Postprint of an Indonesian-Chinese Cross-Language Information Retrieval Model Based on Matrix-Weighted Association Patterns

Authors: Huang Mingxuan

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: To address the query drift problem in cross-language information retrieval, we propose an Indonesian-to-Chinese cross-language information retrieval model that integrates user click-and-download behavior with matrix-weighted association pattern mining.

Methods: Matrix-weighted association pattern mining, query expansion, and user click-and-download behavior are integrated into the Indonesian-to-Chinese cross-language information retrieval model, and the key implementation technologies are presented, namely a matrix-weighted association pattern mining algorithm for cross-language information retrieval, a cross-language query expansion model, and an Indonesian-to-Chinese cross-language information retrieval algorithm.

Results: Experimental results on the NTCIR-5 CLIR dataset demonstrate that the retrieval model' s R_{prec}, p@10, and p@20 values all exceed 60% of the monolingual retrieval baseline, representing improvements of over 37% compared to the cross-language retrieval baseline and over 28% compared to existing pseudo-relevance feedback-based cross-language retrieval algorithms.

Limitations: The model experiments were conducted within a cross-language retrieval system based on the vector space model; further investigation of its specific application in practical search engines is required.

Conclusion: This model can effectively mitigate the query drift problem in cross-language retrieval, improve Indonesian-to-Chinese cross-language retrieval performance, demonstrate superior effectiveness for long queries, and hold significant practical application value.

Full Text

Preamble

Cross-Language Information Retrieval Model for Indonesian-Chinese Based on Matrix-Weighted Association Pattern Mining

Guangxi Key Laboratory Cultivation Base of Cross-border E-commerce Intelligent Information Processing (Guangxi University of Finance and Economics), Nanning 530003
Department of Computer Science, Guangxi University of Finance and Economics, Nanning 530003

Abstract

[**Objective**] To address the query drift problem in cross-language information retrieval, this paper proposes an Indonesian-Chinese cross-language information retrieval model that integrates user click-download behaviors with matrix-weighted association pattern mining. [**Methods**] The model integrates matrix-weighted association pattern mining, query expansion, and user click-download behaviors into an Indonesian-Chinese cross-language information retrieval framework. Key implementation technologies are presented, including a matrix-weighted association pattern mining algorithm for cross-language information retrieval, a cross-language query expansion model, and an Indonesian-Chinese cross-language information retrieval algorithm. [**Results**] Experimental results on the NTCIR-5 CLIR dataset demonstrate that the proposed retrieval model achieves R_{prec}, p@10, and p@20 values exceeding 60% of the monolingual retrieval baseline, representing improvements of over 37% compared to the cross-language retrieval baseline and over 28% compared to existing pseudo-relevance feedback-based cross-language retrieval algorithms. [**Limitations**] The model experiments were conducted within a vector space model-based cross-language retrieval system, requiring further investigation into practical applications in real search engines. [**Conclusions**] The proposed model effectively reduces query drift in cross-language retrieval and improves Indonesian-Chinese cross-language retrieval performance, demonstrating particularly superior results for long queries and showing promising practical application value.

Keywords: Click Behavior; Association Pattern Mining; Indonesian-Chinese Cross-Language Retrieval Model; Cross-Language Information Retrieval; Matrix-Weighted Association Rule
Classification Number: TP311

Cross-language information retrieval refers to the technology of retrieving information resources in other languages using a query in one language. Indonesian-Chinese cross-language information retrieval specifically involves using Indonesian queries to retrieve Chinese documents, where the Indonesian language used for querying is called the source language (SL) and Chinese is the target language (TL). Scholars worldwide have conducted in-depth research on cross-language

information retrieval models and algorithms from various perspectives, yielding rich theoretical results. However, several challenges in cross-language information retrieval remain unresolved, with one of the most pressing and widely discussed issues being that cross-language retrieval faces more severe term mismatch and topic drift problems than monolingual retrieval, often resulting in poor retrieval performance.

