DOI:10.13196/j.cims.2018.07.025

结合文档处理与查询处理技术的 Web 服务搜索方法

赵文玉,周 栋+,曹步清,刘建勋
(湖南科技大学 计算机科学与工程学院,湖南 湘潭 411201)

摘 要:为了提高 Web 服务搜索结果的准确性和改善用户的查询体验,使用文档处理技术与查询处理技术一直是 Web 服务搜索研究中的热点课题。为此,本文提出一种结合两种技术的 Web 服务搜索方法。考虑文档与文档之间的关系,通过正则化框架,融合第一轮检索结果分数,实现搜索结果列表中文档重排序;基于排名靠前文档构建扩展词集合,从构建扩展词集合中选取候选扩展词进行查询词扩展。在 NTCIR 数据集上的实验结果表明,该方法能够有效提高 Web 服务搜索中排名顶端结果的准确率。

关键词:Web 服务搜索;文档处理技术;查询词处理技术;正则化框架;扩展词集合
中图分类号:TP391 **文献标识码**:A

Approach integrating document processing with query processing technique for Web service search

ZHAO Wenyu, *ZHOU Dong*+, *CAO Buqing*, *LIU Jianxun*
(School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China)

Abstract: To improve the accuracy of Web service and the query experience of users, document processing technique and query processing technique are hot research topic in Web service search. For this reason, an approach which integrated document processing technique and query processing technique for Web service search was proposed. By considering the relationships between documents and the normalized framework, the re-rank documents in searching list based on the initial results was realized. The candidate terms from expansion terms set based on top documents were selected to expand the query. The experiments were conducted on NTCIR dataset, and the results showed that the proposed method could better improve precision for top results.

Keywords: Web service search; document processing technique; query processing technique; normalized framework; expansion terms set

---

收稿日期:2017-12-20;**修订日期**:2018-06-02。Received 20 Dec. 2017; accepted 02 June 2018.
基金项目:国家自然科学基金资助项目(61300129,61572187);教育部留学回国人员科研启动基金资助项目(教外司留[2013]1792);湖南省教育厅资助项目(16K030);湖南省教育厅创新平台开放基金项目(17K033);湖南省自然科学基金项目(2017JJ2101,2018JJ2139,2017JJ2098);湖南省研究生科研创新资助项目(CX2016B575);湖南科技大学科学研究基金项目(KJ1839)。**Foundation items**: Project supported by the National Natural Science Foundation, China (No. 61300129, 61572187), the Scientific Research Foundation for the Re-turned Overseas Chinese Scholars of Ministry of Education, China (No. [2013]1792), the Scientific Research Fund of Hunan Provincial Education Department, China (No. 16K030), the Innovation Platform Open Foundation of Hunan Provincial Education Department, China (No. 17K033), the Hunan Provincial Natural Science Foundation, China (No. 2017JJ2101, 2018JJ2139, 2017JJ2098), the Hunan Provincial Innovation Foundation For Postgraduate, China (No. CX2016B575), and the Research Foundation of Hunan University of Science and Technology, China (No. KJ1839).

0　引言

随着 Web 2.0 的爆炸式发展,Web 服务的数量也呈指数增长,如何帮助用户快速高效地在海量数据中搜索到其所需的 Web 服务是 Web 服务研究中的热点[1]。在 Web 服务开发过程中,每个 Web 服务都有与其对应的文档描述信息(如 Web 服务简介或使用说明等),称为 Web 服务文档[2]。Web 服务搜索类似于信息检索系统[3],其目标在于针对用户的特定需求,实时为用户返回一组 Web 服务描述文档,该组结果按照与 Web 服务需求的相关性从大到小排序。由于用户进行 Web 服务搜索通常是基于关键词查询,关键词与 Web 服务的描述文档之间会存在语义信息不匹配的问题,导致返回大量与用户不相关的服务,查询效果很不理想。为此,通常采用文档处理和查询处理两种技术进行改进。在文档处理技术中常采用文档重排序[4-5],文档重排序方法是在第一轮 Web 服务搜索的基础上,对初始的结果进行重新排序;在查询处理技术中常采用查询扩展[6-8],查询扩展是对原始的查询词进行扩展,用于进行第二轮检索。这两种技术都比较成熟,但是较少有研究人员将二者结合,来提高 Web 服务搜索结果中顶端结果的相关性和伪相关反馈的准确率。

