基于SRCSAC评价框架挖掘的跨语言查询译后扩展

黄名选†, 朱丽娜

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

摘 要: 提出一种面向查询扩展的基于评价框架SRCSAC (support-relevancy-chi-square analysis-confidence)的加权关联规则挖掘算法,给出跨语言查询译后扩展模型和新的扩展词权值计算方法,并提出基于SRCSAC框架挖掘的跨语言查询译后扩展算法. 该算法采用支持度-关联度框架和新的剪枝策略挖掘有效频繁项集,通过卡方分析-置信度框架从有效频繁项集中提取加权关联规则,根据扩展模型从关联规则中获取优质扩展词,实现跨语言译后扩展. 实验结果表明: 所提算法能有效遏制查询主题漂移和词不匹配问题; 与基准检索比较,其前件扩展、后件扩展和混合扩展的MAP最低平均增幅分别为86.85 %、86.04 %和86.00 %; 与对比方法比较,其长查询检索的MAP最低平均增幅分别可达12.23 %、9.06 %和12.6 %,都高于短查询检索的增幅; 与后件扩展算法比较,前件扩展和混合扩展的MAP最高增幅可达5.5 %; 置信度有助于提升前件扩展和混合扩展算法的检索性能,关联度有利于后件扩展算法检索性能的提高,支持度和关联度对后件扩展算法的短查询检索更有效.

关键词: 信息检索；查询扩展；跨语言信息检索；自然语言处理

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2018.1647 **开放科学(资源服务)标识码(OSID):**

引用格式: 黄名选,朱丽娜. 基于SRCSAC评价框架挖掘的跨语言查询译后扩展 [J]. 控制与决策, 2020, 35(11): 2787-2796.

Cross language query post-translation expansion based on the SRCSAC evaluation framework mining

HUANG Ming-xuan†, ZHU Li-na

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

Abstract: An algorithm of weighted association rules mining for query expansion is proposed based on the evaluation framework of support-relevancy-chi-square analysis-confidence (SRCSAC). And the models of cross language query post-translation expansion (CLQPTE) are presented and a new computing method of the expansion term weight is given. Then, an algorithm of CLQPTE is proposed forward based on the SRCSAC framework mining. The algorithm uses the support-relevancy framework and the pruning method to mine effective frequent itemsets, and extracts the weighted association rules from the frequent itemsets in terms of the framework of chi-square-confidence. The high quality expansion terms are obtained from the association rules according to the expansion models in order to carry out CLQPTE. The experimental results show that the proposed algorithms can effectively restrain the issue of query topic drift and term mismatch. Compared with the benchmark retrieval, the MAP minimum average increases (MAIs) of the proposed antecedent expansion (AE), consequent expansion (CE) and hybrid expansion (HE) of the association rules are 86.85 %,86.04 % and 86.00 %, respectively. Compared with the contrast methods, the MAP MAIs of the long queries retrieval for the proposed AE, CE and HE algorithms can reach 12.23 %, 9.06 % and 12.6 %, respectively, which are all higher than those of the short queries retrieval. The MAP maximum increase of the AE and HE can be up to 5.5 % compared with the CE algorithm. The confidence is helpful to improve the retrieval performance of the AE and HE algorithms, and the relevancy is more conducive to the improvement of retrieval performance of the CE. The support and relevancy are more effective for short queries retrieval based on the CE algorithm.

Keywords: information retrieval；query expansion；cross language information retrieval；natural language processing

收稿日期: 2018-12-01；**修回日期:** 2019-05-08.

基金项目: 国家自然科学基金项目 (61762006, 61562004)；广西应用经济学一流学科 (培育) 开放性课题项目 (2018MA07)；广西(东盟)财经研究中心开放性课题项目(2018DMCJYB08).

†通讯作者. E-mail: mingxh05@163.com.

0 引 言

跨语言信息检索是自然语言处理应用研究的一个重要内容. 当前,国际上亟待解决的跨语言信息检索问题是查询主题严重漂移、词不匹配以及查询项翻译歧义和多义性等问题. 跨语言查询扩展是解决上述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等3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的研究工作主要是进行比较性和实验性研究. 文献[1]表明跨语言混合查询扩展效果最好,译前扩展比译后扩展能够更有效地提升检索性能. 文献[2]发现译后扩展性能优于译前扩展性能. 2010年以后,学者们提出一些有效的跨语言译前扩展模型[3],这些模型能改善检索性能. 随着机器翻译准确率不断上升,跨语言查询译后扩展得到发展,其研究工作主要集中在基于伪相关反馈[4-10]和基于关联规则挖掘[11-15]的跨语言查询译后扩展.

基于关联规则挖掘的跨语言查询译后扩展研究可归纳为两种:

1) 挖掘那些与原查询相关的扩展词[11-13]. 其基本思想是:采用关联规则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,实现跨语言译后扩展. 挖掘数据源和挖掘技术是该方法的关键,常见挖掘数据源主要是跨语言初次检索的前列文档集[11-13],即相关反馈文档集. 挖掘技术的关键之一是关联模式评价框架的设计.

2) 挖掘那些与源语言查询词对应的目标语言译后查询词[14-15]. 该工作通过对平行语料挖掘关联模式,得到与源语言查询词相对应的目标语言译后查询词项,使得跨语言检索不需要查询翻译即能完成检索任务,实验结果表明该方法是有效的.

当前,基于关联规则挖掘的跨语言查询扩展研究还不是很深入,存在的问题是:1)原查询词项与其他特征词之间的各种隐含关联的挖掘问题还没有得到完全解决,扩展词质量(即与原查询的相关性)有待于提高. 现有研究中,文献[11]使用的评估框架难免导致冗余的或者虚假的关联模式增多,扩展词的挖掘效率和质量难以保证,文献[12-13]构建的评价框架,无法避免项集中高权值项目与低权值项目相关联的虚假项集模式出现等. 2) 在扩展模型设计方面,虽然关联规则后件扩展和前件扩展模型已得到研究[11-13],但其研究还不深入,忽略了对规则混合扩展模型的深入研究.

鉴于上述问题,本文首先构建一种新的关联模式评价框架,提出有序项集的相关定理和基于有序项集的剪枝方法;然后,提出面向跨语言查询扩展的加权关联规则挖掘算法,深入研究和比较加权关联规则混合扩展、前件扩展和后件扩展模型;最后,提出基于加权关联规则挖掘的跨语言查询译后扩展算法. 实验结果表明了所提算法的有效性.

