吴谈,周栋,包恒泽. 基于用户类别兴趣偏好的个性化排序方法[J].湖南科技大学学报(自然科学版),2020,35(1): 104-112.doi:10.13582/j.cnki.1672-9102.2020.01.015

Wu T, Zhou D, Bao H Z. Personalized Ranking Method based on User Preferences of Classes Interests[J]. Journal of Hunan University of Science and Technology(Natural Science Edition), 2020,35(1):104-112.doi:10.13582/j.cnki.1672-9102.2020.01.015

基于用户类别兴趣偏好的个性化排序方法

吴谈,周栋*,包恒泽

(湖南科技大学 计算机科学与工程学院,湖南 湘潭 411201)

摘 要:信息检索中,个性化排序在传统的基于内容匹配的排序算法基础上,结合用户兴趣特征,返回更符合用户需求的检索结果.由于用户数据存在稀疏性和兴趣爱好不均衡等问题,用户兴趣偏好模型构建通常不是很精确,检索效果也不佳.本文在前人研究的基础上,提出了一种基于用户类别偏好的个性化排序方法.该方法首先借助词向量技术计算查询词和文档标签集之间的语义相似程度,其次,考虑到用户对不同兴趣的偏好程度不一,通过构建用户兴趣偏好模型,计算出用户对不同兴趣类别的偏好程度,对待查询文档进行个性化处理,以达到个性化排序的目的.在真实数据集上的实验表明,与传统方法相比,本文提出的方法可以有效地改善用户的个性化检索效果.

关键词:个性化排序;社会化标注;词向量;兴趣偏好模型

中图分类号:TP399 **文献标志码**:A **文章编号**:1672-9102(2020)01-0104-09

Personalized Ranking Method Based on User Preferences of Classes Interests

Wu Tan, Zhou Dong, Bao Hengze

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

Abstract: In information retrieval, the method of personalized sequencing based on the traditional content matching sorting algorithm combined with the interest of users in order to achieve some better search results that were more in line with users' needs. Due to the sparseness of user's data and the imbalance of interests, the construction of user's interest preference model was usually not very accurate, and the retrieval effect was little and poor. Based on those previous studies, a personalized ranking method, based on user category preference, was proposed. Firstly, the word embedding technology was used to calculate the semantic similarity between the query words and the document tag set. Secondly, considering the user´s preference for different interests, the user interest model was constructed to calculate the user´s interest categories. The degree of preference, the query documents were personalized to achieve the purpose of personalized sorting. Experiments on real data sets show that compared with traditional methods, the proposed method effectively improve the user´s personalized retrieval results.

Keywords: personalized ranking; social annotation; word embedding; the preference of interest model

收稿日期:2018-12-07

基金项目:国家自然科学基金资助项目(61876062);湖南省自然科学基金资助项目(2017JJ2101);湖南省教育厅科研项目资助(16K030)

* **通信作者**,E-mail:dongzhou1979@hotmail.com

随着互联网技术的发展和信息时代的到来,海量的数据呈现在人们面前,如何从如此浩瀚且日益增长的数据中快速而精准地获取用户所需要的信息,一直是信息检索研究中的热点问题.而信息检索的返回结果,则是衡量一个信息检索系统性能好坏的重要指标[1].此外,在通常情况下,一次检索返回的结果只有部分满足用户的信息需求,并且大多数用户只会浏览排名靠前的检索结果[2].提高这部分结果的相关性,才能带给用户更好的搜索体验.传统的排序算法虽然在一定程度上能提升信息检索系统的性能,但是却无法满足不同用户在相同条件下的不同需求,难以得到因人而异的检索结果.为了解决这一问题,让排序结果更精确,学者们引进了个性化的思想,研究面向用户的个性化排序算法[3].

个性化排序算法在传统排序算法的基础上,结合用户兴趣特征信息,提高检索结果的准确率,但出于一些隐私保护的机制,很多用户信息是无法获取的.而社会化标注作为一种公开资源,且它们由用户直接提供,这种标注行为在一定程度上也反映了用户想要关注的领域,即用户兴趣,因此用户的兴趣偏好可以通过他/她的标注集来获取[4].

已有的研究中,Xu 等[5]将社会化标注运用到个性化检索中,利用用户标注信息向量和网页的标注信息向量之间的相似度对原始检索结果进行个性化排序.但对于一些文本内容相对较少的网页,不足以生成令人满意的搜索结果.Bouadjene 等[6]对此进行了扩展,认为用户标注信息是网页内容的有力补充,同时考虑了查询信息与网页文本内容、网页标注信息之间的相似程度,用于改善原始检索结果.Xu 等[7]则在文献[6]的基础上提出了一种双重个性化排序方法(Dual Personalized Ranking,记为 D-PR),该方法区分了不同用户的标注信息对网页内容的贡献程度,利用相似用户对标注信息进行扩展,从而使检索结果更具个性化.上述方法都在一定程度上提升了个性化检索的效果,但仍然存在用户数据稀疏性和爱好不均衡的问题.

