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

黄名选[1,2],蒋曹清[1,2]
(1. 广西跨境电商智能信息处理重点实验室培育基地(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: 主题漂移和词不匹配是自然语言处理中一个难题,文本挖掘与信息检索的结合有助于解决该问题. 鉴于此,本文提出一种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算法. 该算法采用新的完全加权正负项集支持度和关联度计算方法以及模式评价框架,对初检用户相关反馈文档集挖掘与原查询词相关的正负关联模式,从模式中提取扩展词实现跨语言查询译后扩展. 与现有基于伪相关反馈、加权关联模式挖掘的跨语言扩展算法比较,本文算法能有效地减少查询主题漂移和词不匹配问题,提高跨语言信息检索性能;本文模式挖掘方法可用于推荐系统,提高其准确性.

关键词: 自然语言处理; 信息检索; 文本挖掘; 模式挖掘; 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 (2018)12-3029-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.12.029

Vietnamese-English Cross Language Query Post-Translation Expansion Based on All-Weighted Positive and Negative Association Patterns Mining

HUANG Ming-xuan[1,2], JIANG Cao-qing[1,2]
(1. *Guangxi Key Laboratory Cultivation Base 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: Topic drift and word mismatch are a difficult problem in natural language processing. The combination of text mining and information retrieval can help to solve the problem. In view of this, this paper proposes an algorithm of Vietnamese-English cross language (VECL) query post-translation expansion based on all-weighted positive and negative association pattern mining. The algorithm utilized a computing method of support and correlation degree of all-weighted positive and negative itemset, and mined the all-weighted positive and negative association pattern related to the original query by the pattern evaluation framework in the user relevance feedback document set from the VECL first retrieval results. The expansion terms were extracted from the patterns in order to carry out VECL query post-translation expansion. A comparison between the proposed algorithm and the existing cross language query expansion algorithms based on pseudo relevance feedback and weighted association pattern mining is made, which shows that the former can effectively reduce the problems of query topic drift and word mismatch, and improve the performance of cross language information retrieval. And moreover, the method of pattern mining in this paper can be used in recommender systems and improve its accuracy.

Key words: natural language processing; information retrieval; text mining; pattern mining; query expansion; recommender system

收稿日期:2017-09-07;修回日期:2018-05-28;责任编辑:覃怀银
基金项目:国家自然科学基金(No. 61762006, No. 61662003, No. 61262028)

1 引言

当前,跨语言信息检索是自然语言处理领域里一个的研究热点,长期受到查询主题漂移、词不匹配以及查询项翻译歧义和多义等困扰. 跨语言查询扩展(Cross-Language Query Expansion,CLQE)是解决该类问题的核心技术之一,分为译前扩展[1,2]、译后扩展[3~11]和混合式扩展[12~14]. 近几年来,基于关联规则挖掘的查询译后扩展得到了研究,例如,文献[9]提出在平行语料全局文档中挖掘译后查询项实现译后扩展检索,文献[10,11]提出基于加权关联规则挖掘的跨语言译后扩展方法,都能提高和改善跨语言检索性能.

基于关联规则挖掘的跨语言译后扩展核心问题是如何计算关联模式支持度. 常见支持度计算主要有四种:(1)将关联模式在事务文档中发生的概率作为该模式的支持度[9];(2)将项目权值总和与无加权支持度的乘积作为加权项集支持度[15];(3)将特征词项目平均权值与无加权支持度的乘积作为完全加权项集支持度[11,16];(4)以项集在事务数据库中项集权值总和占事务数据库中所有项目权值总和的百分比作为完全加权项集支持度[10,17]. 文献[17]表明,方法(4)挖掘效果比方法(3)的好. 然而,方法(4)只考虑特征词项目权值对支持度的影响,忽略特征词频度对支持度的作用.

当前,跨语言查询扩展研究主要针对大语种以及欧洲国家语言等等,而针对东盟语言的报道不多. 鉴于此,本文以东盟语言为研究对象,提出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算法. 该算法采用新的完全加权正负项集支持度和关联度计算方法以及模式评价框架,对跨语言初检用户相关反馈文档集挖掘译后查询扩展词. 与单语言、跨语言检索基准和现有基于伪相关反馈、加权关联模式挖掘的跨语言查询扩展算法比较,本文算法能有效地减少跨语言信息检索中查询主题漂移和词不匹配问题,提高和改善跨语言检索性能. 本文模式挖掘方法在推荐系统具有一定的应用价值,能提高其准确性.

