Using Sentiment Analysis for Pseudo-Relevance Feedback in Social Book Search

Amal Htait
amal.htait@strath.ac.uk
University of Strathclyde
Glasgow, UK

Sébastien Fournier, Patrice Bellot
{firstname.lastname}@univ-amu.fr
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Marseille, France

Leif Azzopardi
leif.azzopardi@strath.ac.uk
University of Strathclyde
Glasgow, UK

Gabriella Pasi
gabriella.pasi@unimib.it
University of Milan Bicocca
Milan, Italy

ABSTRACT

Book search is a challenging task due to discrepancies between the content and description of books, on one side, and the ways in which people query for books, on the other. However, online reviewers provide an opinionated description of the book, with alternative features that describe the emotional and experiential aspects of the book. Therefore, locating emotional sentences within reviews, could provide a rich alternative source of evidence to help improve book recommendations. Specifically, sentiment analysis (SA) could be employed to identify salient emotional terms, which could then be used for query expansion? This paper explores the employment of SA based query expansion, in the book search domain. We introduce a sentiment-oriented method for the selection of sentences from the reviews of top rated book. From these sentences, we extract the terms to be employed in the query formulation. The sentence selection process is based on a semi-supervised SA method, which makes use of adapted word embeddings and lexicon seed-words. Using the CLEF 2016 Social Book Search (SBS) Suggestion Track Collection, an exploratory comparison between standard pseudo-relevance feedback and the proposed sentiment-based approach is performed. The experiments show that the proposed approach obtains 24%-57% improvement over the baselines, whilst the classic technique actually degrades the performance by 14%-51%.

KEYWORDS

Query Expansion, Sentiment Analysis, Pseudo-Relevance Feedback

ACM Reference Format:
Amal Htait, Sébastien Fournier, Patrice Bellot, Leif Azzopardi, and Gabriella Pasi. 2020. Using Sentiment Analysis for Pseudo-Relevance Feedback in Social Book Search. In *Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20), September 14–17, 2020, Virtual Event, Norway.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3409256.3409847

ICTIR '20, September 14–17, 2020, Virtual Event, Norway
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8067-6/20/09…$15.00
https://doi.org/10.1145/3409256.3409847

1 INTRODUCTION

Book search is a difficult Information Retrieval problem due to the vocabulary mismatch between book descriptions and user queries. Pseudo-relevance feedback (PRF), also known as blind relevance feedback, is considered as one of the most effective techniques for improving retrieval performance by query reformulation [13, 16]. As such, PRF provides a way to bridge this semantic gap. Typical PRF methods assume the top retrieved documents in the initial retrieval outcome are relevant, and terms are extracted from the pseudo *relevant* documents based on their term/document statistics. These are then employed in the query reformulation process. However, in the context of book search, where the goal is to rank books given a query [8], the usual application of PRF [4, 18] may not be appropriate, considering that the book's descriptions may not adequately convey the experience and emotion attached to the story line that a reader may be looking to enjoy.

Book social applications (e.g., LibraryThing [1], Goodreads [2]) offer, in addition to the catalogue of books' characteristics and/or their partial content, the information generated by users about the books. This information is typically constituted by reviews, which include opinions/sentiments and personal descriptions about books that can highlight certain aspects not included in the content of the books' representations. Therefore, once extracted, this information may disclose the experiential and emotional aspects of books that can be used to enrich the query with valuable information. However, the extraction of useful information for query expansion from book reviews is problematic. The quantity of reviews associated with a book (especially if it is popular) can be very large – and these reviews can be very noisy as they often include content unrelated to book information. To alleviate these problems, a number of methods have been proposed, not for query expansion, but to reduce the subset of reviews to be presented to users or for selecting which parts of the reviews to present [3, 17]. For example, Yang et al. [17] used Sentiment Analysis (SA) to highlight sentences with positive or negative sentiment polarity in reviews, to reduce the information overload while reading. While Badache et al. [3] made use of SA to identify emotionally loaded characters and entities given their proximity to emotional terms (e.g., *love*, *hate*) for the purpose of extracting interesting aspects from user comments. Those works, and others before [19], deduced that the SA can be a key factor in

[1] https://www.librarything.com/
[2] https://www.goodreads.com/

the mining and summarization of reviews. Given this prior related work, we posit that SA could help by acting as a filter – to limit the set of candidate terms used for expansion – and thus focus on the experiential and emotional aspects of the book.