In response to these challenges, research on query expansion-based cross-language information retrieval has gained increasing attention in recent years, focusing primarily on approaches based on relevance feedback [1-6], latent semantics [7-10], language models [11], and topic models [12-16], with English being the primary language object in most studies investigating cross-language retrieval between English and other languages. Relevance feedback-based cross-language information retrieval utilizes top-ranked documents from initial cross-language retrieval results as sources for expansion terms to achieve query expansion, followed by a second retrieval pass. Latent semantics-based cross-language information retrieval employs latent semantic analysis techniques to establish correspondences between different languages, discovering target language feature terms related to the original query to achieve cross-language query expansion and improve retrieval performance. Research on cross-language information retrieval based on language models and topic models has also become increasingly active.

As evident from relevant literature, research on cross-language information retrieval for ASEAN languages remains scarce. Since Nanning, China became the permanent host city for the China-ASEAN Expo, political, economic, and cultural exchanges between China and ASEAN countries have become more frequent and close, making research on cross-language information retrieval and services for ASEAN languages increasingly urgent and important. Building upon the aforementioned research achievements, this paper investigates cross-language information retrieval for ASEAN languages, focusing on Indonesian and Chinese. The study integrates matrix-weighted association rule mining technology, user click behaviors, and query expansion techniques into Indonesian-Chinese cross-language information retrieval, proposing an Indonesian-Chinese cross-language information retrieval model based on matrix-weighted association pattern mining and its key implementation technologies, including a matrix-weighted association pattern mining algorithm for cross-language information retrieval, a cross-language query expansion model, and an Indonesian-Chinese cross-language information retrieval algorithm.

2. Indonesian-Chinese Cross-Language Information Retrieval Model Based on Matrix-Weighted Association Pattern Mining

2.1 Design Philosophy

The fundamental concept of the Indonesian-Chinese cross-language information retrieval model based on matrix-weighted association pattern mining is as follows: First, Indonesian queries are translated into Chinese queries through a machine translation system and submitted to a search engine to retrieve Chinese documents cross-lingually. User browsing, clicking, and downloading behaviors on initial retrieval documents confirm these documents as user feedback-relevant documents. The matrix-weighted association pattern mining technique proposed in this paper is then applied to mine Chinese query-related expansion terms from these initially retrieved relevant documents to achieve post-translation cross-language query expansion. The expansion terms are combined with the original query and resubmitted to the search engine for retrieval, with the final retrieval results translated into Indonesian documents and returned to users.

2.2 Model Architecture and Module Functions

Based on the design philosophy described above, Figure 1 [Figure 1: see original paper] illustrates the architecture of the Indonesian-Chinese cross-language information retrieval model based on matrix-weighted association pattern mining. The model comprises eight modules and three databases: a machine translation module, search engine module, user click behavior relevance feedback extraction module, document preprocessing module, matrix-weighted association rule mining module for Indonesian-Chinese cross-language retrieval, cross-language query expansion term generation module, cross-language query expansion implementation module, and final result display module, along with an initially retrieved relevant document database, matrix-weighted association rule base, and expansion term base.

(1) **Machine Translation Module**: Utilizes the Bing machine translation interface, specifically the Microsoft Translator API , to translate user-submitted Indonesian queries into Chinese queries and translate final retrieval result Chinese documents into Indonesian documents for user delivery.

(2) **Search Engine Module**: Employs search engines such as Google or Baidu to retrieve Chinese documents on the Internet using the translated Chinese queries, yielding the initial cross-language retrieval result document set.

(3) **User Click Behavior Relevance Feedback Extraction Module**: Captures document download behaviors generated by users when browsing the initial retrieval result document set, extracting user-downloaded

initial retrieval documents to construct a user feedback relevant document set.

(4) **Document Preprocessing Module**: Performs Chinese word segmentation, stop word removal, and feature term extraction on the user feedback relevant document set to construct a user feedback initially retrieved relevant document database.

(5) **Matrix-Weighted Association Rule Mining Module for Indonesian-Chinese Cross-Language Retrieval**: Conducts matrix-weighted association rule mining on the aforementioned user feedback initially retrieved relevant document set, primarily mining matrix-weighted feature term frequent itemsets and association rule patterns containing original query terms to construct a matrix-weighted association rule base.

