DOI:10.3876/j.issn.1000-1980.2013.03.014

一种文献推荐的因子图方法

许卓明,张 岩,林莉莉

(河海大学计算机与信息学院,江苏 南京 210098)

摘要:为克服传统的基于内容过滤或协同过滤的文献推荐方法在缺少用户描述信息(如用户历史兴趣)的情况下无法有效地向用户推荐文献的缺陷,提出一种基于因子图的文献推荐新方法。依赖一种结合了引文网络和网络中文献主题概率分布的因子图模型,进行循环最大和(loopy max-sum)近似推理。基于权威引文网络(Thomson Reuters公司出版的2008年版期刊引证报告)的实验结果表明:该方法能在缺少用户描述信息的情况下有效地推荐主题文献,尤其适合于领域初学者获取主题相关的高质量文献。

关键词:文献推荐;因子图;主题概率分布;引文网络;循环最大和算法

中图分类号:TP311 **文献标志码**:A **文章编号**:1000-1980(2013)03-0271-07

A factor graph approach to literature recommendation

XU Zhuoming, ZHANG Yan, LIN Lili

(*College of Computer and Information Engineering, Hohai University, Nanjing* 210098, *China*)

Abstract: Traditional content-based filtering or collaborative filtering methods cannot effectively recommend literature to users without user profiles, such as the user's historical interests. To solve these problems, a novel, factor graph-based literature recommendation approach is proposed in this paper. The approach relies on a factor graph model combining a citation network and the literature's topic probability distribution and performs a loopy max-sum approximate inference on the model. The experimental results based on an authoritative citation network (Thomson Reuters' 2008 Journal Citation Reports) indicate that the proposed approach can effectively recommend topic literature in the case of missing user profiles, and it is especially suitable for beginners in a research field who want to access high-quality and topic-related literature.

Key words: literature recommendation; factor graph; topic probability distribution; citation network; loopy max-sum algorithm

文献推荐的目的是在一个文献库中向用户推荐满足用户兴趣的文献,以减少用户(尤其是领域初学者)搜索、浏览与筛选海量文献的巨大工作量。用户兴趣最普遍的表达方式是说明所需文献的主题。目前文献推荐主要有两类典型方法:内容过滤和协同过滤。内容过滤法是根据文献内容与用户兴趣的相似性来进行文献推荐,协同过滤法则是先找到与用户兴趣最相似的用户,然后根据最相似用户感兴趣的文献向当前用户推荐文献。

基于内容过滤的文献推荐代表性研究成果包括:Sugiyama等[1]假设某个研究者所发表的所有文献代表其研究兴趣,利用引文网络生成增强型的用户描述文件,并在实验中对初学者和资深研究者进行不同处理,获得了较高的推荐精度。由于传统基于内容过滤的推荐算法无法反映用户对文献的需求变化,曾春等[2]提出一种基于内容过滤的个性化推荐算法,利用领域分类模型上的概率分布来表达用户兴趣的模型,然后给出相似性计算和用户兴趣模型的更新算法。实验表明相对于传统矢量空间模型,这种概率模型能更好地表达

收稿日期:2013-01-14

基金项目:高等学校博士学科点专项科研基金(20090094110015);江苏省自然科学基金(BK2008354,BK2010520);江苏省"六大人才高峰"基金(2008135)

作者简介:许卓明(1965—),男,江苏苏州人,教授,博士,主要从事数据库、信息检索、语义网、机器学习研究。E-mail:zmxu@hhu.edu.cn

用户的兴趣变化。

传统协同过滤算法的核心问题是计算用户兴趣相似度矩阵,但随着用户数目增长,这种矩阵计算将越来越困难,算法时空复杂度的增长和用户数的增长近似成平方关系。为缓解这个性能瓶颈,Sarwar 等[3]提出了基于项的协同过滤算法,通过分析文献矩阵来确定不同文献之间的关系,并据此进行间接文献推荐。为避免传统协同过滤推荐中临近算法可能会引入噪音数据的问题,Naak 等[4]在其研发的研究论文管理系统 Papyres 中提出了基于项的多准则协同过滤方法,允许用户在评价文献时针对文献的特定部分指定兴趣,增加了文献推荐的精度。

