基于主题模型的跨学科协作文献推荐

任 柯 黄智兴 邱玉辉

(西南大学计算机与信息科学学院 重庆 400715) (西南大学语义网格实验室 重庆 400715)

摘 要 Web中存在着海量的各类科技文献,研究人员虽然可以利用各种搜索工具对这些文献进行检索,但是,如何高效地找到与自身研究相关的文献变得越来越困难。最近出现的一系列在线研究者社区为解决这一问题提供了一种新的方案。提出一个基于主题模型的协作文献推荐,此模型将传统的协同过滤和概率主题模型,以及知识协作网络模型相结合,提供了一个可判别的隐语义结构。在考虑不同的用户评价所给出的文献索引率,以及新发表的文献的主题分布的基础上,利用语义相似度的计算工具,提出基于概率的跨学科的检索推荐。采用来自于CiteULike的一组数据,验证了该方法的有效性和可行性。

关键词 协作,主题模型,文献推荐,跨学科

中图法分类号 TP301 **文献标识码** A

Interdisciplinary Collaborative Literature Recommendation Based Topic Modeling

REN Ke HUANG Zhi-xing QIU Yu-hui

(Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China)

(Semantic Grid Laboratory, Southwest University, Chongqing 400715, China)

Abstract There exists plenty of all kinds of literature, although the researchers can use some sorts of searching tools for literature retrieval, but how to seek out more efficiently relevant literature becomes more and more difficult. Recently appearing in a series of online community for the researchers is a new solution. A model based on topic model of interdisciplinary literature recommendation was presented, the combination of traditional collaborative filter and probability topic model, and knowledge collaboration network model has been put forward, so it provides a latent semantic structure which can be distinguishable. In terms of different user's appraisal of the given literature index rate, and topic distribution of newly published literature in the foundation, This paper used semantic similarity calculation tools, and puts forward the retrieval recommendation for interdisciplinary research based on probability. For this reason, we studied a set of data from the CiteULike, and experimental results show the feasibility and effectiveness of this method.

Keywords Collaboration, Topic model, Literature recommendation, Interdisciplinary

1 引言

面对数以百万计的专业文献,研究者已经无法满足于单一的搜索方法,因此如何更有效地将不同期刊和数字图书馆、在线社区中的文献以一种更为结构化的方式进行整合,找到与研究者兴趣相近的相关文献,并按照这些文献的主题进行深入分析,已成为当前研究的热点问题[1-3]。

不同文献间由于结构的不一致,导致不同文献在表达相同或类似语义的表现形式截然不同,因此,研究者需要在快速增长的文献集合中,寻找到符合要求的文献,最大限度地提高文档网络的效用。本文着重介绍如何利用主题模型,引入学科间参数,利用文献间协作来进一步发现文献间的语义关系的方法。

2 基本思想

检索文献,其目的就是为了发现文献间的各种语义关系,一般而言,现有检索各类文献的方法有以下两种:

1. 经由已阅读文献的引文。这样的方法将研究者局限在特定的引用社区中,并且这样的方式更倾向于被引用次数较多的文献。基于统计学的传统方法可能会因此忽略掉其他学科,例如物理学、生物学中的相关文献,而且更有可能的是,所引用的文献作者本身就忽略了这些文章。

2. 利用关键字搜索。这种方法比较有效,但仍然有局限性。因为对于那些不知道应该搜索什么内容的研究者而言,如何形成查询关键词非常困难。关键字搜索是基于内容的搜索,这样的搜索找到的文献是其他研究者认为有价值的文献,这种搜索可以认为是一种有向性搜索,搜索者希望提前获知搜索未知文献的"种子"——也就是搜索关键字。

基于上述分析,我们构建一个新的模型。首先,该模型需要推荐之前的"旧文献",因为,研究者总是希望从过去的文献中发现新的研究领域,并因此获得所在领域的研究基础。在选择旧文献时,其他研究者的评价是有价值的,一篇基础性的

到稿日期:2012-02-08 返修日期:2012-06-12 本文受高校基本业务基金(XDJK2010C035),留学回国人员基金(20091001)资助。

任 柯(1977-),男,博士生,讲师,主要研究方向为语义网格,E-mail:jacky711@swu.edu.cn;黄智兴(1974-),男,博士,副教授,主要研究方向为人工智能、语义网格;邱玉辉(1938-),男,教授,博士生导师,主要研究方向为人工智能、语义网格。

