基于项权值排序挖掘的跨语言查询扩展

黄名选[1,2,3],蒋曹清[3]
(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.广西财经学院广西(东盟)财经研究中心,广西南宁 530003;3.广西财经学院信息与统计学院,广西南宁 530003)

摘 要: 为了改善自然语言处理应用中长期存在的主题漂移和词不匹配问题,本文首先提出一种加权项集支持度计算方法和基于项权值排序的剪枝方法,给出面向查询扩展的基于项权值排序的加权关联规则挖掘算法,讨论关联规则混合扩展、后件扩展和前件扩展模型,最后提出基于项权值排序挖掘的跨语言查询扩展算法.该算法采用新的支持度和剪枝策略挖掘加权关联规则,根据扩展模型从规则中提取高质量扩展词实现跨语言查询扩展.实验结果表明,与现有基于加权关联规则挖掘的跨语言扩展算法比较,本文扩展算法能有效遏制查询主题漂移和词不匹配问题,可用于各种语言的信息检索以改善检索性能,扩展模型中后件扩展获得最优检索性能,混合扩展的检索性能不如后件扩展和前件扩展,支持度对后件扩展更有效,置信度更有利于提升前件扩展和混合扩展的检索性能.本文挖掘方法可用于文本挖掘、商务数据挖掘和推荐系统以提高其挖掘性能.

关键词: 自然语言处理;文本挖掘;信息检索;跨语言检索;查询扩展;推荐系统
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2020)03-0568-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.03.021

Cross Language Query Expansion Based on Item Weight Sorting Mining

HUANG Ming-xuan[1,2,3], JIANG Cao-qing[3]
(1. *Guangxi Key Laboratory of Cross-border E-commerce Intelligent Information Processing, Guangxi University of Finance and Economics, Nanning, Guangxi* 530003, *China*; 2. *Guangxi (ASEAN) Financial Research Center, Guangxi University of Finance and Economics, Nanning, Guangxi* 530003, *China*; 3. *School of Information and Statistics, Guangxi University of Finance and Economics, Nanning, Guangxi* 530003, *China*)

Abstract: To ameliorate the long-standing problems of theme drift and word mismatch in natural language processing applications, this paper first proposes a computing method for weighted itemset support and a pruning method based on item weight sorting (IWS). And then, a weighted association rule mining algorithm for query expansion is presented based on the IWS, and the models such as association rule antecedent and consequent hybrid expansion (RACHE), rule consequent expansion (RCE) along with rule antecedent expansion (RAE) are discussed. Finally, an algorithm of cross-language query expansion (CLQE) is put forward based on the IWS mining. The algorithm utilized the new support and the pruning method to mine the weighted association rules, and extracted high quality expansion terms from the rules according to the expansion models in order to carry out CLQE. A comparison between the proposed expansion algorithm and the existing CLQE algorithms based on weighted association rules mining is made, which shows that the former can effectively restrain the problems of query topic drift and word mismatch, and can be used in information retrieval in various languages to improve retrieval performance. The RCE achieves the optimal retrieval performance in the proposed expansion models, and the retrieval performance of the RACHE is not as good as that of the RAE and the RCE. The support is more effective for the RCE algorithm. The confidence can make the RAE and the RACHE get the best retrieval result. And moreover, the proposed mining method can be used in text mining, business data mining and recommendation system to improve its mining performance.

Key words: natural language processing; text mining; information retrieval; cross language retrieval; query expansion; recommendation system

---

收稿日期:2019-03-18;修回日期:2019-05-17;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61762006,No.61662003);广西应用经济学一流学科(培育)开放性课题(No.2018MA07);广西(东盟)财经研究中心开放性课题(No.2018DMCJYB08)

1 引言

在自然语言处理应用中,跨语言查询扩展是解决查询主题漂移、词不匹配问题的核心技术之一,分为查询译前扩展、查询译后扩展和混合式查询扩展三种. 其关键是扩展词的来源及其扩展模型的设计问题. 跨语言查询扩展早期主要进行比较性和实验性研究[1,2],直到 2010 年以后,查询译前扩展研究[3,4]才得到关注. 随着机器翻译准确率不断提高,查询译后扩展成为一个研究热点,其成果主要集中在基于马可夫链模型的[5],相关反馈的[6,7]和基于关联规则模式挖掘的[8~13]跨语言查询译后扩展. 基于关联规则模式挖掘的跨语言查询译后扩展的关键是如何挖掘到那些与源语言查询对应或相关的优质目标语言查询词项[8],以及那些与原查询相关的高质量扩展词[9~13]. 前者的思路是在平行语料中挖掘关联模式得到目标语言查询词项,实现跨语言检索. 后者是采用关联挖掘技术在目标语言文档集中挖掘与目标语言原查询相关的扩展词,实现跨语言查询译后扩展.

当前,基于关联规则挖掘的跨语言查询扩展研究存在如下问题:(1)扩展词与原查询词之间隐含的各种复杂关联的挖掘问题,到目前为止,还没有找到一种最优的、最普遍适用各种跨语言查询扩展的支持度计算方法和关联模式评估框架;(2)在跨语言查询扩展模型方面,关联规则后件扩展模型受到关注比较多,而关联规则前件扩展模型研究不多,忽略对关联规则前后件混合扩展(简称规则混合扩展)模型的讨论和研究.

针对上述问题,本文聚焦于查询译后扩展研究,首先提出一种新的加权项集支持度计算方法和基于项权值排序的剪枝方法,然后,给出面向查询扩展的基于支持度-置信度-相关系数评价框架和项权值排序的加权关联规则挖掘方法,最后,提出基于项权值排序挖掘的跨语言查询扩展模型及其扩展算法.

