The CYUT System on Social Book Search Track since INEX 2013 to CLEF 2016

Shih-Hung Wu
Assistant Professor
Department of Computer Science and Information Technology
Chaoyang University of Technology, Taiwan (R.O.C.)
E-mail: shwu@cyut.edu.tw

Keywords: Query Type Recognition; Social Book Search; Social Features; Word Embedding Query Expansion

Abstract

The Social Book Search (SBS) Lab is part of Conference and Labs of the Evaluation Forum (CLEF) lab series, which provides query topics on book suggestion in natural language. Participant teams have to build systems that can make suggestions out of 2.8 million books. This paper reports how the CYUT team attends the SBS track from 2013 to 2016. Our system is based on keyword searching and ranking by social features. We also 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. This paper reports the progression of our system.

Introduction

The paper reports our system in the suggestion track of CLEF Social Book Search (SBS) [1] lab (Koolen et al., 2016). Our team attends the SBS track four times since 2013 INEX (Koolen, Kazai, Preminger, & Doucet, 2013), 2014 INEX (Koolen, Bogers, Kazai, Kamps & Preminger, 2014), CLEF 2015(Koolen, Bogers, & Kamps, 2016). SBS provides topics from internet forum which are posted by readers who need book recommendations.

We believe that the search result of traditional information retrieval (IR) technology is not enough for the users who need more personal recommendation in the SBS task. Since the traditional IR system takes the input keywords as search terms and finds only documents contain the search terms. Recommendation from other users are much appealing; it might involve more
personal feelings and cover more subtle reasons that traditional IR system cannot cover. Systems should provide documents that might not contain the search terms but satisfy the users’ information need.

Therefore, we try to integrate more social features into the traditional information retrieval technology to give better recommendation on books. In our system, we integrate some information retrieval techniques such as user-generated metadata as the social features. In addition to social features, 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 (Xiao, 2014). Due to the resource limitation, we only implement a module to recognize one special type of topics and a filtering module to modify the recommendation result. Based on our social feature re-ranking system (Xiao, Wu, Chen, Chiu, & Yang, 2013), we further improve our system by adding a query expansion module which is based on word2vec (Mikolov, Chen, Corrado, & Dean, 2013), a word embedding toolkit.

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.

**Dataset**

**Collection**

The document collection in this task is provided by the CLEF Social Book Search lab. The documents are the XML format metadata from about 2.8 million books, and the data size is 25.9GB. These documents are collected from Amazon.com [2] and LibraryThing [3] which is an online service that users can catalog their books. With data in LibraryThing, users might get suggestions for what to read next. Some of the data are also from the Library of Congress and the British Library. The XML tags used in the data set are listed in Table 1.
Table 1  All the XML tags in SBS dataset

<table>
<thead>
<tr>
<th>tag name</th>
<th>tag name</th>
</tr>
</thead>
<tbody>
<tr>
<td>All the XML tags</td>
<td>in SBS dataset</td>
</tr>
<tr>
<td>tag name</td>
<td>tag name</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Book similarproducts</td>
<td>Title imagecategory</td>
</tr>
<tr>
<td>dimensions Tags</td>
<td>Edition name</td>
</tr>
<tr>
<td>reviews Isbn</td>
<td>Dewey role</td>
</tr>
<tr>
<td>editorialreviews</td>
<td>Ean blurbler</td>
</tr>
<tr>
<td>images Binding</td>
<td>Review dedication</td>
</tr>
<tr>
<td>creators Label</td>
<td>Rating epigraph</td>
</tr>
<tr>
<td>blurbers Listprice</td>
<td>authorid firstwordsitem</td>
</tr>
<tr>
<td>dedications manufacturer totalvotes lastwordsitem</td>
<td></td>
</tr>
<tr>
<td>epigraphs numberofpages helpfulvotes quotation</td>
<td></td>
</tr>
<tr>
<td>firstwords publisher Date seriesitem</td>
<td></td>
</tr>
<tr>
<td>lastwords Height summary award</td>
<td></td>
</tr>
<tr>
<td>quotations Width editorialreview browseNode</td>
<td></td>
</tr>
<tr>
<td>series Length content character</td>
<td></td>
</tr>
<tr>
<td>awards Weight Source place</td>
<td></td>
</tr>
<tr>
<td>browsenodes readinglevel Image subject</td>
<td></td>
</tr>
<tr>
<td>characters releasedate imageCategories similarp PRODUCT</td>
<td></td>
</tr>
<tr>
<td>places publicationdate url tag</td>
<td></td>
</tr>
<tr>
<td>subjects Studio data</td>
<td></td>
</tr>
</tbody>
</table>

