ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201711.01996

A Matrix-Weighted Association Pattern-Based Model for Indonesian-Chinese Cross-Language Information Retrieval (Postprint)

Authors: Huang Mingxuan

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

Abstract

【Purpose】 To address the query drift problem in cross-language information retrieval, we propose an Indonesian-Chinese cross-language information retrieval model that integrates user click and download behavior with matrix-weighted association pattern mining. 【Method】 Matrix-weighted association pattern mining, query expansion, and user click and download behavior are integrated into the Indonesian-Chinese cross-language information retrieval model, and the key technologies for model implementation are presented, namely 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 show that the R_{prec}, p@10, and p@20 values of the retrieval model all achieve more than 60% of the monolingual retrieval baseline, representing an improvement of over 37% compared to the cross-language retrieval baseline and over 28% compared to existing cross-language retrieval algorithms based on pseudo-relevance feedback. 【Limitations】 The model experiments were conducted in a cross-language retrieval system based on the vector space model, and its specific applications in actual search engines need to be explored and studied. 【Conclusion】 The model can effectively reduce the query drift problem in cross-language retrieval, improve and enhance Indonesian-Chinese cross-language retrieval performance, achieves better retrieval effectiveness for long queries, and has good practical application value.

ChinaRxiv 中X

Full Text

Preamble

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

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

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 behavior with matrix-weighted association pattern mining. [**Methods**] The model integrates matrix-weighted association pattern mining, query expansion, and user click-download behavior 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 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 in a vector space model-based cross-language retrieval system, requiring further investigation for practical implementation in real search engines. [**Conclusions**] The proposed model effectively reduces query drift in cross-language retrieval, improves Indonesian-Chinese cross-language retrieval performance, demonstrates better effectiveness for long queries, and holds significant 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 extensive research on cross-language information retrieval models and algorithms from various perspectives, yielding rich theoretical results. However, problems in cross-language information retrieval remain unresolved, with one of the most critical 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, 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. The typical algorithm is the two-step pseudo-relevance feedback method proposed by Gao et al. [1]. Wu et al. [2] conducted in-depth research on pseudo-relevance feedback-based cross-language query expansion, comparing four cross-language information retrieval query translation optimization techniques through pseudo-relevance feedback experiments and achieving promising results. More recently, Chinnakotla et al. [4] proposed using auxiliary language materials different from the query to improve cross-language pseudo-relevance feedback expansion performance and enhance retrieval efficiency. Parton et al. [5] introduced machine learning into the cross-language relevance feedback expansion domain, while Lee et al. [6] proposed a new pseudo-relevance feedback expansion technique for informal texts such as blogs and forums to improve cross-language retrieval performance, both achieving favorable experimental results.

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 implement cross-language query expansion and improve retrieval performance. The typical algorithm is the cross-language query expansion method based on latent semantic analysis proposed by Bi Jianting et al. [7]. Wei et al. [8] improved upon [7] by combining singular value decomposition and non-negative matrix factorization to construct a bilingual space, enhancing cross-language retrieval performance. Subsequently, Ning et al. [9] achieved bilingual abstract cross-language retrieval through improved latent semantic analysis, obtaining good experimental results. Luo et al. [10] constructed latent semantic spaces for each language using bilingual parallel corpora to improve cross-language retrieval performance, with experimental results demonstrating the effectiveness of these methods.

Research on cross-language information retrieval based on language models and topic models has also become active. Rahimi et al. [11] implemented cross-

language query expansion within the language modeling framework, improving retrieval performance. Ganguly et al. [12] utilized latent topics to improve cross-language relevance models, helping enhance target language retrieval effectiveness. Thereafter, Wang et al. [13-16] conducted in-depth research on topic model-based cross-language information retrieval, successively proposing LDA-based cross-language pseudo-relevance feedback expansion [13-14], bilingual topic-based cross-language pseudo-relevance feedback [15], and weakly relevant topic alignment-based cross-language pseudo-relevance feedback expansion [16], with theoretical analysis and experimental results all confirming the effectiveness of these approaches.

Review of relevant literature reveals that 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 behavior, 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 the key technologies for its implementation.

