Retrieval of Research-level Mathematical Information Needs: A Test Collection and Technical Terminology Experiment

Yiannos A. Stathopoulos **Simone Tuefel**
Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge, UK
{yiannos.stathopoulos,simone.teufel}@cl.cam.ac.uk

Abstract

In this paper, we present a test collection for mathematical information retrieval composed of real-life, research-level mathematical information needs. Topics and relevance judgements have been procured from the on-line collaboration website MathOverflow by delegating domain-specific decisions to experts on-line. With our test collection, we construct a baseline using Lucene's vector-space model implementation and conduct an experiment to investigate how prior extraction of technical terms from mathematical text can affect retrieval efficiency. We show that by boosting the importance of technical terms, statistically significant improvements in retrieval performance can be obtained over the baseline.

1 Introduction

Since their introduction through the Cranfield experiments (Cleverdon, 1960; Cleverdon, 1962; Cleverdon et al., 1966a; Cleverdon et al., 1966b), test collections have become the foundation of information retrieval (IR) evaluation.

Recent interest in Mathematical information retrieval (MIR) has prompted the construction of the NTCIR Math IR test collection (Aizawa et al., 2013). Like many general-purpose, domain-specific IR test collections, the NTCIR collection is composed of broad queries intended to test systems over a wide spectrum of query complexity.

In this paper we present a test collection composed of real-life, research-level mathematical topics and associated relevance judgements procured from the online collaboration web-site MathOverflow[1]. The resulting test collection contains 160 atomic questions - material derived from 120 MathOverflow discussion threads.

Topics in our test collection capture specialised information needs that are complex to resolve and often demand collective effort from multiple domain experts. For example[2]:

> The "most symmetric" Mukai-Umemura 3-fold with automorphism group $PGL(2, C)$ admits a Kaehler-Einstein metric according to Donaldson's result. On the contrary, there are some arbitrarily small complex deformations of the above 3-fold which do not admit Kaehler-Einstein metrics, as shown by Tian. All examples considered by Tian seem to have no symmetries at all. *Is it possible to find similarly arbitrarily small complex deformations with C^*-action and which do not admit any Kaehler-Einstein metric?*

Due to their specialised nature, our topics have a relatively small number of relevant documents. Fortunately, there is precedent of this from IR tasks such as QA (Ishikawa et al., 2010) and known-item search (Craswell et al., 2003).

With our test collection, we construct a baseline using Lucene's default implementation of the vector space model (VSM). Additionally, we conduct an experiment designed to investigate the hypothesis that technical terms in mathematics have elevated retrieval significance.

Information in mathematics is communicated by defining, manipulating and otherwise operating on mathematical structures and objects which can be instantiated in the mathematical discourse. In this sense, technical terminology in mathematics has an elevated role. This hypothesis stems from the observation that the mathematical discourse is dense with named mathematical objects, structures, properties and results.

[1] http://mathoverflow.net/

[2] Adapted from MathOverflow post 68096, http://mathoverflow.net/questions/68096/

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 334–340,
Beijing, China, July 26-31, 2015.

In the next section, we present our test collection and discuss the procedure for its construction from crowd-sourced expertise on MathOverflow. In section 3, we discuss related material in the literature and compare it to our work. Our experimental setup and results are discussed in section 4, with a brief summary of our work presented in section 5.

2 The Test Collection

The main motivation behind this work comes from our long-term goal to develop and evaluate MIR models intended to satisfy research-level mathematical information needs. Evaluation is an important final step in the development of IR models and is preconditioned on the availability of a test collection.

A test collection is a resource composed of (1) a *document collection* (or corpus) with uniquely identifiable documents (e.g., scientific papers, news articles), (2) a set of *topics* from which search queries can be produced and (3) a set of *relevance judgements*: pairs connecting individual topics to documents (in the corpus) known to satisfy the corresponding information need.

General-purpose MIR test collections, such as the one produced for NTCIR-10 (Aizawa et al., 2013), are expected to contain both broad and narrow topics capturing a wide range of retrieval complexity. In contrast, we require a collection of topics characterised by a higher lower bound on topic complexity with individual topics capturing highly-specialised, real-world information needs.

Unfortunately, research-level mathematical information needs are hard to source from documents in a way that would not render them artificial. Furthermore, manual construction of topics and relevance judgements is unrealistic due to the large number of experts required to cover the various specialised sub-fields of mathematics. This, coupled with limited access to numerous MIR systems, makes TREC-like pooling (Harman, 1993; Voorhees and Harman, 2005) impractical.