文献会出现在许多研究者的引文库中，相反，较为不重要的文献出现的次数就相对较少。其次，推荐“新文献”也是重要的，因为当前的研究者们总是首先从他们所在学科的最新发表的文献中选择文献。由于这些文献很新，研究者很难及时地将它置于自己的引文库中，而传统的协同过滤方法也就很难及时给予准确的推荐。因此，对于新发表的文献，需要基于其内容的推荐。最后，我们考虑引入探索性变量，利用这些变量来总结和描述每个研究者的基于文献内容的倾向性，虽然 BLEI 在 CTR(Collaborative Topic Regression)[4] 整合了原有历史文献的研究者评价和最新发表文献的文献内容本身，但没有考虑学科交叉对于引文评价值的问题，因此我们在本文中引入知识协作模型，基于隐因素模型协同过滤和基于概率的主题模型的对新发表文献的内容分析基础上，引入研究者之间的学科距离和通信模式这样的超参数，利用语义相似度计量方法来提高协作网络中引文推荐的质量。

3 背景知识

3.1 概率主题模型

概率模型算法[2]用于发现文献集合中的一组主题，主题是基于一组单词的概率分布，而这些单词又是围绕一个主题呈现而构建的。主题模型提供一种多文档的可辨识的低维表示方式[5]，被广泛用于语料库搜索、文档划分以及信息检索等。

本文中采用当前应用较多的概率主题模型：LDA 模型[6,7]。假设有 T 个主题，每一个主题都是基于给定词集合的概率分布，LDA 的生成过程分两步：由文档生成主题，由主题生成对应的词汇，基本步骤如下：

对于给定语料库 D 中的每一篇文档 w：

1. 提取主题比例 $\theta_i \sim \mathrm{Dirichlet}(\alpha)$

2. 对每一个单词 w_i

a)提取主题值 $z_{jn} \sim Mul(\theta_i)$

b)提取单词 $\omega_j \sim Mul(\beta_{z_{jn}})$

其中，$D=\{w_1, w_2, \cdots, w_M\}$ 代表语料库 $(1 \leqslant j \leqslant M)$，$w=(w_1, w_2, \cdots, w_N)$ 代表一篇文档，w_i 代表一篇文档中的第 i 个单词 $(1 \leqslant i \leqslant N)$，一个语料库中有 M 篇文档以及 V 个单词，一篇文档由其中 N 个单词构成 $(1 \leqslant M, 1 \leqslant N \leqslant V)$。

β 是一个 $k \times V$ 的矩阵，代表主题中各个单词的概率，$\beta_{ij}=p(w^j=1 \mid z^i=1)$，$\theta_i$ 是一个 K 维的 Dirichlet 随机变量（K 可预先设定，$\sum_{i=1}^{k}\theta_i=1$），由式(1)：

$$p(\theta \mid \alpha)=\frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_1^{\alpha_k-1} \tag{1}$$

式中，α 是 k 重向量（分量 $\alpha_i>0$），$\Gamma(x)$ 是伽马函数。

这一过程解释了每篇文献的单词是如何从给定主题中提取得到的：主题比例是特定文献决定的，但主题集合是整个语料库共享的。

在给定数据集的前提下，主题的后验概率分布揭示了如何利用该数据集生成对应的文献。与传统的聚类模型不同的是，LDA 允许多个文献包含多个主题。例如，LDA 可以提取一篇文献中涉及计算机和心理学的主题，与此同时，也可以提取另一篇文献中关于心理学和统计学的主题。由于 LDA 是一种无监督学习方法，可以从该语料库中同时发现“计算机”、“心理学”、“统计学”的多个主题。而且，这种混合隶属度也有助于在共现模式下对单词的快速评估。

在给定文献语料库之后，我们可以使用变分 EM[8,18] 算法来完成对主题的学习，利用文献[9]进行文献的分解，针对特定文献，我们可以进一步利用变分推理寻找基于这些主题的内容。我们希望利用主题模型在文献推荐系统中给予基于内容的表示方法。

3.2 推荐目标

在一个推荐系统中的两个基本元素就是用户和物品。在我们的推荐系统中，用户就是研究者，物品就是文章。假设有 I 名研究者及 J 篇文章，评价变量 $r_{ij} \in \{0,1\}$ $(1 \leqslant i \leqslant I, 1 \leqslant j \leqslant J)$，表示研究者 i 将文章 j 收录到他的引文库中[3]。这里，$r_{ij}=0$，表示研究者 i 要么对文章 j 不感兴趣，或者是不知道文章 j 的存在；$r_{ij}=1$，表示研究者 i 对文章 j 感兴趣。我们的推荐目标就是找到那些当前不在研究者自身引文库中，但却是他们可能感兴趣的文章。为了描述推荐目标，文献[4]利用两个推荐矩阵。

