Information Retrieval and Graph Analysis Approaches for Book Recommendation

Chahinez Benkoussas, Patrice Bellot

Hindawi Publishing Corporation
The Scientific World Journal
Volume 2015, Article ID 926418, 8 pages
http://dx.doi.org/10.1155/2015/926418

Research Article

Information Retrieval and Graph Analysis Approaches for Book Recommendation

Chahinez Benkoussas[1,2] and Patrice Bellot[1,2]

[1]*Aix-Marseille Université, CNRS, LSIS UMR 7296, 13397 Marseille, France*
[2]*Aix-Marseille Université, CNRS, CLEO OpenEdition UMS 3287, 13451 Marseille, France*

Correspondence should be addressed to Chahinez Benkoussas; chahinez.benkoussas@lsis.org

Received 15 May 2015; Accepted 24 August 2015

Academic Editor: Mariofanna G. Milanova

A combination of multiple information retrieval approaches is proposed for the purpose of book recommendation. In this paper, book recommendation is based on complex user's query. We used different theoretical retrieval models: probabilistic as InL2 (Divergence from Randomness model) and language model and tested their interpolated combination. Graph analysis algorithms such as PageRank have been successful in Web environments. We consider the application of this algorithm in a new retrieval approach to related document network comprised of social links. We called Directed Graph of Documents (DGD) a network constructed with documents and social information provided from each one of them. Specifically, this work tackles the problem of book recommendation in the context of INEX (Initiative for the Evaluation of XML retrieval) Social Book Search track. A series of reranking experiments demonstrate that combining retrieval models yields significant improvements in terms of standard ranked retrieval metrics. These results extend the applicability of link analysis algorithms to different environments.

1. Introduction

In recent years, document retrieval and recommendation have become more and more popular in many Web 2.0 applications where user can request documents. There has been much work done both in the industry and academia on developing new approaches to improve the performance of retrieval and recommendation systems over the last decade. The interest in this area still remains high to help users to deal with information overload and provide recommendation or retrieval content (books, restaurants, movies, academic publications, etc.). Moreover, some of the vendors have incorporated recommendation capabilities into their commerce services, for example, Amazon in book recommendation.

Existing document retrieval approaches need to be improved to satisfy user's information needs. Most systems use classic information retrieval models, such as language models or probabilistic models. Language models have been applied with a high degree of success in information retrieval applications [1–3]. This was first introduced by Ponte and Croft in [4]. They proposed a method to score documents, called *query likelihood*. It consists of two steps: estimate a language model for each document and then rank documents according to the likelihood scores resulting from the estimated language model. Markov Random Field model was proposed by Metzler and Croft in [5]; it considers query term proximity in documents by estimating term dependencies in the context of language modeling approach. From the existing probabilistic models, InL2 a Divergence from Randomness-based model was proposed by Amati and Van Rijsbergen in [6]. It measures the global informativeness of the term in the document collection. It is based on the idea that "*the more the term occurrences diverge from random throughout the collection, the more informative the term is.*" The limit of such models is that the distance between query terms in documents is not considered.

In this paper, we present an approach that combines probabilistic and language models to improve the retrieval performances and show that the two models combined act much better in the context of book recommendation.

In recent years, an important innovation in information retrieval appeared which consists of algorithms developed

to exploit the relationships between documents. One of the important algorithm is Google's PageRank [7]. It has been successful in Web environments, where the relationships are provided with the existing hyperlinks into documents. We present a new approach for document retrieval based on graph analysis and exploit the PageRank algorithm for ranking documents with respect to a user's query. In the absence of manually created hyperlinks, we use social information to create the Directed Graph of Documents (DGD) and argue that it can be treated in the same manner as hyperlink graphs. Experiments show that incorporating graph analysis algorithms in document retrieval improves the performances in term of the standard ranked retrieval metrics.

Our work focuses on search in the book recommendation domain, in the context of INEX Social Book Search track. The document collection contains Amazon/LibraryThing book descriptions and the queries, called topics, are extracted from the LibraryThing discussion forums.

In the rest of the paper, we presented a summary of related work in document retrieval and recommender systems. Then, we describe briefly the used retrieval models and show the combination method. In Section 5, we illustrate the graph modeling method followed by the different experiments and results.

