基于多语义关系的个性化查询扩展方法*

伍 璇 周 栋

(湖南科技大学 计算机科学与工程学院 湘潭 411201)

摘 要 随着数字内容不断增长,信息检索技术已经不能满足不同用户对高精度信息内容获取的需求. 文中提出基于多语义关系的个性化查询扩展方法,并应用于基于社会化标签的个性化搜索系统. 模型使用标签-主题模型对用户兴趣模型进行建模,能够更有效地表达语义和提升搜索效果. 在此基础上,进一步提出基于多语义关系的个性化查询扩展方法,利用社会化标签的多重语义特征进行扩展词的选择. 在大规模真实社会化标签数据集上的实验表明,文中方法优于非个性化搜索及其它基于社会化标签系统的个性化查询扩展方法.

关键词 主题模型, 社会化标签, 多语义关系, 查询扩展

中图法分类号 TP 391 **DOI** 10.16451/j.cnki.issn1003-6059.201711009

引用格式 伍 璇,周 栋. 基于多语义关系的个性化查询扩展方法. 模式识别与人工智能, 2017, 30(11): 1039-1047.

Personalized Query Expansion Method Based on Multiple Semantic Relationships

WU Xuan, ZHOU Dong

(*School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan* 411201)

ABSTRACT

Due to the ever-increasing amount of digital contents in the internet, the traditional information retrieval technology is unable to meet the demands for high precision information of different users. In this paper, a personalized query expansion method based on multiple semantic relationships is proposed. It is used for personalized search based on social tagging systems. An tag-topic model is utilized to generate the user interesting model. Therefore, more precise semantics can be captured. The performance of the search can also be improved. Based on the user model, a personalized search method based on multiple semantic relationships from social data is further presented to select suitable expansion terms. Experiments conducted on a large social tagging dataset show that the proposed method outperforms

* 国家自然科学基金项目(No. 61300129)、湖南省自然科学基金项目(No. 2017JJ2101)、湖南省教育厅项目(No. 16K030)、湖南省研究生科研创新项目(No. CX2017B650)资助
Supported by National Natural Science Foundation of China (No. 61300129), Natural Science Foundation of Hunan Province (No. 2017JJ2101), Scientific Research Fund of Hunan Provincial Education Department (No. 16K030), Hunan Provincial Innovation Foundation for Postgraduate (No. CX2017B650).
收稿日期:2017-07-15;修回日期:2017-08-31;录用日期:2017-09-14
Manuscript received July 15, 2017; revised August 31, 2017; accepted September 14, 2017

several non-personalized methods as well as the existing personalized search methods based on social tagging systems.

Key Words Topic Model, Social Tagging, Multiple Semantic Relationship, Query Expansion

Citation WU X, ZHOU D. Personalized Query Expansion Method Based on Multiple Semantic Relationships. Pattern Recognition and Artificial Intelligence, 2017, 30(11): 1039-1047.

互联网上的在线数字内容呈现明显指数型增长趋势. 使用个性化 Web 搜索系统已成为搜索用户所需信息的重要方法[1-2]. 个性化搜索和 Web 搜索之间的主要区别在于前者通过获取用户的个人信息和抽取用户查询的上下文等方式构建一个用户兴趣模型,用于持续跟踪用户的信息需求,实现个性化搜索结果. 用户兴趣模型可以分为 2 类:一类模型以个人用户为主体[3-4],而另一类模型使用协作式的群体用户构建[5]. 本文的研究采取以个人用户为主体的方法.

随着数字内容的不断增长,社会化标签系统如 Del. icio. us(http://www. delicious. com)和 Flickr(http://www. flickr. com)等已大众化. 在 Del. icio. us 标签系统中,用户可以收藏并标注感兴趣的网页,使网页分类更清晰明了. 在 Flickr 标签系统中,用户可以标注对自己有价值的图像,使图像检索更简单. 被标记过的资源及标签本身反映标注用户的个性化偏好,为个性化检索提供丰富信息并产生积极效果. 另一方面,由于标注行为的随意性,用户标签的语义模糊给利用社会化标签系统进行个性化搜索带来一定的挑战. 为了解决这一问题,研究者们提出一系列的个性化结果重排序方法[6-9]和个性化查询扩展方法[10-12].

文献[6]、文献[7]、文献[13]和文献[14]利用社会化标签系统建立用户模型,设计关联性反馈框架,利用与用户兴趣相关的隐式信息进行 Web 结果的搜索和排序. Wang 等[15]通过挖掘用户在多个社会媒体上的公开活动进行个性化搜索. Cai 等[16]将查询相关关系和用户相关关系通过模糊满意度函数进行处理. 语言环境[8]和情感分析[17]也用于基于社会化标签系统的个性化搜索结果重排序. 结果重排序虽然有效,但如果不能在第一轮的搜索中抓取到相关的搜索结果,则无论怎样进行重排序,效果也差强人意. 因此,本文主要借助个性化查询扩展方法进行搜索.

个性化查询扩展方法在原查询词的基础上进行扩充,挑选用户兴趣模型中的词项,组成一个更长也能更准确捕捉用户真实查询语义的新查询词. 然而,目前基于社会化标签系统的个性化搜索普遍存在如下问题:1)以往的研究尝试使用标签之间关系[18-20]、词共现关系[10-11,21],以及基于主题模型的语义匹配[22]进行扩展词的选择. 由于标签与被标注资源本身可能在字面上相差甚远,也并非该资源的精确描述,因此借助标签进行查询扩展效率不高[21]. 而基于词共现的方法会忽略用户兴趣模型中的语义关系. 2)虽然基于语义匹配的方法可以改善检索效率,但这一提高有限. 特别是针对那些冷启动用户,即与搜索系统互动较少的用户,由于历史信息太少,对其进行个性化操作还需在少量信息中进一步挖掘更丰富的语义关系. Zhou 等[12]提出基于语义匹配的个性化查询扩展方法,将用户兴趣模型中的词项利用标签-主题模型构建成图,以便准确把握词项与词项间的关系,再借助正则化框架选择扩展词. 这一方法虽然也利用语义信息,但该信息还较浅显.