1 面向跨语言查询扩展的加权关联规则挖掘

1.1 基本概念及相关定理

1.1.1 项集加权支持度

对于跨语言初检相关反馈文档集DS (document set),假设特征词项集为$I, I \subseteq T$,n为DS中总记录数,即总文档篇数,w_I为项集I在DS中的项集权值总和,k_I为项集I的项目个数(即项集长度),则特征词项集I加权支持度(weighted itensets support, WISup)[16]计算如下:

$$\mathrm{WISup}(I) = \frac{w_I}{n \times k_I}. \tag{1}$$

设ms为最小支持度阈值,本文设定$\mathrm{mws} = n \times \mathrm{ms}$为最小权值支持(minimum weight support)阈值,如果$\mathrm{WISup}(I) \geqslant \mathrm{ms}$,即$w_I \geqslant \mathrm{mws} \times k_I$,则称项集$I$为频繁项集. 特别地,如果1_项集的权值不小于mws,则该1_项集是频繁的.

1.1.2 项集关联度

假设$I = (t_1, t_2, \ldots, t_k)$,$I$中各个特征词项目$t_1$,$t_2, \ldots, t_k$单独作为1_项集时为$(t_1), (t_2), \ldots, (t_k)$,$I$中最低的1_项集权值为$w_{\min}$,最高的1_项集权值为$w_{\max}$. 为了避免高权值项目与低权值项目相关联的虚假项集模式出现,提出项集I的关联度(itemSet relrvancy, IRe),计算公式如下:

$$\mathrm{IRe}(I) = \mathrm{IRe}(t_1, t_2, \ldots, t_k) = \frac{w_{\min}}{w_{\max}}. \tag{2}$$

设minIRe为最小项集关联度阈值,$\mathrm{IRe}(I) \geqslant \min \mathrm{IRe}$的特征词频繁项集$I$称为有效频繁项集.

1.1.3 项集卡方分析

假设$I = (I_1, I_2)$,其中$I_1 \bigcup I_2 = I, I_1 \bigcap I_2 = \varnothing$. 借鉴统计学中卡方分析的定义[17],给出项集$I$中$I_1$和$I_2$的卡方(Chi-square, Chis)值计算公式如下所示:

$$\mathrm{Chis}(I_1, I_2) = \frac{(n \times k_1 \times k_2 \times w_I - w_1 \times w_2 \times k_I)^2}{w_1 \times w_2 \times k_1 \times k_2 \times n^2 \times k_I^2}. \tag{3}$$

其中:w_1、w_2分别为项集I_1、I_2在文档集DS中项集权值累加总和,k_1、k_2分别为项集I_1、I_2的长度,n、w_I、k_I定义同式(1).

根据卡方分析的性质,如果$\mathrm{Chis}(I_1, I_2) = 0$,则项集$I_1$与$I_2$相互独立,不存在任何相关性,据此可避免一些虚假相关的关联规则.

1.1.4 加权关联规则置信度

基于传统的置信度定义[18],特征词加权关联规则$(I_1 \rightarrow I_2)$置信度(weighted confidence, WConf)计算如下:

$$\mathrm{WConf}(I_1 \rightarrow I_2) = \frac{\mathrm{WISup}(I_1, I_2)}{\mathrm{WISup}(I_1)} = \frac{w_I \times k_1}{w_1 \times k_I}. \quad (4)$$

其中: $I = I_1 \bigcup I_2, I_1 \bigcap I_2 = \varnothing$, w_I、k_I定义同式(1), w_1、k_1定义同式(3). 设mc为最小置信度阈值,若(I_1, I_2)为有效频繁项集,且$\mathrm{Chis}(I_1, I_2) > 0, \mathrm{WConf}(I_1 \rightarrow I_2) \geqslant \mathrm{mc}$,则$I_1 \rightarrow I_2$是强加权关联规则.

1.1.5 有序项集及其相关定理

将各个项目权值按升序排列后的特征词项集称为有序项集(ordered itemsets, OI), $I_{\mathrm{oi}} = (i_1, i_2, \ldots, i_{k-1}, i_k)$,对应的项目权值集合为$(w_1, w_2, \ldots, w_{k-1}, w_k)$,其中$w_1 \leqslant w_2 \leqslant \ldots \leqslant w_{k-1} \leqslant w_k$. 由此可见, I_{oi}中项目i_k对应项目权值w_k是最高的,称为权值最高项目,简称高权项目. 后续所讨论有序项集I_{oi}的子项集是指按项目权值由低到高从有序项集I_{oi}中抽取项目并组合得到的真子集,即子项集$I_{\mathrm{oi_sub1}} = (i_1), I_{\mathrm{oi_sub2}} = (i_1, i_2), \ldots, I_{\mathrm{oi_sub}(k-1)} = (i_1, i_2, \ldots, i_{k-1})$,这些子项集也是有序项集.

定理1　I_{oi}的有序子项集$I_{\mathrm{oi_sub1}}, I_{\mathrm{oi_sub2}}, \ldots, I_{\mathrm{oi_sub}(k-1)}$的项集权值分别大于或等于$w_1, w_1 + w_2, \ldots, w_1 + w_2 + w_3 + \ldots + w_{k-1}$.

证明过程略.

定理2　对于有序项集$I_{\mathrm{oi}} = (i_1, i_2, \ldots, i_r, i_{r+1}, \ldots, i_{r+k}) = (X_{\mathrm{oi}}, Y_{\mathrm{oi}})$. 其中: $r \neq 0, k \neq 0$, X_{oi}和Y_{oi}是I_{oi}的有序子项集, $X_{\mathrm{oi}} = (i_1, i_2, \ldots, i_r), Y_{\mathrm{oi}} = (i_{r+1}, \ldots, i_{r+k})$. 若有序子项集$Y_{\mathrm{oi}}$是非频繁的,则有序项集$I_{\mathrm{oi}}$一定是非频繁的.