在以往关于文档处理技术的研究中,研究者们充分利用文档间的距离关系[4]、文档间不对称的信息关系[9-11]进行文档重新排序;在以往关于查询处理技术的研究中,Kuzi 等[12]通过词向量为查询词重新分配权重来扩展查询词。少量研究在第一轮检索结果的基础之上,利用伪相关反馈方法[13]、统计关系[14]实现文档处理技术与查询处理技术的结合。

上述将文档处理技术与查询处理技术相结合的研究方法比较简单,存在一定的局限性。本文提出一种结合两种技术实现 Web 服务搜索的方法。该方法使用文档重排技术对第一轮检索结果进行重排序处理,然后从排名靠前文档中选取扩展词进行查询词扩展,使用扩展后的查询词对 Web 服务进行搜索,可以有效提高 Web 服务搜索顶端结果的准确率。

1　相关工作

通常情况下,Web 服务的数量非常庞大,用户需要从海量服务中搜索与其需求相关的 Web 服务。目前关于 Web 服务搜索的研究较多,例如方启明等[15]提出一种基于 P2P 实现 Web 搜索的方法;李伟平等[16]提出一种基于语义的 Web 服务搜索与匹配算法,通过引入功能性与非功能性属性进行服务匹配,可以克服关键字匹配查全率过低的问题。类似地,张桂刚[17]提出一种类自然语言驱动的语义 Web 服务搜索方法;张以文等[18]联合用户的兴趣矩阵和全局偏好为用户进行 Web 服务推荐。因为 Web 服务搜索类似于信息检索系统,所以 Web 服务搜索的本质为信息检索中的查询问题。目前有关查询问题的研究已经比较成熟,这些研究根据处理技术差异性可分为两类:①文档处理技术中的文档重排序[4-5],它根据用户特定的信息需求将结果列表进行重排序,使结果列表中的顶端结果更相关;②查询处理技术中的查询扩展[6-8],它通过从排名靠前文档中抽取与原始查询词相似的词项扩展原始查询词。

根据使用的不同信息资源将文档重排序研究大致分为 4 类:①集中于使用文档间的关系进行检索结果的重排序,例如 Balinski 等[4]基于文本或者超链接不同文档间的距离,再根据这些距离修改最初的相关性权重,实现检索结果重新排序;Plansangke 等[5]根据文档与查询词之间的相关性对文档进行分类,再基于文档分类重新对文档进行排序。②借助外部资源实现文档重排序的目的,如手工词典、受控词汇等外部资源[19-20]。③根据文档和查询词中抽取的特定信息实现文档重排序,例如 Luk 等[21]利用文档标题与查询词标题的信息对文档进行重排序[21]。④基于图分析文档集内部的关系进行搜索结果重排序,例如文献[9-11]利用文档间的相似度建立文档图,然后基于加权的 PageRank 或 HITS 算法,找出图的中心节点,重新计算文档与查询词之间的相似度,从而实现文档的重排序。此外,Yang 等[22]考虑不同文档之间的关系,提出一种基于半监督学习文档重排序的方法;Zhang 等[23]利用改进的向量空间模型计算文档链接之间的相似度,构造文档关系排序图,成功实现文档重排序;Zhou 等[24]在初始检索结果的基础上,利用 LDA(latent dirichlet allocation)模型[25]重新对文档和查询词之间的关系进行建模,实现文档重排序。

在查询处理技术中,根据语料库中词汇之间的关系将查询扩展研究分为两类:①只考虑语料库中词汇之间的统计关系,例如 Zhou 等[5]提出一个查询词扩展的框架,该框架从标签—主题的用户模型

中选取候选扩展词项；②考虑目标词汇的上下文语境，例如 Kuzi 等[12]为语料库中的词汇建立词向量，再挑选与原始查询词语义相关的词项进行查询词扩展。在 Web 服务搜索中，黄瑞等提出一种新的 Web 服务搜索方法，该方法基于 Web 异构，通过语义相关的概率、关键词和语义单元信息进行 Web 搜索中语义的查询扩展[8]。

在信息检索中，文档重排序与查询扩展技术的发展日趋成熟，但是文档重排序与查询扩展技术相结合的研究却比较少。Bouchoucha 等[13]将第一轮检索结果中的顶端结果作为相关文档，利用伪相关反馈方法从相关文档中产生候选扩展词项；Mine 等[14]在搜索顶端结果的基础上，统计文档中句子中的词频进行查询词扩展。