2. Related Work

This work is first related to the area of document retrieval models, more specially language models and probabilistic models. The unigram language models are the most used for ad hoc information retrieval work; several researchers explored the use of language modeling that captures higher order dependencies between terms. Bouchard and Nie in [9] have showed significant improvements of retrieval effectiveness with a new statistical language model for the query based on three different ways: completing the query by terms in the user's domain of interest, reordering the retrieval results, or expanding the query using lexical relations extracted from the user's domain of interest.

Divergence from Randomness (DFR) is one of several probabilistic models that we have used in our work. Abolhassani and Fuhr have investigated several possibilities for applying Amati's DFR model [6] for content-only search in XML documents [10].

This work also relates to the category of graph based document retrieval. There has been increasing use of techniques based on graphs constructed by implicit relationships between documents. Kurland and Lee performed structural reranking based on centrality measures in graph of documents which has been generated using relationships between documents based on language models [11]. In [12], Lin demonstrates the possibility to exploit document networks defined by automatically generated content-similarity links for document retrieval in the absence of explicit hyperlinks. He integrates the PageRank scores with standard retrieval score and shows a significant improvement in ranked retrieval performance. His work was focused on search in the biomedical domain, in the context of PubMed search engine.

3. INEX Social Book Search Track and Test Collection

SBS task (http://social-book-search.humanities.uva.nl/) aims to evaluate the value of professional and user's metadata for book search on the Web. The main goal is to exploit search techniques to deal with complex information needs and complex information sources that include user profiles, personal catalogs, and book descriptions.

The SBS task builds on a collection of 2.8 million book description crawled by the University of Duisburg-Essen from Amazon (http://www.amazon.com/) [13] and enriched with content from LibraryThing (http://www.librarything.com/). Each book is identified by an ISBN and is an XML file. It contains content information like title information, Dewey Decimal Classification (DDC) code (for 61% of the books), category, Amazon product description, and so forth. Amazon records contain also social information generated by users like tags, reviews, and ratings. For each book, Amazon suggests a set of "Similar Products" which represents a result of computed similarity based on content information and user behavior (purchases, likes, reviews, etc.) [14].

SBS task provides a set of queries called topics each year where users describe what they are looking for (books for a particular genre, books of particular authors, similar books to those that have been already read, etc.). These requests for recommendations are natural expressions of information needs for a large collection of online book records. The topics are crawled from LibraryThing discussion forums.

The topic set consists of 680 topics and 208 topics in 2014 and 2015, respectively. Each topic has a narrative description of the information needs. The topic set of 2015 is a subset of that of 2014. Each topic consists of a set of fields. In this contribution we use *title*, *mediated query* (query description), and *narrative* fields. An example of topic is illustrated in Figure 1.

4. Retrieval Models

This section presents brief description and combination method of retrieval models used for book recommendation.

4.1. InL2 of Divergence from Randomness. We used InL2, Inverse Document Frequency model with Laplace after-effect and Normalization 2. This model has been used with success in different works [15–18]. InL2 is a DFR-based model (Divergence from Randomness) based on the Geometric distribution and Laplace law of succession. The DFR models are based on this idea: "The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d" [19]. For this model, the relevance score of a document D for a query Q is given by

$$\text{score}(Q, D) = \sum_{t \in Q} \text{qtw} \cdot \frac{1}{\text{tfn} + 1} \left(\text{tfn} \cdot \log \frac{N + 1}{n_t + 0.5} \right), \quad (1)$$

where qtw is the query term weight given by $\text{qtf}/\text{qtf}_{\max}$; qtf is the query term frequency; $\text{qtf}_{\max}$ is the maximum query

```
<topic id="1116">
    <title>Which LISP?</title>
    <mediated_query>introduction book to Lisp</mediated_query>
    <group>Purely Programmers</group>
    <narrative>It'll be time for me to shake things up and learn a new
         language soon. I had started on Erlang a while back and getting
         back to it might be fun. But I'm starting to lean toward Lisp--
         probably Common Lisp rather than Scheme. Anyone care to
         a good first Lisp book? Would I be crazy to hope that there's
         one out there with an emphasis on using Lisp in a web
         and/or system administration context? Not that I'm unhappy with
         PHP and Perl, but the best way for me to find the time to learn
         new language is to use it for my work... </narrative>
</topic>
```

Figure 1: An example of topic from SBS topics set composed with multiple fields to describe user's need.

term frequency among the query terms. N is the number of documents in the whole collection and n_t is the number of documents containing t. tfn corresponds to the weighted sum of normalized term frequencies tf_f for each field f, known as *Normalization 2* and given by

$$\mathrm{tfn} = \mathrm{tf} \cdot \log\left(1 + c \cdot \frac{\mathrm{avg\_}l}{l}\right) \quad (c > 0), \tag{2}$$

where tf is the frequency of term t in the document D; l is the length of the document in tokens and avg_l is the average length of all documents; c is a hyperparameter that controls the normalization applied to the term frequency with respect to the document length.

