PATENTCAFE®

SuperBoolean™ Patent Search

FOR MORE INFORMATION ON ICO™ PATENT SEARCH, SEE **THIS PAGE**

Heuristic Boolean Patent Search

Comparative Patent Search Quality / Cost Evaluation

"SuperBoolean" vs. Legacy Boolean Search Engines

by Andy Gibbs,
CEO, PatentCafe®

www.PatentCafe.com

Keys: Patent search, Boolean, Heuristic, relevancy rank, expert system, artificial intelligence

Abstract

This paper explores the application of an expert system to Boolean patent searching. Specifically, it will introduce skilled researchers to the next technological evolution of search methodologies that apply Heuristic Boolean methods to reduce cost, increase efficiency, and enhance search results quality.

Boolean search methodology, otherwise known as “keyword searching”, only extracts documents from a database that literally match the search query, but Boolean engines have no capability of determining which of those documents are of highest interest to the researcher. In an effort to overcome this limitation, modifications to simple Boolean engines have emerged, including truncation, proximity searching, nested complex query capability, and wildcarding.

But even with these enhancements, at best, Boolean searching remains little more than an iterative process of applying a query construct to a database in order to (a) extract a reasonable number of documents within (b) a reasonable amount of time, in order to (c) produce the most relevant documents supporting the search objective.

At worst, the restrictive nature of Boolean search methods inordinately increase direct and indirect search costs, and establishes a false confidence in search results quality that increase exposure to long term legal and commercial risks. The patent documents that Boolean engines inherently miss, sometimes discovered years later by other researchers, often establish the true (high) costs attributable to Boolean searches.

This paper examines how next generation Heuristic Boolean search methods can more quickly yield the most relevant documents, mitigate long-term risk associated with poor quality results, and reduce the direct, as well as hidden costs attributable to legacy keyword search engines. When the artificial intelligence of Heuristics is applied to Boolean patent searching, even novice researchers can quickly achieve reliable search results.

A future invalidity search is the ultimate quality test of today’s patentability search.

Relying on the best search tools and processes today is critical. *The future invalidity search performed when millions or 10s of millions of dollars are at risk, and can easily challenge and outperform the earlier patentability search because:*

1) liberal budgets for invalidity searches allow significantly more investment in search labor (higher cost), and

2) invalidity searches rely on search technology advancements which have evolved since completing the patentability search.

Conclusion Summary

The demands to perform a patent search that attempts to identify <u>all</u> of the relevant documents within the scope of available resources (time, budget, computing time, a given patent data quality) keep researchers reliant on the time-honored practice of crafting a lengthy, complex Boolean search string. But it’s been shown that such restrictions, although they produce relevant patents in a final results list, more dangerously drop an increasing number of relevant patents <u>that should have been included</u> in the final search report.

The application of heuristics such as a Latent Semantic Analysis / artificial intelligence **expert system** allows a patent researcher to use a less restrictive Boolean query, and obtain the Best-First search results list containing the highest quantity of documents more relevant to the search. Researchers are then able to manage very large search results lists without filtering the list to a more manageable quantity by using more keywords.

The results of applying heuristics to Boolean patent searching are faster time to identify the most relevant patents, but more importantly, the identification of the largest number of relevant patents that will serve as acceptable prior art.

Introduction

In the age of web-based access to nearly infinite information, the need to scour technical information for patentability of an invention has gone from "luxury" to "mission-critical". The development of more advanced search engines is moving at breakneck speed.

The validity of a patent may not be tested for at least three to five years after filing an application. Although "unfair", that's also the time when a defendant will invest considerable time and money, using the search technology of tomorrow to identify un-cited prior art that the first patentability search missed, often invalidating the patent.

Boolean patent searching, otherwise known as "keyword" searching, has long been the trusted method of ferreting out patents that may teach the present invention. Clearly, with infringement suits being filed at the rate of more than 10-per-day, and infringement awards more frequently hitting the $1/2 billion mark, Boolean search methods as a whole no longer live up to the reputation of being a reliable, long term safeguard to patentability.

Given that even today's best patent search technology in the hands of a skilled researcher will be tested with superior tools mining more data in the future, the hypothesis is that the use of traditional legacy patent search tools, specifically free and commercially available Boolean patent search tools, will result in an increasing number of invalidated or otherwise successfully challenged patents based on the later discovery of prior art that the Boolean engine missed.

In his IEEE paper "The Combinatorics of Heuristic Search Termination for Object Recognition in Cluttered Environments"[1], Grimson illustrates the total cost and performance benefits of applying an *expert system* to help the researcher more effectively find relevant documents in very large data collections. Heuristic processes do carry over to searching the highly-complex data contained in patent documents, but how so?

The performance metrics of research technique primarily include: search quality (discovery of the most relevant patents); speed; cost; reliability; and, durability (probability of the results surviving testing by the next generation of patent search engines).

Reasonable commercial standards are usually applied to patent searching. When is a patent search "good enough"? Usually, when a professional spends the specified time, and consumes the resources allocated to the search. However, while the searcher may find "good patents", good patents are not as important as "good patents not found", especially if those un-found patents actually teach more closely related prior art than those cited in the researcher's final search report.

Boolean search engines are designed to produce positive results (you see only what your keywords you ask for), but they are incapable of returning patents it 'knows you are looking for'.

Search strategies can be classified as "complete" or "heuristic". In "Optimization by learning and simulation of Bayesian and Gaussian networks" [2], Larranaga, Etxeberria, Lozano, and Pena explain that the underlying idea in the *complete search* is the systematic examination of <u>all</u> the possible points of the search space.