针对上述问题,本文利用标签-主题模型方法[12]对用户兴趣模型进行建模. 模型基于词进行隐含主题的学习. 在获得主题的基础上,进一步提出个性化查询扩展方法,利用社会标签的多语义关系进行扩展词的选择. 在本文方法中,扩展词的选择并不仅依赖于词项匹配,而是依赖于基于主题的相关度. 基于大规模真实社会化标签数据集上的实验表明,本文方法优于非个性化搜索方法及基于社会化标签系统的个性化查询扩展方法.

1 问题定义及用户兴趣模型构建

1.1 问题定义

标注是用户对 Web 数字内容使用关键词或自定义标签的方式进行标记和分类,是用于对数字内容进行组织、过滤和搜索的方式. 用户在为 Web 资源进行标签创建的同时,建立自身与资源之间的关系,体现用户的兴趣偏好. 社会化标注的数据可表示为 $P:=(U,D,T,A)$,其中,$A\subseteq U\times D\times T$ 为一个三元关系集,元素为标签标注行为. 用户的所有标注行为表示为

$$A^u := \{(t,d) \mid u \in A, d \in A, t \in A\},$$

其中,t 表示标签,d 表示文档(即 Web 网页),u 表示用户. 用户的所有标签表示为

$$T^u := \{t \mid (t,d) \in A^u\}.$$

用户标注过的所有文档表示为

$$D^u := \{d \mid (t,d) \in A^u\}.$$

本文进一步定义从用户标注过的文档中抽取的所有词项集合为

$$term^{D^u} := \{w \mid w \in D^u\},$$

其中 w 表示词项.

在一个典型的个性化搜索情境下,给定某一查询词 q 及用户兴趣模型中的一组扩展候选词

$$\{w_1, w_2, \cdots, w_x, t_1, t_2, \cdots, t_y\} \in term^{D^u} \cup T^u.$$

本文的目的是从中选取一组按与原查询词相关程度进行排序的扩展词以扩展 q,便于进行第二轮搜索.

1.2　用户模型构建

在提取用户的标签 T^u 以及标注过的文档 D^u 之后,采用标签-主题模型,每篇文档可以从该模型中获得表示主题的概率分布,从而用于对用户的查询历史进行建模. 下面详细描述该模型.

标签主题模型认为每个标签有一个主题概率分布 θ,每个主题有一个词项概率分布 φ. 标签-主题模型不仅可以挖掘与主题相关联的标签("标签-主题" 分布 θ),而且可以发掘与主题关联的词项("主题-词项" 分布 φ),具体步骤如下.

算法　标签-主题模型

Input　tags of a user T^u

documents of a user D^u

for each tag $t \in T^u$

choose $\theta_t \sim Dirichlet(\rho)$

end for

for each topic $e \in E$,

choose $\varphi_e \sim Dirichlet(\upsilon)$

end for

for each document $d \in D^u$, given the vector of tags t_d

for each word w_i indexed by $i = 1,2,\cdots,N_{d_j}$

Conditional on t_d choose a tag

$x_i \sim Uniform(t_d)$

Conditional on x_i choose a topic

$z_i \sim Discrete(\theta_{x_i})$

Conditional on z_i choose a term

$w_i \sim Discrete(\varphi_{z_i})$

end for

end for

Output　θ, φ

根据"标签-主题" 分布和"主题-词项" 分布之间的矩阵计算,在每个标签和词项之间建立"标签-词项" 映射关系.

模型的生成过程是概率抽样过程:1)对于文档中的每个词项,先从标签集合中等概率抽取一个标签;2)根据此标签在主题上的多项概率分布,随机抽取一个主题;3)根据该主题在词项上的多项概率分布,随机抽取词项;4)抽取过程不断迭代,直至抽取完文档中的所有词项.

使用吉布斯抽样,针对每个词项,抽样主题和标签:

$$p(z_i = j, x_i = k \mid w_i = h, z_{\neg i}, x_{\neg i}) \propto \frac{C_{hj}^{WE} + \upsilon}{\sum_{h'} C_{h'j}^{WE} + L\upsilon} \cdot \frac{C_{kj}^{TE} + \rho}{\sum_{j'} C_{kj'}^{TE} + E\rho},$$

其中

$$\varphi_{hj} = \frac{C_{hj}^{WE} + \upsilon}{\sum_{h'} C_{h'j}^{WE} + L\upsilon}$$

表示在主题 j 下抽取一个词项 h 的概率;

$$\theta_{kj} = \frac{C_{kj}^{TE} + \rho}{\sum_{j'} C_{kj'}^{TE} + E\rho}$$

表示对应标签 k 下抽取主题 j 的概率;$z_i = j$ 和 $x_i = k$ 分别表示当前文档中主题 j 下和标签 k 下的第 i 个词项;$w_i = h$ 表示文本语料库中的第 i 个词项为单词 h;$z_{\neg i}$ 和 $x_{\neg i}$ 分别表示除第 i 个词项以外所有词项的主题下标和所有词项的标签下标;ρ 和 υ 表示"标签-主题" 分布 θ 以及"主题-词项" 分布 φ 的先验分布参数;E 为当前文档的主题总数量;T 为当前文档的标签总数量;W 为当前文档的总词项数量;C_{hj}^{WE} 为在不考虑当前词项的情况下,主题 j 生成词项 h 的次数;C_{kj}^{TE} 为在不考虑当前主题的情况下,标签 k 生成主题 j 的次数;L 为所有文本语料库中所有单词的总数.