推荐矩阵如图 1 所示，其中√表示“喜欢”，×表示“不喜欢”，？表示“不知道”。

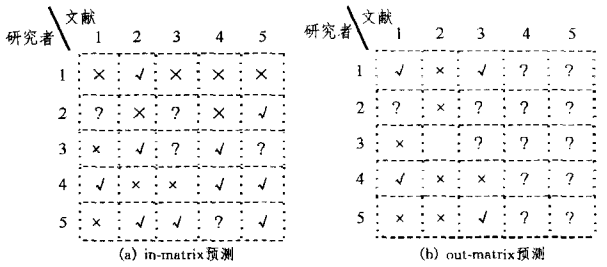


图 1 推荐矩阵

in-matrix 预测，用于说明在网络数据库中被其他用户至少评价过一次的文献，这种推荐在传统的协同过滤系统中经常使用。

out-matrix 预测，图 1(b)中文献[4,5]从未被任何用户评价过，对于这样的文献，传统的协同过滤算法是无能为力的，但对于该领域研究者而言，他们总是希望了解最新的研究成果，所以，推荐系统需要新的方法来处理这些“不知道”的文献。

3.3 基于矩阵分解的推荐

传统的推荐方法是协同过滤，基本思想是分析用户的兴趣，在用户群中找到与指定用户相似(兴趣)的用户，综合这些相似用户对某一信息的评价，形成系统对特定用户对此信息的喜好程度的预测，这里需要注意的是，传统的协同过滤没有考虑到推荐文献自身的内容。近来，出现了一些更为有效的推荐方法，例如，隐因子模型[12-14]以及相邻方法[20]，从实验结果上来看，隐因子模型要优于相邻方法。本文基于隐因子模型提出了隐语义因子模型。

隐因子模型的各种方法中，矩阵分解[12]是比较好的一种。在进行矩阵分解的过程中，提出研究者和推荐文献共存于一个维度为 K 的隐低维空间，研究者 i 用隐向量 $u_i \in R^k$ 表示，文献 j 用隐向量 $v_j \in R^k$ 表示。利用两个向量间的内积预测研究者 i 是否倾向于文献 j。

$$\hat{r}_{ij}=u_i^{\mathrm{T}}v_j \tag{2}$$

为了利用矩阵分解，需要基于给定的已知评价矩阵计算

研究者和文献的隐表征值。一般的方法是将规范化后的平方误差值最小化，这里的$U=(u_i)_{i=1}^{I}$、$V=(v_i)_{j=1}^{J}$。

$$\min_{U,V}\sum_{i,j}(r_{ij}-u_i^{\mathrm{T}}v_j)^2+\lambda_u\|u_i\|^2+\lambda_v\|u_j\|^2 \tag{3}$$

式中，λ_u 和 λ_v 是正则化参数。

用于协同过滤的分解矩阵可由一个概率主题模型获得，在基于概率的语义矩阵分解中，有以下生成过程：

1. 对每一个研究者 i，提取研究者隐向量

$u_i\sim N(0,\lambda_u^{-1}I_K)$

2. 对每一篇文献 j，提取文献隐向量

$v_j\sim N(0,\lambda_v^{-1}I_K)$

3. 对每一个研究者-文献(i,j)，提取对应值

$$r_{ij}\sim N(u_i^{\mathrm{T}}v_j,c_{ij}^{-1}) \tag{4}$$

式中，c_{ij} 是 r_{ij} 的准确度参数，I_K 是一个 K 维的单位矩阵。

当 $c_{ij}=1$，对于任意的 i,j，最大化概率矩阵分解的后验概率期望就是式(2)的计算结果。这里，精确度参数 c_{ij} 作为评估 r_{ij} 的信任度参数。如果 c_{ij} 增加，我们就更加信任 r_{ij}。r_{ij} 可以做两种解释，研究者 i 要么对文献 j 不感兴趣，要么不知道文献 j 的存在。在文献[11,21]中，基于电视节目和新闻推荐讨论了一阶协同过滤问题，不同的作者提出了不同的信任参数 c_{ij}。在这里，使用相同的策略来设定 c_{ij}。

$$c_{ij}=\begin{cases}a, & 若\ r_{ij}=1\\ b, & 若\ r_{ij}=0\end{cases} \tag{5}$$

式中，a,b 都是调节参数，满足，$a>b>0$，$r_{ij}\in\{0,1\}$。

3.4 语义相似度

语义检索方法通常使用本体知识库中的概念来表达用户的查询需求，因而需要分析概念之间的语义相似度，来判断概念与用户需求之间的相关程度[17]。在语义信息检索研究中，语义相似度的计算方法可以分为 3 类：路径长度方法[24]、信息论方法[25]和基于概念特征的方法[26]。传统的计算语义相似度的方法是在单一本体内部计算概念间语义距离，这种单一本体要么是一个领域独立的本体要么是多个已有本体的集成。文献[16]整合了基于边和基于节点的语义相似度计算方法，给出语义相似度的基本计算公式：

$$Dist(w_1,w_2)=\sum_{c\in\{path(c_1,c_2)-LSuper(c_1,c_2)\}}wt(c,parent(c)) \tag{6}$$

