Preliminary study of technical terminology for the retrieval of scientific book metadata records

Larsen, Birger; Lioma, Christina; Frommholz, Ingo; Schütze, Hinrich

Published in:
Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval

DOI:
10.1145/2348283.2348504

Publication date:
2012

Document version
Early version, also known as pre-print

Citation for published version (APA):
Preliminary Study of Technical Terminology for the Retrieval of Scientific Book Metadata Records

Birger Larsen  
Royal School of Library and Information Science, Copenhagen, Denmark  
blar@iva.dk

Christina Lioma  
Computer Science, University of Copenhagen, Denmark  
c.lioma@diku.dk

Ingo Frommholz  
Computer Science and Technology, University of Bedfordshire U.K.  
ingo.frommholz@beds.ac.uk

Hinrich Schütze  
Institute for Natural Language Processing, University of Stuttgart, Germany

ABSTRACT
Books only represented by brief metadata (book records) are particularly hard to retrieve. One way of improving their retrieval is by extracting retrieval enhancing features from them. This work focusses on scientific (Physics) book records. We ask if their technical terminology can be used as a retrieval enhancing feature. A study of 18,443 book records shows a strong correlation between their technical terminology and their likelihood of relevance. Using this finding for retrieval yields >+5% precision and recall gains.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords
Book Records, Technical Terminology

1. INTRODUCTION
Information retrieval (IR) systems often can rely on having full texts available for processing. However, there are cases when full text is not available, e.g. in commercial online bookstores or traditional libraries where material may only be available in print, or where using optical character recognition is difficult. Access to these materials is primarily realised through the supplier or the library catalogue, where documents are represented by short book records of metadata information, e.g. author, title, etc. The problem is that book records provide very little information, and hence they are very hard to retrieve. As a consequence, the accessibility of potentially relevant books is restricted for users. This work focuses on such books from the physics domain.

We ask whether we can increase the retrievability of physics book records by focussing on the special language in this scientific domain, as used by searchers and authors. Specifically, we model separately the technical/non-technical terminology of physics book records, motivated by the intuition that technical terminology may make a good retrieval enhancing feature. Our intuition that this modelling may benefit book record retrieval is experimentally confirmed: we find a strong correlation between the technical terminology contained in book records and their likelihood of relevance. Applying this to the retrieval of book records yields notable improvements in retrieval precision and recall.

2. TECHNICAL TERMINOLOGY AND RELEVANCE
Preprocessing. We use a collection of 18,443 physics book records with 53 queries and relevance assessments (qrels) from the iSearch dataset\footnote{http://itlab.dbit.dk/~isearch}. These book records contain basic Machine-Readable Cataloging information, e.g. title, key phrases. To identify technical terms, we part-of-speech (POS) tag the collection (including queries) with the TreeTagger\footnote{http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger}. We extract all terms tagged as nouns, verbs, adjectives and participles, which are the most salient POS classes, hence the most likely to be technical terms. This results in a list of 12,548 terms, which we submit to Amazon Mechanical Turk (AMT) as isolated tokens (without any context) for classification into technical/non-technical terms. Using 3 AMT users per annotation (≥95% approval rate, paid approximately $0.33 per hour), 34.7% of all terms were annotated as technical, and 65.3% as non-technical, with strong inter-annotator agreement (Fleiss’ $\kappa \approx 0.8$).

Technical terminology density analysis. We count the number of technical terms in each book record (referred to as document henceforth) normalised by their length - this gives the document’s terminological density. We sort all documents by their terminological density, and we divide them into 34 bins: 33 equal-sized bins of 300 documents each, and 1 bin of the remaining 253 documents. We estimate the probability that a randomly selected relevant document belongs to a certain bin as: $p(d \in b_i|\text{rel.}) = \frac{|\text{rel. } d \in b_i|}{|\text{rel. } d|}$ and the probability that a randomly selected retrieved document belongs to a certain bin as: $p(d \in b_i|\text{ret.}) = \frac{|\text{ret. } d \in b_i|}{|\text{ret. } d|}$, where
(rel./ret.)d is a (relevant/retrieved) document, bi is the
ith bin, and | · | denotes cardinality.