为了方便数学公式表述,定义用户模型中标签的数量为 m,词项的数量为 n. 利用特征词分布概率 θ 和 φ 构建用户主题模型中包含的一个大小为 $m \times m$ 的标签-标签矩阵 $\boldsymbol{M}$,$n \times n$ 的词项-词项矩阵 $\boldsymbol{A}$,$m \times n$ 的标签-词项矩阵 $\boldsymbol{R}$.

标签 t_i 与标签 t_j 之间的相似度公式如下:

$$sim(t_i, t_j) = \sum_{e=1}^{E} \theta_{ie} \ln\left(\frac{\theta_{ie}}{\theta_{je}}\right) + \sum_{e=1}^{E} \theta_{je} \ln\left(\frac{\theta_{je}}{\theta_{ie}}\right),$$

词项 w_i 与词项 w_j 之间的相似度公式如下:

$$sim(w_i, w_j) = \sum_{e=1}^{E} \varphi_{ei} \ln\left(\frac{\varphi_{ei}}{\varphi_{ej}}\right) + \sum_{e=1}^{E} \varphi_{ej} \ln\left(\frac{\varphi_{ej}}{\varphi_{ei}}\right),$$

标签 t_i 与词项 w_j 之间的相似度公式如下：

$$sim(t_i,w_j)=\sum_{e=1}^{E}\theta_{ie}\ln\left(\frac{\theta_{ie}}{\varphi_{ej}}\right)+\sum_{e=1}^{E}\varphi_{ej}\ln\left(\frac{\varphi_{ej}}{\theta_{ie}}\right).$$

为了方便描述标签与标签、词项与词项、标签与词项之间的关系，定义 4 个对角矩阵 $\boldsymbol{L}_M$、$\boldsymbol{L}_A$、$\boldsymbol{L}_{R1}$、$\boldsymbol{L}_{R2}$，表示相应矩阵 $\boldsymbol{M}$、$\boldsymbol{A}$、$\boldsymbol{R}$ 中第 i 个对角线上的元素，分别等于 $\boldsymbol{M}$ 中第 i 行的总和、$\boldsymbol{A}$ 中第 i 行的总和、$\boldsymbol{R}$ 中第 i 行的总和、$\boldsymbol{R}$ 中第 i 列的总和. 定义 $\boldsymbol{F}^0\in\mathbf{R}^{m\times c}$、$\boldsymbol{G}^0\in\mathbf{R}^{n\times c}$ 分别表示标签和词项在第一轮信息检索返回的前 c 个文档中的初始加权重(tf-idf) 矩阵. $\boldsymbol{F}\in\mathbf{R}^{m\times c}$、$\boldsymbol{G}\in\mathbf{R}^{n\times c}$ 分别表示通过个性化查询扩展后标签和词项在前 c 个文档中的最终加权重(tf-idf) 矩阵.

本文将用户模型中的词项和标签利用上述矩阵构建成 3 个图，分别表示为词项之间、标签之间、词项和标签之间的关系. 图中的点代表词项和矩阵，点与点之间的边的关系强度通过 $sim(t_i,t_j)$、$sim(w_i,w_j)$、$sim(t_i,w_j)$ 计算.

2 基于多语义关系的查询扩展方法

本节通过 $\boldsymbol{A}$、$\boldsymbol{M}$、$\boldsymbol{R}$、$\boldsymbol{F}^0$、$\boldsymbol{G}^0$ 将标签和词项之间的关系融合在本文的基于多语义关系的个性化查询扩展方法中，并通过 $\boldsymbol{F}$、$\boldsymbol{G}$ 选择查询扩展词. 表示个性化查询扩展后的标签–文档加权矩阵 $\boldsymbol{F}$ 和词项–文档加权矩阵 $\boldsymbol{G}$ 权重分数排名须与 $\boldsymbol{A}$、$\boldsymbol{M}$、$\boldsymbol{R}$、$\boldsymbol{F}^0$、$\boldsymbol{G}^0$ 的相关信息保持一致，且要使 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 的权重分数尽可能最小化，其中，$\boldsymbol{F}^0$，$\boldsymbol{G}^0$ 分别表示标签和词项在第一轮检索返回前 c 个文档中的初始加权矩阵. 目标函数如下所示：

$$\begin{aligned}Q(\boldsymbol{F},\boldsymbol{G})=&\frac{1}{2}\alpha\sum_{\substack{i=1\\j=1}}^{m}M_{ij}\left(\frac{1}{\sqrt{L_{M_{ii}}}}f(t,d_i)-\frac{1}{\sqrt{L_{M_{jj}}}}f(t,d_j)\right)^2+\\&\frac{1}{2}\beta\sum_{\substack{i=1\\j=1}}^{m}A_{ij}\left(\frac{1}{\sqrt{L_{A_{ii}}}}g(w,d_i)-\frac{1}{\sqrt{L_{A_{jj}}}}g(w,d_j)\right)^2+\\&\frac{1}{2}\gamma\sum_{i=1}^{m}\sum_{j=1}^{n}R_{ij}\left(\frac{1}{\sqrt{L_{R1_{ii}}}}f(t,d_i)-\frac{1}{\sqrt{D_{R2_{jj}}}}g(w,d_j)\right)^2+\\&\mu\sum_{i=1}^{m}(f(t,d_i)-f^0(t,d_i))^2+\\&\eta\sum_{i=1}^{n}(g(w,d_i)-g^0(w,d_i))^2.\end{aligned}$$

其中：$f^0(t,d_i)$、$g^0(w,d_i)$ 分别为标签和词项在第一轮检索返回的前 c 个文档 d_i 中的初始权重；$f(t,d_i)$、$g(w,d_i)$ 分别为标签和词项在第一轮检索返回的前 c 个文档 d_i 中的经考虑多语义关系后的权重；α 控制标签之间的关系，β 控制词项之间的关系；γ 控制标签和词项之间的关系；μ、η 为正则化参数.

