A Horizontal Patent Test Collection

Mihai Lupu
Research Studio Data Science, RSA FG
1090 Wien, Austria
mihai.lupu@researchstudio.at

Alexandros Bampoulidis
Research Studio Data Science, RSA FG
1090 Wien, Austria
alexandros.bampoulidis@researchstudio.at

Luca Papariello
Research Studio Data Science, RSA FG
1090 Wien, Austria
luca.papariello@researchstudio.at

ABSTRACT

We motivate the need for, and describe the contents of a novel patent research collection, publicly available and for free, covering multimodal and multilingual data from six patent authorities. The new patent test collection complements existing patent test collections, which are vertical (one domain or one authority over many years). Instead, the new collection is horizontal: it includes all technical domains from the major patenting authorities over the relatively short time span of two years. In addition to bringing together documents currently scattered across different test collections, the collection provides, for the first time, Korean documents, to complement those from Europe, US, Japan, and China. This new collection can be used on a variety of tasks beyond traditional information retrieval. We exemplify this with a task of high-relevance today: de-anonymisation.

KEYWORDS

datasets, patents, de-anonymisation

ACM Reference Format:
Mihai Lupu, Alexandros Bampoulidis, and Luca Papariello. 2019. A Horizontal Patent Test Collection. In *42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3331184.3331346

1 INTRODUCTION

In the field of natural sciences there exist—surely debatable, but still universal—conventions that allow scientists to consistently compare results (e.g. we have an agreed-upon unit of mass [2]). Instead, in fields such as information retrieval (IR) and extraction (IE), natural language processing (NLP), machine learning and translation, etc. we do not have an acknowledged unit of effectiveness.

The current state of the art consists in a statistical assessment of whether a given system A is better than another system B. This requires consistency, with the task and the content on which the two systems are compared remaining unaffected [8]. To put it simply, we should not expect our preference for A versus B to be maintained if we challenge the two systems on a different subject.

All of this wants to show that while we have metrics, we still need a unit to measure them against—this is the purpose of the test collection. For search, numerous test collections have been created in the context of evaluation campaigns [10]. In this context, patent *search* test collections have been created at TREC [5], CLEF [7], and NTCIR [4]. These are search test collections because they were developed with the primary purpose of evaluating search engines. Additional tasks (e.g. image recognition and topic mining) were explored in some of the evaluation campaigns, but their development was secondary to the search considerations. The nature of the patent test collections, when created for search, is vertical (considering time as a vertical axis): there is a filter on either patent authority (which government has issued the patents) or on the technical domain (Figure 1). For instance, the CLEF-IP collections filtered on patents of the European Patent Office, the NTCIR collections filtered on US and Japanese patents, while TREC-CHEM filtered on patents with chemistry-related inventions.

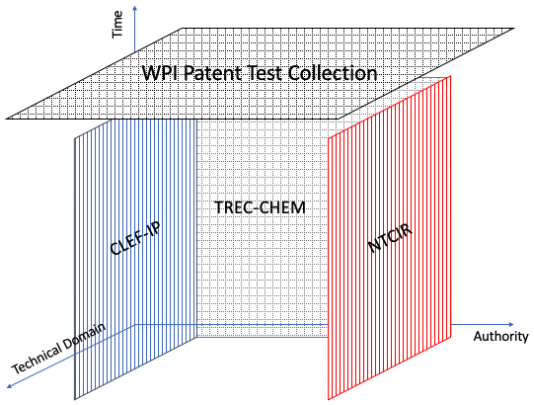

Figure 1: The new collection complements existing patent test collections, providing a horizontal coverage of the global patent corpus

To address additional tasks in information retrieval (e.g. link and content analysis; Clustering, classification, and topic models; or even information privacy and security) we need a collection that covers all patents, across authorities and domains. At the same time, we need to keep the collection fixed and under manageable size. While theoretically all patent data is publicly available, this data is published by different providers differently, both in terms of content and of coverage. For instance, a research paper stating "We tested on US patents from 2011 to 2014" is not reproducible because it is unclear what "US patents" means (i.e. patent applications or patent grants, including or excluding utility models[1]) and the dates (even assuming `2011` refers to `2011-01-01 00:00:00.0000`) are unclear because they do not specify whether they are application, publication, or grant dates.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331346

[1]Utility models are sometimes referred to "petty patents" or "innovation patents" https://www.wipo.int/sme/en/ip_business/utility_models/utility_models.htm

The WPI[2] Test Collection presented in this article covers patents from all major authorities: European Patent Office (EP), United States Patent and Trademark Office (US), World Intellectual Property Organisation (WO), Chinese Patent Office (CN), Japan Patent Office (JP), and Korean Patent Office (KR). It contains full bibliographic information, full text for all documents, as well as all related images and additional material.