4.2. Sequential Dependence Model of Markov Random Field. Language models are largely used in document retrieval search for book recommendation [16, 20]. Metzler and Croft's Markov Random Field (MRF) model [21, 22] integrates multiword phrases in the query. Specifically, we used the Sequential Dependence Model (SDM), which is a special case of MRF. In this model cooccurrence of query terms is taken into consideration. SDM builds upon this idea by considering combinations of query terms with proximity constraints which are single term features (standard unigram language model features, f_T), exact phrase features (words appearing in sequence, f_O), and unordered window features (require words to be close together, but not necessarily in an exact sequence order, f_U).

Finally, documents are ranked according to the following scoring function:

$$\begin{aligned}\mathrm{SDM}(Q, D) = {} & \lambda_T \sum_{q \in Q} f_T(q, D) \\ & + \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i, q_i + 1, D) \\ & + \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i, q_i + 1, D),\end{aligned} \tag{3}$$

where feature weights are set based on the author's recommendation ($\lambda_T = 0.85$, $\lambda_O = 0.1$, $\lambda_U = 0.05$) in [20]. f_T, f_O, and f_U are the log maximum likelihood estimates of query terms in document D as shown in Table 1, computed over the target collection using a Dirichlet smoothing. We applied this model to the queries using Indri (http://www.lemurproject.org/indri/) Query Language (http://www.lemurproject.org/lemur/IndriQueryLanguage.php).

4.3. Combining Search Systems. The use of different retrieval systems retrieves different sets of documents. Combining the output of many search systems, in contrast to using just a single retrieval technique, can improve the retrieval effectiveness as shown by Belkin et al. in [23] where the authors have combined the results of probabilistic and vector space models. In our work, we combined the results of InL2 model and SDM model. The retrieval models use different weighting schemes; therefore we should normalize the scores. We used the maximum and minimum scores according to Lee's formula [24] as follows:

$$\mathrm{normalizedScore} = \frac{\mathrm{oldScore} - \mathrm{minScore}}{\mathrm{maxScore} - \mathrm{minScore}}. \tag{4}$$

It has been shown in [16] that InL2 and SDM models have different levels of retrieval effectiveness; thus it is necessary to weight individual model scores depending on their overall performance. We used an interpolation parameter (α) that we have varied to get the best interpolation that provides better retrieval effectiveness.

5. Graph Modeling

We have studied the INEX SBS collection to link documents. In [25], the authors have used PubMed collection and exploited networks defined by automatically generated content-similarity links for document retrieval. In our case, we exploited a special type of similarity based on several factors. This similarity is provided by Amazon and corresponds to "Similar Products" given generally for each visited book.

Table 1: Language modeling-based unigram and term weighting functions. Here, $tf_{e,D}$ is the number of times term e matches in document D, $cf_{e,D}$ is the number of times term e matches in the entire collection, $|D|$ is the length of document D, and $|C|$ is the size of the collection. Finaly, μ is a weighting function hyperparameter that is set to 2500.

Weighting	Description
$f_T(q_i, D) = \log\left[\frac{tf_{q_i,D} + \mu cf_{q_i}/\|C\|}{\|D\| + \mu}\right]$	Weight of unigram q_i in document D.
$f_O(q_i, q_{i+1}, D) = \log\left[\frac{tf_{\#1(q_i,q_{i+1}),D} + \mu cf_{\#1(q_i,q_{i+1})}/\|C\|}{\|D\| + \mu}\right]$	Weight of exact phrase "$q_i \ \ q_{i+1}$" in document D.
$f_O(q_i, q_{i+1}, D) = \log\left[\frac{tf_{\#uw8(q_i,q_{i+1}),D} + \mu cf_{\#uw8(q_i,q_{i+1})}/\|C\|}{\|D\| + \mu}\right]$	Weight of unordered window "$q_i \ \ q_{i+1}$" (span = 8) in document D.