2 结合文档处理与查询处理技术的 Web 服务搜索方法

2.1 问题定义

在 Web 服务搜索中，通常一个 Web 服务由一个简短的描述文档对该服务提供的服务进行描述，描述文档可表示为 d，Web 服务集合可表示为 $D=\{d_1,d_2,\cdots,d_n\}$。在对某一 Web 服务集合进行搜索的过程中，由于用户的 Web 服务需求（查询词）q 与 Web 服务的描述文档之间存在语义匹配不准确的问题，导致用户第一轮搜索结果 y 不准确。为提高 Web 服务搜索顶端结果的准确率，本文利用文档处理技术重排序结果列表中的文档，产生新结果列表 f。基于列表 f，构建扩展词集合，再从扩展词集合中挑选候选扩展词进行查询词扩展，使用扩展后的查询词进行二次检索。在文档处理技术上结合查询处理技术，不仅可以从文档重排序的 Web 服务文档中挑选与用户信息需求相关的词项，提高查询处理技术的性能，还可以利用文档处理技术避免查询扩展技术可能引入的噪声。本文使用的常用符号及其含义说明如表 1 所示。

表 1 符号及含义说明

符号	含义说明
q	查询词
w	词项
d	文档
n	语料库中文档的数量
m	查询词 q 中词项的数量
D	包含 n 个文档的文档集
f	第一轮检索返回的文档结果列表
x_i	第 i 个文档的空间向量
y	重新排序后的文档结果列表
$\boldsymbol{A}$	$n\times n$ 维的文档—文档矩阵，表示文档—文档之间的关系
$\boldsymbol{D}$	表示文档—文档之间关系的对角矩阵
$\boldsymbol{S}$	表示文档—文档之间关系的归一化矩阵
K	主题的数量
α	先验参数
β	先验参数
θ	表示主题的多项分布
φ	表示词的多项分布
w_j	第 j 个文档中的第 i 个词汇
z_j	第 j 个文档中第 i 个词汇相关的主题
di	词向量的维度
N_c	第 j 个文档中词汇的数量
$\overline{w}$	词项 w 对应的词向量
$\boldsymbol{f}_j^e$	第 j 个文档中第 i 个词汇对应向量维度为 e 的词向量
μ_z	主题 z 下检索分数对应正态分布中的平均值
σ_z	主题 z 下检索分数对应正态分布中的方差值
n_j	第 j 个文档中抽样主题为 k 的次数
v_k	主题为 k 时生成词汇 $w_{j,i}$ 的次数
w	查询词 q 中的词项
d_r	相关文档的集合
N	相关文档集合中文档的数量
η	文档重排序后结果列表中词项逆文档频率分数值最高的数量
τ	候选词的数量

2.2 文档处理技术

本节主要介绍文档重排序过程中邻近图的构造，并描述基于正则化框架的文档重排序过程。

2.2.1 邻近图构造

为构造反映 Web 服务描述文档（简称文档）与 Web 服务描述文档之间关系的邻近图，利用向量空间模型计算文档与文档之间的相似度

$$sim(d_i,d_j)=\frac{x_i,x_j}{\|x_i\|\cdot\|x_j\|}。\tag{1}$$

式中：d_i 和 d_j 分别表示第 i 个文档和第 j 个文档，x_i 和 x_j 分别表示第 i 个文档的空间向量和第 j 个文档的空间向量。反映文档与文档关系邻近图的关系矩阵 $\boldsymbol{A}$ 可表示为 $A_{ij}=sim(d_i,d_j)$，其中 A_{ij} 为关系矩阵 $\boldsymbol{A}$ 中的元素。

为方便描述文档与文档之间的关系，定义 1 个对角矩阵 $\boldsymbol{D}_A$，表示相对应矩阵 $\boldsymbol{A}$ 中第行对角线上

的元素,且等于矩阵 $\boldsymbol{A}$ 中第 i 行的总和。此外,为使所有数据处于同一数量级,还定义了归一化矩阵 $S_A=\boldsymbol{D}_A^{-1/2}\boldsymbol{A}\boldsymbol{D}_A^{-1/2}$。

2.2.2　基于正则化框架的文档重排序方法

为更好地捕捉 Web 服务描述文档之间的关系,本文提出以下目标函数融合文档与文档之间的关系,但是在该正则化框架中,需要最小化重排序后文档的分数,并使重排序后的文档结果列表 f 和 $\boldsymbol{A}$,y 的相关信息保持一致。目标函数为

$$Q(f,g)=\frac{1}{2}\mu_1\sum_{i,j=1}^{p}A_{ij}\left(\frac{1}{\sqrt{D_{A_{ii}}}}f_i-\frac{1}{\sqrt{D_{A_{jj}}}}f_j\right)^2+\mu_2\sum_{i=1}^{p}(f_i-y_i)^2。\tag{2}$$