本文在前人研究的基础上,提出了一种融合用户兴趣偏好的个性化排序方法(Personalized Ranking Integrating User Preference,记为 PRIWuP).在检索过程中,该方法考虑到查询信息和网页的标注信息对检索结果的影响,首先借助 Word2Vec 将查询词和网页标注信息转化成为词向量,然后利用二者的词向量计算出检索结果的分数,其次还考虑了在检索排序时用户对于不同网页的兴趣偏好不同,使用用户的社会化标注信息为其构建兴趣偏好模型,再根据用户对检索排序过程中文档的不同加权值,实现个性化排序.在真实数据集上的实验证明,本文提出的排序方法可以有效地提高个性化检索的准确率.

1　相关工作

近年来,个性化排序算法得到了广泛关注,例如,王晓春等[9]将用户的长短期兴趣结合,利用用户长期兴趣和短期兴趣对查询模型进行改进,来提升用户的个性化检索体验.Zhi 等[10]通过对文档聚类获取查询相关类别,将此类别作为用户可能的查询意图,提高个性化算法的有效性.Dou 等[11]提出了从查询日志、锚文本、文档聚类以及检索结果的网页等 4 种不同的来源中抽取子主题,最后通过与用户查询意图相匹配,实现个性化.Tu 等[12]也考虑利用查询日志和文档集进行分析与挖掘子主题,同时认为有时效性的查询,也应挖掘随时间变化的不同子主题,在文献[11]的基础上,利用随机游走的方法从查询日志中和文档集中获取具有时效性的子主题,以实现个性化排序.

文献[13-15]等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 搜索结果的检索和排序.Wang 等[16]通过挖掘用户在多个社会媒体上的公开活动,来进行个性化搜索.Gao 等人[17]研究用户的标注信息和检索结果之间的关系,使得最后的检索结果能更好地满足用户的兴趣偏好.Li 等人[18]在建立个性化搜索引擎时融入用户的影响力和话题信息.王庆林等[19]针对多义标签问题,应用图聚类算法把语义相近的标签进行聚类,然后以标签类别为中介衡量特定用户和资源的相关度.Ji 等[20]利用用户-资源矩阵计算用户之间相似性,再结合标签-资源矩阵和用户-标签矩阵,计算出用户对特定标签的喜好度,从而返回一组具有个性化的检索结果.文献[21]利用用户的标注信息和用户标注的资源信息构建用户兴趣模型,提出了一种基于用户兴趣模型的个性化查询扩展方法.管毅舟等

人[22]通过结合社会化标注和网页分类筛选出兴趣和偏好相近的用户,进行用户属性的扩展,并在扩展时考虑用户的质量,以实现个性化检索.

以上研究普遍存在一些不足:(1)在匹配查询信息和网页标注信息相似程度得分时,以往所采用的方法是简单的词频向量空间模型,而实际上,查询词一般具有简短、概要以及不精确3个特点,所以这类方法在很多情况下无法精确计算出查询信息和文档标注信息的正确匹配得分;(2)考虑到不同用户对不同类型的文档偏好程度不一样,而大部分研究只是将个性化信息与内容检索得分简单的线性拟合,并没有在内容检索部分融入个性化信息,这在一定程度上无法反映出二者之间的内在联系问题.因此,本文主要做了以下工作:

1)利用Word2vec模型将每个标签词转化为一个包含了上下文信息的词向量,然后将查询信息和文档标注信息转化为向量形式,通过二者词向量的计算,更新查询信息对文档检索结果的分数.

2)考虑到用户兴趣偏好不一的问题,利用用户的社会化标注信息和标注信息之间的相似性构建用户兴趣偏好模型,再根据用户对检索排序过程中文档的不同加权值,以加强文档个性化处理.

2 问题描述

2.1 相关概念

本文的个性化排序方法是基于社会化标注系统构建的.社会化标注系统旨在为用户提供一种给网页资源自由标注标签的途径.由于标签可以与其他人共享,其数据模型可形式化地描述为 $F:(\boldsymbol{U},\boldsymbol{T},\boldsymbol{D})$,其中 $\boldsymbol{U},\boldsymbol{T},\boldsymbol{D}$ 分别表示用户、标签、文档的有限集合.定义在 F 上的三元关系 $F=\{(u,t,d)\mid u\in\boldsymbol{U},t\in\boldsymbol{T},d\in\boldsymbol{D}\}$,表示用户 u 用标签 t 对文档 d 进行了标注.

2.2 问题定义

给定用户 $u\in\boldsymbol{U}$,当用户向搜索引擎提交查询 q 时,搜索引擎会返回一组与 q 相匹配的原始文档集合 $\boldsymbol{D}_q$,并按照相关程度大小进行排序.假设 $\boldsymbol{D}_q$ 遵循的排序序列为 $\tau=[d_1\geqslant d_2\geqslant\cdots\geqslant d_s]$,在这里定义 $d_i\geqslant d_j$:$\mathrm{rankScore}(d_i,q)\geqslant\mathrm{rankScore}(d_j,q)$,其中,$d_i$ 为文档集 $\boldsymbol{D}$ 中的第 i 篇文档,$\mathrm{rankScore}(d,q)$ 是用于计算查询 q 和文档 d 相关程度大小的得分函数.由于用户在通常情况下只关心靠前结果,并且不同用户对于结果的期望值不一样,因此,个性化排序算法可以描述为对第1轮检索结果 $\boldsymbol{D}_q$ 中的文档,针对不同用户,按照个性化相关程度大小进行重新排序,使得与用户相关度大的文档排在靠前的位置,得到新的文档排序集合 $\boldsymbol{D}_{q,u}$,排序序列为 $\tau'=[d_j\geqslant d_s\geqslant\cdots\geqslant d_i]$. 为方便阅读,表1总结了本文所使用符号及其含义.