In the remainder, we structured the article as follows: In Sec. 2 we describe how the proposed collection fits into the existing test collections landscape. The content of the WPI Test Collection is presented in Sec. 3 and some of the key aspects are revealed. Section 4 is devoted to an application of current interest: de-anonymisation. We outline here some of the de-anonymisation challenges that can be undertaken with the help of the test collection. Lastly, in Sec. 5 we summarise the main aspects of the new collection and we provide a short outlook on its future possible applications.

2 RELATION TO OTHER COLLECTIONS

From the early 2000s, a number of new test collections have been created to address patent-related tasks. Each of them has a peculiar objective, which we briefly describe in what follows together with a comment on how the WPI Test Collection differs from and complements the existing resources.

We start our discussion on previous works with the series of test collections generated at NTCIR in Japan [4], as they are the pioneers in the field. Starting in 2001/2002 and over a time period of 12 years, 8 test collections have been produced at NTCIR with an IR task. Two new patent-related evaluations have been introduced in two other international conferences: one in Europe, organised at CLEF and called CLEF-IP [7] and the other in the US, organised at NIST and called TREC-CHEM [5]. All three start with a retrieval task and later diversified into a series of other tasks: translation, classification, text mining, image-based retrieval, and image analysis.

While it is true that these test collections have different characteristics, they all share a peculiar aspect: they all are *vertical* collections. They provide, over a large period, a set of documents filtered in some way. NTCIR filters on Japanese and US patents, while CLEF-IP on EPO patents, and TREC-CHEM on chemistry-related patents. On the contrary, the new test collection is *horizontal*. It includes all technical domains from the major patenting authorities over the relatively short time span of two years. Table 1 shows a brief comparative summary of the new test collection compared to existing ones. It should be noted that the statistics of the existing test collections vary over the years - we have taken the largest values for this summary.

[2]The collection, available on Zenodo as a protected collection, is found at the following link: https://doi.org/10.5281/zenodo.1489994.

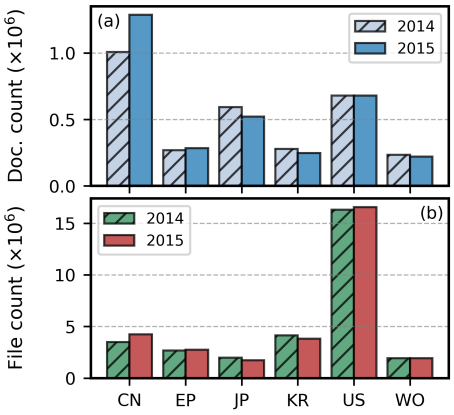

Figure 2: Total number of (a) documents and (b) files for the years 2014 (striped bars) and 2015 (solid bars).

The new collection therefore fills the empty space left behind by the previous collections allowing cross-authority and cross-domain analysis. In Section 4, we provide an example of how this could be used for a new task: de-anonymisation, given the availability of multiple documents from the same inventors across authorities.

3 DATA CONTENT

The content of the WPI Test Collection consists of all patents from 6 authorities over two years. In total there are 6,313,165 patent documents (XML files) and 55,231,022 additional files (images, chemical structures, etc.). In this first version of the collection, we omitted the PDF files due to the very large amount of data – compressed, they exceed 5TB.

In what follows, a *document* refers to a file describing a patent (e.g. EP-1234567-A1.xml). They are in an extended ST-36 format [13]. A *file* denotes instead any file present in the collection, either a document, or auxiliary materials.

Figure 2 shows the number of (a) documents and (b) files filed by each authority in 2014 and 2015. Clear is the dominance of China and the US concerning the volume of documents and files, respectively. The solid predominance of the US authority can be brought back to the practice of the USPTO to issue more auxiliary material.

