Query Expansion by Word Embedding in the Suggestion Track of CLEF 2016 Social Book Search Lab

Shih-Hung Wu[1]*, Yi-Hsiang Hsieh[1], Liang-Pu Chen[2], Ping-Che Yang[2]

[1] Chaoyang University of Technology, Taiwan, R.O.C
{ shwu(*Contact author), s10427617,}@cyut.edu.tw
[2]Institute for Information Industry, Taiwan, R.O.C
{eit, maciaclark}@iii.org.tw

Abstract. The Social Book Search (SBS) Lab is part of CLEF 2016 lab series. This is the fourth time that the CYUT CSIE team attends the SBS track. The content of topics has changed a little bit by the organizer; therefore, we make necessary modification on our system, which is based on keyword searching and ranking by social features. This year, we design a query expansion module which is based on word2vec, a word embedding toolkit. The new module helps our system to get better performance in suggestion track.

Keywords: Query type recognition, social features, social book search, word embedding.

1 Introduction

The paper reports our system in the suggestion track of CLEF 2016 Social Book Search (SBS) lab [10]. This is the fourth time that we attend the SBS track since 2013 INEX [7]. Based on our social feature re-ranking system [1], we improve our system by adding a query expansion module which is based on word2vec [13], a word embedding toolkit.

We believe that the result of traditional information retrieval technology is not enough for the users who need more personal recommendation in the SBS task. Recommendation from other users are more appealing; it might contain more personal feelings and cover more subtle reasons that traditional information retrieval system cannot cover. Our system integrates the social feature into the traditional information retrieval technology to give better recommendation on books. In this task, user-generated metadata is used as the social feature.

According to our observation on the topics in the previous INEX SBS Track, we found that queries can be separated into different types. Simply treating the keywords in the topic as search terms will not get good results. Some queries require higher level of knowledge to deal with. The system needs to understand the information need behind the keyword, for example, the knowledge on the types of literature. We analyze and find several types in them. Due to the time limitation, we only implement a module to recognize one special type of topics and a filtering module to modify the recommendation result.

The structure of this paper is as follows. Section 2 is the data set description, section 3 shows our architecture and the details of our method, section 4 is the experiment results, and final section gives conclusions and future works.