Patent data is voluminous, and as a homogeneous collection constitutes one of the world's largest data sets. There are about 147,000 US Patent Classifications (incorporating classes and subclasses). PatentCafe's database of 25 million patent documents contains about 2.2 million distinct invention concepts, disclosed in about 1/4 billion pages of full text. Performing a complete Boolean search using the number of possible keyword combinations necessary to examine this volume of data to find relevant prior art is economically prohibitive.

1. *W. Eric L. Grimson, "The Combinatorics of Heuristic Search Termination for Object Recognition in Cluttered Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 9, pp. 920-935, Sept., 1991.*
2. *Optimization by learning and simulation of Bayesian and Gaussian networks, by P. Larranaga, R. Etxeberria, J. A. Lozano, J. M. Pena. Technical Report EHU-KZAA-IK-4/99. Intelligent Systems Group, Dept. of Computer Science and Artificial Intelligence, University of the Basque Country http://www.sc.ehu.es/isg 31 December 1999*

Further, a *complete* search requires the crafting of a complex Boolean string to optimize the discovery of all relevant patent documents as instructed by Larranaga et. al., and still cannot reasonably examine all possible points in the space. With almost 100% certainty, with almost every search, one can be assured of missing important documents.

This paper will explore how Latent Semantic Analysis technology can be applied as an “expert system”, raising traditional Boolean patent searching to Heuristic, or “SuperBoolean™” searching. It examines how the cost, performance, and quality metrics overcome the inherent shortcomings of Boolean patent search engines.

Problem: Boolean Patent Search Methods

In order to frame the importance of considering a Heuristic Boolean search process, the critical problems with Boolean searching that hope to be overcome must first be discussed.

Let’s take a look at the fundamentals of Boolean patent searching. Interestingly, the entire process of Boolean searching is pre-disposed to only deliver the results you unwittingly ask for via your Boolean search query string - the computer simply returns the documents that match what you ask for. Clever Boolean search strategies, along with the application of proximity or truncation enhancements may occasionally produce some unanticipated results, but the results still literally correlate to your keyword request. The unanticipated results may or may not have any relevance to the subject matter being searched.

<u>Problem 1: Whether or not a searcher is a subject matter expert or skilled research professional, they cannot reasonably craft a search string that examines all relevant documents in a patent database.</u>

The detail of a Boolean search process includes, to one degree or another, follows this general sequence:

1) Review of the subject matter to be researched, and the development of search keywords or keyword strings that may find responsive documents,
2) Development of a search strategy consistent with the commercial parameters (cost, time, thoroughness, quality, or other directed metrics),
3) Executing the search - and conducting iterative combinations of planned keywords or keyword strings,
4) Obtaining a results “hits list” of sufficient breadth so as to contain documents of interest. This hits list may contain tens of thousands of documents of equally-weighted relevancy (all hits equally satisfy the literal Boolean request).
5) Narrowing of search results to a smaller hits list that can be examined by the researcher. The narrowing process requires the addition of more keywords or additional Boolean operators, filters or limiters.
6) Compiling a smaller, final search results list suitable to incorporate in a final search report.

<u>Problem 2: Vagaries of Selection of Words, Phrases, and Boolean Operators</u>

Since a Boolean engine will only return documents that the researcher asks for, the problem is not what patents match their keyword query, but rather what highly relevant patents are missed because the researcher did not use <u>different</u> words in the search query.

Given the more than 200,000,000 pages of patent text, it’s unlikely that any researcher would know all of the possible words, word sequence, or word combinations to use to discover all patents relevant to the search. This problem is compounded by patent writers who intentionally “submarine” a patent application by using obscure words, or even inventing a new lexicon.

Notwithstanding these practically insurmountable language issues, the researcher will nevertheless craft long and complex queries in the hopes of discovering all relevant patents. But the addition of more words to a query <u>amplifies</u> the statistical probability of missing important patents - *just the opposite of what logical thinking would suggest.*

For instance, a researcher who identified 10 important keywords to incorporate into a search strategy would need to perform a number of search iterations approaching 10 factorial (3,628,880 searches using all combinations of the keywords). Of course, this is unreasonable, as well as economically inefficient. This supports the thesis that a *complete* search can never evaluate the entire space [Grimson, 1991].

The researcher would end the search at a prescribed time (budget), or when they believed their search results list contained a sufficient number of relevant patents.

The end result is a list of documents that the researcher has carefully defined, or in other words, the researcher only sees the documents they have actually asked for. However, it's important to understand that this process does not result in a list of relevant documents *not* requested by the researcher.

A paradox of Boolean searching is that as more keywords and Boolean operators are strung together as a complete "search strategy" to more accurately identify relevant patents, the more patents this process actually excludes from the list of relevant prior art.

Evidence suggests that as the keywords in a search string increase, the inaccuracies increase exponentially [3]. The case study saw a similar exponential increase in inaccuracies. "Inaccuracies" are defined as relevant patents that were eliminated from the search results list because of non-compliance with a larger, more restrictive Boolean query string.

Further, the premise that "AND-ing" keywords will retrieve more pertinent (and thus more relevant) patents is fundamentally flawed. Just because the patent author chose to use one of the words does not assure they will use the other(s). Each author tends to develop their own language or word-selection pattern. If the second term is replaced with a synonym or phrase, then possibly the entire portfolio of that patent author may fail to be discovered. Each "AND" introduced to a Boolean logic increases the probability that one of the terms will fail to occur, so no consistency of increasing relevance can be assumed.

Fig. 1.

