基于有监督主题模型的排序学习算法

丁宇新[1],燕泽权[1],冯 威[1],薛成龙[1],周 迪[2]
(1.哈尔滨工业大学深圳研究生院,广东深圳 518055;2.计算机体系结构国家重点实验室,中科院计算所,北京 100190)

摘 要: 文档表示是排序学习的关键,目前的排序学习算法多采用词袋法表示文档与查询,该方法假设词袋中的词相互独立,忽略了词之间的关系.为了表示文档中词之间的依赖关系,本研究利用文档与查询的主题特征构建排序学习模型,我们将排序函数定义为文档与查询之间的主题关系,提出了基于有监督主题模型的排序学习算法自动学习排序函数.为了评价模型的排序精度,我们在三个标准数据集(OHSUMED,MQ2007,MQ2008)上进行了实验.实验表明基于主题的排序学习算法能够发现文档与查询之间内在的语义关联,并改善排序模型的排序精度.

关键词: 排序学习;机器学习;关系主题模型;主题特征
中图分类号: TP309.2 **文献标识码:** A **文章编号:** 0372-2112 (2015)02-0333-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.02.019

Rank Learning Based on Supervised Topic Model

DING Yu-xin[1], YAN Ze-quan[1], FENG Wei[1], XUE Cheng-long[1], ZHOU Di[2]
(1. *Harbin Institute of Technology Shenzhen Graduate School*, *Shenzhen*, *Guangdong* 518055, *China*;
2. *State Key Laboratory of Computer Architecture*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: One of the key issues in learning to rank is document representation. In most of the learning to rank algorithms documents and queries are represented as a "bag of words", and words are assumed to occur independently. This kind of document representation ignores relationships between different words. To capture the important relationships between words, we try to learn a ranking model using the topic features of documents and queries. We define the ranking function as the topic relations between a document and a query. A novel rank learning algorithm based on supervised topic model is proposed to learn the ranking function. To evaluate the ranking accuracy of the proposed ranking algorithm, experiments are made on three benchmark datasets for information retrieval, OHSUMED, MQ2007, and MQ2008. The experimental results show that the proposed model can find the semantic relation between a document and a query, and can improve the ranking accuracy.

Key words: rank learning; machine learning; relational topic model; topic feature

1 引言

文档表示是排序学习的关键,目前的排序学习算法多采用词袋法表示文档与查询,该方法假设词袋中的词相互独立,忽略了词之间的关系.为了表示文档中词之间的依赖关系,研究者们提出了主题模型[1,2].主题模型已广泛应用于文本分析[3~5]领域.主题模型通过潜在的主题建立词之间的潜在语义关联.传统的主题模型主要有 LSA[1]、LDA[2],这些主题模型是非监督模型.Blei 等人[6]提出了可用于预测的有监督主题模型,之后 MedLDA[7]等有监督主题模型也相继提出.上述模型都是基于单篇文档,不能抽取文档间的关系.为了解决这一问题,论文[8]提出了 Mixed Membership Models,该模型的不足是参数的规模会随着文档数量的增加而增加,且无法应用于语料库之外的文档.Chang 等人[5]提出了 relational topic model(RTM),克服了 Mixed Membership Models 的缺点,不仅保证参数规模不随文档数量的增加而增加,而且能够对未见数据进行分类与预测.

主题模型在文档排序领域也得到了应用.LSI[1]是基于 LSA 的文档排序模型,它将查询与文档投影到潜在的主题空间,在主题空间计算查询与文档的主题距离.Wang 等人[9]提出了 RLSI 排序模型,并对比了不同主题排序模型的性能,包括 LSI、PLSI、LDA.上述主题模型排序算法都是无监督算法.本文尝试利用文档与查询之间的主题关系构建排序函数,以有监督主题模型 RTM 为基础,构建基于主题的排序学习模型.

