关联模式挖掘与词向量学习融合的伪相关反馈查询扩展

黄名选[1,2]
（1. 广西跨境电商智能信息处理重点实验室 (广西财经学院), 广西南宁 530003；
2. 广西财经学院信息与统计学院, 广西南宁 530003）

摘 要： 针对自然语言处理中查询主题漂移和词不匹配问题，提出基于CSC(Copulas-based Support and Confidence)框架的关联模式挖掘与规则扩展算法，并将基于统计学分析的关联模式与具有上下文语义信息的词向量融合，提出关联模式挖掘与词向量学习融合的伪相关反馈查询扩展模型.该模型对伪相关反馈文档集挖掘规则扩展词，对初检文档集进行词嵌入学习训练得到词向量，计算规则扩展词与原查询的向量相似度，提取向量相似度不低于阈值的规则扩展词作为最终扩展词.实验结果表明，所提扩展模型能有效地减少查询主题漂移和词不匹配问题，提高检索性能，与现有基于关联模式的和基于词向量的查询扩展方法比较，MAP（Mean Average Precision）平均增幅最大可达17.52%，对短查询更有效.所提挖掘方法可用于其他文本挖掘任务和推荐系统，以提高其性能.

关键词： 自然语言处理；信息检索；文本挖掘；词嵌入；查询扩展
中图分类号： TP311 **文献标识码：** A **文章编号：** 0372-2112 (2021) 07-1305-09
电子学报URL： http://www.ejournal.org.cn **DOI：** 10.12263/DZXB.20200654

Pseudo-Relevance Feedback Query Expansion Based on the Fusion of Association Pattern Mining and Word Embedding Learning

HUANG Ming-xuan[1,2]
（1. *Guangxi Key Laboratory of Cross-border E-commerce Intelligent Information Processing, Guangxi University of Finance and Economics, Nanning, Guangxi* 530003, *China*;
2. *School of Information and Statistics, Guangxi University of Finance and Economics, Nanning, Guangxi* 530003, *China*）

Abstract: In order to solve the problems of query topic drift and word mismatch in natural language processing, an algorithm of association pattern mining and rule expansion based on CSC(Copulas-based Support and Confidence) framework is proposed. The association patterns based on statistical analysis are fused with the word embedding with context semantic information, and a pseudo-relevance feedback query expansion model is presented based on the fusion of association pattern mining and word embedding learning. In this model, the rule expansion terms are mined from the pseudo-relevance feedback document set, and the word vectors are obtained by word embedding learning training of the initial document set. The vector similarity between the rule expansion term and original query is calculated, and the rule expansion terms whose vector similarity is not lower than the threshold are extracted as the final expansion terms. The experimental results show that the proposed expansion model can effectively reduce the problems of query topic drift and word mismatch, improving the performance of information retrieval. Compared with the existing query expansion methods based on association pattern and word embedding, the average increase of the MAP（Mean Average Precision）of the proposed expansion model is up to 17.52%. The expansion model in this paper is more effective for short queries. The proposed mining method can be used in other text mining tasks and recommendation systems to improve their performance.

Key words: natural language processing; information retrieval; text mining; word embedding; query expansion

1 引言

查询扩展是解决自然语言处理中查询主题漂移和词不匹配问题的核心技术之一，指的是对原查询的修改，弥补查询信息不足，改善检索性能. 其核心问题是

收稿日期：2020-07-05；修回日期：2020-09-25；责任编辑：覃怀银
基金项目：国家自然科学基金(No.61762006)

扩展词的来源及其扩展模型的设计. 伪相关反馈文档是扩展词重要来源之一. 伪相关反馈查询扩展[1~3]指的是从伪相关反馈文档集中抽取扩展词实现查询扩展，其缺陷是严重受限伪相关反馈文档的质量和数量，导致查询主题漂移和词不匹配问题. 基于关联规则挖掘的和基于词向量学习的伪相关反馈查询扩展[4~18]克服上述缺陷，在一定程度上改善和提高信息检索性能.

在基于关联规则挖掘的伪相关反馈查询扩展中，词间关联模式支持度的计算方法直接影响到扩展词质量. 当前面向查询扩展的支持度计算可归纳为3类：第一类支持度[4,7]以项集频度作为计算依据，不考虑项集权值，导致冗余和虚假的关联模式增多，其关联模式难以全面反映特征词之间固有联系；第二类支持度[10~12]只考虑项集权值，忽略项集频度，其关联模式中一些固有关联信息难以被发现；第三类支持度将项集频度和项集权值作为计算因子[13~15]，克服了上述两类的缺陷，取得较好挖掘效果，扩展性能得到提升，但是，该类支持度只是将项集频度和权值简单组合，缺乏理论依据，其扩展性能潜力还需进一步研究.

基于词向量学习的伪相关反馈查询扩展[16~18]对伪相关反馈文档集进行词嵌入学习训练得到词向量，通过词向量相似度发现与原查询语义相关的扩展词. 其缺陷是所得到的扩展词忽略了与原查询词之间基于统计学分析的关联信息.

针对上述缺陷，本文利用Copulas理论[19]融合文档中固有的项集频度和项集权值，提出基于Copulas理论的关联模式支持度和置信度，以及基于CSC(Copulas-based Support and Confidence)框架的关联模式挖掘与规则扩展算法，并将关联模式与具有丰富上下文语义信息的词向量融合，提出关联模式挖掘与词向量语义学习融合的查询扩展模型. 实验结果表明，所提扩展模型及算法能有效地减少查询主题漂移和词不匹配问题，提高检索性能.

2 相关概念与词向量语义学习

假设原查询Q的初检文档集记为FRRDS(Q)(Document Set of First Retrieval Results for Query Q)，FRRDS(Q)=(d_1, d_2, …, d_{num})，其中，Q=(q_1, q_2, …, q_r)，num是初检文档数.