2 面向跨语言查询扩展的完全加权正负关联模式挖掘

2.1 融合项权值和频度的完全加权项集支持度

针对现有支持度计算方法的缺陷,本文提出融合项权值和频度的完全加权项集 I 支持度(New all-weighted Itemset Support,NawISup)的计算方法,如式(1)所示:

$$\mathrm{NawISup}(I)=\alpha\times\frac{n_I}{n}+(1-\alpha)\times\frac{w_I}{W\times k} \tag{1}$$

其中,n 和 W 为文档集中文档总数和所有特征词项目权值总和,n_I 为 I 在文档集中出现的频度,w_I 为 I 在文档集中的项集权值总和,k 为 I 的长度,$\alpha\in(0,1)$ 为插值系数.

在式(1)基础上,本文给出完全加权负项集支持度及正负关联规则($I_1\rightarrow I_2$, $I_1\rightarrow\neg I_2$, $\neg I_1\rightarrow I_2$)置信度(all-weighted Association Rule Confidence,awARConf)的计算公式,如式(2)至(7)所示.

$$\mathrm{NawISup}(\neg I)=1-\mathrm{NawISup}(I) \tag{2}$$

$$\mathrm{NawISup}(I_1\cup\neg I_2)=\mathrm{NawISup}(I_1)-\mathrm{NawISup}(I_1\cup I_2) \tag{3}$$

$$\mathrm{NawISup}(\neg I_1\cup I_2)=\mathrm{NawISup}(I_2)-\mathrm{NawISup}(I_1\cup I_2) \tag{4}$$

$$\mathrm{awARConf}(I_1\rightarrow I_2)=\frac{\mathrm{NawISup}(I_1\cup I_2)}{\mathrm{NawISup}(I_1)} \tag{5}$$

$$\mathrm{awARConf}(\neg I_1\rightarrow I_2)=\frac{\mathrm{NawISup}(I_2)-\mathrm{NawISup}(I_1\cup I_2)}{1-\mathrm{NawISup}(I_1)} \tag{6}$$

$$\mathrm{awARConf}(I_1\rightarrow\neg I_2)=1-\mathrm{awARConf}(I_1\rightarrow I_2) \tag{7}$$

2.2 完全加权正负项集关联度

针对现有关联度的缺陷,本文提出完全加权正项集 PI(Positive Itemset)关联度(all-weighted PI Relevancy,awPIR)的计算如式(8)所示:

$$\mathrm{awPIR}(\mathrm{PI})=\begin{cases}=\dfrac{\mathrm{NawISup}(\mathrm{PI})}{\mathrm{NawISup}(t_k)}, & m=2\\ \dfrac{1}{2}\times\mathrm{NawISup}(\mathrm{PI})\times\left(\dfrac{1}{\mathrm{NawISup}(t_k)}+\dfrac{1}{\mathrm{NawISup}(I_q)}\right), & m>2\end{cases} \tag{8}$$

其中,m 为 PI 的长度,$t_k(1\le k\le m)$ 是 PI 的所有项目中其支持度最大的单项目,I_q 为 PI 的所有 2_子项集至($m-1$)_子项集中其支持度最大的子项集.

同理,完全加权负项集 NI(Negative Itemset)关联度(all-weighted Negative Itemset Relevancy,awNIR)的计算如式(9)所示:

$$\mathrm{awNIR}(\mathrm{NI})=\begin{cases}=\dfrac{\mathrm{NawISup}(\mathrm{NI})}{1-\mathrm{NawISup}(t_p)}, & r=2\\ \dfrac{1}{2}\times\mathrm{NawISup}(\mathrm{NI})\times\left(\dfrac{1}{1-\mathrm{NawISup}(t_p)}+\dfrac{1}{1-\mathrm{NawISup}(I_s)}\right), & r>2\end{cases} \tag{9}$$

其中,r 为 NI 的长度,$t_p(1 \le p \le r)$是 NI 的所有项目中其支持度最大的单项目,I_s为 NI 的所有 2_子项集至($r-1$)_子项集中其支持度最大的子项集.