收稿日期:2013-10-14;修回日期:2014-05-19;责任编辑:马兰英
基金项目:国家自然科学基金(No.61100192);中科院计算机体系结构国家重点实验室开放基金;哈工大学科研创新基金(No.HIT.NSRIF2010123);哈工大深圳研究生院网络智能计算重点实验室资助

2 关系主题模型(RTM)

我们的工作是对 RTM[5]的扩展,为便于理解,对该模型作简要描述.RTM 是一个生成模型,可以通过贝叶斯网络来描述.RTM 生成文档 d_i 的步骤如下:(1)设定文档主题数目 K,选择文档的长度 N(单词个数),N 服从泊松分布;(2)选择 $\boldsymbol{\theta}_i$,$\boldsymbol{\theta}_i$ 服从 Dirichlet($\boldsymbol{\alpha}$)分布,$\boldsymbol{\theta}_i$ 表示文档 d_i 在各个主题上的概率分布,$\boldsymbol{\alpha}$ 是 Dirichlet 分布参数;(3)N 个单词的每一个词 $w_{i(n)}$(文档 d_i 的第 n 个词)的产生过程如下:(a)选择主题 $z_{i(n)}$(文档 d_i 的第 n 个词的主题),$z_{i(n)}$服从 Multinomial($\boldsymbol{\theta}_i$)多项式分布;(b)根据 $p(w_{i(n)}|\ z_{i(n)};\boldsymbol{\beta}_{z_i^{(n)}})$ 选择 $w_{i(n)}$.

图 1 给出了 RTM 的结构图.图的左侧方框表示任意文档 d_i,右侧方框表示标准文档 d_t,响应变量 y_i 表示文档 d_i 和 d_t 的类别,通常以概率方式给出.RTM 由四个参数描述,即($\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},v$),$\beta$ 是一个 $K\times V$ 矩阵(V 表示语料库的词汇量),其行向量表示主题在词汇上的概率分布,$\boldsymbol{\beta}_{z_i^{(n)}}$为 $\boldsymbol{\beta}$ 的行向量,代表主题 $z_{i(n)}$在词汇上的概率分布.$\boldsymbol{\eta}$ 和 v 则是描述 y_i 的概率分布参数.公式(1)定义了 RTM 模型产生文档集合 $\mathbb{D}$ 及类别标签集合 $\boldsymbol{Y}$ 的条件概率,$\boldsymbol{Z}$ 是由 D 中文档(M 个文档)的词主题向量 $\boldsymbol{z}_i$ 和标准文档 d_t 的词主题向量 $\boldsymbol{z}_t$ 组成的$(M+1)\times V$ 矩阵.$\boldsymbol{\theta}$ 是由 M 个文档的主题向量 $\boldsymbol{\theta}_i$ 和 d_t 的主题向量 $\boldsymbol{\theta}_t$ 组成的$(M+1)\times K$ 矩阵.$\boldsymbol{Y}$ 是 M 维向量,每一维 y_i 代表文档 d_i 与 d_t 为相同类别的概率.$\boldsymbol{\theta}$ 与 $\boldsymbol{Z}$ 为隐变量.

$$p(\mathbb{D},\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},v)=\int_{\boldsymbol{\theta}}\sum_{Z}p(\boldsymbol{\theta},\boldsymbol{Z},\mathbb{D},\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},v)\mathrm{d}\boldsymbol{\theta} \tag{1}$$

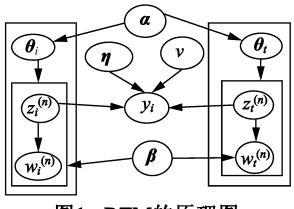

图1 RTM的原理图

3 基于主题的排序学习模型

3.1 问题描述

基于主题的排序学习可描述为:已知训练数据集 $T=\{Q,D,\boldsymbol{Y}\}$,D 是由 M 个文档组成的文档集合,即 $D=\{d_1,d_2\cdots d_M\}$;Q 代表查询;$\boldsymbol{Y}$ 是 D 中各文档的相关度标识集合,其表示为 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_M\}$,其中 y_i 代表文档 d_i 与查询 Q 的相关度.文档 d_i 包含 N_i 个词,可表示为 $d_i=(w_i^{(1)},w_i^{(2)}\cdots\ w_i^{(Ni)})$.我们的目标是学习一个排序函数 f,该函数利用文档与查询之间的主题相关性排序文档.