We propose that topics and relevance judgements be procured from the on-line collaboration website MathOverflow (MO), an online QA site for research mathematicians. A user (information seeker) can post a question on the site, usually relating to a small niche field in mathematics. Colleagues can either post a candidate answer, comment on the question, comment on and/or upvote existing answers. Ultimately, the information seeker decides which answer satisfies the underlying information need by marking it as "accepted".

Prelude	1) Apparently, physicist can calculate the GW invariants of quintic CY 3-fold up to genus 51. 2) For each genus g, there is a lower bound $d(g)$ such that for every $d < d(g)$, all genus g degree d invariants of quintic are zero.
MT-1	I am looking for a reference that has a table of these number for some low degrees (say up to degree 5) and low genera (at least until $g = 3$).
MT-2	Where can I found this lower bound?

Table 1: MO post 14655, prelude and micro-topics

Material on MO is closely aligned with our requirements. Specifically, Tausczik et al. (2014) and Martin and Pease (2013) agree that MO questions (information needs) *arise from doing mathematics research* and are novel to the mathematician involved. The authors conclude that, having been produced by experts, MO answers are *authoritative* and partially credit the website's reward system for their strong *reliability*.

MO questions often have multiple sub-parts, which we refer to as *micro-topics* since they encode atomic information needs. Furthermore, information in MO questions is carried by two types of sentences: *prelude* sentences, which are used to set the mathematical context (introduce mathematical constructs and results) and *query* sentences, which transcribe the information need itself and are semantically bound to the accompanying prelude.

As the underlying document collection, we have used the Mathematical Retrieval Corpus (MREC)[3] (Líška et al., 2011), which contains more than 439,000 mathematical publications, complete with mathematical formulae converted to machine-readable MathML. Similarly, we have made mathematical expressions in our topics accessible to MIR systems by converting all LaTeX embedded in MO questions into MathML using the LaTeXML tool-kit.

For the purpose of constructing our test collection we have adopted a multi-step process. All steps in the process are systematically applicable regardless of the subject material of the topic being considered for inclusion. As such, our test collection can be as diverse, in terms of mathematical subject and sub-fields, as MathOverflow.

[3] version 2011.4.439

Decisions relating to relevance of material to a given topic (MO question) are delegated to experts on the website. However, the information seeker (MO user posting the question) remains the ultimate judge of relevance. This authority is typically exercised by either accepting an answer directly or, by explicitly commenting on the relevance of posted material.

In the first step, all MO discussion threads[4] with at least one citation to the MREC in their accepted answer were collected. Each identified thread was examined by one of the authors for conformance to two ideal-standard criteria: (1) Useful MO questions should not be too broad or vague but rather express an information need that is clear and can be satisfied by describing objects or properties, stating conditions and/or producing examples or counter-examples. (2) MREC documents cited in MO accepted answers should address all sub-parts of the question in a manner that requires minimal deduction and do not synthesise mathematical results from multiple resources.

Subsequently, relevance of documents for each micro-topic is decided using two criteria: *totality* and *directness*. A cited resource is *total* if it contains all necessary information to derive the answer for the micro-topic and *partial* if it only addresses a special case. A cited resource is also said to be *direct*, if the answer can be derived with little intellectual effort from its text, or *indirect* if the same information requires considerable effort (such as mathematical deduction or reasoning) for the information seeker to reproduce.

Making these determinations involves matching the language of arguments and the symbolic context of the answer to the cited resource. As part of this step, we also examine the post-answer (PA) comments for expressions of confirmation of the usefulness of a cited resource from the information seeker.

The completed test collection contains 160 micro-topics with 184 associated relevance judgements (involving 224 unique MREC documents) organised in 120 topics. Topic text in our test collection is sentence tokenised, with relevance judgements being represented conceptually as tuples of the form:

```
(Topic ID, sentenceID, Micro-topic ID,
    relevant MREC document ID)
```

From Table 2 we observe that the vast majority of topics (93.33%) have either 1 or 2 micro-topics, with the average being close to 1 (1.33). The majority of topics (97,80.83%) have only one relevant document while a further 21 (17.5%) have two relevant documents. Two topics have more than 2 relevant documents: one with 3 and another with 4. In terms of micro-topics, this corresponds to 140 micro-topics (87.50%) with 1 relevant document, 17 (10.625%) with 2, 2 micro-topics (1.25%) with 3 and just one (0.625%) with 4 relevant documents.