尽管以上传统文献推荐方法得到较广泛的应用,但仍存在以下不足:(a)基于内容过滤和协同过滤都需要在具有用户描述信息的情况下才能有效地推荐文献,但对于领域初学者,往往不存在用户描述信息;(b)未充分利用文献的主题概率分布和引文网络中文献间引用关系等重要信息,难以针对文献的主题向用户推荐与其兴趣主题相关的高质量文献。

针对现有方法的不足,笔者提出一种基于因子图[5]的文献推荐新方法。该方法将引文网络和文献的主题概率分布相结合,构造主题文献推荐的因子图模型;在该因子图模型上运用循环最大和(loopy maxsum)算法进行近似推理。该方法能发挥因子图简化概率图模型及其推理的优势,可在缺少用户描述信息的情况下向用户有效地推荐若干个主题下的高相关度文献。

1 基于因子图模型的文献推荐

1.1 文献主题概率分布计算

采用 Tang 等[6]提出的 ACT(author-conference-topic)模型计算文献的主题概率分布。ACT 模型中的 conference 可泛指文献发表刊物,包括会议论文集或期刊[6]。ACT 模型是一个能综合建模文献、作者、发表刊物的主题概率模型,它利用主题分布来表示作者、文献和发表刊物之间的相互依赖关系;通过吉布斯采样(Gibbs sampling)进行参数估计,并计算给定作者的主题概率分布、给定主题的词概率分布以及给定主题的发表刊物概率分布。

根据 ACT 模型的特点,笔者将引文网络(包含文献标题、摘要、作者和发表刊物等信息)和主题数目(例如 30)作为 ACT 模型的输入,在因子图建模时自动生成给定数目的主题并计算引文网络中每篇文献的主题概率分布 $p(\boldsymbol{d}_i \mid v_i)$。

1.2 文献推荐的因子图模型

文献推荐的目标是针对输入的引文网络(含计算好的文献主题概率分布),推理产生各主题下按主题相关度从大到小排序的前 k 个文献(本文设 $k=50$),以便用户从中选取文献,此问题可作如下描述。

输入:引文网络 $G(V,E,C,D)$,其中,$V=\{v_i\}$ $(i=1,2,\cdots,n)$ 代表文献 v_i 的集合,n 为引文网络中文献总数;$E=\{e_{ij}\}$ $(j=1,2,\cdots,n)$ 代表引文网络的边集合,e_{ij} 代表文献 v_i 引用文献 v_j;$C=\{c_i\}$ 为文献被引频次的集合,c_i 为文献 v_i 被引文网络中其他文献引用的频次;$D=\{\boldsymbol{d}_i\}$ 为文献主题概率分布的集合,$\boldsymbol{d}_i=(\alpha_{1,i},\alpha_{2,i},\cdots,\alpha_{m,i})$ 为文献 v_i 在全部 m 个主题下的主题概率分布,m 为引文网络中主题数(根据引文网络覆盖的领域主题宽窄来设定,例如可设 $m=30$),主题变量 $z=1,2,\cdots,m$,$\alpha_{z,i}$ 表示文献 v_i 属于主题 z 的概率,显然 $\sum_{z=1}^{m}\alpha_{z,i}=1$。

处理:针对输入建立文献推荐因子图模型,并在该因子图模型上进行近似推理(approximate inference)。

输出:m 个主题下按主题相关度从大到小排序的相关文献列表,从中获取前 k 个文献($k=50$)。

因子图是一个二部图,用来描述某个变量集上的一个全局函数如何因式分解为若干变量子集上的局部函数的乘积[5]:

$$g(x_1,x_2,\cdots,x_n)=\prod_{j\in J}f_j(X_j) \tag{1}$$

式中:J——离散索引集;X_j——$\{x_1,x_2,\cdots,x_n\}$ 的子集,即 $X_j\subseteq\{x_1,x_2,\cdots,x_n\}$;$f_j(X_j)$——以 X_j 中元素为自变量的局部函数。