2.1 关联模式支持度与置信度

Copulas函数描述的是变量间的相关性[19]，在信息检索查询扩展领域得到应用[20, 21]，取得良好效果. 二元Copulas函数基本理论如下：假设给定二维随机向量$\boldsymbol{X}$=(x_1, x_2)，且具有边缘分布函数$F(x_1)$，$F(x_2)$，根据Sklar's定理[22]，存在一个Copula函数$C(F(x_1),F(x_2))$，使得二维联合分布函数$F(x_1, x_2)$=$C(F(x_1), F(x_2))$，同时，存在如式(1)所示的Copulas统一描述二维随机向量的累积分布函数[20].

$$C(F(x_1),F(x_2)) = e^{\log F(x_1)+\log F(x_2)} \tag{1}$$

借鉴Copulas的累积分布函数，本文提出一种基于Copulas函数的特征词项集I支持度(Copulas-based Support, CSup)和关联规则($I_1\rightarrow I_2$)置信度(Copulas-based Confidence, CConf)计算公式，如式(2)和式(3)所示.

$$\mathrm{CSup}(I) = e^{\log(\frac{n_I}{num})+\log(\frac{w_I}{W})} \tag{2}$$

$$\mathrm{CConf}(I_1 \rightarrow I_2) = e^{\log(\frac{n_I}{n_1})+\log(\frac{w_I}{w_1})} \tag{3}$$

其中，$I_1\cup I_2$=I，$I_1\cap I_2$=$\varnothing$，n_I、n_1分别为项集I和I_1在FRRDS(Q)文档集中出现的频度，w_I、w_1分别为项集I和I_1在FRRDS(Q)中的项集权值，W表示FRRDS(Q)中全体特征词权值总和.

给定最小支持度阈值ms和最小置信度阈值mc，将CSup(I)≥ms的项集I称为频繁项集，将CSup($I_1\cup I_2$)≥ms且CConf($I_1\rightarrow I_2$)≥mc的关联规则$I_1\rightarrow I_2$为强关联规则.

2.2 面向查询扩展的词向量语义学习

词向量是一种基于神经网络模型学习得到的分布式语义表示. 当前，词向量有2类：静态编码的词向量和动态编码的词向量，前者典型训练模型有CBOW[23]、Skip-gram[24]以及Glove模型[25]，Skip-gram模型在语义测试中性能更好，对词的语义描述更准确[26]. 后者为BERT预训练语言模型[27].

本文分别以Skip-gram模型和Glove模型为词向量训练模型，提出一种面向查询扩展的词向量语义学习基本思想，即：使用词向量训练模型，对查询Q的全部初检文档集FRRDS(Q)进行词向量语义学习训练，得到初检文档集特征词词向量集FWEC (Q)(Feature Word Embedding Collection for Query Q)，实现对初检文档特征词的语义向量表征，最后，通过词向量余弦相似度的计算，从语义层面得到与原查询相关的扩展词.

3 关联模式挖掘与词向量学习融合的伪相关反馈查询扩展

3.1 基于CSC框架的关联规则扩展

基于CSC框架的关联规则扩展基本思想：首先原查询检索文档集得到初检文档集FRRDS(Q)，提取前列n篇初检文档建立伪相关反馈文档集PRFDS(Q)(Pseudo Relevance Feedback Document Set for Query Q)，在CSC框架下对PRFDS(Q)文档集挖掘含有原查询词项的频繁项集和强关联规则模式，从关联规则模式中提取前件是原查询词项集的后件项集作为规则扩展词RET(Q)(Rule Expansion Term for Query Q)，并将该关联规则置信度CConf()作为扩展词权值(当多个关联规则模式中同时出现相同的扩展词时，取其置信度值最

大的作为该扩展词的权值),扩展词和原查询组合为新查询再次检索文档,实现查询扩展.

上述查询扩展思想形式化为算法QE_ARCSC (Query Expansion based on Association Rules in the framework of Copulas-based Support and Confidence),其中,IL表示候选项集长度阈值(itemset length), NewQ表示规则扩展词和原查询组合后得到的新查询, FRDoc表示扩展检索后最终检索文档(final retrieval document),n表示初检前列伪相关文档数,q为查询词项集, Et为非查询词项集(即扩展词项集),AR为关联规则集合, FIS为频繁项集集合(Frequent Itemset Set).

算法1　算法名称QE_ARCSC

输入:Q, ms, mc, IL, n.
输出:FRRDS(Q), RET(Q), NewQ, FRDoc.
(1)原查询Q预处理后检索原始文档集得到初检结果文档集FRRDS(Q);
(2)PRFDS(Q)←{从FRRDS(Q)中提取前列n篇初检文档};
(3)对伪相关反馈文档集PRFDS(Q)预处理;
(4)PRFDS(Q)特征词的提取及其权值计算,构建文档索引库(PRFDS_Index)和特征词库(PRFDS_TDB);
(5)从PRFDS_TDB库中提取特征词作为1_候选项集C_1;
(6)扫描PRFDS_Index库统计C_1的频度及其权值,计算CSup (C_1);
(7)L_1={ C_1 | CSup (C_1) >= ms };
(8)FIS←FIS∪L_1;
(9)for (k=2; $L_{k-1}\neq\varnothing$; k++){
　(a)C_k←$L_{k-1}\infty L_{k-1}$;
　(b)if (k=2) then
　　if (C_k不含原查询词项) then 删除C_k;
　(c)统计C_k的频度及其权值,计算CSup(C_k);
　(d)L_k={ C_k | CSup(C_k) >= ms };
　(e)FIS←FIS∪L_k;
　(f)if (k> IL) then Break;}
(10)For FIS 中的k_频繁项集L_k do
　For L_k中项集(q, Et) do
　If ((CConf ($q\rightarrow$Et)>=mc) and ($q\cup$Et =L_k) and ($q\cap$Et =$\varnothing$) and ($q\subseteq Q$)) then
　AR←AR∪{ $q\rightarrow$Et };
(11)从AR中提取关联规则$q\rightarrow$Et的后件项集Et作为扩展词RET(Q);
(12)提取置信度CConf($q\rightarrow$Et)作为扩展词RET(Q)的权值w_{Et}(Q);
(13)NewQ= $Q\cup$RET(Q);
(14)NewQ再次检索原文档集得到FRDoc;
(15)Return FRRDS(Q), RET(Q), NewQ, FRDoc;

QE_ARCSC算法中,步骤(5)~(8)挖掘1_频繁项集,步骤(9)挖掘k_频繁项集($k\geq 2$),步骤(10)挖掘含有查询词项的强关联规则$q\rightarrow$Et,步骤(11)~(13)提取扩展词,构建新查询实现查询扩展.