表1 符号及含义说明

符号	含义说明	符号	含义说明
$\boldsymbol{U}$	用户集	$\boldsymbol{U}_d$	标记过文档 d 的用户集
$\boldsymbol{D}$	文档集	$\boldsymbol{T}_d$	文档 d 上的标注信息
$\boldsymbol{T}$	标签集	T_{C_i}	类别 C_i 包含的标注信息
$\boldsymbol{C}$	兴趣类别集	$\boldsymbol{S}_t$	标签 t 的相似标签集
C_i	第 i 个兴趣类别	$\boldsymbol{S}_u$	用户 u 的相似用户集
$\boldsymbol{U}^{t_i}_*$	使用了标签 t_i 的用户集	$\boldsymbol{V}_t$	标签 t 的词向量
$\boldsymbol{C}^{t_i}_*$	含有标签 t_i 的主题域集	w_{u,C_i}	用户 u 对主题域 C_i 的偏好权重
$w^u_{t_i}$	用户 u 对标签 t_i 的偏好权重	$\boldsymbol{Z}_u$	用户 u 对各兴趣类别的偏好向量
$w^{C_i}_{t_i}$	兴趣类别 C_i 中标签 t_i 的偏好权重	$\boldsymbol{P}_u$	用户 u 的标签偏好向量
$\theta_{m,j}$	第 m 篇文档主题 j 的向量权重	P'_u	扩展用户标注信息
$\varphi_{j,term}$	第 m 篇文章中主题 j 下词汇 term 的权重	$P_{u,d}$	文档 d 对于用户 u 的个性化文档属性
ctf_{t_i}	标签 t_i 在兴趣类别 C_i 中出现的次数	α,β	线性拟合参数
utf_{t_i}	用户 u 使用标签 t_i 的次数	K	划分的兴趣类别个数

2.3 双重个性化排序方法(D-PR)

为了便于理解,首先对D-PR[7]做简要介绍,该方法的目的在于计算排序得分 $\mathrm{Rank}(d,q,u)$,由3部

分组成:(1) 查询信息 q 与文档 d 的相似度 Score(q,d),由搜索引擎 terrier 使用 BM 模型得到;(2) 查询信息 q 与文档 d 的标注信息 T_d 的相似度 sim(q,T_d),采用的是基于词频的向量之间的相似度;(3) 扩展用户标注信息 $P_u^{'}$ 与文档对于用户的个性化属性 $P_{u,d}$ 的相似度 Sim($P_u^{'}$,$P_{u,d}$),由用户扩展后的所有标注信息构成的词频向量与用户单个文档扩充的标注信息组成的词频向量之间的相似度所得.如式(1)所示.

$$\mathrm{Rank}(d,q,u) = \alpha \mathrm{Sim}(P_u^{'},P_{u,d}) + (1-\alpha)\left[\beta \mathrm{sim}(q,T_d) + (1-\beta)\,\mathrm{Score}(q,d)\right]. \tag{1}$$

3　结合 Word2vec 和用户偏好模型的个性化排序

3.1　基于 Word2vec 的检索得分计算

在 D-PR 中匹配查询和文档标签集得分时,所采用的方法是简单的词频向量空间模型,这种方法在很多情况下无法精确计算出查询和文档标签集之间的匹配得分.为解决这一问题,本文采用词向量模型,挖掘出词语更深层次的表现形式,而后进行得分计算,具体做法如下:

1)利用 Word2Vec 模型对维基百科语料库进行训练,得到词向量模型 V.

2)然后使用词向量模型将查询信息 q 和文档标注信息 T_{d_i} 转化为向量形式,得到每个标签的词向量 $\boldsymbol{V}_t$.

3)用 averge-pooling 的采样方法[23]进行采样,得到查询信息的向量表现形式 V_q 和文档标注信息表现形式 V_{T_d},具体计算公式为

$$V_q = \frac{1}{|q|}\sum_{t\in q}\boldsymbol{V}_t, V_{T_d} = \frac{1}{|\boldsymbol{T}_d|}\sum_{t\in \boldsymbol{T}_d}\boldsymbol{V}_t. \tag{2}$$

式中:q 为查询信息;$\boldsymbol{T}_d$ 表示文档 d 上的标签集合;$\boldsymbol{V}$ 是由外部语料库训练得到的词向量模型;V_t 表示标签 t 映射到词向量模型中的向量形式;$|q|$ 表示查询信息中标签的个数;$|\boldsymbol{T}_d|$ 表示文档 d 上标注信息中标签的个数.

4)用余弦相似度计算查询信息和文档标注信息之间的相似度,公式为

$$\mathrm{sim}(q,\boldsymbol{T}_d) = \mathrm{Cos}\,(V_q,V_{T_d}). \tag{3}$$

式中:Cos(*)表示常规余弦相似度的计算.

