文章编号：2096-1472(2018)-11-01-05
DOI:10.19644/j.cnki.issn2096-1472.2018.11.001

联邦搜索中基于词向量的多样化信息源选择算法

王雅蓉，李 亮，吴胜利
(江苏大学，江苏 镇江 212013)

摘 要：对支持检索结果多样化任务的信息源选择进行了研究。分析了现有研究的不足，提出利用词向量提取文本的语义特征，在此基础上实现文档建模和信息源选择。采用ClueWeb12b−13数据集构建实验平台和进行实验，基于R方法的评价结果表明，所提出的方法优于现有的方法GLS和MnStD，且在不同条件下性能稳定。

关键词：联邦搜索；信息源选择；检索结果多样化；词向量
中图分类号：TP391.3 **文献标识码**：A

Word Representation-Based Resource Selection for Search Result Diversification in Federated Search

WANG Yarong,LI Liang,WU Shengli
(*Jiangsu University*, *Zhenjiang* 212013,*China*)

Abstract: This article studies the resource selection in supporting of search result diversification,analyzes the shortcomings of the existing researches and proposes to use the distributed word representation to extract the semantic features of the text.Based on this,document modeling and resource selection are achieved.The experimental platform is constructed by using the ClueWeb12b-13 dataset.The evaluation results based on the R-method show that the proposed algorithm is superior to the existing GLS and MnStD and it is stable in various kinds of situations.

Keywords:federated search;resource selection;search result diversification;distributed word representation

1 引言(Introduction)

联邦搜索是一种信息检索形式，主要用于检索多个分布的、独立性较高的信息源[1−3]。目前，一些大型的搜索引擎，如LinkedIn等，通过联邦的方式来完成搜索任务。对于用户提交的特定查询，联邦检索系统分析各子检索系统中的数据源，选择一些相关文档数目较多者，向那些选中的子检索系统移交检索请求，然后回收各子检索系统的检索结果并加以合并，最后将合并的结果返回给用户[3,4]。联邦搜索使得用户通过统一的用户界面同时访问多个独立的信息源，可用性较高。

检索结果多样化[5,6]是信息检索的一项重要任务，其目的是使得结果列表中的文档不仅与查询相关，而且要求这些文档覆盖与查询相关的各个方面。在大多数的检索环境中，特别是对于查询词较少的短查询，用户的查询意图往往不够清晰明确，不同的用户对某一查询项也可能存在不同的查询需求。对于检索结果进行多样化处理可以更好地提升用户的检索体验。

支持检索结果多样化任务的信息源选择研究，要求选择的信息源组合不仅与主查询相关，而且需要与查询的一个或多个子主题相关。即被选的数据源不仅要包含较多的相关文档，而且要覆盖尽可能多的查询子主题。这要求算法不仅需要考虑数据源中文档与查询的相关性，也需要考虑各数据源之间文档内容的冗余度和新颖性[7−9]。

对于那些支持信息检索的结果多样化算法，一般有两种不同的假设。一种假设是：对于用户的任何查询，搜索引擎预先知道与该查询相对应的所有子查询的检索意图。另一种假设是搜索引擎无此信息。要使第一种假设成立，需要做大量的准备工作。尤其是对于一些即席查询(Ad Hoc)而言，准备工作难度和工作量很大。根据假设的不同，结果多样化算法分为显式和隐式两种。类似地，联邦搜索中的多样化数据源选择算法一般也可以分为显式和隐式两种。作为一种显式的信息源选择算法，Hong和Si[9]提出了DivD和DivS。这两者均将一种典型的显式多样化重排算法PM−2[5]用于联邦搜索的信息源选择中。该算法性能较佳。然而由于显式方法本身依赖于子查询信息，同时需要计算文档集与所有子查询的相关性，计算成本很高，因此这类方法的实际应用价值仍有待商榷。针对这个问题，隐式方法在不依赖额外子查询信息的前