基于正则化框架的文档重排序方法主要是在初始检索结果的基础之上,利用正则化框架融合文档之间的关系,实现初始检索结果列表中文档重排序的目的,具体实施过程如算法 1 所示。

算法 1　基于文档关系正则化模型的文档重排序。

输入:查询词 q。

输出:文档重排序后的文档结果列表 f。

1. 进行第一轮检索,返回第一轮检索结果的文档排名列表 y

2. 构造表示文档关系的矩阵 A,并计算相对应的归一化对角矩阵 S_A

3. 使用式(5)计算文档重新排序后的分数

其中,y_i 表示第 i 个文档的第一轮检索结果,f_i 表示第 i 个文档重排序后的结果。在目标函数中,第一项描述文档与文档之间的关系,第二项控制重排序后的结果与原始结果之间的差异。以上两项之间的平衡由归一化参数 μ_1 和 μ_2 控制,$\mu_1=\frac{1}{1+\mu}$,$\mu_1=\frac{\mu}{1+\mu}$,$0<\mu_1,\mu_2<1$,且 $\mu_1+\mu_2=1$。经过一系列简单推导后[22],第一项可表示为等价的矩阵向量形式 $\boldsymbol{f}^{\mathrm{T}}(1-S_{\boldsymbol{A}})\boldsymbol{f}$,使用相同矩阵向量形式对目标函数进行重写,得

$$Q(f,g)=\mu_1\boldsymbol{f}^{\mathrm{T}}(1-S_{\boldsymbol{A}})\boldsymbol{f}+\mu_2(\boldsymbol{f}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f}-\boldsymbol{y})。\tag{3}$$

目标函数对 f 求偏导:

$$\frac{\partial Q}{\partial f}=(I-\mu_1 S_A)\boldsymbol{f}-\mu_2\boldsymbol{y}=0。\tag{4}$$

根据式(4),得到重新排序后文档的排名分数

$$\boldsymbol{f}=(\boldsymbol{I}-\mu_1 S_{\boldsymbol{A}})^{-1}\mu_2\boldsymbol{y}。\tag{5}$$

2.3　查询处理技术

本节主要在文档排序的基础之上,介绍查询扩展过程中扩展词集合的构造以及个性化查询扩展。

2.3.1　扩展词集合的构造

本文提出的个性化方法包括两步:①文档重排序过程,②扩展词集合的构造。在第一步,首先根据用户的 Web 服务需求(简称查询词)q 进行初次检索,得到与查询词相关的初始文档排名列表 y;然后使用正则化框架融合初始检索结果,实现文档重排序,得到文档重排序后的结果列表 f。在第二步,基于排序后结果列表 f 中的文档构造扩展词集合。下面对扩展词集合构造进行详细的介绍。

众所周知,LDA 模型与词向量(Word Embeddings,WE)都可以用来捕捉文本中的词项语义信息。但是 LDA 模型与词向量存在差异,LDA 模型是基于语料库中词项的统计信息,词向量则是利用上下文的语义信息[26]。本文将 LDA 模型与 WE 式结合,基于文档重排序后结果列表中的文档构造扩展词集合,并将其命名为 WE-LDA 模型。将文档重排序后结果列表中的文档作为 Word2Vec 的输入,得到文档重排序后结果列表中所有词项及其词向量。WE-LDA 模型可以学习隐含的主题,并生成文档重排序后结果列表中构成文档的词项以及对应的词向量。

由于 Skip-Gram 模型通常被用来预测目标词汇的上下文[28],本文使用 Skip-Gram 模型学习词向量每一个目标词汇 $\vec{w}$ 都有相应的向量 $\vec{w}\in\mathbb{R}^{dim}$,其中 dim 表示词向量的维度。目标词汇的向量可以用作一个特征来预测上下文的语境,本文采用词向量的正态分布为文档和词汇推断隐含的主题。使用 Skip-Gram 模型训练词向量以后,WE-LDA 模型的生成过程如算法 2 所示。

算法 2　WE-LDA 模型的生成过程。

输入:查询词 q;

文档重排序后的结果列表 f;

基于查询词和重排序后结果列表中文档使用 Skip-Gram 模型训练得到的词向量。

输出:后验估计 θ_j 和 φ。

1. 对于每一个主题 k∈[1,K],根据狄利克雷分布 Dirichlet(β)可得到该主题下对应的多项词分布向量 φ~Dirichlet(β)