3.2　融合用户兴趣偏好的文档加权

考虑到不同用户对不同类型的文档偏好程度不一样,而大部分研究只是将个性化信息与内容检索得分简单的线性拟合,并没有在内容检索部分融入个性化信息,这在一定程度上无法反映出二者之间的内在联系问题.为此,本文利用用户的社会化标注信息和标签之间的相似性构建用户兴趣偏好模型,再根据用户对检索排序过程中文档的不同加权值,以加强文档个性化处理,具体步骤为

1) 对文档集进行分类,将每个类别中的文档上的标注信息组合到一起,形成一个类别标注信息集 $\boldsymbol{T}_{C_i}$,其中文档集分类的方法为

采用 LDA[24]对文档集进行建模,利用 Gibbs 采样算法求解 LDA,得到每个文档下的主题分布 θ_m 和每个主题下的词的分布 φ_j,关于 LDA 模型 Gibbs 采样算法已有很多文献阐述过[24-25],在此不做过多介绍,第 m 篇文章中主题 j 的概率 $\theta_{m,j}$ 和第 m 篇文章中词汇 term 在主题 j 中的概率 $\varphi_{j,\mathrm{term}}$ 的向量权重估计值为

$$\theta_{m,j} = \frac{n_m^j + \alpha_j}{\sum_{j=1}^{k} n_m^j + \alpha_j}, \varphi_{j,\mathrm{term}} = \frac{n_j^{(\mathrm{term})} + \beta_t}{\sum_{\mathrm{term}=1}^{v} n_j^{(\mathrm{term})} + \beta_t}. \tag{4}$$

式中:n_m^j 为第 m 篇文章中主题 j 出现的次数;$n_j^{(\mathrm{term})}$ 为第 m 篇文章中词汇 term 属于主题 j 的次数;k 为隐含主题的个数;v 为文档中词汇的个数;α_j 和 β_t 为对应的狄利克雷分布超参数.

得到文档主题分布 θ_m 后,将其作为聚类的文本特征,然后使用 K-means 聚类算法[26]对文档特征向量进行聚类,最后得到 K 个类别.

2) 基于标签共现原则,采用 Jaccard 相似系数计算标签之间的相似度,构建相似标签集,公式为

$$\mathrm{t\_sim}(t_i,t_j) = \frac{|\boldsymbol{D}(t_i)\cap\boldsymbol{D}(t_j)|}{|\boldsymbol{D}(t_i)\cup\boldsymbol{D}(t_j)|}. \tag{5}$$

式中:$\boldsymbol{D}(t)$ 表示标签 t 标注过的文档集.

3) 用户标注信息集和类别标注信息集中标签权重赋值.将每个用户的标注信息表示为 $\boldsymbol{P}_u = (w_{t_1}^u, w_{t_2}^u, \cdots, w_{t_s}^u)$,每个兴趣类别的标注消息表示为 $p_{C_i} = (w_{t_1}^{C_i}, w_{t_2}^{C_i}, \cdots, w_{t_s}^{C_i})$,向量的权重利用文献[5]中修改的 TF-IDF 来计算,公式为

$$w_{t_i}^u = utf_{t_i} \times \log\left(\frac{|\boldsymbol{U}|}{|\boldsymbol{U}_*^{t_i}|}\right),\ w_{t_i}^{C_i} = ctf_{t_i} \times \log\left(\frac{K}{|\boldsymbol{C}_*^{t_i}|}\right). \tag{6}$$

式中:$w_{t_i}^u$ 为用户 u 对标签 t_i 的偏好权重;$w_t^{C_i}$ 为兴趣类别 C_i 中标签 t_i 的偏好权重;utf_{t_i} 为用户 u 使用标签 t_i 的次数;$|\boldsymbol{U}|$ 为用户的总个数;$|\boldsymbol{U}_*^{t_i}|$ 为使用了标签 t_i 的用户个数;K 为兴趣类别的个数;ctf_{t_i} 为标签 t_i 在兴趣类别 C_i 中出现的次数;$|\boldsymbol{C}_*^{t_i}|$ 为含有标签 t 的兴趣类别个数.

4) 构建用户兴趣偏好模型,计算用户对每个文档类别的偏好.社会化标注作为一种人为的主观性行为,不仅可以视为用户的配置文件信息,还可以看作是对文档内容的精确提炼,因此挖掘用户标签集与文档标签集之间的内在联系,在一定程度上可以得到该用户对该文档的偏好权重.本文将用户的兴趣模型表示为 $\boldsymbol{Z}_u = (w_{u,C_1}, w_{u,C_2}, \cdots, w_{u,C_n})$,$w_{u,C_i}$ 表示用户 u 对第 i 个兴趣类别 C_i 的偏好程度.为得到用户对类别 C_i 的偏好权重 w_{u,C_i},基于文献[27]中计算用户兴趣偏好的思想,计算分 2 步进行:

*Step*1:首先计算用户 u 对类别 C_i 中的各个标签 t 的偏好,然后将该类别中所有标签的偏好权重进行汇总,得到用户 u 对类别 C_i 的偏好,计算公式为

$$w'_{u,C_i} = \sum_{t \in \boldsymbol{T}_u \cap \boldsymbol{T}_{C_i}} w_t^u\, w_t^{C_i}. \tag{7}$$

式中:w_t^u 为用户 u 的标签集中标签 t 的权重;$w_t^{C_i}$ 为兴趣类别 C_i 的标签集中标签 t 的权重.