式中，$c_1=sen(w_1)$，$c_2=sen(w_2)$，$path(c_1,c_2)$是从 c_1 到 c_2 的最短路径所有节点的集合，$LSuper(c_1,c_2)$代表这个集合中的任意一个元素，是 c_1,c_2 的最小上位范畴词。

但是在跨本体库的语义信息检索中，实例往往存在比较复杂的多重继承关系，即一个实例可能同时存在多个父类的情况，如图 2 所示。c_1、c_2 各为一篇“认知心理学文章”，c_2 同时也是一篇“语义网文章”，在计算相似度时，若只考虑单一继承关系，即只考虑 c_1、c_2 作为“认知心理学”文章的特征，那么 $\mathrm{sim}(c_1,c_2)=\mathrm{sim}(c_2,c_3)$，但是，由于 c_2、c_3 均为一篇“语义网文章”，因此 $\mathrm{sim}(c_1,c_2)<\mathrm{sim}(c_2,c_3)$。

因此参考文献[17]，使用实例语义相似度计算公式：

$$\mathrm{sim}(c_1,c_2)=\begin{cases}\theta*\mathrm{sim\_inherit\_relation}(c_1,c_2)+\\ \lambda*\mathrm{sim\_property}(c_1,c_2)+\\ \gamma*\mathrm{sim\_property\_value}(c_1,c_2) & c_1\neq c_2\\ \mathrm{max\_similarity} & c_1=c_2\end{cases} \tag{7}$$

式中，max_similarity 为实例相似度的最大值，sim_inherit_ralation(c_1,c_2)为实例继承关系相似度，sim_property(c_1,c_2)为实例属性相似度，sim_property_value 为实例属性值相似度。$0\leqslant\theta,\lambda,\gamma\leqslant1$，$\theta+\lambda+\gamma=1$。

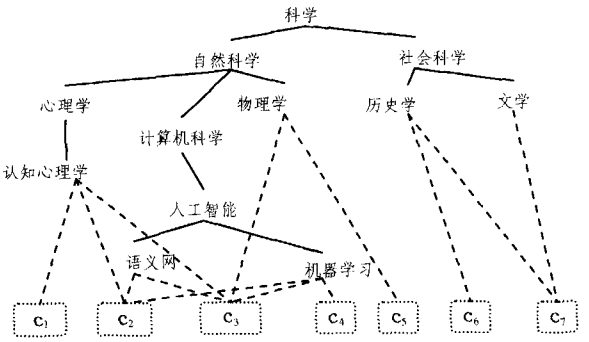


图 2 本体实例多重映射

3.5 协同主题回归

协同主题回归(CTR)模型，整合了传统的协同过滤和主题模型。其整合的思路是构筑一个模型，使用隐主题空间模型解释观察得到的评价值以及词汇表。例如，可以利用主题比例 θ_j 替换式(4)中的隐文献参数，隐向量 v_j：

$$r_{ij}\sim N(u_i^{\mathrm{T}}\theta_j,c_{ij}^{-1}) \tag{8}$$

这种模型的不足在于它不能判断这些主题推荐是基于主题本身的更重要还是基于评估值的更重要。考虑文献 A、B 都是将激活扩散模型应用于社会网，比较相似，因此，它们拥有近似的主题值 θ_A 和 θ_B。而不同的研究者对这两篇文献的感兴趣程度可能是不一致的：文献 A 可能给出了一个可以应用于社会网的新的激活扩散模型的变例，文献 B 应用了一个具体的激活扩散模型，但是给出了基于一组社会网数据的一个重要分析结论。

研究领域是基于认知心理学的研究者会倾向于选择文献 A，研究领域是基于社会网的研究则相反。然而使用式(7)中的主题比值将有可能做出针对两篇文献近似的推荐。协同过滤回归可以检测出其中的差异。

以上是 BLEI 在文献[4]里面提出的 CTR 模型论及的内容，但是没有解决如何度量两个研究者的研究领域的差异对文献推荐质量的影响问题。

4 跨学科协作文献推荐

在原有的 CTR(Collaborative Topic Regression)基础上引入协作网络模型[18]，用以描述基于相同或类似主题兴趣的不同研究者，所讨论的文献均使用一个主题模型构造得到。在此，提出基于主题模型的跨学科协作文献推荐 ICLRBTM (Interdisciplinary Collaborative Literature Recommendation Based Topic Modeling)。

