计算机研究与发展 DOI:10.7544/issn1000-1239.2014.20130339
Journal of Computer Research and Development 51(10):2216-2224,2014

信息检索中的带权邻近度度量研究

薛源海[1,2,3] 俞晓明[1,2] 刘 悦[1,2] 关 峰[1,2,3] 程学旗[1,2]

[1](中国科学院网络数据科学与技术重点实验室 北京 100190)

[2](中国科学院计算技术研究所 北京 100190)

[3](中国科学院大学 北京 100190)

(xueyuanhai@software.ict.ac.cn)

Exploration of Weighted Proximity Measure in Information Retrieval

Xue Yuanhai[1,2,3], Yu Xiaoming[1,2], Liu Yue[1,2], Guan Feng[1,2,3], and Cheng Xueqi[1,2]

[1](*Key Laboratory of Network Data Science and Technology, Chinese Academy of Sciences, Beijing* 100190)

[2](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3](*University of Chinese Academy of Sciences, Beijing* 100190)

Abstract A key problem of information retrieval is to provide information takers with relevant, accurate and even complete information. Lots of traditional information retrieval models are based on the bag-of-words assumption, without considering the implied associations among the query terms. Although term proximity has been widely used for boosting the performance of the classical information retrieval models, most of those efforts do not fully consider the different importance between the query terms. For queries in modern information retrieval, the query terms are not only dependent of each other, but also different in importance. Thus, computing the term proximity with taking into account the different importance of terms will be helpful to improve the retrieval performance. In order to achieve this, a weighted term proximity measure method is introduced, which distinguishes the significance of the query terms based on the collections to be searched. Weighted proximity BM25 model(WP-BM25) that integrating this method into the Okapi BM25 model is proposed to rank the retrieved documents. A large number of experiments are conducted on three standard TREC collections which are FR88-89, WT2G and WT10G. The results show that the weighted proximity BM25 model can significantly improve the retrieval performance, and it has good robustness.

Key words weighted proximity; measure method; BM25; term significance; information retrieval

摘 要 信息检索需要解决的主要问题是为信息索取者提供相关、准确甚至完整的信息.大量的传统检索模型基于词袋假设进行建模,不考虑查询词之间的相互联系.词项邻近度信息在现有的研究中常被用于提升经典信息检索模型的检索效果,但大部分工作没有考虑查询中各个词重要性的差异.在现代信息检索的查询请求中,查询词之间不仅不完全相互独立,而且分别具有不同的重要程度.因此,在计算邻近度信息时对查询词的重要性进行区分,将有助于提高检索效果.带权邻近度 BM25 模型(WP-BM25)使用待检索数据集的背景信息对查询词的重要性进行区分,并将带权邻近度度量方法整合到 BM25 模型

收稿日期:2013-03-27;**修回日期**:2013-07-01

基金项目:国家自然科学基金项目(61100083);国家"八六三"高技术研究发展计划基金项目(2012AA011003)

中.在TREC评测的3个标准数据集FR88-89,WT2G和WT10G上的一系列对比实验表明,该模型具有较好的鲁棒性,且能够使检索效果得到显著提升.

关键词 带权邻近度;度量方法;BM25;查询词重要性;信息检索

中图法分类号 TP391

在过去的20多年里,人类社会迅速地进入了信息时代.人们获取信息的首选途径逐渐由人与人之间的交流发展为在互联网上进行搜索.随着互联网的发展,无穷无尽的内容被网络用户发布在网络上.如何依靠大规模的网络数据,为那些需要获取信息的人们提供相关、准确甚至完整的信息,成为信息检索进行革新的主要推动力[1-2].

各种信息检索模型不断地被提出并投入实际应用,其中主要包括向量空间模型[3]、传统概率模型[4-5]和统计语言模型[6]等.大量的传统检索模型基于“词袋模型”,假设文档以及查询中的词与词之间是完全相互独立的,仅仅利用了诸如词项频率(term frequency)、逆文档频率(inverse document frequency)以及文档长度(document length)等统计量对文档进行打分排序.

直观上讲,使用多个查询词能够更精准地描述用户的查询意图.而当用户输入的查询较短时,搜索引擎普遍采用查询推荐技术帮助用户更好地构造查询[7].大量的信息检索模型已无法高质量地满足用户的信息需求.现代信息检索中的查询请求一般具有以下特征:

1) 查询词之间具有相互联系,并不完全相互独立;

2) 同一查询中的各个词分别具有不同的重要程度.

传统“词袋模型”的词项独立假设与特征1)不相符,但正是由于特征1)已得到了普遍的认同,近年来的一些研究工作将词的邻近度信息引入到检索模型中,通过对查询词之间的相互联系进行建模来提升检索的效果.在此基础上,大部分邻近度相关工作的假设中均未对特征2)进行充分的重视,而传统统计量逆文档频率的提出就是出于对特征2)的考虑,一个词的逆文档频率体现了该词在数据集上的区分能力.本文的主要创新就是针对前人工作的不足,在同时兼顾以上2个特征的情况下,对查询词邻近度的度量展开研究.

下面用一个简单的例子来说明上述2个特征,假设有如下查询请求:

奥运会孙杨参加哪些比赛项目?

首先,各个查询词之间存在着某些潜在的联系,用户既不关心奥运会上其他人的比赛,也不是要查询孙杨在世界游泳锦标赛等其他赛场上参加的项目,反映了特征1);其次,对于该查询来说,“奥运”和“孙杨”这2个词的重要性显然比“参加”和“哪些”等词要高,体现了特征2).