3.2　关联模式挖掘与词向量学习融合的伪相关反馈查询扩展模型

3.2.1　基本思想

关联模式挖掘与词向量语义学习融合的伪相关反馈查询扩展基本思想是:首先原查询检索文档集得到初检文档集FRRDS(Q),然后从初检文档集中提取前列n篇初检文档作为伪相关反馈文档集PRFDS(Q),调用QE_ARCSC算法对PRFDS(Q)文档集挖掘规则扩展词RET(Q),最后在FWEC(Q)向量集或BERT预训练语言模型向量集中计算规则扩展词RET(Q)与Q的全部查询词项的向量余弦相似度,提取不小于相似度阈值minVSim的前列规则扩展词作为基于PRFDS(Q)的或者基于BERT预训练模型的查询Q最终扩展词RVET(Q)(Rule & Vector-based Expansion Term),实现查询扩展.

根据上述扩展基本思想,本文给出关联模式挖掘与词向量语义学习融合的查询扩展模型结构图,如图1所示,图中,$T(t_1)$,…,$T(t_n)$表示初检文档特征词,$v(t_1+1)$,…,$v(t_n-2)$表示语义学习训练后得到的特征词词向量.

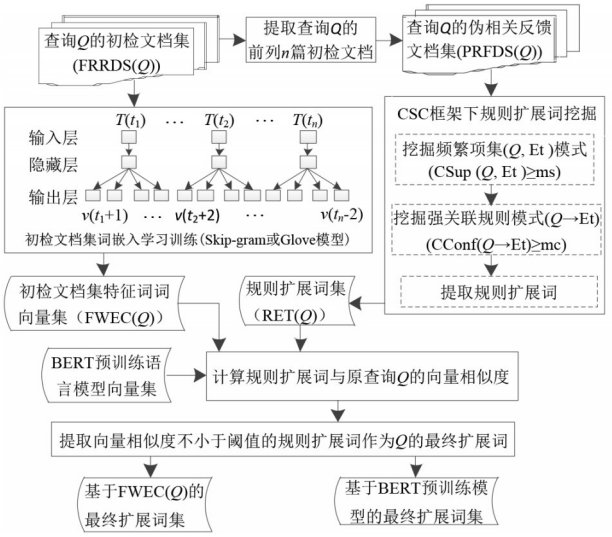

图1　关联模式挖掘与词向量学习融合的伪相关反馈查询扩展模型结构图

3.2.2　查询扩展模型

综上所述,关联模式挖掘与词向量语义学习融合的伪相关反馈查询扩展模型描述如下:

(1)调用QE_ARCSC算法对伪相关反馈文档集PRFDS(Q)挖掘规则扩展词RET(Q),如式(4)所示.

$$\begin{cases}\mathrm{RET}(Q)=\{\mathrm{Ret}_1,\mathrm{Ret}_2,\cdots\mathrm{Ret}_m\}\\ w_{\mathrm{Reti}}(Q)=\mathrm{CConf}(Q\rightarrow \mathrm{Et}_i), i\in 1,2,\cdots,m\end{cases}\tag{4}$$

其中,$w_{\mathrm{Reti}}(Q)$为Q的规则扩展词Rect_i的权值.

(2)对初检文档集FRRDS(Q)进行词向量学习训练,得到初检文档集特征词词向量集FWEC (Q).

(3)在FWEC (Q) 向量集或BERT预训练语言模型向量集中计算规则扩展词与Q各个查询词$\{q_1,q_2,\cdots,q_r\}$的向量余弦相似度,累加其与各个查询词的向量相似度作为该规则扩展词总的向量相似度VecSim(Ret_i,

Q). VecSim(Ret_i, Q)计算公式如式(5)所示.

$$\mathrm{VecSim}(\mathrm{Ret}_i, Q) = \sum_{j=1}^{r} \mathrm{VecSim}(\mathrm{Ret}_i, q_j) \tag{5}$$

(4)根据 VecSim(Ret_i, Q),提取不低于向量相似度阈值 minVSim 的规则扩展词作为原查询 Q 的最终扩展词 RVET(Q),如式(6)所示. RVET(Q)的权值由规则扩展词权值 $w_{\mathrm{Retl}}(Q)$ 及其向量相似度 VecSim(Ret_i, Q)组成,并借鉴式(1),将这两个度量值表示为最终扩展词 RVET(Q)的权值 $w_{\mathrm{RVetl}}(Q)$,如式(7)所示.

$$\begin{aligned}&\mathrm{RVET}(Q) = \{\mathrm{Rvet}_1, \mathrm{Rvet}_2, \cdots, \mathrm{Rvet}_u\}\\&(\mathrm{VecSim}(\mathrm{Rvet}_l, Q) \geqslant \mathrm{minVSim}(l \in (1, 2, \cdots, u)))\end{aligned} \tag{6}$$

$$w_{\mathrm{RVetl}}(Q) = \mathrm{e}^{\log w_{\mathrm{Retl}}(Q) + \log \mathrm{VecSim}\left(\mathrm{Ret}_l, Q\right)} \tag{7}$$

4 实验设计及结果分析

4.1 实验数据及其预处理

实验数据是 NTCIR-5 CLIR(详见:http://research.nii.ac.jp/ntcir/data/data-en.html)中文文本语料,共8个数据集,合计901446篇中文文档,具体信息如表1所示. 该语料有文档集、查询集和结果集,即50个中文查询,4种类型的查询主题和2种评价标准的结果集. 本文采用 Description(简称 Desc)和 Title 查询主题完成检索实验,Title 查询属于短查询,以名词和名词性短语简要描述查询主题,Desc 查询属于长查询,以句子形式简要描述查询主题. 结果集有 Rigid(与查询高度相关,相关)和 Relax(与查询高度相关、相关和部分相关)评价标准.