因子图有两类节点:每个变量 x_i 对应一个变量节点(用小圆圈表示),每个 $f_j(X_j)$ 对应一个因子节点(用小方块表示)。当且仅当 x_i 是函数 $f_j(X_j)$ 的自变量时,相应变量节点与相应因子节点之间才有一条无向边

相连。

笔者构建的因子图模型包含观察变量集合$\{v_i\}$、隐含变量集合$\{y_{z,i}\}$以及特征函数(包括节点函数$g_{z,i}(y_{z,i})$和边特征函数$f_{z,ij}(y_{z,i},y_{z,j})$),其中$y_{z,i}\in\{0,1\}$,是观察变量$v_i$在主题$z$下对应的隐含变量,$y_{z,i}=0$表示文献$v_i$与主题$z$不相关,$y_{z,i}=1$表示文献$v_i$与主题$z$相关。

图 1 是包含 4 个变量节点的文献推荐因子图模型。在该模型中,节点特征函数$g_{z,i}(y_{z,i})$代表文献自身特点对该文献成为被推荐文献的影响程度。由于每篇文献属于某个主题z的概率不同,概率较大的更易于成为主题z下的被推荐文献。当概率大于给定阈值时,被引频次越大的文献就越有可能成为主题z下的被推荐文献;反之,当概率小于给定阈值时,被引频次越大的文献就越有可能成为其他主题下的被推荐文献。基于以上思想,可定义$g_{z,i}(y_{z,i})$为

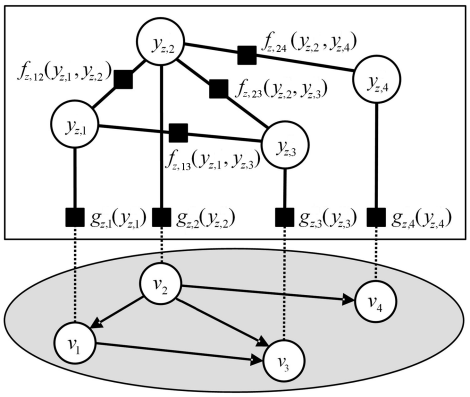


图 1　文献推荐因子图模型

Fig. 1　Factor graph model for literature recommendation

$$g_{z,i}(y_{z,i})=\begin{cases}\exp[(\beta c_i+\mu\alpha_{z,i})y_{z,i}] & \alpha_{z,i}>\lambda_z\\ \exp[-(\beta c_i+\mu\alpha_{z,i})y_{z,i}] & \alpha_{z,i}\leqslant\lambda_z\end{cases}\quad(2)$$

其中　　$\lambda_z=\dfrac{1}{n}\sum_{i=1}^{n}\alpha_{z,i}$

式中:β,μ——权重系数,$\beta,\mu\in[0,1]$,$\beta+\mu=1$;λ_z——文献的主题概率阈值。

可依据c_i(用β表示)和$\alpha_{z,i}$(用μ表示)的相对重要程度进行调整。例如,当c_i占 80%权重,$\alpha_{z,i}$占 20%权重,则$\beta=0.8$,$\mu=0.2$。

$f_{z,ij}(y_{z,i},y_{z,j})$代表文献之间引用关系对文献成为被推荐文献的影响程度。由于$\boldsymbol{d}_i$可看作向量,因此采用余弦相似度来度量文献之间的主题相似度,相似度值越大,表明文献之间的主题相似度越高,反之亦然。基于以上思想,可定义$f_{z,ij}(y_{z,i},y_{z,j})$为

$$f_{z,ij}(y_{z,i},y_{z,j})=\begin{cases}\exp(y_{z,i}\,y_{z,j}\theta_{ij}) & \theta_{ij}>\theta\\ 1 & \theta_{ij}\leqslant\theta\end{cases}\qquad(i,j=1,\cdots,n;z=1,\cdots,m)\qquad(3)$$

其中　　$\theta_{ij}=\cos(\varphi)=\dfrac{\boldsymbol{d}_i\cdot\boldsymbol{d}_j}{|\boldsymbol{d}_i|\times|\boldsymbol{d}_j|}=\dfrac{\sum_{z=1}^{m}\alpha_{z,i}\times\alpha_{z,j}}{\sqrt{\sum_{z=1}^{m}\alpha_{z,i}^2}\times\sqrt{\sum_{z=1}^{m}\alpha_{z,j}^2}}\qquad\theta=\dfrac{1}{|E|}\sum_{e_{ij}\in E}\theta_{ij}$