提下，通过其他方法实现多样化信息源选择。作为一种隐式的信息源选择算法，Naini等人提出了一种利用文档分类来近似子查询的GLS方法[7]。然而由于文档集中不相关文档的数量远远多于相关文档的数量，因此文档分类与子查询之间的差异较大，导致算法多样化性能不足。Benjamin和Wu[8]将投资组合理论应用于联邦检索中的信息源选择，提出了MnStD方法。该方法的主要特点是通过在多样化重排中加入风险因素，降低所选信息源之间的内容冗余度，提升所选信息源列表的新颖性，从而得到较优的多样化信息源选择结果。然而降低风险与多样化性能也没有直接的关系，因此算法性能并不出色。由此可见，如何在不依赖额外子查询的条件下，实现多样化性能较高的隐式信息源选择算法是一项具有挑战性的任务。

本文提出一种利用词向量的隐式信息源选择算法WbRS(Word representation-based Resource Selection for search result diversification)。在此基础上，我们结合文本语义和词项—词频统计两方面的特征于文档建模，这样可以更准确地计算各信息源之间的内容相似度，进一步提高信息源选择的性能。

2 词向量(Distributed word representation)

词向量，也被称为词编码，是词的分布式特征表述(Distributed Word Representation)[10,11]。它通过深度学习方法将单个词项表示为m维语义空间中的一个向量。

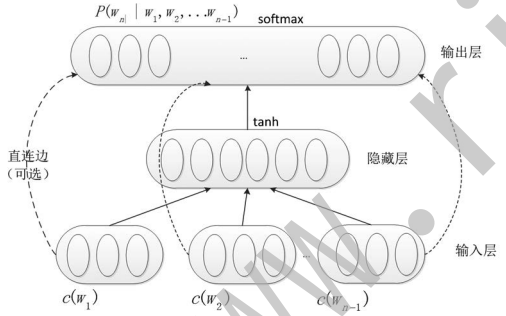

图1 三层神经网络拓扑图

Fig.1 The topology of the three-layer neural network

Bengio[10]利用经典三层神经网络构建n元语言模型，通过在广泛的语料库中训练模型得到广泛认可的词向量，其模型如图1所示。其中，$w_1, w_2, \ldots, w_{n-1}$表示一个词序列的前$n-1$个词。$V$表示语料库中所有词项构成的词汇表，$|V|$表示这个词汇表的大小，即词项数量。$m$为语义空间的维度，即词向量的维度，$c(w_i)$表示词项$w_i$的词向量，是模型的输入层。在整个模型中每个词$w_i$对应着唯一的$1*m$维的词向量$c(w_i)$。模型通过已知的前$n-1$个词，预测第$n$个词$w_n$出现的概率，这$n-1$个词的词向量组成矩阵$C: C = [c(w_1), c(w_2), \ldots, c(w_{n-1})]$。模型的隐藏层由偏置项$d$和激活函数$\tanh$组成。输出层共有$|V|$个节点，每个结点$y_i$的值表示词序列中下一个词为$w_i$的概率，通过激活函数可以得到最终归一化的概率值。因此，模型输出层y的计算如式(1)所示。

$$y = b + WC + U\tanh(d + HC) \tag{1}$$

其中，b为包含$|V|$个元素的输出层偏置项。W表示从输入层到输出层直连边的权重矩阵，当模型中不存在直连边时，将W置为0。U表示从隐藏层到输出层的权重矩阵，d为隐藏层偏置项，包含h个元素。H为隐藏层权重矩阵。最终通过随机梯度下降法优化模型中的参数，得到词汇表中各词项的词向量表示。研究表明[11,12]，词项的这种向量表示可以较为准确地挖掘词项语义，提取文本内容主题特征，本文将词向量技术应用至联邦检索的信息源选择中。

3 WbRS算法(Algorithm WbRS)

本节介绍基于词向量的多样化信息源选择算法WbRS。用词向量对文档建模，得到文档在语义层的相似度得分，将这个得分加入到文档多样化重排的计算中，最终得到多样的信息源选择结果。

WbRS算法第一部分是用词向量技术训练样本相关文档集，得到文档内所有词项的向量表示。样本相关文档集指的是，通过有效的检索系统(如Indri[13]、terrier等)在样本文档集中对用户查询进行检索，得到查询的相关性文档列表D_{rel}，D_{rel}中所有文档组成的文档集即为相关样本文档集。V表示这个文档集的词汇表。将D_{rel}作为语料库，利用词向量技术训练D_{rel}，得到词汇表V中各词项在m维语义空间中的向量表示。V中各词项的词向量表示组成$|V|*m$维的矩阵W。

