SEARCH FOR TOMORROW: SOME SIDE EFFECTS OF PATENT OFFICE AUTOMATION*

ANDREW CHIN**

The United States Patent and Trademark Office's ("Patent Office") move to a paperless search facility and the public's growing involvement in prior art search have recently elevated the role of search engine technology in the patent examination process. This Article reports on an empirical study that examines how this technology has systematically changed not only how patent references are found, but also which patents are cited as prior art.

Publicly available records do not provide information identifying the method by which each of the prior art references cited by a patent was found, such as keyword search, citation tracking, or classification search. The main methodological contribution of this Article is to identify large sets of patent citations that are likely to exhibit characteristics similar to those of citations actually found through a particular search technique. By applying this synthetic approach to a comprehensive citation database, this study compiles a large set of patent citations that can reasonably be imputed to keyword search.

A longitudinal analysis of this imputed data set indicates that examiners became increasingly reliant on keyword full-text search in the late 1990s, as the technology became accessible from their desktop computers. This change in examination practice appears to have had a substantive effect on the choice of patents to be cited as prior art. Specifically, patent citations imputed to keyword search tend to be co-classified (according to the Patent Office classification system) more frequently than patent citations in general and patent citations imputed to citation tracking methods.

** Associate Professor of Law, University of North Carolina School of Law. The author thanks Tony Biller and Kathy Strandburg for helpful suggestions, and Dan Blanchette, Allison Dobson, and Matthew Ruedy for research assistance.

These findings support the concerns of some commentators about Patent Office automation and the outsourcing of prior art search. In particular, it appears that the Patent Office classification system is not being fully utilized to improve the precision of search results. This Article concludes with a survey of some initiatives and techniques that have recently emerged to address this problem.

INTRODUCTION

The increasingly prominent role of innovation in the economy has focused considerable public attention on substantive questions of patentability in recent years.[1] At the same time, the Patent Office's full-text patent database and World Wide Web search engines have greatly extended the public's ability to conduct prior art searches and to draw their own inferences regarding the validity of millions of issued patents and published patent applications.[2] The Patent Office has accommodated these developments with procedural changes that offer unprecedented opportunities for patent applicants and the general public to participate in the preexamination search for prior art.[3] With a world of prior art only a click away, the public is poised to engage the patent system and to challenge the comparative advantage of patent examiners as never before.[4]

The popularization of prior art search has coincided with the emergence of full-text keyword querying as the dominant search methodology. The Patent Office recently replaced most of its

1. *See, e.g.*, *American Innovation at Risk: The Case for Patent Reform*, *Hearing Before the Subcomm. on Courts, the Internet, and Intellectual Property of the H. Comm. on the Judiciary*, 110th Cong. (2007) (hearing public comments in connection with the 2007 Patent Reform Act); ADAM B. JAFFE & JOSH LERNER, INNOVATION AND ITS DISCONTENTS: HOW OUR BROKEN PATENT SYSTEM IS ENDANGERING INNOVATION AND PROGRESS, AND WHAT TO DO ABOUT IT 25–55 (2004) (describing current criticisms of the patent system and proposals for reform in historical perspective); NATIONAL RESEARCH COUNCIL, A PATENT SYSTEM FOR THE 21ST CENTURY (Stephen A. Merrill et al. eds., 2004) (reviewing extensive public comments on evaluating and improving the performance of the patent system).

2. *See infra* Part V.A.

3. *See infra* Part I.B.

4. As recently as 1992,

> the majority of patentability searches [were] conducted by either a patent lawyer or professional searcher manually searching U.S. patents in the public search room in the U.S. Patent Office, located in Crystal City, Virginia, or by use of a computer terminal connected by telephone and modem to one or more proprietary databases.

Louis J. Knobbe, *How to Decide Whether to Obtain a Patent: Legal Framework*, 343 PLI/PAT 9, 25 (1992).

venerable categorized paper file drawers ("shoes")[5] with dedicated search terminals and Web browsers.[6] This move not only represents an important milestone in the agency's transition to a paperless examination system, but also an institutional expectation that examiners, applicants, and the public henceforth will prefer to find prior art references primarily through computer-aided searching of patent documents.

Search engine technology is rapidly taking center stage as the common denominator in the search for prior art by an increasingly diverse set of actors. It is therefore worthwhile to pause at this juncture to examine the ways in which keyword search might be changing not only *how* prior art is found, but *which* prior art is found. While applicants are under a duty to disclose any prior art known to be material to patentability,[7] and examiners are expected to conduct a thorough prior art search,[8] both operate under time and other resource constraints that make it difficult to guarantee the adequacy of the cited prior art for analyzing patentability.[9] Whether search

5. The origin of the equally venerated term "shoes" remains the subject of speculation and dispute, but may be associated with the Patent Office's purchase of "shoe drawers" from Augustus Burgdof in 1879. *See* KENNETH W. DOBYNS, THE PATENT OFFICE PONY: A HISTORY OF THE EARLY PATENT OFFICE 193 (1997).

6. *See* PATENT INFORMATION USERS GROUP, 2005 ANNUAL CONFERENCE REPORT 4 (2005), http://depts.washington.edu/englib/eld/liaisons/piug2005.doc (reporting that the Patent Office's new Public Search Facility in Alexandria "has approximately 300 public workstations that provide access to USPTO internal patent and trademark search systems" and that "[t]he paper collection of classified patents was discarded in 2003–2004"); *see also* U.S. PATENT & TRADEMARK OFFICE, PERFORMANCE AND ACCOUNTABILITY REPORT FOR FISCAL YEAR 2004, at 23 (2004), http://www.uspto.gov/web/offices/com/annual/2004/2004annualreport.pdf (illustrating the new public search facility).

7. *See* 37 C.F.R. § 1.56(a) (2008) ("Each individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability . . . .").

8. *See* 37 C.F.R. § 1.104(a)(1) ("On taking up an application for examination or a patent in a reexamination proceeding, the examiner shall make a thorough study thereof and shall make a thorough investigation of the available prior art relating to the subject matter of the claimed invention.").

9. *See* Mark A. Lemley, *Rational Ignorance at the Patent Office*, 95 NW. U. L. REV. 1495, 1496 n.3 (2001).

> Examiners have astonishingly little time to spend on each application—on average, a total of eighteen hours, including the time spent reading the application, reading the submitted prior art, searching for and reading prior art in databases accessible to the PTO, comparing that prior art to the application, writing an office action, reading and responding to the response to office action, iterating the last two steps at least one and often more times, conducting an interview with the applicant, and ensuring that the diagrams and claims are in form for allowance.

technology is to play an effective role in alleviating these constraints will ultimately depend on whether all parties are able to use the technology to conduct a thorough search of the available prior art.

This Article presents empirical evidence of the rapidly growing reliance on keyword search technology and of the resulting changes in the distribution of patents that are cited as prior art references. It also presents evidence that prior art search results have not reflected a recognition of the changing role of the Patent Office's classification system in a context where keyword search has become the dominant approach to information retrieval. These findings suggest that more advanced search tools should be made available to all concerned parties. This Article also makes a methodological contribution to the empirical literature on patent citations. Namely, the Article develops and validates large imputed data sets that approximate the characteristics of citations found using various search methods where actual data on the provenance of citations are unavailable.

The remainder of this Article is organized as follows. Part I describes the transition to electronic prior art searching in the Patent Office and some changes in Patent Office procedure that have been introduced in the wake of this transition. Part II discusses some of the foreseeable effects of the Patent Office's move to a paperless prior art search facility. Part III then describes the development of the imputed data sets for keyword search and the other search methodologies analyzed in this Article. Part IV summarizes the results of the analysis and validation of the imputed data sets. Finally, Part V discusses potential approaches to improving the precision of automated search results.

I. TECHNOLOGICAL DEVELOPMENTS IN PRIOR ART SEARCH

A. Search Technology at the Patent Office

The Patent Office first instituted full-text patent search capability in 1984 by installing two dedicated terminals to be shared among all

Id.

Patent prior art is also commonly searched in the context of an infringement search, i.e., an inquiry into whether a particular product or process may infringe an issued patent. The scope of this Article, however, is limited to patentability searches, and the term "search," as used herein, refers only to patentability search.

examiners in the office for searching patents issued after 1976.[10] The database, known as USPAT, was expanded in 1991 to include patents issued between 1971 and 1975.[11] The Patent Office connected all of its examiners' desktop computers to the search systems in 1993 and 1994, thereby making the technology more accessible.[12] Even so, according to the Patent Office's automation director Nestor Ramirez, many examiners did not utilize the search capability, preferring to continue the practice of searching through the "shoes."[13] In 1999, however, the Patent Office introduced the Examiner Automated Search Tool ("EAST") and the Web-based Examiner Search Tool ("WEST") software interfaces for the examiners' desktop computers, triggering what Ramirez describes as a "big transition to the system" in 2000.[14] In 2001, a full-text database derived from optical character recognition of scanned paper patents issued between 1920 and 1970, known as USOCR, was made accessible through the EAST and WEST systems.[15]

Public access to the full-text patent databases has historically been limited. Online tools, including the Classification and Search Support Information System ("CASSIS") and Automated Patent Search ("APS") systems, were installed in certain designated Patent Depository Libraries beginning in the early 1980s.[16] Desktop access, however, only became available to the public in 1997 through the introduction of a Web interface to the Patent Full Text ("PatFT") database, which contains the full text of all patents issued on or after January 1, 1976.[17]

Historically, the Patent Office search room's voluminous paper files provided a publicly accessible means of searching U.S. patents by class and subclass. Since the disposal of the paper files in preparation for the agency's move to Alexandria in 2005,[18] on-site access to the patent prior art collections has been almost exclusively via the EAST

10. Telephone Interview with Nestor Ramirez, Dir., Office of Patent Automation, U.S. Patent & Trademark Office (May 15, 2007) (on file with the North Carolina Law Review).

11. *Id.*

12. *Id.*; *see also* Patrick Doody, *The Patent System is Not Broken*, 18 INTELL. PROP. & TECH. L.J. 10, 15 (2006) (stating that all examiners "had the ability to search for prior art electronically" soon after the USPTO issued them desktop computers in 1992).

13. Telephone Interview with Nestor Ramirez, *supra* note 10.

14. *Id.*

15. *Id.*

16. *See* Patent & Trademark Depository Library Association, About PTDLA, http://www.ptdla.org/ptdla (last visited Apr. 25, 2009).