2 面向跨语言查询扩展的基于项权值排序的加权关联规则挖掘

2.1 基本概念及其相关定理

2.1.1 加权项集支持度

设 DS(Document Set)为跨语言初检相关文档集,其中每篇文档特征词及其对应权值分别表示为 $t_1,t_2,\cdots,t_m$ 和 $w_{i1},w_{i2},\cdots,w_{im}$,$n$ 是 DS 的文档总数,W 为 DS 中所有特征词项目权值总和,n_I、w_I分别为加权项集 I 在 DS 中频度和项集权值,则给出加权项集 I 支持度(Weighted Itemset Support,WIS)计算公式,如式(1)所示.

$$\mathrm{WIS}(I)=\frac{w_I\times n_I}{W\times n\times k_I} \tag{1}$$

其中,k_I为项集 I 的长度. 设 ms 为最小支持度阈值,若 $\mathrm{WIS}(I)\geqslant \mathrm{ms}$,则称项集 I 为加权频繁项集.

2.1.2 加权关联规则置信度和相关系数

基于传统的置信度思想[14],给出基于式(1)的特征词加权关联规则置信度(Weighted Association Rule Confidence,WARC)计算公式如式(2)所示.

$$\mathrm{WARC}(I_1\rightarrow I_2)=\frac{w_I\times n_I\times k_1}{w_1\times n_1\times k_I} \tag{2}$$

其中,$I=I_1\cup I_2$,$I_1\cap I_2=\varnothing$,n_1、w_1分别为项集 I_1 在 DS 中出现的频度和项集权值,k_1为 I_1 的长度,n_I、w_I和 k_I同式(1).

借鉴传统的相关系数定义[14],结合式(1),给出加权项集相关系数(Weighted Itemset Correlation Coefficient,WICC)的计算公式,如式(3)所示,其中,k_2 和 k_{12} 分别为项集 I_2和(I_1,I_2)的项集长度,n_2 和 n_{12}分别为项集 I_2和(I_1,I_2)在 DS 中出现的次数,w_2 和 w_{12}分别为项集 I_2和(I_1,I_2)在 DS 中的项集权值.

$$\mathrm{WICC}(I_1,I_2)=\frac{w\times n\times w_{12}\times n_{12}\times k_1\times k_2-w_1\times w_2\times n_1\times n_2\times k_{12}}{k_{12}\times\sqrt{w_1\times w_2\times n_1\times n_2\times(w\times n\times k_1-w_1\times n_1)\times(w\times n\times k_2-w_2\times n_2)}} \tag{3}$$

假设最小置信度阈值为 mc,最小相关系数阈值为 mincc(mincc≥0),若 $\mathrm{WARC}(I_1\rightarrow I_2)\geqslant \mathrm{mc}$,且 $\mathrm{WICC}(I_1,I_2)\geqslant \mathrm{mincc}$,则称 $I_1\rightarrow I_2$为强加权关联规则模式.

2.1.3 基于项权值排序的相关定理

设加权项集 $I=(i_1,i_2,\cdots,i_k)$,将 I 的项目权值排降序后得到的项目权值为$(w_1,w_2,\cdots,w_k)$,其中,$w_1\geqslant w_2\geqslant\cdots\geqslant w_k$,本文将项目权值排降序后的加权项集 I 称为权值排序项集. 后续讨论权值排序子集是指按项目权值由高到低从权值排序项集 I 中抽取项目并组合得到的真子集,即子集 $I_1=(i_1)$,$I_{12}=(i_1,i_2)$,$I_{123}=(i_1,i_2,i_3)$,$\cdots$,$I_{123\cdots(k-1)}=(i_1,i_2,\cdots,i_{k-1})$. 通过对权值排序项集的深入分析,容易证明定理 1 和定理 2,限于篇幅,省略其证明过程.

定理 1　加权项集 I 的权值排序子集 $I_1,I_{12},I_{123},\cdots,I_{123\cdots(k-1)}$的权值分别大于或者等于$(w_1)$,$(w_1+w_2)$,$(w_1+w_2+w_3)$,$(w_1+w_2+w_3+\cdots+w_{k-1})$,其子集项频度 n_{sub}分别大于或者等于加权项集频度 n_I.

(证明过程略)

定理 2　如果权值排序项集 I 的项目权值最高者 $w_1<\min\_\mathrm{w}$,则项集 I 一定是非频繁的,其中,$\min_\mathrm{w}=(W\times n\times \mathrm{ms})/n_I$称为最小权值阈值.

(证明过程略)

定理 3　如果加权项集 I 的权值排序子集 $I_1,I_{12},I_{123},\cdots,I_{123\cdots(k-1)}$中存在非频繁项集,那么项集 I 一定非频繁的.

证明　设权值排序项集 I_{sub}是权值排序项集 I 的任

一子集,其相关参数含义如表 1 所示.

表 1 权值排序项集 I 和 I_{sub} 参数表

参数名称	项集 I 的参数	项集 I_{sub} 的参数
项目组合	$I=(i_1,i_2,\cdots,i_r,i_{(r+1)}\cdots,i_k)$ $(1\leqslant r<k)$	$I_{sub}=(i_1,i_2,\cdots,i_r)$ $(1\leqslant r<k)$
项集权值	$w_I=w_{(1\sim r)}+w_{(1\sim k)}$,其中, $w_{(1\sim r)}=w_1+w_2+w_3+\cdots+w_r$, $w_{(1\sim k)}=w_{r+1}+w_{r+2}+\cdots+w_k$	w_{sub}
项集频度	n_I	n_s
项集长度	$r+k$	r

由定理 1 及已知(I_{sub}是非频繁的)可得,

$w_{(1\sim r)}\leqslant w_{sub}, n_I\leqslant n_s$,

$$\mathrm{WIS}(I_{sub})=\frac{w_{sub}\times n_s}{W\times n\times r}<\mathrm{ms}$$

$$\Rightarrow \frac{w_{(1\sim r)}\times n_I}{W\times n\times r}\leqslant\frac{w_{(1\sim r)}\times n_s}{W\times n\times r}\leqslant\frac{w_{sub}\times n_s}{W\times n\times r}<\mathrm{ms}$$

$$\Rightarrow \frac{w_{(1\sim r)}\times n_I}{W\times n\times r}<\mathrm{ms} \tag{4}$$

$$\mathrm{WIS}(I)-\frac{w_{(1\sim r)}\times n_s}{W\times n\times r}=\frac{(w_{(1\sim r)}+w_{(1\sim k)})\times n_s}{W\times n\times(r+k)}-\frac{w_{(1\sim r)}\times n_s}{W\times n\times r}$$

$$=\frac{n_s}{W\times n\times(1+\frac{r}{k})}\left(\frac{w_{(1\sim k)}}{k}-\frac{w_{(1\sim r)}}{r}\right) \tag{5}$$