2.3　完全加权关联规则提升度

完全加权关联规则($I_1 \rightarrow I_2$)提升度(all-weighted Association Rule Lift,awARL)的计算如式(10)所示.

$$\text{awARL}(I_1 \rightarrow I_2) = \frac{\text{awARConf}(I_1 \rightarrow I_2)}{\text{NawISup}(I_2)} \tag{10}$$

2.4　面向跨语言查询扩展的完全加权正负关联模式挖掘

基本挖掘思想:对跨语言初检用户相关反馈文档集采用支持度-关联度-提升度-置信度评价框架挖掘含有原查询项的特征词正负关联规则模式(即 $I_1 \rightarrow I_2$、$I_1 \rightarrow \neg I_2$和$\neg I_1 \rightarrow I_2$). 上述挖掘思想形式化为算法 AWPNM-CLQE (All-Weighted Positive and Negative Patterns Mining for CLQE),其中,DS 为初检用户相关反馈文档,ms(minimum support)和 mc(minimum confidence)分别为最小支持度和置信度阈值,Q 为用户查询,L_{item}为候选项集长度阈值,PAR(Positive Association Rule)为强正关联规则集,NAR(Negative Association Rule)为强负关联规则集,PIS(Positive ItemSet)为正项集集合,NIS(Negative ItemSet)为负项集集合,minPR 和 minNR 分别为最小正项集关联度阈值和负项集关联度阈值.

算法 1　AWPNM-CLQE

输入:DS,ms,mc,α,minPR,minNR,Q,L_{item}
输出:PAR,NAR
(1) Pretreatment(DS);
(2) L_1 = MiningAWPL1(DS);
(3) for (k = 2;$L_k \neq \varnothing$;k + +)do
  ①PIS←PIS∪L_{k-1};NIS←NIS∪N_{k-1};
  ②C_k←$L_{k-1} \infty L_{k-1}$;
  ③if (k = 2) then C_k←PruningNotQ(C_k,Q);
  ④MiningAWP (DS,ms,α,minPR,minNR,Output L_k,Output N_k);
  ⑤if ($k > L_{\text{item}}$) then Break;
(4) for each positive itemset L_k in PIS do
  for each itemset (qt,PI) in L_k do
  if ((qt∪PI = L_k) and (qt∩PI = ∅) and (qt⊆Q)) then
    begin
    If ((awARL(qt→PI) > 1) and (awARConf (qt→PI) > = mc)) then PAR←PAR∪{ qt→PI };
    If ((awARL(PI→qt) > 1) and (awARConf(PI→qt) > = mc)) then PAR←PAR∪{ PI→qt};
    end;
(5) for each negative itemset N_k in NIS do
  for each itemset (qt,NI) in N_k do
    If ((qt∪NI = N_k) and (qt∩NI = ∅) and (qt⊆Q)) then
      if (awARL(qt→NI) < 1) then
        begin
        If (awARConf (qt→¬NI) > = mc) then NAR←NAR∪{ qt→¬NI };
        If (awARConf (¬ qt→NI) > = mc) then NAR←NAR∪{¬qt→NI };
        end;
      If (awARL(NI→qt) < 1) then
        begin
        If (awARConf (NI→¬qt) > = mc) then NAR←NAR∪{ NI→¬ qt };
        If (awARConf (¬NI→qt) > = mc) then NAR←NAR∪{¬NI→qt };
        end;
(6) Return PAR,NAR;

AWPNM-CLQE 算法中,步骤(3)挖掘 L_k和 N_k,步骤(4)挖掘强正关联规则,步骤(5)挖掘强负关联规则. 过程 Pretreatment( )对文档集进行预处理;函数 MiningAWPL1( )挖掘 L_1;过程 PruningNotQ( )剪除不含查询词项 Q 的候选 2_项集 C_2,以及小于 minPR 的正项集和小于 minNR 的负项集;过程 MiningAWP ( )挖掘 L_k和 N_k,并输出 L_k和 N_k.

3　基于完全加权正负关联模式的越-英跨语言查询译后扩展

3.1　跨语言查询译后扩展模型及其扩展词权值的计算

本文跨语言查询译后扩展模型分为查询译后后件扩展(Post-Translation Consequent Expansion,PTCE)模型和前件扩展(Post-Translation Antecedent Expansion,PTAE)模型,其扩展模型结构如图 1 所示.