2 Dataset

2.1 Collection

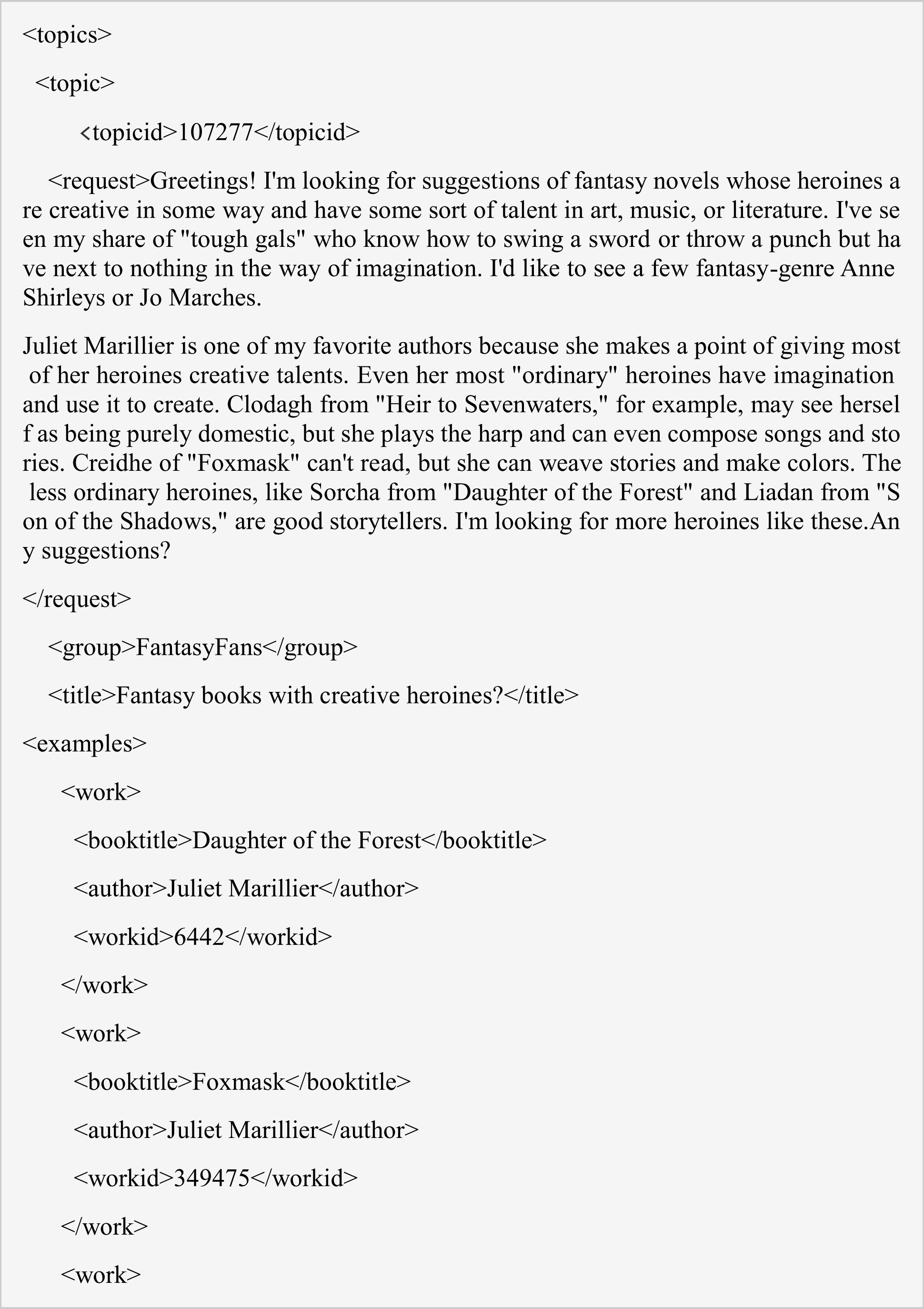
The document collection in this task is provided by the CLEF Social Book Search lab. The documents are the XML format metadata of about 2.8 million books and the data size is 25.9GB. These documents are collected from Amazon.com and LibraryThing [2]. Some of the data is also from the Library of Congress and the British Library. The XML tags used in the data set is listed in Table 1.

Table 1.All the XML tag [2]

tag name			
book	similarproducts	title	imagecategory
dimensions	Tags	edition	name
reviews	Isbn	dewey	role
editorialreviews	Ean	creator	blurber
images	Binding	review	dedication
creators	Label	rating	epigraph
blurbers	Listprice	authorid	firstwordsitem
dedications	manufacturer	totalvotes	lastwordsitem
epigraphs	numberofpages	helpfulvotes	quotation
firstwords	publisher	date	seriesitem
lastwords	Height	summary	award
quotations	Width	editorialreview	browseNode
series	Length	content	character
awards	Weight	source	place
browseNodes	readinglevel	image	subject
characters	releasedate	imageCategories	similarproduct
places	publicationdate	url	tag
subjects	Studio	data	

2.2 Test Topic

The topic format in 2016 is different from the topic format in 2014 and 2015. Figure 1 and Figure 2 show an example. The XML tags related to the query are <topicid>, <request>, <group>, and <title>. Addition XML tags shows the book list of the user: <booktitle>, <author>, and <workid>.



```
<topics>
  <topic>
        <topicid>107277</topicid>
    <request>Greetings! I'm looking for suggestions of fantasy novels whose heroines a
re creative in some way and have some sort of talent in art, music, or literature. I've se
en my share of "tough gals" who know how to swing a sword or throw a punch but ha
ve next to nothing in the way of imagination. I'd like to see a few fantasy-genre Anne
Shirleys or Jo Marches.
Juliet Marillier is one of my favorite authors because she makes a point of giving most
 of her heroines creative talents. Even her most "ordinary" heroines have imagination
and use it to create. Clodagh from "Heir to Sevenwaters," for example, may see hersel
f as being purely domestic, but she plays the harp and can even compose songs and sto
ries. Creidhe of "Foxmask" can't read, but she can weave stories and make colors. The
 less ordinary heroines, like Sorcha from "Daughter of the Forest" and Liadan from "S
on of the Shadows," are good storytellers. I'm looking for more heroines like these.An
y suggestions?
</request>
    <group>FantasyFans</group>
    <title>Fantasy books with creative heroines?</title>
<examples>
      <work>
        <booktitle>Daughter of the Forest</booktitle>
        <author>Juliet Marillier</author>
        <workid>6442</workid>
      </work>
      <work>
        <booktitle>Foxmask</booktitle>
        <author>Juliet Marillier</author>
        <workid>349475</workid>
      </work>
      <work>
```