In this paper we present an exploratory study aimed at exploiting SA to identify, in large collections of book reviews, those sentences from which the query expansion terms can be extracted. Our intuition is that the writers of book reviews are often guided by the experience and emotions provoked by the book's content, characters, plot, etc., which they express in a similar way to how readers of the book express their expectations regarding the book. For example, given the review: *[... my son sat still, absorbing every word.* ***The book has awesome pictures and it delivers a superb message at the end*** *...]*, where the user expresses a sentiment for the book's content with strongly positive expressions, while sharing information about the book. The strongly positive sentence in the example is in bold, since it includes strong positive terms underlined. The dashed, underlined terms within the strong positive sentence, would be the target to expand the initial query. Therefore, we hypothesise that locating sentences within book reviews, which include terms of strong sentiment polarity, may help in identifying useful terms for query expansion. On the other hand, using sentiment or emotion in documents to improve the effectiveness of the system has been explored in the past in other domains, such as *recommender systems*, where [11, 12] used the emotion in reviews to improve the effectiveness of their recommender systems, while in this work we focus on terms' sentiment intensity rather than general emotion. Also, in the *opinion retrieval* domain, [7] employed a query expansion method by adding a set of extracted *opinion words* (e.g., good, like) to the query, extracted from the top retrieved documents in a pseudo relevance feedback method. In their suggestion, [7] served of terms frequency and word weighting techniques, but they did not use any SA methods to classify the documents terms, and they did not go beyond the *opinion words* extraction to explore the extraction of informative words.

We present our work in this paper in two main sections: (1) In Section 2, the description of the *sentence selection* phase from book reviews by SA, including a description of the employed SA method, followed by the description of the *expansion-terms* selection phase. (2) The preliminary experiments in book search domain are reported in Section 3, including a description of the initial retrieval method prior to the query expansion procedure, and the achieved results.

2 SENTIMENT BASED QUERY EXPANSION

In this Section, we describe the method aimed at extracting from all reviews associated with the top ranked book (which we consider, in this preliminary study, as the most relevant to the initial query) the terms to be employed in query expansion. The method is organised into two main phases: the first phase is aimed at identifying, in a review, the sentences with terms characterised by a strong sentiment polarity (Section 2.1). The second phase is focused on extracting from those sentences the terms that will be employed in the query expansion (Section 2.2).

2.1 Sentence Selection Phase

The sentence selection phase is guided by SA. The basic idea is to reduce or eliminate the less important parts of the reviews, by assuming that the sentences expressing strong sentiments, in the reviews, should be considered as carriers of useful information. The selection of sentences from the book reviews relies on the following sub-phases: (1) First, each review is segmented into sentences, where the Sentiment Intensity Score (*SIS(w)*), of each term (w) in the sentence, is calculated by using the SIS method explained below. (2) Then, sentences including terms with very high or very low SIS are the ones selected (reflecting very positive or very negative sentiment, therefore a strong sentiment), considering very high SIS when *SIS(w)* > λ^+, and very low SIS when *SIS(w)* < λ^-.

Words Sentiment Intensity Score:
A semi-supervised method is employed to classify the words in book reviews by their SI. The method relies on the concepts of adapted seed-words and word embedding, where seed-words are words with strong semantic orientation both positive and negative, which are characterised by a lack of sensitivity to context [14]. Adapted seed-words, instead, are seed-words with the characteristic of being used in a certain context or related to a certain subject. For example, the word *touching* can express a positive opinion about a book or a movie, but this may not be the case in other contexts. In this paper, book reviews are the target objects for Sentiment Intensity (SI) classification; therefore, the seed-words are extracted from book reviews [3] [5] and adapted to the books domain. First, two lists of the most frequent words (Positive and Negative) are generated, then, the lists are manually filtered for the selection of words playing the role of seed-words. The following are examples of positive adapted seed-words: $touching$, $insightful$, $masterpiece$, and negative adapted seed-words: $endless$, $waste$, $unnecessary$. These seed-words are used in a word embedding model for the prediction of SI. Therefore, a word embedding model, with a vocabulary size of more than 2.5M words, is created based on Word2Vec [10] with Skip-Gram as training strategy, on more than 22M Amazon's book reviews [4] [9] as training dataset, after applying a pre-processing to the corpora (e.g. tokenization, replacing hyperlinks by $http$, removing special characters and punctuation, etc). The created word embedding model and the extracted seed-words are used to predict the SIS of words in book reviews using the cosine similarity measure to the vectors of words as explained in the following example: to calculate the SIS of the word w, first, the cosine similarities of w with all extracted positive seed-words are calculated, given the created word embedding model. The average of these scores would be the Positive Sentiment Intensity of w ($SI^+(w)$), presented in Equation 1 where $|S^+|$ is the number of words in the positive seed-words list and $CosSim(w,s)$ is the cosine similarity measure between the words w and one of the positive seed-words s. Then, the cosine similarities of w with all extracted negative seed-words are calculated. The average of these scores would be the Negative Sentiment Intensity of w ($SI^-(w)$), as presented in Equation 1 where $|S^-|$ is the number of words in the negative seed-words list. The difference