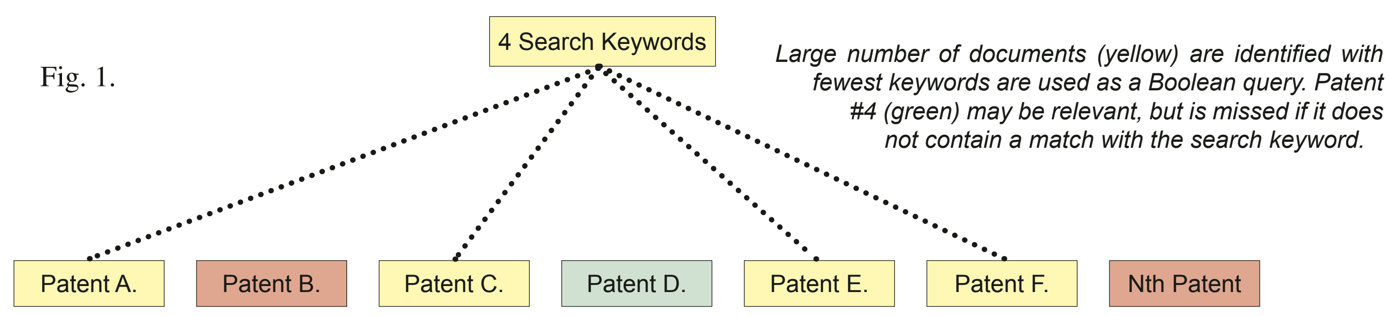
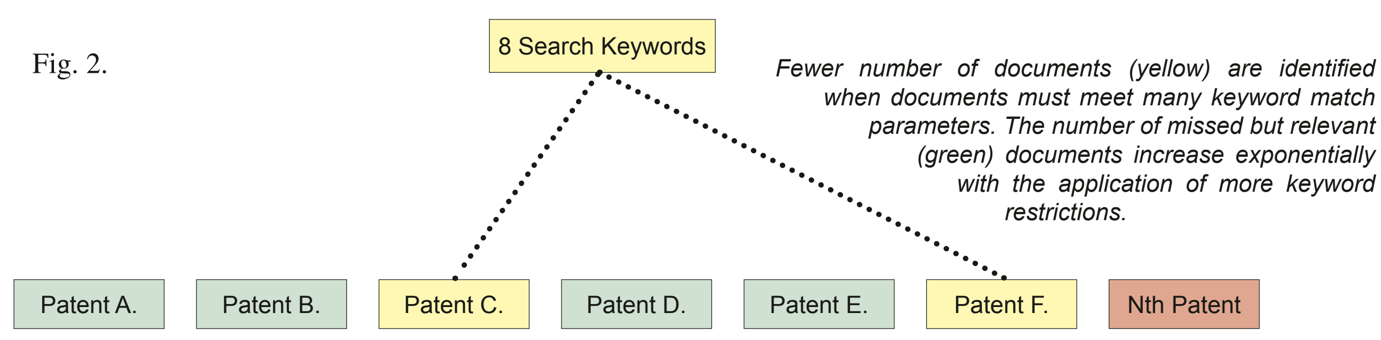


Fig. 2.



Most professional researchers are either unaware of, or ignore these flaws as an inherent drawback of the process over which they have no control - but must use nevertheless.

3. *The Cost of Choosing the Wrong Model in Object Recognition by Constrained Search, W. ERIC L. GRIMSON, MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139 Received May 22, 1990. Revised September 27, 1991.*

Problem 3: Missing Data or Critical Patent Data Errors

The US Patent and Trademark Office indicates that "a particularly relevant document not identified by one search strategy but identified by another can be deemed a 'critical error' for performance and/or quality review".

Because of errors encountered in the digital scanning and conversion to searchable text using the OCR process, many (electronic) patent documents contained in USPTO database have missing claims and other inaccurate, misplaced, or missing data. The USPTO makes its best effort to ensure accuracy of the Full-Text database, but that database is not the official electronic record.[4] Other patent issuing authorities have similar problems (EPO, WIPO, and so forth), and this poor quality data is actually the same data distributed to its commercial patent data customers.

A Boolean search that requires certain words to appear in specific sections of the patent document will miss finding patents for which the data is missing (e.g.: if the word "device" must appear in the claims text, responsive patents that contain no claims text will not be included in the search results list). This is considered a critical error.

Patent data quality that meets 4-Sigma or 5-Sigma quality standards is unsatisfactory when considering the economic investment or risk tied to finding all relevant patents. Patent data meeting 4-Sigma will have up to 6,200 errors per 1,000,000 operations. 5-Sigma can have up to 233 errors in 1,000,000 operations. A database of 3,000,000 patents will have a statistical error volume of 600 to 18,000 errors. There is a high probability of missing one of these patent documents with a Boolean query - and the probability increases exponentially as more words and operators are added to the Boolean search string.

Although at least one word is required to initiate a Boolean query, the probability of encountering missing data (and missing relevant patents) suggest that fewer Boolean filters will find more relevant patents.

The obvious drawback to using *fewer* words is that the results set is extraordinarily large - perhaps 100,000 or more responsive documents. Reading all 100,000 documents is not economically viable - so what process exists wherein fewer Boolean keywords can be used in the query, yet allow the researcher to view the most relevant patents?

There are other conditions in which Boolean-only searching will fail to discover relevant patent documents. They are less likely to be overcome by the use of Heuristics and are therefore not addressed in this paper.

Applying Heuristics to the Boolean Patent Search Process

A Heuristic is a particular technique of directing one's attention in learning, discovery, or problem-solving, otherwise known as an ***expert system***.