Fig. 1 shows the plots of these probabilities (y axis) vs.
the average terminological density in a bin (x axis). We see
that p(d ∈ bi|rel.) varies non-randomly across bins -
terminological density is positively correlated to probability
of relevance (ρ = 0.9008). Hence, boosting the ranking of
documents of higher terminological density may boost retrieval
performance. The differences between how p(d ∈ bi|rel.) and
p(d ∈ bi|ret.) vary across bins shows the document groups
for which the retrieval model underperforms: the retrieval
model has a weaker correlation to the documents’ termino-
logical density (ρ = 0.8773), with some instability in its per-
formance for documents of low terminological density, and
fails for documents of the highest terminological density.

Ranking model. If a user submits the query design of
biexcitonic models, and we know that biexcitonic is a technical
term, we hypothesise that boosting the weight of
documents that contain biexcitonic will improve retrieval
performance. We implement this boosting using Indri’s
combination of the Language Modeling (LM) and inference
network approaches [10], which allows assigning degrees of
belief to different parts of the query. This belief can be
drawn from any suitable external evidence of relevance -
in our case the knowledge that a search term is technical
terminology. Using the #weight and #combine operators
for combining beliefs, the relevance of a document d to a
query q is computed as the probability that d generates q:
p(q|d) = \prod_{t \in q} p(t|d)^{w_t}, where W = \sum_{t \in q} w_t, t is a
term and w_t is the term’s belief weight. The higher w_t is, the
higher the rank of documents containing t. We apply the above
equation separately for non-technical common query terms
with belief weight wcom, and for technical query terms with
belief weight tec, (wcom, tec ∈ wt, wcom + tec = 1). To
boost the ranking of documents containing technical terms,
we increase tec at the expense of wcom.

Experiments. The baseline matches the documents to
queries without any treatment of technical terminology using
LM with Dirichlet smoothing. Our approach boosts the
weight of technical terms using the same retrieval model but
enhanced with belief weights as described above (TEC). We
also use a pseudo-relevance feedback (FB) baseline (Indri’s
default FB implementation), against which we compare our
approach combined with FB (TEC+FB). We measure per-
formance with the standard TREC metrics shown in Table 1
averaged over all queries for the top 100 results (apart from
the number of relevant retrieved documents (REL.RET.)
which is summed). For each metric we tune: Dir’s μ ∈
{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000}; the
belief weights wcom, tec ∈ {0.1 – 0.9} in steps of 0.1 with
wcom + tec = 1 at all times; FB’s number of feedback
documents ∈ {1, 2, 5, 10, 20} and number of feedback terms
∈ {3, 5, 10, 20, 40}.

Table 1 shows that boosting the weight of technical termino-
logy improves retrieval at all times. The biggest
improvement is for REL.RET., indicating that our approach
introduces to the ranking relevant documents that neither
the baseline nor FB retrieve. This finding is positive, con-
sidering that our approach does not add new terms to the
query - it just boosts the weight of existing query (technical)
terms. Average precision benefits more when non-assessed
documents are ignored in the ranking (BPREF) than when
using graded relevance assessments (NDCG), possibly be-
cause NDCG gives a lower score to relevant documents that
occur in the low ranks (and hence ‘penalises’ non-relevant
documents or non-assessed documents that occur in the high
ranks). Our approach benefits early precision for both the
top 100 retrieved documents (P@100) and the first relevant
retrieved document (MRR). We can also report that our
approach outperforms the baseline and FB across the whole
tuning range of μ and for tec = 0.1 – 0.5 (plots not included
for brevity) without any outliers. The values tec = 0.1 – 0.5
practically correspond to applying a moderate boost to the
weight of technical terminology.

3. CONCLUSION

We asked whether the retrieval of scientific books repre-
sented only by limited metadata can be improved by using
their technical terminology, motivated by the empirical
finding that the proportion of technical terms they contained
was positively correlated with their probability of being rele-
vant. To our knowledge this is a novel finding. We integrated
this finding into the retrieval model successfully by boost-
ing the ranking of documents containing technical terms, hence
showing that our approach benefits the retrieval of book records.
Future work includes using the technical term
annotations to train an automatic classifier, comparing our
approach against an automatic way for determining term
relevance (e.g. ontologies, wikipedia pages), and testing the
generality of our approach on more scientific domains.

4. REFERENCES

model and inference network approaches to retrieval.

3Results were not stat. significant when the t-test was used.