(1) $D_{\text{init}} \leftarrow$ Retrieving Documents for each $t_i \in T$
(2) **for** each $D_{t_i} \in D_{\text{init}}$ **do**
(3) $\quad D_{\text{StartingNodes}} \leftarrow$ first β documents $\in D_{t_i}$
(4) $\quad$ **for** each StartingNode **in** $D_{\text{StartingNodes}}$ **do**
(5) $\quad\quad D_{\text{graph}} \leftarrow D_{\text{graph}}$ + neighbors(StartingNode, DGD)
(6) $\quad\quad D_{\text{SPnodes}} \leftarrow$ **all** $D \in$ ShortestPath(StartingNode, $D_{\text{StartingNodes}}$, DGD)
(7) $\quad\quad D_{\text{graph}} \leftarrow D_{\text{graph}} + D_{\text{SPnodes}}$
(8) $\quad\quad$ Delete all duplications from D_{graph}
(9) $\quad D_{\text{final}} \leftarrow D_{\text{final}} + (D_{t_i} + D_{\text{graph}})$
(10) $\quad$ Delete all duplications from D_{final}
(11) $\quad$ Rerank D_{final}

Algorithm 1: Retrieving based on DGD feedback.

Amazon suggests books in "Similar Products" according to there similarity to the consulted book. The degree of similarity is relative to user's social information like clicks or purchases and content based information like book attributes (book description, book title, etc.). The exact formula that combines social and content based information to compute similarity is not delivered by Amazon.

To perform data modeling into DGD, we extracted the "Similar Products" links between documents. The constructed DGD will be used to enrich returned results by the retrieval models; that is, we use the graph structure in the same spirit of pseudorelevance-feedback algorithms. Our method can potentially serve to greatly enhance the retrieval performances.

In this section we use some fixed notations. The collection of documents is denoted by C. In C, each document d has a unique ID. The set of queries called topics is denoted by T, the set $D_{\text{init}} \subset C$ refers to the documents returned by the initial retrieval model. StartingNode indicates document from D_{init} which is used as input to graph processing algorithms in the DGD. The set of documents present in the graph is denoted by S. D_{t_i} indicates the documents retrieved for topic $t_i \in T$.

Each node in the DGD represents document (Amazon description of book) and has set of properties:

(i) ID: book's ISBN;

(ii) content: book description that includes many other properties (title, product description, author(s), users' tags, content of reviews, etc.);

(iii) MeanRating: average of ratings attributed to the book;

(iv) PR: value of book PageRank.

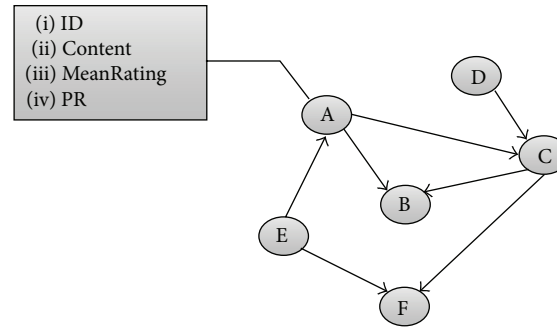

Figure 2: Example of Directed Graph of Documents.

Nodes are connected with directed links; given nodes $\{A, B\} \in S$, if A points to B, B is suggested as Similar Product to A. In Figure 2, we show an example of DGD, network of documents. The DGD network contains 1.645.355 nodes (89.86% of nodes are in the collection and the rest do not belong to it) and 6.582.258 relationships.

5.1. Our Approach. The DGD network contains useful information about documents that can be exploited for document retrieval. Our approach is based first on results of traditional retrieval approach and then on the DGD network to find other documents. The idea is to suppose that the suggestions given by Amazon can be relevant to the user queries.

We introduce the algorithm called "retrieving based on DGD feedback." Algorithm 1 takes the following as inputs: D_{init} returned list of documents for each topic by the