$\because I$ 为权值排序项集,

$$\therefore \frac{w_{r+1}+w_{r+2}+\cdots+w_k}{k}\leqslant\frac{w_1+w_2+\cdots+w_r}{r}$$

$$\Rightarrow \frac{w_{(1\sim k)}}{k}-\frac{w_{(1\sim r)}}{r}\leqslant 0 \tag{6}$$

$\because n_I>0, W\times n\times(1+\frac{r}{k})>0$,由式(4)、(5)和(6)可得 $\mathrm{WIS}(I)<\mathrm{ms}$,即项集 I 是非频繁的.

证毕.

2.2 面向跨语言查询扩展的基于项权值排序的加权关联规则挖掘

2.2.1 初检相关反馈文档特征词权值计算

设 w_{ij}表示文档 d_i中特征词 t_j的权值,df_j表示含有特征词 t_j的文档数量,$\mathrm{tf}_{j,i}$表示特征词 t_j在文档 d_i中的词频,$\max(\mathrm{tf}_i)$ 表示文档 d_i中出现的最大词频,则本文提出 DS 文档集特征词权值计算方法,如式(7)所示.

$$w_{ij}=\frac{\max(\mathrm{tf}_i)+\mathrm{tf}_{j,i}}{2\times\max(\mathrm{tf}_i)\times(\lg n-\lg(\mathrm{df}_j)+1)} \tag{7}$$

2.2.2 基于项权值排序的剪枝方法

基于项权值排序的剪枝方法如下:①剪除那些不含原查询词项的候选 2_项集;②构建权值排序候选 k_项集 $C_k(w_1,w_2,\cdots,w_k)$,其中 $k\geqslant 2, w_1\geqslant w_2\geqslant\cdots\geqslant w_k$,若 $w_1<\min\_w$,据定理 2,剪枝该候选项集 C_k,否则,若存在 C_k的权值排序子集是非频繁的,根据定理 3,剪枝该候选项集 C_k.

2.2.3 加权关联规则挖掘基本思想及算法

基本思想 对 DS 文档集只挖掘包含原查询词项的特征词加权关联规则模式,即采用新的支持度计算方法和剪枝策略挖掘含有原查询词项的加权频繁项集,通过置信度-相关系数评价框架从频繁项集中挖掘特征词加权关联规则模式.

上述挖掘思想形式化为 WARM_IWS_CLQE (Weighted Association Rules Mining Based on Item Weight Sorting for CLQE) 算法.算法中符号含义:Q 为英文查询项集合,L_{item}为候选项集长度阈值,WAR 为强加权关联规则集合,FIS 为(Frequent ItemSet)特征词频繁项集.

算法 1 WARM_IWS_CLQE 算法

输入:DS, ms, mc, Q, L_{item}.
输出:WAR.
(1) Preconditioning (DS);{
①对 DS 进行英文词干提取和剪除英文停用词;
②计算特征词权值,构建相关反馈文档索引库(DS_DB)和总特征词库(DS_Terms);}
(2) L_1 = MiningWISL1(DS_DB, DS_Terms, ms);{
①从 DS_Terms 库中提取特征词作为候选 1_项集 C_1;
②扫描 DS_DB 库统计候选 1_项集 C_1的权值及其频度;
③计算 WIS(C_1);
④$L_1=\{C_1\mid \mathrm{WIS}(C_1)>=\mathrm{ms}\}$;
⑤FIS ← FIS∪L_1;}
(3) for ($k=2; L_k\neq\varnothing; k++$) {
①$C_k\leftarrow L_{k-1}\infty L_{k-1}$;
②if ($k=2$) then $C_k\leftarrow$ C2_PruningNotQ(C_k, Q);
③扫描 DS_DB 库,构建权值排序候选 k_项集 $C_k(w_1,w_2,\cdots,w_k)$;
④if $w_1<\min\_w$ then 剪枝 C_k
else if 存在 C_k的某一子集是非频繁的 then 剪枝 C_k
⑤对于余下的 C_k,计算 WIS(C_k);
⑥$L_k=\{C_k\mid \mathrm{WIS}(C_k)>=\mathrm{ms}\}$;
⑦FIS ← FIS∪L_k;}
⑧if ($k>L_{item}$) then Break; }
(4) For 每一个 k_频繁项集 L_k in FIS do
For 任意项集(qt, Nqt) in L_k do
If (WICC(qt, Nqt)≥mincc and (qt∪Nqt = L_k) and (qt∩Nqt = ∅) and (qt⊆Q)) then {
If (WARC (qt → Nqt)≥mc) then
WAR ← WAR∪{qt → Nqt};
If (WARC(Nqt → qt)≥mc) then
WAR ← WAR∪{Nqt → qt};}
(5) Return WAR;

3 基于项权值排序挖掘的跨语言查询扩展

3.1 跨语言查询扩展模型

本文跨语言查询扩展模型分为基于项权值排序的

关联规则前后件混合扩展(Rule Antecedent and Consequent Hybrid Expansion based on Item Weight Sorting (IWS),RACHE_IWS)、规则前件扩展(Rule Antecedent Expansion based on the IWS,RAE_IWS)和规则后件扩展(Rule Consequent Expansion based on the IWS,RCE_IWS)等三种,各扩展模型的形式化定义及其扩展词权值 w_e 计算公式如表 2 所示.

