From: Andrew Chin [email redacted]
Sent: Saturday, May 16, 2015 1:30 AM
To: WorldClassPatentQuality
Subject: Comment on Pillar 1 - Proposal 2

Please see attached comment.

--

Andrew Chin
Associate Professor
University of North Carolina School of Law
160 Ridge Road, CB #3380
Chapel Hill, NC 27599-3380
AndrewChin. com

Adding to PLUS:
The Promise of Automated Pre-Examination Search

Andrew Chin[*]
University of North Carolina School of Law

Pillar 1, Proposal 2 relates to the U.S. Patent and Trademark Office's development of automated search tools for identifying potentially relevant prior art.[1] The USPTO's Scientific and Technical Information Center currently uses a search tool called the Patent Linguistic Utility Service to provide automated pre-examination searches to examiners upon request.[2] This search consists of two steps: (1) PLUS runs "an algorithm to analyze an application for the presence of frequently-used terms"; and (2) STIC locates and highlights those terms in a database of prior art U.S. patents and U.S. patent application publications.[3]

The USPTO is requesting comments regarding three possible extensions to its automated search practices. First, the office is considering performing an automated pre-examination search in all applications, not just when requested by the examiner.[4] Second, the results of this search may be provided to the applicant and thereby become part of the prosecution history.[5] Finally, the office is evaluating advanced technologies to support a new "custom extraction routine that enables keyword, stemming, concept-semantic, and relational word searching capabilities" and "more modern natural language search queries."[6]

All three of these changes should be encouraged. Examiners have already been heavily relying on keyword searches of the USPTO's full-text patent and patent application databases.[7] To the extent that general strategies for formulating an initial keyword search are amenable to algorithmic implementation or machine learning, automation may be expected to save time and reduce error in examiners' prior art searches.

Institutionalizing an automated pre-examination search may also promote the responsiveness of the USPTO in prosecution. The results of the search could be delivered to the applicant electronically, thereby providing nearly immediate notice of at least some of the references that the examiner will consider. This information, while not comprehensive, may give the applicant early notice of a need to clarify and amend the application — and the

[*] Associate Professor, University of North Carolina School of Law; J.D., Yale; D.Phil. (Computer Science), University of Oxford.

[1] U.S. Patent & Trademark Office, Requests for Comments on Enhancing Patent Quality, 80 Fed. Reg. 6475, 6479 (Feb. 5, 2015).

[2] Id.

[3] Id.

[4] U.S. Patent & Trademark Office, Pillar 1 — Proposal 2: Automated Pre-Examination Search <http: // www .uspto.g ov/sites/default/files/documents/Proposal%202%20final_1.pdf>.

[5] Id.

[6] 80 Fed. Reg. at 6479.

[7] Andrew Chin, *Search for Tomorrow: Some Side Effects of Patent Office Automation*, 87 N.C. L. REV. 1617, 1641 (2009) <http :// www .unclaw. com/chin/scholarship/searchfortomorrow.pdf>; *see also* text accompanying note 11 *infra*.

flexibility to begin such work before the first office action starts the clock.[8] Automated search results may also challenge the applicant's perspective regarding the scope of analogous prior art, impelling more and earlier information disclosures.[9] The benefits of automated search can thereby cascade throughout the prosecution process, promoting and expediting the generation of information material to patentability and improving the quality of the issued patent.

An automated search technology will be able to deliver the above benefits to examiners and applicants only to the extent that it can generate a set of references resembling those actually found and considered by an examiner. The results of this author's longitudinal study of USPTO automation, based on all 3,266,297 patents in the USPTO's PatFT database as of May 1, 2007,[10] provide strong empirical support for the following technological evaluations and recommendations.

1. Examiners are increasingly relying on keyword search. The proportion of citations imputed to keyword search rose consistently between 1990 and 2007, with the sharpest increases accompanying the introduction of desktop search tools for examiners in 1999-2000, the expansion of the databases to include optically scanned pre-1970 patents in 2001, and the elimination of the paper patent collection in 2005.[11] Reliance on keyword search appeared to be heaviest in medicine and chemistry, and lightest in physics and energy.[12] In this respect, PLUS's reliance on "the presence of … terms" (i.e., keywords) reflects the actual practice of examiners in most art fields and might therefore be expected to produce similar results. As the paragraphs 2 and 3 discuss, however, PLUS's focus on frequently-used terms appears to be fundamentally at odds with the aim of emulating contemporary examiner search results.