Figure 1. A topic example in CLEF 2016 Social Book Suggestion track

```
      <booktitle>Son of the Shadows</booktitle>
      <author>Juliet Marillier</author>
      <workid>6471</workid>
    </work>
    <work>
      <booktitle>Heir to Sevenwaters</booktitle>
      <author>Juliet Marillier</author>
      <workid>5161003</workid>
    </work>
  </examples>
  <catalogue>
    <work>
      <tags/>
      <rating>0.0</rating>
      <publication-year>2002</publication-year>
      <booktitle>Blue Moon (Anita Blake, Vampire Hunter, Book 8)</booktitle>
      <cataloging-date>2011-08</cataloging-date>
      <author>Laurell K. Hamilton</author>
      <workid>10868</workid>
    </work>
            <catalogue>
</topic>
```

Figure 2. A topic example in CLEF 2016 Social Book Suggestion track (Continued)

3 CYUT CSIE System Methodology

3.1 System Architecture

Figure 3, 4, 5 shows the architecture of our system. The preprocessing module includes stop-word filtering and stemming. We use an open source search engine, Lucene, as our indexing and search module. Table 3 shows the index building. A keyword expansion module based on the word2vec tool is added into our system. Ranking is based on the social features. Table 4 shows how we train a word2vec model to help query expansion. Table 5 shows the overall architecture of our system.

3.2 Indexing and Query

The index and search engine in use is the Lucene system, which is an open source full text search engine provided by Apache software foundation. Lucene is written in JAVA and can be called easily by JAVA program to build various applications.

Table 1 shows all the tags of the book metadata. According to Bogers and Larsen [3], there are 19 tags more useful in the social book search. They are <isbn>, <title>, <publisher>, <editorial>, <creator>, <series>, <award>, <character>, <place>, <blurber>, <epigraph>, <firstwords>, <lastwords>, <quotation>, <dewey>, <subject>, <browseNode>, <review>, and <tag>. Our system also focuses on the same 19 tags.

In the pre-processing step, the content in the <dewey> tag is restored to strings according to the 2003 list of Dewey category descriptions [9] to make string matching easier. For example: <dewey>004</dewey> will be restored to <dewey>Data processing Computer science</dewey>. The content of <tag> is also expanded according to the count number to emphasize its importance. For example: <tag count="3">fantasy</tag> will be expanded as <tag>fantasy fantasy fantasy</tag>. In additional to the 19 tags, our system also indexes the content of <review> as independent indexes files and names it as reviews.

Fig.1 and 2 shows all the XML tags of the query topics. According to Koolen et al. [4], an Indri [5] based system using all the contents of <Title>, <Query>, <Group>, and <Narrative> as query terms will give better result. We also use the contents of the four fields as our system input queries.