表 2　跨语言查询扩展模型形式化表示

扩展模型	形式化定义
RACHE_IWS	$\{AEt_1, AEt_2, \cdots, AEt_n\} \rightarrow \{qt_1, qt_2, \cdots, qt_m\} \rightarrow \{CEt_1, CEt_2, \cdots, CEt_p\}$ $(WIS \geqslant ms, WARC \geqslant mc, WICC > 0,$ $w_e = \max(WARC) + \max(WICC))$
RAE_IWS	$\{AEt_1, AEt_2, \cdots, AEt_n\} \rightarrow \{qt_1, qt_2, \cdots, qt_m\}$ $(WIS \geqslant ms, WARC \geqslant mc, WICC > 0,$ $w_e = \max(WARC) + \max(WICC))$
RCE_IWS	$\{qt_1, qt_2, \cdots, qt_m\} \rightarrow \{CEt_1, CEt_2, \cdots, CEt_n\}$ $(WIS \geqslant ms, WARC \geqslant mc, WICC > 0,$ $w_e = \max(WARC) + \max(WICC))$

表 2 中,AEt_n 表示第 n 个前件扩展词项,CEt_p 表示第 p 个后件扩展词项,qt_m 表示第 m 个查询词项;w_e 公式中,max(WARC)和 max(WICC)分别表示关联规则置信度和相关系数中的最大值;查询词项权值 W_q 采用文献[15]的计算方法.

3.2　跨语言查询扩展思想及其算法流程图

基于项权值排序挖掘的跨语言查询扩展基本思想是:以印尼语和英语为语言对象,首先将印尼语查询机器翻译为英文并检索英文文档,得到初检前列文档,经过用户相关性判断后获得初检相关反馈文档集,然后,调用 WARM_IWS_CLQE 算法对初检相关反馈文档集挖掘含有原查询词项的加权关联规则模式,最后,按照扩展模型从加权关联规则模式中获取高质量扩展词分别实现规则前后件混合扩展、规则前件扩展和规则后件扩展,扩展词与原查询词组合为新查询再次检索英文文档,并将最后检索结果翻译为印尼语文档返回用户. 根据上述扩展思想,给出图 1 所示的基于项权值排序挖掘的跨语言查询扩展算法流程图.

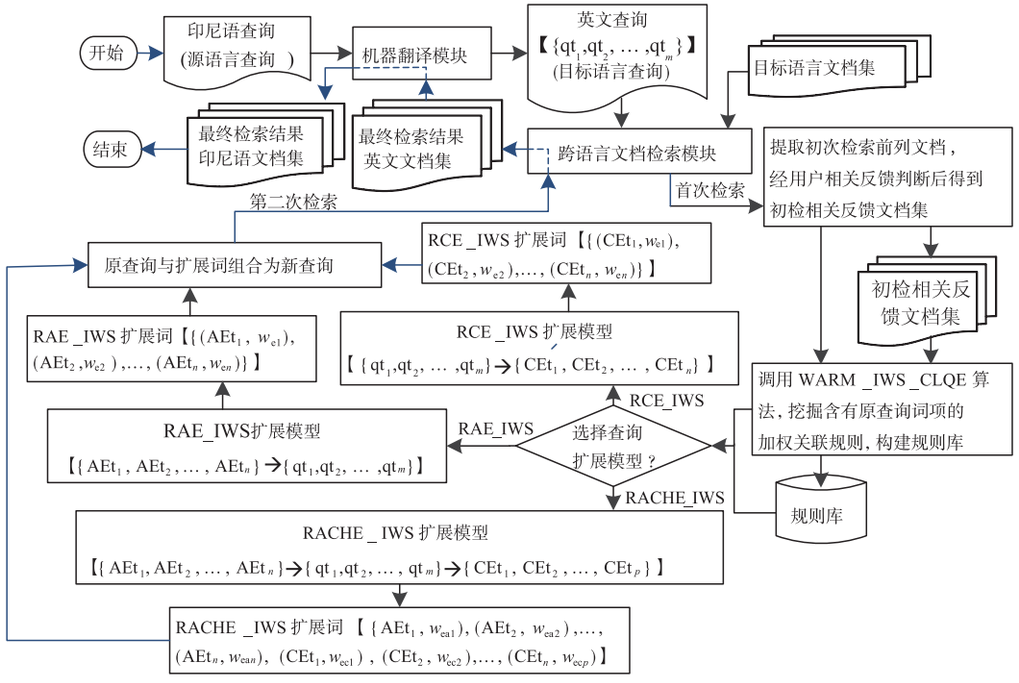

图1　基于项权值排序挖掘的跨语言查询扩展算法流程图

4　实验与结果分析

综上所述,本文基于项权值排序挖掘的跨语言查询扩展分为 RAE_IWS、RCE_IWS 和 RACHE_IWS 等三种扩展算法,为了验证本文扩展算法的检索性能及其有效性,构建基于向量空间检索模型的跨语言信息检索实验平台,在该平台上进行本文检索实验.

4.1　实验数据及其预处理

本文实验数据集是 NTCIR-5 CLIR(详见:http://research.nii.ac.jp/ntcir/data/data-en.html)的英文文本语料,共 26224 篇文档,包括 6608 篇 Mainichi Daily News 新闻媒体 2000 年的新闻文本(简称 md0 数据集),5547 篇 Mainichi Daily News 新闻媒体 2001 年的新闻文本(md1)和 14069 篇 Korea Times 2001 年的新闻文本

(kt1). 该语料有 50 个查询主题,结果集有 Rigid 标准和 Relax 标准两种. 本文实验采用 TITLE(短查询)和 DESC 类型(长查询),通过 Porter 程序(详见:http://tartarus. org/ ~ martin/PorterStemmer)进行预处理,将 NTCIR-5 CLIR 的查询语料人工翻译为印尼语查询语料作为源语言查询. 本文实验评价指标是平均查准率的均值 MAP(Mean Average Precision). 机器翻译系统接口是微软必应机器翻译接口(Microsoft Translator API).