(6) **Cross-Language Query Expansion Term Generation Module**: Extracts expansion terms related to the original query from the matrix-weighted association rule base to construct an expansion term base.

(7) **Cross-Language Query Expansion Implementation Module**: Extracts Chinese expansion terms from the expansion term base, combines them with the original query to form a new query, resubmits it to the search engine for Internet retrieval, and obtains the final retrieved Chinese documents.

(8) **Final Result Display Module**: Submits the final retrieved Chinese documents to the machine translation module for translation into Indonesian documents and returns both the Chinese and Indonesian documents to the user.

2.3 Key Technologies of the Indonesian-Chinese Cross-Language Information Retrieval Model

(1) Matrix-Weighted Association Rule Mining for Indonesian-Chinese Cross-Language Retrieval The fundamental concept of matrix-weighted association rule mining for Indonesian-Chinese cross-language retrieval is as follows: First, the user click behavior relevance feedback information extraction module obtains the initial cross-language retrieval results, i.e., the user feedback target language document set DocTL. The document preprocessing module then preprocesses DocTL to construct a user feedback initially retrieved relevant document database. Combined with the user query, a three-stage itemset pruning strategy is employed to mine matrix-weighted feature term association rules containing user query terms from the initially retrieved relevant document database, constructing a matrix-weighted association rule base. The specific pruning strategies are: The first pruning compares the weight W(Ck) of candidate k-itemsets with KIWT(k, k+1) [17], pruning candidate itemsets Ck where $W(Ck) < KIWT(k, k+1)$. The second pruning, applied when mining

2-itemsets, removes candidate 2-itemsets C2 that do not contain query terms, as the retrieval model only mines frequent itemsets and matrix-weighted association rules related to the original query, considering terms in candidate 2-itemsets without Chinese query terms as irrelevant to the original query. Deleting these at the candidate 2-itemset stage reduces the number of such irrelevant itemsets in subsequent stages and improves mining efficiency. The third pruning removes candidate itemsets Ck with support count equal to zero.

The above mining concept is formalized as the MWARM_{OQT} (Matrix Weighted Association Rule Mining with Original Query Terms) algorithm.

Input: Target language initially retrieved relevant document set (DocTL), minimum support and confidence thresholds [17] (ms, mc), Indonesian user query (QSL).
Output: Target language feature term matrix-weighted association rule set (mwARTL).

Begin
let mwFITL← ; mwARTL← ; //mwFITL is the feature term matrix-weighted frequent itemset collection, mwFITL and mwARTL
(DocTL_{DB})←Preprocessing(DocTL); //Document preprocessing module preprocesses DocTL to construct user feedback initially retrieved relevant document database DocTL_{DB}. In this model, DocTL consists of Chinese documents, and preprocessing includes word segmentation, stop word removal, and Chinese feature term extraction. The word segmentation system used in the model is the Chinese lexical analysis system ICTCLAS developed by the Institute of Computing Technology, Chinese Academy of Sciences .
(C1, w(C1), nc1, KIWT (1, 2))←ScanForC1(DocTL_{DB}); //Scan the initially retrieved relevant document database DocTL_{DB} to extract feature term 1-candidate itemsets C1, calculate support count nc1, weight w(C1), and KIWT (1, 2) values. The KIWT (1, 2) calculation formula is provided in [17].
L1←{C1| mwsupport(C1)≥ms}; //Mine 1-frequent itemsets from 1-candidate itemsets C1, where mwsupport(C1) is the matrix-weighted support of C1, mwsupport(C1)=w(C1) / nc1 [17].
for (k=2; Ck ≠ ; k++){ //Mine matrix-weighted frequent k-itemsets containing query terms (k$ 2)$mwFITL$←$mwFITL$\cup$Lk$-$1; //Addfrequentitemsetstom$wFITL$collectionCk$-$1←FirstPruning($w(Ck-1), KIWT(k-1,k)$); //Comparecandidateitemsetweightswith$KIWT$values, prune$d$1where$W(Ck-1) < KIWT(k-1,k)$.The$KIWT(k-1,k)$calculationformulaisprovidedin[17].Ck←$CJoin(Ck$-1); //PerformApriorijoin[18]oncandidateitemsets$Ck-1$toobtainCkif$(k=2)$thenCk←$SecondPruning(Ck, QSL)$; //Whenmining$2-itemsets$, prunecandidate$2-itemsetswithoutqueryterms(w(Ck), nck, KIWT(k,k+1))$←$ScanForCk(DocTL_{DB})$; //Scantheinitiallyretrie$d$1)values.The$KIWT(k,k+1)$calculationformulaisprovidedin[17].Ck←$ThirdPruning(Ck);//Prune candidate itemsets Ck with support count equal to zero;