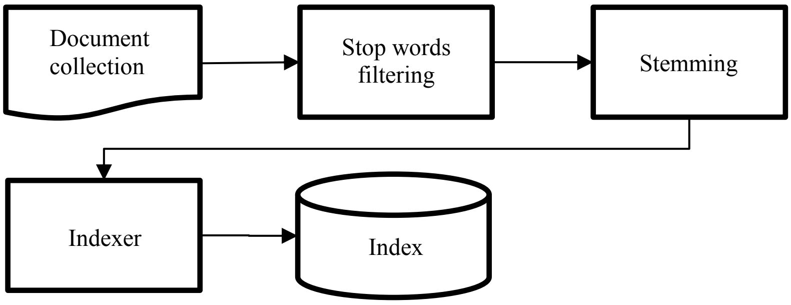


Figure 3. System architecture for index building

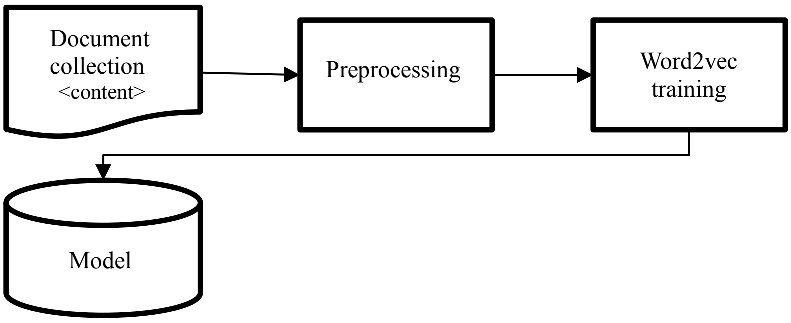


Figure 4. System architecture for query expansion

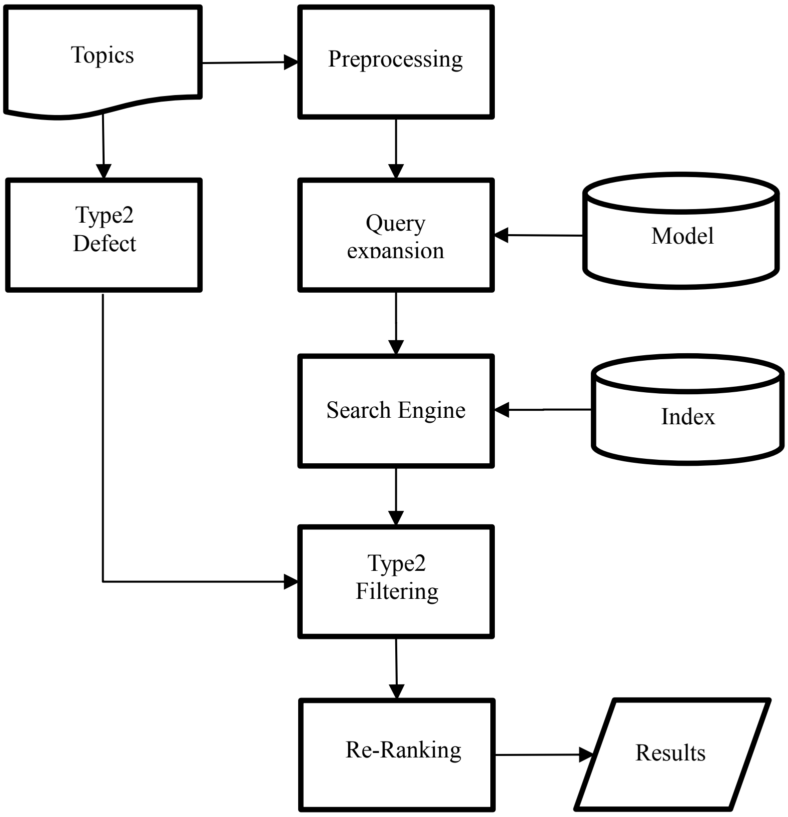
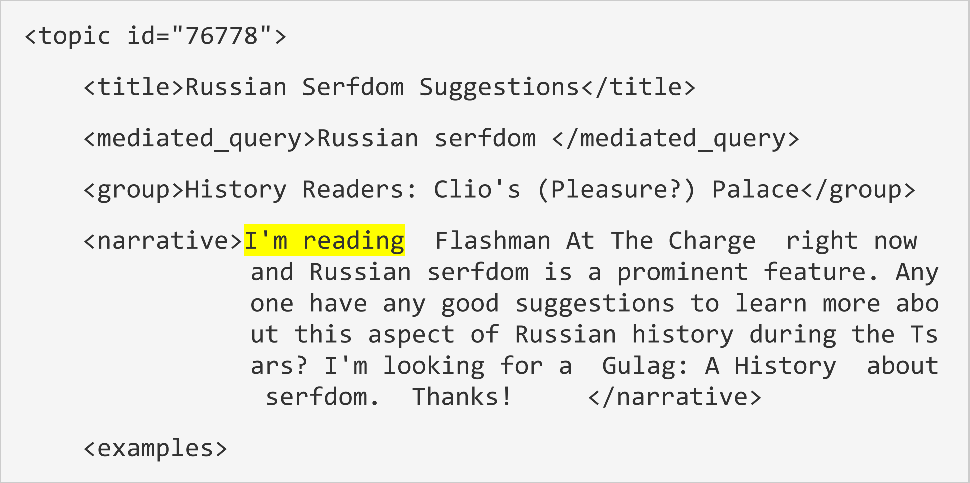