4.2 基准与对比算法

本节从下面 4 个方面对本文扩展方法进行全面考察、实验验证及其实验结果分析:①与单语言检索(Monolingual Retrieval,MR)基准和跨语言检索(Cross-Language Retrieval,CLR)基准进行比较,考察本文扩展算法是否优于基准检索;②与现有基于加权关联模式挖掘的跨语言查询扩展方法[9,10,11,16]对比,考察本文扩展算法是否优于现有同种类型的扩展方法,具体对比算法是:后件扩展对比算法 IECLQE_AWAR[9](mc = 0.1,ms ∈ {0.8,1.0,1.3,1.5,1.7})、PTCE_AWPNP[10](参数同文献的)、IECLQE_WAP[11](mc = 0.01,mi = 0.0001,ms ∈ {0.007,0.008,0.009,0.01,0.011})、IECLCE_AWPNAR[16](mc = 0.5,mi = 0.02,ms ∈ {0.2,0.25,0.3,0.35,0.4}),以及前件扩展对比算法 PTAE_AWPNP[10](参数同文献的)和 IECLAE_AWPNAR[16](mc = 0.5,mi = 0.02,ms ∈ {0.2,0.25,0.3,0.35,0.4});③对本文扩展算法中规则前件扩展、后件扩展和混合扩展的检索性能进行实验性比较,考察其扩展优化的优劣,以便在实际应用中择优选用;④考察本文算法的重要参数及其参数设置对跨语言检索性能的影响,以便在实际应用中方便选用和调整各参数的最优值.

实验时,实验参数值选择原则是:尽量在该参数的有效范围内选择某个比较有效的实验参数值进行实验,带有一定的随机性,例如,$n = 50$,$L_{item} = 3$.

4.3 检索性能比较

4.3.1 本文扩展算法与基准检索、对比算法的检索性能比较

本节比较和分析本文扩展算法与基准检索、对比算法的检索性能. 通过实验,得到检索结果 MAP 的平均值如表 3 至表 5 所示. 实验时,提取印尼-英跨语言初检 n 篇前列英文文档进行用户相关性判断(为了简便,本文实验将初检 n 篇前列文档中含有已知结果集中的相关文档视为用户相关性判断结果文档),构建初检相关文档集. 本节实验参数:RCE_IWS:mc = 0.1,mincc = 0,ms ∈ {0.001,0.002,0.003,0.004,0.005};RAE_IWS 和 RACHE_IWS:ms = 0.001,mincc = 0,mc ∈ {0.1,0.3,0.5,0.7,0.9}.

表 3 本文扩展算法与基准实验的检索结果 MAP 值

查询类型	检索算法	Relax			Rigid		
		md0	md1	kt1	md0	md1	kt1
Title	MR	0.3449	0.5664	0.3288	0.2541	0.4011	0.2363
	CLR	0.3021	0.5152	0.2704	0.2213	0.3582	0.1938
	RAE_IWS	0.7606	0.7162	0.5461	0.5983	0.4673	0.4411
	RCE_IWS	0.7268	0.7517	0.6005	0.5715	0.4944	0.4740
	RACHE_IWS	0.7560	0.6851	0.5692	0.5916	0.4320	0.4543
Desc	MR	0.3021	0.5199	0.2729	0.2357	0.3582	0.2110
	CLR	0.3021	0.4217	0.2612	0.2213	0.3188	0.2150
	RAE_IWS	0.6933	0.6507	0.5281	0.5485	0.5058	0.4281
	RCE_IWS	0.6735	0.7558	0.6276	0.5332	0.5400	0.5086
	RACHE_IWS	0.6933	0.6449	0.5275	0.5485	0.5004	0.4255

表 3 表明:①基准实验中,CLR 的 MAP 值普遍低于 MR 实验的,说明跨语言信息检索出现查询主题漂移问题,使得检索性能下降,而长查询比短查询的 CLR 检索结果下降幅度少些,少数甚至没有下降,说明机器翻译准确率已提高,对长查询检索影响不大;②本文三种扩展算法都取得良好的实验结果,其 MAP 值都比 MR、CLR 的高,性能提升效果显著;③与基准检索比较,RACHE_IWS、RCE_IWS 和 RAE_IWS 算法的 MAP 值最低平均增幅分别为 74.34%、79.14% 和 75.28%(Title).

表 4 本文后件扩展及其对比算法的检索结果 MAP 值

查询类型	检索算法	Relax			Rigid		
		md0	md1	kt1	md0	md1	kt1
Title	IECLQE_AWAR	0.6823	0.6777	0.5183	0.5442	0.4664	0.4101
	IECLQE_WAP	0.6654	0.6466	0.4827	0.5310	0.4301	0.3824
	IECLCE_AWPNAR	0.5233	0.5228	0.3455	0.3883	0.3706	0.2699
	RCE_IWS	0.7268	0.7517	0.6005	0.5715	0.4944	0.4740
Desc	IECLQE_AWAR	0.5934	0.6454	0.5330	0.4757	0.5035	0.4412
	IECLQE_WAP	0.6012	0.6095	0.5349	0.4813	0.4630	0.4613
	IECLCE_AWPNAR	0.4332	0.5267	0.6123	0.3376	0.4198	0.4289
	RCE_IWS	0.6735	0.7558	0.6276	0.5332	0.5400	0.5086

表 4 表明,RCE_IWS 算法的 MAP 值高于后件扩展对比算法的,另外,与后件扩展对比算法(表 4 数据)比较,RCE_IWS、RACHE_IWS 算法的 MAP 值最低平均增幅分别为 9.98%、5.64%(Title)和 13.83%、4.48%(Desc).

表 5 本文前件扩展及其对比算法的检索结果 MAP 值

查询类型	检索算法	Relax			Rigid		
		md0	md1	kt1	md0	md1	kt1
Title	IECLAE_AWPNAR	0.4179	0.5066	0.3409	0.3255	0.3736	0.2669
	RAE_IWS	0.7606	0.7162	0.5461	0.5983	0.4673	0.4411
Desc	IECLAE_AWPNAR	0.3977	0.4995	0.5213	0.3191	0.3990	0.3755
	RAE_IWS	0.6933	0.6507	0.5281	0.5485	0.5058	0.4281

表 5 表明,RAE_IWS 算法的 MAP 值高于其对比算法的,另外,与前件扩展对比算法(表 5 数据)比较,RAE_IWS、RACHE_IWS 算法的 MAP 值增幅分别为 59.62%、58.45%(Title)和 36.43%、35.87%(Desc).