2. 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. Through users' browsing, clicking, and downloading behaviors on initial retrieval documents, a document is confirmed as a user feedback-relevant document. Then, the matrix-weighted association pattern mining technique proposed in this paper is applied to mine Chinese query-related expansion terms from the 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, the architecture of the Indonesian-Chinese cross-language information retrieval model based on matrix-weighted association pattern mining is presented in Figure 1 [Figure 1: see original paper]. The model consists of 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, namely the Microsoft Translator API, primarily to translate user-submitted Indonesian queries into Chinese queries and to translate the final retrieved Chinese documents into Indonesian documents for user presentation.

(2) **Search Engine Module**: Can employ search engines such as Google or Baidu, primarily functioning 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 the 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, and constructing the 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 the expansion term base.

(7) **Cross-Language Query Expansion Implementation Module**: Ex-

tracts Chinese expansion terms from the expansion term base, combines them with the original query to form a new query, and resubmits it to the search engine for retrieval on the Internet to obtain 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 final retrieved Chinese documents and Indonesian documents to users.

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, obtain Indonesian-Chinese cross-language initial retrieval results, i.e., the user feedback target language document set $DocTL$, through the user click behavior relevance feedback information extraction module. The document preprocessing module then preprocesses $DocTL$ to construct the user feedback initially retrieved relevant document database. Subsequently, combined with the user query and employing a three-stage itemset pruning strategy, the system mines matrix-weighted feature term association rules containing user query terms from the initially retrieved relevant document database to construct the matrix-weighted association rule base. The specific pruning strategies are: The first pruning compares the candidate k-itemset weight $W(C_k)$ with $KIWT(k, k+1)$ [17], pruning candidate itemsets C_k where $W(C_k) < KIWT(k, k+1)$. The second pruning, performed when mining 2-itemsets, removes candidate 2-itemsets C_2 that do not contain query terms, primarily because this 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 to be 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 C_k with support count equal to 0.

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 (Q_{SL}).

Output: Target language feature term matrix-weighted association rule set $(mwARTL)$.

Begin

```
let mwFITL← ; mwARTL← ; //mwFITL is the feature term matrix-weighted frequent itemset collect
```

```
(DocTL_DB)←Preprocessing(DocTL); //Document preprocessing module preprocesses DocTL to const

(C1, w(C1), nc1, KIWT(1, 2))←ScanForC1(DocTL_DB); //Scan the initially retrieved relevant do

L1←{C1 | mwsupport(C1) ms}; //Mine 1-frequent itemsets from 1-candidate itemset C1, where mw

for (k=2; Ck ; k++){ //Mine matrix-weighted frequent k-itemsets containing query terms (k 2)
    mwFITL←mwFITL  Lk-1; //Add frequent itemsets to mwFITL collection
    Ck-1←FirstPruning(w(Ck-1), KIWT(k-1, k)); //Compare candidate itemset weight with KIWT v

    Ck←CJoin(Ck-1); //Perform Apriori join [18] on candidate itemsets Ck-1 to obtain Ck

    if (k=2) then Ck←SecondPruning(Ck, QSL); //When mining 2-itemsets, prune candidate 2-ite

    (w(Ck), nck, KIWT(k, k+1))←ScanForCk(DocTL_DB); //Scan initially retrieved relevant docu

    Ck←ThirdPruning(Ck); //Prune candidate itemsets Ck with support count 0

    Lk←{Ck | mwsupport(Ck) ms}; //Mine k-frequent itemsets from k-candidate itemset Ck, wher
}

for each frequent itemset ITL in mwFITL do //Mine feature term matrix-weighted association r
    for each pair of sub-itemsets I1 and I2 in ITL do
        if ((I1  I2=ITL) and (I1  I2= )) then
            Calculate mwconf(I1→I2) and mwconf(I2→I1) values;
            //mwconf(I1→I2) and mwconf(I2→I1) are association rule confidences
            if mwconf(I1→I2) mc then mwARTL←mwARTL  {I1→I2};
            if mwconf(I2→I1) mc then mwARTL←mwARTL  {I2→I1};

output(mwARTL); //Output matrix-weighted strong association rules containing query terms
```