1 相关工作

近些年,许多研究者针对信息检索中词项与词项之间的潜在关系进行了研究.

有人尝试用词组等替代词来建立索引,即将文档表示为词组等大粒度单元的集合[8-9],间接地捕捉词与词之间的依赖关系.文献[8]分别尝试了将具备语法意义的词组和具备统计意义的词组融合到文档建模中.实验证明后一种方案的效果更好,但在某些数据集上有显著的效力提升时,在别的数据集上可能完全没有效果甚至得到负面影响.

有的研究则直接将词与词之间的依赖关系融合到信息检索模型中.随着近年来统计语言建模在信息检索中的兴起,一些相关工作将词项依赖性集成到语言模型中,取得了不错的效果[10-11].文献[11]借助Markov随机场(Markov random fields)提出了一个普适的词项依赖模型,其3种变种形式分别捕获了查询词项之间的不同依赖:第1种假设查询词之间完全独立,作为对照;第2种假设相邻的查询词之间具有某种依赖,类似于二元语言模型;第3种假设所有查询词之间完全依赖,考虑了存在于查询词的所有子集中的潜在依赖.此类工作的主要问题在于,直接考虑词项依赖关系导致模型的参数空间变得异常巨大,往往呈指数型增长.这不仅大大提高了参数估计的难度,而且对数据稀疏性以及噪音也很敏感.

最近几年,一些工作尝试间接地使用词项之间的依赖[12-15].词项邻近度的概念描述了查询词在文档中出现的邻近程度或是密集程度,对词项邻近度进行建模可以间接地捕获词项之间的依赖关系.早期

的研究提出了一些相关的思路,比如文档中包含所有查询词的文本段越短,则该文档与查询越相关.然而那些研究均没有将词项邻近度建模到有效的排序模型中.在近期的研究工作中,2种对于词项邻近度的度量被提出,基于范围(span-based)的度量对查询词分布的范围进行建模,而基于词对(pair-based)的度量在词对的局部距离方面进行建模.核函数(kernel functions)被用于将对词项邻近度的度量转化为得分,最终使用线性加权的形式将词项邻近度合并到传统检索模型中,如式(1)所示:

$$Rank(Q,D)=\lambda Traditional(Q,D)+(1-\lambda)Proximity(Q,D), \tag{1}$$

其中,$Traditional(Q,D)$是传统检索模型的排序得分,而$Proximity(Q,D)$是词项邻近度得分.文献[15]是此类研究中最具有代表性的工作之一,该工作创造了一种伪词项的概念,并取名为交叉词项(cross term),提出了CRTER模型.模型中使用不同的核函数估计查询词对周围文本造成的影响,交叉词项即是2个词相互之间影响的交叠.CRTER模型在大量实验中均取得了不错的效果,是目前最好的模型之一.然而,该工作没有考虑查询词之间的重要性差异,对文档中所有被匹配的查询词平等对待,于是需要对所有词之间的邻近度进行计算,没有取舍,导致模型的计算复杂度较高.

本文在相关工作的基础上,进一步考虑了查询词之间具有不同程度的重要性,提出一种带权邻近度度量方法,不仅使得计算复杂度较低,而且更合理地对邻近度进行了度量.

2 带权邻近度的度量

查询词邻近度描述了在文档空间中2个或多个查询词之间距离相近的程度.显而易见,在检索过程中,需要对查询词出现比较密集的文档给予奖励.为了使用查询词邻近度,就需要通过一定的度量方法得到确切的邻近度统计量.通常1个查询请求包含多个查询词,且每个查询词在文档中将出现不只1次,以往的研究工作在获取邻近度信息时,几乎同等对待各个查询词,而本文将探索如何区别对待各个查询词,更准确地对文档中的查询词邻近度进行度量.

2.1 查询词权重分配

假设有查询$Q=\{q_1,q_2,\cdots,q_n\}$,为了区别对待每个查询词,最简单直接的方法就是给每个查询词赋予权重$w(q_i)$,该权重反映了查询词的重要程度.在无法预先获取查询的相关领域知识的情况下,要想得到查询词的重要程度,能利用的只有待检索数据集以及查询本身.首先,查询词在数据集上的逆文档频率(idf)反映了该词的区分能力;其次,查询词在查询中的词项频率(qtf)反映了用户对该词的关注度.于是对于每个查询词,将idf与qtf组合在一起形成最终的权重.假设待检索数据集为$Collection$,则查询词q_i的权重分配$w(q_i)$如式(2)所示:

$$w(q_i)=\begin{cases} qtf(q_i)\times\ln\dfrac{N}{df(q_i)},\ q_i\in Collection; \\ 0,\ q_i\notin Collection; \end{cases} \tag{2}$$

其中,N为数据集$Collection$中的总文档数,$df(q_i)$为q_i在数据集中出现的文档数.

2.2 度量方法

在给出查询词的带权邻近度度量方法之前,我们首先定义2个与之密切相关的基本概念:文本段和查询覆盖.

假设文档可使用某种切分方式,成为一系列有序的文本单元,文本单元可以是字、词或词组,甚至可以是句子.本文采用中等粒度的文本单元切分方案,以词作为基本的文本单元,则文档表示如式(3)所示:

$$Document=\{Term_1,Term_2,Term_3,\cdots,Term_n\}. \tag{3}$$

定义1. 文本段(segment).文本段是文档中一系列连续出现的文本单元所组成的串.文本段的长度使用其所包含文本单元的总个数进行度量,以$|Segment|$表示.