Relating to or using a problem-solving technique in which the most appropriate solution of several is found by alternative methods, heuristics are applied at successive stages of a program for use in the next step of the program.

Batali explains that the word "heuristic" is not used only to describe cases where a solution might not be found, but to describe cases where we want to find the best solution (according to some way to measure bestness). The measure of "bestness", and the assessment of a heuristic technique, is going to be relative to the domain, and to the specific job that problem solving is going to be applied to in that domain. [5]

An expert system, also known as a knowledge based system, is a computer program that contains some of the subject-

4. *Regarding Patent Data Quality, Donna Cooper, USPTO, http://piug.derwent.co.uk/archive/piug/piug-2003/0597.html*
5. *Batali, John, Associate Professor Department of Cognitive Science, University of California at San Diego. Cogsci 108b Lecture Notes, Fall 2000. http://www.cogsci.ucsd.edu/~batali/108b/lectures/heuristic.html*

specific knowledge approximating the equivalency of a human subject matter expert ***assuming it trained on the subject matter***. This class of program was first developed by researchers in artificial intelligence during the 1960s and 1970s and applied commercially throughout the 1980s.

The most common form of expert systems is a program made up of a set of rules that analyze information (in the current analysis, supplied by a Semantic database) about a specific class of problems, as well as providing analysis of the problem(s), and, depending upon their design, recommend a course of user action in order to implement corrections. It is a system that utilizes reasoning capabilities to reach conclusions. [6]

As it relates to patent searching, heuristics call on an expert system as an alternative to "Boolean-only" searching, Applied to a Boolean search, heuristics assist a researcher in most efficiently reducing the results set to identify the most relevant documents responsive to a search - without adding more keywords that increase the number of missed-but-relevant patents.

What constitutes and expert system? In the patent search world, it is a database that broadens, yet qualifies the patent search results. Expert Systems currently used in association with patent searching include:

- Thesauri or synonym database
- Semantic Analysis Concept Space / Artificial Intelligence

In briefly comparing these two expert systems, we can say that the Thesauri or synonym database relies on a human to create a lookup table of words, then match those words to related words. The reliability of the lookup table will depend on the attentiveness of the person maintaining the database. Unfortunately, patent writers are permitted to invent new words to describe their invention (lexicon), and use these words to intentionally frustrate discovery by Boolean search engines. The person maintaining a synonym database will not even know what words they don't know, so there is always a probability that even a thesaurus or synonym lookup table will be incomplete.

On the other hand, semantic engines incorporate artificial intelligence that actually learns a new lexicon on the fly. It has been proven to be more knowledgeable about finding documents that are closely related to the search query, even though none of the important keywords in the semantic query are contained in the most relevant documents.

The "problems" to be solved in the patent research process are many, depending on the researcher's objectives. Heuristics can address each of these problems, although some problems will be better solved than others. The problems, defined at the highest level, include:

1) Obtaining the highest quality search results (not missing relevant patents because of search process limitations),
2) Completing the research project while consuming less than the total available resources (time, cost)

Heuristics allow the researcher to obtain the Best-First Search results - a key process component to capturing the maximum number of relevant documents, while ensuring that the fewest number of relevant documents are missed. As we have discussed in the previous section, adding keywords to a complex Boolean search query increases exponentially the number of documents that will be missed by the researcher.

Conversely, using shorter Boolean search queries, comprised of fewer keywords, widens the potential search results "hits". A larger results set significantly increases the capture rate of relevant documents. (Refer to Fig. 1. above)

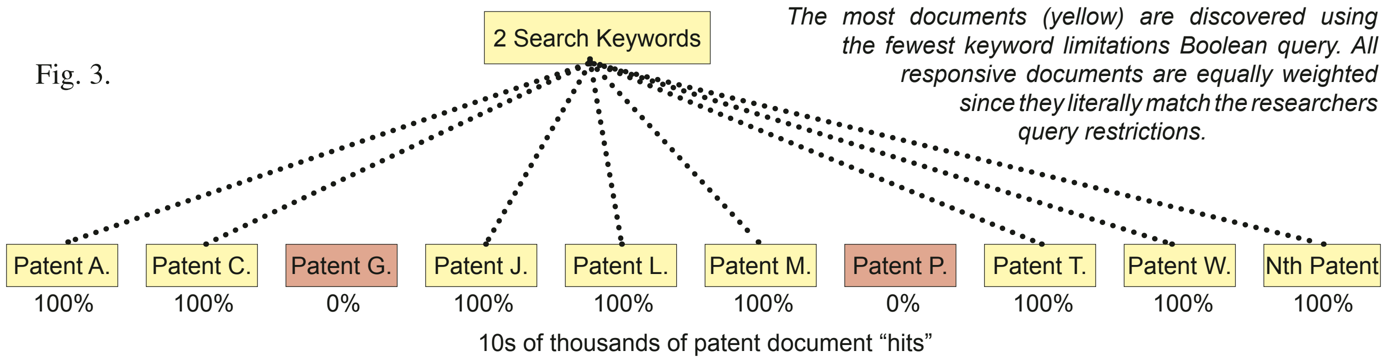
As most patent researchers know, a search that results in 10s of 100s of thousands of hits is quite useless. That is, unless an expert system is applies to the results set to bring the most relevant patent documents to the top of the results list.