假定有 N 个研究者(其中有一个是具有代表性的研究者)用 i,j 表示，这里的研究者可以是一个独立的个体，也可能是一个特定的学科或者他所代表的学科的教育、科研经验的综合。研究者之间可以通过某些特定的通信模式 P_i 进行协作，其中有一种典型模式，用 K 表示。图 3 描述了知识协作模型，不同的通信模式用不同的弧线或者直线表示。

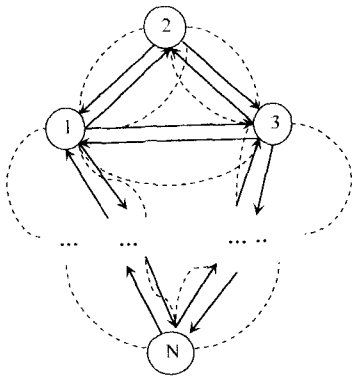


图 3　协作网络结构

在 Beckmann(1995)[19,20]中提出了对于协作网络模型的一组约定：

1. 研究者仅仅以成对的形式协作。

2. 每对研究者的协作必须相互达成一致。

与 Beckman 的假设(即两个研究者之间的物理距离决定了协作时间)不同的是，我们在给定了某种通信模式的前提下，协作时间也受到虚拟距离的影响，可以将两个研究者的虚拟距离看作是研究者们所从事的学科间距离，这其实就是研究者之间的知识距离。对于特定主题的知识，使用主题模型来提取。当两个高度分离的研究领域相比较于两个非常相关的研究领域时，虚拟距离的变化是显著的。在表 1 中基于文献[18]，给出了协作知识推荐网络模型的相关参数。

表 1　协作知识推荐网络模型的参数

符号	定义
a_{ijp}	研究者 i 和研究者 j 通过模式 p 交流的系数
r_{1ij}	研究者 i 和研究者 j 的虚拟距离
r_{2ij}	研究者 i 和研究者 j 通过模式 p 交流的通信距离
r_{ijp}	$r_{ijk}=r_{1ij}+r_{2ij}$。研究者 i 和研究者 j 通过模式 p 沟通的总距离
t_{ijp}	$t_{ijk}=1+a_{ijp}\,r_{ijp}$。研究者 i 和研究者 j 在模式 p 中为了达到一个有效协作的单位时间所花费的实际时间

前面已经提到，已有的协作主题推荐利用主题关注度，并认为所有文档均有一个主题模型生成。在此基础上，协作文献推荐模型引入研究者之间的知识距离 r_{ijp}，用以作为构建研究者评价时对主题比值的修正值，以此说明学科差异对于推荐主题的影响。

为了将语义距离引入到 LDA 中，在本模型中定义了一个马尔科夫随机场，随机场中的条件概率 $P(z,\omega,\theta|\alpha,\beta,w,d,OB)$：

$$\exp\left(\sum_{p}^{P} r_{ijp}\right)\left(\prod_{t}^{T} p(\omega_t|\beta)\right)\left(\prod_{j}^{D} P(\theta_j'|\alpha)\right)\times \left(\prod_{i}^{N} \omega_{z_i}(w_i)\theta'_{di}(z_i)\right) \tag{9}$$

这里给出 r_{ijp}，对于原有 θ_j 进行调整：

$$r_{ijp}=\sum_{p=1}^{L}\varphi_p \mathrm{sim}(i,j) \tag{10}$$

$$\theta_j'=\theta_j+r_{ijp} \tag{11}$$

$OB=\{\varphi_1,\varphi_2,\cdots,\varphi_L\}$。$OB$ 是合取范式，$\varphi_i(1\leqslant i\leqslant L)$代表在协作网络中各个学科具体分支所对应节点的权重值，用以描述领域专家对于特定领域内学科交叉时具体研究分支间的联系紧密程度。

协作模型的生成过程如下：

1. 对每一个研究者，提取用户隐向量
$u_i\sim N(0,\lambda_u^{-1}I_K)$。

2. 对每一篇文献 j，

a) 提取主题比值 $\theta_i\sim$Dirichlet(α)。

b) 提取文献隐修正值 $\varepsilon_j\sim N(0,\lambda_v^{-1}I_K)$，$v_j=\varepsilon_j+\theta_i$。

c) 对每一个单词 ω_{jn}，

i. 提取主题值 $z_{jn}\sim$Mult(θ_i)，

ii. 提取单词 $\omega_{jn}\sim$Mult($\beta_{z_{jn}}$)。

3. 对每一个研究者-文献对(i,j)，提取对应值
$r_{ij}\sim N(u_i^{\mathrm{T}}v_j,c_{ij}^{-1})$ (12)