定义2. 查询覆盖(cover).给定查询$Q=\{q_1,q_2,\cdots,q_m\}$以及文档$D=\{t_1,t_2,\cdots,t_n\}$,则文档D上对于查询Q的查询覆盖$Cover(Q,D)$是满足下列条件的文本段S:

1) $S\subseteq D$;

2) 对于所有的$1\leqslant i\leqslant m$,有$q_i\in S$;

3) 除S自身外,S的任意子文本段S'均不满足条件2).

根据定义1,2可知,对于查询Q以及文档D,当多个查询词在文档中均出现多于1次时,则可能存在多个查询覆盖$Cover(Q,D)$,此时,称所有的$Cover(Q,D)$所组成的集合为查询覆盖集,以$CSet(Q,D)$表示.

下面用一个简单的例子给出直观的展示,给定文档D以及查询Q:

$$D=\{t_3,t_5,t_4,t_2,t_3,t_3,t_3,t_1,t_5,t_2\},$$

$$Q=\{q_1,q_2\,|\,q_1=t_2,q_2=t_3\},$$

那么对应的查询覆盖$Cover(Q,D)$为$S_1=\{t_3,t_5,$

$t_4,t_2\}$,$S_2=\{t_2,t_3\}$和 $S_3=\{t_3,t_1,t_5,t_2\}$,所以查询覆盖集 $CSet(Q,D)=\{S_1,S_2,S_3\}$.

在查询词权重以及查询覆盖集的基础上,本文通过构造一个特殊的集合来实现带权邻近度的度量,该集合以原始查询的子查询为元素,称其为特征查询集.对于查询 Q 以及文档 D,特征查询集用 $R(Q,D)$表示:

$$R(Q,D)=\arg\max_{Q'\in R'(Q,D)}\sum_{q_i\in Q'}w(q_i), \tag{4}$$

$$R'(Q,D)=\{Q'\mid Q'\subset Q\cap D,|Q'|=N\}, \tag{5}$$

其中,q_i 为原始查询中的查询词,Q'为包含 N 个查询词的子查询,且 Q'中的查询词均在文档 D 中出现.

最终,式(6)给出了查询 Q 在文档 D 上的带权邻近度度量,该统计量用 $WP\text{-}Measure(Q,D)$表示:

$$WP\text{-}Measure(Q,D)=\begin{cases}\dfrac{1}{|R(Q,D)|}\times\\ \displaystyle\sum_{Q'\in R(Q,D)}\min_{Cover(Q',D)\in CSet(Q',D)}\{|Cover(Q',D)|\},\\ \quad R(Q,D)\neq\varnothing;\\ \infty,\ R(Q,D)=\varnothing.\end{cases} \tag{6}$$

显然,对于查询 Q,文档 D 的 $WP\text{-}Measure(Q,D)$越小越好.而式(5)中的参数 N 在一定程度上控制了该方法的计算复杂度,N 越大计算复杂度也越大.特别地,当 $N=1$ 时,该方法相当于不考虑邻近度信息,本文取 $N=2$ 进行实验验证.

3 带权邻近度 BM25 模型

本文选用经典的 BM25 模型作为基础,探究第 2 节中提出的带权邻近度度量方法能够带来多大程度的检索效果提升.

3.1 BM25 模型

BM25 模型最先由 Okapi 信息检索系统所实现,是最为经典的基于词袋假设的检索模型之一,其完全忽略文档中查询词与词之间的内在联系.给定查询 $Q=\{q_1,q_2,\cdots,q_n\}$,文档 D 的 BM25 得分为

$$BM25(Q,D)=\sum_{w\in Q\cap D}\left(\ln\frac{N-df(w)+0.5}{df(w)+0.5}\times\frac{(k_1+1)\times tf(w,D)}{k_1\times\left((1-b)+b\times\dfrac{|D|}{avdl}\right)+tf(w,D)}\times\frac{(k_3+1)\times qtf(w,Q)}{k_3+qtf(w,Q)}\right), \tag{7}$$

其中,N 为数据集的总文档数,$|D|$为文档长度,$avdl$ 为数据集的平均文档长度.本文中仅将 b 保留为可调参数,而 k_1 和 k_3 分别根据文献[12]中的推荐值默认设置为 1.2 和 1 000.

3.2 WP-BM25 模型

本文将查询词的带权邻近度度量整合到 BM25 模型中,得到带权邻近度 BM25 模型(WP-BM25).给定查询 Q,文档 D 的 WP-BM25 得分为

$$WP\text{-}BM25(Q,D)=(1-\lambda)BM25(Q,D)+\lambda Proximity(Q,D), \tag{8}$$

其中:

$$Proximity(Q,D)=Kernel(WP\text{-}Measure(Q,D))\times BM25(Q,D). \tag{9}$$

λ 的取值范围为[0,1],当 $\lambda=0$ 时,该模型退化为经典的 BM25 模型,λ 反映了 WP-BM25 模型中对邻近度信息的使用程度.$Kernel(\cdot)$是用于将度量映射到得分的核函数,被广泛应用于邻近度相关研究中.