*Step*2:由于用户标注信息存在稀疏性的问题,会使得式(7)计算不够准确,为解决这一问题,利用标签相似网络对计算框架进行扩充,扩充部分计算公式为

$$w''_{u,C_i} = \sum_{t \in \boldsymbol{T}_u \cap \boldsymbol{T}_{C_i}} w_t^u \Big(\sum_{t' \in \boldsymbol{S}_t \text{and } t \neq t'} w_{t'}^{C_i} \times t\_\mathrm{sim}(t, t') \Big). \tag{8}$$

式中:$\boldsymbol{T}_u$ 为用户 u 标注的标签集合;$\boldsymbol{T}_{C_i}$ 为主题域 C_i 中标签集合;t' 为标签 t 的相似标签;$\boldsymbol{S}_t$ 为标签 t 的相似标签集.

最终用户 u 对兴趣类别 C_i 的兴趣偏好权重 w_{u,C_i} 的计算方法为

$$w_{u,C_i} = w'_{u,C_i} + w''_{u,C_i}. \tag{9}$$

图 1 为计算用户 u 对每个文档类别偏好的示例图.用户 u 标签集中含有红色标签(偏好为 0.13),类别 C_1 标签集中也有红色标签(偏好为 0.11),因而计算用户 u 对类别 C_1 中红色标签的偏好为 0.13×0.11=0.014 3.用户 u 标签集中有浅蓝色标签(偏好为 0.27),并没有包含深蓝色标签,但类别 C_1 标签集中有深蓝色标签(偏好为0.26),而浅蓝色和深蓝色标签相似度为 0.82,所以用户 u 对类别 C_1 中深蓝色标签的偏好为 0.27×0.26×0.82=0.057 6.最后,综合用户 u 对类别 C_1 中所有标签偏好权重,得到其对类别 C_1 的偏好.

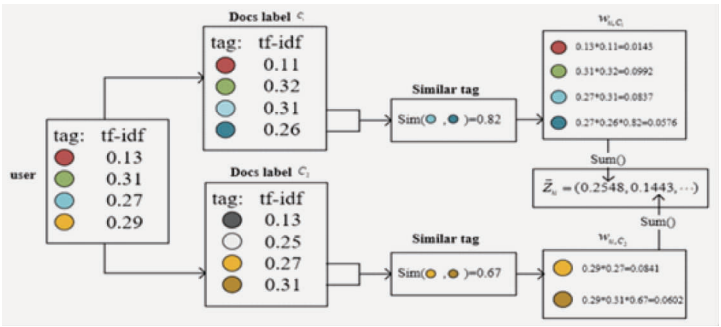

图 1 用户对文档类别偏好计算示例

5) 利用上一步得到的用户对文档类别的偏好程度,对不同文档加权,以加强文档的个性化处理:

$$\mathrm{Rank1}(q,d) = \underset{(d\in C_i)}{w_{u,C_i}}[\beta \cdot \mathrm{Cos}(\boldsymbol{V}_q, V_{T_d}) + (1-\beta)\ \mathrm{Score}(q,d)\ . \tag{10}$$

式中: $\underset{(d\in C_i)}{w_{u,C_i}}$ 为用户检索到文档 d 时,对文档 d 所属于的兴趣类别 C_i 的偏好权重 w_{u,z_i}.

3.3 PRIWA 方法

D-PR 方法经过前 2 步的处理后,即为本文所提出的个性化排序方法 PRIWA,计算排序得分公式为

$$\mathrm{Rank}(d,q,u) = \mathrm{Sim}(P_u^{'}, P_{u,d}, P_{u,d}) + (1-\alpha)\ \mathrm{Rank1}(q,d). \tag{11}$$

式中: $p_u^{'}$ 为用户的 u 的扩展属性,计算公式为

$$p_u^{'} = \sum_{d\in D_u} p_{u,d}. \tag{12}$$

式中: $P_{u,d}$ 为文档 d 对于用户 u 的个性化文档属性,其基本思想是对于每一篇文档采用基于用户的协同过滤算法扩充用户标注信息,公式为

$$p_{u,d} = \sum_{i=1}^{|\boldsymbol{U}_d \cap \boldsymbol{S}_u|} (\boldsymbol{v}_{u_i,d} \cdot u\_\mathrm{sim}(u,u_i))\ . \tag{13}$$

式中: $\boldsymbol{v}_{u_i,d}$ 为用户 u_i 对文档 d 的标签经词频统计得到的向量; $\boldsymbol{U}_d$ 表示标注过文档 d 的用户集合; $u\_\mathrm{sim}(u,u_i)$ 为用户 u 和 u_i 的相似度(基于用户的标注信息集); $\boldsymbol{S}_u$ 表示与用户 u 的相似的用户集.

4 实验评估

4.1 数据集

实验评估的数据采用公开数据集 socialbm0311 和 deliciousT140 通过网址匹配得到.实验中对匹配后的数据集进行了以下预处理:(1)清洗标签数据,剔除无意义和非英文标签,同时对标签进行去停用词和词干化处理;(2)清洗网页数据,对网页内容进行解析,去停用词和词干化处理后得到相应的文档;(3)对于标签出现次数过低的标签,没有过多的代表性,在实验中将其剔除.而且标签太少的用户,体现不了用户的兴趣偏好,标签太多的用户没有代表性,基于标签个数的累积分布图,在实验中,取标签个数多余 5 个少于 200 个的用户.最后得到数据集中包含 121 184 个用户(Users),29 131 272 个标注词(Tags),116 855 篇文档(Docs).