5　实验结果

设计基本的实验来验证本文提出的跨学科文档推荐发现方法的有效性和可行性。分析了一个真实的研究者社区[1)]及归属于他们的各类索引文献。

实验数据集来源于 CiteULike，在 CiteULike 上，注册用户可以设定多个不同的研究领域以及与之对应的引文库，每篇引文都包含文章标题、摘要以及所属讨论群组(本文所需要使用的信息主要来自于这 3 个方面)。

我们从 CiteULike 获取的 3 类数据集(Who-posted-what data、Who-posted-what data、Group membership data)中，通过预处理(移去空白文章，合并重复文章，移去引文数少于 10 篇的用户信息)，最后得到 5855 个注册用户、17825 篇文章以及 22364 个用户-商品对。

在后续的实验中将分析一个用户引文和发表文章的集合。使用所提取样本集合中的引文评价和文章来评价之前提出的推荐算法。我们的推荐系统为每位用户推荐 X 篇文章，并基于每位用户引文库中的所有文章进行评价。

一般而言，推荐系统采用准确度矩阵和召回率矩阵来评价推荐质量，然而如同第 3 节所述，评价值为零不能给出确切推荐，它可能代表一个研究者对该文献“不喜欢”或者“不知道”。这使得我们很难计算精确度矩阵，然而，由于 $r_{ij}=1$ 是正向激励的，因此需更多地考虑召回率。召回率只考虑前 X 篇文章中那些正相关的文章。对每一个研究者而言，召回率的定义[21]：

$$R=\frac{N_n}{N_r}$$

式中，N_n 是所有相关文章中被推荐系统选中的文章数量，N_r 是所有与研究者感兴趣领域相关文章的数量。这里使用的召回率是整个推荐系统给所有研究者的平均召回率。

实验考虑两个推荐目标：In-matrix 预测和 Out-of-matrix 预测。其中 In-matrix 预测可以分析出在当前研究者访问该组文章之前，那些至少被评价过一次的“老文章”，而 Out-of-matrix 则预测那些新近发表的还从未被访问过的“新文章”。

In-matrix 预测：我们将数据分为训练数据和测试数据，并进行 10 倍交叉验证[22]。对于那些出现次数少于 10 次的文章，我们将其置于训练集中，这就可以保证所有在测试集中的文章一定在训练集中。

Out-of-matrix 预测：我们仍然进行 10 倍的交叉验证。首先，将所有文章分为 10 份，轮流将其中 9 份用作训练集，1 份做测试，10 次的结果均值作为算法召回率的估计。

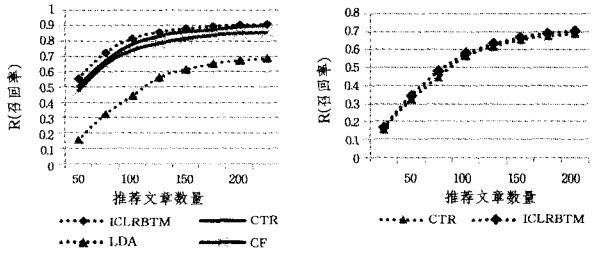
1)　http://www.citeulike.org/faq/data.adp

对于协同过滤(CF)中的矩阵分解,使用网格搜索,确定参数 $K=180,\lambda_u=\lambda_v=0.01,a=2,b=0.02$。

对于协同主题回归(CTR)模型,$K=180,\lambda_u=0.01,a=2,b=0.02$ 与 CF 相同,$\lambda_v\in\{10,100,1000,10000\}$。

对于基于主题模型的跨学科协作文献推荐(ICLRBTM),学科间语义距离利用文献[24]中对多学科距离的定义。

图 4 说明了 in-matrix 预测和 out-of-matrix 预测对于不同模型的整体性能,实验中,设定随机推荐的文档值 $Q=30,60,\cdots,180$。



(a) 针对 in-matrix 的召回率对比 (b) 针对 out-of-matrix 的召回率对比

图 4 不同模型召回率对比(基于不同的推荐文章数量值)

实验结果显示,对于 in-matrix 预测而言,ICLRBTM 与矩阵分解方法和 CTR 类似,相比于 LDA,其在原有的基于文档自身内容的主题模型分析基础上,加入文章评价机制后,模型的召回率性能有所提升,随着提供的测试的文档数量的增加,预测性能还有所增强。

而对于 out-of-matrix 预测,ICLRBTM 和 CTR 的召回率都略有下降,但是相比于 CTR,ICLRBTM 仍然获得了较高的召回率,因此,我们认为基于语义相似度的跨学科空间文献推荐模型是有效的、可行的和实用的。