[3] Book reviews from Multi-Domain Sentiment Dataset by http://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html

[4] October 2018: http://jmcauley.ucsd.edu/data/amazon/

between $SI^{+}(w)$ and $SI^{-}(w)$ represents the SIS of the word w, as shown in the following equation: $SIS(w) = SI^{+}(w) - SI^{-}(w)$.

$$SI^{+/-}(w) = \frac{\sum_{s \in S^{+/-}} (CosSim(w, s))}{|S^{+/-}|} \quad (1)$$

In Figure 1, we present an example of the proposed SI classification method, in the book domain, of the word *youth* where its SIS is equal to 0.02 (neutral). Meanwhile, *boring* has a score equal to -0.126 (negative) and *exceptional* has a score equal to 0.232 (positive).

2.2 Term Selection

To expand the query, the selected subset of sentences from the reviews of the top k books are then taken as pseudo-relevance feedback. To decide which terms to then select, two main factors were considered in the process: (1) how important a word is in a collection of reviews (i.e. tf-idf), and (2) how intense the sentiment is within the word (i.e. $|SIS(w)|$). Thus the intuition was to select expansion terms that were important and not intense (i.e. $|SIS(w)|.\textit{tf-idf}$).

3 EXPERIMENTS

The document collection used in this study is provided by the Suggestion Track [5] of the Social Book Search Lab (SBS) at CLEF [8]. The provided collection of 2.8M book records contains professional book meta-data from Amazon books, based on controlled vocabularies (e.g, book title, author name) with a set of subject headings or classification information (e.g., cover type). In addition, the collection is extended with user-generated content and social meta-data from LibraryThing, presented as *reviews* that can contain information on how engaging, educational or well-written a book is, and *tags* which are single terms added by users describing the book. In addition, search queries are provided by SBS. In this work, a set of 120 search queries of the 2016's SBS Suggestion Track has been used; for each query an ideal list of books is provided by SBS for evaluation purposes. Furthermore, three different indexing strategies have been applied with each selected search model: [*All-Data*] an indexing of all books' meta-data in the books collection, it includes the professional and the user-generated/social meta-data (reviews and tags), [*Reviews*] an indexing of only the *reviews* (with a total number of 14M reviews in the collection), and [*Tags*] an indexing of only the *tags* (with a total number of 32M tags in the collection).

For the initial retrieval pass (baselines), we used two retrieval models: BM25 from Indri and InL2 from Terrier [1]. For comparison purposes, and to highlight the impact of SA on query expansion, we apply the classic tf-idf words weighting method, *QE* for Query Expansion, to extract terms from the first retrieved book reviews, based on their importance in a total collection of 22M reviews from Amazon's book reviews [9]. To select the number of retrieved terms, we apply a tuning from 1 to 10 terms. Finally, the proposed sentiment based Query Expansion method *SQE*, as explained above, was used, where we set λ^{+}=0.2 and λ^{-}=-0.1. We chose these values to exclude words that were not at least one standard deviation away from the mean ($\mu = 0.05, \sigma = 0.15$). For the terms extraction phase, as explained in section 2.2, terms were selected based on intensity and importance. However, we found that we could globally identify a subset of words that could be removed given the term scoring process, and thus create an additional stoplist to make the selection process more efficient. Thus, 500 additional words were identified, which had a high sentiment intensity such as *great*, *good*, *bad*, and/or low importance, due to the domain, such as *book*, *read*, *story*. Once these terms were removed, the remaining terms are added to the initial query for the query expansion. In addition to the retrieval models, we also employed a meta-search strategy based on Borda Count [6], it was included since previous studies have shown that combining results has been effective for book search [2, 15]. The metric used in this study is *Normalised Discounted Cumulative Gain* (*nDCG*), and we have reported *nDCG* at 10 retrieved elements (i.e. *nDCG*@10). To determine the statistical significance between runs, we performed t-test where significance was denoted when p-value is lower than 0.01. To facilitate reproducing the method, the SI classification code and tool is available on Github (link to be added).