17. *See id.*

18. *See supra* note 6 and accompanying text.

and WEST interfaces, through which users access the USPAT and USOCR databases on LiveLink Discovery servers supplied by OpenText Corporation.[19]

EAST and WEST support keyword searches ranging from simple single-word queries to highly complex structured queries combining keywords and phrases with class and subclass restrictions and Boolean and proximity operators. The Patent Office provides extensive training to examiners and members of the public in the proper use of EAST and WEST. In addition to text searches, users are trained to retrieve and browse patent drawings and other images in the agency's LiveLink Discovery databases. Image search queries, however, are limited to individual patent numbers and specific classes and subclasses.[20]

The Patent Office also continues to support off-site searching of the PatFT database via the agency's website. The Web interface supports a somewhat narrower range of search queries than is available on EAST and WEST. For example, proximity operators are not accepted, and the results from one search cannot be used to build a subsequent search. The Web-based and EAST/WEST search engines are similar, however, in that they both support Boolean queries that combine keywords and phrases with class and subclass restrictions.[21]

B. Institutionalization of Off-Site Searching

The Patent Office has further established its commitment to electronic search through various initiatives that will move prior art search activities to off-site locations. In recent years, the agency has explored new procedures for prior art search, including outsourcing the task to contractors and other third parties, encouraging applicants to conduct more rigorous and well-documented searches, and allowing many examiners to work from home. As a result of these initiatives, the quality of prior art searches increasingly depends on the performance of online patent databases in supporting search queries.

19. U.S. PATENT & TRADEMARK OFFICE, EAST TRAINING FOR PUBLIC USERS 2 (Oct. 2004) (describing EAST as an interface to BRS databases) [hereinafter EAST TRAINING MANUAL]; Wikipedia, BRS/Search, http://en.wikipedia.org/wiki/BRS/Search (last visited Apr. 25, 2009) (explaining that BRS databases have been re-branded as OpenText's Live Link Directory Servers).

20. *See generally* EAST TRAINING MANUAL, *supra* note 19, at 15–73.

21. *See* U.S. Patent & Trademark Office, Patent Full-Text and Full-Page Image Databases, http://www.uspto.gov/patft/ (last visited Apr. 25, 2009).

1. Outsourcing of Prior Art Search

In 2003, the Patent Office published its *21st Century Strategic Plan*,[22] announcing a new "multi-track" process in which the procedure for examining a patent application would vary according to how the accompanying prior art search was to be performed.[23] The plan expressly allows for the performance of prior art search by various parties other than the patent examiner, including contractor search services, foreign patent offices, and patent search and examination agencies acting on applications filed under the Patent Cooperation Treaty ("PCT").[24]

While the Patent Office plans to "continue to conduct in-house searches of practically all applications in the near term," the agency plans to conduct pilot studies on outsourcing to contractor search services.[25] If such outsourcing proves successful, the agency expects that the use of contractors "would gradually increase over time and eventually predominate" over searches by examiners.[26] One such pilot study began in 2005, with the outsourcing of searching for a number of its PCT applications to two firms.[27] The study is to "determine whether searches by commercial entities can maintain the accuracy and quality standards for searches conducted by the USPTO during the patent examination process while remaining cost effective."[28]

Apart from the issues to be addressed in the Patent Office's pilot study, some commentators have raised broader concerns about the outsourcing initiative. Ronald Stern, president of the Patent Office Professional Association, testified to Congress that an important "synergy" between the search and examination functions would be lost if the two processes were separated.[29] Susan Walmsley Graf has

22. U.S. PATENT & TRADEMARK OFFICE, THE 21ST CENTURY STRATEGIC PLAN 10 (2003), http://www.uspto.gov/web/offices/com/strat21/stratplan_03feb2003.pdf.

23. *See* U.S. PATENT & TRADEMARK OFFICE, MULTI-TRACK PATENT EXAMINATION PROCESS, http://www.uspto.gov/web/offices/com/strat21/action/p2p01.htm (last visited Apr. 25, 2009).

24. *See id.*

25. *Id.*

26. *Id.*

27. *See* Press Release, U.S. Patent & Trademark Office, USPTO Contracts International Patent Application Searches to Commercial Firms (Sept. 21, 2005), http://www.uspto.gov/web/offices/com/speeches/05-48.htm.

28. *Id.*

29. *U.S. Patent & Trademark Office: Fee Schedule Adjustment and Agency Reform, Hearing Before the Subcomm. on Courts, the Internet and Intellectual Property of the H. Comm. on the Judiciary*, 107th Cong. 87–94 (2002) (statement of Ronald A. Stern, President, Patent Office Professional Association).

suggested that outsourcing search services may be an inefficient use of Patent Office funds.[30] John Jeffery has argued that a major shift toward outsourcing would constitute an abdication of the Patent Office's congressionally authorized and inherently governmental function in determining patentability, thereby undermining the presumption of validity in issued patents and potentially disrupting the constitutional fidelity of the patent system.[31] To avoid these problems, Jeffery proposes limiting the outsourcing of prior art search to non-patent literature.[32]

2. Accelerated Examination Procedure

In August 2006, the Patent Office introduced an "Accelerated Examination" procedure whereby patent applicants who satisfy certain additional procedural requirements can expect to have their applications processed within twelve months instead of the more typical twenty-four to thirty months.[33] These procedural requirements include a preexamination prior art search by the applicant and the filing of a statement identifying: (1) the field of search by class and subclass, and (2) the databases searched and the logical queries used to search those databases.[34] The applicant must search U.S. patents and patent applications, as well as foreign patent documents and non-patent literature, unless she can provide a justification for omitting one of these sources.[35] The applicant's search must encompass every feature of the invention as either claimed or disclosed in the patent specification.[36] The applicant must also file an "accelerated examination support document" explaining in detail how each of the references found bears on the patentability of each of the claims.[37] The applicant's request for accelerated examination takes the form of a "petition to make special," which previously had been limited to inventions promoting environmental

30. Susan Walmsley Graf, *Improving Patents by Identifying Prior Art*, 11 LEWIS & CLARK L. REV. 495, 513 (2007) ("Since most patents are never asserted, it can be argued that money spent on prior art searches for the vast majority of patents will be wasted.").

31. John A. Jeffery, *Preserving the Presumption of Patent Validity: An Alternative to Outsourcing the U.S. Patent Examiner's Prior Art Search*, 52 CATH. U. L. REV. 761, 778–96 (2003).

32. *Id.* at 799.

33. *See* Changes to Practice for Petitions in Patent Applications To Make Special and for Accelerated Examination, 71 Fed. Reg. 36,323 (June 26, 2006) (announcing accelerated examination procedures and effective date of August 25, 2006).

34. *Id.* at 36,324, pt. 1, ¶ 8.

35. *Id.* at 36,324, pt. 1, ¶ 8(A).

36. *Id.* at 36,324–25, pt. 1, ¶ 8(B).

37. *Id.* at 36,325, pt. 1, ¶ 9.

quality, energy development or conservation, countering terrorism, or to applicants of advanced age or failing health.[38]

The advantage of accelerated examination was illustrated by the issuance of the first patent under the new program—for an ink cartridge to Brother Kogyo Kabushiki Kaisha on March 13, 2007—less than six months after the September 29, 2006, filing date.[39] Many applicants may decline to pursue this approach, however, because of the additional burdens and costs of satisfying the procedural requirements[40] and the potential estoppel effects of the representations made in the search statement and support document.[41]

3. Patent Hoteling Program for Patent Examiners

In 2005, the Patent Office introduced the Patent Hoteling Program, which offered up to 500 patent examiners the option of working from home.[42] The program, modeled after the Trademark Work-at-Home Program that began in 1997, is aimed at freeing up office space and allowing examiners to reside outside the Washington, D.C., region.[43] Examiners in the program are issued home computer equipment, given special training, and connected to the Patent Office's systems via a virtual private network.[44] Teleworkers may reserve shared on-site workspace for use during occasional periods when they prefer to work in the office.[45]

As examiners, applicants, and contractors work in isolation, remotely from the Patent Office, they all must rely heavily on search technology to identify and retrieve the relevant prior art that is needed for patentability determinations. Whether this reliance is warranted is yet to be determined, but it is possible in the remainder of this Article to identify and examine some of the foreseeable effects of the transition to search technology that should be considered in such an assessment.

38. 37 C.F.R. § 1.102 (2008).

39. *See* David Schaeffer, *USPTO's Accelerated Examination Program: Speed at a Price*, STROOCK CLIENT MEMORANDUM, (Stroock & Stroock & Lavan LLP, New York, N.Y.), Mar. 26, 2007, at 1, *available at* http://www.stroock.com/SiteFiles/Pub501.pdf.

40. *Id.*

41. *Id.* at 2 ("Such statements become a part of the application record and an adversary might later try to rely on those statements to challenge the patent.").

42. Daniel Pulliam, *Patent Office Launching Massive Telework Program*, GOV'T EXECUTIVE, Dec. 16, 2005, http://www.govexec.com/dailyfed/1205/121605p1.htm.

43. *Id.*

44. *Id.*

45. *Id.*

II. EXPECTED EFFECTS OF THE TRANSITION

A. Eliminating Paper

By discontinuing its paper U.S. patent archive, the Patent Office committed its examiners and search room patrons to prior art searching in a largely paperless environment. While digital automation has certainly eased the storage and retrieval of millions of patent documents, the effect of the transition on the *overall usability* of those documents appears to be more ambiguous. As a general matter, paper documents often prove to be more user-friendly than electronic documents. Researchers at Microsoft have empirically confirmed the advantages of paper in facilitating such activities as navigating through and around documents, reading more than one document at a time, marking up documents, and interweaving reading and writing.[46]

Concerns about the usability of an all-electronic public search facility came to a head in June 2002, when the agency requested comments and conducted a public hearing on the decision to go paperless.[47] Dozens of comments were submitted in opposition to the plan, including one from the American Bar Association's Section of Intellectual Property Law.[48] The comments were generally anecdotal but indicative of systemic problems. For example, various commentators noted that many records in the database appeared to be missing, inaccurate, or not readily accessible,[49] and that text files were unavailable for patents issued prior to 1971.[50] Representatives

46. ABIGAIL J. SELLEN & RICHARD H.R. HARPER, THE MYTH OF THE PAPERLESS OFFICE 145–47 (2002).

47. *See* United States Patent & Trademark Office, Proposed Plan for an Electronic Public Search Facility (May 7, 2002), http://www.uspto.gov/go/og/2002/week19/patsrch.htm.

48. *See* United States Patent & Trademark Office, Public Comments Resulting From: Notice of Public Hearing and Request for Comments on the Proposed Plan for an Electronic Public Search Facility, 67 Fed. Reg. 17055 (proposed Apr. 9, 2002), *available at* http://www.uspto.gov/web/offices/com/sol/comments/epubsearch/index.html [hereinafter Public Comments on Electronic Search] (comments of Hayden Gregory, Legislative Consultant, American Bar Association's Section of Intellectual Property Law) (opposing, "at least until an equivalent or better electronic system is demonstrated, the removal of the paper patent files from the PTO facilities, on the grounds that the paper files continue to be an important tool for searching patents").

49. *See id.* (comments of Joseph Clawson, the National Intellectual Property Researchers Association, Robert B. Weir, Randy Rabin, and David Testardi).