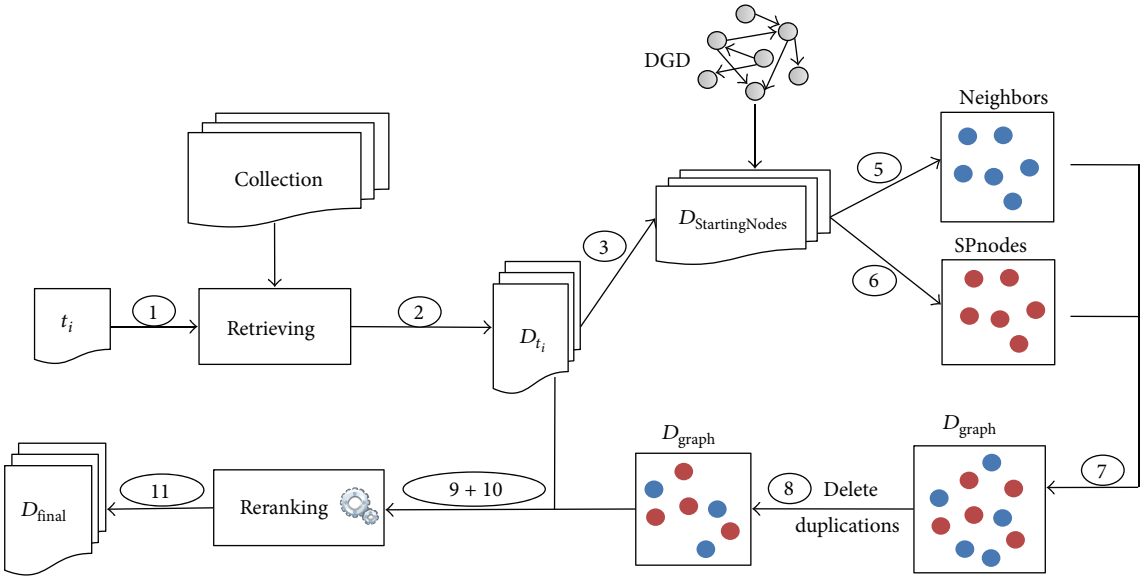

Figure 3: Architecture of book retrieval approach based on DGD feedback.

retrieval techniques described in Section 3, DGD network, and parameter β which is the number of the top selected StartingNode from D_{init} denoted by $D_{\text{StartingNodes}}$. We fixed β to 100 (10% of the returned list for each topic). The algorithm returns list of recommendations for each topic denoted by "D_{final}". It processes topic by topic and extracts the list of all neighbors for each StartingNode. It performs mutual Shortest Paths computation between all selected StartingNode in DGD. The two lists (neighbors and nodes in computed Shortest Paths) are concatenated; after that all duplicated nodes are deleted. The set of documents in returned list is denoted by "D_{graph}". A second concatenation is performed between initial list of documents and D_{graph} (all duplications are deleted) in new final list of retrieved documents; "D_{final}" reranked using different reranking schemes.

Figure 3 shows the architecture of the document retrieval approach based on DGD feedback. The numbers on arrows represent instructions in Algorithm 1.

6. Experiments and Results

In this section, we describe the experimental setup we used for our experiments. Furthermore, we present the different reranking schemes used in previously defined approaches.

6.1. Experiments Setup. For our experiments, we used different tools that implement retrieval models and handle the graph processing. First, we used Terrier (Terabyte Retriever) (http://terrier.org/) Information Retrieval framework developed at the University of Glasgow [26–28]. Terrier is a modular platform for rapid development of large-scale IR applications. It provides indexing and retrieval functionalities. It is based on DFR framework and we used it to deploy InL2 model described in Section 4.1. Further information about Terrier can be found at http://ir.dcs.gla.ac.uk/terrier.

A preprocessing step was performed to convert INEX SBS corpus into Trec Collection Format (http://lab.hypotheses.org/1129), by considering that the content of all tags in each XML file is important for indexing; therefore the whole XML file was transformed on one document identified by its ISBN. Thus, we just need two tags instead of all tags in XML, the ISBN and the whole content (named text).

Secondly, *Indri* (http://www.lemurproject.org/indri/), *Lemur Toolkit for Language Modeling and Information Retrieval* was used to carry out a language model (SDM) described in Section 4.2. Indri is a framework that provides state-of-the-art text search methods and a rich structured query language for big collections (up to 50 million documents). It is a part of the Lemur project and developed by researchers from UMass and Carnegie Mellon University. We used Porter stemmer and performed Bayesian smoothing with Dirichlet priors (Dirichlet prior $\mu = 1500$).

In Section 5.1, we have described our approach based on DGD which includes graph processing. We used NetworkX (https://networkx.github.io/) tool of Python to perform shortest path computing, neighborhood extraction, and PageRank calculation.

To evaluate the results of retrieval systems, several measurements have been used for SBS task: Discounted Cumulative Gain (nDCG), the most popular measure in IR [29], Mean Average Precision (MAP) which calculates the mean of average precisions over a set of queries, and other measures: Recip Rank and Precision at the rank 10 (P@10).