4.2 评估方法

虽然个性化搜索结果的相关性判断主观上取决于终端用户,但是一些研究[4]已经证明,用户在社交网络上的标签行为与其在线搜索行为密切相关,如果文档由具有标签的用户注释,则该文档很可能被同一用户访问.这一发现为本文的自动评估框架提供了理论基础:如果某个查询是由某个用户发布的,那么相关的文档就是这个用户使用与查询相同的标签标注过的文档.

因此,为了生成 1 组合成用户查询,实验时从数据集中随机选择一组书签,对于每个书签 (u,t,d),创建一个查询 $q=t$,它由用户 u 发布,目的是查找文档 d,然后删除所有选定书签,以减小在使用标签 t 作为查询词时,文档 d 对结果的影响.此外,为了减少删除书签的影响,本文每次只随机创建 100 个合成用户查询,独立进行 10 次评估,然后报告平均结果.本文实验采用的评价标准如下:(1)归一化折损累积增益(Normalized Discounted Cumulative Gain,NDCG):该评价方法不仅考虑二值评分(即相关或者不相关),而且考虑针对查询结果的等级评分;(2)平均倒数排名(Mean Reciprocal Rank,MRR):查询结果的倒数排名是第一个相关文档的倒数积,该排名通常被用来评价结果重排序.两种评价方法具体计算过程可参考[28].

4.3 实验结果和分析

为了评估本文提出的改进后的个性化排序方法的有效性及合理性,采用 SoPRa[6] 和 D-PR[7] 两种方法作为基线方法,进行重复对比实验并对结果进行统计分析.实验中,在构建主题域模型时,因为文档集很大,本文设置的隐语义主题 k 的取值范围为[50,200],递增区间长度为 25,当获得最优模型后,对文档-主题矩阵表示的文档特征,使用 K-means 模型对文档进行聚类,设定 K 的范围为[55,150],递增区间长度为 15.训练后,最优参数 k 的取值为 125,参数 K 的取值为 85,词向量维度 m 的取值为 150.

本文的实验结果如图 2 所示.从图 2 的结果可以看出:(1)本文提出的方法可以有效地提高个性化检

索效果,在 α 取大部分值的情况下效果都优于这 2 种基线方法,这说明考察词义深度向量形式的得分比单纯的内容匹配得到的结果更准确,而加入用户偏好权重之后的提升,则说明在个性化检索中用户偏好对于检索质量有较大影响;(2)相较于 Baseline 中较好的 D-PR,本文提出的改进点可以发现在最大值上表现出更好的结果,当 α 从 0.3 到 0.6 变化时,效果提升是比较大的,其中 α = 0.6 时,PRIWuP 提升值为 0.086 1(MRR),提升了 36.1%和 0.106 8 (NDCG@10),提升了 39.5% ;(3)当 α 取值大于 0.9 时,结果都呈现出下降的趋势,说明过度依赖用户配置文件结果也不理想,内容匹配部分同样不可忽视;(4)就整体而言,本文提出的 PRIWuP 方法和 D-PR 方法都优于 SoPRa 方法,说明基于扩展社会化标注的方法也可以一定程度上减少数据稀疏性带来的检索不精确的问题.

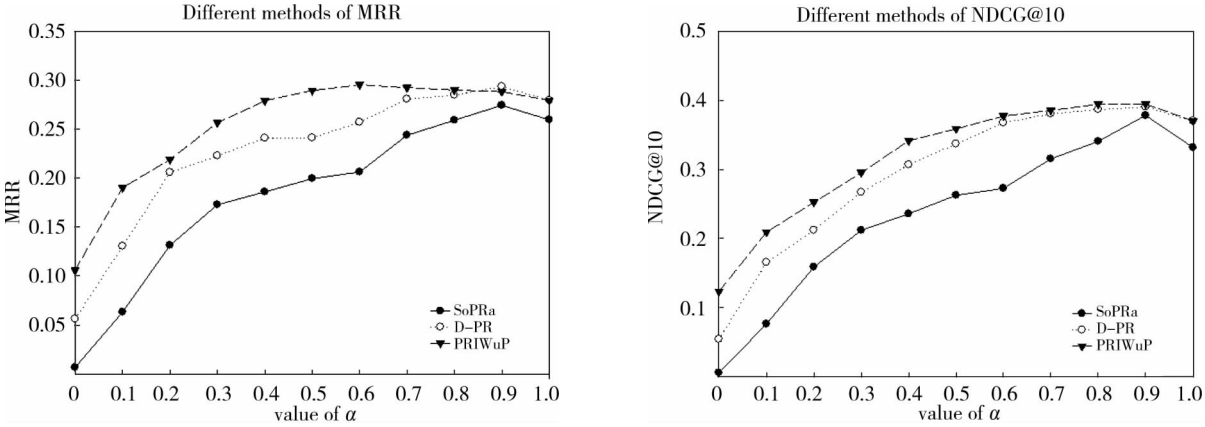

图 2 实验结果对比

为进一步说明本文提出改进方法的有效性和通用性,本文还进行了如下实验:单独在基线方法 D-PR 上利用词向量技术计算查询词和文档的语义相似度(记作 D-PR_W2V);单独在基线方法 D-PR 的基础上利用用户对文档的偏好程度对文档加权(记作 D-PR_uP);单独在基线方法 SoPRa 上利用词向量技术计算查询词和文档的语义相似度(记作 SoPRa_W2V);单独在基线方法 SoPRa 的基础上利用用户对文档的偏好程度对文档加权(记作 SoPRa_uP);在基线方法 SoPRa 上结合 Word2vec 和用户兴趣偏好的个性化排序方法(记作 SoPRa_WuP).结果如图 3 和图 4 所示.