4 RESULTS

Table 1 presents the results for each retrieval model and indexing strategy. The scores for the baselines, classic query expansion and the SA are shown along with the percentage changes over the baselines, in addition to the aggregation of retrievals by the Borda Count method.

Classic Query Expansion with tf-idf: The Borda Count aggregation method, applied on the baseline retrievals, reaches the best *nDCG*@10 score (0.121). Therefore, its retrieved list is used to extract the new terms for each search query, from the reviews of the top ranked book. In this tf-idf terms extraction method, we limit the terms extraction to 6 terms per query since it achieved the best result when tuning from 1 to 10 terms. The results are shown in the column *QE* (as Query Expansion) of Table 1, where the tf-idf method was not able to improve the baseline results, with any of the used search models and indexing strategies – on the contrary – it decreased the *nDCG*@10 values between 14% and 51%.

Proposed Query Expansion: For the proposed sentiment-based Query Expansion method, as for the classic tf-idf one, the retrieved list of the Borda Count aggregation method applied on the baseline is used to extract the new terms. The results are presented in column *SQE* (as for Sentiment Query Expansion) of Table 1, where the symbols (*,†) next to the scores indicate a p-value lower than 0.01, therefore a statistically significant result (vis-a-vis baselines and QE). The percentage of improvement, with all the used search models and indexing strategies, is presented in the last column of Table 1, %Δ, showing an improvement between 24% and 57%.

Analysis of Results: In a more detailed observation of the results, of the 120 tested queries of SBS 2016, the sentiment based method was able to improve the retrieval performance of 27 queries, but also decreased the retrieval performance of 14 queries. As for the tf-idf based method, it improved the retrieval performance of 18 queries, but decreased the retrieval performance of 25 queries. Therefore, in this study, our proposed method was able to increase the number of queries with an improved retrieval performance, and decrease the number of queries with a worse retrieval performance compared to the classic tf-idf method.

[5] http://social-book-search.humanities.uva.nl/#/suggestion

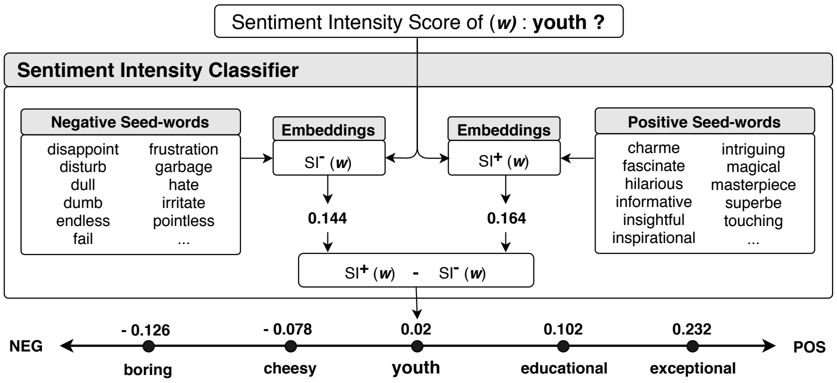

Figure 1: Sentiment Intensity Score, in book domain, where the score of the word is the difference between its positive SI and its negative SI, and where the SIs are calculated based on cosine similarity measure between the word and the seed-words.

Table 1: The *nDCG*@10 for each retrieval model and each indexing strategy, at baseline, at QE using tf-idf (QE), and QE by SA (SQE). * and † denote significant differences, consecutively with baselines and QE, with a p-value less than 0.01.