后件扩展模型 PTCE 指的是正扩展词和负扩展词来自正负关联规则模式的后件,正扩展词中去除负扩展词后余下的正扩展词即为最终扩展词,其模型形式化为式(11)和(12)所示.

$$\text{qt}^{\text{tl}} \rightarrow \text{PEt}^{\text{tl}}, \text{qt}^{\text{tl}} \rightarrow \neg \text{NEt}^{\text{tl}}, \neg\text{qt}^{\text{tl}} \rightarrow \text{NEt}^{\text{tl}} \quad (\text{ms},\text{mc},\alpha,\text{minPR},\text{minNR}) \tag{11}$$

$$\text{PEt}^{\text{tl}} - \text{NEt}^{\text{tl}} \rightarrow \text{CEt}^{\text{tl}} \tag{12}$$

其中,qt^{tl}表示目标语言查询词项集,PEt^{tl}表示正扩展词项集,NEt^{tl}表示负扩展词项集,CEt^{tl}表示最终后件扩展词项集. 同理,前件扩展模型 PTAE 指的是正扩展词和负扩展词来自正负关联规则模式的前件,其模型形式化表示为式(13)和(14)所示,其中,AEt^{tl}表示目标语言最终前件扩展词项集.

$$\text{PEt}^{\text{tl}} \rightarrow \text{qt}^{\text{tl}}, \text{NEt}^{\text{tl}} \rightarrow \neg \text{qt}^{\text{tl}}, \neg \text{NEt}^{\text{tl}} \rightarrow \text{qt}^{\text{tl}} \quad (\text{ms},\text{mc},\alpha,\text{minPR},\text{minNR}) \tag{13}$$

$$\text{PEt}^{\text{tl}} - \text{NEt}^{\text{tl}} \rightarrow \text{AEt}^{\text{tl}} \tag{14}$$

本文给出译后扩展词权值 W_{exp} 计算公式如式(15)所示.

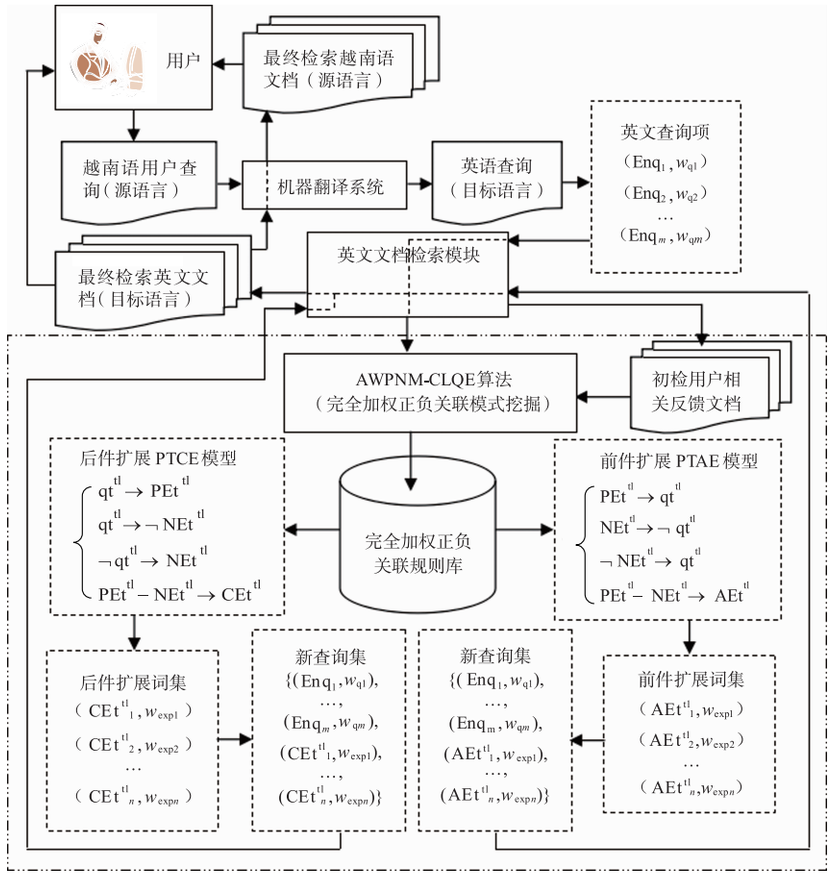

图1 越–英跨语言查询译后扩展模型结构图

$$W_{\exp}=\frac{1}{2}\times[\max(\text{awARConf})+0.5\times(\max(\text{awARL})-1)] \tag{15}$$

另外，原查询项的权值计算采用传统的 tf-idf 方法，具体计算公式见文献[18].