PatentCafe's patent database has been indexed using a Latent Semantic Analysis ("LSA") search engine. Through the indexing process, the LSA engine has learned more than 2.5 million distinct invention concepts, and has mapped every patent against these concepts. The mapping involves the development of N-Dimensional Vectors for each concept - some vectors being long with many variants of the concept positioned along its length, with other vectors being shorter to express an invention concept containing fewer variants. The index is best defined as a database "Concept Space" containing only

6. Wikipedia; definition of Expert System, http://en.wikipedia.org/wiki/Expert_system mathematical expressions of each concept, and no human-readable words (as are found in traditional patent databases). Without expounding on the details of LSA or PatentCafe's Semantic database [7], suffice to say that the LSA database, when applied to a Boolean search process, serves as the trained *expert system*.

Fig. 3., when compared to Fig. 1. above, shows an exponentially larger search results set when only 2 keywords (fewer keywords) are used in the Boolean query. Obviously, the number of "hits" can be extraordinarily large since very few Boolean filters were applied to the database.

Fig. 3.



Since all responsive documents are equally weighted, no patent stands out as more relevant to satisfying the researcher's objective than any other.

This "problem" has been historically solved by continuing with a sequence of search iterations, each iteration beginning with an expanded Boolean search string. Each iteration defines more restrictive Boolean filters that results in fewer hits - until the number of patent documents is small enough so that the research can begin the manual review of the documents to identify those most relevant.

The application of heuristics results in an organization of the Boolean "hits' based on relevance to the Semantic query. Of course, all of the hits also satisfy the literal Boolean restrictions as well.

This heuristic approach is illustrated in the following Fig. 4.

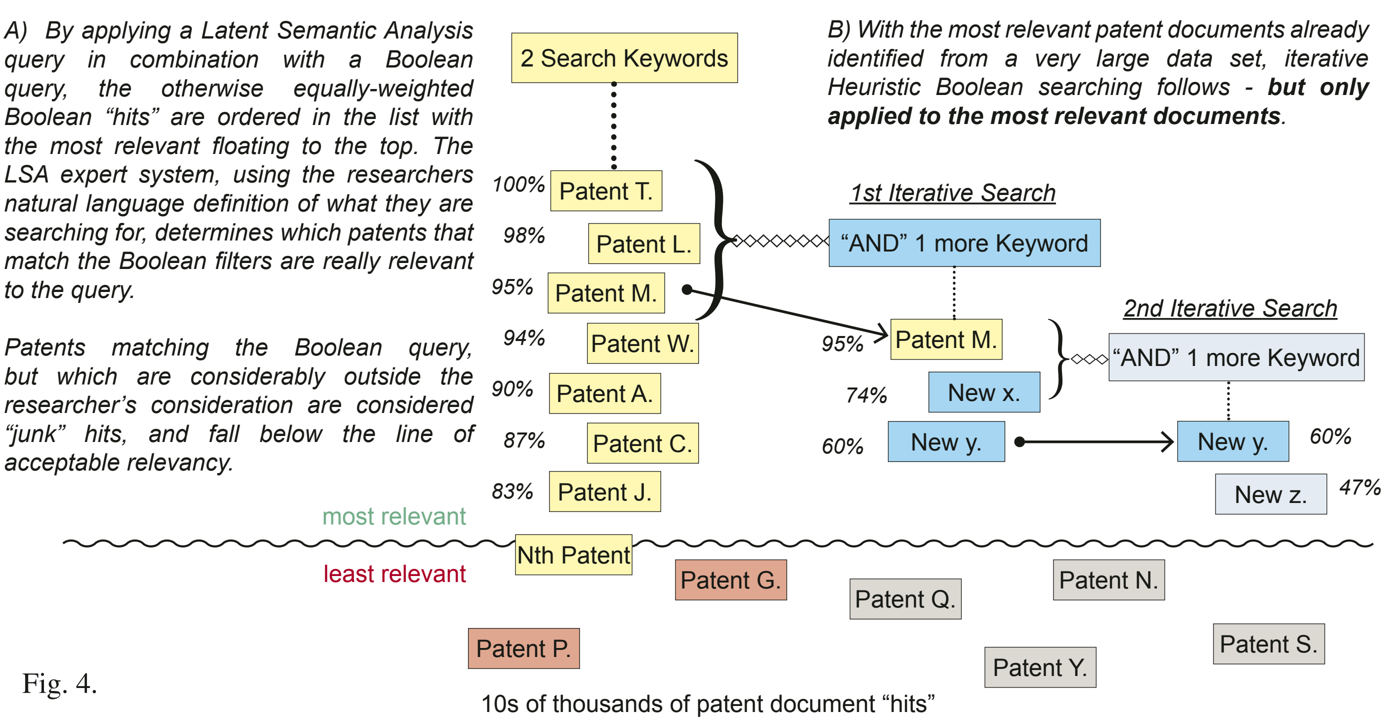


Fig. 4.

7. *Latent Semantic Analysis Search Engine – Conceptual Search and Discovery, Engenium Corporaton, http://www.patentcafe.com/actionitems/whitepapers/semantic_engine_whitepaper.pdf*

In this Case Study, the number of hits from the first search result exceeded 14,000 (only two Boolean keywords were used). Contained in the results set were highly relevant, somewhat relevant, and the expected inclusion of a large number of "junk" patent documents. A results list of 14,000 documents is, or course, too large for the researcher to begin manually viewing and qualifying the more appropriate patents - so the researcher would begin the iterative process of expanding the Boolean search strategy by adding more keywords of other restrictions.

Following the application of heuristics to the first results set as outlined above in Fig. 4., the process shows the speed and efficiency in narrowing the results to a final prior art list.