已有研究利用词向量技术，在语料库中训练得到性能较优的词项向量表示。其中，Mikolov和Chen等人[14]用循环神经网络模型来训练语言模型，提出并开源词向量技术word2vec。Google利用word2vec在广泛语料库上训练模型，得到并公开了词向量表示。大量研究表明[12,15,16]，Google采用word2vec模型训练得到的词向量表示，在大部分性能评价指标中均取得较优的结果。这些词向量表示可以准确地挖掘词项语义[11,12]，直接将其应用至本算法，可以提升算法运行效率，同时保证较好的实验性能。本文直接将Google通过word2vec技术训练得到的词向量表示应用至WbRS算法中。

算法的第二部分，利用word2vec技术得到的词向量表示计算文档相似度，并对样本文档相关性排序列表D_{rel}进行多样化重排。首先，利用词向量矩阵W实现文档建模，得到各文档的向量表示。然后，利用这些文档向量计算文档间的相似度。最后实现样本文档的多样化重排。

本节拟采用简洁高效的加权平均方法，根据算法第一部分中词项在语义空间中的向量表示，将文档表示为语义空间中的向量。已有研究证明[11,12]，用加权平均的方法处理文档中的词向量，具有较高的可行性。WbRS首先统计一篇文档d_i内所有词项$w_{i,1}, w_{i,2}, \ldots, w_{i,n_i}$的出现频率$f_{i,1}, f_{i,2}, \ldots, f_{i,n_i}$和文档$d_i$内的词项数量$n_i$，得到文档的词项-词频向量表示$\overrightarrow{\theta_{d_i}} = [f_{i,1}, f_{i,2}, \ldots, f_{i,n_i}](1 * n_i$维)。另一方面，对照词向量矩阵$W$，可以得到文档$d_i$内各个词项的词向量表示

$\overrightarrow{w_{i,1}}, \overrightarrow{w_{i,2}}, \ldots, \overrightarrow{w_{i,n_i}}$。并由这些词向量组成文档$d_i$的词向量矩阵$W_i = (\overrightarrow{w_{i,1}}, \overrightarrow{w_{i,2}}, \ldots, \overrightarrow{w_{i,n_i}})(n_i * m$维)。此时，文档$d_i$基于词向量的向量表示$\overrightarrow{\varphi_{d_i}}(1 * m$维)，如式(2)所示。

$$\overrightarrow{\varphi_{d_i}} = \frac{1}{n_i} * (\overrightarrow{\theta_{d_i}} \times W_i) \tag{2}$$

其中，n_i表示文档d_i中的词项数量。因此文档d_i与d_j的相似度计算可以用向量$\overrightarrow{\varphi_{d_i}}$与$\overrightarrow{\varphi_{d_j}}$的向量夹角余弦值表示。

本方法中$\overrightarrow{\theta_{d_i}}$是文档内词项−词频得到的文档向量表示。$W_i$为由word2vec得到的由文档$d_i$内各词项在语义空间中的词向量组成的映射矩阵。其中$\overrightarrow{\theta_{d_i}}$考察了传统的词项的词频特征，即文档中出现频率越高的词项，对文档内容越为重要。$\overrightarrow{\theta_{d_i}}$经过词向量矩阵$W_i$映射，将仅考虑词频的文档向量，转化为在$m$维语义空间中的文档向量$\overrightarrow{\theta_{d_i}}$，通过$\overrightarrow{\theta_{d_i}}$可以分析文档的语义相似度。

然而，式(2)中的$\overrightarrow{\theta_{d_i}}$是仅考虑词频的向量，忽略了语料库中包含某词项的文档数，即反文档频率IDF。因而，对本节提出的式(2)做出进一步改进，采用TF−IDF加权的词向量表示。

TF值表示文档中词项的出现频率，是文档中词项、词频的归一化度量，描述了词项在文档中的重要程度。TF的计算如式(3)。

$$tf_{i,t} = \log\left(f_{i,t} / \sum_{t \in d_i} f_{i,j}\right) \tag{3}$$

其中，$tf_{i,t}$表示词项t在文档d_i中的TF值，$f_{i,t}$表示文档d_i中词项t的出现频数。式(3)中对词项的出现频率取对数，使得文档中某些高频词汇和低频词汇的TF值更具有可比性。