50. *See id.* (comments of Randy Rabin, Michael H. Minns, and Mark A. Watkins). The USOCR database, containing text files for patents issued between 1920 and 1971, is now accessible to the public from the computers in the Patent Office Search Room, *see*

of the National Intellectual Property Researchers Association ("NIPRA") were especially critical of the state of the database, noting that identical search queries often returned different results, numerous patents that had been reclassified in the paper files had not been reclassified in the database, the number of patents in a particular subclass in the paper files often did not match the corresponding number in the database, and more than 100,000 patents issued since 1971 were not yet text-searchable.[51]

Other commentators expressed concerns more specifically about the effectiveness of keyword search. They noted that keyword searches might miss references where patent applicants and searchers use different terms to describe the same concept,[52] or where searchers needed to examine the details of patent drawings[53] or chemical formulae.[54] One commentator felt that an overreliance on keyword search was leading to false positives as well as false negatives, as he had received office actions citing references "that have little to do with the invention but do contain appropriate keywords."[55]

It is reasonable to expect that the commentators' concerns regarding the integrity of the patent database will be resolved in time as the database is revised and expanded.[56] Concerns regarding overreliance on keyword search results, however, are likely to persist at least as long as the Patent Office maintains its patent classification system as a collection of knowledge (metadata) that may be searched

supra text accompanying note 16, but still not via the Web. *See supra* text accompanying note 17.

51. *See id.* (comments of Robert B. Weir).

52. *See id.* (comments of Allan M. Lowe, Esq., Michael H. Minns, and Mark A. Watkins); *cf.* Dale L. Carlson & Robert A. Migliorini, *Patent Reform at the Crossroads: Experience in the Far East with Oppositions Suggests an Alternative Approach for the United States*, 7 N.C. J.L. & TECH. 261, 264 (2006) ("[T]here are certain more recently developed technologies, such as computer software and business methods, where identifying the relevant prior art is often difficult with current computerized search tools.").

53. *See* Public Comments on Electronic Search, *supra* note 48 (comments of Allan M. Lowe, Esq., Michael H. Minns, and Mark H. Watkins).

54. *See id.* (comments of Charlotte M. Kraebel). *But see* U.S. Patent & Trademark Office, Public Hearing on Issues Related to the Identification of Prior Art During the Examination of a Patent Application 193 (July 14, 1999), http://www.uspto.gov/web/offices/com/hearings/priorart/0714pato.doc [hereinafter Public Hearings on Prior Art] (comments of Glenn E. Wise, Registered Patent Agent) (stating that keyword searching is relatively more useful in "the chemical area where the terms are better defined").

55. Public Comments on Electronic Search, *supra* note 48 (comments of Lee Grantham, the search department manager at a mid-size patent firm).

56. On-site searchers can now electronically search the U.S. patent collection dating back to 1920. *See supra* text accompanying note 15.

instead of, or in combination with, the full text of patent documents (the underlying data).

B. Changing the Role of Patent Classification

Examiners, practitioners, and the public have historically found patent prior art through a variety of techniques other than keyword searching. With or without the aid of paper files, people have commonly searched through entire subclasses of patents and patent applications.[57] The Patent Office's classification system continues to be maintained with this practice in mind.[58] Once found, a prior art patent can identify other prior art references, both those citing it and those cited by it. Examiners, particularly those within the same practice group, often direct each other to prior art references they have cited in previous office actions.[59]

The emergence of keyword search represents a significant departure from procedures that rely (directly or indirectly) on the Patent Office's classification system. From the beginning, the Patent Office's search engines have supported query terms that limit search results to specific classes or subclasses, but search queries do not contain such limitations by default. Thus, keyword search results often include patents dispersed throughout the Patent Office's classification system.

57. A small but well-known empirical study illustrates the variety of search approaches that have historically been used. In 1997, NIPRA's then-president James Cottone reviewed the records of 421 patentability searches his firm had conducted between 1988 and 1994 to determine how the resulting 787 prior art references had been found. James F. Cottone, *Online Patent Searching: A Good News Story, But Not the Whole Story*, 79 J. PAT. & TRADEMARK OFF. SOC'Y 233, 233–34 (1997). The study found that 358, or 45%, of the references had been found through manual searching in the Patent Office's search room; 294, or 37%, had been found through the Patent Office's online search facilities; 84, or 11%, had been found through manual searches of foreign patents and non-patent publications; and 51, or 6%, had been suggested by a Patent Office examiner. *See id.* at 234–35. Cottone presented these findings during the Patent Office's July 1999 public hearing on the identification of prior art at the examination stage. *See* Public Hearings on Prior Art, *supra* note 54, at 75.

58. *See* U.S. PATENT & TRADEMARK OFFICE, EXAMINER HANDBOOK TO THE U.S. PATENT CLASSIFICATION SYSTEM (1997), *available at* http://uspto.gov/web/offices/pac/dapp/sir/co/examhbk/one.htm (noting the goal of "subdivid[ing] our classification files into searchable units"); U.S. PATENT & TRADEMARK OFFICE, MANUAL OF PATENT EXAMINING PROCEDURE § 904.02 (8th ed. 2006), *available at* http://www.uspto.gov/web/offices/pac/mpep/mpep_e8r5_0900.pdf [hereinafter MANUAL OF PATENT EXAMINING PROCEDURE] ("The traditional method of browsing all patent documents in one or more classifications will continue to be an important part of the search strategy when it is difficult to express search needs in textual terms.").

59. Telephone Interview with Nestor Ramirez, *supra* note 10.

Given that the Patent Office's classification system, like all such systems, is a useful but imperfect aid to information retrieval, commentators have disagreed on whether the rise of keyword search is more promising or problematic. Some speakers at the Patent Office's June 2002 hearing argued that keyword search is an inadequate substitute for class- and subclass-wide search in identifying relevant prior art,[60] and predicted that reliance on keyword search would cause the classification system to fall into obsolescence and disuse.[61] One commenter, however, took the position that keyword searching was beneficial in identifying prior art that a classification-based search might miss.[62]

A more radical view, famously espoused by David Weinberger in his recent book *Everything is Miscellaneous*,[63] is that classification systems have largely become obsolete in the digital age.[64] According to Weinberger, systems for organizing information can be described as "first-order" (organizing physical documents), "second-order" (organizing metadata about physical documents), and "third-order" (gathering, but not organizing, documents and metadata to be processed later in response to a search query).[65] Weinberger argues that digital media and search engine technology obviate the need for first-order and second-order organization systems.[66] Thus, the time and effort required to maintain a useful classification system would be

60. *See* Public Comments on Electronic Search, *supra* note 48 (comments of Calvin E. VanSant, Lee Grantham, Charlotte M. Kraebel, and Donal B. Tobin).

61. *See id.* (comments of Randy Rabin and Lee Grantham).

The concern that updates to the U.S. patent classification schedule are failing to keep up with technological developments has recurred in the literature. *See, e.g.*, Leah S. Larkey, *A Patent Search and Classification System*, PROC. OF THE FOURTH ACM CONF. ON DIGITAL LIBRARIES 179, 180 (1999) (describing the difficulty of training classifiers and updating schedule).

62. Public Hearings on Prior Art, *supra* note 54, at 47–48 (comment of Mary Helen Sears).

> [I]f the examiner who is classifying particular claims in connection with allowing the application happens to make a mistake or two, it makes it very easy to miss U.S. patent references if you're relying on the classification system to search only a particular class and subclass, and today I do believe the computer word searches that are carefully carried out even in U.S. patents can help to alleviate that problem.

Id.; *see also id.* at 178–89 (comment of Glenn E. Wise) (commenting on shortage of staff in Office of Patent Classification).

63. DAVID WEINBERGER, EVERYTHING IS MISCELLANEOUS: THE POWER OF THE NEW DIGITAL DISORDER (2007).

64. *See id.* at 46–63 (discussing obsolescence of Dewey decimal system).

65. *See id.* at 17–23.

66. *See id.* at 84–85.

more efficiently spent on gathering rich metadata for query-time processing,[67] such as metatags,[68] recommendations,[69] and discussions.[70] Weinberger also describes collaborative technologies as potentially powerful tools for harnessing collective knowledge in the production and refinement of such metadata.[71]

In terms of Weinberger's taxonomy, the Patent Office's move to a paperless patent collection effected a transition from a first-order approach to a second-order approach to organization. For the foreseeable future, however, the Patent Office does not appear likely to abandon its classification system in favor of a third-order approach to organization. This is because the classification system serves not merely as an aid to information retrieval, but as the basis for assigning incoming applications to examiners.[72] While particular classifications can sometimes be erroneous[73] and are subject to change,[74] the system represents a vast and unique body of collective knowledge that is immediately applicable to the Patent Office's information retrieval functions. Thus, classes and subclasses continue to figure heavily in the formulation of search queries, even though the "shoes" they once represented have now been retired.

C. Reassessing the Patent Classification System

Since the Patent Office's search engines allow requests for all patents in a specified class,[75] searchers today can still treat the classification system as a system of "shoes" to be browsed.[76] Such an approach to prior art search—in which the searcher undertakes to review the entire contents of a class—assumes that all or almost all of the relevant patent references can be found within a relatively small number of classes that can be browsed in their entirety, with a particular focus on the class to which the patent application has been

67. *See id.* at 173–98.

68. *See id.* at 84–128.

69. *See id.* at 129–33.

70. *See id.* at 133–47.

71. *See id.* at 57–63 (describing collaborative filtering on Amazon.com); *id.* at 129–47 (discussing emergence of "social knowing" in the creation of Wikipedia content).

72. *See* MANUAL OF PATENT EXAMINING PROCEDURE, *supra* note 58, § 903.08.

73. *See supra* note 62.

74. *See generally* MANUAL OF PATENT EXAMINING PROCEDURE, *supra* note 58, §§ 903.02(a), 903.04–.08 (describing procedures for defining new classes and reclassifying patent applications after assignment to an examiner).

75. This Section uses the term "class" generically to refer to a class or subclass defined by the Patent Office's classification system.

76. *See supra* note 58 and accompanying text.

assigned.[77] To support searches of this kind, the patent classification system should be designed with high "recall." In the study of information retrieval, "recall" refers to the fraction of relevant items that are retrieved.[78] If the patent classification system has high recall, searchers can expect on average to find a high percentage of relevant patents by retrieving and reviewing a small number of classes.

More typically, however, patent searchers tend to rely primarily on the text search capabilities of automated search tools,[79] using classifications as necessary to resolve ambiguities that may arise from the imprecision of language. The Patent Office's *Manual of Patent Examining Procedure* acknowledges that examiners will attempt to express their "search needs" primarily through text search queries, but notes that lexical ambiguities may require the use of classification terms in those queries.[80] The agency also appears to have similar expectations regarding applicants who perform their own prior art searches in hopes of obtaining accelerated examination, if the Patent

77. *See* MANUAL OF PATENT EXAMINING PROCEDURE, *supra* note 58, § 904.02(a) ("A proper field of search normally includes the subclass in which the claimed subject matter of an application would be properly classified. It is not necessary to search areas in which it could reasonably have been determined that there was a low probability of finding the best reference(s).").

78. *See* CHARLES T. MEADOW ET AL., TEXT INFORMATION RETRIEVAL SYSTEMS 329 (2007).