While it's shown that heuristics increases the overall quality of the search results, the process also shows that even with heuristics, the increase in the number of Boolean keywords in each successive iteration still results in some relevant documents being dropped from the final list, albeit fewer drops than with Boolean alone.

Case Study

In the following case study, one can readily see that the application of LSA Heuristics found the most relevant documents and brought them to the top of the 14,000 hits, even when only a small number of keyword terms were used.

The problems solved with the addition of heuristics were:

- Faster search: by percolating the most relevant documents to the top, the researcher found the best documents within minutes. The **Best-First Search** achieved results that otherwise could have literally taken hours,
- Reduced number of missed patents that were relevant to the search: by using fewer keywords (shorter Boolean query), more documents were included in the search results that would have otherwise been eliminated by a more extensive Boolean query.

But even with the additional heuristic qualifiers, some relevant patents were eliminated from the results set each time another Boolean keyword filter was introduced. Nevertheless, heuristics reduced the number of good patents that were eliminated by keeping the Boolean queries very short, and by reducing the total number of iterations.

Case Study Objective: The task called for the executing of a search process to identify the most relevant patents that would constitute a Prior Art objection to the patentability of an invention related to an improved nozzle design for a digital ink jet printer. This required the researcher to:

A) develop a natural language search query (heuristic) that will be applied to the Semantic database. This description defines the actual invention, and constitutes the instructions for the heuristic process. The query used was:

The improved nozzle geometry of a digital ink jet printer, such nozzle improvement providing for a more precise control of the dispersion pattern of the liquid ink. This improved control of the dispersion spray pattern provides for the printing of high resolution graphics, specifically including high definition photographic prints.

B) develop a keyword list that could be used in the crafting of the Boolean search strategy. The large number of possible words contained in the relevant documents were assessed, and the keywords that were simple and most obvious were selected as the starting point. The objective was to obtain the Best-First Search results list with the fewest number of relevant documents being eliminated from the results set.

C) conduct the search, initially using only two broad keywords + the heuristic query, and refine the results by expanding the Boolean query string with additional keywords.

D) refine the results list by adding another keyword to the Boolean query during each iteration.

Case Study Details:

The following search was conducted to illustrate the correlation between the increase in the number of keywords used in a Boolean search string with the rate at which highly relevant patents are dropped from the list or search results.

Keywords considered: A partial list of other relevant keywords discovered during the case study, but which were not used in the search process.

diameter	deposit dots	orifice	ejecting stream	compensation	output device
sharp edge	dot spread	volume	pixel printing	nozzle bore	chimney
chamfered edge	vertical banding	deflection angle	venturi spittoon	inkdrop generator	spitting
architecture	orificeless	shaped	channel	plus - others not recorded	

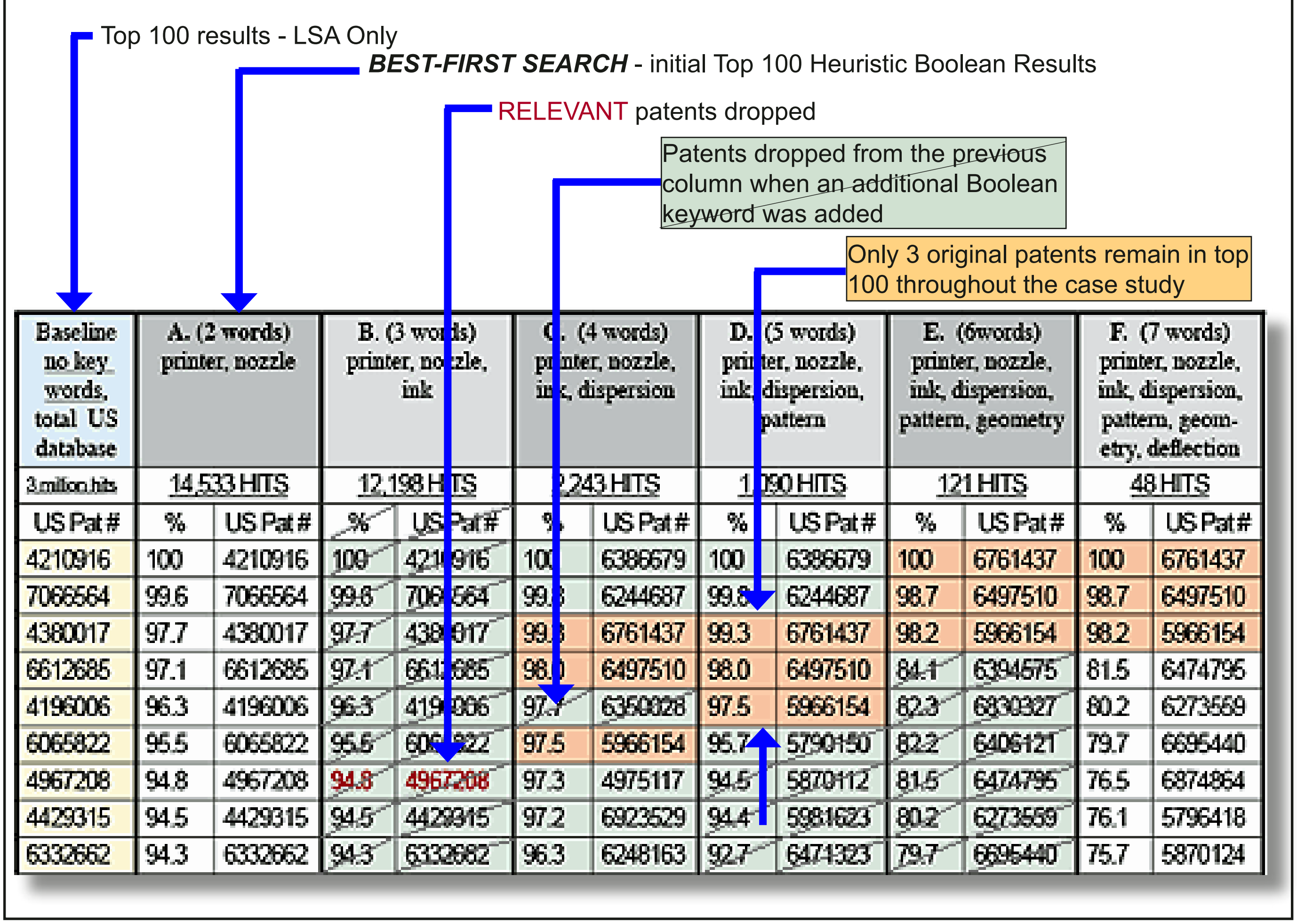
* *Examination of all possible keyword combinations would be N Factorial variations (worst case).*