Figure 2 also shows that, apart from the Chinese authority that has published about 27% more documents in 2015, there is no substantial difference between the two years. In fact, 30,504,423 files have been filed in 2014 and 31,039,764 in 2015, which corresponds to an increase of about 1.75%. For the documents, the increase amounts

	WPI	**CLEF-IP**	**TREC-CHEM**	**NTCIR**
Authorities	EP, JP, CN, KR, US, and WO	EP (WO[3])	EP,WO,US	JP, US, CN
Document types	Applications and grants, no utility models			
Time period	Publication years 2014 and 2015	until 2007	until 2007	1993-2003
Language	AR, DE, EN, ES, FR, JA, JO, PT, RU, ZH	DE, EN, FR	EN	JP, EN, ZH
Beyond text	All referenced images or files	text-only	all referenced images or files	text-only

[3]only part of WO documents, corresponding to a specific case of EP documents

Table 1: The WPI patent collection factsheet, compared with existing patent test collections.

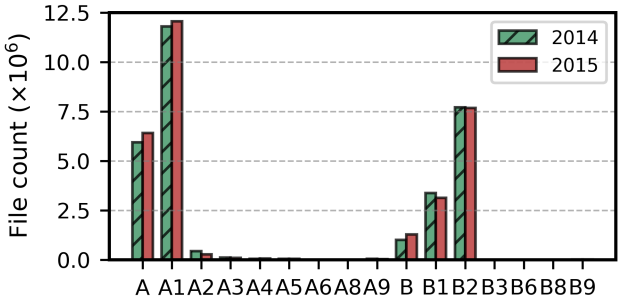

Figure 3: Total number of files per kind-code and year (green striped bar for 2014 and red solid bars for 2015).

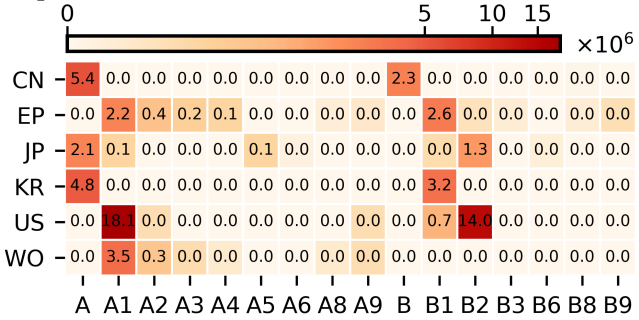

Figure 4: Total number of files per kind-code and country. Values across disparate orders of magnitude are compensated by a power-law scale x^γ, with $\gamma = 1/4$.

to 5.8% and goes back to the Chinese step forward. The authorities CN, EP, JP, KR, and WO produce a similar amount of files, ranging from 1.7 to 4.2 million files per year. In contrast, the US authority produces the vast majority of files, reaching ~16 million per year.

The distribution of the number of files exhibits a great inhomogeneity over the kind codes[4] (Fig. 3). Out of the 16 existing kind codes, only 5 contribute in a substantial way. The kind codes A1 (~12 million files per year), B2 (~7.7), A (~6), B1 (~3), and B (~1) gather indeed ~98% of the files.

Figure 4 provides further details about the distribution of the number files over the different kind codes and authorities. The trend previously observed, establishing the US authority as the richest source (see Fig. 2(b)) and the kind codes A1 and B2 as the most populated (see Fig. 3), is here clearly explained: 97.7% of the files published by them belong to either one of these two kind codes.

A feature shared by all the patenting authorities concerns the most widely spread file extension: *TIF*. This file type is used for images (drawings, chemical structures, etc.) and reaches 39,711,852 files. Figure 5 shows indeed that the third column from the right is largely populated by all authorities. The last column, corresponding to the *XML* file extension, follows with 6,350,664 files. This is expected because there is at least one such file per patent. The US and Korean authorities are the only ones with a substantial number of chemical structure files, of *CDX* and *MOL* type, and *JPG* type, respectively.

As one would expect, the two languages that prevail in the test collection are English (EN) and Chinese (ZH), followed by Japanese (JA) and Korean (KR) (Fig. 6). This consistently holds true for all the

[4]The kind code broadly indicates the phase of the granting process in which the document was published. A typically represents application and B granted patents.

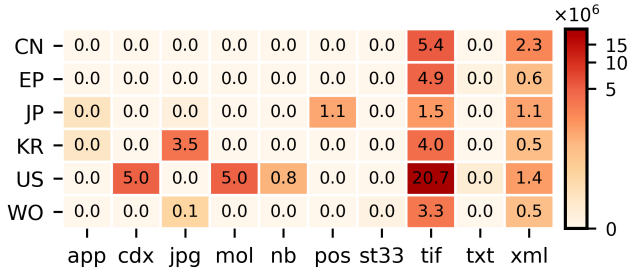

Figure 5: Total number of files per file extension and country. Values across disparate orders of magnitude are compensated by a power-law scale x^γ, with $\gamma = 1/4$.