证明　设有序项集$I_{\mathrm{oi}} = (i_1, i_2, \ldots, i_r, i_{r+1}, \ldots, i_{r+k})$对应的项目权值为$(w_1, w_2, \ldots, w_r, w_{r+1}, \ldots, w_{r+k})$,则有

$$w_1 \leqslant w_2 \leqslant \ldots \leqslant w_r \leqslant w_{r+1} \leqslant \ldots \leqslant w_{r+k} \Rightarrow \frac{w_1 + w_2 + \ldots + w_r}{r} \leqslant \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{k}. \quad (5)$$

设Y_{oi}的项集权值为w_Y, Y_{oi}是非频繁的,有

$$\mathrm{WISup}(Y_{\mathrm{oi}}) = \frac{w_Y}{n \times k} < \mathrm{ms}. \quad (6)$$

由定理1

$$w_{r+1} + w_{r+2} + \ldots + w_{r+k} \leqslant w_Y \Rightarrow \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k} \leqslant \frac{w_Y}{n \times k}. \quad (7)$$

由式(6)和(7),可得

$$\frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k} < \mathrm{ms}, \quad (8)$$

$$\mathrm{WISup}(I_{\mathrm{oi}}) - \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k} = \frac{w_1 + w_2 + \ldots + w_r + w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times (r+k)} - \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k} \Rightarrow$$

$$\mathrm{WISup}(I_{\mathrm{oi}}) - \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k} = \frac{1}{n \times (1 + k/r)} \times \left(\frac{w_1 + w_2 + \ldots + w_r}{r} - \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{k} \right). \quad (9)$$

因为$n \times \left(1 + \dfrac{k}{r}\right) > 0$,由式(5)和(9)可得

$$\mathrm{WISup}(I_{\mathrm{oi}}) - \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k} \leqslant 0 \Rightarrow \mathrm{WISup}(I_{\mathrm{oi}}) \leqslant \frac{w_{r+1} + w_{r+2} + \ldots + w_{r+k}}{n \times k}. \quad (10)$$

由式(8)和(10)可得$\mathrm{WISup}(I_{\mathrm{oi}}) < \mathrm{ms}$,即有序项集$I_{\mathrm{oi}}$是非频繁的. □

定理2表明,对于有序项集$I_{\mathrm{oi}} = (i_1, i_2, \ldots, i_k)$,如果$I_{\mathrm{oi}}$的高权项目$i_k$对应的1_项集$(i_k)$是非频繁的,则有序项集$I_{\mathrm{oi}}$是非频繁的.

定理3　有序项集$I_{\mathrm{oi}} = (i_1, i_2, \ldots, i_k)$对应的项目权值为$(w_1, w_2, \ldots, w_k)$,若$w_k < \mathrm{mws}$,则有序项集$I_{\mathrm{oi}}$一定是非频繁项集;若$w_1 \geqslant \mathrm{mws}$,则有序项集$I_{\mathrm{oi}}$一定是频繁项集.

证明过程略.

1.2 SRCSAC评价框架及候选项集剪枝

本文提出的SRCSAC评价框架是: 支持度-关联度-卡方分析-置信度(support-relevancy-Chi_square analysis-confidence, SRCSAC)评价框架.

本文使用SRCSAC评价框架评估含有原查询词项的关联模式,即采用支持度-关联度评价框架评估特征词频繁项集模式,使用卡方分析-置信度评价框架衡量关联规则前件与后件的关联程度,避免高权值项目与低权值项目相关联的虚假项集模式出现,减少无趣和无效的关联模式产生,获取优质扩展词.

本文提出一种基于有序项集的剪枝方法,该方法分为候选2_项集的剪枝和候选k_项集$(k > 2)$的剪枝,前者主要剪除不含原查询词项的候选2_项集,后者的剪枝方法是: 构建候选k_有序项集$(i_1, i_2, \ldots, i_k)$,根据定理1和定理2,如果存在如下两种情况之一,则可以剪除该候选k_项集: 1) 候选k_有序项集$(i_1, i_2, \ldots, i_k)$高权项目i_k对应的1_项集(i_k)是非频

繁的;2) 候选$k_$有序项集$(i_1, i_2, \ldots, i_k)$高权项目i_k对应的项目权值$w_k < \text{mws}$.

1.3 基于SRCSAC评价框架的加权关联规则挖掘算法

挖掘算法基本思想是:采用支持度-关联度框架和项集剪枝策略挖掘含有译后原查询词项的有效频繁项集,采用卡方分析-置信度框架从特征词频繁项集中提取强加权关联规则模式,这些关联规则的前件或后件含有原查询词项.

上述挖掘思想形式化为算法WARM_SRCSAC_CLQE(weighted association rules mining based on SRCSAC for cross language query expansion). 算法符号含义如下:Qt为用户查询词项集,NQt为不含查询词项的特征词项集,ILen为候选项集长度阈值,WAR为强加权关联规则集合,FIS (frequent itemset)为特征词频繁项集集合,L_k为含有原查询词项的有效$k_$频繁项集,$w(C_k)$代表候选$k_$项集C_k的项集权值.

算法1 WARM_SRCSAC_CLQE.

input: DS, Qt, ILen, ms, mc, minIRe;

output: WAR.

begin

1) (DS_DB, DS_Terms, mws) ←TextPreprocessing (DS, ms);

2) mining L_1(DS_DB, DS_Terms, mws); {

2.1) C_1 ←ScanDSTerms (DS_Terms);

2.2) $w(C_1)$ ←ScanDSDB(DS_DB);

2.3) $L_1 \leftarrow \{C_1 | w(C_1) >= \text{mws}\}$;

2.4) FIS ← FIS$\bigcup L_1$;}

3) mining L_2(DS_DB, FIS, mws, Qt, minIRe); {

3.1) $L_1 \leftarrow \text{ExtractL1(FIS)}$;

3.2) $C_2 \leftarrow L_1 \infty L_1$;

3.3) $C_2 \leftarrow \text{PruningNotQ}(C_2)$;

3.4) $w(C_2)$ ← ScanDSDB(DS_DB);

3.5) $L_2 \leftarrow \{C_2 | w(C_2) \geqslant \text{mws} \times 2 \text{ and } \text{IRe}(C_2) \geqslant \min \text{IRe}\}$;

3.6) FIS ← FIS$\bigcup L_2$;}

4) for$(k = 3; L_k \neq \varnothing; k++)${

4.1) $C_k \leftarrow L_{k-1} \infty L_{k-1}$;

4.2) miningWAR_SRCSAC (DS_DB, mws, minIRe); {

① $C_k(w_1, w_2, \ldots, w_k)$ ← ScanDSDB(DS_DB);

② if (Exist1_Unfrequent(C_k) or $w_k < \text{mws}$) then $C_k \leftarrow \text{Pruning}C_k(C_k)$;

③ $(w(C_k), \text{IRe}(C_k))$ ← ScanDSDB(DS_DB);

④ $L_k \leftarrow \{C_k | w(C_k) \geqslant \text{mws} \times k \text{ and } \text{IRe}(C_k) \geqslant \min \text{IRe}\}$;

⑤ FIS ← FIS$\bigcup L_k$;}

4.3) if$(k > \text{ILen})$ then Break;}

5) for each effective frequent itemset L_k in FIS do

for each itemset (qt, NQt) in L_k do

if (Chis(qt, NQt) > 0 and (qt$\bigcup$NQt $= L_k$) and

(qt$\bigcap$NQt $= \varnothing$) and (qt $\subseteq$ Qt)) then {

if (WConf (qt → NQt) >= mc) then

WAR←WAR$\bigcup$ {qt→NQt;

if (WConf (NQt→qt) >= mc) then

WA←WAR$\bigcup${NQt → qt};}

6) return WAR;

end.