实验数据预处理是:将文档集和查询集繁体中文转换为简体中文,实验文档集和查询集进行分词和去除停用词.

表1 实验原始语料集及其数量

语料集	简称	文档数量
udn2000	UN0	244038
udn2001	UN1	222526
ude2000	UE0	40445
ude2001	UE1	51851
mhn2000	MN0	84437
mhn2001	MN1	85302
edn2000	EN0	79380
end2001	EN1	93467

4.2 实验设计、基准检索与对比算法

本文关联模式挖掘与词向量语义学习融合的伪相关反馈查询扩展模型实验分2种情况:①采用Skip-Gram模型和Glove模型对初检文档集特征词进行词向量语义学习训练,实现本文查询扩展,分别记为 QE_AP&SG(Query Expasion based on the fusion of Association Pattern mining and Skip-Gram learnig)和 QE_AP&GL算法;②基于BERT预训练语言模型的查询扩展实验,记为 QE_AP&BT,即直接使用现有的BERT预训练语言模型词向量集计算规则扩展词与原查询的向量余弦相似度.

实验基本检索环境采用开源的全文检索引擎开发包 Lucene. Net 搭建,将 Lucene. Net 提供的向量空间检索模型作为基准检索(BaseLine Retrieval, BLR),本文将50个原始查询提交到 Lucene. Net 进行初次检索得到的检索结果作为基准检索结果.

对比方法选择依据是:本文扩展模型涉及关联模式挖掘和词向量语义学习,选择与这两方面相关的近年同类文献作为对比方法,即如下4种对比方法:QE_WAPM,采用文献[12]的加权关联模式挖掘技术挖掘规则扩展词(mc=0.1, mi=0.0001, ms∈(0.004, 0.005, 0.006, 0.007);QE_WPNPM,采用文献[14]的完全加权正负关联模式挖掘技术挖掘规则扩展词(mc=0.1,α=0.3,minPR=0.1,minNR=0.01,ms∈(0.10,0.11,0.12, 0.13));QE_WMSM,采用文献[15]的基于多支持度阈值的加权频繁模式挖掘技术挖掘规则扩展词(mc=0.1,LMS=0.2,HMS=0.25,WT=0.1,ms∈(0.1, 0.15,0.2,0.25));QE_W2Vec,采用文献[18]"Word2Vec 查询扩展方法1",扩展词权值按文献[18]式(9)计算(k=60,α=0.1).

实验参数值选择原则是:本文算法参数 ms 和 mc 通过实验来确定其最佳参数值,其他参数在其有效范围内尽量选择比较有效的参数值进行实验,实验参数值的选择有一定随机性,例如,n=20, IL=2, minVSim=0.1. Skip-gram 参数:batch_size=128, embedding_size=300, skip_window=2, num_skips=4, num_sampled=64. Glove参数:Window-size=10,Vector size=300,menory=4,iter=25,其余为默认值. BERT 预训练语言模型采用 Google 发布的预训练好的中文 BERT 预训练语言模型:"BERT-Base, Chinese"(详见:https://github.com/google-research/bert).

4.3 检索性能比较

本文利用 Google 开源词向量工具 word2vec 实现 Skip-gram 模型,GloVe-win 程序(详见:https://github.com/anoidgit/GloVe-win)实现 Glove 模型,完成初检文档集词向量语义学习训练任务,结合 Lucene. Net,利用 C#+SQL Server 编写实验源程序,完成本文扩展检索实验.

4.3.1 参数 ms 和 mc 的扩展检索性能分析

本节分析参数 ms 和 mc 的不同阈值设置对 QE_ARCSC 扩展性能的影响,以便发现最佳实验参数值. ms 和 mc 分别变化时 QE_ARCSC 算法在 EN0 和 UN0 上实验,结果如图2和图3所示. 图中,后缀"Re"代表 Relax 评价,后缀"Ri"代表 Rigid 评价,"average"代表所有结果值累加后的平均值.

图2和图3表明,随着 ms 逐渐增大,大部分评价指

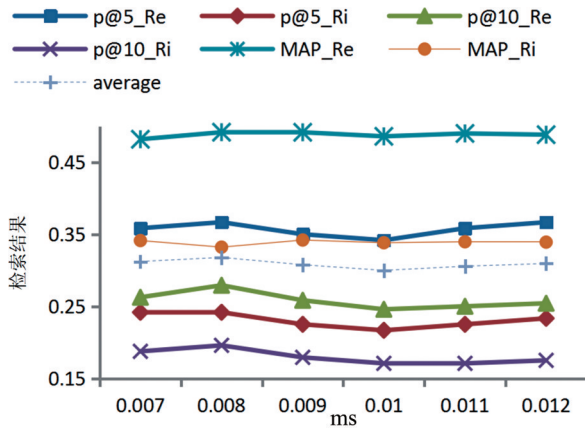

图2　本文算法 ms 设置的扩展检索结果(Title)

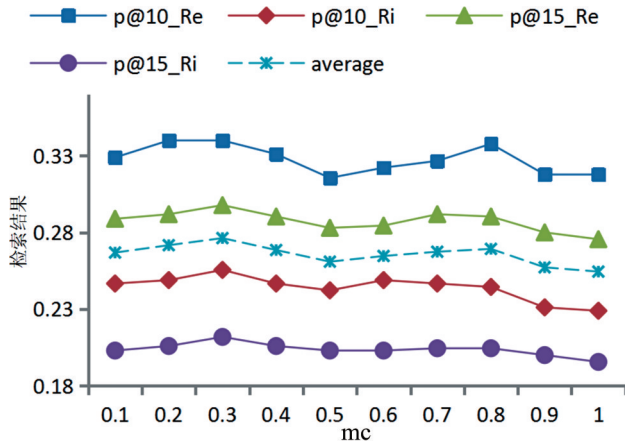

图3　本文算法 mc 设置的扩展检索结果(Desc 查询)

标值开始呈现缓慢上升,当 ms 为 0.008 时,指标值达到较好的状态,ms 大于 0.008,开始出现缓慢下降趋势.对于 mc,情况与上述类似,当 mc 为 0.3 时,其扩展检索效果比较好,随后呈现缓慢下降趋势.综上所述,本文后续实验设置:ms=0.008,mc=0.3.