79. Patent Office procedures and training materials appear to acknowledge that text searches now constitute the predominant use of the agency's search technologies. *See generally* EAST TRAINING MANUAL, *supra* note 19 (devoting the bulk of the training materials to techniques for text search).

80. The manual states:

> Text search can be powerful, especially where the art includes well-established terminology and the search need can be expressed with reasonable accuracy in textual terms. However, it is rare that a text search alone will constitute a thorough search of patent documents. Some combination of text search with other criteria, in particular classification, would be a normal expectation in most technologies.
>
> Examiners will recognize that it is sometimes difficult to express search needs accurately in textual terms. This occurs often, though not exclusively, in mechanical arts . . . . In such situations, text searching can still be useful by employing broader text terms, with or without classification parameters. The traditional method of browsing all patent documents in one or more classifications will continue to be an important part of the search strategy when it is difficult to express search needs in textual terms.

MANUAL OF PATENT EXAMINING PROCEDURE, *supra* note 58, § 904.02.

Office's example of a pre-examination search report is anything to go by.[81]

For a patent classification system to function effectively as an adjunct to keyword search terms in the face of lexical ambiguities, it is helpful for the system to have high recall, but high "precision" can often be of equal or greater importance. "Precision" refers to the fraction of retrieved items that are relevant.[82] A classification system that offers high precision can significantly reduce the number of documents that need to be reviewed for relevance where lexical ambiguities might otherwise lead to an overbroad set of keyword search results.

For example, a keyword search for prior art on "cell phone" might be underinclusive, failing to find documents containing the synonymous terms "mobile phone" or "hand phone."[83] If the classification system has sufficiently high recall, then the use of a classification term in place of underinclusive keywords may be expected to result in the retrieval of a significantly greater number of relevant references. On the other hand, a keyword search might be overinclusive, producing a result set that includes documents about cell phone holders, cell phone shields, or cell phone mice; and further afield, perhaps even terrorist cells, jail cells, electrolytic cells, the Sony Cell™ microprocessor, or stem cells.[84] If the classification system has sufficiently high precision, then the use of a classification term in conjunction with an overbroad keyword query may be expected to narrow the result set to those fields of technology in which the relevant references can be found.

For all the imprecision that attends the formulation of keyword queries, it is probably best that searchers no longer have to rely exclusively on the classification system to identify sets of references to be retrieved and browsed. Given that classification errors can occur,[85]

81. *See* U.S. Patent & Trademark Office, Pre-Examination Search Document, http://www.uspto.gov/web/patents/accelerated/ae_presearch_sample.doc (last visited Apr. 25, 2009) (providing a sample letter of a pre-examination search).

82. *See* MEADOW et al., *supra* note 78, at 328–31.

83. *See* James Ryley, Using Conceptual Search in Scientific, Financial and Intellectual Property Databases, http://www.infonortics.eu/chemical/ch07/slides/ryley-2.pdf (last visited Apr. 25, 2009).

84. *See id.*; *see also* Arti K. Rai, John R. Allison & Bhaven N. Sampat, *University Software Ownership and Litigation: A First Examination*, 87 N.C. L. REV. 1519 nn. 59–60 and accompanying text (reporting findings supporting the researchers' concern that keyword search for software-related terms "could produce a data set that has both false positives and false negatives").

85. *See supra* note 62.

the ability of searchers to find patent prior art using a wide variety of approaches—such as classification terms to resolve ambiguities resulting from both underinclusive and overinclusive keyword terms—makes it more likely that misclassified references will eventually be found, and that erroneous patentability determinations that result from failures to find such references will not propagate indefinitely.

In summary, the Patent Office's transition from paper patents to search engines requires a shift in the way we evaluate the patent classification system as an aid to information retrieval. Recall is no longer the paramount performance measure; depending on the nature of the lexical ambiguity involved, precision often assumes greater importance.

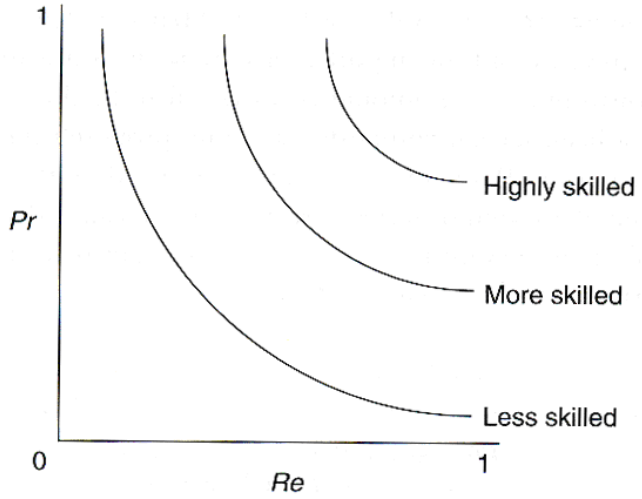
D. The Recall-Precision Tradeoff

The need for the patent classification system to exhibit both high recall and high precision in support of automated search creates a potential tension, because there is typically an inverse relationship between the two performance measures. This tradeoff between recall and precision has been demonstrated in both empirical studies of human searching behavior[86] and theoretical studies of automated systems for aiding or performing information retrieval.[87] As Figure 1 illustrates, a highly skilled searcher may be able to formulate search queries that achieve higher levels of recall and precision than a less skilled searcher, but for any given individual, greater recall can be achieved only at the expense of a loss in precision, and vice versa.

86. *See, e.g.*, MEADOW et al., *supra* note 78, at 330 (citing CYRIL CLEVERDON & MICHAEL KEEN, 2 ASLIB CRANFIELD RESEARCH PROJECT: FACTORS AFFECTING THE PERFORMANCE OF INDEXING SYSTEMS 37 (1966)).

87. *See, e.g.*, Michael Buckland & Fredric Gey, *The Relationship Between Recall and Precision*, 45 J. AM. SOC'Y FOR INFO. SCI. 12, 16–19 (1994); Michael Gordon & Manfred Kochen, *Recall-Precision Trade-Off: A Derivation*, 40 J. AM. SOC'Y FOR INFO. SCI. 145, 146–50 (1989); Sergio A. Alvarez, An Exact Analytical Relation Among Recall, Precision, and Classification Accuracy in Information Retrieval 15–21 (2002), http://www.cs.bc.edu/~alvarez/APR/aprformula.pdf.

Figure 1. Typical Relationship Between Precision (Pr), Recall (Re), and User Skill[88]



Similarly, information retrieval systems (including classification systems) may vary according to their accuracy in classifying documents as relevant or irrelevant. For any given system, however, there is a tradeoff between recall and precision, as shown in Figure 2.

The Patent Office's classification system is expressly required by statute to be maintained for the purpose of supporting accurate determinations regarding the relevance of prior art patents to patentability.[89] Given this requirement, it is reasonable to regard the classification system's level of accuracy as an invariant, and the attained levels of recall and precision as parameters that can vary according to how the classification system is being used in the context of various prior art search techniques, including automated search.[90]

88. *See* MEADOW et al., *supra* note 78, at 331.

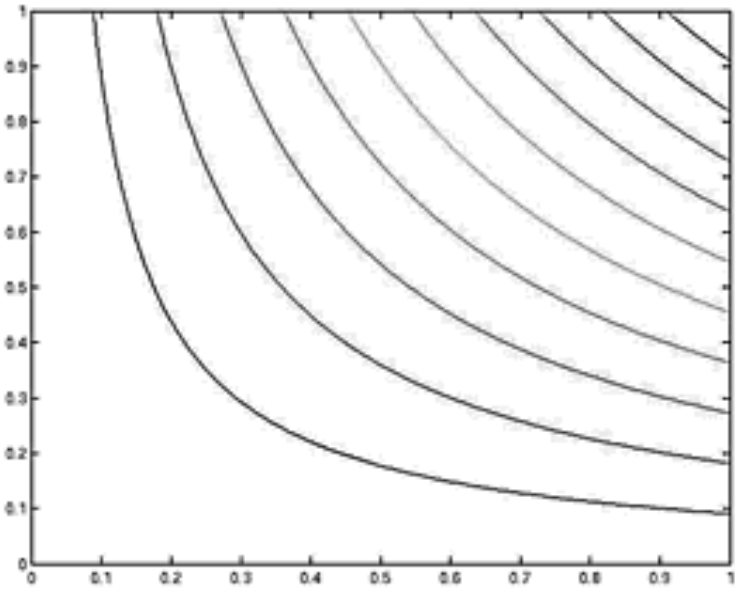
89. Section 8 of the Patent Act provides:

> The Director may revise and maintain the classification by subject matter of United States letters patent, and such other patents and printed publications as may be necessary or practicable, for the purpose of determining with readiness and accuracy the novelty of inventions for which applications for patent are filed.

35 U.S.C.A. § 8 (West 2001 & Supp. 2008).

90. See *supra* text accompanying notes 57–62 for a description of various patent prior art search methods.

Figure 2. Recall-precision Tradeoffs at Varying Levels of Classification Accuracy[91]



Accordingly, for the classification system when used in support of automated search, it is not possible simultaneously to achieve higher recall and higher precision instead of traditional search methods. Thus, if the classification system is adequately performing its new function of resolving lexical ambiguities in text searches, some trading off of recall for precision should be evident in the results of those searches. Part IV presents empirical evidence that this trade off does not occur: recall actually seems to be significantly higher for search results generated through text search than those generated through other search methods.

III. METHODOLOGY AND DATA

The primary source data for this study were extracted from the Patent Office's PatFT database, which contains the full text of all patents issued on or after January 1, 1976, and supports keyword full-text search via the Web.[92] The study includes all U.S. utility patents issued on or before May 1, 2007, covering patent numbers 3,930,271

91. *See* Alvarez, *supra* note 87, at 18.

92. *See* U.S. Patent & Trademark Office, Patent Full-Text and Full-Page Image Databases, http://www.uspto.gov/patft/ (last visited Apr. 25, 2009).

through 7,213,269, inclusive. Excluding withdrawn patent numbers, the full-text patent data set includes 3,266,297 patents.

The limitations on the full-text database impose some further limitations on the set of patent citations that can be analyzed in this study. While patents of any vintage can be cited as prior art, this study covers only citations to patents within the database itself (i.e., those issued on or after January 1, 1976). Thus, for a citation to be included in this study, both the citing patent and the cited patent must be numbered between 3,930,271 and 7,213,269 inclusive. The base citation data set includes 23,729,900 citations of this form.