算法1中,步骤1) TextPreprocessing( ) 对DS预处理;步骤2) mining L_1( ) 挖掘1_频繁项集;步骤3) mining L_2( ) 挖掘含有原查询词项的有效2_频繁项集;步骤4) 构建候选$k_$有序项集$(w_1, w_2, \ldots, w_k)(k \geqslant 3)$,然后剪枝,挖掘出含有原查询词项的有效$k_$频繁项集$L_k$;步骤5) 挖掘含有原查询词项的特征词强加权关联规则模式.

2 跨语言查询译后扩展

2.1 跨语言查询译后扩展模型

本文将跨语言查询译后扩展模型分为基于SRCSAC框架的关联规则前件扩展(association rule antecedent expansion based on SRCSAC, ARAE_SRCSAC)、后件扩展(association rule consequent expansion, ARCE_SRCSAC)和规则前后件混合扩展(association rule antecedent and consequent hybrid expansion, ARACHE_SRCSAC)三种模型,如下所示:

$$\{\text{AEt}_1, \text{AEt}_2, \ldots, \text{AEt}_n\} \rightarrow \{\text{Qt}\} \Rightarrow (\text{AEt}_1, w_{e1}), (\text{AEt}_2, w_{e2}), \ldots, (\text{AEt}_n, w_{en}); \quad (11)$$

$$\{\text{Qt}\} \rightarrow \{\text{CEt}_1, \text{CEt}_2, \ldots, \text{CEt}_n\} \Rightarrow (\text{CEt}_1, w_{e1}), (\text{CEt}_2, w_{e2}), \ldots, (\text{CEt}_n, w_{en}); \quad (12)$$

$$\left.\begin{array}{l}\{\text{AEt}_1, \text{AEt}_2, \ldots, \text{AEt}_n\} \rightarrow \{\text{Qt}\} \\ \{\text{Qt}\} \rightarrow \{\text{CEt}_1, \text{CEt}_2, \ldots, \text{CEt}_p\}\end{array}\right\} \Rightarrow \left\{\begin{array}{l}(\text{AEt}_1, w_{e1}), (\text{AEt}_2, w_{e2}), \ldots, (\text{AEt}_n, w_{en}), \\ (\text{CEt}_1, w_{e1}), (\text{CEt}_2, w_{e2}), \ldots, (\text{CEt}_n, w_{ep}).\end{array}\right. \quad (13)$$

其中:Qt为原查询项集,AEt_i为第i $(1 \leqslant i \leqslant n)$个前件扩展词项,$\text{CEt}_j$为第$j$ $(1 \leqslant j \leqslant p)$个后件扩展词

项, w_e 为扩展词权值.

2.2　扩展词权值计算

将关联度、卡方值和置信度等参数值作为扩展词权值的计算依据,得到

$$w_e = 0.5\max(\mathrm{WConf}(\cdot)) + 0.3\max(\mathrm{Chis}(\cdot)) + 0.2\max(\mathrm{IRe}(\cdot)). \tag{14}$$

其中: $\max(\mathrm{WConf}(\cdot))$、$\max(\mathrm{Chis}(\cdot))$ 和 $\max(\mathrm{IRe}(\cdot))$ 分别为置信度、卡方值和关联度的最大值. 另外, 公式中系数表示各个参数对扩展词的贡献程度, 是本文经过分析后得出的一个估计值.

2.3　初检相关文档特征词权值计算

给出跨语言首次检索得到的相关反馈文档特征词权值计算公式如下所示:

$$w_{ij} = \frac{\mathrm{tf}_{j,i} + 1}{2 \times (\mathrm{idf}_j + 1)}. \tag{15}$$

其中: w_{ij} 为跨语言相关反馈文档 d_i 中特征词 t_j 的权值, idf_j 为逆文档频度 (inverse document frequency), $\mathrm{tf}_{j,i}$ 为特征词 t_j 在文档 d_i 中的词频, 需要进行标准化处理.

2.4　跨语言查询译后扩展算法

本文提出的跨语言查询译后扩展基本思想是: 首先将源语言查询词项通过机器翻译为目标语言, 并检索目标语言文档, 调用 WARM_SRCSAC_CLQE 挖掘算法对相关反馈文档集挖掘含有目标语言查询词项的有效频繁项集和强加权关联规则, 根据扩展模型从强加权关联规则集合中提取译后目标语言扩展词实现译后扩展, 扩展词与原查询词组合为新查询再次检索目标语言文档.

上述查询扩展思想形式化为 CLQPTE_SRCSAC (cross language query post-translation expansion based on SRCSAC) 算法. 算法中符号含义如下: Q_{SL} 为源语言用户查询, Q_{TL} 为目标语言查询, n 为跨语言初检前列文档数, ET 为译后扩展词集合, NewQ_{TL} 为扩展后的目标语言新查询, 其余的同算法 1.

算法 2　CLQPTE_ SRCSAC.

input : $Q_{\mathrm{SL}}, n, \mathrm{ILen}, \mathrm{ms}, \mathrm{mc}, \min \mathrm{IRe}$;

output: NewQ_{TL}, 最终检索结果源语言文档.

begin

1) $Q_{\mathrm{TL}} \leftarrow \{Q_{\mathrm{SL}}$ 机器翻译为 $Q_{\mathrm{TL}}\}$;

2) DFirstR $\leftarrow \{Q_{\mathrm{TL}}$ 检索目标语言文档集, 提取前列 n 篇初检文档};

3) DUserJ ←{将 DFirstR 文档经用户相关反馈后得到相关反馈文档 DUserJ};

4) WAR ← WARM_SRCSAC_CLQE(DUserJ, $Q_{\mathrm{TL}}, \mathrm{ms}, \mathrm{mc}, \min \mathrm{IRe}, \mathrm{ILen}$);

5) Switch (扩展模型){

Case ARAE_SRCSAC: ET ←{从 WAR 提取形如 $\mathrm{NQt} \rightarrow \mathrm{qt}$ 的关联规则前件 NQt 作为扩展词};

Case ARCE_SRCSAC: ET←{从 WAR 提取形如 $\mathrm{qt} \rightarrow \mathrm{NQt}$ 的关联规则后件 NQt 作为扩展词};

Case ARACHE_SRCSAC: ET←{从 WAR 提取形如 $\mathrm{qt} \rightarrow \mathrm{NQt}_1$ 和 $\mathrm{NQt}_2 \rightarrow \mathrm{qt}$ 的关联规则, 提取规则后件项集 NQt_1 和前件项集 NQt_2 作为扩展词};}

6) 计算扩展词 ET 的权值;

7) New$Q_{\mathrm{TL}} = Q_{\mathrm{TL}} \bigcup \mathrm{ET}$;

8) 最终检索结果目标语言文档←{NewQ_{TL} 检索目标语言文档集}

9) 最终检索结果源语言文档←{最终检索结果目标语言文档机器翻译为源语言文档};

10) retrun NewQ_{TL} 和最终检索结果源语言文档;

end.