在目标函数中，前 3 项为平滑约束因子. 第 1 项挖掘标签与标签之间的关系，表示相似标签应具有相同的权重分数；第2项挖掘第1轮检索返回的文档结果列表中最靠前的 c 个文档中词项与词项的关系，表示语义相似的词项应具有相似的权重分数；第 3 项挖掘标签和前 c 个文档中词项之间的关系，对于某一文档，词项的权重值越大，意味着该词项语义越接近文档的主题，词项和标签之间应具有相似的权重分数. 第 4 项和第 5 项用于最小化权重分数和第 1 轮检索排名靠前文档集之间的差异性. 上述项之间的平衡都由归一化参数 α、β、γ、μ、η 控制，在本次实验中，设置 $0<\alpha<1,0<\beta<1,0<\gamma<1,0<\mu<1,0<\eta<1$. 为使所有数据处于同一数量级，对 $\boldsymbol{A}$、$\boldsymbol{M}$、$\boldsymbol{R}$ 进行归一化处理，定义

$$\boldsymbol{S}_A=\boldsymbol{L}_A^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{L}_A^{-\frac{1}{2}},\ \boldsymbol{S}_M=\boldsymbol{L}_M^{-\frac{1}{2}}\boldsymbol{M}\boldsymbol{L}_M^{-\frac{1}{2}},\ \boldsymbol{S}_R=\boldsymbol{L}_{RA}^{-\frac{1}{2}}\boldsymbol{R}\boldsymbol{L}_{RM}^{-\frac{1}{2}}.$$

在经过一系列简单的推导[23]后，可将第 1 项替换成矩阵的表现形式 $\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S}_A)\boldsymbol{F}$，第2项可写成 $\boldsymbol{g}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S}_A)\boldsymbol{G}$，第 3 项可转化为 $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}+\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}-2\boldsymbol{F}^{\mathrm{T}}\boldsymbol{S}_R\boldsymbol{G}$.

然后通过使用等价的矩阵向量可重写目标函数：

$$\begin{aligned}Q(\boldsymbol{F},\boldsymbol{G})=&\alpha\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S}_M)\boldsymbol{F}+\beta\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{S}_A)+\\&\gamma(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}+\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}-2\boldsymbol{F}^{\mathrm{T}}\boldsymbol{S}_R\boldsymbol{G})+\\&\mu(\boldsymbol{F}-\boldsymbol{F}^0)^{\mathrm{T}}(\boldsymbol{F}-\boldsymbol{F}^0)+\\&\eta(\boldsymbol{G}-\boldsymbol{G}^0)^{\mathrm{T}}(\boldsymbol{G}-\boldsymbol{G}^0).\end{aligned}$$

目标函数分别对 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 求偏导：

$$\frac{\partial\boldsymbol{Q}}{\partial\boldsymbol{F}}=[(1-\beta-\eta)\boldsymbol{I}-\alpha\boldsymbol{S}_A]\boldsymbol{F}-\gamma\boldsymbol{S}_R\boldsymbol{G}-\mu\boldsymbol{F}^0=0,\quad(1)$$

$$\frac{\partial\boldsymbol{Q}}{\partial\boldsymbol{G}}=[(1-\alpha-\mu)\boldsymbol{I}-\beta\boldsymbol{S}_M]\boldsymbol{G}-\gamma\boldsymbol{S}_R^{\mathrm{T}}\boldsymbol{F}-\eta\boldsymbol{G}^0=0.\quad(2)$$

为简化公式，令

$$\boldsymbol{\chi}=[(1-\beta-\eta)\boldsymbol{I}-\alpha\boldsymbol{S}_A],$$

$$\boldsymbol{\psi}=[(1-\alpha-\mu)\boldsymbol{I}-\beta\boldsymbol{S}_M],$$

根据式(2)可得

$$\boldsymbol{G}=\boldsymbol{G}^{-1}(\gamma\boldsymbol{S}_R^{\mathrm{T}}\boldsymbol{F}+\eta\boldsymbol{G}^0),\qquad(3)$$

然后将 $\boldsymbol{G}$ 代入式(1)中，得到文档重新排序的权重分数：

$$\boldsymbol{F}=(\boldsymbol{\chi}-\gamma^2\boldsymbol{S}_R\boldsymbol{\psi}^{-1}\boldsymbol{S}_R^{\mathrm{T}})^{-1}(\gamma\eta\boldsymbol{S}_R\boldsymbol{\psi}^{-1}\boldsymbol{G}^0+\mu\boldsymbol{F}^0),$$

在求解 $\boldsymbol{F}$ 之后，可将其代入式(3)，并求解 $\boldsymbol{G}$.

对于某一给定的查询词，给定个性化查询扩展

后的加权矩阵 $\boldsymbol{F}$、$\boldsymbol{G}$,每个标签 t 和词项 w 的最终加权得分可通过

$$t = \sum_{i=1}^{m} F(t,d_i),\ w = \sum_{i=1}^{c} G(w,d_i)$$

进行计算排序,选择靠前的 δ 个标签和 ζ 个词项用于扩展初始查询词.

3　实验及结果分析

3.1　实验数据

为在真实数据集上验证方法,本文选择来自 Del. icio. us 的 2 个公开数据集:socialbm0311、deliciousT140. socialbm0311 包含从 2003 年至 2011 年约 200 万名用户完整的社会化标签数据. deliciousT140 包含 2008 年 6 月 144 547 条 URL 对应的网页,这 2 个数据集在文献[24]和文献[25]中都有详细描述. deliciousT140 数据集包含具体的 Web 网页,即文档,而 socialbm0311 包含用户的标注数据. 对 2 个数据集进行合并处理后,得到 5 153 720 条标注记录,259 511 名用户,131 283 个 Web 网页,137 870 个标签. 本文使用公开的工具 htmlparser (http://htmlparser. sourceforge. net) 对 Web 网页进行文本提取.