式中:θ_{ij}——文献v_i与v_j的主题相似度;φ——向量$\boldsymbol{d}_i$与$\boldsymbol{d}_j$间的夹角。

为了简化模型,可假设主题之间是相互独立的[7]。基于特征函数以及概率建模的特点,考虑到因子图与概率图模型的对应关系[8],可认为因子图模型对应的概率图模型是一个马尔可夫网络。根据概率论中马尔科夫网络的因式分解特性原理[8],概率图中所有节点的联合概率分布$p(y_{z,1},y_{z,2}\cdots,y_{z,n})$是图中所有最大团上势函数的乘积;而局部最大团上的势函数可定义为该团所包含的全部节点特征函数与边特征函数的乘积,因此可得出主题z下的联合概率分布为

$$p(y_{z,1},y_{z,2}\cdots,y_{z,n})=\frac{1}{\Delta}\prod_{i=1}^{n}g_{z,i}(y_{z,i})\prod_{e_{ij}\in E}f_{z,ij}(y_{z,i},y_{z,j})\qquad(i,j=1,\cdots,n;z=1,\cdots,m)\qquad(4)$$

其中　　$\Delta=\sum_{y_{z,1},\cdots,y_{z,n}}[\prod_{i=1}^{n}g_{z,i}(y_{z,i})\prod_{e_{ij}\in E}f_{z,ij}(y_{z,i},y_{z,j})]$

式中:Δ——归一化常数,亦称配分函数;$1/\Delta$——在变量空集上的一个因子,为常数。

基于因子图的推理问题属于动态规划问题,即针对式(4)定义的目标函数$p(y_{z,1},y_{z,2}\cdots,y_{z,n})$找到一个隐含变量配置空间$(y_{z,1},y_{z,2}\cdots,y_{z,n})$,使得联合概率值达到最大。

1.3　基于因子图的文献推荐推理

因子图模型是一个有环因子图,由于传统的和-积算法[5]不能有效解决有环因子图的推理问题,因此笔者选用循环最大和算法[8]。为保证推理算法收敛,选用串行消息调度机制[8],并给定循环最大和算法推理

算法中两类消息函数为

$$\mu_{f_{z,ij}(y_{z,i},y_{z,j})\to y_{z,i}}(y_{z,i}) = \max_{y_{z,1},\cdots,y_{z,n}} \left[\ln f_{z,ij}(y_{z,i},y_{z,j}) + \sum_{y_{z,s}\in \mathrm{ne}(f_{z,ij}(y_{z,i},y_{z,j}))\setminus y_{z,i}} \mu_{y_{z,s}\to f_{z,ij}(y_{z,i},y_{z,j})}(y_{z,s})\right] \tag{5}$$

$$\mu_{y_{z,i}\to f_{z,ij}(y_{z,i},y_{z,j})}(y_{z,i}) = \sum_{f_{z,s}\in \mathrm{ne}(y_{z,i})\setminus f_{z,ij}(y_{z,i},y_{z,j})} \mu_{f_{z,s}\to y_{z,i}}(y_{z,i}) \tag{6}$$

式中:$\mathrm{ne}(f_{z,ij}(y_{z,i},y_{z,j}))\setminus y_{z,i}$——与$f_{z,ij}(y_{z,i},y_{z,j})$相邻的除$y_{z,i}$外的其余变量节点的集合;$\mathrm{ne}(y_{z,i})\setminus f_{z,ij}(y_{z,i},y_{z,j})$——与$y_{z,i}$相邻的除$f_{z,ij}(y_{z,i},y_{z,j})$外的其余因子节点的集合。