3.2 跨语言查询译后扩展算法

基于完全加权正负关联模式的跨语言查询译后扩展基本思想：将越南语查询机器翻译为英文并检索英文文档，提取跨语言初检前列 n 篇文档通过用户相关性判断后构建初检相关英文文档集，调用 AWPNM-CLQE 算法对初检相关文档集挖掘含有原查询词项的完全加权正负关联规则模式，根据 PTCE 模型实现译后后件扩展，根据 PTAE 模型实现译后前件扩展. 上述扩展思想形式化为后件扩展算法 PTCE_AWPNP（PTCE Based on All-weighted Positive and Negative Patterns）和前件扩展算法 PTAE_AWPNP（PTAE Based on All-weighted Positive and Negative Patterns），其中，Q_{Vi} 为越南语用户查询，n 为初检前列文档数，ConsequentET 为后件扩展词集，AntecedentET 为前件扩展词集，NewQ_{En} 为扩展后新查询，其余同算法 1.

算法 2 PTCE_AWPNP

输入：Q_{Vi}，n，L_{item}，ms，mc，α

输出：ConsequentET，New Q_{En}

(1) AWPNRules = CreatePNRules(Q_{Vi}，n，ms，mc，minPR，minNR，α，L_{item})；

(2) CandidateET = CreateCET(AWPNRules，qt→PEt)；

(3) NegativeET = CreateNegativeCET(AWPNRules，qt→¬NEt，¬qt→NEt)；

(4) Consequent ET = CreateFinalET(CandidateET，NegativeET)；

(5) New Q_{En} = $Q_{\text{Vi}}\cup$ConsequentET；

(6) Return New Q_{En}；

算法 3 PTAE_AWPNP

输入：Q_{Vi}，n，L_{item}，ms，mc，α

输出：AntecedentET，New Q_{En}

(1) AWPNRules = CreatePNRules(Q_{Vi}，n，ms，mc，minPR，minNR，α，L_{item})；

(2) CandidateET = CreateAET(AWPNRules，PEt→qt)；

(3) NegativeET = CreateNegativeAET(AWPNRules，NEt→¬qt，¬NEt→qt)；

(4) AntecedentET = CreateFinalET(CandidateET, NegativeET);
(5) NewQ_{En} = Q_{Vi} ∪ AntecedentET;
(6) Retrun NewQ_{En};

上述算法 2 和 3 中,函数 CreatePNRules( ) 调用 AWPNM-CLQE 算法对初检用户相关反馈文档集挖掘正负关联规则,函数 CreateCET( ) 提取候选后件扩展词,函数 CreateAET( ) 提取候选前件扩展词,函数 CreateNegativeCET( ) 提取后件负扩展词,函数 CreateNegativeAET( ) 提取前件负扩展词,函数 CreateFinalET( ) 获取最终扩展词.

4　实验设计及结果分析

为了验证本文扩展算法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,并在该平台进行本文越-英跨语言查询译后扩展实验. 实验所用的机器翻译接口是 Microsoft Translator API.

4.1　实验数据及预处理

选择跨语言标准数据集语料 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)作为本文实验数据,即数据集是 $m0$、$m1$ 和 $k1$,其中 $m0$ 数据集来源于 Mainichi Daily News 新闻媒体 2000 年的新闻文本,共 6608 篇文本文档,$m1$ 来源于 Mainichi Daily News 新闻媒体 2001 年的新闻文本,共 5547 篇文本文档,$k1$ 来源于 Korea Times 2001 年的新闻文本,共 14069 篇文本文档. 本文采用 TITLE 和 DESC 查询进行检索实验. 数据预处理时采用 Porter 程序(详见:http://tartarus.org/~martin/PorterStemmer) 进行词干提取. 本文源语言越南语查询由专业翻译人员对 NTCIR-5 CLIR 语料的 50 个中文版查询语料人工翻译得到.

4.2　评价指标及基准算法

采用平均查准率的均值 MAP(Mean Average Precision) 作为检索评价指标. 本文基准对比算法如下:

基准实验:(1) 单语言检索(Monolingual Retrieval, MR) 基准:使用 NTCIR-5 CLIR 的英文版查询直接检索英文文档;(2) 越-英跨语言检索(Vietnamese-English Cross-Language Retrieval, VECLR) 基准:即将越南语查询机器翻译为英文后检索英文文档.