4.3.2 本文加权关联规则混合扩展、后件扩展和前件扩展的检索性能比较

本节进一步对比和分析本文扩展算法的检索性能,根据表3实验结果得到本文三种算法之间的增幅情况,如表6所示,其中,“RCE_IWS vs. RACHE_IWS”表示RCE_IWS算法的检索结果MAP值比RACHE_IWS算法的提高幅度,其余类似.

表6　本文混合扩展、后件扩展和前件扩展的检索结果比较

查询类型	算法描述	Relax			Rigid		
		md0	md1	kt1	md0	md1	kt1
Title	RCE_IWS vs. RACHE_IWS	-3.86%	9.72%	5.50%	-3.40%	14.44%	4.34%
	RAE_IWS vs. RACHE_IWS	0.61%	4.54%	-4.06%	1.13%	8.17%	-2.91%
	RCE_IWS vs. RAE_IWS	-4.44%	4.96%	9.96%	-4.48%	5.80%	7.46%
Desc	RCE_IWS vs. RACHE_IWS	-2.86%	17.20%	18.98%	-2.79%	7.91%	19.53%
	RAE_IWS vs. RACHE_IWS	0.00%	0.90%	0.11%	0.00%	1.08%	0.61%
	RCE_IWS vs. RAE_IWS	-2.86%	16.15%	18.84%	-2.79%	6.76%	18.80%

表6表明,RCE_IWS算法在2个数据集上的MAP值比RACHE_IWS算法的高,其中,长查询检索的增幅高些,最高可达19.53%;RAE_IWS也有类似情况,比RACHE_IWS的增幅最高为8.17%,增幅效果不如RCE_IWS的;RCE_IWS的MAP值高于RAE_IWS的,增幅最大可达18.84%.由此可见,后件扩展的检索性能优于前件扩展的,混合扩展的检索性能不如后件扩展和前件扩展的.

4.3.3 本文算法与对比算法[10]的检索性能比较

由于文献实验的源语言查询是越南语,而本文的是印尼语查询,所以本节比较本文扩展算法与对比算法(PTCE_AWPNP、PTAE_AWPNP)[10]的检索结果MAP值比各自对应对比算法的提高幅度情况,根据文献的原始实验数据以及表4、表5的实验结果,得到其具体增幅,如表7所示,其中,“PTCE_AWPNP vs. VECLQE_WAR”表示PTCE_AWPNP的MAP值比VECLQE_WAR的提高幅度,其余类似.

表7表明,与对比算法比较,RCE_IWS算法的提高幅度高于PTCE_AWPNP算法的,短查询检索的最高提高幅度可达75.62%,而其长查询检索也有类似情况;RAE_IWS算法比PTAE_AWPNP算法的提高幅度稍微有一定优势,即在2个数据集上短查询(Title)检索获得比较高的提高幅度,而其长查询检索MAP的提高幅度不如PTAE_AWPNP算法的,即在2个数据集上的MAP的提高幅度低于PTAE_AWPNP算法的.上述这些情况说明本文算法的检索性能还存在一定的不稳定性,需要进一步深入研究.

表7　本文算法与对比算法[10]的检索性能对比

查询类型	算法描述	Relax			Rigid		
		md0	md1	kt1	md0	md1	kt1
Title	PTCE_AWPNP vs. VECLQE_WAR[10]	6.73%	9.50%	1.22%	7.59%	19.04%	-3.04%
	RCE_IWS vs. IECLQE_WAP	9.23%	16.25%	24.40%	7.63%	14.95%	23.95%
	PTCE_AWPNP vs. CE_AWPNAR[10]	28.77%	30.83%	50.05%	29.13%	27.08%	46.79%
	RCE_IWS vs. IECLCE_AWPNAR	38.89%	43.78%	73.81%	47.18%	33.41%	75.62%
	PTAE_AWPNP vs. AE_AWPNAR[10]	68.05%	46.08%	58.48%	72.75%	37.83%	75.37%
	RAE_IWS vs. IECLAE_AWPNAR	82.01%	41.37%	60.19%	83.81%	25.08%	65.27%
Desc	PTCE_AWPNP vs. VECLQE_WAR[10]	8.76%	4.80%	0.52%	10.09%	3.77%	1.11%
	RCE_IWS vs. IECLQE_WAP	12.03%	24.00%	17.33%	10.78%	16.63%	10.25%
	PTCE_AWPNP vs. CE_AWPNAR[10]	22.53%	52.15%	39.62%	23.75%	50.90%	91.56%
	RCE_IWS vs. IECLCE_AWPNAR	55.47%	43.50%	2.50%	57.94%	28.63%	18.58%
	PTAE_AWPNP vs. AE_AWPNAR[10]	33.33%	68.52%	39.31%	30.04%	56.26%	42.77%
	RAE_IWSvs. IECLAE_AWPNAR	74.33%	30.27%	1.30%	71.89%	26.77%	14.01%

4.3.4 本文算法各参数的扩展性能分析

本节分析本文算法参数ms、mc和mincc对查询扩展性能的影响.本文三种扩展算法参数ms、mc和mincc分别变化时的检索结果MAP的平均值如图2和图3所示,图中符号含义:前缀“A”代表RAE_IWS,前缀“C”代表RCE_IWS,前缀“AC”代表RACHE_IWS,后缀“s”代表ms,后缀“c”代表mc,后缀“cc”代表mincc,此外,其他字符“t”表示TITLE,“d”表示是DESC,例如,图例中“Ats”表示RAE_IWS算法TITLE查询在ms变化时的MAP值.图中横坐标表示数据集,其中,后缀“e”代表Relax,后缀“i”代表Rigid,例如,“md0e”表示扩展算法在md0运行后使用Relax标准得到检索结果.

图2和图3表明,RAE_IWS和RACHE_IWS算法在mc变化时获得最优的检索结果,而在ms和mincc变

化时这两种扩展算法的检索结果比较接近,甚至一致;RCE_IWS 算法在 ms 变化时获得比较好的检索结果.