End

The confidence calculation formula for association rules [17] is as follows:

$$mwconf(I_1 \rightarrow I_2) = \frac{mwsupport(I_1, I_2)}{mwsupport(I_1)} \quad (1)$$

$$mwconf(I_2 \rightarrow I_1) = \frac{mwsupport(I_1, I_2)}{mwsupport(I_2)} \quad (2)$$

(2) Indonesian-Chinese Cross-Language Query Expansion Model in the Retrieval Model In this retrieval model, the source of post-translation query expansion terms is the matrix-weighted association rules mined by the MWARM_OQT algorithm from the initially retrieved user-relevant document set. The antecedent of these rules is the translated target language original query

term set (QTL), while the consequent is the target language expansion term set ($ETTL$). The degree of association between query terms and expansion terms is determined by the matrix-weighted association rule confidence $mwconf$ value. Therefore, the cross-language query expansion model (CLQEM) is described by formula (3):

$$QTL = \{q_1, q_2, \dots, q_n\}, q_n(n \geq 1) \text{ are query terms}$$

$$ETTL = \{t_1, t_2, \dots, t_m\}, t_m(m \geq 1) \text{ are expansion terms}$$

$$QTL \rightarrow ETTL(mwsupport \geq ms, mwconf \geq mc)$$

$$W_{ET} = \max(mwconf)$$

In the above expansion model, W_q represents the weight of query term q in the translated original query QTL, tf_q is the initial frequency of query term q in the query, $\max(tf_q)$ denotes the highest initial frequency among all query terms, df_q is the number of initially retrieved documents containing query term q, and N is the total number of initially retrieved relevant documents. W_{ET} represents the weight of target language query expansion terms derived from matrix association rules $QTL \rightarrow ETTL$, with its value equal to the confidence value of the matrix association rule. The W_{ET} expression indicates that when expansion terms appear repeatedly in different matrix association rules with different confidence values, the highest confidence value is taken as the weight for that expansion term.

(3) Indonesian-Chinese Cross-Language Information Retrieval Algorithm Based on Matrix-Weighted Association Pattern Mining In this cross-language retrieval model, the fundamental concept of Indonesian-Chinese cross-language information retrieval based on matrix-weighted association pattern mining is as follows: Employ a two-stage retrieval strategy. First, translate the Indonesian query into Chinese via a machine translation system and submit it to a search engine to retrieve Chinese documents on the Internet. Obtain the cross-language user feedback initially retrieved relevant document set through user click-download behavior. Invoke the MWARM_OQT algorithm to mine the user feedback initially retrieved relevant document set to obtain matrix-weighted association rules related to the original query. Extract expansion terms from the association rules to achieve post-translation cross-language query expansion. Combine the expansion terms with the original query to form a new query and resubmit it to the search engine to retrieve Chinese documents. Translate the final retrieval results into Indonesian documents via the machine translation system and return them to users. This concept is formalized as the

ICCLIR_MWAR (Indonesian-Chinese Cross Language Information Retrieval Based on Matrix-Weighted Association Rules) algorithm.

Input: Indonesian user query (Q_{SL}), minimum support and confidence thresholds (ms, mc).

Output: Cross-language retrieval results after query expansion (Indonesian documents and Chinese documents).

Begin

```
QTL←ExecMTranslate(QSL); //Submit Indonesian user query QSL (source language query) to machi

FirstRDoc←FirstRetrieval(QTL, Wq); //Submit translated Chinese query to search engine (e.g.,

DocTL←UserClickDownload(FirstRDoc); //Based on user click, browse, and download behaviors or

mwARTL←MWARM_OQT(DocTL, ms, mc, QTL); //Invoke MWARM_OQT algorithm to mine target language f

(ETTL, WET)←GetExpTerm(mwARTL); //Extract target language expansion terms ETTL from mwARTL c

TL_Doc←SecondRetrieval(QTL, ETTL); //Combine original query with expansion terms and retriev

SL_Doc←ExecMTranslate(TL_Doc); //Translate target language documents TL_Doc (Chinese documen

outputToUser(TL_Doc, SL_Doc); //Return post-query-expansion retrieval results (Chinese docum
```