当$y_{z,i}$和$f_{z,ij}(y_{z,i},y_{z,j})$作为叶子节点时,需要初始化消息:$\mu_{y_{z,i}\to f_{z,ij}(y_{z,i},y_{z,j})}(y_{z,i})=0$,$\mu_{f_{z,ij}(y_{z,i},y_{z,j})\to y_{z,i}}(y_{z,i})=\ln f_{z,ij}(y_{z,i},y_{z,j})$。当算法收敛时,可得节点$y_{z,i}$的边缘概率为

$$p(y_{z,i}) = \prod_{f_{z,s}\in \mathrm{ne}(y_{z,i})} \mu_{f_{z,s}\to y_{z,i}}(y_{z,i}) \tag{7}$$

推理任务包括:利用式(5)、式(6)进行迭代计算,使式(4)定义的联合概率值达到最大,得到此时的隐含变量配置空间$(y_{z,1},y_{z,2},\cdots,y_{z,n})$以及式(7)定义的每个隐含变量的边缘概率$p(y_{z,i})$。根据节点特征函数定义(式(2)),在某个主题$z$下,文献的边缘概率值越大,表明该文献与主题$z$越相关。完成推理计算后,根据联合概率值最大时的$(y_{z,1},y_{z,2},\cdots,y_{z,n})$选取所有满足$y_{z,i}=1$ $(i\in\{1,2,\cdots,n\})$的文献v_i构成与主题z相关的文献列表,对该列表中文献按其边缘概率值降序排序,从中选择前k个文献($k=50$)构成主题z下的被推荐文献。

2 实 验

2.1 实验数据

实验数据集由笔者访学和合作研究的机构——美国印第安纳大学图书馆与信息科学学院提供。数据集是该学院学者已发表文献[9]中实验数据集的一部分,来自于Thomson Reuters公司出版的2008年版期刊引证报告(Journal Citation Reports, JCR)。笔者从该JCR报告中选取"信息科学和图书馆学"类属中共计59个SCI期刊从2006年1月到2010年3月的全部引文数据构成一个引文网络,其中包含10 508篇文献和20 221个文献引用关系。

2.2 因子图方法的实现

2.2.1 主题概率分布计算

采用清华大学提供的ACT原型建模工具[6]进行实验。根据文献计量学领域的研究经验,一般将文献库(即引文网络)划分为30~50个主题,本文指定主题数目为30(即$m=30$,主题编号为1~30)。对实验数据集文件进行预处理,转换为满足ACT模型输入格式要求的文献库文件。将预处理后的文献库连同主题数目(30)一起输入到ACT原型建模工具中,从该原型建模工具的输出结果中选取需要的计算结果,即文献库中给定数目的主题(其编号为1, 2, …, 30)以及每篇文献在这些主题上的概率分布。例如,表1为其中8个主题及其条件概率值排序前10的词。

表1 8个主题及其条件概率值排序前10的词

Table 1 Eight topics and their top-ten words ordered by conditional probability value

主题编号	主题下条件概率值排序前10的词
1	document, interlending, electronic, digital, health, service, education, literature, university, international
2	systems, technology, model, approach, design, empirical, theory, effects, performance, product
3	productivity, knowledge, networks, web, academic, universities, network, growth, references, patents
4	science, organizational, introduction, collections, reference, services, quality, access, role, implications
5	classification, retrieval, citation, search, indexing, impact, subject, information, documentation, documents
6	publishing, journal, scholarly, care, clinical, journals, scientific, medical, uk, learning
7	behavior, humanities, characteristics, scientists, learn, lobby, dna, fingerprinting, colleges, recognition
8	relevance, evaluation, query, searching, comparison, interactive, peer, interdisciplinary, multiple, evaluating

2.2.2 因子图建模及文献推荐推理

a. 因子图建模的实现。按经验选取$\beta=0.8,\mu=0.2$;$\beta=\mu=0.5$;$\beta=0.2,\mu=0.8$共3组不同参数值。通过Java编程读入引文网络,根据文献间的引用关系计算引文网络中每篇文献的被引频次。基于已计算获得