2. 对于每一个文档 $d_j\in f$,根据狄利克雷分布 Dirichlet(α)可得到该文档对应的主题分布向量 θ_j~Dirichlet(α)

3. 对于文档 d_j 中的每一个词汇 w_i:

1)从文档 d_j 对应的多项主题分布向量中 θ_j,为该词抽样一个主题 $z_{j,i}\sim\mathrm{Mult}(\theta_{d_j})$

2)从主题 $z_{j,i}$ 中抽样一个词汇 $w_{j,i}\sim\mathrm{Mult}(\varphi_{z_{j,i}})$

3)对于每一维词 $w_{j,i}$的向量,从正态分布中抽样出词向量 $f^e_{j,i}\sim N(\mu^e_{z_{j,i}},\sigma^e_{z_{j,i}})$

在算法 2 中,$f^e_{j,i}$表示词向量;α 和β 为狄利克雷的先验参数,θ_j 和 φ 是隐含变量的后验分布,μ 和 σ 表示正态分布的平均值和方差值。本文使用吉布斯抽样法抽样隐含变量的特定主题 $z_{j,i}$ 和后验分布[28]。在整个吉布斯抽样过程中,使用词汇以及相对应词向量的隐含主题信息抽样每一个单词的主题,更新规则如下:

$$P(z_{j,i}=k)\alpha\frac{n_{j,k,i}+\alpha}{n_{j,i}+Ka}\times\frac{v_{k,w_{j,i}}+\beta}{v_{k,i}+V\beta}\times\prod_{e=1}^{dim}\frac{1}{\sqrt{2\pi}\sigma_{z_{j,i}}}\exp\left(-\frac{(f^e_{j,i}-\mu_{z_{j,i}})^2}{2\sigma^2_{z_{j,i}}}\right)。\quad(6)$$

式中:对于每一个抽样文档 $d_j\in f$,$n_{j,k,i}$表示主题 k 抽样于文档 d_j 的多项分布的次数,并不包含当前的主题 $z_{j,i}$;$v_{k,w_{j,i}}$表示主题 k 生成词 $w_{j,i}$的次数,也不包含当前的词 $w_{j,i}$。通过吉布斯抽样迭代后,可以得到后验估计 θ 和 φ。图 1 所示为 WE-LDA 模型的物理生成过程。本文所提方法的时间复杂度为 $O(mKL)$,其中 m 为查询词中词项的数量,K 为主题数量,L 为 WE-LDA 模型中词项的数量。

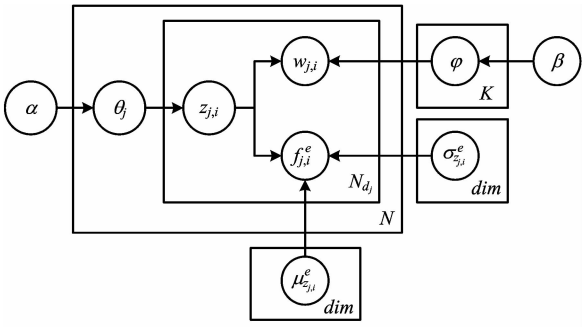

图1 WE-LDA模型

2.3.2 个性化查询扩展

查询扩展的方式类似 Zhou 等[7]提出的个性化扩展方法。假设存在一个隐含的相关文档集合,该集合 $\{d_r\}_{r=1}^N$ 中的文档既与查询词中的词项有相关性,又与扩展词集合中的词项有相关性。此外,假设每一个查询词 q 是由 m 个独立的词项 $\{w'_1,w'_2\cdots w'_m\}$ 组成,查询词 q 生成每一个词 w 的概率可以定义如下:

$$P(w\mid q)=P(w\mid w'_1,w'_2,\cdots,w'_m)\propto\prod_{e=1}^{m}\frac{1}{N}\sum_{r=1}^{N}\left(\sum_{k=1}^{K}P(w\mid Z_k)P(Z_k\mid d_r)\right)\times\left(\sum_{k=1}^{K}P(w_e\mid Z_k)P(Z_k\mid d_r)\right)。\quad(7)$$