Micro-topics	1	2	3	≥ 4
Instances (topics)	88	24	8	0
Percentage	73.33%	20.00 %	6.67%	0%

Table 2: Topic/Micro-topic break-down

3 Related Work

Test collections over scientific publications were first introduced for the Cranfield experiments (Cleverdon, 1960; Cleverdon, 1962; Cleverdon et al., 1966a; Cleverdon et al., 1966b). Despite criticism for sourcing queries from collection documents, the Cranfield experiments highlighted the importance of jointly reporting recall and precision, pioneered the practice of using authors and citations for augmenting relevance judgements and established the test collection paradigm.

Expert citations have already been exploited for procuring relevance judgements. For example, Ritchie et al. (2006) elicited relevance judgements for citations in papers accepted in a scientific conference from their authors and used these judgements as part of their test collection of scientific publications.

In terms of domain, our work is related to the NTCIR-10 Math IR test collection (Aizawa et al., 2013). Furthermore, the topics in our collection are analogous to those in the NTCIR full-text search, in the sense that they take the form of coherent text interspersed with mathematical expressions. Rather than being focused on accommodating information needs of varying complexity, however, our test collection has been designed to facilitate retrieval of highly specialised, mathematical information needs of uniformly high complexity.

Similar use of crowd-sourced expertise has been proposed in the context of QA. For example, Gyongyi et al. (2008), examined 10 months-worth of "Yahoo! Answers" material as part of an investigation of QA data, which was later used for the

[4] from MathOverflow data-dump of 20/01/2014

NTCIR-8 Community QA pilot task (Ishikawa et al., 2010; Sakai et al., 2011). Characterisation of crowd-sourced answers in terms of totality (section 2) has also been considered in the context of QA. In particular, Sakai et al. (2011) describe a relevance grading scheme of crowd-sourced answers based on the total/partial/irrelevant scale, but highlight that answers on "Yahoo! Answers" vary in quality (e.g., due to instances of bias or obscenity).

Finally, the idea of sourcing relevance judgements from expert citations is an established practice in IR. In the context of patent search, for example, Graf and Azzopardi (2008) utilised citations in patent office expert reports as relevance judgements, while Fujii et al. (2006) automatically extracted patent office expert citations used to reject patent applications.

4 Experiments

In this section we conduct an experiment to demonstrate the usefulness of our test collection by investigating the impact of terminology boosting on MIR effectiveness. An important assumption of this experiment is that the retrieval of each micro-topic is dependent only on the attached prelude.

4.1 Experimental Setup

We first produced a Lucene index over all documents in the MREC. In order to normalise processing of XHTML+MathML, topics and MREC documents were passed through the Tika framework[5]. Lucene's `StandardAnalyzer` was modified to preserve stop-words since frequent words such as the preposition "of" can be important parts of technical terms (e.g., "set of vectors"). The analyzer was also modified to preserve dashes, which are common in technical terms (e.g., "Calabi-Yau manifold"). This analyzer is used during both indexing and query processing for consistency.

4.2 Building Queries

For each micro-topic in a given topic, we emit a query string by concatenating all sentences in the prelude with sentences associated with the micro-topic. For example, query string for micro-topic MT-1 in Table 1 is generated by concatenating its text with that of the prelude. Using this strategy, consistency with the assumption outlined at the beginning of the section is achieved since no overlap beyond the prelude is introduced between queries generated for micro-topics attached to a given topic.

[5] `https://tika.apache.org/`

4.3 Systems

Using Lucene as the indexing and searching backend, we compare the performance of two retrieval methods. Underpinning both methods is Lucene's default similarity (project, 2013), which is based on cosine similarity:

$$sim(q, d) = \frac{V(q).V(d)}{|V(q)||V(d)|}$$

where $V(q)$ and $V(d)$ are weighted vectors for the query and candidate document respectively. As a performance measure, we use mean average precision (MAP):

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$$

4.3.1 Baseline

Lucene's VSM implementation with default TF-IDF weighting and scoring is used as the baseline. This is intended to emulate a general-purpose information retrieval scenario, which is the motivation behind the design of Lucene's default configuration.

4.3.2 Boosted Technical Terms

The alternative model is designed to give more weight to technical terminology common to both documents and queries. In order to construct this model, all technical terms are extracted from the document collection using an implementation of the C-Value multi-word technical term extraction method (Frantzi et al., 1996; Frantzi et al., 1998). Given an input corpus, the C-Value method extracts multi-word terms by making use of a linguistic and a statistical component.