Test Topic

Topics are queries collected by the organizers from internet forum, where readers post what they want to read and request recommendations from other users. Figure 1 shows an example, one reader posted a request in natural language. The XML tags related to the query are <topicid>, <request>, <group>, and <title>. Addition XML tags show the book list that the reader has read: <booktitle>, <author>, and <workid>. These books are examples of suggestion that the reader wants.

<topics>
<topic>
	<topicid>107277</topicid>
	:request>Greetings! I'm looking for suggestions of fantasy novels whose heroines are creative in some way and have some sort of talent in art, music, or literature. I've seen my share of "tough gals" who know how to swing a sword or throw a punch but have 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 herself as being purely domestic, but she plays the harp and can even compose songs and stories. 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 "Son of the Shadows," are good storytellers. I'm looking for more heroines like these. Any suggestions?

</request></topic>
</topics>
<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>
      <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>
      <booktitle>Blue Moon (Anita Blake, Vampire Hunter, Book 8)</booktitle>
      <author>Laurell K. Hamilton</author>
      <workid>10868</workid>
    </work>
  </catalogue>
</topic>

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

System Architecture

Figures 2, 3, and 4 show the architecture of our system. The preprocessing module includes stop-word filtering and stemming. We use an open source search engine, Lucene [4], as our indexing and search module. Figure 2 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. Figure 3 shows how we train a word2vec model to help query expansion. Figure 4 shows the overall architecture of our system.

Indexing and Query

The index and search engine in use are worked under 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 (2012), there are 19 tags 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 the original terms according to the 2003 list of Dewey category descriptions [5] to make string matching easier. For example: <dewey>004</dewey> will be restored as <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 index files and names it as reviews.

Not all the XML tags of the query topics in Figure 1 are used. According to a previous work, an Indri based system (Strohman, Metzler, Turtle, & Croft, 2005), using all the contents of <Title>, <Query>, <Group>, and <Narrative> as query terms will give better result (Koolen, Huurdeman, & Kamps, 2013). We also use all the terms in the four fields as our system input queries.
Figure 2  System architecture for index building

Figure 3  System architecture for query expansion

Figure 4  System architecture for topic processing
Word Embedding for Keyword Expansion

Word embedding is based on an open source toolkit, word2vec, which is developed by Google in 2013 (Mikolov et al., 2013). Word2vec is a neural network that is trained on a given corpus and can transfer the representation of a word into a vector. 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 in similar context for the keywords that we extract from the topics. These words are our expanded keywords. For example, a query term “City” is expanded with “times” and “new”. This may due to the fact that in the training corpus, there are many phrases like “New York City” and “New York times”, therefore, “times” and “new” are related terms according to the word2vec. Another example, “Book” is expanded with “encyclopedia” and “poetry”, with similar reason. Since the related terms were extracted from training corpus, the query expansion results by our method will be very different from the results by traditional dictionary-based method.

Type2 Query Recognition and Result Filtering

According to our observation on the topics in INEX 2012 SBS Track (Koolen, Kazai, Kamps, Preminger, Doucet, & Landoni, 2012), we find that there are some queries that are different from others (Xiao, 2014). Type1 queries are queries with specific information, users who posted the queries provided precise information on desired books, such as keywords of books or authors. Type2 queries are queries with examples, users provided read books and wanted related books. Type3 queries are queries with imprecise information on books, such as the books that gives certain feelings.