4.3.2　本文算法与基准、对比方法的检索性能对比

50 个中文查询在 8 个数据集上实验,得到基准检索、对比方法以及本文算法 QE_ARCSC、QE_AP&SG、QE_AP&GL 和 QE_AP&BT 的检索结果 MAP 和 P@5 平均值,限于篇幅,只列举实验结果 MAP 平均值(Title 查询),如表 2 所示.

表 3 和表 4 列举本文算法 MAP 和 P@5 值相对于基准和对比方法在 8 个数据集上的平均增幅.平均增幅计算方法如下:首先计算本文算法检索结果比基准、对比方法在各数据集上的增幅,然后累加各个数据集上的增幅再除以 8 即为平均增幅.例如,表 2 中本文 QE_AP&SG 算法 Title 查询检索结果 MAP 值(Relax)比 QE_WMSM 算法的平均增幅为 11.48%,其计算过程是:{[(0.2939 − 0.2831)/0.2831+(0.3186 − 0.2993)/0.2993+(0.5344 − 0.4375)/0.4375+(0.3208 − 0.2842)/0.2842+(0.3706 − 0.3264)/0.3264+(0.3955 − 0.3783)/0.3783+(0.5060 − 0.4615)/0.4615+(0.2735 − 0.2301)/0.2301]/8}×100%=11.48%.

表 2~4 实验结果表明,本文 QE_ARCSC、QE_AP&SG、QE_AP&GL 和 QE_AP&BT 算法实验结果 MAP 和

表 2　本文算法与基准、对比算法的检索性能 MAP 值(Title 查询)

算法	数据集								评价标准
	UN0	UN1	UE0	UE1	MN0	MN1	EN0	EN1	
BLR	0.2180	0.2679	0.3701	0.2497	0.3049	0.3144	0.4278	0.1992	Relax
QE_WAPM	0.2777	0.2963	0.4130	0.2694	0.3447	0.3517	0.4777	0.2551	
QE_WPNPM	0.2699	0.2714	0.4622	0.2815	0.3481	0.3370	0.4631	0.2010	
QE_WMSM	0.2831	0.2993	0.4375	0.2842	0.3264	0.3783	0.4615	0.2301	
QE_W2Vec	0.2713	0.3122	0.5003	0.3006	0.2871	0.3570	0.4628	0.2122	
QE_ARCSC	**0.2890**	**0.3077**	**0.4983**	**0.3264**	**0.3691**	**0.3742**	**0.4875**	**0.2647**	
QE_AP&SG	**0.2939**	**0.3186**	**0.5344**	**0.3208**	**0.3706**	**0.3955**	**0.5060**	**0.2735**	
QE_AP&GL	**0.2704**	**0.3079**	**0.5362**	**0.3259**	**0.3623**	**0.3675**	**0.5038**	**0.2711**	
QE_AP&BT	**0.2940**	**0.3178**	**0.5316**	**0.3228**	**0.3752**	**0.3787**	**0.5136**	**0.2728**	
BLR	0.1253	0.1839	0.2075	0.1795	0.2089	0.1850	0.2814	0.1200	Rigid
QE_WAPM	0.1597	0.1954	0.2165	0.1661	0.2016	0.1976	0.3313	0.1690	
QE_WPNPM	0.1496	0.1776	0.2452	0.1668	0.2172	0.1856	0.3359	0.1398	
QE_WMSM	0.1596	0.1906	0.2531	0.1787	0.1987	0.2142	0.3145	0.1310	
QE_W2Vec	0.1474	0.2054	0.3056	0.1997	0.1894	0.2147	0.3218	0.1383	
QE_ARCSC	**0.1608**	**0.2043**	**0.2712**	**0.2026**	**0.2335**	**0.2074**	**0.3338**	**0.1728**	
QE_AP&SG	**0.1613**	**0.2155**	**0.3061**	**0.2036**	**0.2412**	**0.2038**	**0.3514**	**0.1752**	
QE_AP&GL	**0.1507**	**0.2036**	**0.3057**	**0.2023**	**0.2281**	**0.1971**	**0.3501**	**0.1717**	
QE_AP&BT	**0.1615**	**0.2092**	**0.3059**	**0.2004**	**0.2385**	**0.2064**	**0.3553**	**0.1749**	

P@5值都高于基准检索BLR;相对于4种对比方法,本文算法MAP和P@5值绝大部分都得到提升;本文算法中,QE_AP&SG和QE_AP&BT算法MAP和P@5值绝大部分略高于QE_ARCSC和QE_AP&GL算法的,具体情况描述如下:

(1)相对基准BLR,获得最好的MAP值平均增幅(%)是QE_AP&SG算法(其增幅范围为24.93~28.69)和QE_AP&BT(23.36~28.36),其次是QE_AP&GL,最后是QE_ARCSC. P@5平均增幅(%)最好的是QE_AP&SG(增幅范围是14.9~23.58)和QE_AP&BT(16.38~23.54),其次是QE_ARCSC,最后是QE_AP&GL.

(2)相对3种基于关联模式挖掘的同类对比方法(QE_WAPM, QE_WPNPM和QE_WMSM),获得最好MAP平均增幅(%)是算法QE_AP&SG(增幅范围是8.26~17.52),其次是QE_AP&BT(8.03~16.13),再次是QE_ARCSC,最后是QE_AP&GL. P@5平均增幅(%)最好的是算法QE_AP&SG(2.67~13.66)和QE_AP&BT(4.06~12.65),其次是QE_AP&GL,最后是QE_ARCSC.