对比实验:(1) 基于文献[14]伪相关反馈的越-英跨语言查询译后扩展[14](Post-Translation Expansion based on Pseudo Relevance Feedback, PTE_PRF);(2) 基于加权关联模式挖掘的越-英跨语言查询扩展[10](VECLQE_WAR[10]),实验参数:n = 50, ms ∈ {0.001, 0.002, 0.003, 0.004, 0.005}, L_{item} = 2, λ = 100, mc = 0.01, mi = 0.0001;(3) 基于文献[16]完全加权正负关联规则挖掘技术的越-英跨语言查询译后后件扩展(Consequent Expansion based on All-Weighted Positive and Negative Association Rule (AWPNAR), CE_AWPNAR) 和前件扩展(Antecedent Expansion based on AWPNAR, AE_AWPNAR),实验参数:n = 50, L_{item} = 2, mc = 0.5, minInt = 0.02, ms ∈ {0.1, 0.2, 0.3, 0.4, 0.5}.

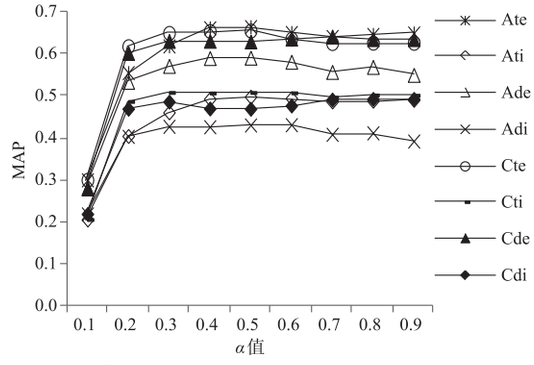

图2　本文算法在$m0$数据集的检索结果

4.3　检索性能比较

本节考察α对本文算法检索性能的影响,通过实验分析比较本文算法和基准对比算法的跨语言检索性能. 实验中参数选择依据是:尽量在该参数的有效范围内选择比较有效的实验参数值进行实验,带有一定的随机性.

4.3.1　插值系数 α 对本文算法检索性能的影响

本节分析比较参数 α 对本文算法检索性能的影响. 在 α 有效范围内,50 个查询在 3 个数据集 $m0$、$m1$ 和 $k1$ 中进行本文算法实验,其检索结果 MAP 值如图 2 至图 4 所示. 其中,实验参数:ms = 0.2, mc = 0.8, L_{item} = 2, minPR = 0.1, minNR = 0.01. 图例中,“t”和“d”分别表示 TITLE 和 DESC 查询,“e”代表 Relax 值,“i”代表 Rigid 值,“A”代表 PTAE_AWPNP 算法,“C”代表 PTCE_AWPNP 算法.

图 2 至图 4 表明,当 α = 0.1 且 ms = 0.2 时,没有挖掘出扩展词,该 α 值的 MAP 值最小,但实验时发现,在

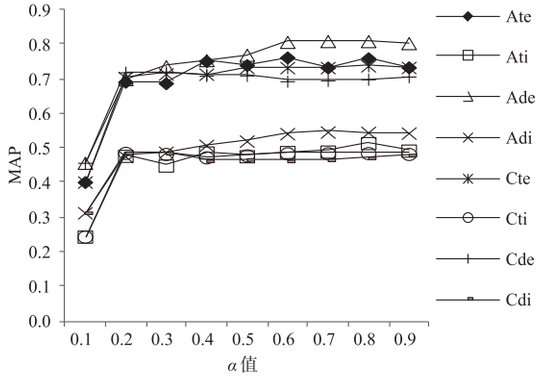

图3　本文算法在$m1$数据集的检索结果

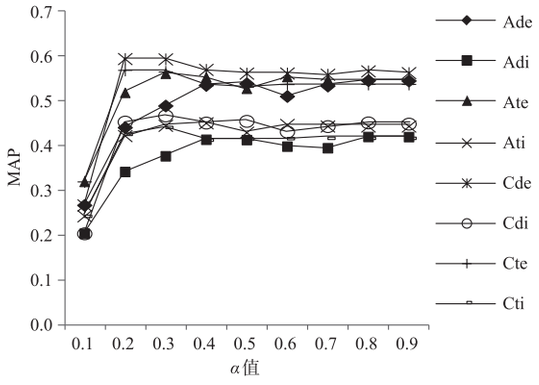

图4 本文算法在$k1$数据集的检索结果

该种情况下,如果选择合适的 ms 值也能挖掘出扩展词. 当 $\alpha >= 0.2$,MAP 值开始急剧上升,当 $\alpha >= 0.3$,变化趋于平缓,不随 α 值的变化而大幅度震荡,而是呈现缓慢变化的趋势,说明本文算法的检索性能具有一定的鲁棒性,由此可见,参数α是有效的,能改善检索性能. 另外,参数 α 在 3 个数据集的检索性能表现不一致,呈现多样化.