Type2 queries usually contain the names of some books that the original users who posted the queries want to find similar books, not the ones posted. Therefore, the books in the topics should not be part of the recommendation. Since the book names are given explicitly, our system originally found 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 (Xiao, 2014). Figure 5 gives an example of Type2 queries taken from SBS topics, 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 filter out the books mentioned in the topics.

<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. Anyone have any good suggestions to learn more about this aspect of Russian history during the Tsars? I'm looking for a Gulag: A History about serfdom. Thanks! </narrative>
<examples>

Figure 5  A type2 query example that we defined in 2015 SBS track

Re-ranking

The Re-ranking part is similar to that in our previous work. 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 uses them to give more weight on certain books. Three variables 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 the helpful or not.

We design 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):

\[
Score_{\text{re-ranked}}(i) = \alpha \times Score_{\text{org}}(i) + (1 - \alpha) \times Score_{\text{user rating}}(i)
\]  (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):

\[
Score_{\text{re-ranked}}(i) = Score_{\text{org}}(i) + Score_{\text{average user rating}}(i)
\]  (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):

\[
\text{Score}_{\text{Weights User Rating}} = \text{User rating} \times \frac{\text{helpful vote}}{\text{total vote}}
\]  
(3)

\[
\text{Score}_{\text{re-ranked}}(i) = \alpha \times \text{Score}_{\text{org}}(i) + (1 - \alpha) \times \text{Score}_{\text{Weights User Rating}}(i)
\]  
(4)

Find the Best \( \alpha \) Value by Experiment

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

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>P@10</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.0221</td>
<td>0.0193</td>
</tr>
<tr>
<td>0.60</td>
<td>0.0221</td>
<td>0.0193</td>
</tr>
<tr>
<td>0.70</td>
<td>0.0224</td>
<td>0.0195</td>
</tr>
<tr>
<td>0.80</td>
<td>0.0226</td>
<td>0.0196</td>
</tr>
<tr>
<td>0.90</td>
<td>0.0237</td>
<td>0.0204</td>
</tr>
<tr>
<td>0.95</td>
<td><strong>0.0245</strong></td>
<td><strong>0.0220</strong></td>
</tr>
</tbody>
</table>

Experimental Results

We sent our six runs in the 2016 formal run. Three different settings are the ones used in year 2015 without query expansion; three corresponding settings with query expansion are proposed.

Run 1: CYUT - 0.95AverageType2TGR. This is the best setting in year 2015.

Run 2: CYUT-0.95Averageword2vecType2TGR. In this run, our system, with the help of word2vec, incorporate with keyword expansion module.

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. Rank query result by rating of users’ reviews.

Run 6: CYUT - 0.95Ratingword2vecType2TGR. After query expansion, ranking query result by rating in users’ reviews.
Table 3 shows the official evaluation results of our six runs. Among them the CYUT - 0.95AverageWord2vecType2TGR run gives the best NDCG@10 (Järvelin & Kekäläinen, 2002) result, while the CYUT - Type2QTGN run gives a similar result on NDCG@10 but a better result on MAP and R@1000. This result shows that word2vec can help overall performance.

Since the test items changed every year, the difficulty of the task also changed. The performance between two different years will be different even for the same system. Here we list that some ideas which work in each year. Table 4, 5, 6 show the SBS official evaluation results in 2015, 2014, and 2013 respectively. The first three settings in 2014 were also used in year 2015 and 2016. Table 6 shows the official evaluation results in 2013 SBS. Run4.query.RW gave the best result, the setting is searching the index file built from 19 tags with content-based search and re-ranking with the Weights User Rating by equation (4). The second best run is Run2.query.Rating. The setting is searching the index file built from 19 tags with content-based search and re-ranking with the User Rating by equation (1).