(3)相对基于词向量的查询扩展对比方法(QE_W2Vec),MAP平均增幅(%)最好的是算法QE_AP&SG(7.67~12.75)和QE_AP&BT(7.13~12.51),其次是QE_AP&GL和QE_ARCSC. P@5平均增幅(%)最好的是QE_AP&SG(2.02~14.76)和QE_AP&BT(2.49~14.12),其次是QE_AP&GL,最后是QE_ARCSC.

综上所述,本文QE_ARCSC、QE_AP&SG、QE_AP&GL和QE_AP&BT算法扩展检索性能优于基准检索和4种对比方法,具有如下特点:(1)QE_AP&SG和QE_AP&BT算法获得最好的扩展检索性能,其次是QE_AP&GL,最后是QE_ARCSC,说明关联模式挖掘与词向量语义学习融合后能获得比较好的查询扩展性能;(2)Title查询的检索结果MAP和P@5平均增幅高于Desc查询,说明本文扩展算法对短查询(Title)的扩展检索性能更有效;(3)MAP平均增幅普遍高于P@5的,说明本文扩展算法的整体扩展检索性能比较好;(4)QE_AP&SG和QE_AP&BT算法MAP和P@5平均增幅都高于QE_ARCSC的,QE_AP&GL算法短查询检索结果MAP和P@5平均增幅略高于QE_ARCSC的,说明关联模式挖掘与词向量语义学习融合后的查询扩展检索性能优于单纯基于关联模式挖掘的查询扩展.

存在不足之处:与对比算法比较,本文QE_AP&GL算法长查询(Desc)检索结果P@5值还存在负值,说明该算法的扩展性能存在不提升反而降低现象,即扩展性能不稳定,QE_ARCSC算法也存在类似情况. 这些问题是本文后续要重点研究的问题.

表3 本文算法检索结果MAP值在8个数据集上的平均增幅(%)

查询类型(评价标准)	本文算法	基准检索和对比算法				
		BLR	QE_WAPM	QE_WPNPM	QE_WMSM	QE_W2Vec
Title(Relax)	QE_ARCSC	24.96	8.62	12.28	8.28	9.59
	QE_AP&SG	28.69	11.86	15.63	**11.48**	12.75
	QE_AP&GL	25.49	9.18	12.78	8.84	9.95
	QE_AP&BT	28.36	11.58	15.31	11.25	12.51
Title(Rigid)	QE_ARCSC	21.18	9.52	12.1	10.09	5.91
	QE_AP&SG	25.41	13.46	15.86	13.87	9.32
	QE_AP&GL	21.78	10.22	12.45	10.56	5.99
	QE_AP&BT	24.93	12.95	15.32	13.35	8.85
Desc(Relax)	QE_ARCSC	23.88	6.62	14.15	12.95	6.22
	QE_AP&SG	25.67	8.26	15.71	14.7	7.67
	QE_AP&GL	21.74	5.02	12.34	11.31	4.46
	QE_AP&BT	25.99	8.55	16.13	15.04	7.99
Desc(Rigid)	QE_ARCSC	22.21	7.02	14.46	14.99	6.26
	QE_AP&SG	24.93	9.45	16.9	17.52	8.43
	QE_AP&GL	19.71	4.96	12.28	12.86	4.19
	QE_AP&BT	23.36	8.03	15.51	16.11	7.13

表4 本文算法检索结果P@5值在8个数据集上的平均增幅(%)

查询类型(评价标准)	本文算法	基准检索和对比算法				
		BLR	QE_WAPM	QE_WPNPM	QE_WMSM	QE_W2Vec
Title(Relax)	QE_ARCSC	22.21	4.95	10.41	9.96	12.73
	QE_AP&SG	23.58	6.30	11.62	11.04	13.53
	QE_AP&GL	23.48	6.13	11.78	11.35	14.16
	QE_AP&BT	23.54	6.23	11.90	11.31	14.12
Title(Rigid)	QE_ARCSC	16.70	4.25	8.36	11.39	12.93
	QE_AP&SG	19.81	7.29	10.98	13.66	14.76
	QE_AP&GL	17.99	5.50	9.32	12.16	13.55
	QE_AP&BT	18.08	5.59	9.66	12.65	14.05
Desc(Relax)	QE_ARCSC	12.77	0.84	5.15	4.57	−0.71
	QE_AP&SG	14.9	2.67	7.08	6.38	1.08
	QE_AP&GL	11.51	−0.29	4.14	3.32	−1.70
	QE_AP&BT	16.38	4.06	8.50	7.91	2.49
Desc(Rigid)	QE_ARCSC	14.06	1.10	4.05	3.81	−0.99
	QE_AP&SG	17.61	4.33	7.32	7.28	2.02
	QE_AP&GL	15.16	1.86	4.98	4.66	0.00
	QE_AP&BT	19.11	5.70	8.63	8.72	3.20

4.3.3 查询实例扩展检索效果分析

本节以本文算法QE_AP&SG和QE_ARCSC为例,列举NTCIR-5 CLIR语料No. 4(查询美国国防部长柯恩

于2000年6月访问北京的相关报导)和No. 42(查询被认为对于经济复苏有贡献的美国联邦准备理事会主席葛林斯班所提出的货币政策)查询Desc主题实例在UE0数据集上进行基准检索、对比方法和本文扩展算法等检索实验的具体结果,进一步说明本文扩展算法能有效地遏制查询主题漂移和词不匹配问题. 表5列举的是基准检索和各扩展算法对于查询实例的检索结果MAP值.