从图 3 和图 4 中可以看出:(1)相较于基线系统,无论是使用词向量技术还是对文档按用户偏好程度对文档进行加权,对个性化的检索结果都有一定程度的提高,说明了本文方法的合理性;(2)在使用词向量技术和对文档加权的两种方法中,评价指标的值是交替的,说明了这两种方法没有绝对的谁优谁劣,都是可行的;(3)在 α 取值大于 0.5 时,对文档按用户偏好对文档进行加权的方法效果在更多情况下比使用词向量技术更好一点,说明了融合用户偏好信息的方法在个性化排序中是非常关键的一环;(4)从结果中可以发现,两个方法在一起使用时,检索效果是最佳的,说明了本文提出的两个改进的地方是相辅相成的.

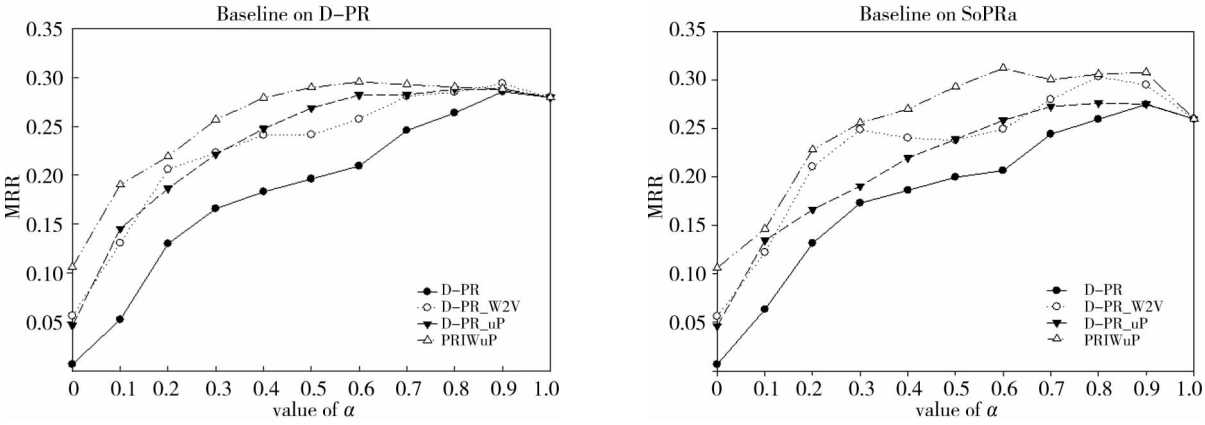

图 3 MRR 评价指标的对比实验

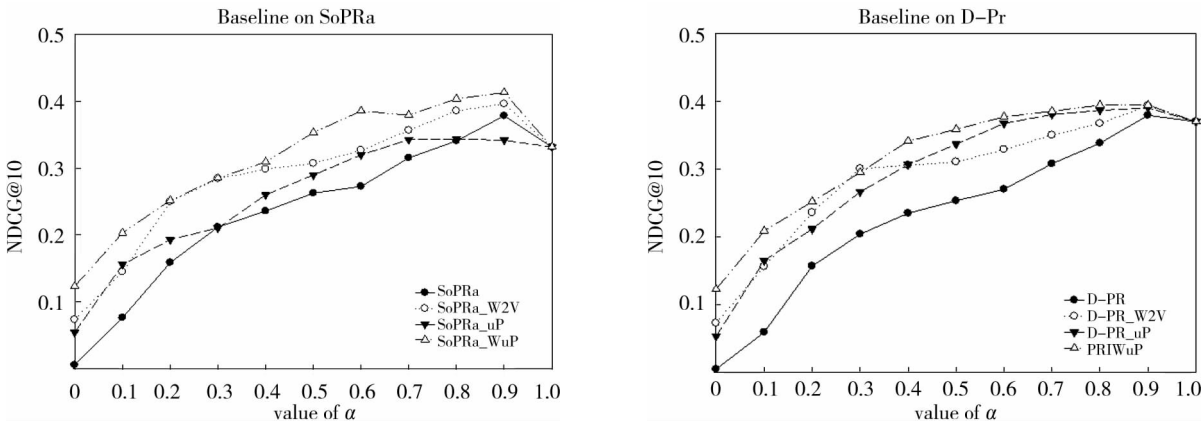

图 4　NDCG@ 10 评价指标的对比实验

综上所述,实验结果一定程度上说明了本文提出的 PWIuP 方法对于提升个性化检索结果的质量有一定的效果.

5　结论

1)利用用户共同标注过的文档构建用户相似网络,之后采用协同过滤的思想对用户的标注信息进行扩充,从而使得用户能更多地表现出对于文档的兴趣信息.