Table 3  Official evaluation results in 2016 SBS

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@10</th>
<th>MRR</th>
<th>MAP</th>
<th>R@1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYUT - 0.95AverageWord2vecType2TGR</td>
<td><strong>0.1158</strong></td>
<td>0.2563</td>
<td>0.0563</td>
<td>0.1603</td>
</tr>
<tr>
<td>CYUT - 0.95AverageType2TGR</td>
<td>0.1137</td>
<td><strong>0.2718</strong></td>
<td><strong>0.0572</strong></td>
<td><strong>0.1626</strong></td>
</tr>
<tr>
<td>CYUT - word2vecType2TGR</td>
<td>0.1107</td>
<td>0.2479</td>
<td>0.0542</td>
<td>0.1614</td>
</tr>
<tr>
<td>CYUT - Type2TGR</td>
<td>0.1060</td>
<td>0.2545</td>
<td>0.0550</td>
<td>0.1635</td>
</tr>
<tr>
<td>CYUT - 0.95RatingType2TGR</td>
<td>0.0392</td>
<td>0.1363</td>
<td>0.0145</td>
<td>0.1089</td>
</tr>
<tr>
<td>CYUT - 0.95RatingWord2vecType2TGR</td>
<td>0.0373</td>
<td>0.1265</td>
<td>0.0136</td>
<td>0.1055</td>
</tr>
</tbody>
</table>

Table 4  Official evaluation results in 2015 SBS

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@10</th>
<th>MRR</th>
<th>MAP</th>
<th>R@1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSIE - 0.95AverageType2QTGN</td>
<td><strong>0.082</strong></td>
<td><strong>0.194</strong></td>
<td>0.050</td>
<td>0.319</td>
</tr>
<tr>
<td>CSIE - Type2QTGN</td>
<td>0.080</td>
<td>0.191</td>
<td><strong>0.052</strong></td>
<td><strong>0.325</strong></td>
</tr>
<tr>
<td>CSIE - 0.95RatingType2QTGN</td>
<td>0.032</td>
<td>0.113</td>
<td>0.019</td>
<td>0.214</td>
</tr>
<tr>
<td>CSIE - 0.95WRType2QTGN</td>
<td>0.023</td>
<td>0.072</td>
<td>0.015</td>
<td>0.216</td>
</tr>
</tbody>
</table>

Table 5  Official evaluation results in 2014 INEX SBS

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@10</th>
<th>MRR</th>
<th>MAP</th>
<th>R@1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYUT - Type2QTGN</td>
<td><strong>0.119</strong></td>
<td>0.246</td>
<td><strong>0.086</strong></td>
<td><strong>0.340</strong></td>
</tr>
<tr>
<td>CYUT - 0.95AverageType2QTGN</td>
<td><strong>0.119</strong></td>
<td>0.243</td>
<td>0.085</td>
<td>0.332</td>
</tr>
<tr>
<td>CYUT - 0.95RatingType2QTGN</td>
<td>0.034</td>
<td>0.101</td>
<td>0.021</td>
<td>0.200</td>
</tr>
<tr>
<td>CYUT - 0.95WRType2QTGN</td>
<td>0.028</td>
<td>0.084</td>
<td>0.018</td>
<td>0.213</td>
</tr>
</tbody>
</table>
Table 6  Official evaluation results in 2013 INEX SBS

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG@10</th>
<th>MRR</th>
<th>MAP</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run1.query.content-base</td>
<td>0.0265</td>
<td>0.0418</td>
<td>0.0153</td>
<td>0.0147</td>
</tr>
<tr>
<td>Run2.query.Rating</td>
<td>0.0376</td>
<td>0.0792</td>
<td>0.0178</td>
<td>0.0284</td>
</tr>
<tr>
<td>Run3.query.RA</td>
<td>0.0170</td>
<td>0.0352</td>
<td>0.0107</td>
<td>0.0087</td>
</tr>
<tr>
<td>Run4.query.RW</td>
<td><strong>0.0392</strong></td>
<td><strong>0.0796</strong></td>
<td><strong>0.0201</strong></td>
<td><strong>0.0287</strong></td>
</tr>
<tr>
<td>Run5.query.reviews.content-base</td>
<td>0.0254</td>
<td>0.0359</td>
<td>0.0137</td>
<td>0.0153</td>
</tr>
<tr>
<td>Run6.query.reviews.RW</td>
<td>0.0378</td>
<td>0.0772</td>
<td>0.0165</td>
<td>0.0284</td>
</tr>
</tbody>
</table>