表5　查询实例Desc主题的MAP值(UE0)比较

查询	算法	Relax	Rigid
No. 4	BLR	0. 1257	0. 1051
	QE_WAPM	0. 2870	0. 2316
	QE_WPNPM	0. 3722	0. 2400
	QE_WMSM	0. 3299	0. 2128
	QE_W2Vec	0. 4364	0. 2887
	QE_APMCC	0. 3607	0. 2033
	QE_AP&SG	0. 4426	0. 2924
No. 42	BLR	0. 5000	0. 5000
	QE_WAPM	0. 5833	0. 5833
	QE_WPNPM	0. 5000	0. 5000
	QE_WMSM	0. 5833	0. 5833
	QE_W2Vec	0. 5000	0. 5000
	QE_APMCC	0. 5833	0. 5833
	QE_AP&SG	0. 7500	0. 7500

从表5可知,两种查询实例DESC主题的检索实验中,本文QE_AP&SG算法获得最高的MAP值,均高于基准检索和所有对比方法,本文QE_ARCSC的MAP值高于基准检索和部分对比方法,存在低于一些对比方法的MAP值的情况. 由此可见,本文QE_AP&SG和QE_ARCSC算法确实能有效地遏制查询主题漂移和词不匹配问题,而QE_AP&SG算法表现出最好的效果,说明关联模式挖掘与词向量语义学习的融合更能有效地遏制查询主题漂移和词不匹配问题.

4.4　实验结果分析

综上所述,本文查询扩展模型及算法是有效的,能改善信息检索性能,有效遏制查询主题漂移和词不匹配问题,具体表现为:(1)本文扩展算法检索性能优于基准检索和同类近年出现的对比方法;(2)短查询的扩展检索性能优于长查询,说明本文算法对于短查询的检索性能提升更有利;(3)MAP值平均增幅普遍高于P@5值的平均增幅,说明本文扩展算法的整体扩展检索性能比较好;(4)QE_AP&SG、QE_AP&BT和QE_AP&GL算法的扩展检索性能优于QE_ARCSC算法,说明单纯基于关联模式的查询扩展检索性能不如关联模式挖掘与词向量语义学习融合后的查询扩展检索性能;(5)QE_AP&SG、QE_AP&BT的扩展检索性能优于QE_AP&GL算法,说明对于初检文档集(小样本数据集)来说,面向查询扩展的Skip-Gram模型词向量语义训练效果比Glove模型训练效果好,以及关联模式挖掘与BERT预训练语言模型的融合能取得较好的扩展性能. 由此可见,关联模式挖掘与BERT预训练语言模型的结合更有优势,因为,采用预训练语言模型可以节省初检文档集的词向量训练时间,具有较高的应用价值,因此,QE_AP&BT算法比QE_AP&SG更有优势.

本文查询扩展模型的有效性得益于如下3个方面的改进:(1)提出了一种基于Copulas理论的关联模式支持度和置信度,使得支持度和置信度更能客观地反映项集之间的固有关联,提高扩展词质量;(2)提出了基于CSC框架的关联规则扩展方法,该方法采用基于Copulas理论的关联模式支持度和置信度,扩展词质量得到有效提升,扩展检索性能高于对比方法;(3)对于扩展词与原查询之间的关联分析,本文不仅考虑来自统计学分析和挖掘的关联信息,还考虑具有上下文语义信息的词向量信息,将加权关联模式挖掘与词向量语义学习融合,提出一种关联模式挖掘与词向量语义学习融合的查询扩展模型,提高了扩展词的质量,扩展检索性能得到较好的改善. 以上3个方面共同作用,使得本文查询扩展模型优于基准算法和近年同类的对比扩展方法,有效地改善信息检索性能,减少伪相关反馈引起的查询主题漂移和词不匹配问题.

5　结论

本文将基于统计学分析的关联模式与具有上下文语义信息的词向量融合,提出一种关联模式挖掘与词向量语义学习融合的伪相关反馈查询扩展模型. 该模型挖掘伪相关反馈文档集中规则扩展词,计算规则扩展词与原查询的向量相似度,根据向量相似度提取前列扩展词作为最终扩展词. 实验结果表明,本文扩展模型及算法有效,关联模式挖掘与词向量语义学习或者预训练语言模型的融合能遏制伪相关反馈引起的查询主题漂移和词不匹配问题,提高检索性能,关联模式挖掘与BERT预训练语言模型的结合有很高的实际应用价值. 本文扩展模型对短查询更有效,可用于跨语言检索系统,提升系统检索性能,所提出的支持度计算方法可用于其他文本挖掘任务和推荐系统,提高推荐系统准确性. 本文后续工作是继续优化扩展模型及其参数,探讨将本文扩展方法应用于实际的信息检索系统.

致谢　感谢匿名外审专家的修改意见以及编辑部老师的辛勤工作.

参考文献

[1] Vaidyanathan R, Das S, Srivastava N. Query expansion strategy based on pseudo relevance feedback and term

weight scheme for monolingual retrieval[J]. International Journal of Computer Applications, 2015, 105(8):1 – 6.

[2] Keikha A, Ensan F, Bagheri E. Query expansion using pseudo relevance feedback on Wikipedia[J]. Journal of Intelligent Information Systems, 2018, 50(3): 455 – 478.

[3] Pan M, Huang J, He T, et al. A simple kernel co-occurrence-based enhancement for pseudo-relevance feedback[J]. Journal of the Association for Information Science and Technology (JASIST), 2020, 71(3): 264 – 281.

[4] Latiri C, Haddad H, Hamrouni T. Towards an effective automatic query expansion process using an association rule mining approach[J]. Journal of Intelligent Information Systems, 2012, 39(1): 209 – 247.

[5] Bouziri A, Latiri C, Gaussier E, et al. Learning query expansion from association rules between terms[A]. Fred A, Dietz J, Aveiro D, et al. Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)[C]. Lisbon, Portugal: Scitepress, 2015. 525 – 530.

[6] Bouziri A, Latiri C, Gaussier E. Efficient association rules selecting for automatic query expansion[A]. Gelbukh A. Proceedings of the 18th International Conference on Computational Linguistics & Intelligent Text Processing[C]. Budapest, Hungary: Springer, 2017. 563 – 574.

[7] Bouziri A, Latiri C, Gaussier E. LTR-expand: Query Expansion Model Based on Learning to Rank Association Rules [EB/OL]. https://doi.org/10.1007/s10844-020-00596-8, 2020.03.21/2020.08.15.

