# Developing a Concept-Oriented Search Engine for Isabelle Based on Natural Language: Technical Challenges

Yiannos A. Stathopoulos, Angeliki Koutsoukou-Argyraki, and Lawrence C. Paulson*

Department of Computer Science and Technology, University of Cambridge, UK
[yas23,ak2110,lp15]@cam.ac.uk

The Isabelle libraries and the Archive of Formal Proofs (AFP) contain thousands of formally checked facts (theorems, lemmata, corollaries, propositions, definitions etc.). Current efforts for indexing and searching collections of facts revolve around two approaches. The first approach is mathematical knowledge management (MKM), which involves abstracting mathematical knowledge in the libraries using a semantic markup language, such as OMDoc [6, 2] or a formal meta language, such as MMT [10, 8]. The second approach is *online search* (i.e., searching libraries loaded in the active session in real-time) of Isabelle libraries using symbolic pattern matching of strings. For instance, the Isabelle command `find_theorems` [19] takes a set of criteria (e.g., keywords that must be present in fact names) as input and returns a list of facts that explicitly match these criteria.

In certain cases, `find_theorems` may be limiting for the users. Inexperienced users might have an idea of what kind of material is needed to complete their proof but not enough knowledge of the Isabelle library organisation and naming conventions to construct effective queries for `find_theorems` [7]. This limitation is exacerbated by the fact that new users are more familiar with search interfaces akin to Google's search box: they expect their search query to be a "bag-of-words" describing in natural language the concepts or topic of their enquiry. Furthermore, in response to their query, users expect to be presented with a list of results ordered by relevance.

The aforementioned user expectations are presently not always fulfilled by `find_theorems`. First, it is not straightforward to rank by relevance results produced using strict pattern matching: many facts may match the input criteria exactly. Second, `find_theorems` only matches queries to facts in libraries and theories loaded in the active session. This may be counter-intuitive to new users who might be looking for facts in unloaded theories and are accustomed to searching the entire web in fractions of a second. At the same time, users may not know in which theory the material they are searching for is located; note that classifying mathematical knowledge is non-trivial in principle. Third, as `find_theorems` is based on pattern matching (and is even case-sensitive) it does not find results that are associated conceptually if their names do not exactly match.

We have been investigating a new approach to indexing and searching Isabelle libraries based on natural language. In our approach, each fact is represented by a "bag-of-words" and a set of textual "mathematical concepts" [15, 13] (natural language phrases that refer to mathematical objects, structures and ideas) rather than formal abstractions. Our goal is to develop and evaluate a search engine that (1) enables efficient, *offline search* (search is performed on an index with pre-computed representations so that it does not depend on the loaded theories at each session) of facts in the Isabelle libraries and the AFP; (2) allows Isabelle users to search the libraries using a search box (3) supports "conceptual search" by allowing Isabelle users to search the libraries for desired facts or definitions by describing them using a bag-of-words

*The authors were supported by the ERC Advanced Grant ALEXANDRIA (Project 742178) funded by the European Research Council and led by Professor Lawrence Paulson at the University of Cambridge, UK.

and associated textual mathematical concepts; (4) presents results in order of relevance so that Isabelle users can quickly assess the usefulness of each listed fact or definition. In this presentation we focus on the technical challenges encountered while working towards the above goals and introduce promising solutions.

The first challenge is offline indexing of Isabelle theories. Isabelle users interact with the theorem prover using Isabelle's rich syntax, which includes outer syntax commands, structured Isar proofs and an inner syntax term language [19]. An important step in offline indexing of Isabelle theories is extracting information from the syntax and internal state of the theorem prover. This task is complicated for two reasons. First, it is non-trivial to write an external parser of Isabelle's syntax mainly because the syntax is ambiguous and valid parse trees can only be selected after type-checking [19]. Second, useful information about facts in an Isabelle session, such as types, can only be retrieved from the internal state of the prover, which is not easily achieved using external tools. In order to produce an offline index of the Isabelle libraries we developed an information extraction pipeline for Isabelle. The first stage of our pipeline involves interpreting the PIDE [16, 17] message exchange between Isabelle and jEdit (obtained from `isabelle-dump` [18]). Next, the interpreted messages are transformed into a sequence of tokens representing Isabelle commands. The sequence of tokens is then chunked into constructs such as theorems (and proofs), lemmata and definitions. Our pipeline supports extraction of arbitrary feature sets from Isabelle theories using an interface akin to Map-Reduce [4].