4.3.2 检索性能比较

在数据集 $m0$、$m1$ 和 $k1$ 上分别进行基准对比算法和本文算法的越-英跨语言检索实验. 实验时,提取跨语言初检前列 n 篇英文文档进行用户相关性判断(为了简便,本文实验将初检前列 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. 本文算法和基准对比算法的实验结果 MAP 如表 1 和表 2 所示,其中,本文算法实验参数:$n=50$,$L_{item}=2$,minPR = 0.1,minNR = 0.01,ms = 0.2,mc = 0.8, $m0$ 数据集:$\alpha=0.5$,$m1$:$\alpha=0.8$,$k1$:$\alpha=0.3$.

表 1 和表 2 表明,本文算法 MAP 值比基准对比算法的高,而且大多数 MAP 值提高的幅度比较大,性能提升效果显著,但也存在少数 MAP 值低于对比算法的,说明本文算法还存在不稳定性,需要进一步研究.

表 1 本文算法与基准对比算法的检索结果 MAP 值(TITLE 查询)

检索算法	Relax			Rigid		
	$m0$	$m1$	$k1$	$m0$	$m1$	$k1$
MR	0.4438	0.5664	0.3288	0.3552	0.4011	0.2305
VECLR	0.2999	0.4031	0.3213	0.2055	0.2451	0.2443
PTE_PRF	0.3149	0.3656	0.2233	0.2222	0.2246	0.1414
CE_AWPNAR	0.5086	0.5654	0.3982	0.3931	0.3833	0.3020
AE_AWPNAR	0.3953	0.5230	0.3555	0.2881	0.3555	0.2611
VECLQE_WAR	0.6136	0.6755	0.5903	0.4718	0.4092	0.4572
PTCE_AWPNP	0.6549	0.7397	0.5975	0.5076	0.4871	0.4433
PTAE_AWPNP	0.6643	0.7640	0.5634	0.4977	0.4900	0.4579

表 2 本文算法与基准对比算法的检索结果 MAP 值(DESC 查询)

检索算法	Relax			Rigid		
	$m0$	$m1$	$k1$	$m0$	$m1$	$k1$
MR	0.2927	0.5664	0.2548	0.2301	0.4011	0.2548
VECLR	0.2798	0.4589	0.2686	0.2187	0.3142	0.2686
PTE_PRF	0.2549	0.2515	0.1583	0.1996	0.1710	0.1583
CE_AWPNAR	0.5229	0.4748	0.4265	0.3966	0.3226	0.3127
AE_AWPNAR	0.4429	0.4815	0.3933	0.3309	0.3507	0.2955
VECLQE_WAR	0.5891	0.6893	0.5924	0.4458	0.4691	0.5924
PTCE_AWPNP	0.6407	0.7224	0.5955	0.4908	0.4868	0.5990
PTAE_AWPNP	0.5905	0.8114	0.5479	0.4303	0.5480	0.4219

4.3.3 查询实例检索效果分析

为了说明本文算法能有效地减少跨语言查询主题漂移和词不匹配现象,列举 DESC 查询类型的 No.25 和 No.40 查询主题在 $k1$ 数据集检索的实验结果(如表 3 和表 4 所示)加以说明.

表 3　No. 025 和. 040 查询及其后件扩展词(ms = 0. 2, mc = 0. 8, L_{item} = 3, minPR = 0. 1 minNR = 0. 01)

查询语言版本	No. 025 查询主题	No. 040 查询主题
越文版	Truyền thông Thể thao hoặc ngành thể thao liên quan tuyên duong Tiger Woods là ngôi sao thể thao.	Tình hình tiêu thể của sách bán chạy nhất (best seller), Truyện Harry Potter trên toàn thế giới.
英文版 (机器翻译结果)	Sports media or sports related industries commend Tiger Woods as sports stars.	Consumption situation of best selling books (best seller), theHarry Potter books all over the world.
英文版 (语料中的原版)	Find documents about sports media or relatedenterprises recognizing Tiger Woods as a sports star.	Find documents describing worldwide circulation of the super best seller, Harry Potter.
扩展词词干 (PTCE_AWPNP)	david, surpris, tournament, win, won, victory, tom, tour, shot, par, golferchalleng, consist, bob, duval, finish, golf, major, master, mickelson, player, pga, perform, play, open, name, game, championship.	seri, movi, theater, base, author, boy