A. Imputation of Citations to Search Methods

To characterize the influence of technology on the search for patent prior art, it would be helpful to have data identifying the search method that was used to locate each reference cited in the patent. The patent's prosecution history file provides a good deal of this information, including references cited by the examiner and disclosed by the applicant, patent classes and subclasses searched by the examiner, and logical keyword queries used by the examiner to search the full-text databases. Moreover, this information is now more widely available than ever, as the Patent Office's move to a paperless examination system has led to the publication of scanned prosecution history files ("image file wrappers") on the agency's Web site since August 2004.[93] There is nothing in these files, however, to indicate which of the cited prior art references were found through keyword searching or the use of other search technologies. The agency generally does not make such nonpublic information regarding prior art search available even for research purposes.[94]

A study by NIPRA's James Cottone[95] illustrates one possible approach to identifying sets of citations that were found through various search methods. Cottone identified a data set of 294 citations that were actually known to have been found through the Patent Office's online search facilities.[96] His study was based on the nonpublic records of searches conducted by his firm,[97] however, and is

93. *See* Joseph D. Cohen, *What's Really Happening in Inter Partes Reexamination*, 87 J. PAT. & TRADEMARK OFF. SOC. 207, 212 (2005); Press Release, U.S. Patent & Trademark Office, Internet Access to Patent Application Files Now Available (Aug. 2, 2004), http://www.uspto.gov/web/offices/com/speeches/04-13.htm.

94. *See* Telephone Interview with Nestor Ramirez, *supra* note 10.

95. *See* Cottone, *supra* note 57.

96. *See id.* at 234.

97. *See id.* at 233.

therefore neither repeatable nor extensible. Moreover, it is unclear whether the 421 patentability searches conducted by his firm were representative of prior art searches in general.

To support more general observations about the impacts of search technology, it would be desirable to generate a much larger data set based on a comprehensive analysis of the available underlying data. Accordingly, this study relaxes the requirement of actual knowledge, and instead attempts to impute patent citations to various search methods based on other known information about the relationships between the citing and cited patents. Each of the resulting imputed data sets consists of those citations in the basic data set that share a particular property in common with the citations that would actually have been found through the method under study. The properties are chosen so as to be characteristic of the method under study and weakly correlated with the characteristic properties of other methods.

For keyword search, this study's imputed data set consists of all citations in the base citation data set where both the citing and cited patents contain the same "low-frequency" keyword in both their detailed description and claims sections. A keyword is defined as low-frequency if it appears in these fields in fifty or fewer patents in the public PatFT database, as determined by a structured single-keyword query to the Patent Office's Web server. Queries were conducted for each of the 354,984 words in the Moby Words II SINGLE.TXT word list, a widely-used public domain text file,[98] and found 29,050 low-frequency words. This analysis produced a list of 61,221 citations imputed to keyword search. For each of these citations, there is a corresponding low-frequency keyword, which is assumed to have appeared in a logical query during the prior art search for the citing patent whereby the cited patent was found.

This study also examined the methods of searching through forward citation tracking (i.e., locating the patents that also cite a cited patent) and backward citation tracking (i.e., locating the patents cited by a cited patent). To produce the imputed data sets, the study identified all citations in the base citation data set where the citing and cited patents both cited a third patent, or where the citing patent cited a third patent that also cited the cited patent. The 7,405,952 citations of the first type were imputed to forward citation tracking, and the 7,624,501 citations of the second type were imputed to

98. *See* Wikipedia, *Moby Project*, http://en.wikipedia.org/wiki/Moby_Project (last visited Apr. 25, 2009).

backward citation tracking. Note that while backward citation tracking is amenable to manual (paper-based) searching, forward citation tracking is not.

Finally, this study examined the method of searching through the entire primary subclass to which the citing patent was ultimately assigned. This method is amenable to manual searching[99] and corresponds to the time-honored tradition of browsing the "shoes" in the Patent Office. The imputed data set for classification search consists of 2,631,901 citations where the citing and cited patents were both assigned to the same class and subclass, per the Patent Office's February 2006 classification schedule.[100]

B. Further Assumptions and Limitations

The imputed data sets omit several other potentially relevant considerations, reflecting further simplifying assumptions and limitations on the scope of this study.

Pre-1976 data. In confining its analysis to patents available in the PatFT database, the present study does not incorporate other data that the Patent Office has made available through its public search facilities. The USPAT database, which contains the full-text of U.S. patents issued since 1971, can be accessed by examiners and the public on Patent Office workstations that run the EAST and WEST software interfaces. While additional data from patents issued between 1971 and 1975 would no doubt yield more informative results, the difficulty of conducting such an extensive study on-site in the Patent Office made it necessary to utilize the more widely available PatFT database.

Changes to the USPTO classification schedule. This study did not account for changes in the Patent Office's classification schedule, which has been amended from time to time, generally in the direction of further refinement. While the renumbering of classes and subclasses over time does not affect the validity of the imputed data set for classification search, the refinement of subclasses may have led to the systematic omission of many earlier citations.

99. *See* CLASSIFICATION HANDBOOK, *supra* note 58, ch.1 (noting the goal of "subdivid[ing] our classification files into searchable units").

100. The classification schedule is maintained at http://www.uspto.gov/web/offices/opc. This study used the schedule as updated through Classification Order 1854. U.S. Dept. of Commerce, Patent & Trademark Office, Classification Order 1854 (Feb. 7, 2006), http://www.uspto.gov/web/offices/opc/documents/1854.pdf.

Examiner- vs. applicant-generated references. Since 2001, the paper versions of U.S. patents have distinguished between prior art references cited by the examiner and those cited by the applicant for patent; however, the PatFT database does not draw this distinction. This study's citation data sets are based on data extracted from the PatFT database and therefore do not distinguish between examiner- and applicant-generated references. It is therefore not possible here to determine the extent to which this study's conclusions relate to reliance on keyword search by examiners rather than applicants, or vice versa. Such a determination would certainly be of considerable interest, particularly in assessing the increasing involvement of applicants and the general public in the search process. Considerable additional resources would, however, be required to perform the necessary data entry tasks, and so this subject is left for future study.[101]

Multiple-word queries. In contrast to the single-word queries used to generate the imputed data set for keyword search, most search queries are more complex, combining words and phrases with class and subclass limitations, as well as Boolean and proximity operators. Even so, low-frequency keywords, by their nature, contribute disproportionately to the discriminatory power of a search query when taken in conjunction with other keywords.[102] Recognizing this fact, the Patent Office's training manuals advise users of EAST and WEST to "[s]earch for *unique* words first" and to

101. For empirical studies of the characteristics of examiner- and applicant-identified citations, see generally, Juan Alcacer & Michelle Gittelman, *How Do I Know What You Know? The Role of Inventors and Examiners in the Generation of Patent Citations* (Working Paper, 2004), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=548003; Bhaven N. Sampat, *Examining Patent Examination: An Analysis of Examiner and Applicant Generated Prior Art* (unpublished manuscript), available at http://www.stiy.com/MeasuringInnovation/Sampat.pdf.

102. *See generally* Antoine Blanchard, *Understanding and Customizing Stopword Lists for Enhanced Patent Mapping*, 29 WORLD PATENT INFO. 308, 309–12 (2007) (showing that precision in patent retrieval is improved when high-frequency "stopwords" in queries are ignored, but stopword lists may be technology-specific); H.P. Luhn, *The Automatic Creation of Literature Abstracts*, 2 IBM J. RES. & DEV. 159, 160 (1958) ("Within a technical discussion, there is a very small probability that a given word is used to reflect more than one notion. The probability is also small that an author will use different words to reflect the same notion."); Liz Price & Mike Thelwall, *The Clustering Power of Low Frequency Words in Academic Webs*, 56 J. AM. SOC'Y FOR INFO. SCI. & TECH. 883, 886–87 (2005) (concluding that "a significant proportion of low frequency words contain subject-related information" and aid in the creation of similar clusters); *cf.* Jeremy Pickens & W. Bruce Croft, *An Exploratory Analysis of Phrases in Text Retrieval* 16 (Working Paper, 2000), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.8810 (follow "Download PDF" link) (showing that the structure of phrases used in text queries influences the precision of search results).

build more complex queries from there.[103] While low-frequency keywords need not play a role in every keyword search result, there does not appear to be a loss of generality in restricting the imputed data set to citations imputed to single-word queries.[104]

Non-patent prior art. While the influence of patent search technology on the search for non-patent prior art was excluded from the present study, it is a subject worthy of further investigation, particularly in fields such as software and business methods.[105]

IV. ANALYSIS OF THE DATA

A. Longitudinal Data on Keyword Searching

1. Prevalence of Citations Imputed to Keyword Search

Table 1 shows the trend in the relative prevalence of keyword search over time, based on the imputed data set for low-frequency keywords (2–50 hits) and medium-frequency keywords (51–500 hits). It is necessary to normalize the number of previously issued patents that could be identified as prior art through keyword search. Accordingly, in each case this study applies a sliding window of 1,000,000 patent numbers (i.e., a citation is included in the count if the cited patent was among the 1,000,000 patents issued immediately prior to the citing patent).

The observed gradual upward trend is consistent with the Patent Office's transition to search technology during this period. The most dramatic increases were in 1999–2002 and in 2006–2007, which correspond closely to the introduction of desktop search tools for examiners in 1999–2000, the expansion of the searchable patent database in 2001, and the elimination of the Patent Office's paper files in 2005.

103. EAST TRAINING MANUAL, *supra* note 19, at 180.

104. *See infra* Part IV.D.1.

105. *See, e.g.*, John R. Allison & Mark A. Lemley, *The Growing Complexity of the United States Patent System*, 82 B.U. L. REV. 77, 102 (2002); Julie E. Cohen, *Reverse Engineering and the Rise of Electronic Vigilantism: Intellectual Property Implications of "Lock-Out" Programs*, 68 S. CAL. L. REV. 1091, 1179 (1995) (noting difficulty of finding software prior art); *see also infra* notes 124–27 and accompanying text (noting identification of non-patent prior art in software field).