The linguistic component is responsible for eliminating multi-term strings that are unlikely to be technical terms through the use of a stop-list (composed of high-frequency corpus terms) and linguistic filters (regular expressions) applied on sequences of part-of-speech tags. The statistical component assigns a "termhood" score to a candidate term sequence based on corpus-wide statistical characteristics of the sequence itself and those of term sequences that contain it. The output of

class/Form 1	Form 2 ...	Form 8	C-Value
riemannian manifold	Riemannian manifold	RIEMANNIAN MANIFOLDS	13236.6

Table 3: C-Value technical-term list entry

Original Text
a Riemannian manifold is a smooth manifold
Original Term vector
(a,2), (Riemannian,1),(manifold,2),(is,1),(smooth,1)
Technical terms
Riemannian manifold, smooth manifold
Re-Attributed Term Vector
(a,2), (Riemannian_manifold,1),(is,1),(smooth_manifold,1)
Re-generated delta index text
a a a a Riemannian_manifold Riemannian_manifold is is smooth_manifold smooth_manifold

Table 4: Example of re-attribution and delta index

the algorithm is a list of candidate technical terms in the corpus, ordered by their C-Value termhood score.

As shown in Table 3, each entry in the resulting list represents a single technical term (the class) and enumerates all forms of the candidate term as observed in the input corpus. In total, 3 million classes of technical terms have been detected in the MREC. Using Lucene's positional indexing mechanism, we retrieved the position of each technical term (all forms), recorded its term frequency (TF) and produced a new technical term index. This technical term index contains 426 million tuple entries of the form

```
<class, form, MREC docid, TF, position
    -of-occurrence list>
```

The same re-indexing process is repeated for the queries and the result is stored in a separate query table (10,433 entries).

Subsequently, the indexed document and query term vectors were modified by (1) adding new tokens to represent technical term phrases and (2) re-attributing the TF of component terms to the term vector of the phrase.

Finally, the text for each MREC document and query is re-generated from the term vectors and stored in a "delta index". At this stage, the number of technical term instances emitted is twice that recorded by the original term vector. This has the effect of boosting the significance of technical terms and phrases. An example of the application of this process, from original text to delta index generation is presented in Table 4. Rankings for the alternative model can be obtained by searching the delta index using the re-generated query.

	Baseline	Tech-Term boosting	Difference
MAP	0.0602	0.0732	**0.013*** (17.7%)

Table 5: Difference in MAP performance between models (* **statistically significant at** $\alpha = 0.05$)

Although the choice of boosting factor 2 is arbitrary, our intention is to demonstrate the presence of a difference in retrieval efficiency, rather than optimising the effect of boosting.

4.4 Results

The MAP scores obtained for the models are presented in Table 5. We observe that the difference in MAP is in favour of the alternative model. This difference is statistically significant at $\alpha = 0.05$ using the Wilcoxon signed-rank test ($p < 0.05$). Therefore, we have sufficient evidence to conclude that, in the context of the VSM, boosting technical terms improves retrieval efficiency of research mathematics.

When compared to MAP scores produced by the same systems in more traditional IR tasks, the scores in Table 5 may seem poor. We attribute this phenomenon to the fact that sense in written mathematics is communicated via a complex interaction of text and mathematical expressions and is thus hard to extract using shallow methods.

5 Conclusions and Further Work

We have constructed a Math IR test collection for real-life, research-level mathematical information needs. As part of the work of constructing our test collection, we have developed a methodology for compiling domain-specific test collections that requires minimal expertise in the domain itself.

Using 160 micro-topics in our test collection, we have shown experimentally that the performance of VSM-based retrieval models with research mathematics can be improved by boosting the importance of technical terminology. Furthermore, our experimental work suggests that our test collection can be used to identify statistically significant differences between MIR systems. It is our intention to make our collection available to the IR community.

As part of on-going and future work, we will be incorporating additional retrieval models, such as the Okapi BM25, in our evaluation framework. In addition, we are looking into investigating the statistical properties of our test collection along the lines of Harman (2011) and Soboroff et al. (2001).

References

Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. 2013. Ntcir-10 math pilot task overview. In *Proceedings of the 10th NTCIR Conference*, June.

C.W. Cleverdon, J. Mills, M. Keen, and Aslib. Cranfield Research Project. 1966a. *Factors Determining the Performance of Indexing Systems, Vol. 1: Design*. Number v. 1 in Factors Determining the Performance of Indexing Systems. College of Aeronautics.

