推荐标签的用户和物品较高权重,并以此二顶点为出发点进行随机游走,以一定概率走到当前顶点的邻居顶点,以一定概率重新回到出发顶点.最后,顶点权重分布趋于收敛,该权重即为标签排序权重.

(3) OptRank[9].该算法为目前三维数据标签推荐最优算法,其基于 RWR 算法,引入异构网络信息,对顶点评分状态转移函数的参数进行学习.

Meetup 数据集

(1) CF.协同过滤算法,在本数据集中采用〈用户,标签〉二维数据信息.计算方法与豆瓣数据集基本相同.

© 中国科学院软件研究所 http://www.jos.org.cn

(2) RWR.在本数据集中基于用户和标签关系图,用到⟨用户,标签⟩二维数据信息,若用户被某标签标记,则将二者相连.对于需进行标签推荐的用户,初始时其权重为 1,其余用户权重为 0,构成查询向量进行求解,该算法属于基于同构网络方法.

(3) full_RWR.基于 RWR 进行改进,在整个网络上运行 RWR 算法,不区分不同类型的顶点和边,是基于同构网络的算法.

(4) HeteRS[8].适用于基于活动的社会网络数据,是目前的最优算法.该算法构建异构网络模型,对每种类型的顶点分别进行建模,并基于多参数马尔可夫链对参数进行学习.

4.3 实验结果

本节描述针对 3 种不同标签数据集的实验结果,并对结果进行量化分析和解释,对各种算法优劣进行比较.由于 TF-IDF 和 CF 算法不是基于图的算法,运行时间与基于图的算法不处于同一数量级,不具有可比性,故运行时间仅对基于图的算法进行比较.

4.3.1 二维异构网络实验结果

实验将豆瓣数据集随机分为 8 份,其中 1 份作为测试集,其余作为训练集.最终训练集记录有 23 378 条,包含 3 315 本图书信息和 6 879 个标签信息;测试集记录数为 2 892 条,包含 2 012 本图书信息.该数据集包含⟨图书,标签⟩二维信息,故无需对信息进行同空间映射,直接采用推荐模型进行标签推荐.

HnMTR 算法与其他 3 种算法的对比结果如图 1 所示,对准确率、召回率、平均准确率和运行时间 4 个指标进行评测.由结果可知,HnMTR 算法在 4 个指标上优于其他非基于图的算法 70%以上,总体上优于 RWR 算法.在二维数据集中,不进行异构网络同空间映射.HnMTR 算法与 RWR 算法的区别在于:HnMTR 算法引入多参数马尔可夫链进行状态转换建模,并对参数进行学习;而 RWR 基于单参数马尔可夫链建模.可见,多参数马尔可夫链可以更精确地描述顶点的状态转换,提高推荐结果质量.