3.2 基于主题的排序函数设计

针对排序问题,我们将排序学习中的查询 Q 看作 RTM 中的标准文档 d_t,并按同样的过程产生查询 Q.响应变量 y_i 代表文档 d_i 与查询 Q 的相关性.本文用相同符号表示查询相关参数与文档相关参数,只是下标不同,与查询有关的参数下标为 t,例如:$w_{t(n)}$表示查询 Q 的第 n 个词.RTM 模型中文档 d_i 与 d_t 之间的主题关系由函数$f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)$定义,该函数值决定文档 d_i 与 d_t 属于相同类别 y_i 的概率,$f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)$代表了两文档之间的主题相似性.在排序学习模型中,我们将该函数转换为文档与查询 Q 的主题关系函数,该函数也就是要学习的排序函数.排序函数 $f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)$的定义见公式(2).其中,$\eta$ 表示每个主题的权重,符号$\circ$表示 Hadamard 积[10],其意义是文档与查询在每一维上的相似度,$\bar{\boldsymbol{z}}_i$ 为文档 d_i 的均值主题指示向量,$\bar{\boldsymbol{z}}_t$ 为查询 Q 的均值主题指示向量.$\bar{\boldsymbol{z}}_i$ 与 $\bar{\boldsymbol{z}}_t$ 的定义见公式(3)与(4),其中 $\boldsymbol{z}_{i,n}$是文档 d_i 关于词 $w_{i(n)}$的指示向量,若第 j 个主题中包含 $w_{i(n)}$,则 $\boldsymbol{z}_{i,n}$的第 j 维为 1,否则为 0;同理 $\boldsymbol{z}_{t,n}$是查询 Q 关于词 $w_{t(n)}$的指示向量.$\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t$ 的定义见公式(5),$\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t$ 的每一分量大于、等于 0 且小于 1.v 是$f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)$的截距.

$$f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)=\boldsymbol{\eta}(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)+\nu \tag{2}$$

$$\bar{\boldsymbol{z}}_i=(1/N_i)\sum_{n=1}^{N_i}\boldsymbol{z}_{i,n} \tag{3}$$

$$\bar{\boldsymbol{z}}_t=(1/N_t)\sum_{n=1}^{N_t}\boldsymbol{z}_{t,n} \tag{4}$$

$$\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t=(1/(N_iN_t))\sum_{n=1,m=1}^{(N_i,N_t)}\boldsymbol{z}_{i,n}\circ\boldsymbol{z}_{t,m} \tag{5}$$

$$p(y_i|\boldsymbol{\eta},\nu,\bar{\boldsymbol{z}}_i,\bar{\boldsymbol{z}}_t)=\exp(f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)) \tag{6}$$

3.3 二分类模排序型及回归 RTM 排序模型

RTM 是一个单分类模型,训练时只考虑了正例($y_i=1$),公式(6)表示 d_i 与 d_t 属于同一类别的概率.排序学习是一个多分类或回归问题,y_i 可取不同的值.针对排序学习问题,我们将 RTM 扩展为二分类模型.在二分类模型中,训练样本由正例与反例组成,正例为查询相关文档,反例为查询不相关文档.针对二分类模型,我们重新定义公式(6),见公式(7),(8).公式(7)定义了当训练样本为正例时,文档与查询的相关概率,公式(8)定义了当训练样本为反例时,文档与查询的非相关概率.在训练过程中,对于正例,令公式(7)趋向 1(最大概率),即 $f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)$趋向于 0;对于反例,令公式(8)趋向 1,在实际中可令公式(7),(8)趋向于大于 0.5 小于 1 的某个值 c(本文实验设定 c 为 0.9).

$$p(y_i=1|\eta,v,\bar{\boldsymbol{z}}_i,\bar{\boldsymbol{z}}_t)=\exp(f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)) \tag{7}$$

$$p(y_i=-1|\eta,v,\bar{\boldsymbol{z}}_i,\bar{\boldsymbol{z}}_t)=1-\exp(f(\bar{\boldsymbol{z}}_i\circ\bar{\boldsymbol{z}}_t)) \tag{8}$$