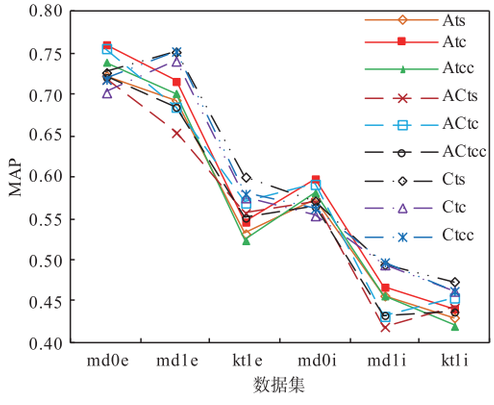

图2 本文扩展算法Title查询的检索结果

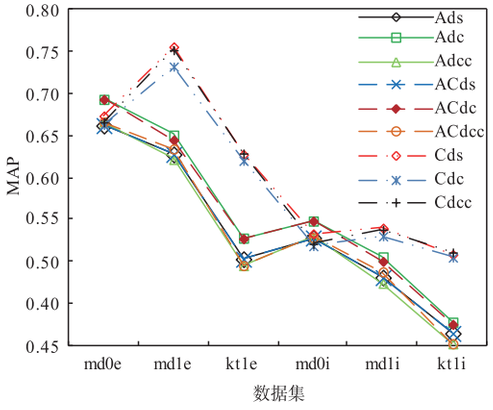

图3 本文扩展算法Desc查询的检索结果

4.3.5 本文算法参数设置的检索性能分析

本节分析和比较参数阈值设置对本文扩展算法检索性能的影响,在每个参数不同的阈值设置下将本文扩展算法在三个数据集的检索结果 MAP 进行平均,得到每个算法的 MAP 平均值,如图 4 至图 6 所示. 图例中字符含义:后缀“e”代表 Relax,后缀“i”代表 Rigid,其余字符含义同图 2 和图 3 的,例如,“Ct_e”就代表 RCE_IWS 算法的 TITLE 查询的检索结果 MAP 值(Relax 标准),其余类似.

图 4 至图 6 表明:①随着 ms、mc 和 mincc 阈值的增大,本文三个扩展算法的 MAP 值呈现逐渐减少的趋势,大部分值减少幅度比较快,少数个别变化比较缓慢,其主要原因是参数阈值的增大导致每个查询获得的扩展词数量减少,扩展性能随之下降;②无论参数阈值如何变化,RCE_IWS 的 MAP 值绝大多数高于 RAE_IWS 和 RACHE_IWS 的,获得最优的检索效果,另外,RAE_IWS 和 RACHE_IWS 的 MAP 值表现比较一致,特别是 ms 变化情况下,其 MAP 值相同;③三个参数变化时,除了 RCE_IWS 的 Rigid 标准的 Title 查询检索结果 MAP 值低于 Desc 查询的外,本文其他算法的 Title 查询检索结果 MAP 值高于 DESC 查询的检索结果,说明本文扩展算法对于短查询的跨语言扩展性能提升程度略高于长查询.

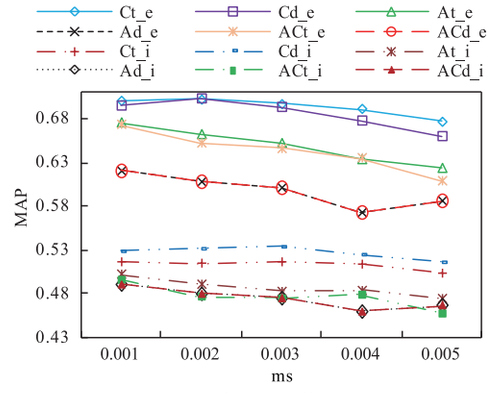

图4 本文算法参数ms阈值设置的检索结果

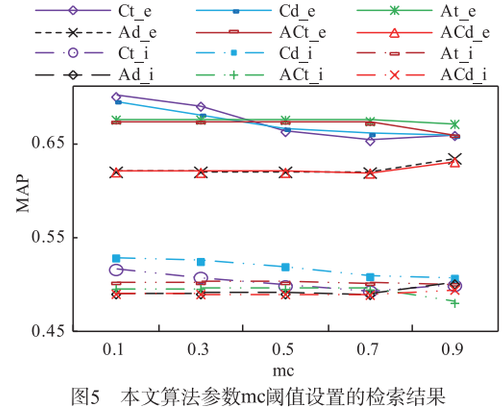

图5 本文算法参数mc阈值设置的检索结果

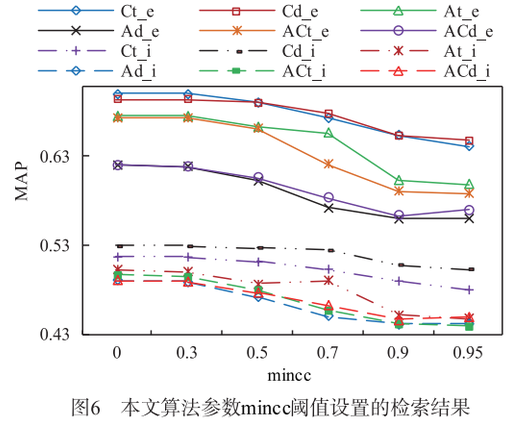

图6 本文算法参数mincc阈值设置的检索结果

4.4 实验结果分析

综上所述,本文提出的基于项权值排序挖掘的跨语言查询扩展是有效的,能减少跨语言检索中查询主题漂移和词不匹配等问题,改善和提升跨语言检索性

能,具体表现为:①与基准检索和现有同类对比算法比较,本文扩展算法都获得最优的检索结果;②各个参数都有利于改善和提高本文扩展算法的跨语言检索性能,支持度参数对后件扩展最有效,置信度更有利于提高前件扩展和混合扩展的检索性能. 对于前件扩展和混合扩展,支持度和相关系数影响的程度比较接近;③本文扩展算法的短查询跨语言检索性能略高于长查询的;④后件扩展获得最好的检索效果,其检索性能优于前件扩展和和混合扩展,这个结果说明本文提出的加权关联规则混合扩展模型还不是最优的,因此,如何改进关联规则混合扩展模型,是本文后续研究之一.