本文将尝试 4 种核函数,其中高斯核(Gaussian kernel)早已在统计与机器学习领域得到高度评价,文献[14]与文献[15]均使用余弦核(cosine kernel)和三角核(triangle kernel)取得了不错的效果,而反比例核(inverse kernel)尚未被普遍认可.

1) 高斯核:

$$Kernel(x)=\exp\left[\frac{-x^2}{2\sigma^2}\right]; \tag{10}$$

2) 余弦核:

$$Kernel(x)=\frac{1}{2}\left[1+\cos\left(\frac{\pi x}{\sigma}\right)\right]\times 1_{\{x\leqslant\sigma\}}; \tag{11}$$

3) 三角核:

$$Kernel(x)=\left(1-\frac{x}{\sigma}\right)\times 1_{\{x\leqslant\sigma\}}; \tag{12}$$

4) 反比例核:

$$Kernel(x)=\frac{\sigma}{x+\sigma}; \tag{13}$$

其中,σ 为标准化参数,而 $1_{\{x\leqslant\sigma\}}$ 为指示函数:

$$1_{\{x\leqslant\sigma\}}(x)=\begin{cases}1,\ x\leqslant\sigma;\\ 0,\ x>\sigma.\end{cases} \tag{14}$$

这些核函数所得到的检索得分均落在区间[0,1]之内,随着 x 的增加,相应的得分减少,也就意味着查询词项在文档中出现越稀疏,邻近度度量越大,则检索得分越低.4 种核函数分别代表了 4 种不同的随着度量变化的得分变化趋势,通过调整参数 σ 能够对其进行微调,图 1 展示了 $\sigma=20$ 时的各函数图像.其中,随着 x 的增加,高斯核的得分先平缓下

降，后快速下降；余弦核的得分先平缓下降，后快速下降，再趋于平缓；三角核的得分线性下降；反比例核的得分先迅速下降，后趋于平缓.

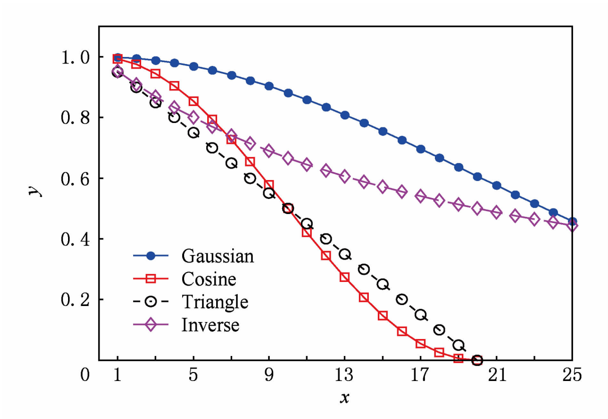


Fig. 1 Curves of four kernel functions.

图 1 4 种核函数的图像

4 实验结果

4.1 测试数据集与评价指标

本文在测试数据集的选取与构成方面，对数据集的规模大小、内容同质性以及噪音程度进行了综合考虑，最终选择在 TREC(text retrieval conference)评测的 3 个标准数据集上进行了一系列的实验. 其中，FR88-89 总体规模较小，单文档规模较大且文档之间长度差异较大，内容同质，噪音很少，文档质量高；WT2G 作为中等规模的内容异质文档集被选取；WT10G 则代表了规模较大的 Web 采集数据集，其平均文档长度较短，文档噪音较大. 查询取自与各个数据集所对应的标准 TREC topics 的 title 域. 各数据集以及查询的相关信息如表 1 所示：

Table 1 Overview of Testing Data Sets

表 1 测试数据集概况

Data Sets	Topics	# of Used Topics	Average Query Length	# of Docs	Average per Doc Length
FR88-89	51～100	21	4.19	45 820	1 498
WT2G	401～450	50	2.44	247 491	1 056
WT10G	451～550	100	4.15	1 692 096	399

For FR88-89, although there are 50 topics, only the 21 with relevant documents were used in experiments.

为了尽量避免数据预处理的好坏程度对于实验结果所造成的影响，本文对于所有数据集中的文档，仅去除了 HTML 标签以及“the”和“a”等限定词. 关于文档以及查询中每个词的词干提取，本文采用 Porter's English Stemmer[16] 进行处理. 对于每一次查询，先使用基准 BM25 模型获取 2000 篇文档，然后再使用 WP-BM25 模型对这些文档进行重排序，最终排名前 1 000 的结果将使用平均准确率均值(mean average precision，*MAP*)，*Precision*@5(*P*@5)以及 *Precision*@10(*P*@10)进行评价. 其中 *MAP* 是 TREC 中最常规的评价指标，而后 2 个指标对于 Web 搜索来说意义更大.

4.2 WP-BM25 的总体效果

本文使用总体效果最好的 BM25 作为基准参照，考察 WP-BM25 对其检索效果的提升情况. WP-BM25 分别使用 Gaussian，Cosine，Triangle 以及 Inverse 这 4 个核函数，在 FR88-89，WT2G 以及 WT10G 这 3 个数据集上进行了实验，并全部通过 *MAP*，*P*@5 以及 *P*@10 这 3 个指标进行了评价，详细的实验结果展示在表 2 中，其中各数据集上针对各指标的最好结果被加粗显示. 如表 2 所示，在各数据集的各指标上，WP-BM25 均带来了显著的效果提升；而针对不同的数据集以及不同的评价指标，4 个核函数各有千秋.