对于排序学习问题,文档与查询之间的相关度类别标签一般多于两种,为解决这一问题,我们将 RTM 扩展为回归模型.在回归 RTM 模型中,y_i 是一个绝对数

值,它代表文档的排序值.我们将公式(6)定义为公式(9).在训练过程中,同样让公式(9)趋向 1(最大概率).

$$p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t) = \exp(f(\bar{z}_i \circ \bar{z}_t) - y_i) \tag{9}$$

3.4 二分类及回归 RTM 模型参数估计

参数估计与 RTM 模型[5]的参数估计相似.本节主要针对当涉及概率分布 $p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$时参数估计公式的推导.

3.4.1 后验推导

RTM 的参数估计算法就是使 $p(\mathcal{D}, \boldsymbol{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, v)$(公式(1))的概率最大化,由 Jensen 不等式,公式(1)的下界如公式(10)所示.文献[5]采用 EM 算法使公式(10)的下界最大化,以不断接近 $\log p(\mathcal{D}, \boldsymbol{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, v)$的真实值.

$$\begin{aligned}&\log p(\mathcal{D}, \boldsymbol{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, v)\\ &\geq E_{q(\theta, Z)}[\log p(\boldsymbol{\theta} \mid \alpha) p(\boldsymbol{Z} \mid \boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{Z}, \boldsymbol{\beta})]\\ &\quad - E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log q(\boldsymbol{\theta}, \boldsymbol{Z})] + E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\eta}, v)]\end{aligned} \tag{10}$$

公式(10)中,参数 $\boldsymbol{\theta}$ 与 $\boldsymbol{Z}$ 的均值场变分分布如公式(11)所示.$\boldsymbol{\gamma}$ 由 γ_i 组成,γ_i 表示文档 d_i 的 Dirichlet 分布参数向量;$\boldsymbol{\varphi}$ 由 $\boldsymbol{\phi}_i$ 组成,$\boldsymbol{\phi}_i$ 由 $\boldsymbol{\phi}_{i,n}$组成,$\boldsymbol{\phi}_{i,n}$表示文档 d_i 的词 $w_{i(n)}$在 K 个主题上的多项式分布参数向量,该词在第 k 个主题上的分布用 $\phi_{i,n}^{(k)}$表示.为方便讨论,查询对应的参数下标 $t=0$.公式(10)的下界可展开为公式(12),其中 γ_i 与 $\varphi_{i,n}$是要估计的参数.

$$q(\boldsymbol{\theta}, \boldsymbol{Z} \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod_{i=0}^{M} q(\boldsymbol{\theta}_i \mid \boldsymbol{\gamma}_i) \prod_{i=0}^{M} \prod_{n=1}^{N_i} q(z_i^{(n)} \mid \boldsymbol{\phi}_{i,n}) \tag{11}$$

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\phi}_{1:M})\\ &= \sum_{i=0}^{M} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(\boldsymbol{\theta}_i \mid \boldsymbol{\alpha})]\\ &\quad + \sum_{i=0}^{M} \sum_{n=1}^{N_i} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(z_i^{(n)} \mid \boldsymbol{\theta}_i)]\\ &\quad + \sum_{i=0}^{M} \sum_{n=1}^{N_i} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(w_i^{(n)} \mid z_i, \beta)]\\ &\quad - E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log q(\boldsymbol{\theta}, \boldsymbol{Z})]\\ &\quad + \sum_{i=1}^{M} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(y_i \mid z_i, z_t, \boldsymbol{\eta}, v)]\end{aligned} \tag{12}$$

隐变量 $\boldsymbol{\gamma}_i$ 与概率分布 $p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$无关,可直接依文献[5]给出的公式进行计算,见公式(13).参数 $\boldsymbol{\varphi}_{i,n}$的估计较繁琐.从公式(12)抽出所有包含 $\boldsymbol{\varphi}_{i,n}$的项,对 $\boldsymbol{\varphi}_{i,n}$求偏导,并令其等于 0,可得 $\boldsymbol{\varphi}_{i,n}$的更新公式,见公式(14).其中 $\boldsymbol{\beta}_{*, w_{i,n}}$是 K 维列向量,表示词 $w_{i(n)}$在 K 个主题的概率分布.