[8] Jabri S, Dahbi A, Gadi T, et al. Improving retrieval performance based on query expansion with Wikipedia and text mining technique[J]. International Journal of Intelligent Engineering & Systems, 2018, 11(4): 283 – 292.

[9] Jabri S, Dahbi A, Gadi T. A graph-based approach for text query expansion using pseudo relevance feedback and association rules mining[J]. International Journal of Electrical&Computer Engineering, 2019, 9(6): 5016 – 5023.

[10] 黄名选, 严小卫, 张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报, 2009, 20(7): 1854 – 1865.
Huang MX, Yan XW, Zhang SC. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining[J]. Journal of Software, 2009, 20(7): 1854 – 1865. (in Chinese)

[11] 黄名选. 完全加权模式挖掘与相关反馈融合的印尼汉跨语言查询扩展[J]. 小型微型计算机系统, 2017, 38(8): 1783 – 1791.
HUANG Ming-xuan. Indonesian-Chinese cross language query expansion based on all-weighted patterns mining and relevance feedback[J]. Journal of Chinese Computer Systems, 2017, 38(8): 1783 – 1791. (in Chinese)

[12] 黄名选. 基于加权关联模式挖掘的越-英跨语言查询扩展[J]. 情报学报,2017,36(3): 307 – 318.
HUANG Ming-xuan. Vietnamese-English cross language query expansion based on weighted association patterns mining[J]. Journal of the China Society for Scientific and Technical Information, 2017, 36(3): 307 – 318. (in Chinese)

[13] 黄名选, 蒋曹清. 基于项权值排序挖掘的跨语言查询扩展[J]. 电子学报, 2020,48(3): 568 – 576.
HUANG Ming-xuan, JIANG Cao-qing. Cross language query expansion based on item weight sorting mining[J]. Acta Electronica Sinica, 2020,48(3): 568 – 576. (in Chinese)

[14] 黄名选, 蒋曹清. 基于完全加权正负关联模式挖掘的越-英跨语言查询译后扩展[J]. 电子学报, 2018, 46(12): 3029 – 3036.
HUANG Ming-xuan, JIANG Cao-qing. Vietnamese-English cross language query post-translation expansion based on all-weighted positive and negative association patterns mining[J]. Acta Electronica Sinica, 2018, 46(12): 3029 – 3036. (in Chinese)

[15] Zhang H R, Zhang J W, Wei X Y, et al. A new frequent pattern mining algorithm with weighted multiple minimum supports[J]. Intelligent Automation & Soft Computing, 2017, 23(4): 605 – 612.

[16] Roy D, Ganguly D, Mitra M, et al. Word vector compositionality based relevance feedback using kernel density estimation[A]. Proceedings of the 25th ACM International Conference on Information and Knowledge Management [C]. New York, USA: ACM Press, 2016. 1281 – 1290.

[17] Kuzi S, Shtok A, Kurland O. Query expansion using word embeddings[A]. Proceedings of the 25th ACM International Conference on Information and Knowledge Management [C]. New York, USA: ACM Press, 2016. 1929 – 1932.

[18] 许侃, 林原, 曲忱, 等. 专利查询扩展的词向量方法研究[J]. 计算机科学与探索, 2018, 12(6): 972 – 980.
XU Kan, LIN Yuan, QU Chen, et al. Research on patent query expansion methods using word embedding[J]. Journal of Frontiers of Computer Science and Technology, 2018, 12(6): 972 – 980. (in Chinese)

[19] Sklar A. Fonctions de repartition à n dimensions et leurs marges[J]. Publication de l'Institut de Statistique l'Universite Paris, 1959, 8(1): 229 – 231.

[20] Eickhoff C, De Vries A P, Collins-Thompson K. Copulas

for information retrieval[A]. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'13) [C]. New York, USA: ACM Press, 2013. 663 – 672.

[21] 张书波, 张引, 张斌, 等. 基于Copulas框架的混合式查询扩展方法[J]. 计算机科学, 2016,43(6A):485 – 488 496.
ZHANG Shu-bo, ZHANG Yin, ZHANG Bin, et al. Combined query expansion method based on copulas framework[J]. Computer Science, 2016,43(6A): 485 – 488, 496. (in Chinese)

[22] Nelson R B. An Introduction to Copulas(Second Edition) [M]. New York, USA: Springer Science+Business Media, Inc, 2006. 17 – 22.

[23] Mikolov T, Chen K, Corradog G, et al. Efficient Estimation of Word Representations in Vector Space[EB/OL]. https://arxiv.org/pdf/1301.3781v3.pdf. arXiv:1301.3781v3[cs.CL] 7 Sep 2013/2020.08.15.

[24] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [A]. Burges C J C, Bottou L, Welling M. Proceedings of Advances in Neural Information Processing Systems(NIPS 2013)[C]. New York, USA: Curran Associates Inc, 2013. 3111 – 3119.

[25] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[A]. Moschitti A, Pang B, Daelemans W. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) [C]. Doha, Qatar: Association for Computational Linguistics, 2014. 1532 – 1543.

[26] 张剑, 屈丹, 李真. 基于词向量特征的循环神经网络语言模型[J]. 模式识别与人工智能, 2015, 28(4):299 – 305.
ZHANG Jian, QU Dan, LI Zhen. Recurrent neural network language model based on word vector features[J]. PR & AI, 2015, 28(4): 299 – 305. (in Chinese)

[27] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. https://arxiv. org/pdf/1810.04805. pdf, arXiv:1810. 04805v2 [cs.CL] 24 May 2019/2020.08.15.

作者简介

黄名选　男,1966年出生于广西乐业县,硕士,现为广西财经学院教授,硕士研究生导师,主要研究方向为数据挖掘、信息检索、机器学习,主持国家自然科学基金项目2项,主持完成广西自然科学基金项目1项和广西教育厅科研项目3项,获2011年广西高校优秀人才资助计划项目1项,参与完成国家自然科学基金项目1项,发表学术论文60余篇,其中,中文核心期刊论文40余篇,期刊EI收录7篇,ISTP收录1篇,发明专利授权15件.
E-mail: mingxh05@163.com