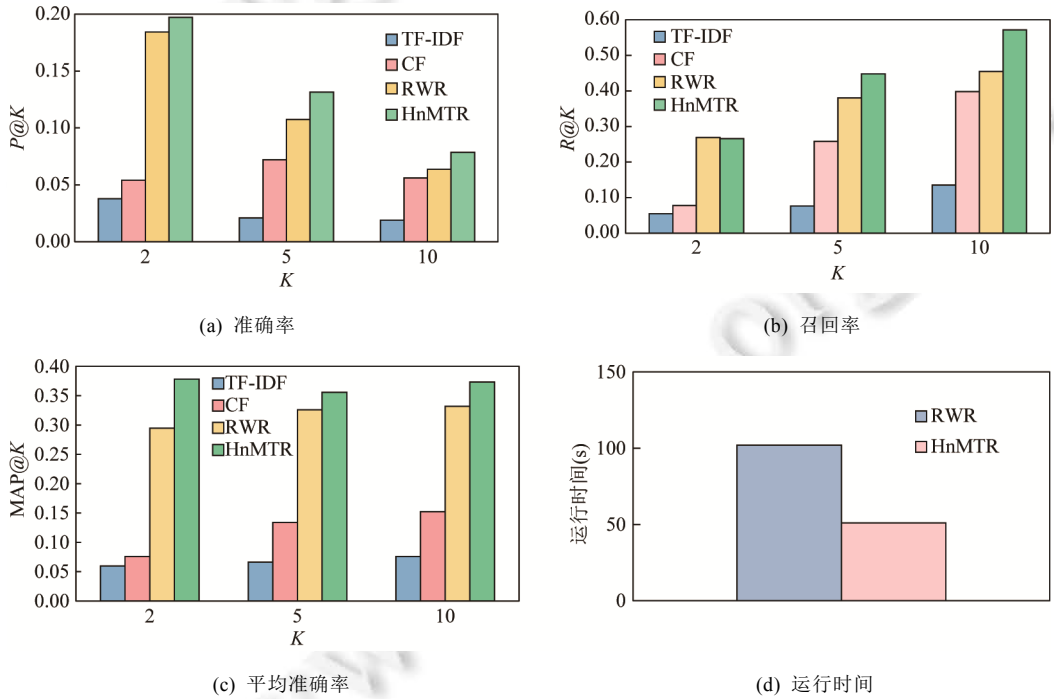


Fig.1 Douban dataset result

图 1 豆瓣数据集实验结果

© 中国科学院软件研究所 http://www.jos.org.cn

从图 1(b)可以发现,当 K=2 时,HnMTR 算法的召回率与 RWR 算法相当,这可能是由于 RWR 算法在该数据集上的召回率有较好表现.在 K=5 和 K=10 时,以及后续在其他数据集上的实验发现,HnMTR 算法的召回率优于 RWR 算法.在 K=5 时,HnMTR 算法在准确率、召回率和平均准确率上优于 CF 算法分别为 82.58%、73.76%和 165.79%;优于 RWR 算法分别为 22.28%、17.79%和 9.22%.故 HnMTR 算法在二维数据集上通过引入多参数马尔可夫链进行建模而提高了算法的精度和效率.

在 K=2 时,无论是推荐的准确度还是 MAP,RWR 和 HnMTR 基于图模型的算法推荐质量要远远高于 TF-IDF 和 CF 算法.对用户而言,往往仅关注前几个推荐的标签,所以 K 较小时的推荐质量更重要.基于图模型的算法虽然运算时间较长,但推荐结果质量得到大幅提高.此外,RWR 和 HnMTR 都是基于图模型的算法,由于 HnMTR 基于多参数马尔可夫链,参数增多使计算时邻接矩阵规模减小,进行权重迭代时参数收敛更快,故运行效率高于 RWR,图 1(d)所示为两种算法运行 74 个训练实例时间的对比情况(由于部分算法运行时间较长,本文均采用运行 74 个训练实例结果进行对比).

4.3.2 三维异构网络实验结果

实验用 Delicious 数据集,将数据随机分为 8 份,其中 1 份作为测试集,其余作为训练集.最终训练集记录有 136 883 条,包含 9 220 个用户,83 422 个网页和 35 934 个标签;测试集记录数为 17 119 条.对于 Delicious 数据,包含〈用户,网页,标签〉3 个维度信息,需对<用户,网页>两种类型顶点进行同空间映射,使顶点间边的权重可在同一空间进行比较.本实验中,用户特征信息表示为(访问网页数,标记标签数),网页特征信息表示为(访问该网页用户数,被标记标签数),被标记顶点类型为网页.将两个特征向量映射到维空间,由于参数决定顶点间关系的重要程度,本例中只有〈用户,网页〉一种类型关系,无〈用户,用户〉和〈网页,网页〉类型的边,故通过多次实验获取参数,在训练用户顶点转换向量时,设置为 1;在训练网页顶点转换向量时,设置为 0.01.转换向量的学习结果见表 1.