Performance Indexes

The performance indexes used in this task are nDCG@10, P@10, MRR, and MAP. These indexes are used widely in the field of Information Retrieval (Manning, Raghavan, & Schutze, 2008). For details, please see the overview reports of the organizers (Koolen, Bogers, & Kamps, 2016).

Related Work

Each year, there are many teams participated the lab. The general approaches to SBS are similar, and each system has some variations on detailed system integration. Here we briefly introduce some related works. In year 2016, (Feng et al., 2016) gave the best run, which was made by a searching-re-ranking process. The system retrieves the top 1000 results by a Language Model based search engine with default parameters. Then the system re-ranks results by a combination of several strategies (number of people who read the book from profile, similar-product from amazon.com, popularity from LT forum, etc.). In year 2015, MIIB team gave the best run. Their queries were generated from all topic fields. The system used the famous BM25 retrieval model with all textual document fields merged into a single field. A Learning-to-rank framework is also applied. Results are re-ranked based on tags and ratings (Imhof, Badache, & Boughanem, 2015). In 2014, USTB team gave the best run, which used all topic fields combined against an index containing all available document fields. The results were re-ranked with 12 different re-ranking strategies, which were then combined adaptively using learning-to-rank (Zhang et al., 2014).

Conclusions and Future Work

This paper reports our system and results in SBS since INEX 2013 to CLEF 2016 Social Book Suggestion track and share the experience on improving a system gradually each year. The 2016 formal run results of our system are listed in Table 3. In the six runs, the new proposed run, CYUT - 0.95Averageword2vecType2TGR, gives the best nDCG@10, which is searching through
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 phrases 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 special cases in the topics. We used word2vec to expand query terms as new queries. In the future, more resources might also be used to expand the queries. For example, a system might check Wikipedia to find more authors in the same genre and make better recommendation. Books that won some awards might also be a good list for recommendation.

Acknowledgements

The CYUT team includes students from Chaoyang University of Technology: Wei-Lun Xiao, Pei-Kai Liao, Hua-Wei Lin, Li-Jen Hsu, Yi-Hsiang Hsieh, and the support from Institute for Information Industry. They support all the works in the four years.

Notes

[2] Amazon.com
[5] https://www.library.illinois.edu/ugl/about/dewey.html

References


Xiao, Wei-Lun. (2014). A social book search system that integrating the social features and content information, Master Thesis, CYUT.


CYUT 團隊參加社群書籍搜尋國際公開評測系統從 INEX2013 到 CLEF2016

吳世弘
朝陽科技大學資訊工程系助理教授
E-mail: shwu@cyut.edu.tw

關鍵詞：查詢類型識別；社群圖書推薦；社群特徵；詞嵌入查詢擴張

【摘要】
本論文報告朝陽科技大學參加 INEX 2013, 2014 以及後續歐洲資訊檢索論壇 CLEF 2015～2016 舉辦的 SBS 開發經驗。這個評測的目標是由系統從 280 萬本書中自動推薦相關圖書給使用者。主辦單位蒐集了許多使用者在社群媒體上對書籍推薦的需求，形成一個測試集。所有參加公開評測的系統回傳推薦圖書之後，由主辦單位確認是否是恰當的推薦，藉此評估系統的優劣。我們的系統整合了搜尋、排序、語意理解等技術，逐年改良了系統的效能。