Figure 5. System architecture for topic processing

3.3 Word Embedding

Word embedding is based on an open source toolkit, word2vec, which is developed by Google in 2013 [13]. Word2vec is a neural network that trained on a given corpus and can transfer the representation of the words into a vector space. The new representation can be used to find words with similar context. The toolkit is used on various natural language processing applications, such as document clustering, similar word finding, sentiment analysis, and machine translation.

In this year, we use the word2vec as a way for keyword expansion. We extract the contents of the 2.8 million books as the training corpus. The word2vec toolkit is used to find words with similar context to the keywords that we extract from the topics. These words are our expanded keywords.



```
<topic id="76778">

    <title>Russian Serfdom Suggestions</title>

    <mediated_query>Russian serfdom </mediated_query>

    <group>History Readers: Clio's (Pleasure?) Palace</group>

    <narrative>I'm reading  Flashman At The Charge  right now
               and Russian serfdom is a prominent feature. Any
               one have any good suggestions to learn more abo
               ut this aspect of Russian history during the Ts
               ars? I'm looking for a  Gulag: A History  about
                serfdom.  Thanks!     </narrative>

    <examples>
```

Figure 4. A type2 query example that we defined in 2015 SBS track

3.4 Type2 Query Recognition and Result Filtering

According to our observation on the topics in INEX 2012 SBS Track, we find that there are some queries that are different from others, we call them the Type2 queries [11]. Type2 queries are the queries that contain the names of some books that the original users want to find similar ones. Therefore, the books in the topics should not be part of the recommendation. Since the book names are given explicitly, our system originally will find exactly the same books as the top recommendation. To recognize type2 queries, we define a list of phrases to identify such queries and filter out the books in the queries from the recommendation lists. The phrases are listed in the appendix in the previous paper [11]. Figure 4 gives an example of Type2 queries taken from INEX 2013 SBS topics, in which contains a key phrase “I’m reading”. We find that there are 174 queries in the INEX 2013 SBS track that can be classified as Type2 queries. Therefore, we add a module in our system to identify the Type2 queries and filtering out the books mentioned in the topics.

3.5 Re-ranking

The Re-ranking part is similar to that in our previous work [1]. We integrate the user-generated metadata into the traditional content-based search result by re-ranking the results. The social features are provided by the amazon users, and our system use them to give more weight on certain books. Three numbers are available:

- User rating: users might evaluate a book from 1 to 5, the higher the better.
- Helpful vote: other users might endorse one comment by voting it as helpful.
- Total vote: the total number of helpful or not.

We designed 3 different ways to use these social features in re-ranking.

1) User rating method

Increase the weight of content-based retrieval result by adding the summation of user rating. As shown in formula (1):

$$\mathrm{Score}_{\mathrm{re-ranked}}(\mathrm{i}) = \alpha * \mathrm{Score}_{\mathrm{org}}(\mathrm{i}) + (1-\alpha) * \mathrm{Score}_{\mathrm{user\ rating}}(\mathrm{i}) \tag{1}$$

2) Average User rating method

Increase the weight of content-based retrieval result by adding the average of user rating. As shown in formula (2):

$$\mathrm{Score}_{\mathrm{re-ranked}}(\mathrm{i}) = \mathrm{Score}_{\mathrm{org}}(\mathrm{i}) + \mathrm{Score}_{\mathrm{average\ user\ rating}}(\mathrm{i}) \tag{2}$$