Table 2 Comparison between WP-BM25 and BM25 on Different Data Sets

表 2 带权邻近度 BM25 模型与 BM25 模型在不同数据集上的对比

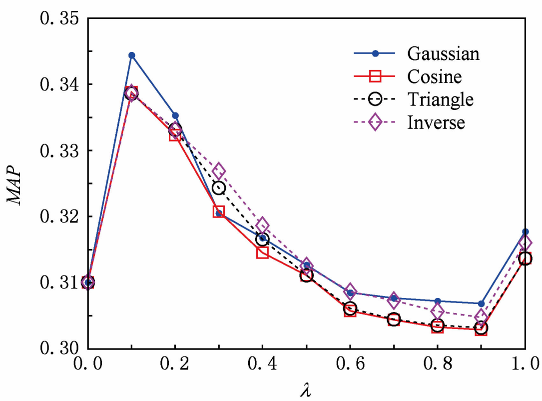
Model	FR88-89			WT2G			WT10G		
	MAP	*P*@5	*P*@10	*MAP*	*P*@5	*P*@10	*MAP*	*P*@5	*P*@10
BM25	0.3100	0.2190	0.1762	0.3735	0.5240	0.4900	0.2422	0.3560	0.3040
WP-BM25 Using Gaussian	0.3363	**0.2476**	**0.1905**	0.3847	0.5560	0.5080	**0.2514**	0.3860	0.3130
WP-BM25 Using Cosine	0.3365	**0.2476**	**0.1905**	**0.3875**	**0.5680**	0.5080	0.2511	0.3860	0.3120
WP-BM25 Using Triangle	0.3413	0.2381	**0.1905**	**0.3875**	**0.5680**	0.5080	0.2512	**0.3900**	0.3140
WP-BM25 Using Inverse	**0.3428**	0.2381	**0.1905**	0.3859	0.5520	**0.5100**	0.2501	**0.3900**	**0.3180**

The parameter b of BM25 is 0.25, and the parameter λ of WP-BM25 is 0.2.

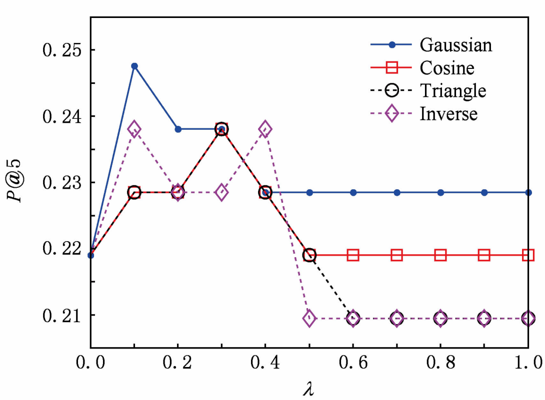
4.3 模型参数对 WP-BM25 的影响

在 WP-BM25 模型中,存在 2 个可调参数 λ 和 σ,其中 λ 控制了模型对词项邻近度的使用程度,而 σ 则影响了模型中带权邻近度度量的得分变化趋势. 本节将分别探究参数 λ 和 σ 的变化对 WP-BM25 造成的影响.

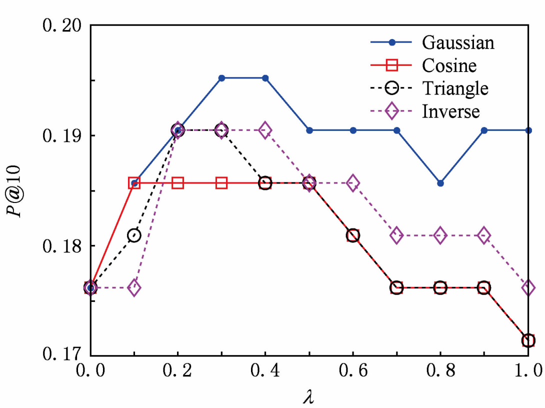
首先将 σ 固定为 20,使 λ 从 0.0 递增至 1.0, WP-BM25 在 FR88-89, WT2G 以及 WT10G 这 3 个数据集上的效果变化如图 2 所示. 总体上看,当 λ 在(0.0,0.5)范围内变化时,WP-BM25 在各个数据集上均能获得不错的效果提升.



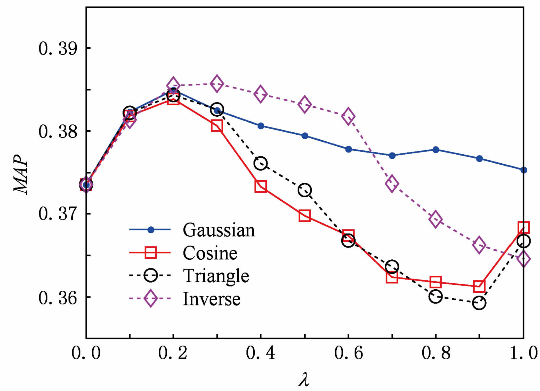
(a) *MAP* of WP-BM25 with different λ on FR88-89



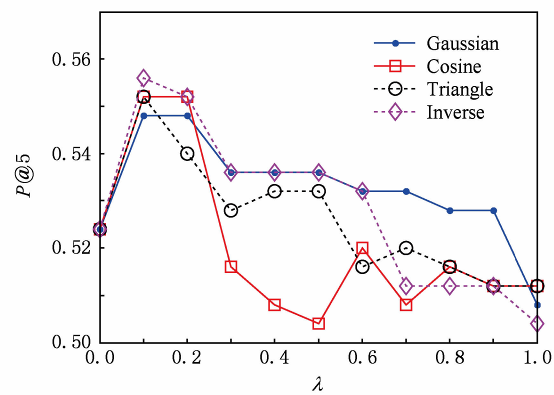
(b) *P*@5 of WP-BM25 with different λ on FR88-89