$$\boldsymbol{\gamma}_i = \boldsymbol{\alpha} + \sum_{n=1}^{N_i} \boldsymbol{\varphi}_{i,n} \tag{13}$$

$$\begin{aligned}&\boldsymbol{\varphi}_{i,n}^{new} \propto\\ &\exp\left(\begin{array}{l}\Psi(\boldsymbol{\gamma}_i) - \Psi(\sum_{m=1}^{K} \boldsymbol{\gamma}_{i,m}) + \log \boldsymbol{\beta}_{*, w_{i,n}}\\ + \left(\sum_{i=1}^{M} \sum_{n=1}^{N_i} \nabla_{\boldsymbol{\varphi}_{i,n}} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)]\right)\end{array}\right)\end{aligned} \tag{14}$$

公式(14)依赖于 $p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$, $\nabla_{\boldsymbol{\varphi}_{i,n}} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\log p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)]$的推导如下:

对于二分类模型,$p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$的定义见公式(7),(8).对于正例,$\log p(y_i = 1 \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$的期望可表示为公式(15).其中 $E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\bar{z}_i \circ \bar{z}_{t_t}]$可展开为公式(16),其中 $\bar{\boldsymbol{\varphi}}_i = (1/N_i) \sum_{n=1}^{N_i} \boldsymbol{\varphi}_{i,n}$, $\bar{\boldsymbol{\varphi}}_t = (1/N_t) \sum_{n=1}^{N_t} \varphi_{t,n}$.$\mathcal{L}(y_i = 1)$(公式(15))关于 $\boldsymbol{\varphi}_{i,n}$的偏导为公式(17).

$$\mathcal{L}(y_i = 1) = \boldsymbol{\eta}^{\mathrm{T}} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\bar{z}_i \circ \bar{z}_t] + v \tag{15}$$

$$E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\bar{z}_i \circ \bar{z}_t] = \bar{\varphi}_i \circ \bar{\boldsymbol{\varphi}}_t \tag{16}$$

$$\nabla_{\varphi_{i,n}} \mathcal{L}(y_i = 1) = \boldsymbol{\eta} \circ \frac{\bar{\boldsymbol{\varphi}}_t}{N_i} \tag{17}$$

二分类模型中,对于反例,$p(y_i = -1 \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$等于公式(8),$\log p(y_i = -1 \mid \eta, v, \bar{z}_i, \bar{z}_t)$的期望可表示为公式(18),$\mathcal{L}(y_i = -1)$关于 $\boldsymbol{\varphi}_{i,n}$的偏导可表示为公式(19).

$$\mathcal{L}(y_i = -1) = \log(1 - \exp(\boldsymbol{\eta} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\bar{z}_i \circ \bar{z}_t] + v)) \tag{18}$$

$$\nabla_{\boldsymbol{\varphi}_{i,n}} \mathcal{L}(y_i = -1) = \frac{-\exp(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) + v)}{1 - \exp(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) + v)} \boldsymbol{\eta} \circ \frac{\bar{\boldsymbol{\varphi}}_t}{N_i} \tag{19}$$

同理,对于回归模型,$p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$的定义见公式(9),$\log p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$的期望可表示为公式(20).公式(20)对 $\boldsymbol{\varphi}_{i,n}$求偏导,可得公式(21).

$$\mathcal{L}(y_i) = (\boldsymbol{\eta} E_{q(\boldsymbol{\theta}, \boldsymbol{Z})}[\bar{z}_i \circ \bar{z}_t] + v) - y_i \tag{20}$$

$$\nabla_{\varphi_{i,n}} \mathcal{L}(y_i) = \boldsymbol{\eta} \circ \frac{\bar{\boldsymbol{\varphi}}_t}{N_i} \tag{21}$$