6.2. Reranking Schemes. In this paper, we have proposed two approaches. The first (see Section 4.3) consists of merging the results of two different information retrieval models which are the language model (SDM) and DFR model (InL2). For topic t_i, each of the models gives 1000 documents and each retrieved document has an associated score. The linear combination method uses the following formula to calculate

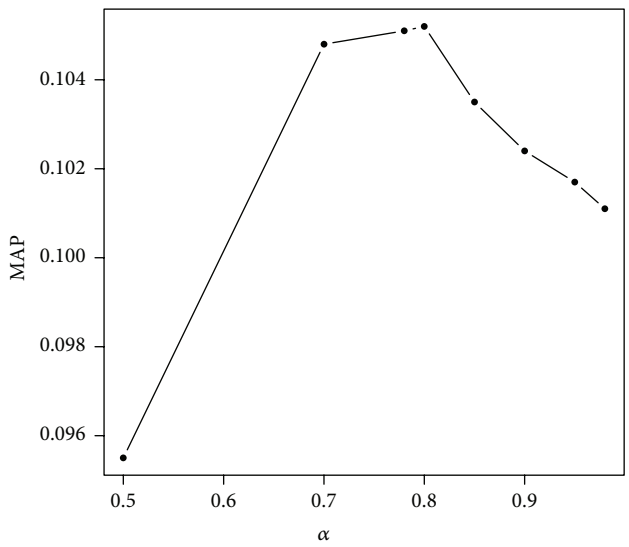

Figure 4: MAP distribution for INEX SBS 2014 when varying the interpolation parameter α.

final score for each retrieved document d by SDM and InL2 models:

$$\text{final}_{\text{score}(d,t_i)} = \alpha \cdot \text{score}_{\text{InL2}}(d,t_i) + (1-\alpha) \cdot \text{score}_{\text{SDM}}(d,t_i), \quad (5)$$

where $\text{score}_{\text{InL2}}(d,t_i)$ and $\text{score}_{\text{SDM}}(d,t_i)$ are normalized scores. α is the interpolation parameter set up at 0.8 after several tests on topics 2014, according to the MAP measure shown in Figure 4.

The second approach (described in Section 5.1) uses the DGD constructed from the "Similar Products" information. The document set returned by the retrieval model are fused to the documents in neighbors set and Shortest Path results. We tested different reranking methods that combine the retrieval model scores and other scores based on social information. For each document in the resulting list, we calculated the following scores:

(i) *PageRank*, computed using NetworkX tool: it is a well-known algorithm that exploits link structure to score the importance of nodes in a graph. Usually, it has been used for hyperlink graphs such as the Web [30].

(ii) *MeanRatings*, information generated by users: it represents the mean of all ratings attributed by users for a book.

The computed scores were normalized using this formula: normalizedScore = oldScore/maxScore. After that, to combine the results of retrieval system and each of normalized scores, an intuitive solution is to weight the retrieval model scores with the previously described scores (normalized PageRank and MeanRatings). However, this would favor documents with high PageRank and MeanRatings scores even though their content is much less related to the topics.

6.3. Results. In this section, we discuss the results we achieved by using the InL2 retrieval model, its combination to the SDM model, and retrieval system proposed in our approach that uses the graph structure DGD.

We used two topic sets provided by INEX SBS task in 2014 (680 topics) and 2015 (208 topics). The systems retrieve 1000 documents per topic. The experimental results, which describe the performance of the different retrieval systems on INEX document collection, are shown in Table 2.

As illustrated in Table 2, the system that combines probabilistic model InL2 and the language model SDM (InL2_SDM) achieves a significant improvement comparing to InL2 model (baseline). The two systems do not provide similar level of retrieval effectiveness (as proved in [16]). The results of run InL2_DDG_PR using the 2015 topic set confirm that incorporating PageRank scores using our approach based on DGD network improves ranked retrieval performance but decreases the baseline performances when using the 2014 topic set. The run that uses the MeanRatings property (InL2_DGD_MnRtg) to rerank retrieved documents is the lowest performer in terms of all measurements for 2014 topic set. That means that ratings given by users do not help to improve the reranking performances for 2014 topic set.

Notice that 2015 topic set contains a subset of 2014 topic set. We can clearly observe the difference between results of systems using the two topic sets. We think that the main reason is the evaluation processes that are not the same. Where analysing the qrels of common topics between 2014 and 2015 sets, we found that relevancy values are not the same in most cases. More details on the evaluation processes used in 2014 and 2015 can be found in [31] and http://social-book-search.humanities.uva.nl/#/suggestion.