C.W. Cleverdon, J. Mills, M. Keen, and Aslib. Cranfield Research Project. 1966b. *Factors determining the performance of indexing systems, Vol 2: Test Results*. Number v. 2 in Factors Determining the Performance of Indexing Systems. College of Aeronautics.

C. W. Cleverdon. 1960. Report on the first stage of an investigation into the comparative efficiency of indexing systems. Technical report.

C. W. Cleverdon. 1962. Aslib cranfield research project: Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, October.

Cyril W. Cleverdon. 1991. The significance of the cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '91, pages 3–12, New York, NY, USA. ACM.

Nick Craswell, David Hawking, Ross Wilkinson, and Mingfang Wu. 2003. Overview of the trec-2003 web track. In *Proceedings of TREC-2003*, Gaithersburg, Maryland USA, November.

K. Frantzi, S. Ananiadou, and J. Tsujii. 1996. Extracting terminological expressions. In *The Special Interest Group Notes of Information Processing Society of Japan , IPSJ SIG Notes*, number 112 in Natural Language, page 8388.

Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '98, pages 585–604, London, UK, UK. Springer-Verlag.

Atsushi Fujii, Makoto Iwayama, and Noriko K. 2006. Test collections for patent retrieval and patent classification in the fifth ntcir workshop.

E. Graf and L. Azzopardi. 2008. A methodology for building a patent test collection for prior art search.

Zoltan Gyongyi, Georgia Koutrika, Jan Pedersen, and Hector Garcia-Molina. 2008. Questioning yahoo! answers.

Donna Harman. 1993. Overview of the first trec conference. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 36–47, New York, NY, USA. ACM.

Donna Harman. 2011. *Information Retrieval Evaluation*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Daisuke Ishikawa, Tetsuya Sakai, and Noriko Kando. 2010. Overview of the ntcir-8 community qa pilot task (part i): The test collection and the task.

Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Akihiko Takano. 2003. Overview of patent retrieval task at ntcir-3. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20*, PATENT '03, pages 24–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

K.S. Jones. 1981. *Information retrieval experiment*. Butterworths.

Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec. 2011. Web interface and collection for mathematical retrieval: Webmias and mrec. In Petr Sojka and Thierry Bouche, editors, *Towards a Digital Mathematics Library.*, pages 77–84, Bertinoro, Italy, Jul. Masaryk University.

Ursula Martin and Alison Pease. 2013. What does mathoverflow tell us about the production of mathematics? *CoRR*, abs/1305.0904.

Lucene project. 2013. Lucene's tf-idf similarity function.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006. Creating a test collection for citation-based ir experiments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 391–398, New York City, USA, June. Association for Computational Linguistics.

Tetsuya Sakai, Daisuke Ishikawa, Noriko Kando, Yohei Seki, Kazuko Kuriyama, and Chin-Yew Lin. 2011. Using graded-relevance metrics for evaluating community qa answer selection. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 187–196, New York, NY, USA. ACM.

Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 66–73, New York, NY, USA. ACM.

K. Sprck Jones and C.J. Van Rijsbergen. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. Technical report, British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.

K. Sprck Jones and C.J. Van Rijsbergen. 1976. Information retrieval test collections. *Journal of Documentation*, 32(1):59–75.

Heinrich Stamerjohanns and Michael Kohlhase. 2008. Transforming the ariv to xml. In Serge Autexier, John Campbell, Julio Rubio, Volker Sorge, Masakazu Suzuki, and Freek Wiedijk, editors, *Intelligent Computer Mathematics*, volume 5144 of *Lecture Notes in Computer Science*, pages 574–582. Springer Berlin Heidelberg.

Heinrich Stamerjohanns, Michael Kohlhase, Deyan Ginev, Catalin David, and Bruce R. Miller. 2010. Transforming large collections of scientific publications to xml. *Mathematics in Computer Science*, 3(3):299–307.

Yla R. Tausczik and James W. Pennebaker. 2011. Predicting the perceived quality of online mathematics contributions from users' reputations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1885–1888, New York, NY, USA. ACM.

Yla R. Tausczik and James W. Pennebaker. 2012. Participation in an online mathematics community: Differentiating motivations to add. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 207–216, New York, NY, USA. ACM.

Yla R. Tausczik, Aniket Kittur, and Robert E. Kraut. 2014. Collaborative problem solving: A study of mathoverflow. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 355–367, New York, NY, USA. ACM.

Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 315–323, New York, NY, USA. ACM.

Ellen M. Voorhees. 2004. Overview of the trec 2001 question answering track. In *Proceedings of the Thirteenth Text REtreival Conference (TREC 2004)*.