3.4.2 参数估计

该步骤估计 LDA 的两个模型参数 $\boldsymbol{\alpha}$、$\boldsymbol{\beta}$,以及排序函数的参数 $\boldsymbol{\eta}$、v.参数 $\boldsymbol{\alpha}$ 与 $\boldsymbol{\beta}$ 的估计与传统的 LDA 模型相同,可参考文献[5],本文不作赘述.我们主要讨论如何估计 $\boldsymbol{\eta}$、v.对于二分类模型,当 $p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$等于公式(7)时,公式(12)对 $\boldsymbol{\eta}$ 的偏导见公式(22),其对 v 的偏导见公式(23).当 $p(y_i \mid \boldsymbol{\eta}, v, \bar{z}_i, \bar{z}_t)$等于公式(8)时,公式(12)对 $\boldsymbol{\eta}$ 的偏导见公式(24),其对 v 的偏导见公式(25).

$$\nabla_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\varphi}_{1:M}) = \sum_{i=1}^{M} \bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t \tag{22}$$

$$\nabla_{v} \mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\varphi}_{1:M}) = M \tag{23}$$

$$\nabla_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\varphi}_{1:M}) = \sum_{i=1}^{M} \frac{-\exp(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) + v)}{1 - \exp(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) + v)} (\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) \tag{24}$$

$$\nabla_v \mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\varphi}_{1:M}) = \sum_{i=1}^{M} \frac{-\exp(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) + v)}{1 - \exp(\boldsymbol{\eta}^{\mathrm{T}}(\bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t) + v)} \tag{25}$$

同理,对于回归模型,可采用相同的方式分别计算公式(12)对 $\boldsymbol{\eta}$ 与 v 的导数,各偏导公式如公式(26)、(27)所示.

$$\nabla_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\varphi}_{1:M}) = \sum_{i=1}^{M} \bar{\boldsymbol{\varphi}}_i \circ \bar{\boldsymbol{\varphi}}_t \tag{26}$$

$$\nabla_v \mathcal{L}(\boldsymbol{\gamma}_{1:M}, \boldsymbol{\varphi}_{1:M}) = M \tag{27}$$

4 实验

我们采用 OHSUMED[11]、MQ2007[12]、MQ2008[12]三个评测集测试排序算法的性能.文献[11,12]给出了上述数据集的特征,从中选取 25 个基于词的特征,与链接相关的特征没有选取,如 PageRank、HITS、HostRank 等.

为了评价算法的排序性能,我们将其与 RankSVM[13]、BM25、LDA[2]、原始 RTM[5]等模型进行了对比.RankSVM 与 BM25 为词特征排序模型.LDA 为非监督主题模型,我们用文档与查询的主题向量相似性排序文档,主题向量的相似性定义为向量的余弦相似性.实验中的三个数据集合各被划分为五个子集[11,12].本文采用五重交叉验证验评估模型的排序性能.模型的排序性能用 NDCG 值来评价[14].

对于单分类及二分类模型,训练数据中的相关文档及部分相关文档作为正例.对于回归模型,不相关文档的排序值设为 0,部分相关文档的排序值为 1,相关文档的排序值为 2.对于主题模型,主题数目选为 40.

表 1、2、3 给出了各排序模型在不同实验数据集上的排序结果,表中结果为各模型的 NDCG@1 至 NDCG@10 排序值(表中标识为 N@1 至 N@10).各表中,TRTM 代表二分类 RTM 模型,RRTM 代表回归 RTM 模型,RankSVM-2 表示由两类训练数据训练的 RankSVM 模型,RankSVM-3 表示由三类训练数据训练的 RankSVM 模型.表中最后一列是各行的平均值.