3. Experimental Design and Results Analysis

Based on the theoretical analysis and model architecture described above, source code was implemented for the Indonesian-Chinese cross-language information retrieval model based on vector space model and matrix-weighted association pattern mining. The experimental hardware environment consisted of an Intel(R) Core(TM) i7-3770 CPU @3.4GHz desktop computer with 8.0GB memory and 1TB hard disk. The software environment was Windows 7+VC#+SQL Server.

3.1 Dataset and Preprocessing

The experiment utilized the Chinese news text from the Economic Daily News 2000 corpus in the NTCIR-5 CLIR standard test dataset from the International Evaluation Conference on Multilingual Information Processing organized by the National Institute of Informatics, Japan, comprising 79,380 Chinese text documents. NTCIR-5 CLIR includes query sets, document test sets, and result sets. The query set contains 50 query topics with four types: TITLE, DESC, NARR, and CONC. This experiment selected TITLE and DESC types. TITLE-type query topics provide brief descriptions using nouns and noun phrases, representing short queries, while DESC-type provides brief descriptions in sentence form, representing long queries. The result sets include two evaluation standards: Rigid and Relax. The Rigid standard includes only documents highly relevant or relevant to the original query, while the Relax standard includes highly relevant, relevant, or partially relevant documents.

To conduct experiments on the proposed Indonesian-Chinese cross-language information retrieval model, professional translators from translation agencies were invited to manually translate the 50 Chinese query topics from NTCIR-5 CLIR into Indonesian, followed by further processing.

3.2 Baseline Experiments and Evaluation Metrics

To validate the effectiveness of the proposed Indonesian-Chinese cross-language information retrieval model, three baselines were selected for performance comparison and analysis: Chinese monolingual retrieval baseline (Monolingual Retrieval Baseline, MRB), Indonesian-Chinese cross-language retrieval without query expansion (Cross-language Retrieval Baseline, CLRB), and the traditional pseudo-relevance feedback-based Indonesian-Chinese cross-language information retrieval algorithm [2] (Cross-Language Retrieval Using Pseudo Relevance Feedback, CLR_{PRF}).

The retrieval results for these three baselines are as follows: MRB baseline results were obtained by directly retrieving Chinese documents with Chinese queries; CLRB baseline represents traditional cross-language retrieval results obtained by translating Indonesian queries into Chinese queries via machine translation and retrieving Chinese documents; CLR_{PRF} baseline results were obtained by implementing cross-language query expansion under the following parameter settings (consistent with [2]): extracting the top 20 initially

retrieved documents to construct the initially retrieved relevant document set and selecting the top 20 feature terms by weight (in descending order) as expansion terms.

The evaluation metrics employed were R-precision (R_{prec}), P@10, and P@20. R-precision is the precision calculated after R documents have been retrieved, where R refers to the number of relevant documents for a given query in the document collection, without emphasizing document ranking in the result set. Since the number of relevant documents varies significantly across different query topics in the NTCIR-5 CLIR test set, this metric is particularly meaningful and valuable for evaluation.