2. Examiners are increasingly relying on low-frequency keywords. The proportion of citations attributed to searches on keywords with fewer than 50 hits in the PatFT database grew much more quickly between 1990 and 2007 than the corresponding statistic for keywords with 51 to 500 hits.[13] In this respect, the actual practice of examiners is trending away from PLUS's focus on "the presence of *frequently-used* terms."

3. Focusing on high-frequency terms does not help PLUS produce results that resemble examiners' cited references. High-frequency keyword search terms do not offer significantly more distinguishing power than low-frequency terms for identifying the patents actually cited by the examiner from the PatFT database. For a given *n* between 2 and 500, the expected information content of a keyword search result consisting of *n* hits remains nearly constant, in a range between 19 and 23 bits.[14] PLUS may have been designed to use high-frequency keywords in order to guarantee a substantial number of search results, but these results provide no more information regarding examiners' actual practices than do the results of low-frequency keyword searches.

[8] *See* 37 C.F.R. § 1.134.
[9] *See* 37 C.F.R. § 1.56.
[10] *See* Chin, *supra* note 7, at 1636.
[11] *See id.* at 1641.
[12] *See id.* at 1645.
[13] *See id.* at 1642 & 1650 ("the low-frequency keyword set accounts for most of the observed increase").
[14] *See id.* at 1649-50 ("The study found that the search engine results for higher-frequency keywords contain on average only slightly more information than could be obtained from search engine results for lower-frequency keywords.").

4. Overreliance on keyword search tends to sacrifice precision for recall. Efficient information retrieval requires both high *recall* and high *precision*. Recall refers to the fraction of relevant items that are retrieved;[15] precision refers to the fraction of retrieved items that are relevant.[16] The information retrieval literature has recognized an empirical tradeoff between recall and precision in both human and automated search.[17] As paragraph 3 concludes, PLUS's use of high-frequency keywords is an example of a system optimized for high recall at the expense of precision. In light of examiners' growing reliance on keyword search, this is broadly consistent with the study's finding that the PTO classification system's recall was higher (and precision lower) for search results generated through keyword search than for results generated through other search methods.[18]

5. Auxiliary search technologies should target linguistic imprecision. Unlike PLUS, examiners need not rely exclusively on keyword search. To date, however, the tools available to examiners for improving the precision of their search queries have been limited in their effectiveness. In some cases, search terms from the USPTO classification system can helpfully disambiguate keywords whose usage patterns do not conform to the boundaries of art fields.[19] As a more general matter, however, the study found that citations imputed to keyword search are more frequently co-classified than citations imputed to other search methods, demonstrating that the classification system's power is relatively attenuated when it comes to discriminating among keyword search results.[20] The proposed technologies, particularly concept-semantic and natural language search, are promising avenues for advances that might directly address the inherent imprecision of keyword search.

6. Auxiliary search technologies should harness the vast information within the existing citation network. The study found that forward and backward citation tracking — once an essential tool in a USPTO without keyword search — maintained a stable presence in examiners' search practices throughout the transition from paper file drawers ("shoes") to searchable databases.[21] Given the robustness of even this rudimentary use of the patent citation network, citation network analysis should be recognized as an essential and potentially powerful element of any automated search technology aiming to emulate examiner search results. Citation network data can — and should — be fully incorporated in the proposed relational search technologies to improve both recall and precision in search results.[22]

[15] *See id.* at 1632.

[16] *See id.* at 1633.

[17] *See id.* at 1634-35.

[18] *See id.* at 1646, 1650-51.

[19] *See id.* at 1633 (giving the example of cell phone holders, cell phone shields, cell phone mice, terrorist cells, jail cells, electrolytic cells, the Sony Cell microprocessor, and stem cells).

[20] *See id.* at 1645-46.

[21] *See id.* at 1644.

[22] *See, e.g.*, Parvaz Mahdabi & Fabio Crestani, *Patent Query Formulation by Synthesizing Multiple Sources of Relevance Evidence*, 32 ACM TRANS. ON INFO. SYS. 16:1 (2014); Atsushi Fujii, *Enhancing Patent Retrieval by Citation Analysis*, PROC. OF THE 30TH ANNUAL INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 793 (2007) <http: //if-lab .slis. tsukuba .ac. jp/fujii/paper/sigir2007.pdf>.