表 1 OHSUMED 数据集排序精度

	N@1	N@2	N@3	N@4	N@5	N@6	N@7	N@8	N@9	N@10	Avg.
TRTM	0.51	0.46	0.45	0.44	0.44	0.43	0.45	0.43	0.43	0.42	0.446
RRTM	0.53	0.49	0.48	0.47	0.47	0.46	0.46	0.44	0.43	0.44	0.467
RTM	0.48	0.43	0.42	0.42	0.4	0.41	0.41	0.4	0.41	0.38	0.416
LDA	0.4	0.38	0.39	0.36	0.37	0.36	0.34	0.35	0.35	0.35	0.365
RankSVM-2	0.49	0.43	0.43	0.43	0.42	0.42	0.41	0.41	0.41	0.41	0.426
BM25	0.43	0.4	0.41	0.41	0.41	0.39	0.39	0.4	0.39	0.4	0.403
RankSVM-3	0.51	0.49	0.47	0.45	0.45	0.44	0.45	0.44	0.44	0.43	0.457

表 2 MQ2007 数据集排序精度

	N@1	N@2	N@3	N@4	N@5	N@6	N@7	N@8	N@9	N@10	Avg
TRTM	0.34	0.37	0.38	0.38	0.38	0.35	0.36	0.37	0.35	0.33	0.361
LDA	0.13	0.12	0.15	0.16	0.17	0.16	0.14	0.16	0.15	0.14	0.148
BM25	0.14	0.16	0.2	0.22	0.21	0.21	0.19	0.18	0.2	0.18	0.189
RankSVM-2	0.33	0.34	0.35	0.36	0.37	0.36	0.35	0.34	0.33	0.32	0.345
RTM	0.34	0.32	0.34	0.32	0.34	0.33	0.33	0.34	0.31	0.29	0.326
RankSVM-3	0.36	0.4	0.4	0.39	0.41	0.38	0.39	0.4	0.41	0.35	0.389
RRTM	0.38	0.41	0.42	0.4	0.43	0.4	0.39	0.41	0.41	0.38	0.403

表 3 MQ2008 数据集排序精度

	N@1	N@2	N@3	N@4	N@5	N@6	N@7	N@8	N@9	N@10	Avg.
TRTM	0.3	0.3	0.32	0.33	0.32	0.34	0.3	0.34	0.16	0.17	0.288
LDA	0.17	0.16	0.15	0.15	0.14	0.15	0.13	0.13	0.12	0.13	0.143
BM25	0.21	0.19	0.18	0.17	0.16	0.17	0.16	0.17	0.13	0.13	0.167
RankSVM-2	0.3	0.3	0.31	0.34	0.32	0.35	0.35	0.33	0.15	0.16	0.291
RTM	0.27	0.28	0.28	0.31	0.32	0.3	0.31	0.3	0.14	0.15	0.266
RankSVM-3	0.31	0.32	0.35	0.37	0.39	0.41	0.42	0.37	0.15	0.16	0.325
RRTM	0.33	0.33	0.34	0.38	0.4	0.41	0.42	0.39	0.18	0.18	0.336

实验总结如下:(1)主题模型与词模型对比:实验选取的词排序模型为 RankSVM 及 BM25 模型.由表 1、2、3 可看出,二分类主题模型 TRTM 的排序精度优于 RankSVM-2,回归主题模型 RRTM 的排序精度优于 RankSVM-3.非监督 BM25 模型的排序性能最差.

与基于主题的排序模型相比,基于词的排序模型假设词之间是相互独立的,忽略了文档内词之间的关系,例如不同的词可表示同样的意义,这就导致基于词的排序模型不能较准确地表示文档内容.在基于 RTM 排序模型中,主题被定义为一组词的概率分布,一个主

题可用不同的词去描述,一个文档用不同的主题描述,因此,主题模型能够通过潜在的主题描述词之间以及文档之间的关系,较词袋模型具有更强的表达能力.因此,基于主题的排序模型有较好的排序性能.(2)各主题模型对比:三种基于 RTM 的主题模型中,回归 RTM 模型的性能优于二分类 RTM 模型,二分类 RTM 模型的性能优于单分类 RTM 模型,无监督 LDA 模型的排序性能最差.