为了验证方法的有效性,本文选取 4 组用户作为测试用户:1)标注记录少于 50 条的用户(记为 BK50),2)标注记录介于 50 至 100 条之间的用户(记为 BK100),3)标注记录介于 100 至 500 条之间的用户(记为 BK500),4)标注记录多于 500 条的用户(记为 BKG500). 这一分组充分表示较活跃的用户和不太活跃的用户.

针对每组用户,随机选取 50 名作为测试用户. 对于每名用户,使用 75% 的标签及被标注资源构建用户兴趣模型,剩余的 25% 作为测试数据. 该选择与以往的研究相同[6,12]. 另外随机选择 200 名来自数据集中的用户,训练模型所需的参数,该 200 名用户与测试用户并无任何重叠.

3.2　评价标准

与以往基于社会化标签系统的个性化搜索研究相同[6,12,15],本文采取如下方法评价相关性. 若某一用户使用某一标签 t 作为查询词(u,t) 进行搜索,标记的文档 u 视为相关文档. 值得注意的是,该标签与该文档的关系在搜索时对于系统不可见.

本文实验采用如下 4 种评价标准.

1)平均精度均值(Mean Average Precision, MAP),每篇相关文档检索后的准确率的平均值.

2)平均倒数排名(Mean Reciprocal Rank, MRR),查询结果的倒数排名是第 1 个相关文档的倒数积.

3)归一化折损累积增益(Normalized Discounted Cumulative Gain, NDCG),对检索结果列表中相关文档有效性的度量.

4)前 5 个文档准确率(Precision of Top 5 Documents, P@5),返回检索结果列表中前 5 名的相关文档的准确率.

计算结果为每组所有用户的平均表现. 显著差异由配对样本 t 检验测定.

3.3　基线方法

本文选取一系列非个性化和个性化搜索方法作为基线系统与本文方法进行对比,这些基线方法及本文方法分别说明如下.

1)基于 BM25 模型的方法(Method Based on BM25 Model, BM25). 一种通用且具有鲁棒性的基于语言模型的搜索方法,以往的实验表明该方法能取得不错的搜索效果[26].

2)基于 BM25 模型的伪相关反馈方法(Pseudo-Relevance Feedback Method Based on BM25, BM25PRF)[27]. 相比 BM25,取得更好效果,可作为一个有竞争性的非个性化搜索基线方法.

3)基于词共现方法(Method Based on Terms Cooccurrence, COOC). 为个性化的查询扩展方法,以往有很多研究者采用该方法[4,10]. COOC 计算查询词词项与用户兴趣模型中的字面共现关系以选择扩展词. 以往的研究表明,方法可以取得不错效果.

4)基于词项的标签主题方法(Method Based on Tag-Topic of Terms, TT_terms). 作为文献[12]方法的变体之一,该方法将用户兴趣模型中的词项利用标签-主题模型构建成图,以便准确把握词项与词项间的关系,再借助正则化框架进行扩展词的选择,具有较高关联权重的词项用于扩展原始查询词.

5)基于标签的标签主题方法(Method Based on Tag-Topic of Tags, TT_tags). 作为文献[12]方法的变体之一,该方法将用户兴趣模型中的标签利用标签-主题模型构建成图,以便准确把握标签与标签关系,再借助正则化框架进行扩展词的选择,具有较高关联权重的标签用于扩展原始查询词.

6)基于词项与标签的混合方法(Method Based on Mixed Terms and Tags, TT_mix). Zhou 等[12]提出基于语义匹配的个性化查询扩展方法,将用户兴趣模型中的词项和标签利用标签-主题模型构建成图,以便准确把握词项与标签间的关系,再借助正则

化框架进行扩展词的选择(标签和词项). 该方法在社会化标签系统中的实验效果优于 TT_terms 和 TT_tags.

7) 基于词项的语义方法(Method Based on Semantics of Terms, MLE_terms). 作为本文方法之一,该方法基于个人用户兴趣模型,将用户模型中的词项和标签利用标签-主题模型构建 3 个图,分别表示词项之间、标签之间、词项和标签之间的关系,再借助社会标签的多语义关系,选择具有较高关联权重的词项,用于扩展原始查询词.

8) 基于标签的语义方法(Method Based on Semantics of Tags, MLE_tags). 作为本文方法之一,该方法基于个人用户兴趣模型,将用户模型中的词项和标签利用标签-主题模型构建 3 个图,分别表示词项之间、标签之间、词项和标签之间的关系,再借助社会标签的多语义关系,选择具有较高关联权重的标签用于扩展原始查询词.

9) 基于词项与标签的混合语义方法(Method Based on Mixed Semantics of Terms and Tags, MLE_mix). 作为本文方法之一,该方法基于个人用户兴趣模型,将用户模型中的词项和标签利用标签-主题模型构建 3 个图,分别表示词项之间、标签之间、词项和标签之间的关系,再借助社会标签的多语义关系,选择具有较高关联权重的标签和词项,扩展原始查询词.

上述个性化查询扩展方法均采用社会化标签系统作为测试数据集.

由于本文提出的个性化搜索方法基于查询扩展,因此无法与重排序方法[7]进行直接对比.

3.4 参数设定

本节介绍在查询扩展框架中使用的参数设置. E 为主题模型的主题数目,α 控制标签之间信息的重要性,β 控制文档中词项之间信息的重要性,γ 控制标签与前 c 个文档中词项之间信息的重要性,μ 控制标签权重排序的效果,η 控制词项权重排序的效果.