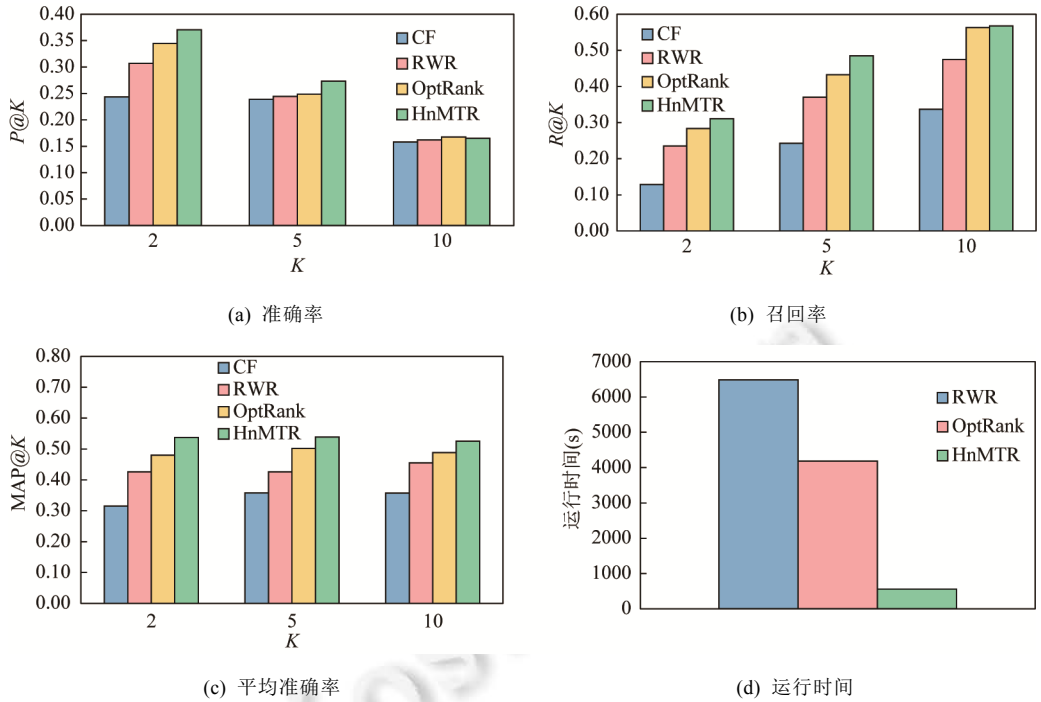
Table 1 Result of U

表 1 U 学习结果

顶点类型	U
用户	$(0.0396568,0.0366193)^{T}$
网页	$(0.0686875,0.1104794)^{T}$

HnMTR 算法与其他 3 种算法的对比结果如图 2 所示,分别对准确率、召回率、MAP 和运行时间 4 个指标进行评测.从图中可以看出,不论是准确率、召回率还是 MAP 指标,HnMTR 算法总体上都优于其他算法.如图 2(a)所示,HnMTR 算法的准确率比 CF 算法提高 14%以上,比 RWR 算法提高 11%以上.K 值越小,HnMTR 算法比其他算法的优势越明显,因为随着 K 值的增加,推荐标签个数大于网页实际标签个数,准确率整体下降.结果说明,顶点进行同空间映射后,各个顶点间关系可直接进行量化比较,使各个边的权重与实际重要程度更符合,从而取得更好的推荐结果.HnMTR 算法与 OptRank 算法的结果相当,而在现实推荐中,用户更看重 K 较小时推荐的标签,而当 K=2 和 K=5 时,HnMTR 算法的结果优于 OptRank 算法,故 HnMTR 算法的准确率优于目前最先进的算法 OptRank.从图 2(b)可以发现,HnMTR 算法的召回率优于 CF 算法 68%以上,优于 RWR 算法 19%以上,优于 OptRank 算法 1%以上.这说明,HnMTR 算法的推荐结果覆盖率高于其他算法,引入多参数马尔可夫链进行建模后,对参数的学习更加细化,使查询向量的设置更加合理,顶点状态转化概率更加符合实际,提升了推荐结果召回率.图 2(c)展示了不同算法平均准确率的对比情况,从图中可以发现,不论 K 是否大于实际标签个数,HnMTR 算法的平均准确率均高于其他算法,这说明,该算法推荐标签的平均质量高于其他算法.图 2(d)展示了算法运行时间,从图中可以看出,RWR 和 OptRank 算法由于需要维护计算全部顶点和边的状态转移矩阵,参数收敛较慢,计算时间较长;HnMTR 算法由于对数据进行同空间映射,网络中有两种类型顶点,且采用多参数马尔可夫链进行建模,每个状态转移函数对应状态转移矩阵相对较小,参数收敛迅速,大大提高了运行效率.