3　实验与分析

3.1　实验数据及其预处理

实验数据采用 NTCIR-5 CLIR 的英文文本语料, 包括 6 608 篇 Mainichi Daily News 2000 年的新闻文本 (简称 m_0 数据集), 5 547 篇 Mainichi Daily 2001 年的新闻文本 (简称 m_1), 14 069 篇 Korea Times 2001 年的新闻文本 (简称 k_1). 该语料有 50 个查询主题, 本文采用 title 和 desc 查询主题进行检索实验. title 查询以名词和名词性短语简要描述查询主题, 属于短查询, 而 desc 查询以句子形式简要描述查询主题查询, 属于长查询. 语料结果集有 rigid 标准 (与查询高度相关, 相关) 和 relax 标准 (与查询高度相关、相关和部分相关) 两种标准.

采用 Porter 程序对实验数据进行词干提取, 实验的源语言印尼语查询由翻译机构专业人员对 NTCIR-5 CLIR 语料的 50 个中文版查询主题语料人工翻译而得到, 实验所用的机器翻译工具接口是微软必应机器翻译接口 (microsoft translator API), 检索评价指标是平均查准率的均值 MAP (mean average precision).

3.2　实验设计及其对比方法

实验设计总体思想: 构建基于向量空间检索模型的跨语言信息检索实验平台; 在相同的实验环境下, 以印尼语为源语言、以英语为目标语言进行本文实验, 验证本文扩展算法的检索性能及其有效性. 考察内容如下: 1) 与单语言检索 (monolingual retrieval, MLR) 基准和跨语言检索 (cross-language retrieval,

CLR)基准进行比较,考察本文扩展算法的检索结果评价指标值是否高于基准检索的评价指标值;2)与经典的基于伪相关反馈的跨语言查询扩展方法(PTE_PRF)[2]对比,考察本文扩展算法的检索性能是否优于现有不同扩展类型的方法;3)与现有基于加权关联模式挖掘的跨语言查询扩展方法[11-12,16,19]对比,即PTE_AWAP[11](ms $\in \{0.8, 1.0, 1.3, 1.5, 1.7\}$, mc $= 0.1$), PTE_WAP[12](ms $\in \{0.007, 0.008, 0.009, 0.01, 0.011\}$, mc $= 0.01$, mi $= 0.000\,1$), PTE_AWPNAR[16] (mc $= 0.5$, mi $= 0.02$, ms $\in \{0.2, 0.25, 0.3, 0.35, 0.4\}$), PTE_WMMSM[19](ms $\in \{0.9, 1.1, 1.3, 1.5, 1.7\}$, mc $= 0.1$,最低支持度下界LMS $= 0.1$,最低支持度上界HMS $= 0.15$,最低权值阈值WT $= 0.1$),考察本文扩展算法的检索性能是否优于现有同种类型的扩展方法;4)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;5)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;6)最后进行查询实例的检索效果分析,进一步表明本文扩展算法是否能有效地遏制跨语言查询主题漂移和词不匹配问题.

扩展算法CLQPTE_SRCSAC的实验分为前件扩展(ARAE_SRCSAC)、后件扩展(ARCE_SRCSAC)和混合扩展(ARACHE_SRCSAC)三种算法进行. 实验参数有两大类:一类是实验环境参数,即n和ILen,这类参数在各实验算法取值相同,本文实验环境设置为$n = 50$, ILen $= 3$;另一类参数是算法参数,即各算法特有的参数,如ms、mc、mi等,这类参数在各实验算法的实验取值难以完全一致,主要原因是,各个实验算法的挖掘方法和关联模式的评价框架不同,以及各个算法存在的参数及其参数计算公式也不完全相同,导致算法参数有效的取值范围不完全一致. 因此,本文算法实验参数取值原则是在各自参数的有效范围内取值,通过反复实验,将获得较好检索结果的参数值作为本文实验参数值,带有一定的随机性. 如何确定最优的算法参数取值后续会进一步研究.

3.3 检索性能比较

3.3.1 本文算法与基准、对比方法的检索性能对比

通过实验平台在数据集m_0, m_1和k_1上进行实验,得到检索结果MAP的平均值如表1和表2所示. 为了简便,实验过程中将初检前列n篇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. 表1和表2的实验参数如下. ARCE_SRCSAC: ms $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, mc $= 0.1$, min IRe $= 0.4$ ARAE_SRCSAC和ARACHE_SRCSAC: ms $= 0.5$, mc $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, min IRe $= 0.4$.

表1 title查询的检索结果MAP

检索算法	relax			rigid		
	m_0	m_1	k_1	m_0	m_1	k_1
CLR	0.302 1	0.515 2	0.270 4	0.221 3	0.358 2	0.193 8
MLR	0.344 9	0.566 4	0.328 8	0.254 1	0.401 1	0.230 5
PTE_PRF	0.238 2	0.336 8	0.189 9	0.162 3	0.224 3	0.137 2
PTE_WAP	0.665 4	0.646 6	0.482 7	0.531 0	0.430 1	0.382 4
PTE_AWAP	0.682 3	0.677 7	0.518 3	0.544 2	0.466 4	0.410 1
PTE_WMMSM	0.696 4	0.752 0	0.558 0	0.531 0	0.488 9	0.450 3
PTE_AWPNAR	0.523 3	0.522 8	0.345 5	0.388 3	0.370 6	0.269 9
ARAE_SRCSAC	0.749 6	0.767 3	0.635 5	0.579 5	0.504 4	0.509 9
ARCE_SRCSAC	0.769 4	0.751 9	0.619 1	0.595 1	0.495 0	0.494 5
ARACHE_SRCSAC	0.750 1	0.771 0	0.634 1	0.568 9	0.500 6	0.509 1