的文献主题概率分布,每组参数值下的因子图文件生成过程如下:(a) 利用式(2)计算节点特征函数值;(b)利用式(3)的 θ_{ij} 计算文献间的主题相似度,并计算阈值 θ,再由阈值 θ 和文献间的主题相似度用式(3)计算边特征函数值;(c)利用计算得到的节点特征函数值和边特征函数值,建立满足循环最大和推理算法要求的因子图文件。

b. 基于因子图的文献推荐推理的实现。选用 libDAI 软件包[10-11]中已实现的循环最大和推理算法来完成推理任务。libDAI 软件包是一个免费的开源 C++库,能实现由离散值变量构成的各种概率图模型——包括有向图模型(贝叶斯网络)、无向图模型(马尔可夫网络、因子图)的多种精确和近似推理算法。将已建立的因子图文件作为输入,在 Ubuntu 系统下运行 libDAI 软件包中的循环最大和算法。为了保证算法收敛,设定算法最大迭代次数为 10 000 次,选择串行消息调度机制。在该算法的每次迭代中,不断更新式(5)(6)两类消息。当算法收敛时,获得的推理结果包括式(4)定义的联合概率最大值、隐含变量配置空间、式(7)定义的隐含变量的边缘概率值。

2.3　对比方法及评价基准

2.3.1　对比方法

为比较本文所提因子图算法与传统算法的优劣,选择了 2 个对比方法。由于协同过滤方法在没有用户历史描述信息的情况下无法进行文献推荐,因此本文选择了没有利用引文网络中文献间引用关系和文献主题概率分布的基于内容过滤方法,以及仅利用引文网络中文献主题概率分布的朴素方法作为 2 个对比方法。

a. 基于内容过滤方法。方法思想:通过计算引文网络中每篇文献与用户兴趣主题的相似性,向用户推荐相似性较高的文献。

方法实现:由于缺少用户兴趣主题,假设每个主题下条件概率值排名前 10 的词代表用户的兴趣主题。通过 ACT 模型输入引文网络和主题数目,计算给定每个主题下的词的条件概率分布 $p(\boldsymbol{d}_w \mid w)$($\boldsymbol{d}_w$ 为词的主题概率分布),对每个主题下的词按照该条件概率值进行降序排序,得到排序前 10 的词集合(见表 1)。利用 Java 编程计算引文网络中每个主题条件概率值排名前 10 的词在每篇文献上的 TF-IDF 值,按照 TF-IDF 值对文献进行降序排名,选取前 $k(k=50)$ 个文献组成文献推荐列表。

b. 朴素方法。方法思想:基于主题概率分布和文献被引频次进行文献推荐,在文献主题概率值大于给定阈值 λ_z 的情况下,按照文献被引频次从高到低向用户推荐文献。

方法实现:依据上述 ACT 模型生成的文献主题概率分布,利用 Java 编程将主题概率值大于 λ_z 的文献加入候选文献列表,再按照文献被引频次对候选文献列表进行降序排名,选取前 $k(k=50)$ 个文献组成文献推荐列表。

2.3.2 有效性评价基准

选择归一化折扣增益值(normalized discounted cumulative gain, nDCG)[12]作为方法有效性评价基准。nDCG 值是信息检索领域中衡量搜索算法或相关应用有效性的常用度量。由于基于主题的文献推荐可以看作是一种给定主题的文献检索,因此 nDCG 值可作为评价基准来衡量文献推荐方法获得主题高相关性文献的能力(nDCG 值越高表明方法的有效性越高)。

在计算 nDCG 值之前必须首先计算折扣增益值(discounted cumulative gain, DCG)。通过对排在文献推荐列表前 k 个文献的分级相关度累计加权计算,可得到 DCG 值。用 τ 表示 DCG 值,τ_k 的计算公式[12]为

$$\tau_k = r_1 + \sum_{\varepsilon=2}^{k} \frac{r_\varepsilon}{\log_2 \varepsilon} \quad (k = 50) \tag{8}$$