在原始 RTM 模型中,训练样本只有正例,因此,模型中正例、反例的分类边界较为模糊,此外正例包含相关及部分相关两类样本,导致模型不能准确区分相关与部分相关;在二分类 RTM 模型中,训练样本被标记为两类,其中相关与部分相关文档同样被标记为一类,因此,二分类模型也不能准确区分相关与部分相关文档;对于回归 RTM 模型,训练样本被标记为三类,因此,回归 RTM 模型能够较好区分查询相关文档及部分相关文档.这表明在有监督模型中,提供准确的训练样本相关度信息可改善模型的排序性能.无监督 LDA 排序模型依据文档与查询的余弦距离排序文档,由于这一距离是人为定义,不是经训练样本学习获得,因此,与实际存在一定误差,这也是非监督 LDA 排序模型性能较差的原因.

5　结论

本文尝试利用文档与查询的主题关系特征构建排序学习模型,该排序模型在一定程度上能够发现与查询语义相关的文档.实验结果表明扩展后的主题模型较原模型在排序精度上有较大改善.

参考文献

[1] Deerwester S, Dumais SS T, Furnas G W, Landauer T K, Harshman R. Indexing by latent semantic analysis[J]. JAM SOC INFORM SCI, 1990, 41(6): 391 – 407.

[2] Blei D, Ng A, Jordan M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993 – 1022.

[3] 王李冬,等.基于概率主题模型的文档聚类[J].电子学报,2012,40(11):2346 – 2350.
Wang Lidong, et al. Document clustering based on probabilistic topic model[J]. Acta Electronica Sinica, 2012, 40(11): 2346 – 2350. (in Chinese)

[4] 吴永辉,等.基于主题的自适应、在线网络热点发现方法及新闻推荐系统[J].电子学报,2010,38(11):2620 – 2634.
Wu yonghui, et al. Adaptive on-line web topic detection method for web news recommendation system[J]. Acta Electronica Sinica, 2010, 38(11): 2620 – 2634. (in Chinese)

[5] Chang J, Blei D. Hierarchical relational models for document networks[J]. The Annals of Applied Statistics, 2010, 4(1): 124 – 150.

[6] Blei D, McAuliffe J. Supervised topic models[A]. Neural Inforamtion Processing System Conference[C]. Vancouver, Canada: MIT Press, 2007. 1 – 8.

[7] Zhu J, Ahmed A, Xing E P. MedLDA: maximum margin supervised topic models for regression and classification[A]. International Conference on Machine Learning[C]. Montreal, Canada: ACM, 2009. 158 – 1264.

[8] Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications[J]. National Academy of Sciences, 2004, 97(22): 11885 – 11892.

[9] Wang Q, et al. Regularized latent semantic indexing[A]. SIGIR'2011[C]. Beijing, China: ACM, 2011. 685 – 694.

[10] Nallapati R, Cohen W. Link-PLSA-LDA: a new unsupervised model for topics and influence of blogs[A]. Proceedings of Association for the Advancement of Artificial Intelligence[C]. Chicago, USA: AAAI Press, 2008. 84 – 92.

[11] Xu J, et al. LETOR: Benchmark letor dataset for research on learning to rank for information retrieval[A]. Proceding of SIGIR Workshop on Learning to Rank for Information Retrieval[C]. Amsterdam, Holland: ACM, 2007. 201 – 206.

[12] Qin T, Liu T, Xu J, Li H. LETOR: Benchmark letor dataset for research on learning to rank for information retrieval[J]. Information Retrieval, 2010, 13(4): 346 – 374.

[13] Joachims T. Training Linear SVMs in Linear Time[A]. the ACM Conference on Knowledge Discovery and Data Mining[C]. Philadelphia, USA: ACM, 2006. 217 – 226.

[14] Jarvelin K, Kekalainen, J. IR evaluation methods for retrieving highly relevant documents[A]. SIGIR'2000[C]. Athens, Greece: ACM, 2000. 41 – 48.

作者简介

丁宇新　男,1972 出生,天津人,博士,哈尔滨工业大学深圳研究生院副教授,研究方向为自然语言处理,机器学习.
E-mail: yxding@hitsz.edu.cn

燕泽权　男,1989 出生,河北唐山人,哈尔滨工业大学深圳研究生院硕士研究生,研究方向为自然语言处理,机器学习.
E-mail: saiboyan@163.com