Table 1. Citations Imputed to Keyword Search as a Percentage of All Citations Based on Results of Low- and Medium-frequency Single-word Queries

Issue Year	Total Citations	Citations Imputed to Keyword Search			
		2–50 hits		51–500 hits	
		Number	%	Number	%
1990[106]	74,516	248	0.33	3,171	4.26
1991	458,747	1,418	0.31	20,122	4.39
1992	480,705	1,514	0.31	20,868	4.34
1993	500,352	1,587	0.32	22,201	4.44
1994	543,618	1,811	0.33	24,230	4.46
1995	566,289	1,843	0.33	24,465	4.32
1996	635,450	2,223	0.35	27,468	4.32
1997	655,047	2,228	0.34	28,926	4.42
1998	864,396	3,055	0.35	38,555	4.46
1999	865,653	3,677	0.42	39,917	4.61
2000	772,609	3,492	0.45	36,291	4.70
2001	817,032	3,953	0.48	38,916	4.76
2002	822,401	4,319	0.53	40,072	4.87
2003	847,952	4,808	0.57	40,950	4.83
2004	811,115	4,654	0.57	36,412	4.49
2005	713,267	4,153	0.58	32,155	4.51
2006	866,806	5,418	0.63	39,968	4.61
2007[107]	268,107	1,905	0.71	12,418	4.63

2. Effect of Keyword Search on Years of Patents Cited

Each row in Table 2 summarizes the respective estimates for the coefficient B in linear regression models of the form

$$p = Ad + Bk + C,$$

where for each patent (observation), p is the fraction of cited patents issued during the indicated five-year interval, d is the issue year of the patent, and k is the number of times the patent appears as a citing patent in the imputed data set for keyword search, restricted to patents issued after the terminal year of the interval. The resulting regression estimates show that the distribution of ages in a patent's list of prior art references is significantly associated with the prevalence of citations imputed to keyword search in that list. Specifically, citing patents in the imputed data set for keyword search tend to cite more post-1976 references and fewer pre-1976 references than other patents issued in the same year. Given that patent examiners have until recently been unable to perform keyword

106. Partial year.
107. Partial year.

searches for older references, these results are as expected, provided that the imputed data set for keyword search is a valid proxy for citations actually found through keyword search. Various approaches to the validity analysis are presented in detail below.[108]

Table 2. Linear Regression Estimates Indicating Associations Between Prevalence of Citations Imputed to Keyword Search and Issue Years of Cited Patents

Issue Year of Patent Reference	*B* coefficient Estimate	Standard Error	*t* statistic	*p* value
Pre-1956	–0.00144	0.000430	–3.34	0.0008
1956–60	–0.00080	0.000190	–4.22	< 0.0001
1961–65	–0.00123	0.000234	–5.25	< 0.0001
1966–70	–0.00186	0.000307	–6.05	< 0.0001
1971–75	–0.00151	0.000428	–3.53	0.0004
1976–80	0.00940	0.000451	20.87	< 0.0001
1981–85	0.00602	0.000476	12.64	< 0.0001
1986–90	0.00432	0.000587	7.36	< 0.0001
1991–95	0.00351	0.000670	5.24	< 0.0001
1996–2000	–0.00061	0.000809	–0.75	0.4510

B. Other Imputed Search Methods

1. Cross-Tabulations of Citations Between Imputed Data Sets

As discussed above,[109] a total of four imputed data sets were created to compare the proliferation and performance of keyword search with other search methods. The imputations do not cover all citations in the base data set, and some citations are imputed to more than one search method. Table 3 shows the number of citations contained in each of these sets and in their pairwise intersections.

Table 3. Numbers of Citations Imputed to Individual Search Methods and Pairs of Search Methods

	Keyword	Forward	Backward	Classification	All
Keyword	61,221	32,250	18,140	13,997	61,221
Forward	32,250	7,405,952	2,910,858	1,126,645	7,405,952
Backward	18,140	2,910,858	7,624,501	840,098	7,624,501
Classification	13,997	1,126,645	840,098	2,631,901	2,631,901
All	61,221	7,405,952	7,624,501	2,631,901	23,729,900

108. *See infra* Part IV.D.
109. *See supra* Part III.A.

2. Prevalence of Citations Imputed to Citation Tracking

Table 4 shows the trends in the relative prevalence of backward and forward citation tracking over time. To calculate these trends, it is necessary to normalize the number of previously issued patents that could either identify or be identified as prior art through citation tracking. Accordingly, this study applies a sliding window of 1,000,000 patent numbers to the base and imputed data sets, i.e., a citation is included in the count if the cited patent was among the 1,000,000 patents issued immediately prior to the citing patent. Unlike Table 1, Table 4 shows no clear overall trend in the prevalence of citations imputed to these methods. There is one noteworthy discontinuity—a decrease from 1999 to 2000, which may reflect the introduction of automated search tools on examiners' desktops in those years.

Table 4. Citations Imputed to Backward and Forward Citation Tracking as a Percentage of All Citations

Issue Year	Total Citations	Citations Imputed to Citation Tracking					
		Backward		Forward		Total	
		Number	%	Number	%	Number	%
1990[110]	74,516	15,964	21.42	22,356	30.00	38,320	51.43
1991	458,747	99,335	21.65	140,858	30.70	240,193	52.36
1992	480,705	106,631	22.18	152,731	31.77	259,362	53.95
1993	500,352	113,092	22.60	162,355	32.45	275,447	55.05
1994	543,618	125,248	23.04	181,307	33.35	306,555	56.39
1995	566,289	135,169	23.87	194,739	34.39	329,908	58.26
1996	635,450	155,869	24.53	223,244	35.13	379,113	59.66
1997	655,047	163,012	24.89	238,797	36.45	401,809	61.34
1998	864,396	207,015	23.95	316,619	36.63	523,634	60.58
1999	865,653	201,281	23.25	321,753	37.17	523,034	60.42
2000	772,609	152,526	19.74	251,194	32.51	403,720	52.25
2001	817,032	156,405	19.14	269,621	33.00	426,026	52.14
2002	822,401	154,244	18.76	278,689	33.89	432,933	52.64
2003	847,952	151,010	17.81	295,369	34.83	446,379	52.64
2004	811,115	138,044	17.02	278,618	34.35	416,662	51.37
2005	713,267	124,342	17.43	249,671	35.00	374,013	52.44
2006	866,806	158,851	18.33	313,772	36.20	472,623	54.52
2007[111]	268,107	51,015	19.03	103,676	38.67	154,691	57.70

110. Partial year.
111. Partial year.

C. Performance of Imputed Search Methods

1. Imputed Search Method by Technological Field

Table 5 confirms and quantifies the observations of various commentators that the utilization of keyword search varies by field of technology.[112] Prior art searches in medicine and chemistry appear to rely more heavily on keywords than average; searches in physics, energy, and tools appear to rely less on keywords. In contrast, there appears to be considerably less variation in the usage of forward and backward citation tracking methods across technological fields.

Table 5. Relative Prevalence of Citations Imputed to Keyword Search and to Backward and Forward Citation Tracking by Technological Field

	Citations Imputed to Keyword Search		Citations Imputed to Citation Tracking Forward		Backward	
	Percent of Category	Multiple of Overall	Percent of Category	Multiple of Overall	Percent of Category	Multiple of Overall
Overall	0.258%		31.2%		32.1%	
Chemistry	0.430%	1.667	31.2%	1.000	31.1%	0.967
Communications	0.170%	0.659	24.9%	0.798	28.4%	0.884
Construction	0.237%	0.919	36.2%	1.160	33.4%	1.040
Energy	0.129%	0.500	27.7%	0.889	28.2%	0.877
Engineering	0.203%	0.789	31.1%	0.997	32.5%	1.010
Medicine	0.489%	1.897	40.0%	1.282	42.6%	1.324
Household	0.230%	0.893	33.4%	1.069	31.9%	0.992
Industrial	0.190%	0.735	33.5%	1.073	32.0%	0.996
IT	0.191%	0.739	24.3%	0.778	27.9%	0.867
Material Science	0.280%	1.085	32.6%	1.046	32.3%	1.007
Optics	0.197%	0.764	28.4%	0.910	31.7%	0.985
Packaging	0.185%	0.717	38.1%	1.222	38.7%	1.204
Physics	0.078%	0.304	20.9%	0.671	26.6%	0.827
Tools	0.166%	0.644	35.5%	1.138	34.3%	1.067
Transportation	0.183%	0.708	33.5%	1.073	31.8%	0.990

2. Co-Classified Prior Art by Technological Field

Given the wide variation in the utilization of keyword search across technological fields, Table 6 presents the striking finding that the citations imputed to keyword search tend to be disproportionately between patents in the same PTO class and/or subclass, regardless of technological field. In every field, citations imputed to keyword search are more frequently co-classified than citations imputed to

112. *See supra* notes 52–54 and accompanying text.

either forward or backward citation tracking and citations overall. In terms of the PTO classification system's performance, the system's recall appears to be significantly higher for search results generated through keyword search than results generated through other search methods.

Table 6. Relative Prevalence of Co-classified Prior Art by Technological Field and Imputed Search Method

	All Citations		Keyword		Citation Tracking Forward		Citation Tracking Backward	
	Same Class	Same Sub	Same Class	Same Sub	Same Class	Same Sub	Same Class	Same Sub
Overall	47.9%	11.1%	61.0%	22.9%	53.4%	15.3%	47.3%	11.0%
Chemistry	46.6%	11.4%	62.9%	22.9%	49.7%	14.4%	44.3%	10.2%
Communications	46.3%	7.5%	57.6%	16.7%	51.3%	10.2%	45.2%	6.8%
Construction	49.0%	12.6%	61.4%	23.9%	56.3%	17.5%	51.6%	13.6%
Energy	50.5%	13.2%	59.7%	26.5%	55.4%	17.2%	49.9%	13.1%
Engineering	34.7%	9.0%	54.6%	19.8%	42.4%	13.8%	33.2%	9.1%
Medicine	49.8%	11.7%	63.2%	23.2%	54.9%	15.4%	49.0%	11.1%
Household	57.4%	14.1%	67.2%	25.5%	64.1%	19.9%	58.9%	15.7%
Industrial	45.1%	11.7%	57.3%	22.3%	51.6%	16.0%	45.4%	11.9%
IT	41.6%	8.3%	53.3%	17.1%	46.9%	11.5%	39.7%	7.9%
Material Science	37.0%	8.8%	53.2%	21.5%	41.3%	11.8%	35.6%	8.3%
Optics	44.8%	9.2%	53.4%	22.1%	48.8%	12.2%	41.6%	8.4%
Packaging	53.3%	12.8%	62.3%	26.4%	59.9%	16.9%	54.6%	13.0%
Physics	61.5%	7.2%	67.2%	22.5%	63.3%	10.3%	56.7%	6.1%
Tools	45.1%	10.3%	52.9%	17.4%	50.4%	13.7%	45.5%	10.2%
Transportation	59.8%	17.8%	71.2%	29.7%	65.0%	22.9%	60.3%	18.4%

D. Validation of the Imputed Keyword Search Data Set

As noted above, the imputed keyword search data set is not derived from actual knowledge of the search method used to find each of the cited references,[113] and citations are included in the data set only if they are attributed to searches involving low-frequency keywords.[114] To validate the relevance of the data set to the characteristics studied in this Article, I studied actual prior art search records for a smaller sample of patents and compared the performance of single-keyword searches using keywords of different frequencies.

1. Sample Analysis of Examiner Search Strategy and Results Reports

In the examination of actual prior search records, this study utilized the image file wrappers that have become available on the

113. *See supra* text accompanying note 93–94.
114. *See supra* text accompanying note 98.

Patent Office website for the most recently issued patents.[115] The study compared a random sample of 633 citations from patents issued between January 1, 2006, and May 1, 2007,[116] and their associated conjectural keywords, with the logical search queries listed in the citing patent's Examiner's Search Strategy and Results ("ESSR") reports. The ESSR reports list each of the logical queries sent to the search engine and the number of hits returned in response in connection with the prior art search for a given patent application. As shown in Table 7, in 223 (35.2%) of the cases found, the conjectural keyword as an essential term in at least one of the search queries listed in the ESSR report(s) for the citing patents. This is evidence of the unsurprising fact that there is a substantial but not conclusive association between membership in the imputed keyword search data set and use of the keyword as a search term by an examiner. As Table 7 also shows, this association is reflected in the similar prevalence of co-classification among citations imputed to keyword searching, whether on the basis of single-keyword query results alone or in conjunction with the ESSR reports.