3.3 Experimental Results and Analysis

The source code of the proposed model was executed to conduct Indonesian-Chinese cross-language retrieval experiments using the 50 Indonesian query topics from NTCIR-5 CLIR (TITLE and DESC sections). Retrieval performance was compared and analyzed with the three baselines (MRB, CLRB, and CLR_{PRF}) under varying support and confidence threshold conditions. The R_{prec}, p@10, and p@20 values of the retrieval results are shown in Table 2 and Table 3 . The experimental parameters for the proposed model were set as follows: the top 100 initially retrieved cross-language documents were presented to users, who determined relevant documents through clicking, browsing, and downloading behaviors. For experimental convenience, the top 100 documents containing known result sets in the initial retrieval were treated as user feedback-relevant document information obtained through clicking, browsing, and downloading. Additionally, the mined itemset length was set to 3. For support variation experiments, confidence was fixed at mc=0.01 while support ms varied among 0.5, 0.55, 0.6, 0.65, 0.7, and 0.75, with average values reported in Table 2. For confidence variation experiments, support was fixed at ms=0.5 while confidence mc varied among 0.008, 0.01, 0.05, 0.08, and 0.1, with results shown in Table 3.

(1) Baseline Experimental Results and Analysis To compare with the proposed retrieval model' s performance, the three baseline source programs (MRB, CLRB, CLR_{PRF}) were first executed. Chinese queries from the TITLE and DESC sections of NTCIR-5 CLIR were submitted for the Chinese monolingual retrieval baseline experiment, while Indonesian queries were used for Indonesian-Chinese cross-language retrieval and traditional pseudo-relevance feedback-based Indonesian-Chinese cross-language retrieval baseline experiments. The baseline experimental results are presented in Table 1 .

Table 1 shows that the traditional cross-language retrieval CLRB baseline achieved only 32.32% to 75.14% of the monolingual retrieval baseline MRB performance. After query translation, severely affected by translation quality, query topic drift was substantial, resulting in fewer relevant documents retrieved and more non-relevant documents. The traditional pseudo-relevance feedback-

based Indonesian-Chinese cross-language information retrieval CLR_{PRF} performed even worse, achieving only 15.59% to 58.73% of the monolingual baseline MRB. Compared with the CLRB baseline, most evaluation metrics of CLR_{PRF} decreased, with a maximum reduction of 70.62% (R_{prec} value for DESC-type queries under Relax evaluation type). Only a few metrics increased, with the maximum improvement being the p@20 metric for TITLE-type queries under Rigid evaluation type, reaching 43.84%.

These baseline results indicate that Indonesian-Chinese cross-language retrieval (traditional cross-language retrieval) performance is significantly lower than monolingual baseline retrieval performance, with some metrics reaching as low as 15.59%. This demonstrates that in traditional cross-language information retrieval, Indonesian queries translated into Chinese via machine translation suffer from severe query topic drift. Performing pseudo-relevance feedback query expansion under such severe topic drift conditions further degrades retrieval performance, making CLR_{PRF} inferior to CLRB.

(2) Performance Comparison Between Proposed Model and Baselines The proposed model source code was executed to conduct Indonesian-Chinese cross-language retrieval experiments using Indonesian queries from the TITLE and DESC sections of NTCIR-5 CLIR. Performance was compared with the three baselines (MRB, CLRB, and CLR_{PRF}) under varying support and confidence conditions, with R_{prec}, p@10, and p@20 values shown in Table 2 and Table 3 .

Table 2 results indicate that under support variation, the proposed model' s evaluation metrics range from 59.72% (minimum) to 124.32% (maximum) of the monolingual retrieval baseline MRB, showing improvements of 41.19% (minimum) to 97.79% (maximum) over the cross-language baseline CLRB, and 30.19% (minimum) to 573.16% (maximum) over the pseudo-relevance feedback baseline CLR_{PRF}, demonstrating significant effectiveness. Additionally, Table 2 shows that long query type DESC achieves better retrieval effectiveness than short query type TITLE. For DESC-type queries, the proposed model' s R_{prec} value under Rigid type exceeds monolingual retrieval by 24.32% (i.e., (0.2321-0.1867)/0.1867).

Table 3 results demonstrate that under confidence threshold variation, the proposed model's metrics range from 60.72% to 126.78% of the monolingual baseline MRB, with the best case being a 14.25% improvement over monolingual retrieval for long query type DESC (Rigid type R_{prec} value: (0.2133-0.1867)/0.1867). Compared with CLRB, the proposed model improves metrics by 37.08% to 90.44%, and compared with CLR_{PRF} by 28.51% to 548.25%, showing significant improvements. Again, long query type DESC outperforms short query type TITLE.