© 中国科学院软件研究所 http://www.jos.org.cn



(a) 准确率 (b) 召回率

(c) 平均准确率 (d) 运行时间

Fig.2 Delicious dateset result

图 2 Delicious 数据集实验结果

4.3.3 五维异构网络实验结果

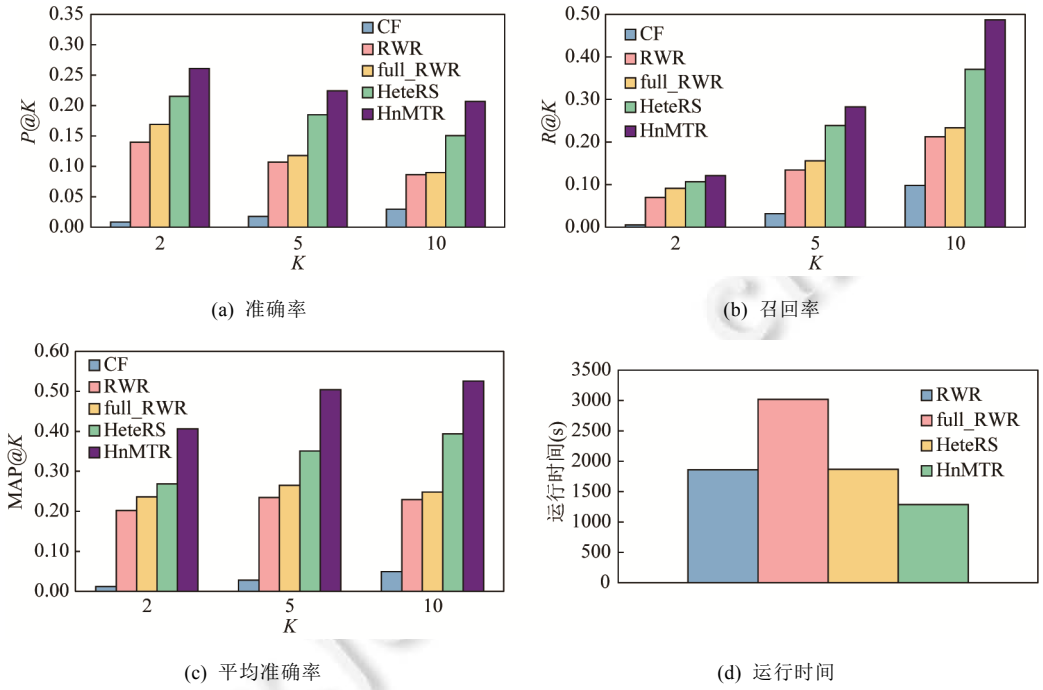
实验用 Meetup 数据集,该数据集包含纽约和加利福尼亚两个地点数据,本文选择纽约市数据进行研究.数据集中包含群体(group)标签和用户(user)标签两种,本文选择用户作为标记对象进行实验.最终实验数据训练集包含 32 443 个用户,15 514 个标签,9 551 个事件,398 个群体;测试集包含 2 999 个用户.

该数据集包含〈用户,活动,群体,标签,地点〉5 种顶点类型,故称其为五维异构网络.需要对〈用户,活动,群体,地点〉4 个顶点类型进行同空间映射.本文用(参加群体数,标记标签数,参与活动数)特征信息标识用户,用(地点,时间,参与群体)特征信息标识活动,用(成员数量,举办活动数)特征信息标识群体,用(举办活动数)特征信息标识地点.将以上特征向量映射到二维空间,4 类顶点信息包含 6 个顶点对元组,故设置参数学习方法与三维异构网络一致,本数据集的结果分别为 3×2 维、3×2 维、2×2 维和 1×2 维矩阵.HnMTR 算法与其他 4 种算法的对比结果如图 3 所示.