3) Weights User rating method

Increase the weight of content-based retrieval result by adding the book which gets more helpful votes. As shown in formula (3) and (4):

$$\mathrm{Score}_{\mathrm{Weights\ User\ Rating}} = \mathrm{User\ rating}\ * \frac{\mathrm{helpfulvote}}{\mathrm{totalvote}} \tag{3}$$

$$\mathrm{Score}_{\mathrm{re-ranked}}(\mathrm{i}) = \alpha * \mathrm{Score}_{\mathrm{org}}(\mathrm{i}) + (1-\alpha) * \mathrm{Score}_{\mathrm{Weights\ User\ Rating}}(\mathrm{i}) \tag{4}$$

3.6 Find the Best α Value by Experiment

Since there is no theoretical reference on how to set the α value, in our official runs, the value is selected via a series experiments that we conduct on the 2013 dataset. Table 2 shows the results, we find that the system gets the best result when α is 0.95.

Table 2. Experimental Result for different α on 2013 data set

A	P@10	MAP
0.50	0.0221	0.0193
0.60	0.0221	0.0193
0.70	0.0224	0.0195
0.80	0.0226	0.0196
0.90	0.0237	0.0204
0.95	**0.0245**	**0.0220**

4 Experimental results

We sent our six runs in the formal run. Three different settings are as the ones used last year without query expansion; three corresponding settings with query expansion is proposed.
Run 1: CYUT - 0.95AverageType2TGR. This is the best setting in last year.
Run 2: CYUT - 0.95Averageword2vecType2TGR. In this run, our system incorporate with the keyword expansion module with the help of word2vec.
Run 3: CYUT - Type2TGR. Use type 2 filter to filter out some candidates.
Run 4: CYUT - word2vecType2TGR. After query expansion, use type 2 filter to filter out some candidates.
Run 5: CUYT - 0.95RatingType2TGR. Ranking query result by the rating in user review.
Run 6: CYUT - 0.95Ratingword2vecType2TGR. After query expansion, ranking query result by the rating in user review.

Table 3 shows the official evaluation results of our four runs. Among them the CSIE - 0.95AverageType2QTGN run gives the best NDCG@10 [8] result, while the CSIE - Type2QTGN run gives similar result on NDCG@10 but give better result on MAP and R@1000. Comparing to the 2013 INEX SBS results in Table 5 and 2015 SBS results in Table 4, our system performance improved significantly. However, comparing to the result of INEX SBS 2014 in Table 4, our system performance decreased.

Table 3. Official evaluation results in 2016 SBS

Run	*nDCG@10*	*MRR*	*MAP*	*R@1000*
CYUT - 0.95Averageword2vec-Type2TGR	**0.1158**	**0.2563**	**0.0563**	**0.1603**
CYUT- 0.95AverageType2TGR	0.1137	0.2718	0.0572	0.1626
CYUT - word2vecType2TGR	0.1107	0.2479	0.0542	0.1614
CYUT - Type2TGR	0.1060	0.2545	0.0550	0.1635
CYUT - 0.95RatingType2TGR	0.0392	0.1363	0.0145	0.1089
CYUT -0.95Ratingword2vec-Type2TGR	0.0373	0.1265	0.0136	0.1055

Table 4. Official evaluation results in 2015 SBS

Run	*nDCG@10*	*MRR*	*MAP*	*R@1000*	*Profiles*
CSIE - 0.95AverageType2QTGN	**0.082**	**0.194**	0.050	0.319	no
CSIE - Type2QTGN	0.080	0.191	**0.052**	**0.325**	no
CSIE - 0.95RatingType2QTGN	0.032	0.113	0.019	0.214	no
CSIE - 0.95WRType2QTGN	0.023	0.072	0.015	0.216	no

Table 5. Official evaluation results in 2014 INEX SBS

Run	*nDCG@10*	*MRR*	*MAP*	*R@1000*
CYUT - Type2QTGN	**0.119**	**0.246**	**0.086**	**0.340**

CYUT - 0.95AverageType2QTGN	**0.119**	0.243	0.085	0.332
CYUT - 0.95RatingType2QTGN	0.034	0.101	0.021	0.200
CYUT - 0.95WRType2QTGN	0.028	0.084	0.018	0.213