2)利用词向量技术将标签词映射到向量空间中,使得意思相近的词语之间向量距离更短,进而改进数据稀疏时查询词与文档标注信息匹配不精确的问题,使得最终的个性化文档得分更准确.

3)利用偏好信息对查询与文档整体内容检索部分进行加权改进,以增大检索结果的个性化特征.

最终实验结果表明,本文所提的几个改进之处是相互独立,且在作用上是相辅相成的,都可以对个性化检索效果进行优化,可以改善用户的检索体验.同时本文并未考虑用户消极标签所带来的负面影响,在后续工作中,会对用户标签质量进行提取分析,使得本文提出的个性化排序方法能更好地满足用户需求.

参考文献:

[1] Manning C D, Raghavan P, Schütze H. 信息检索导论[M].王斌,译.北京:人民邮电出版社, 2010.

[2] Zhang Y, Jansen B J, Spink A. Time series analysis of a Web search engine transaction log[J]. Information Processing & Management, 2009, 45(2):230-245.

[3] Jiang Y, Lv M, Sun J, et al. A Bayesian personalized ranking algorithmbased on Tag Preference[C]//2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE, 2018: 465-471.

[4] Krause B, Hotho A, Stumme G. A comparison of social bookmarking with traditional search[C]//European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2008: 101-113.

[5] Xu S, Bao S, Fei B, et al. Exploring folksonomy for personalized search[C]//International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2008:155-162.

[6] Hacid H, Bouzeghoub M. Sopra: a new social personalized ranking function for improving web search[C]//International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013: 861-864.

[7] Xu Z, Lukasiewicz T, Tifrea-Marciuska O. Improving personalized search on the social web based on similarities between users [C]//International Conference on Scalable Uncertainty Management. Springer, Cham, 2014: 306-319.

[8] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in Vector space [J]. Computer Science, 2013.

[9] 王晓春, 李生, 杨沐昀,等. 一种长短期兴趣结合的个性化检索模型[J]. 中文信息学报, 2016, 30(3):172-177.

[10] Li Z C, Chen F, Xing Q L, et al. Thuir at trec 2009 web track: Finding relevant and diverse results for large scale web search [R]. TSINGHUA UNIV BEIJING (CHINA) NATIONAL LAB FOR INFORMATION SCIENCE AND TECHNOLOGY, 2009: 145-153.

[11] Dou Z, Hu S, Chen K, et al. Multi-dimensional search result diversification [J]. Ge Portuguese Journal of Gastroenterology, 2011(6): 475-484.

[12] Tu N N, Kanhabua N. Leveraging dynamic query subtopics for time-aware search result diversification[C]//European Conference on Information Retrieval. Springer International Publishing, 2014: 222-234.

[13] Hacid H, Bouzeghoub M, et al. Using social annotations to enhance document representation for personalized search[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013: 1049-1052.

[14] Zhao T, McAuley J, King I. Leveraging social connections to improve personalized rankingfor collaborative filtering[C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014: 261-270.

[15] 李鹏, 王斌, 晋薇. 一种基于社会化标签的信息检索方法[J]. 中文信息学报, 2013, 27(1):39-46.

[16] Wang Q, Jin H. Exploring online social activities for adaptive search personalization[C]//ACM International Conference on Information & Knowledge Management. ACM, 2010: 999-1008.

[17] GaoY, Wang M, Zha Z J, et al. Visual-textual joint relevance learning for tag-based social image search[J]. IEEE Transactions on Image Processing, 2013, 22(1): 363-376.

[18] Li J, Liu C, Yu J X, et al. Personalized influential topic search via social network summarization[J]. IEEETransactions on Knowledge and Data Engineering, 2016, 28(7): 1820-1834.

[19] 王庆林, 薛惠锋, 林波. 基于图聚类的协同标注系统资源个性推荐[J]. 计算机工程与应用, 2010, 46(11): 10-13.

[20] Ji A T, Yeon C, Kim H N, et al. Collaborative tagging in recommender systems [C]//Proceedings of 20th Australian Joint Conference on Artificial Intelligence. Berlin: Springer-Verlag, 2007: 377-386.

[21] Zhou D , Séamus L, Wade V. Improving search via personalized query expansion using social media[J]. Information Retrieval, 2012, 15(3/4):218-242.

[22] 管毅舟, 徐博, 林原,等. 基于社会化标注和网页分类的个性化检索方法[J]. 山东大学学报(理学版), 2016, 51(7): 35-42.

[23] Xu J, Mei T, Yao T, et al. Msr-vtt: A large video description dataset for bridging video and language[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5288-5296.

[24] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3): 993-1022.

[25] Papanikolaou Y, Foulds J R, Rubin T N, et al. Dense distributions from sparse samples: improved Gibbs sampling parameter estimators for LDA[J]. Journal of Machine Learning Research, 2017, 18(1): 2058-2115.

[26] Botía J A, Vandrovcova J, Forabosco P, et al. An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks[J]. BMC systems biology, 2017, 11(1): 47.

[27] 赵海燕,郭娣,陈庆奎,等.一种融合相似网络的多主题域混合推荐算法[J].计算机应用研究,2015,32(10): 2901-2904.

[28] Baezayates-Yates R, Ribeiro-Neto B. Modern information retrieval: the concepts and technology behind search [M]. New Jersey: Pearson, 2010.