根据给定的原始查询词生成每一个词的概率,得到扩展词集合中所有词项的分数排名列表,将前 τ 个词项作为候选词项进行查询词扩展。

3 实验评估

3.1 实验数据

为验证该方法的效果,本文从 *ProgramableWeb*(*http://research.nii.ac.jp/ntcir/index-en.html*)网站上爬取了 12 920 个 *Web* 服务及相应的标签。由于该数据集缺乏查询词和相关性评估标准,本文在实验过程中随机选取了 10 个标签作为查询词进行 *Web* 服务搜索,并邀请实验室 40 位本科生和研究生主观判断搜索结果与训练查询词之间的相关性,生成相关性评估标准文档。该模型在 *Web* 服务数据集上的性能测试结果如表 2 所示。由表 2 结果可知,本文所提方法在 *Web* 服务数据集上有所提高,因为 *Web* 服务数据集具有短文本等特点,所以实验结果提高得不是很明显。

表 2 Web Services 数据集上不同方法的评估结果

	P@5	MRR	NDCG@1
FirstResult	0.158	0.351	0.278
Aff	0.158	0.354	0.281
Structline	0.161	0.345	0.276
Yang	0.160	0.355	0.282
LDA	0.163	0.359	0.284
RRQE	0.165	0.361	0.288

考虑到 Web 服务搜索中的数据集缺乏相关性评估,本文选取经典信息检索的数据集 NTCIR-5,采用 NTCIR-5 CLIR 任务提供的中文单语检索语料库对方法进行验证。中文单语检索语料库中的文档格式与 Web 服务搜索中的数据集相同,却提供了查询词与文档之间的相关性评估文档。本文先使用分词器对所有文档进行分词,去停用词等预处理,然后使用开源软件 Terrier* 为语料库中的文档建立索引。

本文采用的评估方法为 P@5,MRR,NDCG@1,3 种评估的具体计算过程可参考文献[29],给出的结果均为某一用户所有查询词的平均表现。显著差异由配对样本检验测定。

3.2 基线系统

因为近期相关文献所提方法与经典方法的性能

* http://www.terrier.org

相差不大,所以本文使用以下经典基线系统与本文所提方法进行比较:

FirstResult:该方法使用 TF－IDF 模型作为检索模型进行第一轮检索,它是一种没有考虑文档处理技术与查询处理技术的方法。

Aff:该方法是基于文档之间非对称性的内容结构为整个语料库构建邻近图实现文档重排序。

Structline:该方法是基于语言模型对文档间不对称的信息关系进行建模,然后基于加权的 PageRank 算法,找出图的中心节点,以达到文档重排序的目的[9]。

LDA:该方法是基于第一轮检索结果,利用 LDA 模型对第一轮检索结果列表中的文档重新建模,实现文档重排序[9]。

Yang:该方法是利用第一轮检索结果列表中文档与文档之间的关系进行重排序[22]。

RRQE:该方法是本文提出的一种结合文档重排序技术与查询扩展技术的方法,它是在文档重排序的基础之上,然后基于排名靠前的文档构造扩展词集合,从扩展词集合中挑选词项扩展原始查询词。

在文档处理过程中,参数 μ_1 控制 Web 服务描述文档之间的信息重要性,参数 μ_2 控制重排序后的文档分数与第一轮检索结果中文档分数的比值,它们由参数共同设置。在查询处理过程中,α 和 β 为 WE-LDA 模型的先验参数,可分别设置为 $K/50$,0.01,候选词数量 δ 的最佳值为 45,WE-LDA 模型的主题数目 $K=5$,词向量的维度 $dim=50$,文档重排序后结果列表中词项逆文档频率分数值最高的数量参数 $\eta=2$。在其他基线系统中,参数的设置与原文中具有最优表现的参数保持一致。在实验中,查找单个查询词的平均时间为 100 s,完成 50 个查询词的搜索需要 83.3 min,实验的时间耗时与主题数量、查询词中词项数量成正比例关系。

表 3　不同方法的评估结果

	NTCIR-5		
	P@5	MRR	NDCG@1
FirstResult	0.404	0.61	0.5
Aff	0.396	0.622	0.52
Structline	0.356	0.6071	0.5
Yang	0.408	0.6216	0.52
LDA	0.408	0.6576	0.58
RRQE	0.436*	0.684 1*	0.6*

注:* 表示与最好基线系统结果之间的显著差异。

3.3　实验结果

本文方法与其他的基线系统的结果比较如表 3 所示。实验表明,对于不同的评估方法,InitialResult 的效果均最差。在基于图的基线系统中,Aff 与 Structline 的方法都比初始结果 InitialResult 的效果更好,在一定程度上说明基于图的方法可以提高检索效果结果的效率。在考虑 Web 服务描述文档之间关系的结果重排序方法中,LDA 和 Yang 的方法也取得了非常好的效果。但本文提出将文档重排序技术与查询扩展技术结合的搜索方法,经不同评估方法对进行评估,结果表明该方法优于本文其他的基线系统。这可能是由于查询扩展技术是基于重排序后的文档,其与查询词之间的相关性更强,其中包含更多与查询词相似的词项,有助于候选词项的挑选。