文档倒置频率IDF，表示包含某词项的文档数量，它反映了文档集中某词项的重要性。因为包含词项t的文档越少，则表明这些文档对词项t越为重要。词项t的文档倒置频率idf_t的经典计算如式(4)所示。

$$idf_t = \log\left(\frac{N}{n_t + 1}\right) \tag{4}$$

其中，N表示文档集中的文档总数，n_t为包含词项t的文档数量。基于TF−IDF的文档词项向量$\overrightarrow{\theta_{d_i}'}$可以表示为$\overrightarrow{\theta_{d_i}'} = [tf_{i,1} * idf_1, tf_{i,2} * idf_2, \ldots, tf_{i,n} * idf_n]$。

用$\overrightarrow{\theta_{d_i}'}$替代式(2)中基于词项−词频的文档向量表示$\overrightarrow{\theta_{d_i}}$，可以得到文档基于TF−IDF权重和词向量的文档向量表示$\overrightarrow{\varphi_{d_i}'}$。

$$\overrightarrow{\varphi_{d_i}'} = \frac{1}{n_i} * (\overrightarrow{\theta_{d_i}'} \times W_i) \tag{5}$$

文档d_i，d_j相似度可以转化为求解文档对应向量$\overrightarrow{\varphi_{d_i}'}$与$\overrightarrow{\varphi_{d_j}'}$夹角的余弦值，文档$d_i, d_j$在语义空间$\Phi$中的相似度计算，如式(6)所示。

相似度计算，如式(6)所示。

$$sim(d_i, d_j, \Phi) = sim(\overrightarrow{\varphi_{d_i}'}, \overrightarrow{\varphi_{d_j}'}) = \frac{\sum_{t=1}^{m} \overrightarrow{\varphi_{d_i}'}_t \overrightarrow{\varphi_{d_j}'}_t}{\sqrt{\sum_{t=1}^{m} \overrightarrow{\varphi_{d_i}'}_t^2} * \sqrt{\sum_{t=1}^{n} \overrightarrow{\varphi_{d_j}'}_t^2}} \tag{6}$$

其中，$\overrightarrow{\varphi_{d_i}'}_t$表示向量空间中$\overrightarrow{\varphi_{d_i}'}$在第$t$维语义空间中的特征。

结合文档相似度$sim(d_i, d_j, \Phi)$和文档的查询相关度$rel(d, q)$，WbRS方法用经典的贪心选择策略，对样本文档相关性排序列表D_{rel}进行多样化重排，得到多样化的文档排序列表D_{div}。重排过程中，依次贪心地选择使得目标函数$GoodnessDiv(d_i, Q, D_{div}, \lambda, \Phi)$得分最高的文档$d_i$，加入到已排序列表$D_{div}$末尾，直到所有文档被加入重排列表$D_{div}$中。目标函数$GoodnessDiv(d_i, Q, D_{div}, \lambda, \Phi)$要求候选文档$d_i$与查询保持足够的相关性，同时要求$d_i$与$D_{div}$中的所有文档的相似度最小，如式(7)所示。

$$GoodnessDiv(d_i, Q, D_{div}, \lambda, \Phi) \leftarrow \lambda rel(d_i, Q) - (1 - \lambda) max_{d_j \in D_{div}} sim(d_i, d_j, \Phi) \tag{7}$$

其中，$rel(d_i, Q)$表示文档d_i与查询项Q的相关度得分，$sim(d_i, d_j, \Phi)$为文档在向量空间Φ中的内容相似度，λ为平衡参数。

与传统的多样化重排方法不同，WbRS方法在计算文档相似度过程中，采用基于word2vec和TF−IDF权重的文档向量得到较为准确的文档内容相似度。这种方法结合了传统的TF−IDF权重，综合考虑了文档中词项的出现频率和文档倒排频率，同时基于词向量技术，将传统的统计语言模型映射为语义空间中的文档向量模型，因而可以提高重排列表的多样化性能。算法中使用加权平均方法，由文档中各词项的词向量计算得到文档向量，方法简洁高效且能取得较好的实验性能。