End

3. Experimental Design and Results Analysis

Based on the theoretical analysis and model architecture diagram presented above, source code for the Indonesian-Chinese cross-language information retrieval model based on vector space model and matrix-weighted association pattern mining was implemented for experiments. The hardware environment for experiments was: Intel(R) Core(TM) i7-3770 CPU @3.4GHz 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 Chinese news texts from the Economic Daily News 2000 corpus in the NTCIR-5 CLIR standard test set from the National Institute of Informatics (NII) Cross-Lingual Information Retrieval evaluation campaign were used as the experimental corpus, totaling 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 are briefly described with nouns and noun phrases, representing short queries;

DESC-type queries are briefly described in sentence form, representing long queries. The result sets have two evaluation standards: Rigid and Relax. The Rigid standard includes only answers highly relevant or relevant to the original query, while the Relax standard includes highly relevant, relevant, or partially relevant answers.

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.

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 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 are obtained by directly retrieving Chinese documents with Chinese queries; CLRB baseline results are obtained by translating Indonesian queries into Chinese queries via machine translation system to retrieve Chinese documents, representing traditional cross-language information retrieval results; CLR_PRF baseline results are obtained by implementing cross-language query expansion under the following parameter settings (consistent with literature [2]): extracting 20 top-ranked cross-language initially retrieved documents to construct the initially retrieved relevant document set, extracting 20 top-weighted feature terms (sorted in descending order) as expansion terms, with confidence threshold $mc = 0.01$ and support threshold $ms = 0.5$.

The experimental evaluation metrics adopted are 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. This metric does not emphasize 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, and Indonesian-Chinese cross-language retrieval experiments were conducted on the 50 query topics from NTCIR-5 CLIR (both TITLE and DESC portions). Retrieval performance was compared and analyzed with the three baselines (MRB, CLRB, and CLR_PRF)

under varying support and confidence threshold conditions. The experimental parameter settings for the proposed model were as follows: 100 top-ranked cross-language initially retrieved documents were presented to users, and documents were determined as initially retrieved relevant documents based on user click, browse, and download behaviors. For experimental convenience, 100 known relevant documents from the top-ranked initially retrieved documents were treated as user feedback relevant document information obtained through click and browse behaviors. 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 model's retrieval performance, baseline source codes for MRB, CLRB, and CLR_PRF were first executed. Chinese queries from the TITLE and DESC portions of NTCIR-5 CLIR's 50 query topics were submitted for Chinese monolingual retrieval baseline experiments, while Indonesian queries were submitted for Indonesian-Chinese cross-language retrieval and traditional pseudo-relevance feedback-based Indonesian-Chinese cross-language retrieval baseline experiments. The baseline experimental results are shown in Table 1 .

As shown in Table 1, traditional cross-language retrieval CLRB baseline achieved only 32.32% to 75.14% of 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. Traditional pseudo-relevance feedback-based Indonesian-Chinese cross-language information retrieval CLR_PRF performed even worse, achieving only 15.59% to 58.73% of monolingual baseline MRB performance. Compared with CLRB baseline, most evaluation metrics of CLR_PRF retrieval results decreased, with the maximum reduction reaching 70.62% (R_prec value for DESC-type queries under Relax evaluation). Only a few metrics increased, with the maximum improvement being the p@20 metric for TITLE-type queries under Rigid evaluation, reaching 43.84%.

The baseline experimental results in Table 1 demonstrate that Indonesian-Chinese cross-language baseline (traditional cross-language retrieval) performance is significantly lower than monolingual baseline performance, with some metrics reaching as low as 15.59%. This indicates that in traditional cross-language information retrieval, Indonesian queries translated into Chinese queries via machine translation suffer from severe query topic drift, with poor initial retrieval quality compared to monolingual retrieval. Under such severe topic drift conditions, conducting pseudo-relevance feedback query expansion for cross-language retrieval leads to even worse performance, making CLR_PRF inferior to CLRB.