(3) Impact of Support and Confidence on Model Retrieval Performance The retrieval performance of the proposed Indonesian-Chinese cross-

language retrieval model under different matrix-weighted support thresholds ms and confidence thresholds mc is shown in Table 4 (with matrix-weighted confidence mc=0.01) and Table 5 (with matrix-weighted support ms=0.5).

Tables 4 and 5 reveal that for both TITLE and DESC query types, as matrix-weighted support or confidence thresholds continuously increase, the proposed model' s R_{prec}, p@10, and p@20 values change slowly, with some showing a downward trend. The primary reason is that under severe query topic drift conditions, as support or confidence thresholds increase, fewer expansion terms are obtained from matrix-weighted inter-term association rules, leading to degraded cross-language retrieval performance. Conversely, when support or confidence thresholds decrease, the retrieval system obtains more expansion terms, improving cross-language retrieval performance. However, as expansion terms increase, the chance of false expansion terms (noise) also increases, which can degrade retrieval performance. Therefore, determining appropriate support and confidence thresholds is a worthwhile research question.

(4) Experimental Results Analysis Theoretical analysis and experimental results demonstrate that compared with the monolingual retrieval baseline MRB, traditional cross-language retrieval baseline CLRB, and traditional pseudo-relevance feedback-based cross-language query algorithm CLR_{PRF}, the proposed Indonesian-Chinese cross-language retrieval model effectively reduces query topic drift and achieves substantial improvements in retrieval performance. Tables 2 and 3 show that the model's R_{prec}, p@10, and p@20 values all exceed 60% of the monolingual retrieval baseline MRB, with the best case showing a 24.32% improvement in R_{prec} over monolingual retrieval. Particularly, the model' s retrieval results outperform both CLRB and CLR_{PRF} baselines, with maximum improvements reaching 548.25%. These results confirm that the proposed Indonesian-Chinese cross-language information retrieval model is effective in improving cross-language retrieval performance.

The main reasons for these improvements are as follows: In cross-language information retrieval, query translation results significantly impact retrieval outcomes, often causing severe query topic drift that makes initial cross-language retrieval results inferior to monolingual results. Integrating user browsing, clicking, and downloading behaviors with matrix-weighted association pattern mining and query expansion technologies into the Indonesian-Chinese cross-language information retrieval model enables acquisition of the most relevant feedback information related to the original query. Mining matrix-weighted association rules to obtain query-related expansion terms for cross-language query expansion substantially reduces severe topic drift in cross-language retrieval and improves Indonesian-Chinese cross-language retrieval performance.

Meanwhile, matrix-weighted support and confidence thresholds affect the retrieval performance of the proposed Indonesian-Chinese cross-language information retrieval model. Excessively high support or confidence thresholds may omit some query-relevant expansion terms, reducing cross-language query ex-

pansion performance. Conversely, excessively low thresholds may introduce or increase query-irrelevant expansion terms, potentially causing new query topic drift in severe cases. Therefore, determining appropriate support and confidence thresholds is a worthwhile research topic.

As exchanges between China and ASEAN countries deepen across various fields, research on cross-language information retrieval and services for ASEAN languages has become urgent and important. This study focuses on Indonesian and Chinese, integrating user click behaviors with matrix-weighted association pattern mining into an Indonesian-Chinese cross-language information retrieval model. The experimental results demonstrate that the proposed model effectively reduces query topic drift, addresses the long-standing severe topic drift problem in cross-language information retrieval, and improves Indonesian-Chinese cross-language retrieval performance, with particularly better results for long queries.

Due to the broad scope of search engine research and numerous factors to consider, this study' s experiments were conducted in a vector space model-based cross-language retrieval system as simulation experiments. Future research will focus on: (1) Practical implementation of the retrieval model to develop a practical Indonesian-Chinese cross-language information retrieval system in a real search engine environment, and (2) In-depth investigation of the impact of matrix-weighted association pattern mining parameters on Indonesian-Chinese cross-language retrieval performance to identify variation patterns for application in practical systems.