Search Time - 1 hour aprox.: for purposes of the case study, the 48 hits shown in Column F. were obtained in about one hour. A researcher would have taken considerably more time in qualifying the appropriate keywords and Boolean search string, but the one hour to conduct the case study search resulted in a good reportable results set.

Patents Reviewed: The SuperBoolean search allowed us to focus only on the top 100 results at each results stage, those patents in the results list being based on relevancy to the Heuristic Query.

Patents Eliminated: Upon obtaining a new results list / column with each additional keyword, those patents that were dropped from the current column (when compared to the preceding column) were highlighted in green. All eliminated patents were then individually examined to determine whether they were actually relevant to the search. Dropped, but still relevant patents are shown with a **red** patent number.

LEGEND - CASE STUDY RESULTS



Baseline no key words, total US database	A. (2 words) printer, nozzle		B. (3 words) printer, nozzle, ink		C. (4 words) printer, nozzle, ink, dispersion		D. (5 words) printer, nozzle, ink, dispersion, pattern		E. (6words) printer, nozzle, ink, dispersion, pattern, geometry		F. (7 words) printer, nozzle, ink, dispersion, pattern, geometry, deflection	
3 millon hits	14,533 HITS		12,198 HITS		2,243 HITS		1,090 HITS		121 HITS		48 HITS	
US Pat #	%	US Pat #	%	US Pat #	%	US Pat #	%	US Pat #	%	US Pat #	%	US Pat #
4210916	100	4210916	100	4210916	100	6386679	100	6386679	100	6761437	100	6761437
7066564	99.6	7066564	99.6	7066564	99.8	6244687	99.8	6244687	98.7	6497510	98.7	6497510
4380017	97.7	4380017	97.7	4380017	99.3	6761437	99.3	6761437	98.2	5966154	98.2	5966154
6612685	97.1	6612685	97.1	6612685	98.0	6497510	98.0	6497510	84.1	6394675	81.5	6474796
4196006	96.3	4196006	96.3	4196006	97.7	6350028	97.5	5966154	82.3	6830327	80.2	6273559
6065822	95.5	6065822	95.6	6065822	97.5	5966154	95.7	5790150	82.2	6406121	79.7	6695440
4967208	94.8	4967208	94.8	4967208	97.3	4975117	94.5	5870112	81.5	6474796	76.5	6874864
4429315	94.5	4429315	94.5	4429315	97.2	6923529	94.4	5981623	80.2	6273559	76.1	5796418
6332662	94.3	6332662	94.3	6332662	96.3	6248163	92.7	6471323	79.7	6695440	75.7	5870124