结束语 本文探讨了一种基于主题模型的跨学科科技文档推荐模型,使得研究人员跨学科相互利用彼此的学科研究成果成为可能。其主要的目的是为了增强信息检索的质量,自动推荐不同学科间语义相关的科技文档,使得研究人员对文献的搜索工作质量更高。本文介绍了主题模型的定义、协作网络的基本架构,以及隐因子模型中的矩阵分解方法,并基于主题模型讨论了度量不同层级的文档间相关度的方法,进一步根据文档相关度,利用语义相似度计算工具,将语义相关的文档映射在分级结构的不同层次上,从而逐步发现不同层级上不同学科间的文档语义关系。

由于不同期刊和数字图书馆、在线社区中的文献之间复杂的相互关系,文档之间的关系会不时地动态变化,而且不同学科间的语义关系也不一定只能是本体中的分层关系,如何更好地定义学科间的关联关系是值得探讨的另一个重要问题,是将来研究的进一步工作。

参 考 文 献

[1] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions[J]. IEEE Trans. Knowl. Data Eng., 2005: 17(6): 734-749

[2] Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering[C]// SIGI '99 Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, USA, 1999: 230-237

[3] Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets[C]// Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 2008: 263-272

[4] Wang Chong, Blei D M. Collaborative topic modeling for recommending scientific articles[C]// KDD '11 Proceeding of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Datamining. 2011: 448-456

[5] Hofmann T. Probabilistic Latent Semantic Indexing[C]// SIGIR 1999 Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieva. Berkeley, CA, USA. ACM, 1999: 50-57

[6] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[C]// NIPS 2001. Vancouver, British Columbia, Canada, 2001: 601-608

[7] Blei D M, John D. Lafferty: Dynamic topic models[C]// ICML. 2006: 113-120

[8] Dempster A, Laird N, Rubin D. Maximam likelihood from imcomplete data via the EM algorithm[J]. Joarnal of the Royal Statistical Society, SeriesB, 1977, 39(1): 1-38

[9] Cohen S B, Blei D M, Smith N A. Variational Inference for Adaptor Grammars[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings. Los Angeles, California, USA. The Association for Computational Linguistics, 2010: 564-572

[10] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[11] Koren Y, Bell R M, Volinsky C. Matrix Factorization Techniques for Recommender Systems[J]. IEEE Computer, 2009, 42(8): 30-37

[12] Salakhutdinov R, Minih A. Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo[C]// Proceedings of the 25th International Conference on Machine Learning. ACM, 2008: 880-887

[13] Salakhutdinov R, Mnih A Probabilistic matrix factorization[C]// Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-first Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, December 3-6, 2007. Curran Associates, Inc. 2008: 1257-1264

[14] Yu K, Lafferty J, Zhu S, et al. Large-scale collaborative prediction using a nonparametric random effects model[C]// Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA, ACM, 2009: 1185-1192

[15] Rodríguez M A, Egenhofer M J. Determining Semantic Similarity among Entity Classes from Different Ontologies[J]. IEEE Trans. Knowl. Data Eng., 2003, 15(2): 442-456

[16] Yu Yuang-xin. Study of Information Retrieval Based on Semantic Similarity[J]. Journal of Intelligence, 2009, 28(9): 172-175

[17] Nagurney A, Qiang Qiang. A Knowledge Collaboration Network Model across Disciplines[C]// Advances in Social Computing, Third International Conference on Social Computing. Behavioral Modeling, and Prediction, SBP 2010, Bethesda, MD, USA, 2010: 138-148

(下转第 261 页)

在文献[18]中，还利用特定的参数对分割结果做出了一个综合评测，用百分制的形式进行表示，以便对分割效果进行快速的判断，涉及到的参数有体积重叠错误率(VOE)、体积差异(RVD)、平均面距离(ASD)、平方根面距离(RMS)、最大面距离(MSD)。表1是两组肝脏和肿瘤GMMAC分割结果与临床专家手工分割结果的比较。肝脏的分割误差比肿瘤的误差大，这是因为肝脏背景组织复杂，主动轮廓演化过程中容易溢出。整体上，自动分割结果还是可以接受的。

结束语 本文给出了基于混合高斯分布的区域竞争主动轮廓模型。基于医学图像统计概率分布和水平集的三维目标分割模型，把能量函数表示为目标和背景的子类属于该类高斯概率的积分，在水平集框架下使能量函数最小化。附加的速度约束项使得主动轮廓越过目标边缘时速度降低，从而提高分割的准确性。把该模型应用于肝脏的建模，包括肝脏、血管和肿瘤的分割。通过与GAC模型、C-V模型、GMM分类器以及手工分割的试验比较，表明GMMAC模型是一种快速、收敛、准确的分割模型。由于先验混合高斯分布的控制，因此该模型可以灵活机动地从复杂背景下提取出多重目标。