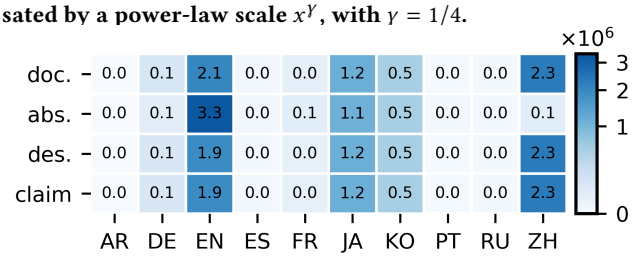

Figure 6: Languages in the corpus, divided by patent section. "Doc." is the language for the bibliographic data, while "abs.", "des.", and "claim" to the languages of each section. Values across disparate orders of magnitude are compensated by a power-law scale x^γ, with $\gamma = 1/2$.

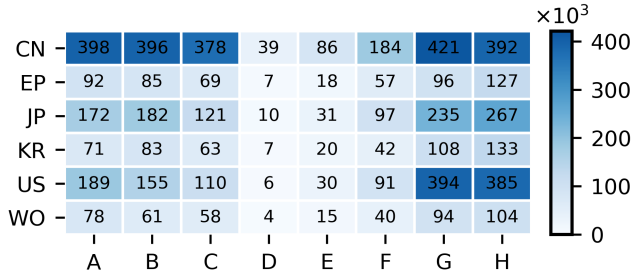

Figure 7: Number of documents per patenting authority and IPC section.

sections of the documents, that is abstract, description, and claims. It is worth mentioning the fact that, while the numbers in Fig. 6 are rather constant across the sections of the documents, for the abstract there is a strong dominance of English, motivated by the publication of English abstracts in China and Japan.

Lastly, in Figure 7 we show the number of documents per country and International Patent Classification (IPC) [12] section. The latter, labelled by letters ranging from A to H, gives a broad indication of the patent content. Confirming the trend observed in Fig. 2, China clearly stands out in almost every IPC section. The US authority is the only one that can compete in the sections G and H (Physics and Electricity, respectively). What all the patenting authorities have in common, is a large number of documents in sections A-C and F-H and much less documents in sections D and E (Textiles-Paper and Fixed Constructions, respectively).

4 DE-ANONYMISATION USE CASE

To provide an example of how this collection can be used, we chose a relatively unusual task in our community: de-anonymisation. We

do this because we observe that, while there exists some academic interest, which we will list below, and even more public interest, there are no test collections offering, on one hand, the richness of information that a patent collection with world-wide coverage offers, and, on the other hand, explicit voluntary personal information offered by, in this case, the owners of the intellectual property rights described by patents.

De-anonymisation is the process of identifying individuals in an anonymised dataset and it has become a task of high relevance today due to the vast availability of datasets that can be used as an aid in de-anonymising anonymised datasets, and due to the privacy concerns that have been raised worldwide in the recent years and which led to the establishment of regulations such as the GDPR[5].

One de-anonymisation task that has been studied, not extensively however, that casts a doubt on the anonymity of the double-blind review process in academia, is the *author de-anonymisation* of scientific publications. The de-anonymisation of authors relies on their citation behaviour [1, 3], writing style [9], the content of the paper [6], and a combination of the aforementioned [6]. To the best of our knowledge, there is no publicly available test collection, that contains both citations and full text, and spans over different fields, on which prior research may be applied. Hill and Provost [3] used the KDD Cup 2003 dataset which contains both citations and full text, but is limited to the field of physics, Bradley et al. [1] used the CiteSeerX data which contains only citations, and Payer et al. [6] and Sarwar et al. [9] crawled and parsed their datasets themselves.

Due to the similarity between the nature of the description section of patents and scientific publications [11], this test collection can serve as a test collection for the task of author de-anonymisation of scientific publications. All methods applied in the literature of this task may be applied to patents as well, but instead of authors, patents have inventors and assignees. The test collection itself may be split into training and test set, with the test set containing patents whose information about inventors and assignees has been removed. Alternatively, the whole test collection may serve as a test set, with the training set being patents from other resources such as the Open Patent Services (OPS)[6] or the existing collections mentioned earlier in this paper. Specifically, the methods of the literature may be applied in the following ways:

Citation behaviour: In order to apply methods that exploit the citation behaviour of authors, the identifiers of cited patents and the names of the inventors and assignees may be extracted from the fields *patent-citations*, *inventors* and *assignees*, respectively. This information may be used to build a vector-space model or a citation graph and have various methods applied to them: Attributing authorship to the most cited author of an anonymous paper [6], matching by cosine similarity of citation vectors [1, 3, 6], Latent Semantic Analysis with authors and papers as terms and documents, respectively [1].