算法第三部分通过样本文档多样化排序列表D_{div}，对信息源进行多样化排名。D_{div}列表中的各文档在与查询相关的同时，文档之间存在足够的差异度，这些文档按照最优顺序排列，构成了样本文档的多样化列表。另一方面，D_{div}由来自各信息源的样本文档组成，这些样本相关文档在对应的信息源中都存在着一些相似文档与之对应。D_{div}中文档的多样化排名，反映了各信息源整体的查询相关性和内容新颖度。各信息源根据样本文档在D_{div}中的排名，获得不同的分值。并按照各信息源得分对它们进行排名，得到信息源的排名列表，这个得分由公式$G(r) = 1/(r + 60)$得到。Cormack等人[17]的研究表明，通过倒数模型，可以将文档排名转换为有效的文档得分。

算法1(基于词向量的多样化信息源选择算法WbRS)

输入：查询Q，中央样本文档集相关性排序列表D_{rel}，平衡参数λ，词向量矩阵W，目标函数$GoodnessDiv$，计分函数G。

输出：多样化信息源结果列表R

/*第一部分，构建文档向量空间模型Φ*/

1 用词向量技术训练D_{rel}中文档，得到词向量矩阵W

/*第二部分，执行文档列表多样化*/

2 for each $d_i \in D$ do

3 $\overrightarrow{\theta_{d_i}'} \leftarrow [tf_{i,1} * idf_1, tf_{i,2} * idf_2, \ldots, tf_{i,n_i} * idf_{n_i}]$, $n_i \leftarrow \sum_{f_{i,t} \neq 0} 1$

4 $\overrightarrow{\varphi_{d_t}'} \leftarrow \frac{1}{n_i} * (\overrightarrow{\theta_{d_t}'} \times W_i)$
5 end for
6 $D_{div} \leftarrow \emptyset$
7 while $|D_{rel}| > 0$ do
8 for each $d \in D_{rel}$ do
9 $v(d) \leftarrow GoodnessDiv(d, Q, D_{div}, \lambda, \Phi)$
10 end for
11 $d^* \leftarrow \text{argmax}\{ v(d) \}$
12 $D_{rel} \leftarrow D_{rel} / \{d^*\}, D_{div} \leftarrow D_{div} \cup \{d^*\}$
13 end while
/*第三部分，信息源排序*/
14 $s_1, s_2, \ldots, s_{n_C} \leftarrow 0$
15 for each $i \in [1, n_C]$ do
16 for each $d \in D_{div}$ and $d \in C_i$ do
17 $s_i \leftarrow s_i + G(r_d)$
18 end for
19 end for
20 按s值从大到小顺序为信息源排序，得到信息源多样化排序列表R
21 return R

算法中，$tf_{i,t}$表示词项t在文档d_i中的TF权重，idf_t表示t的IDF权重。n_i为文档d_i中的词项数量。$\overrightarrow{\theta_{d_t}'}$为$1 * n_i$维向量，表示$d_i$中词项的TF-IDF权重。$W_i$为$n_i * m$维矩阵，由$d_i$中各词项的词向量表示组成。$\overrightarrow{\varphi_{d_t}'}$为$1*m$维向量，表示文档$d_i$在$m$维语义空间中的向量表示。$s_i$为信息源$C_i$的多样化得分。$n_C$为候选信息源数量。$d \in C_i$表示文档$d$包含在信息源$C_i$中。

4 实验(Experiment)

4.1 数据集

本文实验数据来自国际信息检索会议(Text Retrieval Conference, TREC)在网络检索任务中提供的Clueweb12-B13英文数据集2。Clueweb12-B13数据集解压缩后约为1.95T，包含52,343,021篇网页。本文在ClueWeb12-B13上构建100个子检索系统，这100个子检索系统内包含了ClueWeb12-B13的全体文档。本实验首先在ClueWeb12-B13中随机选取1%的网页文本得到约52万个网页文档，使用K-means算法对这些网页文档作简单的文本聚类，迭代50次得到100个文本聚类中心，这100个聚类中心就对应着联邦检索系统的100个子系统。分别计算ClueWeb12-B13上的52,343,021篇文档与这100个聚类中心的距离，为每篇文档选择与之最近的聚类中心，则该文档被分配至这个聚类中心对应的子检索系统中。重复这个过程，直至ClueWeb12-B13中的所有文档都被分配完毕，得到最终的100个子检索系统。最后，从各子检索系统中，随机选取1%的文档作为各子检索系统的样本文档，这100个子系统的样本文档构成中央样本文档集。这100个子检索系统和样本文档集，共同构成本文的联邦检索实验环境。表1给出了更多的统计数据。