The second challenge is that of automatically modelling mathematical knowledge by assigning concepts (keyword and phrase clouds) to Isabelle facts. Mapping mathematical concepts to Isabelle facts enables linking natural language descriptions of the mathematical knowledge being sought by the user to facts in the Isabelle libraries. Constructing this mapping automatically at scale is challenging because mathematical knowledge in the libraries is almost exclusively expressed in Isabelle's formal language. Our approach is to construct this mapping by linking Isabelle facts to Wikipedia articles that describe mathematical results, structures and objects. We represent each Isabelle fact using two vectors extracted from linked Wikipedia articles. The first representation is a vector of associated words constructed from the body and title of linked Wikipedia articles. The second representation is a vector of associated mathematical concepts discovered in linked articles. We discover mathematical concepts in linked Wikipedia articles using a dictionary of 1.23 million phrases that name mathematical concepts [13]. Searching and ranking facts and definitions using natural language representations enables us to use the Vector Space Model (VSM) [12] to approximate topical similarity to bag-of-words queries. The VSM is an established model of topical similarity in natural language that is known to produce reliable rankings of search results [11, 3].

The third challenge is evaluating the effectiveness of our search engine at retrieving Isabelle facts. The main challenge for evaluation is building a *test collection* for Isabelle search composed of real-life Isabelle queries, complete with expert decisions on which facts in the libraries are relevant to each query (also referred to as *relevance judgements*). In Mathematical information retrieval (MIR), evaluation resources such as the Cambridge University MathIR Test Collection (CUMTC) [14] and the NTCIR math track test collection [1] have facilitated comparisons between systems [5].

We have implemented some promising solutions to the above challenges in the form of a prototype search engine (SErAPIS: Search Engine by the Alexandria Project [9] for ISabelle) and performed a preliminary evaluation. It is our intention to make our search engine publicly available online[1], and procure real-life search queries and relevance judgements from the Isabelle community to produce a resource much like the CUMTC and NTCIR test collections.

[1] The search engine will be available online at `behemoth.cl.cam.ac.uk/serapis`

# References

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

[2] Jonas Betzendahl and Michael Kohlhase. Translating the IMPS Theory Library to MMT/OMDoc. In *Intelligent Computer Mathematics - 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, Proceedings*, pages 7–22, 2018.

[3] Chris Buckley and Ellen M. Voorhees. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 33–40, New York, NY, USA, 2000. ACM.

[4] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004.

[5] Ferruccio Guidi and Claudio Sacerdoti Coen. A Survey on Retrieval of Mathematical Knowledge. *CoRR*, abs/1505.06646, 2015.

[6] Michael Kohlhase. *OMDoc - An Open Markup Format for Mathematical Documents [version 1.2]*, volume 4180 of *Lecture Notes in Computer Science*. Springer, 2006.

[7] Angeliki Koutsoukou-Argyraki. Formalising Mathematics – in Praxis ; A Mathematician's First Experiences with Isabelle/HOL and the Why and How of Getting Started (submitted preprint), 07 2019.

[8] Dennis Müller, Thibault Gauthier, Cezary Kaliszyk, Michael Kohlhase, and Florian Rabe. Classification of Alignments Between Concepts of Formal Mathematical Systems. pages 83–98, 06 2017.

[9] Lawrence Paulson. ALEXANDRIA: Large-Scale Formal Proof for the Working Mathematician. https://www.cl.cam.ac.uk/~lp15/Grants/Alexandria/.

[10] Florian Rabe. The MMT Language and System.

[11] S. E. Robertson. The Probability Ranking Principle in IR. *The Journal of Documentation*, 33:294–304, 1977.

[12] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, November 1975.

[13] Yiannos Stathopoulos, Simon Baker, Marek Rei, and Simone Teufel. Variable Typing: Assigning Meaning to Variables in Mathematical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 303–312, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[14] Yiannos Stathopoulos and Simone Teufel. Retrieval of research-level mathematical information needs: A test collection and technical terminology experiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 334–340, 2015.

[15] Yiannos Stathopoulos and Simone Teufel. Mathematical information retrieval based on type embeddings and query expansion. In *Proceedings of the 26th International Conference on Computational Linguistics, Coling 2016, December 11-16, 2016, Osaka, Japan*, pages 334–340, 2016.

[16] Makarius Wenzel. PIDE as front-end technology for Coq. *CoRR*, abs/1304.6626, 2013.

[17] Makarius Wenzel. Isabelle/PIDE after 10 years of development. In *UITP 2018 (International Workshop on User Interfaces for Theorem Provers 2018).*, 2018. https://sketis.net/wp-content/uploads/2018/08/isabelle-pide-uitp2018.pdf.

[18] Makarius Wenzel. The Isabelle System Manual. 2019. https://isabelle.in.tum.de/doc/system.pdf.

[19] Makarius Wenzel. The Isabelle/Isar Reference Manual. 2019. https://isabelle.in.tum.de/doc/

isar-ref.pdf.