式中 $r_\varepsilon(\varepsilon=1,\cdots,k)$ 为排在文献推荐列表中第 ε 位的文献的分级相关度。

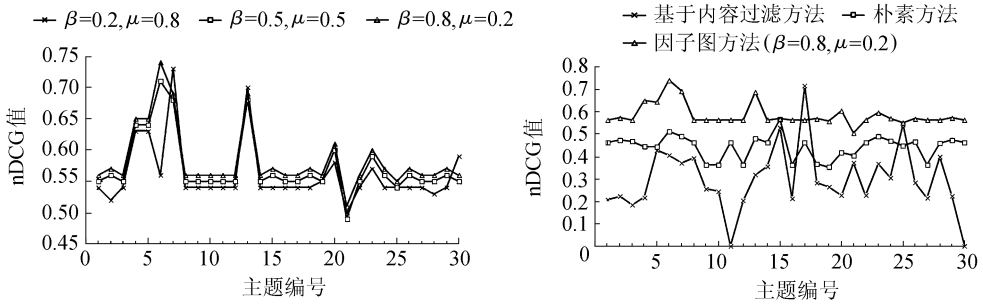
完成 DCG 值计算后,还需计算理想折扣增益值(ideal discounted cumulative gain, IDCG)。IDCG 值是指对文献推荐列表中 k 个文献按照它们的分级相关度 $r_\varepsilon(\varepsilon=1,\cdots,k)$ 降序排列后,重新计算得到的 DCG 值。用 η 表示 nDCG 值,用 ω 表示 IDCG 值,得到 η_k 计算公式[12]为

$$\eta_k = \frac{\tau_k}{\omega_k} \ (k = 50) \tag{9}$$

采用 Java 编程实现 nDCG 值的计算。通过 ACT 模型设定 λ_z,获得每个主题的综述文献,文献的分级相关性定义为其被综述文献引用的频次。

2.4 实验结果与讨论

图2是因子图方法与基于内容过滤方法及朴素方法在30个主题下分别推荐前50篇文献所得到nDCG值的比较。如图2(a)所示,在因子图方法中,当参数取值为$\beta=0.8$,$\mu=0.2$时,得到的nDCG值优于另外2组参数的nDCG值,表明此参数值下因子图方法的文献推荐效果最佳。由于β代表文献被引频次所占权重,μ代表文献主题概率所占权重,因此实验结果进一步说明因子图方法中文献被引频次对文献成为被推荐文献的影响比文献主题概率分布的影响要大。如图2(b)所示,与朴素方法相比,因子图方法($\beta=0.8$,$\mu=0.2$)在30个主题上的nDCG值均大于朴素方法;由于朴素方法没有使用引文网络,进一步表明引文网络的引入可以提高文献推荐效果。与基于内容过滤方法相比,因子图方法在其中29个主题上完全优于基于内容过滤方法,说明因子图方法综合考虑了文献本身特点以及引文网络,对推荐效果有很大提升;其中一个主题(编号为17)下的推荐效果差于基于内容过滤的方法,说明文献内容因素对文献成为被推荐文献也有一定影响。



(a) 因子图方法在3组参数值下的nDCG值 (b) 3种方法的nDCG值比较

图2 nDCG值计算结果比较

Fig. 2 Comparison of calculated nDCG values

实验结果表明,在缺少用户描述信息时,因子图方法总体上明显优于传统的基于内容过滤方法和朴素方法。

3 结 语

为了克服传统文献推荐方法在缺少用户描述信息(如用户历史兴趣或当前兴趣主题)时无法有效地向用户推荐文献的缺陷,笔者提出一种基于因子图模型的文献推荐新方法。该方法通过将引文网络信息(包括文献发表信息、文献引用关系及被引频次等)及文献本身特点(文献的主题概率分布)进行统一建模,实现利用主题概率模型自动生成引文网络中给定数目的主题,并计算出各主题下按主题相关度从大到小排序的若干个相关文献,供用户选择。该方法还能充分发挥因子图简化概率图模型及推理的优势,提高主题文献推荐的效果。基于权威引文网络(Thomson Reuters公司出版的2008年版期刊引证报告)的实验结果表明,在缺少用户描述信息的情况下,因子图方法的推荐效果明显优于基于内容过滤方法和朴素方法等传统方法,从而为文献推荐研究提供了新思路。因子图方法涉及多个参数(β,μ),为了简化实验,笔者根据经验选择了几组不同的参数值进行实验,下一步可考虑通过参数学习的方法来训练出具有最佳文献推荐效果的最优参数值组合。

致谢:感谢美国印第安纳大学图书馆与信息科学学院博士研究生Yan Erija为本文研究提供了实验数据集;感谢清华大学计算机科学与技术系唐杰副教授为本文实验提供了ACT原型建模工具。

参考文献:

[1] SUGIYAMA K, KAN M Y. Scholarly paper recommendation via user's recent research interests [C]//HUNTER J, CARL L, GILES C L, et al. Proceedings of the 10th annual joint conference on digital libraries. New York: ACM, 2010: 29-38.