表1 实验数据集统计信息

Tab.1 Statistics of the experimental dataset

统计信息	数量
信息源	100
文档总计	52,343,021
最小信息源包含的文档数	17,835
最大信息源包含的文档数	6,653,521
所有信息源平均文档数	523,430

本实验选用了TREC Web Track 2013[18]和TREC Web Track 2014[19]中的100个查询。

4.2 评价方法

经典的基于R多样化评价方法[8,9]，是联邦检索信息源选择应对检索结果多样化常用的评价标准。

$$R_M(C_K) = \frac{M(C_K\text{上理想检索结果})}{M(\text{全集上理想检索结果})} \tag{8}$$

其中，C_K表示由算法选出的K个信息源构成的组合，M代表常用的检索结果多样化的评价标准，如：ERR-IA、nERR-IA、NRBP、nNRBP、α-nDCG、MAP-IA、P-IA等。

4.3 WbRS算法性能评估与分析

本节采用ReDDE+MMR[7,9]算法作为实验基线，对比现有的隐式信息源选择算法研究中最新的研究成果GLS[7]和MnStD[8]算法。设置信息源选择数N_c=3、5和10，分别考察信息源选择数量较少，数量适中和数量较多，三种情形下算法的多样化性能。采用R(ERR-IA@20)、R(nERR-IA@20)、R(α-nDCG@20)、R(NRBP)和R(nNRBP)作为评价标准，实验对比结果分别如表2—表4所示。

表2 N_c=3时各算法性能比较

Tab.2 Comparison of performance when N_c=3

算法	R(ERR-IA)	R(nERR-IA)	R(α-nDCG)	R(NRBP)	R(nNRBP)
WbRS	0.561	0.554	0.550	0.587	0.513
GLS	0.506	0.499	0.504	0.514	0.502
MnStD	0.507	0.499	0.483	0.534	0.520
ReDDE+MMR	0.400	0.386	0.381	0.410	0.390

表3 N_c=5时各算法性能比较

Tab.3 Comparison of performance when N_c=5

算法	R(ERR-IA)	R(nERR-IA)	R(α-nDCG)	R(NRBP)	R(nNRBP)
WbRS	0.622	0.607	0.629	0.613	0.590
GLS	0.592	0.587	0.613	0.589	0.583
MnStD	0.592	0.607	0.610	0.590	0.584
ReDDE+MMR	0.492	0.476	0.478	0.497	0.472

表4 N_c=10时各算法性能比较

Tab.4　Comparison of performance when N_c=10

算法	R(ERR−IA)	R(nERR−IA)	R(α−nDCG)	R(NRBP)	R(nNRBP)
WbRS	0.711	0.709	0.727	0.712	0.695
GLS	0.646	0.641	0.667	0.641	0.634
MnStD	0.645	0.634	0.665	0.640	0.624
ReDDE+MMR	0.583	0.568	0.492	0.577	0.557

从表2—表4中可以发现，在基于R评价方法的各种多样化指标中，WbRS算法的实验性能明显优于ReDDE+MMR方法，多样化性能较好。在R(ERR−IA@20)，R(nERR−IA@20)，R(α−nDCG@20)，R(NRBP)和R(nNRBP)多项指标中，WbRS算法性能均表现最佳。其中，在R(ERR−IA@20)评价指标中，N_c=3时，WbRS算法相较于GLS和MnStD算法，多样化性能分别提升了10.9%和10.7%。N_c=5时，WbRS算法相较于GLS和MnStD算法，多样化性能分别提升了5.1%和5.1%。N_c=10时，WbRS算法相较于GLS和MnStD算法，多样化性能分别提升了10.1%和10.2%。在各项指标中，WbRS都有着较优的表现。实验表明，基于词向量和TF−IDF权重实现的信息源选择算法WbRS，可以有效提高信息源选择的多样化性能。