为了提高个性化查询扩展的效率,MLE_terms、MLE_tags 和 MLE_mix 中参数设置如下:$E=5$,$\alpha=0.55$,$\beta=0.06$,$\gamma=0.1$,$\mu=0.4$,$\eta=0.1$. 确定上述 6 个参数的取值方法为,固定 5 个参数不变,仅调整 1 个参数的取值范围,直到找到使 MRR 达到最优的取值.

3.4.1 c 取值对算法的影响

为了提高检索的效率,设置 MLE_mix 中第 1 轮检索返回的文档结果列表中最靠前的 c 个文档数目的调整范围为[1,10],递增区间长度为 1. 如图 1 所示,使 MRR 达到最优的参数值 $c=4$. 当 $c\leqslant 4$ 时,随着文档数目的增加,搜索性能提升. 当 $c\geqslant 4$ 时,随着文档数目的增加,搜索性能开始下降. 这是因为,网页本身存在噪音,需要选择合适的文档数目以实现最佳扩展性能. 类似地,MLE_terms 和 MLE_tags 使 MRR 达到最优的参数值分别为 $c=6$ 和 $c=3$.

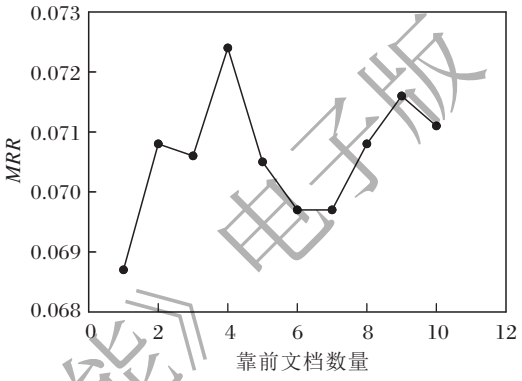


图 1 靠前文档数目对算法性能的影响

Fig. 1 Influence of the number of top-ranked documents on the performance of algorithm

3.4.2 ζ 取值对算法的影响

在 MLE_mix 中扩展词项 ζ 数目的调整范围为[10,100],递增区间长度为 10. 如图 2 所示,使 MRR 达到最优的参数值 $\zeta=40$. 当 $\zeta\leqslant 40$ 时,随着扩展词数目的增加,搜索性能达到持平状态. 类似地,在 MLE_terms 中,扩展词项的数目设置为 $\zeta=40$,在 MLE_tags 中,扩展标签的数目均设置为 $\sigma=5$,在 MLE_mix 中,扩展词项和扩展标签的数目设置为 $\sigma+\zeta=40$,$\sigma=5$. 基线系统中的参数均按照原始文献在训练用户上进行训练并使 MRR 达到最优的参数取值作为测试参数值.

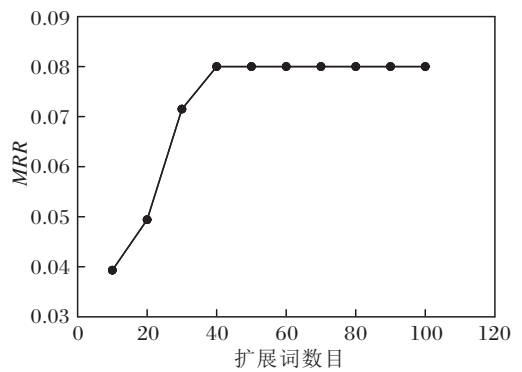


图 2 扩展词数目对算法性能的影响

Fig. 2 Influence of the number of expansion terms on the performance of algorithm

3.5　总体性能分析

表 1 为所有方法在总体测试用户中的实验结果. 与非个性化搜索方法 BM25PRF 结果之间的显著差异使用 * 表示,与个性化搜索方法 TT_mix 结果之间的显著差异使用#表示.

从实验可看出,BM25 在用户中的搜索效果均最差. 借助伪相关反馈(BM25PRF)的帮助,搜索结果取得一定提高. 然而,几乎所有的非个性化搜索方法都比个性化搜索方法,包括本文方法的搜索效果更差,而且结果对比都具有显著差异. 实验表明,相比非个性化的方法,个性化的方法可以大幅提高搜索的准确率.

表 1　各方法在总体测试用户中的实验结果

Table 1　Experimental results of different methods on overall test users

方法	MAP	MRR	NDCG	P@5
BM25	0.0348	0.0388	0.0533	0.010
BM25PRF	0.0361	0.0401	0.0546	0.0111
COOC	0.0355	0.0393	0.0565	0.0120 *
TT_terms	0.0311	0.0408 *	0.0523	0.0110
TT_tags	0.0348 *	0.0390	0.0510	0.0114 *
TT_mix	0.0384 *	0.0462 *	0.0555 *	0.0124 *
MLE_terms	0.0410 * #	0.0613 * #	0.0648 * #	0.0187 * #
MLE_tags	0.0370 * #	0.0416 * #	0.0557 * #	0.0119 * #
MLE_mix	0.0512 * #	0.0817 * #	0.0759 * #	0.0159 * #

对比本文方法与个性化搜索基线方法,可得如下结论.

1)MLE_mix 在评价标准 MAP、MRR 和 NDCG 上搜索效果明显优于所有的基线方法,结果都具有显著差异. 这一结果说明使用语义匹配方法进行扩展词的选择优于字面匹配的方法.

2)MLE_mix 在评价标准 P@5 上搜索效果差于 MLE_terms,说明用户标签可能存在歧义性,对文本描述的不准确性会影响检索的准确率.