表2 desc查询的检索结果MAP

检索算法	relax			rigid		
	m_0	m_1	k_1	m_0	m_1	k_1
CLR	0.302 1	0.421 7	0.261 2	0.221 3	0.318 8	0.215 0
MLR	0.292 7	0.522 5	0.254 8	0.230 1	0.364 9	0.195 1
PTE_PRF	0.225 3	0.357 5	0.149 4	0.157 5	0.258 6	0.118 8
PTE_WAP	0.601 2	0.609 5	0.534 9	0.481 3	0.463 0	0.461 3
PTE_AWAP	0.593 4	0.645 4	0.533 0	0.475 7	0.503 5	0.441 2
PTE_WMMSM	0.612 7	0.732 3	0.550 5	0.477 5	0.511 9	0.448 8
PTE_AWPNAR	0.433 2	0.526 7	0.612 3	0.337 6	0.419 8	0.428 9
ARAE_SRCSAC	0.715 4	0.709 3	0.676 3	0.554 6	0.499 6	0.552 7
ARCE_SRCSAC	0.696 9	0.696 5	0.651 9	0.545 2	0.492 5	0.523 8
ARACHE_SRCSAC	0.714 6	0.715 2	0.678 3	0.555 0	0.505 4	0.552 6

表1和表2的实验结果表明,与基准检索和对比算法比较,规则前件、后件和混合扩展算法在3个数据集上的MAP值都比基准检索和对比算法高,检索性能提升效果显著. 具体表现为:

1) 与基准检索相比,本文3种扩展算法的检索结果MAP值最低平均增幅分别为86.85 % (title,平均增幅 (%) = [(0.749 6 − 0.344 9)/0.344 9 + (0.767 3 − 0.566 4)/0.566 4 + (0.635 5 − 0.328 8)/0.328 8 + (0.579 5 − 0.254 1)/0.254 1 + (0.504 4 − 0.401 1)/0.401 1+(0.509 9−0.230 5)/0.230 5]×100/6 = 86.85, 后面类似)、86.04 % (title)和86.00 % (title).

2) 与对比方法相比,本文3种扩展算法的MAP值最低平均增幅分别为8.18 %、7.43 %、7.74 %(title)和12.23 %、 9.06 %、 12.60 %(desc). 由此可见,本文规则前件、后件和混合扩展算法的长查询检索结果MAP最低平均增幅高于短查询检索.

表1和表2表明,MLR的MAP值绝大多数均高于CLR,表明跨语言检索结果受查询翻译等因素影响较大,检索性能不如单语言检索. 对比方法中,PTE_WMMSM、PTE_WAP和PTE_AWAP方法获得了较好的检索结果,其MAP值均高于基准检索. 其中PTE_WMMSM的检索效果最好,PTE_PRF的实验结果并不理想,MAP值低于基准检索,表明直接从跨语言初检文档提取扩展词,导致扩展词噪音比较多,易产生查询主题漂移,检索性能反而降低.

3.3.2　规则前件、后件和混合扩展的检索性能对比

为了比较规则前件、后件和混合扩展的检索性能,由表1和表2的实验结果,将ARACHE_SRCSAC检索结果与ARCE_SRCSAC、ARAE_SRCSAC进行对比,ARAE_SRCSAC检索结果与ARCE_SRCSAC进行比较,其MAP增幅 (%) 如表3所示. 表3中,“ARACHE vs. ARCE”表示ARACHE_SRCSAC算法MAP值较ARCE_SRCSAC算法的增幅,其余类似.

表 3　混合扩展、后件扩展和前件扩展检索结果比较

查询类型	算法描述	relax			rigid		
		m_0	m_1	k_1	m_0	m_1	k_1
title	ARACHE vs. ARCE	−2.51	2.54	2.42	−4.40	1.13	2.95
	ARAE vs. ARCE	−2.57	2.05	2.65	−2.62	1.90	3.11
	ARAE vs. ARACHE	−0.07	−0.48	0.22	1.86	0.76	0.16
desc	ARACHE vs. ARCE	2.54	2.68	4.05	1.80	2.62	5.50
	ARAE vs. ARCE	2.65	1.84	3.74	1.72	1.44	5.52
	ARACHE vs. ARAE	−0.11	0.83	0.30	0.07	1.16	−0.02

表3表明,对于短查询 (title) 和长查询 (desc) 检索,ARAE_SRCSAC和ARACHE_SRCSAC检索结果的MAP值绝大部分比ARCE_SRCSAC高. 与ARCE_SRCSAC比较,ARAE_SRCSAC和ARACHE_SRCSAC的MAP值最大增幅分别为5.5 % 和5.52 %; ARAE_SRCSAC的MAP值与ARACHE_SRCSAC相差不明显. 对于短查询检索,ARAE_SRCSAC的MAP值大部分比ARACHE_SRCSAC略高,对于长查询检索,这一情况刚好相反.

3.3.3　算法参数对检索性能的影响

图1和图2给出了扩展算法在各个参数阈值状态下的检索结果MAP平均值. 图中: 前缀“A”表示ARAE_ SRCSAC,前缀“C”表示ARCE_SRCSAC,前缀“H”表示ARACHE_SRCSAC,后缀“s”表示ms,后缀“c”表示mc,后缀“ir”表示minIRe,“t”表示title查询,“d”表示desc查询,横坐标后缀“e”表示relax标准,后缀“i”表示rigid标准. 实验参数如下: $\mathrm{ms} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}, \mathrm{mc} = 0.1, \mathrm{minIRe} = 0.4; \mathrm{mc} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}, \mathrm{ms} = 0.5, \mathrm{minIRe} = 0.4; \mathrm{minIRe} \in \{0.4, 0.45, 0.5, 0.55, 0.6\}, \mathrm{ms} = 0.5, \mathrm{mc} = 0.1$.

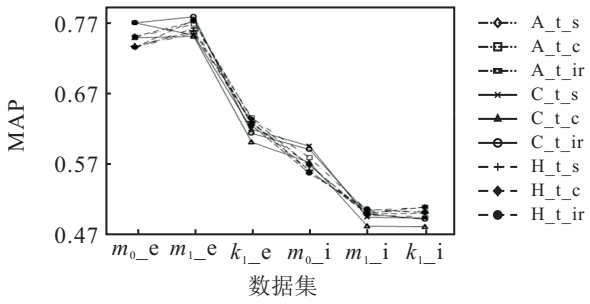

图 1　参数对本文扩展算法检索性能的影响(title查询)

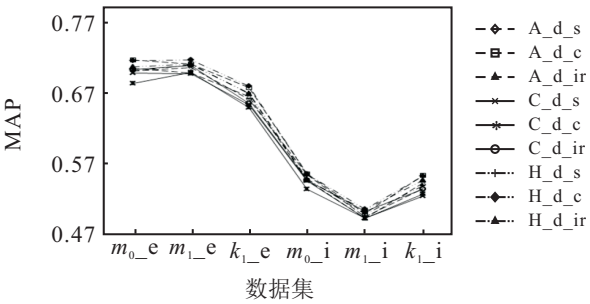

图 2　参数对本文扩展算法检索性能的影响(desc查询)

图1和图2表明,mc参数对前件扩展和混合扩展的检索性能影响最大,有助于提升前件扩展和混合扩展的检索性能,后件扩展在minIRe参数阈值状态下能够获得更好的检索结果. 在ms和minIRe参数阈值状态下,规则前件扩展与混合扩展的检索结果比较接近,甚至相等.