Nevertheless, the depicted results confirm that we are starting with competitive baseline, the improvements contributed by combining the retrieval systems' outputs and social link analysis are indeed meaningful.

7. Conclusion and Future Work

This paper proposed and evaluated two approaches of document retrieval in the context of book recommendation. We used the test collection of INEX Social Book Search track and the proposed topics in 2014 and 2015. We presented the first approach that combines the outputs of probabilistic model (InL2) and language model (SDM) using a linear interpolation after normalizing scores of each retrieval system. We have shown a significant improvement of baseline results using this combination.

This paper also proposed a novel approach based on Directed Graph of Documents (DGD). It exploits social link structure to enrich the returned document list by traditional retrieval model (InL2, in our case). We performed a reranking method using PageRank and ratings of each retrieved document.

For future work, we would like to test the proposed approaches in this paper on another test collection which consists of OpenEdition Portal. It is dedicated to electronic resources in the humanities and social sciences. We would like to explore citation links between scientific documents

Table 2: Experimental results. The runs are ranked according to nDCG@10. (∗) denotes significance according to Wilcoxon test [8]. In all cases, all of our tests produced two-sided p value, $\alpha = 0.05$.

		2014 topic set		
Run	nDCG@10	Recip Rank	MAP	P@10
InL2	**0.128**	**0.236**	**0.101**	**0.067**
InL2_SDM	0.136 (+6%*)	0.249 (+5%*)	0.1052 (+4%*)	0.070 (+4%*)
InL2_DGD_PR	0.122 (−4%*)	0.239 (+1%*)	0.090 (−9%*)	0.0695 (+2%*)
InL2_DGD_MnRtg	0.105 (−17%*)	0.192 (−18%*)	0.081 (−18%*)	0.057 (−15%*)
		2015 topic set		
Run	nDCG@10	Recip Rank	MAP	P@10
InL2	**0.063**	**0.147**	**0.046**	**0.044**
InL2_SDM	0.069 (+9%*)	0.166 (+12%*)	0.051 (+10%)	0.050 (+13%*)
InL2_DGD_PR	0.068 (+7%*)	0.157 (+6%*)	0.048 (+4%*)	0.052 (+18%*)
InL2_DGD_MnRtg	0.066 (+4%)	0.148 (+0.6%)	0.042 (−8%)	0.052 (+18%*)

extracted using Kim et al. [32]. Using the traversal algorithms we will develop a new way to retrieve documents. Another interesting extension of this work would be using the learning to rank techniques to automatically adjust the settings of reranking parameters.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the French program "*Investissements d'Avenir—Développement de l'Economie Numérique*" under Project Inter-Textes no. O14751-408983.

References

[1] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pp. 279–280, ACM, Berkeley, Calif, USA, August 1999.

[2] T. Tao, X. Wang, Q. Mei, and C. Zhai, "Language model information retrieval with document expansion," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, R. C. Moore, J. A. Bilmes, J. Chu-Carroll, and M. Sanderson, Eds., pp. 407–414, Association for Computational Linguistics, Nagoya, Japan, 2006.

[3] C. X. Zhai, *Statistical Language Models for Information Retrieval*, Synthesis Lectures on Human Language Technologies, Morgan and Claypool Publishers, 2008.

[4] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pp. 275–281, Melbourne, Australia, August 1998.

[5] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, pp. 472–479, ACM, Salvador, Brazil, August 2005.

[6] G. Amati and C. J. Van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357–389, 2002.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," Tech. Rep. 1999-66, The Stanford University InfoLab, 1999, Previous number: SIDL-WP-1999-0120.

[8] W. B. Croft, *Organizing and searching large files of document descriptions [Ph.D. thesis]*, Cambridge University, 1978.

[9] H. Bouchard and J.-N. Nie, "Modèles de langue appliqués à la recherché d'information contextuelle," in *Conférence en Recherche d'Information et Applications (CORIA '06)*, pp. 213–224, Université de Lyon, Lyon, France, March 2006.

[10] M. Abolhassani and N. Fuhr, "Applying the divergence from randomness approach for content-only search in XML documents," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, pp. 409–419, Springer, Berlin, Germany, 2004.

[11] O. Kurland and L. Lee, "PageRank without hyperlinks: structural re-ranking using links induced by language models," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, pp. 306–313, ACM, Salvador, Brazil, August 2005.