4.4　WbRS算法的稳定性观测与分析

本节对比不同信息源选择数N_c时WbRS算法的性能，并分析算法的稳定性。设置N_c的范围为{3,4,5,6,7,8,9,10}，图2和图3分别给出在R(MAP−IA@20)与R(P−IA@20)两种评价指标中，WbRS、GLS和MnStD算法的性能对比。

从图2和图3可以发现，随着N_c的增加，各种算法的R(MAP−IA@20)、R(P−IA@20)评价值均有明显提升。这表明随着信息源选择数的增加，各种算法选择的信息源组C_K均能覆盖更多的查询子主题。其中，WbRS算法的性能始终优于GLS和MnStD，这表明对于信息源选择数的变化，WbRS算法均能保持较优的多样化性能，算法性能相对稳定。

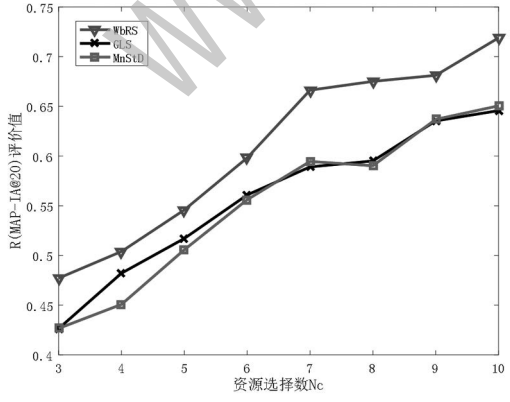

图2 在R(MAP−IA@20)指标中算法性能对比

Fig.2 Comparison of algorithms performance in R(MAP−IA@20)

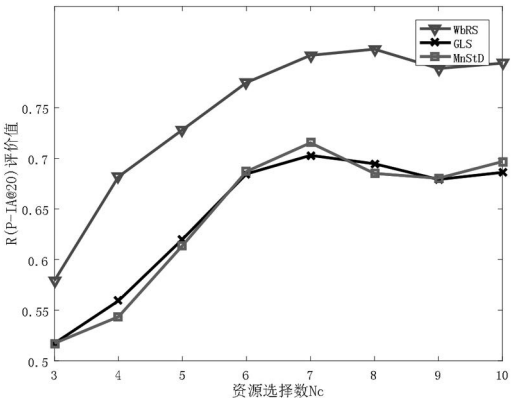

图3 在R(P−IA@20)指标中算法性能对比

Fig.3 Comparison of algorithms performance in R(P−IA@20)

5　结论(Conclusion)

本文将自然语言处理的最新研究成果——词向量技术应用至联邦搜索中多样化的信息源选择研究中。提出了一种隐式的多样化信息源选择方法WbRS。实验结果表明，与最新的隐式多样化信息源选择方法MnStD和GLS算法相比，WbRS算法能有效提高信息源选择结果的多样化性能，且算法性能较为稳定。

结果合并和结果显示是联邦检索中另两项重要的任务，也是我们下一步的工作。在结果合并任务中，我们继续探讨基于词向量技术的方法。在结果显示任务中，我们将探讨层次化的多样化文档检索结果显示方法。

参考文献(References)

[1] 杨海锋,陆伟.联邦检索研究综述[J].图书情报工作,2015,59(1):134−143.

[2] 耿骞,刘畅.分布式检索系统及其体系结构[J].国家图书馆学刊,2004(2):69−73.

[3] 万常选,邓松,刘喜平,等.Web数据源选择技术[J].软件学报,2013,24(4):781−797.

[4] Shokouhi M,Si L.Federated Search[J].Foundations & Trends in Information Retrieval,2011,5(1):1−102.

[5] Dang V,Croft W B.Diversity by proportionality:an election−based approach to search result diversification[C].International ACM SIGIR Conference on Research and Development in Information Retrieval.ACM,2012:65−74.

[6] Markowitz H M.Foundations of Portfolio Theory[J].Journal of Finance,1991,46(2):469−477.

[7] Naini K D,Siberski W,Siberski W.Scalable and Efficient Web Search Result Diversification[J].ACM Transactions on the Web,2016,10(3):1−30.

[8] Benjamin Ghansah,Shengli Wu.A Mean−Variance Analysis Based Approach for Search Result Diversification in Federated Search[J].International Journal of Uncertainty, Fuzziness and Knowledge−Based Systems,2016,24(02):195−211.