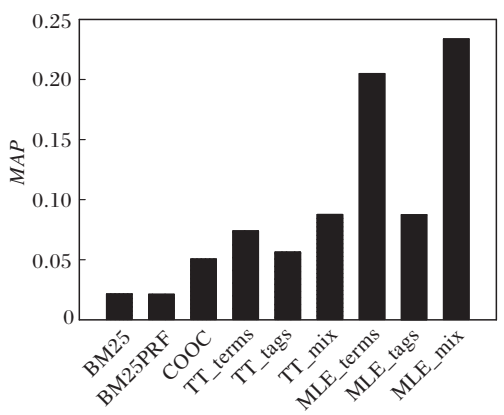
3)仅基于共现统计和词项匹配的查询扩展技术检索效率的提高有限. 在所有评价指标中,相比基线系统 BM25PRF,COOC 的性能更低. 这是因为本次实验是基于真实社交网络数据集,可以使用任意数量任意形式的标签标注各个网页资源,导致网页带有过多噪音.

4)本文的 3 种方法表现均好于文献[12]中只考虑标签(TT_tags)和词项(TT_terms)之间有限语义关系的方法. 相比 TT_terms,MLE_terms 的 MRR 提高 50.24%,相比 TT_tags,MLE_tags 提高 6.67%. 相比 TT_mix,MLE_mix 的 MRR 提高 51.29%. 这一结果表明,本文提出的通过结合标签-标签、词项-词项、标签-词项语义关系的个性化查询扩展方法是有效的.

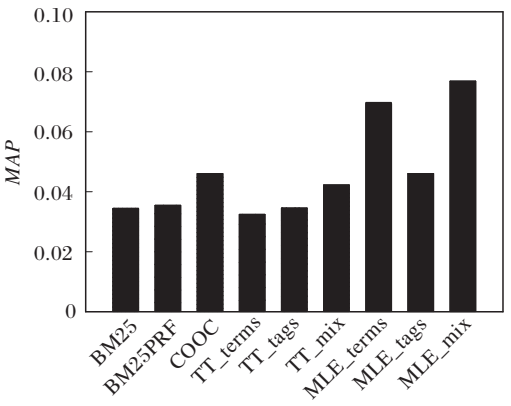
3.6　分组测试用户性能分析

图 3 为各方法在分组测试用户中的实验结果. 从实验可看出,BM25 在 BK50、BK100、BK500 这 3 组用户中的搜索效果均最差(使用 MAP 的评价标准). 由此表明,仅基于统计语言模型的信息检索改进有限,而个性化查询扩展方法可以大幅提高检索有效性.

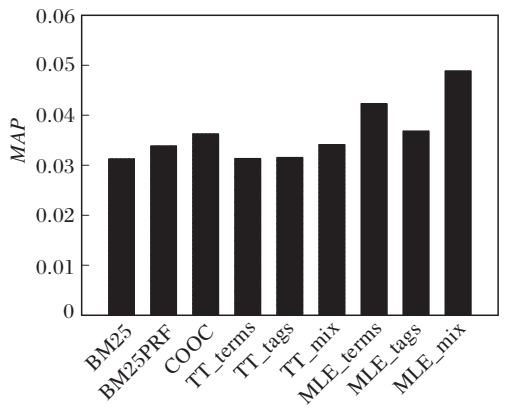
然而,并非所有非个性化基线都超越个性化方法,如 BM25PRF 在 BKM500 中均好于除 MLE_mix 方法之外的所有个性化方法的搜索效果(使用 MAP 的评价标准),而且所有的结果对比都具有显著差异. 因此用户可以使用任意数量的自由形式标签注释每个网页资源,虽然它可能无法对应网页精确描述,导致存储在用户模型中的关键字带有噪声. MLE_mix 在不同组别的用户中搜索效果明显优于所有个性化查询扩展方法和非个性化方法(使用 MAP 的评价标准). 这说明将社会标签用于多重语义关系模型可以提升搜索效果. 由于本文方法包括标签-标签、词项-词项、词项-标签之间的相似性,在不同组别的用户中表现稳定,特别是在 BK50 中,效果提升程度大于其它组. 实验表明本文方法不仅对较不活跃用户可以取得不错效果,对于活跃的用户,也可以取得不错的效果提升.



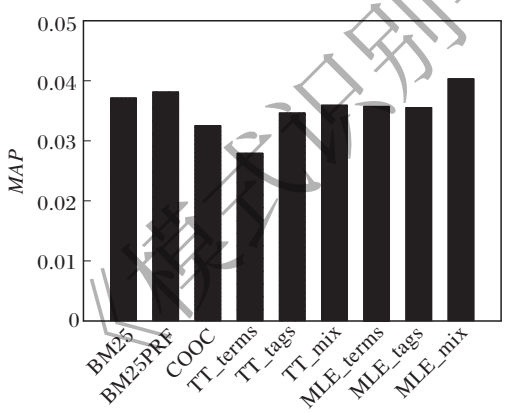
(a)BK50



(b) BK100



(c) BK500



(d) BKM500

图 3 各方法在分组测试用户中的实验结果

Fig . 3 Experimental results of users from different groups

4 结 束 语

利用社交数据进行个性化搜索的相关研究现在已经成为研究热点. 本文提出基于多语义关系的个性化查询扩展方法,使用 3 个信息来源,即标签之间、词项之间及标签和词项之间的关系,选择扩展词项. 在大规模真实数据集上的实验表明,本文方法明显优于非个性化搜索方法及其它基于查询扩展的个性化搜索方法. 今后将考虑糅合更多与社会标签数据相关的信息,并融合其它高效的算法和模型,提高个性化查询扩展的性能及检索结果的准确率.

参 考 文 献

[1] 李树青. 个性化信息检索技术综述. 情报理论与实践, 2009, 32(5): 107-113.
(LI S Q. Personalized Information Retrieval Review. Information Studies: Theory and Application, 2009, 32(5): 107-113.)

[2] GHORAB M R, ZHOU D, O'CONNOR A, *et al*. Personalised Information Retrieval: Survey and Classification. User Modeling and User-Adapted Interaction, 2013, 23(4): 381-443.