Acknowledgments: The authors thank the anonymous reviewers and editorial board for their revision suggestions.

References

[1] Gao J F, Nie J Y, Zhang J, et al. TREC-9 CLIR Experiments at MSRCN [C]//Proceedings of the 9th Text Retrieval Evaluation Conference. 2001.
[2] Wu Dan, He Daqing, 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.
[3] Wu Dan, He Daqing, Wang Huilin. A Relevance Feedback Based Query Translation Enhancement Technique in Cross Language Information Retrieval [J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(4): 398-406.
[4] 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. Association for Computational Linguistics, 2010: 1346-1356.
[5] Parton K, Gao J. Combining Signals for Cross-Lingual Relevance Feedback [C]//Proceedings of the 8th Asia Information Retrieval Societies Conference (AIRS 2012), Tianjin, China. Springer Berlin Heidelberg. 2012.

[6] Lee C J, Croft W B. Cross-Language Pseudo-Relevance Feedback Techniques for Informal Text [C]//Proceedings of the 36th European Conference on IR Research (ECIR 2014), Amsterdam, The Netherlands. Springer International Publishing, 2014.
[7] Bi Jianting, Su Yidan. Expansion Method for Language-crossed Query Based on Latent Semantic Analysis [J]. Computer Engineering, 2009, 35(10): 49-50.
[8] Wei Lu, Li Shuqin, Li Weinan, et al. Optimization of Cross-language Query Expansion [J]. Computer Engineering and Design, 2014, 35(8): 2785-2803.
[9] Ning Jian, Lin Hongfei. Cross-Language Information Retrieval Based on Improved Latent Semantic Indexing [J]. Journal of Chinese Information Processing, 2010, 24(3): 105-111.
[10] Luo Yuansheng, Wang Mingwen, Le Zhongjian, et al. Bilingual Topic Correlation Model in Cross-lingual Information Retrieval [J]. Journal of Chinese Computer Systems, 2013, 34(12): 2758-2763.
[11] Rahimi R, Shakery A, King I. Multilingual Information Retrieval within the Language Modeling Framework [J]. Information Retrieval Journal, 2015, 18(3): 246-281.
[12] Ganguly D, Leveling J, Jones G J F. Cross-lingual Topical Relevance Models [C]//Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012).
[13] Wang X W, Zhang Q, Wang X J, et al. LDA Based PSEUDO Relevance Feedback for Cross Language Information Retrieval [C]//Proceedings of the 2nd International Conference on Cloud Computing and Intelligence Systems. IEEE, 2012.
[14] Wang X W, Wang X J, Zhang Q, et al. A Web-Based CLIR System with Cross-Lingual Topical Pseudo Relevance Feedback [C]//Proceedings of the 4th International Conference and Labs of the Evaluation Forum (CLEF) Initiative, Valencia, Spain. 2013.
[15] Wang Xuwen, Wang Xiaojie, Sun Yueping. Cross-lingual Pseudo Relevance Feedback Based on Bilingual Topics [J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(4): 81-84.
[16] Wang X W, Zhang Q, Wang X J, et al. Cross-lingual Pseudo Relevance Feedback Based on Weak Relevant Topic Alignment [C]//Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China. 2015: 529-534.
[17] Huang Mingxuan, Yan Xiaowei, Zhang Shichao. Query Expansion of Pseudo Relevance Feedback Based on Matrix-Weighted Association Rules Mining [J]. Journal of Software, 2009, 20(7): 1854-1865.
[18] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Database[C]//Proceedings of 1993 ACM SIGMOD International Conference on Management of Data. 1993.
[19] Salton G, Buckley C. Term-weighting Approaches in Automatic Text Retrieval [J]. Information Processing & Management, 1988, 24(5): 513-523.

Conflict of Interest Statement: The authors declare no conflict of interest.

Supporting Data: Supporting data is available in the journal' s online version at http://www.infotech.ac.cn.

Received Date: 2016-09-18
Revised Date: 2016-11-09

Note: Figure translations are in progress. See original paper for figures.