3.3.4 算法参数设置对检索性能的影响

本节分析和比较算法参数的不同阈值设置对跨语言检索性能的影响. 图3～图5给出了各个参数不同阈值设置下,3种扩展算法在3个数据集上检索得到的各自MAP平均值. 图例中字符含义同图1.

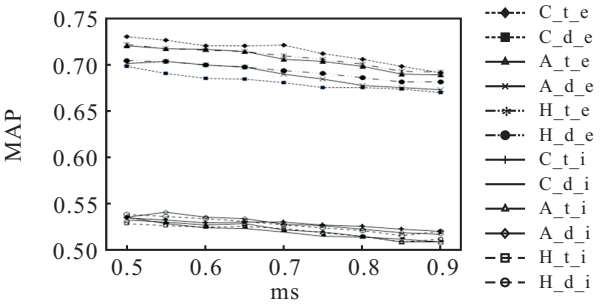

图 3 ms对本文扩展算法检索性能的影响

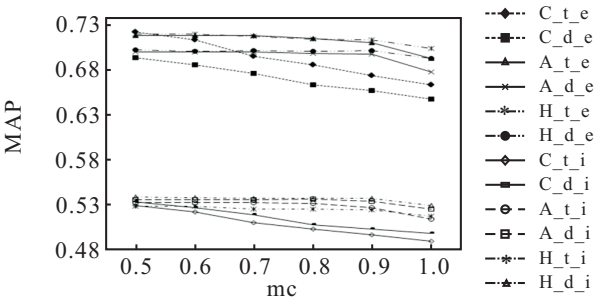

图 4 mc对本文扩展算法检索性能的影响

图3～图5表明,随着参数阈值取值的增大,本文算法的检索结果MAP值呈下降趋势,有的下降较快. 其主要原因是参数阈值增大,挖掘出的每个查询扩展词数量减少,扩展性能随之降低. 另外,本文扩展算法对长查询和短查询均有效,而ARCE_SRCSAC算法在ms和minIRe参数不同阈值设置时,title查询的检索结果比desc效果好,表明支持度和关联度参数对基于规则后件扩展的短查询检索更有效.

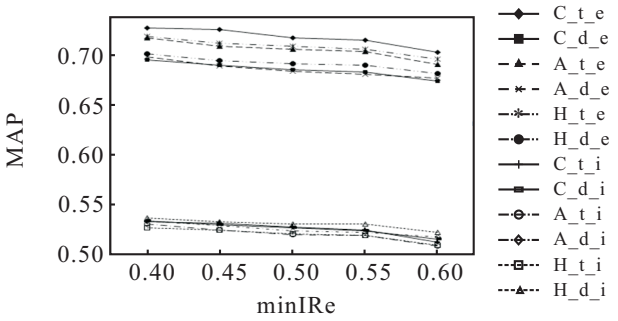

图 5 minIRe对本文扩展算法检索性能的影响

3.3.5 查询实例检索效果分析

本节列举实验语料中No.16和No.34查询desc主题印尼文和英文实例及其在m_1数据集上跨语言检索结果. 表4给出了查询主题实例原文及其前件扩展词词干实例,扩展词后面括号内的数值为该扩展词的权值;表5给出了查询实例的检索结果P@5、P@10和MAP值$(\mathrm{ms}=0.5, \mathrm{mc}=0.8, \min \mathrm{IRe}=0.4)$.

由表5可见,对于No.34查询主题,其基准检索CLR的P@5、P@10和MAP值都较MLR要低,表明跨语言检索过程中发生了查询主题漂移和词不匹配现象,导致检索性能下降. 本文扩展算法运行后得到如表4所示的扩展词,执行跨语言查询译后扩展后得到的P@5、P@10和MAP值都比CLR高，甚至接近和高于基准MLR,对于No.16查询也有类似趋势. 由此可见,本文扩展算法确实能有效地遏制查询主题漂移和词不匹配问题.

表 4 查询实例原文及其扩展词

查询版本	查询编号	实例描述
印尼语版	No.16	Bagaian kebudayaan dan pendidikan Jepang melaksanakan siklus pemeriksaan buku pelajaran yang berhubungan dengan sejarah mengakibatkan perselisihan.
	No.34	Strategi yang AS mengejar dan menangkap Bin Laden, kepala unsur terorisme anti-AS.
英文版(机器翻译结果)	No.16	The various cultures and education of Japan implement cycle examination textbooks related to history resulted in a dispute.
	No.34	The U.S. strategy of pursuing and capturing Bin Laden, head of the U.S. anti-terrorism element.
前件扩展词(ARAE_SRCSAC)	No.16	controversi(0.85), seoul(0.82), asian(0.67), revis(0.77), beij(0.65), sport(0.75), ministry(0.74), technolog(0.78), south(0.84), history(0.85), publish(0.77), korea(0.84), approv(0.7), distort(0.76).
	No.34	terrorist(0.84), afghanistan(0.81), osama(0.82), dissid(0.63), millionair(0.74), saudi(0.75), rumsfeld(0.67), taliban(0.77), al(0.65) , mastermind(0.73).

表 5　查询实例DESC主题的检索性能比较

查询	检索算法	relax			rigid		
		p@5	p@10	MAP	p@5	p@10	MAP
No.16	MLR	0.4	0.3	0.403 1	0.2	0.1	0.146 9
	CLR	0.2	0.3	0.228 9	0	0.1	0.090 7
	ARAE_SRCSAC	0.2	0.5	0.495 7	0	0.3	0.223 6
	ARCE_SRCSAC	0.2	0.5	0.483 2	0	0.3	0.213 9
	ARACHE_SRCSAC	0.2	0.5	0.495 7	0	0.3	0.223 6
No.34	MLR	0.4	0.4	0.307 2	0.2	0.3	0.234 8
	CLR	0.2	0.2	0.211 9	0.2	0.2	0.184 3
	ARAE_SRCSAC	0.6	0.4	0.375 2	0.4	0.3	0.258 9
	ARCE_SRCSAC	0.2	0.2	0.265 9	0.2	0.2	0.224 4
	ARACHE_SRCSAC	0.6	0.4	0.375 2	0.4	0.3	0.258 9