Table 7. Relative Prevalence of Co-classification

	All Imputed to Keyword		Matching in ESSR Sample		All Citations		Spurious in ESSR Sample	
Total Citations	7,313		223		3,397,179		410	
Same Class	4,080	55.8%	130	58.3%	1,427,130	42.0%	182	44.4%
Same Subclass	1,329	18.2%	28	12.6%	272,228	8.0%	40	9.8%

Relative prevalence of co-classification among (1) citations imputed to keyword searching; (2) citations imputed to a keyword that appears in an Examiner's Search Strategy and Results (ESSR) report; (3) all citations from patents issued between 1/1/2006 and 5/1/2007; and (4) citations imputed to a keyword that does not appear in any ESSR report

The ESSR reports do not identify any of the patents that were read and cited by the examiner as a consequence of the keyword

115. *See* United States Patent & Trademark Office, Patent Application Information Retrival (PAIR), http://portal.uspto.gov/external/portal/pair (last visited Apr. 25, 2009).

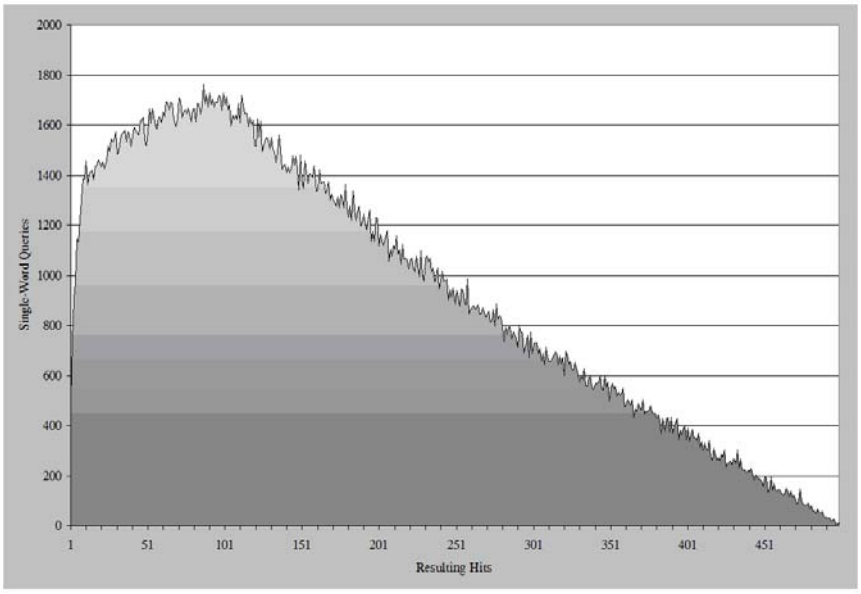
116. I focused on the most recently issued citing patents because many of the image file wrappers for patents issued in 2004 and 2005 appeared to be incomplete. *Cf.* Cohen, *supra* note 93, at 213 n.39 (noting inaccuracies in and omissions from online image file wrappers).

search.[117] Even in the absence of such data, however, it seems reasonable to assume that an examiner's use of the keyword in a search term and the examiner's subsequent citation of one of the hits resulting from that search are often causally related events.[118] Given this assumption, the imputed keyword search data set can be accepted as evidence that co-classified prior art is more prevalent among references found through keyword search than those found through other methods.

2. Distinguishing Power of Single-Keyword Queries

The histogram in Figure 3 illustrates the distribution of hit counts when searching the PatFT database for single-keyword queries using each of the 354,984 words in the Moby SINGLE.TXT dictionary.[119] As Figure 3 indicates, the vast majority of words in the English language appear in between 51 and 500 patents.

Figure 3. Distribution of Hit Counts (i.e., frequency of occurrence in patents in the PatFT database) Among the 354,984 Words in the Moby SINGLE.TXT Dictionary



117. *See supra* text accompanying note 94.

118. In particular, the time constraints forced by examiners are likely to discourage redundant search strategies. *See supra* note 9 and accompanying text.

119. Project Gutenberg, http://www.gutenberg.org/dirs/etext02/mword10.zip (last visited Apr. 25, 2009) (SINGLE.TXT available after extracting from .zip archive).

My focus on low-frequency keywords (i.e., those having 2–51 hits) was motivated by the general observations that search engine users tend to browse only the first part of a list of results when the list is lengthy,[120] and that short search engine queries tend to be effective only when the keywords are very specific.[121] To test these observations with respect to searches in the PatFT database, I evaluated the expected performance of a single-keyword search in locating patent prior art as a function of the keyword's frequency in PatFT.

To quantify the performance of a search, this study uses a standard information-theoretic measure of the partial information provided by the search result about the identities of the patents that were actually cited. The information content of a keyword search result for patent P in which k of the n keyword hits to earlier-issued patents were actually cited by P is given by

$$H = \log_2 \frac{\binom{N}{K}}{\binom{n}{k}\binom{N-n}{K-k}},$$

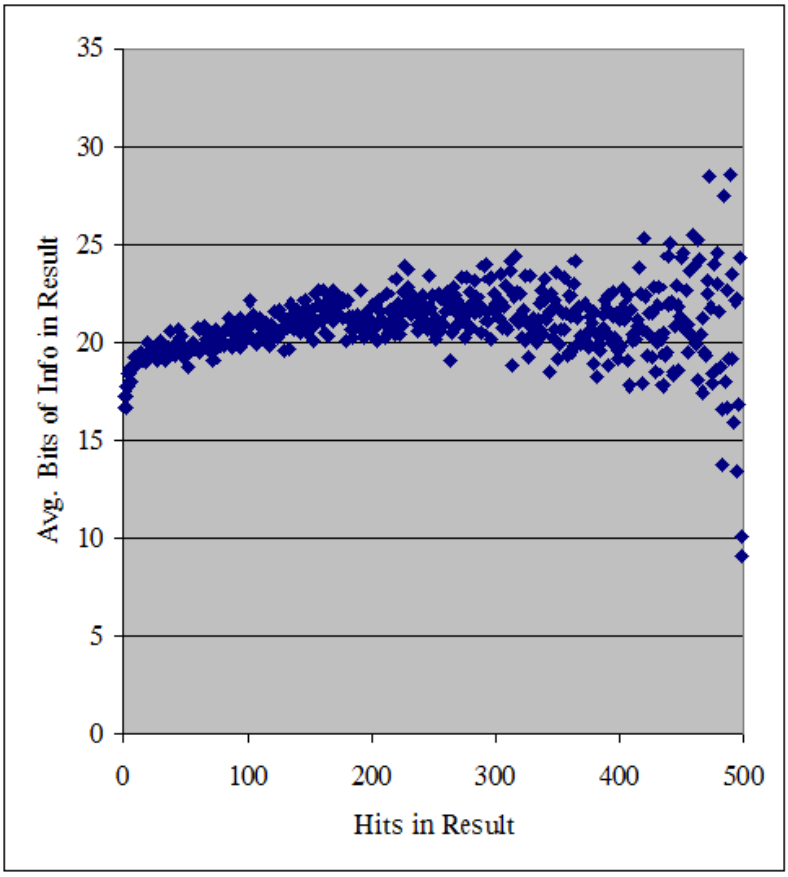
where N is the number of earlier-issued patents represented in the base citation data set and K is the number of citations in the base citation data set in which P is the citing patent.

Figure 4 shows the average information content of keyword search results for each value of n, $2 \le n \le 500$. 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. This finding indicates that concerns regarding the precision of search results arising from this study of low-frequency keywords apply with similar force to more general classes of keyword searches.

120. *See, e.g.*, B.J. Jansen et al., *Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web*, 36 INFO. PROCESSING & MGMT. 207 (2000) (finding that 58% of search engine users view only the first page of results).

121. *See, e.g.*, Nega Alemayehu, *Analysis of Performance Variation Using Query Expansion*, 54 J. AM. SOC'Y INFO. SCI. & TECH. 379, 380 (2003); K.L. Kwok, *Higher Precision for Two-Word Queries*, PROC. OF THE 25TH ANNUAL INT'L ACM SIGIR CONF. ON RES. & DEV. IN INFO. RETRIEVAL 395, 395 (2002). *But see* Caroline M. Eastman, *30,000 Hits May Be Better Than 300: Precision Anomalies in Internet Searches*, 53 J. AM. SOC'Y INFO. SCI. & TECH. 879, 880 (2002) (describing "anomalies" where the first of a large set of search results is more precise than the smaller set of results from a more focused query).

Figure 4. Expected Number of Bits of Information Contained in a List of Results Obtained Through a Single-keyword Query to PatFT, Given the Number of Hits Appearing in the List



The approach of focusing on low-frequency keywords is also supported by the correspondence between the sharpest increases in the percentage of citations imputed to keyword search and the critical periods of search technology implementation in the Patent Office. As Table 1 indicates, this trend is exhibited by both imputed data sets, but the low-frequency keyword set accounts for most of the observed increase.

V. POTENTIAL IMPROVEMENTS TO KEYWORD SEARCH

This study data indicate that the Patent Office's classification system is not being utilized in accordance with its new role as an adjunct to keywords in the formulation of search queries. As discussed above, such a role requires at least some tradeoff of recall

for precision in search results,[122] but Table 6 shows a net increase in the recall of classes and subclasses when used in conjunction with keyword search. Given the severity of the time constraints facing examiners in browsing the patents retrieved through automated search,[123] there is a pressing need to develop and implement auxiliary information retrieval systems to improve the precision of automated search results. Fortunately, in recent years a considerable number of public initiatives and research findings have emerged with the potential to improve the performance of the Patent Office's search technology.

A. Community-Based Prior Art Search Programs

In the past decade, the Patent Office's provision of Web access to the PatFT database has coincided with, and some cases facilitated, the formation of various Web-based communities of interest around a shared desire for improvements in patent quality.

Since 1991, the Software Patent Institute, "a nonprofit corporation formed to provide prior art related to software technology with the intention of improving the patent process,"[124] has sought to address longstanding concerns about the underutilization of non-patent prior art in the examination of software patent applications.[125] Through the collaborative contributions of its software industry, academic members, and other copyright owners, the organization has assembled an extensive online collection of old software documentation, academic literature, and defensive disclosures that could serve as software prior art.[126] The database opened for free public access on the organization's website in 1995.[127]

In 2000, software book publisher Tim O'Reilly and Amazon.com founder Jeff Bezos founded a private company, BountyQuest, which offered cash rewards to anyone who could find prior art invalidating

122. *See supra* Parts II.C–D.

123. *See* Graf, *supra* note 30, at 502 (describing patent examiners as "overburdened"); *supra* note 9 and accompanying text.

124. *See* Software Patent Institute, About SPI, http://spi.org/about-spi.jsp (last visited Apr. 25, 2009).

125. *See* Andrew Chin, *Computational Complexity and the Scope of Software Patents*, 39 JURIMETRICS 17, 28 & n.51 (1998); David R. Syrowik & Roland J. Cole, *The Software Patent Institute and the Challenge of Software-Related Patents*, 73 MICH. BAR. J. 544, 544 (1994) (discussing the failure to fully utilize academic literature and technological advances over using previously issued patents when examining software patent applications).