Table 6. Official evaluation results in 2013 INEX SBS

Run	*nDCG@10*	*P@10*	*MRR*	*MAP*
Run1.query.content-base	0.0265	0.0147	0.0418	0.0153
Run2.query.Rating	0.0376	0.0284	0.0792	0.0178
Run3.query.RA	0.0170	0.0087	0.0352	0.0107
Run4.query.RW	**0.0392**	**0.0287**	**0.0796**	**0.0201**
Run5.query.reviwes.content-base	0.0254	0.0153	0.0359	0.0137
Run6.query.reviews.RW	0.0378	0.0284	0.0772	0.0165

5 Conclusions and Future work

This paper reports our system and result in CLEF 2016 Social Book Suggestion track. We sent six runs and the formal run results are list in Table 3. In the six runs, the new proposed run, CSIE - 0.95Averageword2vecType2TGR, gives best nDCG@10, which is searching with content-based search engine with the help of a keyword expansion module based on word2vec, then applying a set of filtering rules based on a list of key phrase and re-ranking with Average User Rating. In the future, we will implement more modules with literature knowledge on the writers, genre of books, geometric categories of the publishers, and temporal categories of the authors that can deal with the special cases in the topics.

From last year, user profiles are available, which can be used to give better recommendation. A system might use the user profiles to expand the queries or to suggest more books that the user read before for other similar users. Outside resources might also be used to expand the queries. For example, a system might check Wikipedia to find more authors of the books in the same genre, and make better recommendation. Books that won some awards might also be a good list for recommendation.

Acknowledgement

“This study is conducted under the "Online and Offline integrated Smart Commerce Platform(3/4)" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

1. Wei-Lun Xiao, Shih-Hung Wu, Liang-Pu Chen, Hung-Sheng Chiu, and Ren-Dar Yang, *Social Feature Re-ranking in INEX 2013 Social Book Search Track*, CLEF 2013 Evaluation Labs and Workshop Online Working Notes, 23 - 26 September, Valencia, Spain.
2. Marijn Koolen, Gabriella Kazai, Jaap Kamps, Michael Preminger, Antoine Doucet, and Monica Landoni, *Overview of the INEX 2012 Social Book Search Track*, INEX'12 Workshop Pre-proceedings,P.77-P.96, 2012.
3. Toine Bogers and Birger Larsen, *RSLIS at INEX 2012: Social Book Search Track*, INEX'12 Workshop Pre-proceedings,P.97-P.108, 2012.
4. Marijn Koolen, Hugo Huurdeman and Jaap Kamps, *Comparing Topic Representations for Social Book Search*, CLEF 2013 Evaluation Labs and Workshop Online Working Notes, 23 - 26 September, Valencia – Spain.
5. T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, *Indri: a language-model based search engine for complex queries*, In Proceedings of the International Conference on Intelligent Analysis, 2005.
6. Lucene, https://lucene.apache.org
7. Marijn Koolen, Gabriella Kazai, Michael Preminger, and Antoine Doucet, *Overview of the INEX 2013 Social Book Search Track*, CLEF 2013 Evaluation Labs and Workshop Online Working Notes, 23 - 26 September, Valencia – Spain.
8. Järvelin, K., Kekäläinen, *J.: Cumulated Gain-based Evaluation of IR Techniques*, ACM Transactions on Information Systems 20(4) (2002) 422–446.
9. 2003 list of Dewey category descriptions, https://www.library.illinois.edu/ugl/about/dewey.html
10. CLEF 2015 Social Book Search Track, http://social-book-search.humanities.uva.nl/#/suggestion
11. Shih-Hung Wu, Pei-Kai Liao, Hua-Wei Lin, Li-Jen Hsu, Wei-Lun Xiao, Liang-Pu Chen, Tsun Ku, and Gwo-Dong Chen Query Type Recognition and Result Filtering in INEX 2014 Social Book Search Track
12. Marijn Koolen, Toine Bogers, Gabriella Kazai, Jaap Kamps, and MichaelPreminger Overview of the INEX 2014 Social Book Search Track
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: *Efficient estimation of word representations in vector space*. CoRR abs/1301.3781 (2013)