由表 3 可得以下结论:①采用不同评价指标时,本文方法优于其他基线系统,且与其他基线系统之间的差异显著,这种差异主要来自文档重排序技术与查询扩展技术相结合。②本文考虑 Web 服务描述文档之间的关系,利用正则化框架融合不同文档之间的关系,达到了文档重新排序的目的,从而在一定程度上提高检索结果列表中顶端结果的相关性。考虑到重排序后的文档与查询词之间的相关性更强,重排序后的文档中包含更多与查询词相似的词项,有助于查询扩展技术的最终效果。③本文提出的 WE-LDA 模型将词向量与 LDA 模型结合,可充分利用二者之间的优势,因此能够更加有效地表达上下文的语义,有利于候选词项的选择。

4　结束语

为提高 Web 服务搜索中排名顶端结果的准确率,本文将文档处理技术与查询处理技术相结合,提出一种结合两种技术的 Web 服务搜索方法。该方法基于文档与文档关系的正则化框架,对初始检索结果列表中的文档重新排序,在重排序文档的基础上使用 WE-LDA 模型构建扩展词集合,从中挑选逆文档频率分数值最高的词项作为候选词进行查询词扩展。在 Web 服务数据集和信息检索数据集上的测试和实验结果表明,该方法可以有效提高 Web 服务搜索的效率及顶端结果的相关性。下一步将考虑改进文档重排序与查询扩展的方法,提高 Web 服务搜索的性能。

参考文献:

[1] LI Shuqing, CUI Beiliang. Web search engine based on personalized information recommendation service: a survey[J]. Journal of Information, 2007(8): 98-101(in Chinese). [李树青, 崔北亮. 基于个性化信息推荐服务的 Web 搜索引擎技术综述[J]. 情报杂志, 2007(8): 98-101.]

[2] WANG Lijie, LI Meng, CAI Sibo, et al. Internet information search based approach to enriching textual descriptions for public Web services[J]. Journal of Software, 2012(6): 1335-1349(in Chinese). [王立杰, 李 萌, 蔡斯博, 等. 基于网络信息搜索的 Web Service 文本描述信息扩充方法[J]. 软件学报, 2012(6): 1335-1349.]

[3] LI Wei, LI Li. Web search engine and full-text retrieval technique[J]. Information Science, 2003(5): 558-560(in Chinese). [李 玮, 李 利. Web 搜索引擎与全文检索技术[J]. 情报科学, 2003(5): 558-560.]

[4] BALINSKI J, DANILOWICZ C. Re-ranking method based on inter-document distances[J]. Information processing & management, 2005, 41(4): 759-775.

[5] PLANSANGKET S, GAN J Q. Re-ranking Google search returned Web documents using document classification scores [J]. Artificial Intelligence Research, 2016, 6(1): 59-68.

[6] ZHOU D, LAWLESS S, WADE V. Improving search via personalized query expansion using social media[J]. Information Retrieval, 2012, 15(3/4): 218-242.

[7] ZHOU D, LAWLESS S, WU X, et al. Enhanced personalized search using social data[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Tokyo, Japan: NII Press, 2016, 700-710.

[8] HUANG Rui, SHI Zhongzhi. A new approach to heterogeneous semantic search on the Web[J]. Journal of Computer Research and Development, 2008(8): 1338-1345(in Chinese). [黄瑞, 史忠植. 一种新的 Web 异构语义信息搜索方法[J]. 计算机研究与发展, 2008(8): 1338-1345.]

[9] KURLAND O, LEE L. PageRank without hyperlinks: structural re-ranking using links induced by language models[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Salvador. New York, N. Y., USA: ACM, 2005: 306-313.

[10] KURLAND O, LEE L. Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Seattle. New York, N. Y., USA: ACM, 2006, 83-90.

[11] KURLAND O. KRIKON E. The opposite of smoothing: a language model approach to ranking query specific document clusters[J]. Journal of Artificial Intelligent Research, 2011, 41: 367-395.

[12] KUZI S, SHTOK A, KURLAND O. Query expansion using word embeddings[C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York, N. Y., USA: ACM, 2016: 1929-1932.

[13] BOUCHOUCHA A, NIE J Y, LIU X. Université de Montréal at the NTCIR-11 IMine Task[C]//Proceedings of the 11th NTCIR Conference. Tokyo, Japan: NTCIR, 2014: 28-35.