[12] J. Lin, "Pagerank without hyperlinks: reranking with pubmed related article networks for biomedical text retrieval," *BMC Bioinformatics*, vol. 9, no. 1, article 270, 2008.

[13] T. Beckers, N. Fuhr, N. Pharo, R. Nordlie, and K. N. Fachry, "Overview and results of the INEX 2009 interactive track," in *Research and Advanced Technology for Digital Libraries: 14th European Conference, ECDL 2010, Glasgow, UK, September 6–10, 2010. Proceedings*, vol. 6273 of *Lecture Notes in Computer Science*, pp. 409–412, Springer, Berlin, Germany, 2010.

[14] M. Koolen, T. Bogers, J. Kamps, G. Kazai, and M. Preminger, "Overview of the INEX 2014 social book search track," in *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014*, pp. 462–479, 2014.

[15] G. Amati and C. J. van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357–389, 2002.

[16] C. Benkoussas, H. Hamdan, S. Albitar, A. Ollagnier, and P. Bellot, "Collaborative filtering for book recommandation," in *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014*, pp. 501–507, 2014.

[17] R. Guillén, "GIR with language modeling and DFR using terrier," in *Evaluating Systems for Multilingual and Multimodal Information Access*, C. Peters, T. Deselaers, N. Ferro et al., Eds., vol. 5706 of *Lecture Notes in Computer Science*, pp. 822–829, Springer, Berlin, Germany, 2009.

[18] V. Plachouras, B. He, and I. Ounis, "University of glasgow at trec 2004: experiments in web, robust, and terabyte tracks with terrier," in *Proceedings of the 13th Text Retrieval Conference (TREC '04)*, E. M. Voorhees and L. P. Buckland, Eds., Special Publication 500-261, National Institute of Standards and Technology (NIST), Gaithersburg, Md, USA, November 2004.

[19] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter, "Probabilistic models of indexing and searching," in *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '80)*, pp. 35–56, Kent, UK, 1980.

[20] L. Bonnefoy, R. Deveaud, and P. Bellot, "Do social information help book search?" in *CLEF (Online Working Notes/Labs/Workshop)*, P. Forner, J. Karlgren, and C. Womser-Hacker, Eds., 2012.

[21] D. Metzler and W. B. Croft, "Combining the language model and inference network approaches to retrieval," *Information Processing and Management*, vol. 40, no. 5, pp. 735–750, 2004.

[22] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, Eds., pp. 472–479, ACM, Salvador, Brazil, August 2005.

[23] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw, "Combining the evidence of multiple query representations for information retrieval," *Information Processing and Management*, vol. 31, no. 3, pp. 431–448, 1995.

[24] J. H. Lee, "Combining multiple evidence from different properties of weighting schemes," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95)*, pp. 180–188, ACM, New York, NY, USA, 1995.

[25] J. Lin, "PageRank without hyperlinks: reranking with PubMed related article networks for biomedical text retrieval," *BMC Bioinformatics*, vol. 9, no. 1, article 270, 2008.

[26] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma, "Terrier: a high performance and scalable information retrieval platform," in *Proceedings of the ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR '06)*, Seattle, Wash, USA, August 2006.

[27] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson, "Terrier information retrieval platform," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings*, vol. 3408 of *Lecture Notes in Computer Science*, pp. 517–519, Springer, Berlin, Germany, 2005.

[28] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras, "Research directions in terrier: a search engine for advanced retrieval on the web," *Novatica/UPGRADE Special Issue on Next Generation Web Search*, vol. 8, no. 1, pp. 49–56, 2007.

[29] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*, E. Yannakoudakis, N. Belkin, P. Ingwersen, and M.-K. Leong, Eds., pp. 41–48, ACM, Athens, Greece, July 2000.

[30] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," in *Proceedings of the 7th International World Wide Web Conference*, pp. 161–172, Brisbane, Australia, April 1998.

[31] P. Bellot, T. Bogers, S. Geva et al., "Overview of inex 2014," in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, E. Kanoulas, M. Lupu, P. Clough, and etal, Eds., vol. 8685 of *Lecture Notes in Computer Science*, pp. 212–228, Springer, Berlin, Germany, 2014.

[32] Y.-M. Kim, P. Bellot, E. Faath, and M. Dacos, "Automatic annotation of bibliographical references in digital humanities books, articles and blogs," in *BooksOnline*, G. Kazai, C. Eickhoff, and P. Brusilovsky, Eds., pp. 41–48, ACM, 2011.