[2] 曾春, 邢春晓, 周立柱. 基于内容过滤的个性化搜索算法[J]. 软件学报, 2003, 14(5): 999-1004. (ZENG Chun, XING Chunxiao, ZHOU Lizhu. Personalized search algorithm using content-based filtering[J]. Journal of Software, 2003, 14(5): 999-1004. (in Chinese))

[3] SARWAR B M, KARYPIS G, KONSTAN J A, et al. Item-based collaborative filtering recommendation algorithms[C]//SHEN

V Y, SAITO N, LYU M R, et al. Proceedings of the 10th international world wide web conference. New York: ACM, 2001: 285-295.

[4] NAAK A, HAGE H, AÏMEUR E. A multi-criteria collaborative filtering approach for research paper recommendation in Papyres [C]//BABIN G, KROPF P, WEISS M. Proceedings of the 4th International MCETECH Conference on E-technologies, as Lecture Notes in Business Information Processing (LNBIP) vol. 26. Heidelberg: Springer Berlin Heidelberg, 2009: 25-39.

[5] KSCHISCHANG F R, FREY B J, LOELIGER H-A. Factor graphs and the sum-product algorithm[J]. IEEE Transactions on Information Theory, 2001, 47(2): 498-519.

[6] TANG Jie, JIN Ruoming, ZHANG Jing. A topic modeling approach and its integration into the random walk framework for academic search[C]//GIANNOTTI F, GUNOPULOS D, TURINI F, et al. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society, 2008: 1055-1060.

[7] TANG Jie, SUN Jimeng, WANG Chi, et al. Social influence analysis in large-scale networks[C]//ELDER J, FOGELMAN F S, FLACH P, et al. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009: 807-816.

[8] BISHOP C M. Pattern recognition and machine learning[M]. Berlin: Springer, 2006: 359-422.

[9] YAN Erjia, DING Ying. Library and information science (LIS) as we see it: an overview at the state and country level from 1965-2010[C]//Proceedings of the American Society for Information Science and Technology. Malden MA, USA: Wiley Subscription Services, Inc,2011:1-8.

[10] MOOIJ J M. libDAI: A free and open source C++library for discrete approximate inference in graphical models[J]. Journal of Machine Learning Research, 2010, 11: 2169-2173.

[11] MOOIJ J M. libDAI-A free and open source C++library for discrete approximate inference in graphical models[EB/OL]. [2013-01-14]. http://cs.ru.nl/~jorism/libDAI/.

[12] JÄRVELIN K, KEKÄLÄINEN J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems, 2002, 20(4): 422-446.

·简讯·

第三届江苏省海洋湖沼学会研究生论坛将在南京举办

为提高海洋湖沼领域科学研究和社会服务水平,促进青年学者的成长,江苏省海洋湖沼学会将于 2013 年 9 月上旬在南京举办第三届江苏省海洋湖沼学会研究生论坛。

本次论坛由江苏省海洋湖沼学会主办,河海大学青年科协、河海大学研究生科协、国际水利与环境工程协会(IAHR)河海大学学生分会承办,南京师范大学研究生会、南京大学研究生会、南京信息工程大学研究生会、中国科学院南京地理与湖泊研究所研究生会联合协办。

论坛以“聚叙湖海情,共铸生态梦”为主题,设河湖水环境与生态修复、海岸带资源开发与保护、海岸灾害预测与防治、水信息学与海洋遥感 4 个方向。

联系人:河海大学研究生科协 钱学生

联系电话:15850664066

通信地址:210098 南京市西康路 1 号　河海大学港口海岸与近海工程学院

(本刊编辑部供稿)