3.4　实验结果分析

综上所述,本文提出的扩展算法有效,能改善和提高跨语言信息检索性能,遏制查询主题漂移和词不匹配问题,具有如下特点: 1)扩展算法检索结果MAP值普遍高于单语言基准检索、跨语言基准检索和对比方法; 2)支持度、置信度和关联度参数对扩展算法检索性能有较大的影响,随着参数阈值的增大,扩展算法的检索性能呈下降趋势,另外,置信度参数更有助于提高和改善规则前件扩展和规则混合扩展的检索性能,规则后件扩展在关联度参数的影响下可获得更好的检索结果; 3)扩展算法对长查询和短查询的检索均有效,而支持度和关联度参数对规则后件扩展算法的跨语言短查询检索更有效; 4)规则后件扩展的检索性能不如规则前件扩展和规则混合扩展,规则前件扩展对跨语言短查询检索效果更有效,规则混合扩展对跨语言长查询检索性能更有利.

本文扩展算法的有效性得益于如下3个方面的改进: 一是改进了加权项集挖掘方法,即采用支持度-关联度评价框架和基于有序项集的剪枝策略挖掘含有译后原查询词项的有效频繁项集,得到比较合理的特征词频繁项集,剪除更多无效的项集,挖掘效率得到提升; 二是改进了特征词关联规则评价框架,即采用卡方分析-置信度评价框架评估特征词关联规则,通过这些关联规则获得与原查询相关的优质译后扩展词; 三是改进了跨语言译后扩展模型,即提出基于SRCSAC评价框架挖掘的跨语言规则前件扩展、后件扩展和规则前后件混合扩展模型,以及扩展词权值和初检相关反馈文档特征词权值的计算方法. 以上3个方面共同作用,得到了与原查询关联性更高、更为有效的优质扩展词,提升了扩展词质量,提高和改善了检索性能,使得本文扩展算法的检索性能优于基准检索和对比方法.

4　结　论

本文主要研究了加权关联模式挖掘在跨语言查询译后扩展中的应用. 首先构建SRCSAC关联模式评价框架,提出基于该评价框架的加权关联规则挖掘算法,给出有序项集的相关理论及基于有序项集的新剪枝策略,研究关联规则混合扩展、规则前件扩展和规则后件扩展模型,给出扩展词权值计算新方法,最后给出基于SRCSAC评价框架挖掘的跨语言查询译后扩展算法,分析和比较算法参数及其设置对跨语言检索性能的影响. 实验结果表明,本文扩展算法有效,能遏制查询主题漂移和词不匹配问题,提高和改善跨语言信息检索性能. 另外,所提出的关联模式挖掘方法在文本挖掘、中国-东盟贸易商务数据挖掘以及推荐系统领域有着较高的应用价值. 本文不足之处是各个算法参数最优有效取值的数学模型没有得到深入讨论,下一步研究将深入探讨这些问题,并将本文算法应用到实际的跨语言搜索引擎中,以改善和提高实际跨语言信息检索系统性能.

参考文献(References)

[1] Mcnamee P, Mayfield J. Comparing cross-language query expansion techniques by degrading translation resources[C]. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2002: 159-166.

[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 D, He D Q, Wang H L. Cross-Language query expansion using pseudo relevance feedback[J]. Journal

of the China Society for Scientific and Technical Information, 2010, 29(2): 232-239.)

[3] 魏露, 李书琴, 李伟男, 等. 跨语言查询扩展优化[J]. 计算机工程与设计, 2014, 35(8): 2785-2788.
(Wei L, Li S Q, Li W N, et al. Optimization of cross-language query expansion[J]. Computer Engineering and Design, 2014, 35(8): 2785-2788.)

[4] Adriani M, Hayurani H, Sari S. Indonesian-English transitive translation for cross-language information retrieval[C]. Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum. Berlin: Springer Heidelberg, 2007: 127-133.

[5] Adriani M, Wahyu I. The performance of a machine translation-based English-Indonesian CLIR system[C]. Proceedings of the 6th International Conference on Cross-Language Evalution Forum. Berlin: Springer Heidelberg, 2005: 151-154.

[6] Hayurani H, Sari S, Adriani M. Query and document translation for English-Indonesian cross language IR[C]. Proceedings of the 7th Workshop of the Cross-Language Evaluation Forum. Berlin: Springer Heidelberg, 2006: 57-61.

[7] Chinnakotla M K, Raman K, Bhattacharyya P. Multilingual pseudo-relevance feedback: Performance study of assisting languages[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 1346-1356.

[8] Aditi Agrawal, 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.

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

[10] Chandra G, Dwivedi S K. Query expansion based on term selection for Hindi-English cross lingual IR[EB/OL]. [2019-04-17]. https://ac.els-cdn.com/S131915781730129 5/1-s2.0-S1319157817301295-main.pdf?tid=9e157d6a-8729-42b0-adc7-28d2f4483dce&acdnat=1555471922846df30c4a368fa2ea9fa61b3f27e401.

[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 M X. 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.)

[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 M X. 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.)

[13] 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 M X, Jiang C Q, He D L. Cross language query post-translation expansion based on matrix-weighted association rules[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(10): 887-898.)

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

[15] Cao G, Gao J, Nie J Y, et al. Extending query translation to cross-language query expansion with markov chain models[C]. Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York: ACM, 2007: 351-360.

[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 X M, Huang M X. All-weighted positive and negative association rules mining based on dynamic item weight[J]. Acta Electronica Sinica, 2015, 43(8): 1545-1554.)

[17] 张云涛, 龚玲. 数据挖掘原理与技术[M]. 北京: 电子工业出版社, 2004: 29-31.
(Zhang Y T, Gong L. Data mining principles and techniques[M]. Beijing: Electronics Industry Press, 2004: 29-31.)

[18] 黄名选, 黄发良, 严小卫, 等. 基于项权值变化和SCCI框架的加权正负关联规则挖掘[J]. 控制与决策, 2015, 30(10): 1729-1741.
(Huang M X, Huang F L, Yan X W, et al. Weighted positive and negative association rules mining based on dynamic item weight and SCCI framework[J]. Control and Decision, 2015, 30(10): 1729-1741.)

[19] 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 and Soft Computing, 2017, 23(4): 605-612.

作者简介

黄名选(1966−), 男, 教授, 硕士, 从事数据挖掘、信息检索和机器学习等研究, E-mail: mingxh05@163.com;

朱丽娜(1981−), 女, 副教授, 博士, 从事网络安全等研究, E-mail: zhulina81@163.com.

(责任编辑：郑晓蕾)