	Indexing	Baselines	QE	%Δ	SQE	%Δ
Okapi	All-data	0.105	0.072	−45%	**0.141**	+34%
	Reviews	0.108	0.085	−27%	**0.134**	+24%
	Tags	0.071	0.047	−51%	**0.107**	+51%
InL2	All-data	0.101	0.079	−27%	**0.144***†	+42%
	Reviews	0.093	0.078	−19%	**0.118**	+27%
	Tags	0.054	0.047	−14%	**0.085**	+57%
	Borda Count	0.121	0.101	−20%	**0.159***†	+31%

5 CONCLUSION

This paper presented an exploratory study on employing SA to query expansion by pseudo-relevance feedback, applied to the book search domain. The initial retrieval was an aggregation of the results of multiple retrieval models, with three different indexing strategies. Based on the results of the initial retrieval, the SI classification was used to filter information, by sentence selection from the reviews of the first initially retrieved book, for extracting terms to be exploited in query formulation. The preliminary experiments showed a ranking quality improvement (*nDCG*@10) between 24% and 57% after query expansion with the proposed approach compared to the initial query results. Such promising results indicate a potential of SA in query expansion.

Additional research and testing are required, such as (1) analysing and comparing the characteristics of queries in the case of improvement or deterioration by the retrieval performance of our suggested method, (2) investigating new potential for the suggested method, such as the effect of including the reviews of the *N top* retrieved books in the term extraction process, and not only the first retrieved book, (3) collecting online posts about the books with no reviews from micro-blogs (e.g., tweets) to eliminate any possible bias, (4) exploring new domains of application (e.g. movie, video, product, etc.) where the new terms for query expansion could be extracted from customers' reviews, in addition to social media and micro-blogs.

ACKNOWLEDGMENTS

The authors would like to thank the followings for their support:
- Project *EquipEx DILOH* (ANR-11-EQPX-0013)
- Project *Cumulative Revelations in Personal Data* (EP/R033897/1)

REFERENCES

[1] Giambattista Amati. 2003. *Probabilistic Models for Information Retrieval based on Divergence from Randomness. University of Glasgow, UK.* Ph.D. Dissertation.
[2] Javed A Aslam and Mark Montague. 2001. Models for metasearch. In *SIGIR.* 276–284.
[3] Ismail Badache, Sébastien Fournier, and Adrian-Gabriel Chifu. 2018. Predicting Contradiction Intensity: Low, Strong or Very Strong?. In *SIGIR.* 1125–1128.
[4] Jing Bai, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao. 2005. Query expansion using term relationships in language models for information retrieval. In *CIKM.* 688–695.
[5] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL.* 440–447.
[6] Jean-Charles de Borda. 1995. On elections by ballot. *Classics of social choice, eds. I. McLean, AB Urken, and F. Hewitt* (1995), 83–89.
[7] Xuanjing Huang and W Bruce Croft. 2009. A unified relevance model for opinion retrieval. In *CIKM.* 947–956.
[8] Marijn Koolen, Toine Bogers, Maria Gäde, Mark Hall, Iris Hendrickx, Hugo Huurdeman, Jaap Kamps, Mette Skov, Suzan Verberne, and David Walsh. 2016. Overview of the CLEF 2016 Social Book Search Lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction.* 351–370.
[9] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR.* 43–52.
[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop* (2013).
[11] Yashar Moshfeghi, Benjamin Piwowarski, and Joemon M Jose. 2011. Handling data sparsity in collaborative filtering using emotion and semantic based features. In *SIGIR.* 625–634.
[12] Cataldo Musto, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2019. Justifying recommendations through aspect-based sentiment analysis of users reviews. In *UMAP.* 4–12.
[13] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*, 313–323.
[14] Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *TOIS* 21, 4 (2003), 315–346.
[15] Merijn Van Erp and Lambert Schomaker. 2000. Variants of the borda count method for combining ranked classifier hypotheses. In *IWFHR.*
[16] Jinxi Xu and W Bruce Croft. 1996. Query Expansion Using Local and Global Document Analysis. In *SIGIR.* 4–11.
[17] Heng-Li Yang and August FY Chao. 2018. Sentiment annotations for reviews: an information quality perspective. *Online Information Review* 42, 5 (2018), 579–594.
[18] Zheng Ye and Jimmy Xiangji Huang. 2014. A simple term frequency transformation model for effective pseudo relevance feedback. In *SIGIR.* 323–332.
[19] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *CIKM.* 43–50.