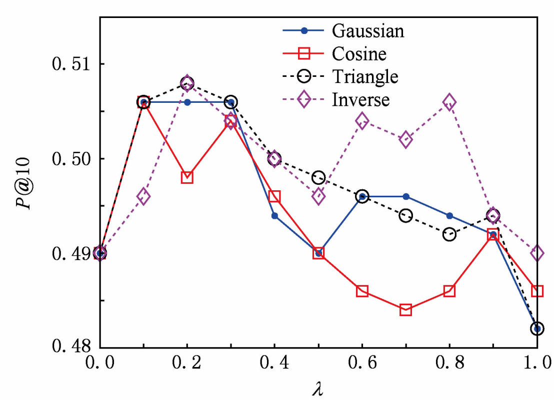
(c) *P*@10 of WP-BM25 with different λ on FR88-89



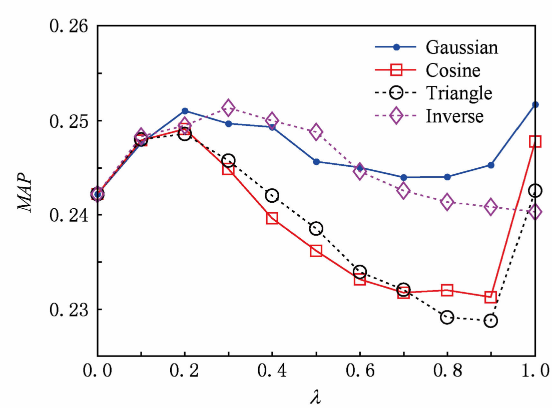
(d) *MAP* of WP-BM25 with different λ on WT2G



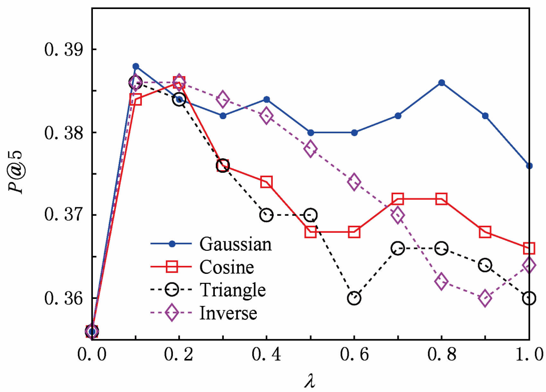
(e) *P*@5 of WP-BM25 with different λ on WT2G



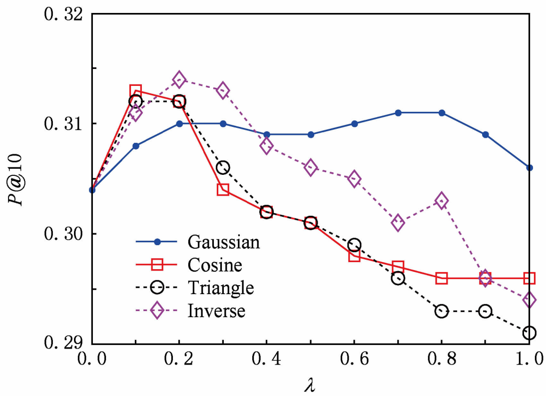
(f) *P*@10 of WP-BM25 with different λ on WT2G



(g) *MAP* of WP-BM25 with different λ on WT10G



(h) $P@5$ of WP-BM25 with different λ on WT10G

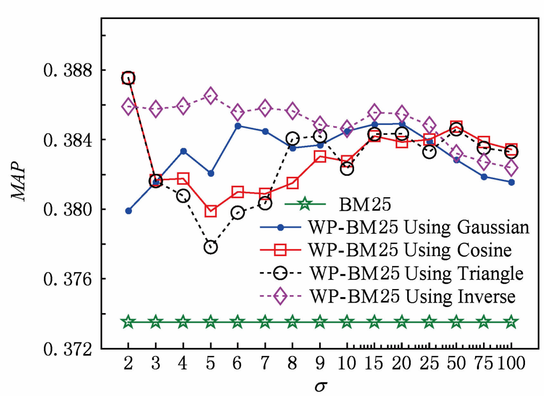


(i) $P@10$ of WP-BM25 with different λ on WT10G

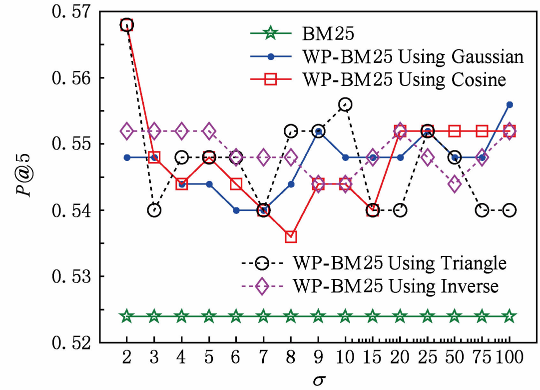
Fig. 2 Sensitivity to WP-BM25 parameter λ on different data sets.