(2) Retrieval Performance Comparison Between Proposed Model and Baselines The proposed model source code was executed, and Indonesian-Chinese cross-language retrieval experiments were conducted using Indonesian queries from the TITLE and DESC portions of NTCIR-5 CLIR. Retrieval performance was compared and analyzed 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 show that with support variation, the proposed model's retrieval metrics range from 59.72% (minimum) to 124.32% (maximum) of monolingual retrieval baseline MRB, representing improvements of 41.19% (minimum) to 97.79% (maximum) over cross-language baseline CLRB, and 30.19% (minimum) to 573.16% (maximum) over pseudo-relevance feedback baseline CLR_PRF, demonstrating significant effectiveness. Additionally, Table 2 indicates that long query type DESC achieves better retrieval effectiveness than short query type TITLE. For long query type DESC, the proposed model's R_prec value under Rigid evaluation exceeds monolingual retrieval by 24.32% (i.e., (0.2321–0.1867)/0.1867).

Table 3 results demonstrate that with confidence threshold variation, the proposed model's retrieval metrics range from 60.72% to 126.78% of monolingual retrieval baseline MRB, with the best case being a 14.25% improvement over monolingual retrieval for long query type DESC (R_prec value under Rigid evaluation: (0.2133–0.1867)/0.1867). Compared with cross-language baseline CLRB, the proposed model improves retrieval metrics by 37.08% to 90.44%, while improving over CLR_PRF baseline by 28.51% to 548.25%, showing significant effectiveness. Table 3 also indicates that long query type DESC achieves better retrieval effectiveness than 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 threshold ms and confidence threshold mc values 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 show 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 main reason is that under severe query topic drift conditions, as matrix-weighted support or confidence thresholds increase, fewer expansion terms are obtained from matrix-weighted inter-word 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 or confidence thresholds is a worthwhile research question.

(4) Experimental Results Analysis Theoretical analysis and experimental results demonstrate that compared with 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 problems, with 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 monolingual retrieval baseline MRB, with the best case showing a 24.32% improvement in R_prec over monolingual retrieval. Particularly, its retrieval results outperform cross-language baselines CLRB and CLR_PRF, with maximum improvement reaching 548.25%. These experimental results demonstrate that the proposed Indonesian-Chinese cross-language information retrieval model is effective and can improve cross-language information retrieval performance. The main reasons are analyzed as follows: In cross-language information retrieval, query translation results significantly impact retrieval outcomes, often causing severe query topic drift and poorer initial retrieval quality than monolingual retrieval. Integrating user browsing, clicking, and downloading behaviors, matrix-weighted association pattern mining, and query expansion techniques into the Indonesian-Chinese cross-language information retrieval model can obtain feedback information most relevant to the original query. Mining matrix-weighted association rules to obtain expansion terms related to the original query for cross-language query expansion can greatly reduce severe topic drift problems in cross-language retrieval and improve 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 matrix-weighted support or confidence may omit some expansion terms relevant to the original query, reducing cross-language query expansion performance. Conversely, excessively low thresholds may introduce or increase expansion terms irrelevant to the original query, 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 becomes increasingly urgent and important. This study focuses on Indonesian and Chinese, integrating user click behavior and matrix-weighted association pattern mining into an Indonesian-Chinese cross-language information retrieval model. Key implementation technologies are elaborated, and experimental results demonstrate that the proposed model is effective, 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 better effectiveness for long queries.

Due to the broad research scope of search engines and numerous factors to consider, this study's experimental work was conducted in a vector space model-

based cross-language retrieval system as a simulation experiment. Future research priorities include: practical implementation of the retrieval model, development of a practical Indonesian-Chinese cross-language information retrieval system in a search engine environment, and in-depth investigation of matrix-weighted association pattern mining parameters' impact on Indonesian-Chinese cross-language retrieval performance to identify variation patterns for practical system deployment.

Acknowledgments: We 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. (in Chinese)

[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. (in Chinese)

[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. (in Chinese)

[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. (in Chinese)

[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. (in Chinese)

[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. (in Chinese)

[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 on 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. (in Chinese)

[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. (in Chinese)

[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 online version of the journal 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.

Source: ChinaXiv — Machine translation. Verify with original.