从图 3 可以看出,HnMTR 算法的准确率、召回率和平均准确率优于目前最先进的 HeteRS 算法 21%、18%和 43%以上.特别地,在平均准确率评测算法的平均推荐质量方面,HnMTR 算法的平均推荐质量比 HeteRS 的优势更加明显,因为 HnMTR 算法经过顶点同空间映射后,不同类型顶点间的关系可进行量化比较,在迭代计算顶点权重时,不同类型顶点的权重更能表征其在整个网络中的实际权重;而不进行同空间映射的模型在顶点权重迭代计算时,无法准确比较不同类型顶点权重间的关系.full_RWR 算法的推荐质量优于 RWR 算法,由于 full_RWR 算法虽无法对不同顶点类型进行区分,也无法度量其权重,但其引入了不同类型顶点间的关联信息,从而其推荐质量优于传统的 RWR 算法.故引入更多数据信息,可提高推荐质量.图 3(d)对各种基于图模型算法的运行时间进行了比较.发现 full_RWR 算法最耗时,由于该算法将所有数据信息维护在一个邻接矩阵中,每次查询和状态更新都需对整个矩阵进行更新;且其状态转换方程基于单参数马尔可夫链,参数收敛较慢.RWR 算法由于比 full_RWR 算法维护较少的数据信息,其速度较快但推荐质量下降.HeteRS 算法基于多参数马尔可夫链,状态转移邻接矩阵规模较小,参数收敛较快,其运行时间小于 full_RWR 算法.但其需要维护大量的异构顶点信息,对不同类型的顶点采用不同的状态转换方程计算权重,权重数值差异较大且需要学习较多参数,故运行时间大

© 中国科学院软件研究所 http://www.jos.org.cn

于 HnMTR 算法.HnMTR 算法的运行时间最短,由于其对异构顶点信息进行了同空间映射,仅需要学习较少的参数且权重数值可比较,参数收敛速度快于 RWR 算法,故该算法效率较高.



(a) 准确率 (b) 召回率

(c) 平均准确率 (d) 运行时间

Fig.3 Meetup dataset result

图 3 Meetup 数据集实验结果

4.3.4 结果分析

HnMTR 模型的精确度之所以优于以往模型,主要包含以下两方面原因.

(1) 通过同空间映射模型在同一维度全面考虑异构网络不同类型顶点信息.

将异构网络不同类型顶点特征信息通过同空间映射方式引入推荐顶点排序过程,直接影响了模型中不同顶点状态转移的概率,从而提高了模型精度.

以 Delicious 数据集为例,网络中存在 3 种类型顶点(用户,网页,标签),用户顶点与网页顶点间的权重是用户为该网页打标签的个数,用户顶点与标签顶点间的权重为标注该标签的次数,网页顶点与标签顶点间的权重为被该标签标注的次数.用户 104、网页 26496 和标签“photograpy”3 个顶点,权重分别为 8,1,23,则在进行状态转移计算时,不同类型顶点间的权重缺乏可比性,转移函数系数差异较大.基于第 4.3.2 节中顶点特征进行顶点同空间映射,经过参数学习后,其权重分别为 0.2 683 499,0.0 887 831,0.6 739 876,则将其映射到同一度量空间且权重可比较,状态转移函数计算误差减小,提高了推荐精度.

(2) 采用多参数马尔可夫链进行标签推荐建模.

多参数马尔可夫链比单参数马尔可夫链可更精确地描述不同类型顶点间的状态转移,且由于本文针对标签推荐应用建模,对标签类型顶点进行特殊处理,凸显其重要性.故能提高标签推荐精度.

以 HnMTR 模型与 OptRank 模型在豆瓣数据集运行为例,网络中存在两种类型顶点(图书,标签),图书与标签之间的权重为图书被标注此标签的次数.存在图书《罪责》、《简爱》和标签“外国文学”3 个顶点,经过同空间映射后权重为 $w_{v_1t}=0.7564688, w_{v_1v_2}=0.8733234$, 在 OptRank 模型中,顶点状态转移函数为 $\boldsymbol{v}^{(t+1)}=(1-\alpha)(\boldsymbol{P}_{VV}\boldsymbol{v}^{(t)}+\boldsymbol{P}_{VT}\boldsymbol{t}(t))+\alpha\boldsymbol{q}$,学习后$\alpha$值为 0.4 447 323,然而在 HnMTR 模型中,顶点状态转移函数为 $\boldsymbol{v}^{(t+1)}=\alpha_{TV}\boldsymbol{P}_{TV}\boldsymbol{t}^{(t)}+\alpha_{VV}\boldsymbol{P}_{VV}\boldsymbol{v}^{(t)}+(1-\alpha_{TV}-\alpha_{VV})\boldsymbol{q}$,学习后 α_{TV}=0.4987676,α_{VV}=0.1003541.可以看到,两个参数在进行网络中顶点权重迭代时,不同类