[14] MINE S, MATSUMOTO T, YOSHIDA T, et al. Interactive media MINE at the NTCIR-11 IMine search task[C]// Proceedings of the 11th NTCIR Conference. Tokyo, Japan: NTCIR, 2014: 84-87.

[15] FANG Qiming, YANG Guangwen, WU Yongwei, et al. P2P Web search technology[J]. Journal of Software, 2008 (10): 2706-2719(in Chinese). [方启明, 杨广文, 武永卫, 等. 基于 P2P 的 Wleb 搜索技术[J]. 软件学报, 2008 (10): 2706-2719.]

[16] LI Weiping, GAO Fuliang, ZHU Xuwei, et al. Semantic based on service discovery and matching method[J]. Journal of Chinese Computer Systems, 2011(9): 1728-1733(in Chinese). [李伟平, 高福亮, 祝旭巍, 等. 一种基于语义的服务搜索与匹配方法[J]. 小型微型计算机系统, 2011(9): 1728-1733.]

[17] ZHANG Guigang. A kind of semantic service search method driven by natural-like language[J]. Computer Science, 2009 (7): 107-112(in Chinese). [张桂刚. 一种类自然语言驱动的语义服务搜索方法[J]. 计算机科学, 2009(7): 107-112.]

[18] ZHANG Yiwen, AI Xiaofei, CUI Guangming, et al. Recommendation algorithm with user's interest matrix and global preference[J/OL]. Journal of Frontiers of Computer Science and Technology. http://kns.cnki.net/kcms/detail/11.5602.TP.20161207.0922.016.html(in Chinese). [张以文, 艾晓飞, 崔光明, 等. 联合用户兴趣矩阵及全局偏好的推荐算法[J/OL]. 计算机科学与探索. http://kns.cnki.net/kcms/detail/11.5602.TP.20161207.0922.016.html.]

[19] QU Y, XU G. WANG J. Rerank method based on individual thesaurus[C]// Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization. Tokyo, Japan: NII Press, 2001: 1-6.

[20] KAMPS J. Improving retrieval effectiveness by reranking documents based on controlled vocabulary[C]// Proceedings of the 26th European Conference on Information Retrieval. Berlin, Germany: Springer-Verlag, 2004: 283-295.

[21] LUK R W P, WONG K F. Pseudo-relevance feedback and title re-ranking for Chinese information retrieval[C]// Proceedings of the Working Notes of the 4th NTCIR Workshop Meeting Tokyo, Japan: NII Press, 2004: 1-8.

[22] YANG L, JI D, ZHOU G, et al. Document re-ranking using cluster validation and label propagation[C]// Proceedings of the 15th ACM International Conference on Information and Knowledge Management. New York, N. Y., USA: ACM, 2006: 690-697.

[23] ZHANG B Y. LI H, LIU Y, et al. Improving Web search

results using affinity graph[C]// Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. New York, N. Y., USA: ACM, 2005: 504-511.

[24] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2005, 3: 993-1022.

[25] ZHOU D, WADE V. Latent document re-ranking[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Trier, Germany: DBLP, 2009, 1571-1580.

[26] VULIC I, MOENS M F. Monolingual and cross-lingual information retrieval models based on (Bilingual) word embeddings[C]// Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, N. Y., USA: ACM, 2015: 363-372.

[27] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the IEEE Conference on neural information processing systems. Washington, D. C., USA: IEEE, 2013: 3111-3119.

[28] HEINRICH G. Parameter estimation for text analysis[R]. Leipzig, Germany: University of Leipzig, 2008.

[29] BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval: the concepts and technology behind search [M]. 2nd ed. Indianapolis, Ind., USA: Addison-Wesley Professional, 2011: 94.

[30] BIANCALANA C, GASPARETTI F, MICARELLI A, et al. Social semantic query expansion[J]. ACM Transactions on Intelligent Systems and Technology, 2013, 4(4): 1-43.

作者简介:

赵文玉(1993—),女,湖南衡阳人,硕士研究生,研究方向:信息检索、自然语言处理,E-mail: 719727262@qq.com;

＋周　栋(1979—),男,湖南长沙人,副教授,博士,研究方向:信息检索、自然语言处理,通信作者,E-mail: dongzhou1979@hotmail.com;

曹步清(1979—),男,湖南湘潭人,副教授,博士,研究方向:服务计算、云计算、社会网络;

刘建勋(1970—),男,湖南衡阳人,教授,博士,研究方向:服务计算、工作流。