图 2 各数据集上参数 λ 对 WP-BM25 的影响

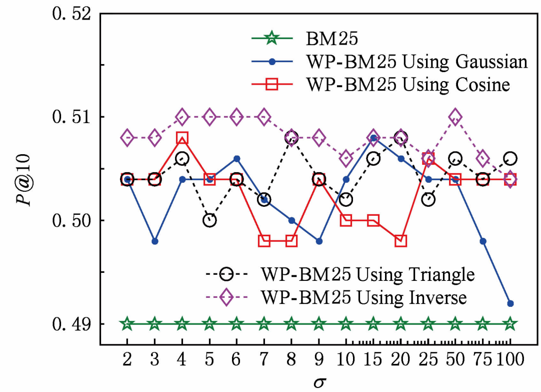
其次固定 λ 为 0.2，使 σ 分别等于{2,3,4,5,6,7,8,9,10,15,20,25,50,75,100}，图 3 展示了分别使用 Gaussian，Cosine，Triangle 以及 Inverse 这 4 个核函数时，WP-BM25 在 WT2G 数据集上的效果变化. 由于 Cosine 和 Triangle 这 2 个核函数对应的效果曲线变化较剧烈，可以看出其明显对 σ 的变化更敏感，在 σ 较小时尤为明显. 无论 σ 取值多少，WP-BM25 在各指标上均能获得效果提升，而不同的核函数倾向于不同的 σ 取值范围.



(a) MAP of BM25 and WP-BM25 with different σ



(b) $P@5$ of BM25 and WP-BM25 with different σ



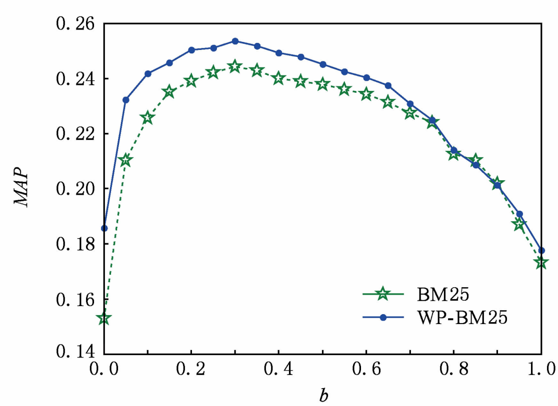
(c) $P@10$ of BM25 and WP-BM25 with different σ

Fig. 3 Sensitivity to WP-BM25 kernel parameter σ on WT2G.

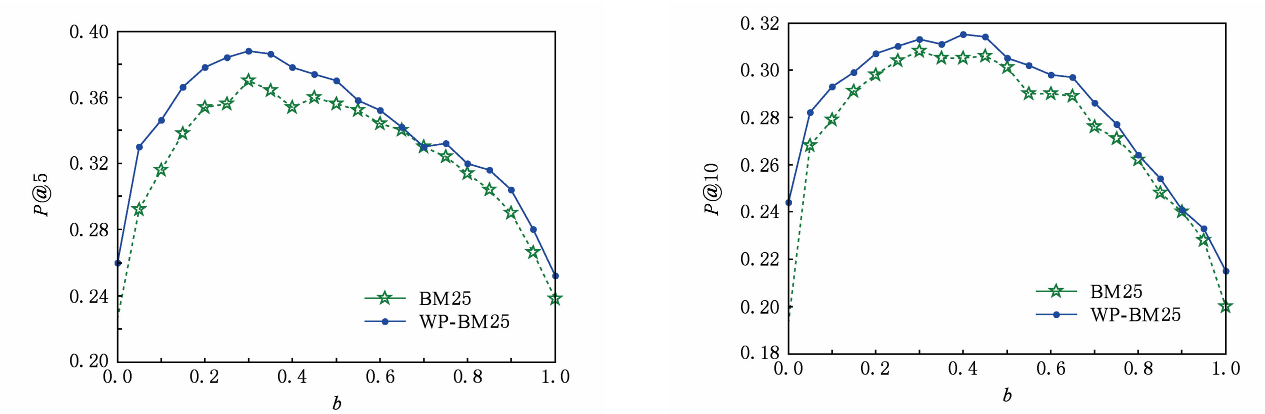
图 3 WT2G 上 σ 对 WP-BM25 的影响

4.4 WP-BM25 模型的鲁棒性

模型的鲁棒性直接决定了其是否可以被应用到实际生产环境中. 本节使用高斯核函数，且将邻近度模型参数 λ 和 σ 分别固定为 0.2 和 20，通过调节 b 来改变基础检索模型 BM25 的检索效果. 在 WT10G 上的实验结果如图 4 所示，WP-BM25 几乎在所有情况下都胜过 BM25.



(a) MAP of BM25 and WP-BM25 with different b



(b) $P@5$ of BM25 and WP-BM25 with different b

(c) $P@10$ of BM25 and WP-BM25 with different b

Fig. 4 Comparison between WP-BM25 and BM25 with the change of b on WT10G.

图 4 b 变化时 WP-BM25 与 BM25 在 WT10G 上的对比

4.5 WP-BM25 模型与 CRTER 模型的对比

最后,将 WP-BM25 模型与当前最好的邻近度检索模型之一 CRTER[15] 进行对比. 文献[15]与本文均以 BM25 模型作为基准,且都在数据集 WT2G 以及 WT10G 上进行了实验,由于这 2 个数据集均为 Web 数据集,使用 Web 搜索更关注的 $P@5$ 指标进行结果比较,如表 3 所示. 可以看出,WP-BM25 至少取得了与 CRTER 可比的效果. 除此之外,本文提出的 WP-BM25 模型在计算复杂度方面,比 CRTER 模型更具优势,例如当查询词个数为 N 时,CRTER 模型需要计算($N(N-1)/2$)个词对的邻近度信息,而 WP-BM25 几乎总是只需计算 1 个词对的邻近度信息.