© 中国科学院软件研究所 http://www.jos.org.cn

型的顶点,状态转移概率有所不同,比单一参数模型可描述更多的数据特征,更贴近真实世界状态转移机制,故精度更高.

另一方面,采用多参数马尔可夫链建模的模型(HnMTR 和 HeteRS)计算效率总体上高于采用单参数马尔可夫链的模型(RWR 和 OptRank),这是由于状态转移函数参数增多,每个函数进行矩阵相乘时所需搜索的矩阵规模缩小,效率提高.此外,HnMTR 模型的运行效率比未进行顶点同空间映射的模型 HeteRS 要高,这是因为经过同空间映射后,计算所需邻接矩阵数值范围得到有效控制,多次相乘后数值差异相对较小,使得计算效率有所提高.

5 总结与展望

本文提出一种基于异构网络的可适用于多种标签数据的标签推荐模型,该模型的优势在于既能全面包含异构网络不同顶点和边的信息,又可适用于不同类型不同维度的异构网络.本文将算法应用到具有不同顶点类型数量(本文称其为维度)的标签数据,并将应用该模型的推荐结果与当前主流算法进行对比,对比算法包括目前商用最经典的算法、基于同构网络的最优算法及其改进算法和针对特定顶点类型种类的基于异构网络算法.本模型的推荐准确率、召回率和平均准确率指标均优于商用经典算法和基于同构网络的推荐算法,在绝大多数情况下优于基于异构网络的推荐算法,从总体上达到较好的推荐结果.此外,基于本模型的推荐算法运行效率也比其他基于网络模型的推荐算法要高.未来,一方面将进一步对模型其他参数进行学习,使模型计算结果更加准确;另一方面对算法基于数据中心集群进行并行设计,进一步提高运行效率,争取获得实时的推荐结果.

References:

[1] Dong ZH. Research on the system structure and key issues of community labelling recommendation system [Ph.D. Thesis]. Tianjin: Nankai University, 2012 (in Chinese with English abstract).

[2] Begelman G, Keller P, Smadja F. Automated tag clustering: Improving search and exploration in the tag space. In: Proc. of the Collaborative Web Tagging Workshop at WWW. 2006. http://www2006.org/

[3] Xiang L. Recommended System Practice. Beijing: Post and Telecom Press, 2012 (in Chinese).

[4] Jäschke R, Marinho L, Hotho A, Schmidt-Thieme L, Stumme G. Tag recommendations in folksonomies. In: Proc. of the 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases. 2007. 506–514. [doi: 10.1007/978-3-540-74976-9_52]

[5] Vig J, Sen S, Riedl J. Tagsplanations: Explaining recommendations using tags. In: Proc. of the 14th Int'l Conf. on Intelligent User Interfaces. 2009. 47–56. http://www.iuiconf.org/

[6] Song Y, Zhuang Z, Li H, Zhao Q, Li J, Lee WC, Giles CL. Real-Time automatic tag recommendation. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2008. 515–522. [doi: 10.1145/1390334.1390423]

[7] Yang D, Xiao Y, Tong H, Zhang J, Wang W. An integrated tag recommendation algorithm towards Weibo user profiling. Database Systems for Advanced Applications, 2015,9049:353–373. [doi: 10.1007/978-3-319-18120-2_21]

[8] Phamta N, Li X, Cong G, Zhang Z. A general graph-based model for recommendation in event-based social networks. In: Proc. of the 31st Int'l Conf. on Data Engineering. 2015. 567–578. [doi: 10.1109/ICDE.2015.7113315]

[9] Feng W, Wang J. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2012. 1276–1284. [doi: 10.1145/2339530.2339729]