[3] SHEN X H, TAN B, ZHAI C X. Implicit User Modeling for Personalized Search // Proc of the 14th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2005: 824-831.

[4] CHIRITA P A, FIRAN C S, NEJDL W. Personalized Query Expansion for the Web // Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2007: 7-14.

[5] AGICHTEIN E, BRILL E, DUMAIS S. Improving Web Search Ranking by Incorporating User Behavior Information // Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2006: 19-26.

[6] XU S L, BAO S H, FEI B, *et al*. Exploring Folksonomy for Personalized Search // Proc of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2008: 155-162.

[7] BOUADJENEK M R, HACID H, BOUZEGHOUB M, *et al*. Using Social Annotations to Enhance Document Representation for Personalized Search // Proc of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2013: 1049-1052.

[8] XIE H R, LI X D, WANG T, *et al*. Personalized Search for Social Media via Dominating Verbal Context. Neurocomputing, 2016, 172: 27-37.

[9] DOU Z C, SONG R H, WEN J R. A Large-Scale Evaluation and Analysis of Personalized Search Strategies // Proc of the 16th International Conference on World Wide Web. New York, USA: ACM, 2007: 581-590.

[10] BIANCALANA C, MICARELLI A. Social Tagging in Query Expansion: A New Way for Personalized Web Search // Proc of the International Conference on Computational Science and Engineering. Washington, USA: IEEE, 2009, IV: 1060-1065.

[11] 张志强,孟庆海,谢晓芹. 个性化的社会标签查询扩展技术研究. 计算机科学与探索, 2010, 4(9): 812-829.

(ZHANG Z Q, MENG Q H, XIE X Q. Research on Personalized Social Tag Query Expansion Techniques. Journal of Frontiers of Computer Science and Technology, 2010, 4(9): 812-829.)

[12] ZHOU D, LAWLESS S, WADE V. Improving Search via Personalized Query Expansion Using Social Media. Information Retrieval, 2012, 15(3/4): 218-242.

[13] BOUADJENEK M R, HACID H, BOUZEGHOUB M. SoPRa: A New Social Personalized Ranking Function for Improving Web Search // Proc of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2013: 861-864.

[14] 李 鹏,王 斌,晋 薇. 一种基于社会化标签的信息检索方法. 中文信息学报, 2013, 27(1): 39-46.
(LI P, WANG B, JIN W. An Information Retrieval Method Based on Social Annotations. Journal of Chinese Information Processing, 2013, 27(1): 39-46.)

[15] WANG Q H, JIN H X. Exploring Online Social Activities for Adaptive Search Personalization // Proc of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2010: 999-1008.

[16] CAI Y, LI Q. Personalized Search by Tag-Based User Profile and Resource Profile in Collaborative Tagging Systems // Proc of the 19th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2010: 969-978.

[17] XIE H R, LI X D, WANG T, *et al*. Incorporating Sentiment into Tag-Based User Profiles and Resource Profiles for Personalized Search in Folksonomy. Information Processing & Management, 2016, 52(1): 61-72.

[18] BENDER M, CRECELIUS T, KACIMI M, *et al*. Exploiting Social Relations for Query Expansion and Result Ranking // Proc of the 24th IEEE International Conference on Data Engineering Workshop. Washington, USA: IEEE, 2008: 501-506.

[19] BERTIER M, GUERRAOUI R, LEROY V, *et al*. Toward Personalized Query Expansion // Proc of the 2nd ACM EuroSys Workshop on Social Network Systems. New York, USA: ACM, 2009: 7-12.

[20] 冯 勇,刘 瑶,徐红艳. 一种基于标签用户模型的个性化信息检索方法. 小型微型计算机系统, 2014, 35(9): 2004-2008.
(FENG Y, LIU Y, XU H Y. Personalized Information Retrieval Method Based on Tag User Model. Journal of Chinese Computer Systems, 2014, 35(9): 2004-2008.)

[21] BIANCALANA C, GASPARETTI F, MICARELLI A, *et al*. Social Semantic Query Expansion. ACM Transactions on Intelligent Systems and Technology, 2013, 4(4): 1-43.

[22] ZHOU D, LAWLESS S, WADE V. Web Search Personalization Using Social Data // Proc of the 2nd International Conference on Theory and Practice of Digital Libraries. Berlin, Heidelberg: Springer, 2012: 298-310.

[23] YANG L P, JI D H, ZHOU G D, *et al*. Document Re-ranking Using Cluster Validation and Label Propagation // Proc of the 15th ACM International Conference on Information and Knowledge Management. New York, USA: ACM, 2006: 690-697.

[24] ZUBIAGA A, GARCIA-PLAZA A P, FRESNO V, *et al*. Content-Based Clustering for Tag Cloud Visualization // Proc of the International Conference on Advances in Social Network Analysis and Mining. New York, USA: ACM, 2009: 316-319.

[25] ZUBIAGA A, FRESNO V, MARTINEZ R, *et al*. Harnessing Folksonomies to Produce a Social Classification of Resources. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(8): 1801-1813.

[26] ROBERTSON S, ZARAGOZA H. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

[27] AMATI G, VAN RIJSBERGEN C J. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transactions on Information Systems (TOIS), 2002, 20(4): 357-389.

作者简介

伍 璇,女,1993 年生,硕士研究生,主要研究方向为信息检索、自然语言处理. E-mail: xuanwu0708@ gmail. com.
(**WU Xuan**, born in 1993, master student. Her research interests include information retrieval and natural language processing.)

周 栋(通讯作者),男,1979 年生,博士,副教授,主要研究方向为信息检索、自然语言处理. E-mail: dongzhou1979@ hotmail. com.
(**ZHOU Dong** (Corresponding author), born in 1979, Ph. D., associate professor. His research interests include information retrieval and natural language processing.)