2-years after Reliance on Science:

Discovering patent-to-article citing sentences by supervised classification

Mei Yun Lai (lai@uni-bremen.de), University of Bremen, Institute of Project Management and Innovation (IPMI)
Thomas Pastuska (thomas.m.pastuska@student.hs-anhalt.de), Anhalt University of Applied Sciences, Department Computer Science and Languages
Korinna Bade (korinna.bade@hs-anhalt.de), Anhalt University of Applied Sciences, Department Computer Science and Languages
Martin G. Moehrle (moehrle@uni-bremen.de) University of Bremen, Institute of Project Management and Innovation (IPMI)

Extended Abstract (2-pages)

Scientific discoveries have led to many great inventions for societal and economic growth. Arguably, inventors use science as a "map" to search in unfamiliar or difficult knowledge domain, thus directing them to exploit more useful combinations and to eliminate less-promising research efforts [1]. Tracing how an invention is linked to scientific knowledge helps researchers to understand how much inventors rely on scientific knowledge to build follow-on research and innovation. Today, two ways are known for this purpose. The first way is based on front-page citations, namely on articles explicitly cited on the front-page of patents. Such citations from patent to scientific articles proxy better on scientific knowledge in inventions, rather than citations to prior-art patents. The second way is based on in-text citations [2]. There is a remarkable discrepancy between front-page and in-text citations. The front-page citations serve mostly legal purposes by patent attorneys to set claim boundary to scientific prior-art knowledge. Contrarily, in-text citations within patent specification sections are provided by inventors as supporting evidence to enable their claims [3]. Using a hybrid approach from machine learning methods [4] and manual heuristics, [5,6] publish an open-access data of 16.8 million in-text citations from worldwide patents.

These studies and the provided open-access data help trace scientific knowledge-invention linkages. However, they examine citations by dummy variables (i.e. cite=1; not cite=0) and differentiate these "in-text" citations from the regular front-page citations solely by statistics. Doing so, they miss the

opportunity of qualitative, semantically contextual reasoning for temporally, geographical and topical diverse; or less self-referential characteristics of patent citations. This opportunity bridges the existing understanding gap particularly, in which way a patent claim is supported by context of different "intext" citation within patents. The context extraction from such "in-text" citations requires the identification of the exact citing sentences (following [7], henceforth: citances) within patents, which refer using "in-text" citations. These in-text citations are embedded in sentences and paragraphs, often formatted in non-standardized citation styles, as contrasted with the established citation formats in scientific articles (e.g. IEEE, APA, MLA). Since there is still no viable automated approach, we attempt to leverage supervised classification techniques to solve this challenge. We ask, how could we use the open-access bibliometric data, to discover each patent-to-article citance? In which way does the disclosed scientific knowledge from the cited articles relate to the patent claim-supporting context of each citance?

Given our domain expertise in natural language processing software, we focus in this technological area and start with a pool of 604 USPTO-patents, adapted from [8]. Then, we proceed in three steps. First, we extract all textual information from the patent specifications, which contain in-text patentto-article citations. Second, following [9], we assemble a preliminary feature-based Support Vector Machine (SVM) to distinguish citances into two classes: 1) with supporting context and 2) without supporting context (c.f. [10]). Third, we customize a preliminary SVM-classifier by including novel patent-specific features (both lexical and semantic) to improve performance. We then, re-train the SVM-classifier, using manually annotated citances extracted based on in-text citations in patents. We successfully extract citances from 3914 in-text citations and distinguish the extracted citances with supporting contexts from those without supporting contexts. Through the integration of nine novel patent-specific features, the performance of SVM-classifier is improved and resulting to an increment of 20%, in terms of precision and F1-score.

Our work contributes to the research and practice in two ways. First, using our method and data, techminers can differentiate between patent-to-article citances and integrate the measured nuances

of contextual information into their existing studies (e.g. bibliographic or network analysis) , which previously evaluated by simple absolute citation count. Second, the supporting evidence extracted from citances help support legal scholars and practitioners alike to examine the patentability based on the requirement of written description and enablement.

References

[1] L. Fleming and O. Sorenson (2004), "Science as a map in technological search", Strat. Mgmt. J., 25: 909-928. https://doi.org/10.1002/smj.384

[2] M. Roach and W. M. Cohen (2012) "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research", Management Science 59(2):504-525. https://doi.org/10.1287/mnsc.1120.1644

[3] B. K. A., Ozcan and B. Sampat (2020), "In-text patent citations: A user's guide", Research Policy 49(4). https://doi.org/10.1016/j.respol.2020.10394

[4] C. Verluise, G. Cristelli, K. Higham and G. de Rassenfosse (2020), "The Missing 15 Percent of Patent Citations", SSRN Electronic Journal. http://dx.doi.org/10.2139/ssrn.3754772

[5] M. Marx and A. Fuegi, (2021), "Reliance on science: Worldwide front-page patent citations to scientific articles", Strat. Mgmt. J., 41: 1572-1594. https://doi.org/10.1002/smj.3145

[6] M. Marx and A. Fuegi, (2022), "Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations", Journal of Economics & Management Strategy, 31, 369– 392. https://doi.org/10.1111/jems.12455

[7] P. Nakov, S. S. Ariel and A. H. Marti (2004), "Citances: Citation Sentences for Semantic Analysis of Bioscience Text", https://biotext.berkeley.edu/papers/citances-nlpbio04.pdf

[8] B. S. Haney (2020), "Patents for nlp software: An empirical review", The IUP Journal of Knowledge Management 18(4), 27–58. http://dx.doi.org/10.2139/ssrn.3594515

[9] A. Athar and S. Teufel (2012), "Context-enhanced citation sentiment detection", Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies pp. 597–601. https://aclanthology.org/N12-1073

[10] J. Freilich and Soomi K. (2021), “Is the Patent System Sensitive to Information. Quality?”, working paper. https://sites.bu.edu/tpri/files/2021/05/Freilich_Kim_Information-Quality.pdf