参考文献

[1] Caselles V, Kimmel R, Sapiro G. Geodesic active contours[J]. International Journal of Computer Vision, 1997, 22: 61-79

[2] Malladi R, Sethian J A, Vemuri B C. Shape Modeling with Front Propagation - a Level Set Approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17: 158-175

[3] Salah M B, Ayed I B, Mitiche A. Active Curve Recovery of Region Boundary Patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34: 834-49

[4] Chan T F, Vese L A. Active contours without edges[J]. IEEE Transaction Image Processing, 2001, 10(2): 266-277

[5] Zhu Song-cun, Yuille A. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18: 884-900

[6] Precioso F, Barlaud M, Blu T, et al. Robust real-time segmentation of images and videos using a smooth-spline snake-based algorithm[J]. IEEE Transactions on Image Processing, 2005, 14: 910-924

[7] Mukherjee D P, Ray N, Acton S T. Level set analysis for leukocyte detection and tracking[J]. IEEE Transactions on Image Processing, 2004, 13: 562-572

[8] Cremers D, Tischhauser F, Weickert J, et al. Diffusion snakes: Introducing statistical shape knowledge into the Mumford-Shah functional[J]. International Journal of Computer Vision, 2002, 50: 295-313

[9] Chen Y M, Tagare H D, Thiruvenkadam S, et al. Using prior shapes in geometric active contours in a variational framework [J]. International Journal of Computer Vision, 2002, 50: 315-328

[10] Shang Yan-feng, Yang Xin, Zhu Ming, et al. Prior based cardiac valve segmentation in echocardiographic sequences: geodesic active contour guided by region and shape prior[C]// Second Iberian Conference, IbPRIA 2005. 2005: 447-54

[11] Suri J S, Liu K C, Singh S, et al. Shape recovery algorithms using level sets in 2-D/3-D medical imagery: A state-of-the-art review [J]. IEEE Transactions on Information Technology in Biomedicine, 2002, 6: 8-28

[12] Heimann T, Meinzer H P. Statistical shape models for 3D medical image segmentation: A review[J]. Medical Image Analysis, 2009, 13: 543-563

[13] Cremers D, Rousson M, Deriche R. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape[J]. International Journal of Computer Vision, 2007, 72: 195-215

[14] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, Series B, 1977, 39(1): 1-38

[15] 李新仕，王天江，刘芳. 基于高斯混合模型的视频运动对象自动分割算法[J]. 计算机科学，2009，36(1)：205-207

[16] Soler L, Delingette H, Malandain G, et al. Fully Automatic Anatomical, Pathological, and Functional Segmentation from CT Scans for Hepatic Surgery[J]. Comp. Aid. Surg., 2001, 6(3): 131-142

[17] Markova A, Deklerck R, Nyssen E, et al. Comparison of the trust-region and the expectation-maximization algorithm for the application of automatic liver segmentation[C]// IEEE Benelux EMBS Symposium. Brussels, Belgium, 2006

[18] Heimann T, van Ginneken B, Styner M A, et al. Comparison and evaluation of methods for liver segmentation from CT datasets [J]. IEEE Trans Med Imaging, 2009, 28(8)

(上接第239页)

[18] Beckmann M J. On knowledge networks in science: collaboration among equals[J]. The Annals of Regional Science, 1994(28): 233-242

[19] Beckmann M J. Economic models of knowledge networks[M]// Bat- Ten D, Casti J, Thord R, eds. Networks in Action, Springer-Verlag, Berlin, Germany, 1995: 159-174

[20] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Trans. Inf. Syst, 2004, 22(1): 5-53

[21] Cleverdon C W. Introduction[J]. Information Storage and Retrieval, 1968, 4(2): 85

[22] Geisser S. Predictive Inference: an Introduction[M]. CRC Press, 1993

[23] Rafols I, Meyer M. Diversity and Network Coherenceas Indicators of Interdisciplinarity[J]. Case Studyies in bionanoscience, Scientometrics, 2010, 82(2): 263-287

[24] Lin D. An Information-Theoretic Definition of Similarity[C]// Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998). Madison, Wisconsin, USA, 1998: 296-304

[25] Budanitsky A, Hirst G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness [J]. Computational Linguistics, 2006, 32(1): 13-47

[26] Tversky A. Features of similarity [J]. Psychological Review, 1977, 84(2): 327-352