126. *See* Software Patent Institute, *supra* note 124.

127. *See id.*

any of twenty-three patents.[128] The company invited other companies and individuals interested in invalidating specific patents to post bounties at bountyquest.com, paying fees and commissions for the privilege.[129] The venture was eventually abandoned,[130] but served as a proof of concept that helped inspire subsequent "open-source" efforts to involve the public in the search for prior art to invalidate issued patents, including the Electronic Frontier Foundation's Patent Busting Project.[131]

"Peer-to-Patent," a pilot project spearheaded by Beth Noveck launched in 2007,[132] is an effort to develop a public online community around the task of assisting examiners in prior art search.[133] The program has obtained the cooperation of the Patent Office and the consent of various high-volume applicants for software patents to implement a "community patent review process" that supplements the agency's usual patent examination procedure.[134] Consenting software companies may submit their patent applications simultaneously for Patent Office examination and for posting on the Peer-to-Patent website, where the public can view the applications and submit prior art.[135] Applications so submitted are entitled to accelerated examination.[136] Communities of Peer-to-Patent users may form around particular applications or groups of related applications, facilitating the sharing of comments, related references, tags, ratings, and other metadata.[137] For example, the system may inform users that "people who submitted prior art for this patent also read patent X," or that a previous user labeled a particular device classified under Class 482 Exercise Devices as an "elliptical

128. Sabra Chartrand, *A Web Site Invites Bounty Hunters to Disprove Ownership of Ideas, Even Those of Its Founders*, N.Y. TIMES, Oct. 23, 2000, at C8.

129. *See id.*

130. *See* Anne Marie Squeo, *Old Records May Turn Up to Kill Patent*, TORONTO STAR, Jan. 26, 2006, at D16 (stating that BountyQuest "shut down its service in late 2002").

131. Electronic Frontier Foundation (EFF), The Patent Busting Project, http://w2.eff.org/patent/wp.php (last visited Apr. 25, 2009). EFF invites prior art contributions from the public with the aim of challenging the validity of "the worst offending patents" through reexamination proceedings. *See id.*

132. Peer to Patent, Community Patent Review, http://www.peertopatent.org (last visited Apr. 25, 2009).

133. *See* Beth Simone Noveck, *"Peer to Patent": Collective Intelligence, Open Review, and Patent Reform*, 20 HARV. J.L. & TECH. 123, 144 (2006).

134. *Id*. at 146.

135. *Id.*

136. *Id.* at 145.

137. *Id.* at 146.

machine."[138] Prior art submitted by the public in this way is presented to the examiner for consideration in the same manner as Rule 99 third-party submissions,[139] with the added advantage that no fee is required.[140] After one year of the site's operation, Noveck noted: "Though it is too early in the program to contend that these encouraging results prove the utility of extending open peer review to the patenting process, these cases appear to support the notion that Peer-to-Patent participants are qualified to provide relevant information to the system."[141]

Peer-to-Patent appears to have been developed independently of WikiPatents.com, a Web portal established in August 2006 by Peter Johnson and Kevin Hermansen.[142] WikiPatents provides access to a privately maintained database of U.S. patents and prosecution histories, as well as an online community that allows the public to comment on patents, post prior art references, and add search tags.[143] While WikiPatents's coverage seems more comprehensive than Peer-to-Patent—it includes all patents since 1976[144]—it does not include newly filed applications and there is no agreement by the Patent Office to review prior art submitted through the site. WikiPatents's founders express the hope, however, that the information on their site will be useful to examiners and other interested parties in evaluating patents and patent applications.[145]

138. *Id.*

139. 37 C.F.R. § 1.99 (2008).

140. *See* Noveck, *supra* note 133, at 145. The fee for a Rule 99 submission is currently $180. *See* 37 C.F.R. § 1.17(p) (2008).

141. Beth Simone Noveck, *Peer-to-Patent: Collaborative Patent Examination*, TOKUGIKON, May 21, 2008, at 77, 89, http://dotank.nyls.edu/communitypatent/TokugikonEnglish.pdf.

142. *See* Kevin Hermansen, *WikiPatents Enables Community Patent Review*, ARTICLECITY.COM, Sept. 19, 2006, http://www.articlecity.com/articles/legal/article_711.shtml.

143. WikiPatents, Patent Reviews, PDFs, and File Histories, http://www.wikipatents.com/ (last visited Apr. 25, 2009).

144. The WikiPatents database begins with U.S. Patent No. 3,930,270, which was granted in the final days of 1975. *See* WikiPatents, Community Patent Review, http://www.wikipatents.com/faq.php (last visited Apr. 25, 2009) (listing the earliest patent in the database on the bottom left of the page); *see also* U.S. Patent & Trademark Office, United States Patent Database Search, http://patft.uspto.gov/netahtml/PTO/srchnum.htm (search for "3,930,270") (last visited Mar. 20, 2009) (showing that the patent was granted on Dec. 30, 1975).

145. *See* WikiPatents, Community Patent Review, http://www.wikipatents.com/faq.php (last visited Apr. 25, 2009).

B. Advances in Information Retrieval Research

The information retrieval research community has long regarded the collection of U.S. patents as a subject of special interest, due to the critical economic and scientific importance of efficient and accurate search, as well as the patent document's distinctive use of structured metadata that is amenable to novel data processing approaches.[146] In addition, research on information retrieval from the Web presents techniques that appear to be applicable in the patent prior art search context. The research literature in this area is far too vast to review here, but a few promising techniques are worth highlighting.

Drawing on the classification powers of humans, collaborative filtering systems (also referred to as "recommender systems") accumulate the preferences of a multitude of individual users to produce a list of items that the seeker may like.[147] As noted above, Noveck hopes that her Peer-to-Patent project will serve as a proof of concept in support of the use of recommender systems in connection with prior art search both inside and outside the Patent Office.[148] Citations between patent documents are a particularly significant and stable collection of examiner recommendations. Citation analysis, such as the PageRank method employed by Google[149] in identifying the most authoritative sites on the Web, has been shown to be helpful in refining search results.[150]

"Lexical semantic indexing" is an approach to searching that attempts to retrieve texts that match the meaning of the query, not just those that match the literal text of the query.[151] The technique

146. *See generally* Noriko Kando, *What Shall We Evaluate?—Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in the Patent Searching and Parent Attorneys*, http://research.nii.ac.jp/ntcir/sigir2000ws/sigirprws-kando.pdf (last visited Apr. 25, 2009) (describing the use of searchable abstracts by the United States, Japanese, and European patent offices).

147. For surveys, see, for example, Loren Terveen & Will Hill, *Beyond Recommender Systems: Helping People Help Each Other*, *in* HUMAN-COMPUTER INTERACTION IN THE NEW MILLENNIUM 487, 487–509 (John H. Carroll ed., 2001); Paul Resnick & Hal R. Varian, *Recommender Systems*, COMM. ACM, Mar. 1997, at 56–58.

148. *See supra* text accompanying notes 132–41.

149. Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 30 COMPUTER NETWORKS & ISDN SYSTEMS 107, 109 (1998).

150. *See* 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, 793 (2007), *available at* http://if-lab.slis.tsukuba.ac.jp/fujii/paper/sigir2007.pdf.

151. *See* James F. Ryley et al., *Advanced Document Retrieval Techniques for Patent Research*, 30 WORLD PATENT INFO. 238, 238 (2008); Christopher G. Lucas, Patent Semantics: Analysis, Search and Visualization of Large Text Corpora 17–21 (Aug. 20,

uses clustering techniques to locate documents in a very high-dimensional vector space on which a search query can generate a measure of relevance.[152] Other approaches to patent clustering have also achieved improvements in search performance, including hierarchical Bayesian clustering,[153] Naive Bayes, K-nearest neighbors and support vector machine clustering.[154] A survey and comparison of these promising techniques was recently published.[155] Patent examiners, however, have yet to be convinced of the value of automated clustering tools.[156] Beyond the technical challenges involved in tailoring these advanced methods to prior art search, the greater difficulty may lie in changing the habits of patent examiners and other end users.[157]

CONCLUSION

The Patent Office's classification system no longer governs the physical organization of paper documents in the search room, but it is a permanent feature of the patent system and represents an important body of collective knowledge that can powerfully aid a prior art search. This study indicates that users of text search have not yet been able to take full advantage of the classification system's ability to resolve lexical ambiguities that result in overinclusive search results. As an information retrieval system, the classification system continues to be used primarily as an aid to recall, rather than to enhance precision.

2004) (unpublished M.E. thesis, MIT) (on file with Barker Library, MIT), *available at* http://dspace.mit.edu/ handle/1721.1/33146.

152. *See* Ryley, *supra* note 151, at 239–41; Lucas, *supra* note 151, at 22. Implicit in this approach is the hypothesis that closely clustered documents are relevant to the same queries. C.J. VAN RIJSBERGEN, INFORMATION RETRIEVAL 45–47 (2d ed. 1979) (discussing the "cluster hypothesis").

153. Naomi Inoue et al., Speaker, ACM SIGIR 2000 Workshop on Patent Retrieval, Patent Retrieval System Using Document Filtering Techniques (July 28, 2000), http://research.nii.ac.jp/ntcir/sigir2000ws/sigirprws-inoue.pdf.

154. C.J. Fall et al., *Automated Categorization in the International Patent Classification*, 37 ACM SIGIR FORUM 10, 10 (2003); Leah S. Larkey, *Some Issues in the Automatic Classification of U.S Patents* (1998), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.210 (select PDF icon on right side of page) (exploring k-nearest-neighbors classification and Bayesian classifiers).

155. Yuen-Hsien Tseng et al., *Text Mining Techniques for Patent Analysis*, 43 INFO. PROCESSING & MGMT. 1216, 1216–43 (2007).

156. *See* Harold Smith, *Automation of Patent Classification*, 24 WORLD PATENT INFO. 269, 271 (2002) (reporting that most USPTO examiners found automated classification tools too time-consuming and difficult to use).

157. *Cf.* text accompanying note 13 (noting resistance of some examiners to search automation).

If, as patent examiners are advised, keyword search is unreliable as an exclusive method for locating patent prior art, there appears to have been a systemic failure to utilize the classification system fully to address this problem. And if, as some commentators suggest, there are mounting deficiencies in the classification system,[158] keyword search is not adequately enabling searchers to transcend them. Finally, if both systems are flawed, then at least to some extent, the blind are leading the blind. New approaches are needed to achieve a better search for tomorrow.[159]

158. *See supra* notes 61–62 and accompanying text.

159. *See also Edited & Excerpted Transcript of the Symposium on Ideas Into Action: Implementing Reform of the Patent System*, 19 BERKELEY TECH. L.J. 1053, 1071 (2004) (comments of former USPTO Director Q. Todd Dickinson) ("[T]he examiners . . . need greater access to prior art, and they need better search tools. They have great search tools and they need even better search tools.").