[9] Hong D,Si L.Search result diversification in resource selection for federated search[C].International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2013:613-622.

[10] Bengio Y,Ducharme R,Vincent P,et al.A neural probabilistic language model[J].Journal of Machine Learning Research,2006,3(6):1137-1155.

[11] Mnih A,Hinton G.A scalable hierarchical distributed language model[C].International Conference on Neural Information Processing Systems.Curran Associates Inc,2008:1081-1088.

[12] Kusner M J,Sun Y,Kolkin N I,et al.From word embeddings to document distances[C].International Conference on International Conference on Machine Learning.JMLR.org,2015:957-966.

[13] Metzler D,Croft W B.Combining the language model and inference network approaches to retrieval[J].Information Processing & Management,2004,40(5):735-750.

[14] Mikolov T,Chen K,Corrado G,et al.Efficient Estimation of Word Representations in Vector Space.CoRR abs/1301.3781(2013).

[15] Goldberg Y,Levy O.word2vec Explained:deriving Mikolov et al.negative-sampling word-embedding method[J].Eprint Arxiv,2014,9:1402-1407.

[16] Levy O,Goldberg Y.Neural word embedding as implicit matrix factorization[J].Advances in Neural Information Processing Systems,2014,3:2177-2185.

[17] Cormack G V,Clarke C L A,Buettcher S.Reciprocal rank fusion outperforms condorcet and individual rank learning methods[C].International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2009:758-759.

[18] Collins-Thompson K,Bennett P,Diaz F,et al.Overview of the TREC 2013 webtrack[C].TREC,2013.

[19] Collins-Thompson K,Macdonald C,Bennett P,et al.TREC 2014 web track overview[C].TREC,2014.

作者简介：

王雅蓉(1994-)，女，硕士生.研究领域：信息检索.

李 亮(1994-)，男，硕士生.研究领域：信息检索.

吴胜利(1965-)，男，博士，教授.研究领域：信息检索.

(上接第9页)

表2 特征匹配实验数据

Tab.2 Experimental data of feature matching

特征算子类型	特征点数量		匹配对数(未提纯)	匹配对数(提纯后)
SIFT	875	1318	538	236
改进SIFT	708	1178	623	373

5.2.2 特征匹配实验分析

从图2和表2可以看出，经RANSAC算法剔除错误匹配(提纯)后[8]，已经去除大量的错误匹配点，SIFT算子还存在少量的错误匹配，而基于改进的SIFT算法的特征点提取匹配基本没有了错误匹配。

6 结论(Conclusion)

通过物体特征提取和匹配实验将Harris、SIFT、改进SIFT算法进行对比，改进的SIFT算法通过RANSAC剔除错误匹配后，克服了其他两种算法的缺点，在特征提取和匹配上效果更好。

参考文献(References)

[1] 李鹏程,曾毓敏,张梦.一种改进的Harris角点检测算法[J].南京师大学报(自然科学版),2014,37(02):49-54.

[2] 周颖.基于SIFT算法的图像特征匹配[J].现代计算机,2015(5):63-68.

[3] 曾峦,王元钦,谭久彬.改进的SIFT特征提取和匹配算法[J].光学精密仪器,2011,19(06):1392-1396.

[4] S.Bell,C.L.Zitnick,K.Bala,et al.Inside-outside net:Detecting objects in context with skip pooling and recurrent neural net works[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Reco}nition_2016:2874-2883.

[5] Fang Xianyong,Zhang Mingmin,Pan Zhigeng,et al.A new method of manifold mosaic for large displacement images[J].Journal of Computer Science and Technology,2006,21(2):214-223.

[6] 锥伟群,高屹.基于改进RANSAC算法的图像拼接方法[J].科技创新与应用,2015,26(5):21-22.

[7] Kasar T,Ramakrishnaa A G.Block-based feature detection and matching for mosaicing of camera-captured document images [C].Pros of IEEE region 10 conference.Taipei:IEEE.

[8] 周建平,杨金坤,郑宇.基于改进SIFT特征匹配的视频拼接—在倒车系统中的应用[J].企业技术开发,2011,30(22):70-71.

作者简介：

张春林(1987-)，男，硕士生.研究领域：智能制造.

陈劲杰(1969-)，男，硕士，副教授.研究领域：智能制造.