Writing style: Following Payer's et al. [6] and Sarwar's et al. [9] approach and since the description of a patent is the part of a patent that corresponds the most to a scientific publication [11], stylometric features, such as average word length and frequency of punctuation, may be extracted from the field *description*. These features may be used as in [6] for training a Support Vector Machine (SVM), an ML-kNN classifier and calculating cosine similarities, or as in [9] for building a graph linking stylistically similar fragments of different documents.

Content: Following Payer's et al. [6] approach, the bag-of-words of each patent's description may be created and then have the same methods as for the writing style, as in [6], applied.

Finally, the evaluation methods and measures of the literature can be applied to this test collection as well: Success rate in predicting at least one or all authors of a paper within top-K predictions [1, 3, 6, 9], or the number of correctly predicted authors of a paper divided by its total number of authors [9].

5 CONCLUSIONS

In this article, we motivated the need for a novel patent research collection, and presented the WPI Patent Test Collection, which contains data from six of the major patent authorities. We have provided an overview of the collection by discussing some of its key aspects. We expect that this collection will stimulate information retrieval and data mining research further, as it complements existing test collections by providing complete coverage over authorities and domains, albeit for a reduced time-period.

The field of patent applications is still very prosperous, especially for what concerns the Asian languages, and this is the first collection to provide comparable corpora for all major east-Asian languages, including Korean. Apart from the traditional and well-known tasks of information retrieval, we consider a concrete task of current interest, namely de-anonymisation and argue that, in addition to all the traditional tasks of relevance to the IR community, this task can also be investigated under the same solid empirical framework we are accustomed to.

Acknowledgements. The authors received funding from the EU's Horizon 2020 research and innovation programme under grant agreement No 825225, and from Project Data Market Austria, funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under program "ICT of the Future".

[5] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[6] https://www.epo.org/searching-for-patents/data/web-services/ops.html

REFERENCES

[1] J. Bradley, P. Kelley, and A. Roth. 2008. Author identification from citations. *Dept. Comput. Sci., CMU, Technical Report* (2008).
[2] R. Davis. 2003. The SI unit of mass. *Metrologia* 40, 6 (2003).
[3] S. Hill and F. Provost. 2003. The Myth of the Double-blind Review?: Author Identification Using Only Citations. *SIGKDD Explor. Newsl.* 5, 2 (2003).
[4] M. Lupu, A. Fujii, D. Oard, M. Iwayama, and N. Kando. 2017. *Patent-Related Tasks at NTCIR.* Springer.
[5] M. Lupu, J. Huang, J. Zhu, and J. Tait. 2011. TREC chemical information retrieval - An initial evaluation effort for chemical IR systems. *WPI Journal* 33, 3 (2011).
[6] M. Payer, L. Huang, N. Z. Gong, K. Borgolte, and M. Frank. 2015. What You Submit Is Who You Are: A Multimodal Approach for Deanonymizing Scientific Publications. *IEEE Transactions on Information Forensics and Security* 10, 1 (2015).
[7] F. Piroi and A. Hanbury. 2017. *Evaluating Information Retrieval Systems on European Patent Data: The CLEF-IP Campaign.* Springer.
[8] M. Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010).
[9] R. Sarwar, C. Yu, S. Nutanong, N. Urailertprasert, N. Vannaboot, and T. Rakthanmanon. 2018. A Scalable Framework for Stylometric Analysis of Multi-author Documents. In *Database Systems for Advanced Applications.*
[10] E. M. Voorhees and A. Ellis. 2019. The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018). In *TREC Proceedings.*
[11] R. Walker. 1995. *Patents as scientific and technical literature.* Scarecrow Press.
[12] World Intellectual Property Organisation. 2019. International Patent Classification. Online. https://www.wipo.int/classifications/ipc/en/
[13] World Intellectual Property Organisation. 2019. WIPO XML Standard ST36. Online. http://www.wipo.int/standards/en/xml_material/st36/