基于项权值排序挖掘的跨语言查询扩展的有效性得益于如下四个方面的改进:一是提出一种新的加权项集支持度,使得频繁项集更为合理;二是提出基于项权值排序的项集剪枝方法,提升挖掘效率和加权关联规则的质量;三是对关联规则评价框架进行改进,即采用置信度-相关系数框架评估关联规则,由此获得更能反映特征词实际相关的关联规则;四是改进跨语言扩展模型,给出新的扩展词权值计算方法,由此得到与原查询相关性更高、更为优质的扩展词,有效地提升扩展词质量. 以上四个方面共同作用,使得本文跨语言查询扩展方法有效地提升信息检索性能,优于基准算法和对比算法.

5　结束语

本文对基于项权值排序挖掘的跨语言查询扩展进行深入研究,提出一种新的加权项集支持度及基于项权值排序的剪枝策略,给出面向查询扩展的基于项权值排序的加权关联规则挖掘方法,深入研究和比较基于项权值排序挖掘的加权关联规则前后件混合扩展、前件扩展和后件扩展模型及其算法. 实验结果表明,本文扩展算法能提升检索性能,能遏制查询主题漂移和词不匹配问题. 本文所提出的查询扩展方法也能适用于其他语言的跨语言信息检索和单语言信息检索,以及关联模式挖掘方法可用于中国-东盟贸易商务数据挖掘和推荐系统. 下一步研究是继续研究如何优化规则混合扩展模型,并探讨将本文算法应用到实际的跨语言搜索引擎中.

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] Ballesteros L, Croft W B. Phrasal translation and query expansion techniques for cross-language information retrieval [A]. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York, NY, USA: ACM, 1997. 84 – 91.

[2] 吴丹,何大庆,王惠临. 基于伪相关反馈的跨语言查询扩展[J]. 情报学报,2010,29(2):232 – 239.
WU Dan, HE Daging, WANG Huilin. Cross-language query expansion using pseudo relevance feedback[J]. Journal of the China Society for Scientific and Technical Information, 2010,29(2):232 – 239. (in Chinese)

[3] Gaillard B, Bouraoui J L, Neef E G D, et al. Query expansion for cross language information retrieval improvement [A]. Proceedings of the Fourth IEEE International Conference on Research Challenges in Information Science [C]. Nice, France: IEEE, 2010. 337 – 342.

[4] 魏露,李书琴等. 跨语言查询扩展优化[J]. 计算机工程与设计,2014,35(8):2785 – 2803.
WEI Lu, LI Shu-qin, et al. Optimization of cross-language query expansion [J]. Computer Engineering and Design, 2014,35(8):2785 – 2803. (in Chinese)

[5] Cao G, Gao J, Nie J Y, et al. Extending query translation to cross-language query expansion with Markov chain models [A]. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management [C]. New York, NY, USA: ACM, 2007. 351 – 360.

[6] Agrawal A, Agrawal D A J. Improving performance of Hindi-English based cross language information retrieval using selective documents technique and query expansion [J]. International Journal of Science and Research, 2016, 5(5): 1964 – 1967.

[7] Tang P, Zhao J, Yu Z, et al. A method of Chinese and Thai cross-lingual query expansion based on comparable corpus [J]. Journal of Information Processing Systems, 2017, 13(4): 805 – 817.

[8] Geraldo A P, Moreira V P. UFRGS@ CLEF2008: using association rules for cross-language information retrieval [A]. Proceedings of the 9th Cross-Language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access [C]. Berlin, Germany: Springer-Verlag, 2009. 66 – 74.

[9] 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)

[10] 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-Eng-

lish 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)

[11] 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)

[12] 黄名选,蒋曹清,何冬蕾. 基于矩阵加权关联规则的跨语言查询译后扩展[J]. 模式识别与人工智能,2018,31(10):887 –898.
HUANG Ming-xuan, JIANG Cao-qing, HE Dong-lei. Cross language query post-translation expansion based on matrix-weighted association rules[J]. Pattern Recognition and Artificial Intelligence,2018,31(10):887 –898. (in Chinese)

[13] 黄名选. 基于矩阵加权关联模式的印尼中跨语言信息检索模型[J]. 数据分析与知识发现,2017(1):26 –36.
HUANG Ming-xuan. Indonesian-Chinese cross language information retrieval model based on matrix-weighted association patterns mining[J]. Data Analysis and Knowledge Discovery,2017,1(1):26 –35. (in Chinese)

[14] 周秀梅,黄名选. 有效的矩阵加权正负关联规则挖掘算法——MWARM-SRCCCI[J]. 计算机应用,2015,34(10):2820 –2826.
ZHOU Xiu-mei, HUANG Ming-xuan. MWARM-SRCCCI:efficient algorithm for mining matrix-weighted positive and negative association rules[J]. Journal of Computer Application,2015,34(10):2820 –2826. (in Chinese)

[15] Salton G,Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management,1988,24(5):513 –523.

[16] 周秀梅,黄名选. 基于项权值变化的完全加权正负关联规则挖掘[J]. 电子学报,2015,43(8):1545 –1554.
ZHOU Xiu-mei, HUANG Ming-xuan. All-weighted positive and negative association rules mining based on dynamic item weight[J]. Acta Electronica Sinica,2015,43(8):1545 –1554. (in Chinese)

作者简介

黄名选 男,1966 年出生于广西乐业县,硕士,现为广西财经学院信息与统计学院教授,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,主持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,被期刊 EI 收录 5 篇,ISTP 收录 1 篇,发明专利授权 15 件.
E-mail:mingxh05@163.com

蒋曹清 男,1973 年出生于湖南省永州市,博士,现为广西财经学院教授,主要研究方向为形式化方法,程序分析,数据挖掘.
E-mail:jcqng@163.com