表 4　No. 025 和 040 查询的检索性能比较

查询实例	检索算法	MAP
No. 025 查询	MR	0. 4770
	VECLR	0. 3875
	PTCE_AWPNP	0. 7659
No. 040 查询	MR	0. 6515
	VECLR	0. 5184
	PTCE_AWPNP	0. 8151

从表 4 可看出,VECLR 检索结果 MAP 值比 MR 检索的低,说明跨语言检索过程中查询主题经过机器翻译后发生了主题漂移和词不匹配现象,而本文算法的 MAP 值比 VECLR 的高,由此可见,本文算法能有效地遏制查询主题漂移和词不匹配现象.

4.4　实验结果分析

综上所述,本文算法能改善和提高跨语言信息检索性能. 其主要原因分析如下:本文将完全加权正负关联模式挖掘应用于跨语言查询译后扩展,采用一种新的融合项权重和频度的支持度以及正负项集关联度计算方法,结合用户相关反馈技术,在新的支持度-关联度-提升度-置信度评价框架下挖掘译后扩展词实现译后查询扩展,跨语言检索性能得到改善和提升. 而对比算法 VECLQE_WAR、CE_AWPNAR 和 AE_AWPNAR 在挖掘扩展词时,没有考虑项集关联度和提升度,使得挖掘出来的扩展词质量(即与原查询的相关性)不如本文算法的;对比算法 PTE_PRF 的扩展词直接来自初检前列文档前列特征词,难以避免虚假扩展词,其检索性能不如本文算法的.

5　结论

本文提出基于完全加权正负关联模式挖掘的越-英跨语言查询译后后件扩展和前件扩展算法. 该算法采用新的完全加权正负项集支持度和关联度计算方法,对跨语言初检用户相关反馈文档采用新的完全加权支持度-关联度-提升度-置信度评价框架挖掘译后扩展词实现跨语言译后扩展. 实验表明,本文算法能有效地减少跨语言检索中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. 本文所提出的模式挖掘方法在推荐系统具有较好的应用价值,可提高其准确性. 本文存在的不足是:用户相关反馈的文档数量不多,通过负扩展词发现候选扩展词中虚假扩展词的效果不是很明显,参数 α, ms, mc, minPR, minNR 对本文算法的影响没有得到深入讨论,等等,这些问题需要进一步深入研究.

参考文献

[1] 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.

[2] 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)

[3] 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.

[4] 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.

[5] Bellaachia A, AmorTijani G. Enhanced query expansion in English-Arabic CLIR[A]. Proceedings of the 19th International Conference on Database and Expert Systems Application[C]. Washington, DC, USA: IEEE Computer Society, 2008. 61 – 66.

[6] Chinnakotla M K, Raman K, Bhattacharyya P. Multilingual pseudo-relevance feedback: performance study of assisting languages[A]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics[C]. Stroudsburg, PA, USA: ACL, 2010. 1346 – 1356.

[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] Chandra G, Dwivedi S K. Query expansion based on term selection for Hindi-English cross lingual IR[J]. Journal of King Saud University-Computer and Information Sciences, 2017, 29(1): 1 – 10.

[9] 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.

[10] 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)

[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] 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.

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

[14] 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 and 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)

[15] Cai C H, Da A, Fu W C, et al. Mining association rules with weighted items[A]. Proceedings of the 1998 International Symposium on Database Engineering & Applications[C]. Washington, DC, USA: IEEE Computer Society, 1998. 68 – 77.

[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)

[17] 周秀梅, 黄名选. 基于项权值变化的矩阵加权关联规则挖掘[J]. 计算机应用研究, 2015, 32(10): 2918 – 2923.
ZHOU Xiu-mei, HUANG Ming-xuan. Matrix-weighted association rules mining based on dynamic weight of item [J]. Application Research of Computers, 2015, 32(10): 2918 – 2923. (in Chinese)

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

作者简介

黄名选 男, 1966 年出生于广西乐业县, 工学硕士, 现为广西财经学院计算机系教授, 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 4 篇, ISTP 收录 1 篇, 授权的发明专利 9 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.
Email: jcqng@163.com