[10] Jenatton R, Roux NL, Bordes A, Obozinski GR. A latent factor model for highly multi-relational data. In: Advances in Neural Information Processing Systems. 2012. 3167–3175. https://papers.nips.cc/book/advances-in-neural-information-processing-systems-25-2012

[11] Zhu S, Yu K, Chi Y, Gong Y. Combining content and link for classification using matrix factorization. In: Proc. of the 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2007. 487–494. [doi: 10.1145/1277741.1277825]

[12] Shaw B, Jebara T. Structure preserving embedding. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. 2009. 937–944. [doi: 10.1145/1553374.1553494]

[13] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 701–710. [doi: 10.1145/2623330.2623732]

© 中国科学院软件研究所 http://www.jos.org.cn

[14] Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. In: Proc. of the 28th Int'l Conf. on Machine Learning. 2011. 809–816. http://www.icml-2011.org/

[15] Yuan Z, Sang J, Liu Y, Xu C. Latent feature learning in social media network. In: Proc. of the 21st ACM Int'l Conf. on Multimedia. 2013. 253–262. [doi: 10.1145/2502081.2502284]

[16] Chang S, Han W, Tang J, Qi GJ, Aggarwal CC, Huang TS. Heterogeneous network embedding via deep architectures. In: Proc. of the 21th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2015. 119–128. [doi: 10.1145/2783258.2783296]

[17] Meng X, Shi C, Li Y, Zhang L, Wu B. Relevance measure in large-scale heterogeneous networks. In: Proc. of the 16th Asia-Pacific Web Conf. 2014. 636–643. [doi: 10.1007/978-3-319-11116-2_61]

[18] Shi C, Zhang Z, Luo P, Yu PS, Yue Y, Wu B. Semantic path based personalized recommendation on weighted heterogeneous information networks. In: Proc. of the 24th ACM Int'l on Conf. on Information and Knowledge Management. 2015. 453–462. [doi: 10.1145/2806416.2806528]

[19] Wu B, Suo L. LiterMiner: A literature visual analytic system. In: Proc. of the 1st Int'l Conf. on Information Science and Engineering. 2009. 891–894. [doi: 10.1109/ICISE.2009.717]

[20] Liu X, Yu Y, Guo C, Sun Y. Meta-Path-Based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation. In: Proc. of the 23rd ACM Int'l Conf. on Information and Knowledge Management. 2014. 121–130. [doi: 10.1145/2661829.2661965]

[21] Zhang J, Yu PS, Zhou ZH. Meta-Path based multi-network collective link prediction. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 1286–1295. [doi: 10.1145/2623330.2623645]

[22] Zhou Y, Liu L, Buttler D. Integrating vertex-centric clustering with edge-centric clustering for meta path graph analysis. In: Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2015. 1563–1572. [doi: 10.1145/2783258.2783328]

[23] Wan C, Li X, Kao B, Yu X, Gu Q, Cheung D, Han J. Classification with active learning and meta-paths in heterogeneous information networks. In: Proc. of the 24th ACM Int'l on Conf. on Information and Knowledge Management. 2015. 443–452. [doi: 10.1145/2806416.2806507]

[24] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: Data mining, inference and prediction. The Mathematical Intelligencer, 2005,27(2):83–85.

[25] DadaPig. Douban reading content label data (in Chinese). http://www.datatang.com/data/43977

[26] Delicious. http://www.datatang.com/data/44194

附中文参考文献:

[1] 董振华.群落标签推荐系统体系结构及关键问题研究[博士学位论文].天津:南开大学,2012.

[3] 项亮.推荐系统实践.北京:人民邮电出版社,2012.

[25] DadaPig.豆瓣读书内容标签数据.http://www.datatang.com/data/43977

王瑜(1987－),女,河北秦皇岛人,工程师,主要研究领域为数据挖掘,机器学习,推荐系统,图模型.

吴敬征(1982－),男,博士,高级工程师,主要研究领域为信息安全,漏洞挖掘,隐蔽信道,操作系统安全.

武延军(1979－),男,博士,正高级工程师,博士生导师,CCF 高级会员,主要研究领域为系统软件与安全.

刘晓燕(1993－),女,学士,主要研究领域为基础软件与应用.

© 中国科学院软件研究所 http://www.jos.org.cn