Table 3 Direct $P@5$ Comparison with CRTER

表 3 在 $P@5$ 指标上与 CRTER 的直接对比

Model	WT2G			WT10G		
	Baseline $P@5$	$P@5$	Outperformance/%	Baseline $P@5$	$P@5$	Outperformance/%
CRTER[15]	0.5280	0.5480	+3.788	0.3800	0.4080	+7.368
WP-BM25	0.5240	0.5680	+8.397	0.3560	0.3900	+9.551

5 总结与展望

本文对信息检索中的词项邻近度问题进行了研究,认为在同样范围内出现的不同词集合承载着不一样的邻近度信息量,在此基础上提出了一种带权邻近度度量方法. 以经典的 BM25 检索模型为例,建模得到带权邻近度 BM25 模型(WP-BM25),并在 TREC 评测的 3 个标准数据集上进行了一系列的实验与分析. 结果表明,WP-BM25 在各个数据集上均显著优于 BM25. 同时,本文也对 WP-BM25 对模型参数的敏感程度进行了讨论.

在今后的工作中,希望能够将带权邻近度检索模型应用到现有的其他检索模型上,如语言模型等. 此外,不同粒度的文档切分方案以及更多的邻近度度量方式对带权邻近度检索模型的影响也是未来的研究方向.

参 考 文 献

[1] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval [M]. Cambridge: Cambridge University Press, 2008

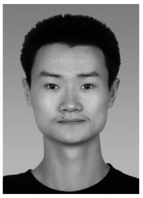
[2] Cheng Xueqi, Lü Jianming, Zhou Zhaotao. P2P full text information retrieval based on centroid method [J]. Journal of Computer Research and Development, 2004, 41(12): 2148-2155 (in Chinese)
(程学旗, 吕建明, 周昭涛. 基于对等网络的全文信息检索[J]. 计算机研究与发展, 2004, 41(12): 2148-2155)

[3] Salton G, Wong A, Yang C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620

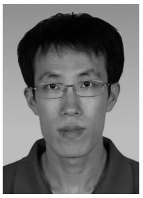
[4] Robertson S E, Jones K S. Relevance weighting of search terms [J]. Journal of the American Society for Information Science, 1976, 27(3): 129-146

[5] Robertson S, Zaragoza H. The Probabilistic Relevance Framework [M]. Hanover, MA: Now Publishers Inc, 2009

[6] Ponte J M, Croft W B. A language modeling approach to information retrieval [C] //Proc of the 21st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1998: 275-281
[7] Li Yanan, Wang Bin, Li Jintao. A survey of query suggestion in search engine [J]. Journal of Chinese Information Processing, 2010, 24(6): 75-84 (in Chinese)
(李亚楠，王斌，李锦涛. 搜索引擎查询推荐技术综述[J]. 中文信息学报，2010，24(6)：75-84)
[8] Fagan J. Automatic phrase indexing for document retrieval [C] //Proc of the 10th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1987: 91-101
[9] Croft W B, Turtle H R, Lewis D D. The use of phrases and structured queries in information retrieval [C] //Proc of the 14th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1991: 32-45
[10] Gao J, Nie J Y, Wu G, et al. Dependence language model for information retrieval [C] //Proc of the 27th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2004: 170-177
[11] Metzler D, Croft W B. A Markov random field model for term dependencies [C] //Proc of the 28th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2005: 472-479
[12] Tao T, Zhai C X. An exploration of proximity measures in information retrieval [C] //Proc of the 30th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2007: 295-302
[13] Zhao J, Yun Y. A proximity language model for information retrieval [C] //Proc of the 32nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2009: 291-298
[14] Lü Y, Zhai C X. Positional language models for information retrieval [C] //Proc of the 32nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2009: 299-306
[15] Zhao J, Huang J X, He B. CRTER: Using cross terms to enhance probabilistic information retrieval [C] //Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 155-164
[16] Porter M F. An algorithm for suffix stripping [J]. Program: Electronic Library and Information Systems, 1980, 14(3): 130-137

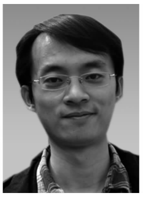


Xue Yuanhai, born in 1987. PhD candidate. His main research interests include information retrieval and data mining.

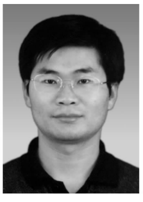


Yu Xiaoming, born in 1977. PhD and associate professor. His main research interests include information retrieval and big data (yuxiaoming@software.ict.ac.cn).

Liu Yue, born in 1971. PhD and associate professor. Her main research interests include information retrieval, data mining and social computing (liuyue@ict.ac.cn).



Guan Feng, born in 1984. PhD candidate. His main research interests include information retrieval and data mining (guanfeng@software.ict.ac.cn).



Cheng Xueqi, born in 1971. PhD and professor. Member of China Computer Federation. His main research interests include network information security, large-scale information retrieval and knowledge mining (cxq@ict.ac.cn).