Case Study Search Results

Baseline no key words, total US database	A. (2 words) printer, nozzle		B. (3 words) printer, nozzle, ink		C. (4 words) printer, nozzle, ink, dispersion		D. (5 words) printer, nozzle, ink, dispersion, pattern		E. (6 words) printer, nozzle, ink, dispersion, pattern, geometry		F. (7 words) printer, nozzle, ink, dispersion, pattern, geometry, deflection	
3 million hits	14,533 HITS		12,198 HITS		2,243 HITS		1,090 HITS		121 HITS		48 HITS	
US Pat #	%	US Pat #	%	US Pat #	%	US Pat #	%	US Pat #	%	US Pat #	%	US Pat #
4210916	100	4210916	100	4210916	100	6386679	100	6386679	100	6761437	100	6761437
7066564	99.6	7066564	99.6	7066564	99.8	6244687	99.8	6244687	98.7	6497510	98.7	6497510
4380017	97.7	4380017	97.7	4380017	99.3	6761437	99.3	6761437	98.2	5966154	98.2	5966154
6612685	97.1	6612685	97.1	6612685	98.0	6497510	98.0	6497510	84.1	6394575	81.5	6474795
4196006	96.3	4196006	96.3	4196006	97.7	6350028	97.5	5966154	82.3	6830327	80.2	6273559
6065822	95.5	6065822	95.5	6065822	97.5	5966154	95.7	5790150	82.2	6406121	79.7	6695440
4967208	94.8	4967208	**94.8**	**4967208**	97.3	4975117	94.5	5870112	81.5	6474795	76.5	6874864
4429315	94.5	4429315	94.5	4429315	97.2	6923529	94.4	5981623	80.2	6273559	76.1	5796418
6332662	94.3	6332662	94.3	6332662	96.3	6248163	92.7	6471323	79.7	6695440	75.7	5870124
6705699	94.0	6705699	**94.0**	**6705699**	95.7	5790150	92.3	6779865	79.4	6585369	75.5	6491376
4184881	93.9	4184881	93.9	4184881	95.6	5382963	92.0	6533851	78.3	5114477	75.5	6012799
6561609	93.9	6561609	**93.9**	**6561609**	95.0	5781214	90.6	5808637	**77.7**	**5892524**	75.1	5815178
6896357	93.6	6582055	93.6	6582055	94.5	6352340	90.1	7029095	76.5	6874864	73.8	5781205
6582055	93.4	6863384	**93.4**	**6863384**	94.5	5870112	89.3	6137507	76.1	5796418	73.1	5856836
6863384	93.3	4343013	93.3	4343013	94.4	5981623	**88.7**	**5407136**	75.7	5870124	72.4	5838339
4343013	93.3	6241333	93.3	6241333	93.7	6474781	88.6	6702419	75.5	6491376	71.8	6796641
6241333	92.8	6908171	92.8	6908171	93.5	5837046	87.9	6193361	75.	6012799	71.8	6971739
6908171	92.7	6203140	92.7	6203140	93.3	6827429	86.9	6834927	75.4	5920331	71.7	5871656
6203140	92.6	6908178	92.6	6908178	92.7	6471323	86.1	6966643	75.1	5815178	70.4	6672702
6908178	92.5	6905552	92.5	6905552	92.4	6254670	**85.9**	**6336708**	74.2	5897695	70.3	6217155
6905552	92.5	6565180	92.5	6565180	92.3	6550889	85.2	6074052	73.8	5781205	70.3	5850241
6565180	92.4	4413268	**92.4**	**4413268**	92.3	6779865	85.1	6450098	**73.5**	**5371527**	70.3	6126846
4413268	92.4	6027203	92.4	6027203	92.0	6533851	84.8	6618066	73.1	5801739	70.0	6045710
6027203	92.2	6394569	92.2	6394569	91.9	6428157	84.5	6425331	73.1	5880759	69.2	5781202
6394569	92.2	6050675	92.2	6050675	**91.0**	**5963235**	84.4	6665091	73.1	5856836	68.5	5916358
6050675	92.2	5880758	92.2	5880758	90.6	5808637	84.3	6475271	72.4	5838339	68.3	5905517
5880758	92.1	6786975	92.1	6786975	90.5	6059869	83.8	6422698	71.8	6796641	67.7	5812162
6786975	92.0	5710581	92.0	5710581	90.5	6746108	83.5	5948150	71.8	6971739	67.1	5984446
5710581	92.0	6663222	92.0	6663222	90.4	6550882	83.5	6394575	71.7	5871656	66.8	5825385
6663222	91.8	5751312	91.8	5751312	90.3	6793328	83.3	6964700	70.7	5936008	66.5	6002847
5751312	91.8	6158838	91.8	6158838	90.1	7029095	**82.8**	**6439703**	70.4	6672702	64.0	6849308
6158838	91.8	6565191	91.8	6565191	89.8	6361161	82.7	6733111	**70.3**	**6217155**	62.1	6509085
6565191	91.8	6273542	91.8	6273542	**89.8**	**6554410**	82.5	6056812	70.3	5850241	61.3	6074725
6273542	91.7	5410342	91.7	5410342	89.5	6350014	82.4	6899426	70.3	6126846	59.6	6659598
5410342	91.7	5929876	91.7	5929876	89.4	6945628	82.3	6913353	70	6045710	49.8	6713389
5929876	91.3	7036901	91.3	7036901	**89.4**	**6509917**	82.2	6737109	69.5	6485134	49.3	6503831
7036901	91.2	5640183	91.2	5640183	**89.4**	**6364470**	82.0	4421559	69.2	5781202	46.3	6761758
5640183	91.2	5091005	91.2	5091005	89.3	6137507	81.8	6805736	68.5	5916358	45.8	6811595

5091005	91.2	6394585	**91.2**	**6394585**	89.2	6716278	81.8	6830327	68.3	5905517	45.1	6835833
6921150	91.1	6709084	91.1	6709084	88.9	6030439	81.6	6406121	67.7	5812162	42.6	6860928
6394585	91.0	6161918	91.0	6161918	88.7	5407136	**81.4**	**6491385**	67.1	5984446	42.4	7087752
6709084	90.9	6158835	90.9	6158835	**88.6**	**6450628**	81.3	4668533	67.1	5389131	38.7	6818276
6161918	90.8	5654744	90.8	5654744	88.6	6702419	**80.9**	**6382782**	67.0	6030072	34.0	6720519
6158835	90.8	5559540	90.8	5559540	88.4	6379440	80.9	6474795	66.8	5825385	32.9	6991706
6997533	90.8	4961785	90.8	4961785	88.3	6012805	80.7	6478394	66.5	6002847	21.5	6769969
5654744	90.8	5581284	90.8	5581284	**88.2**	**6079821**	80.5	7029112	64.6	5389133	16.3	5679145
5559540	90.7	6386679	90.7	6386679	88.0	6491362	80.5	6631983	64.0	6849308	16.2	5709827
4961785	90.7	6502912	90.7	6502912	87.9	6193361	80.4	5221332	62.1	6509085	15.8	5683772
5581284	90.7	6705702	90.7	6705702	87.2	6030438	80.4	4352691	61.6	5790156		
5971518	90.6	6854829	**90.6**	**6854829**	86.9	6834927	80.4	6471347	61.3	6074725		
6386679	90.5	6244687	90.5	6244687	86.8	6488370	**80.3**	**6450619**	61.1	6595630		
6502912	90.5	7004571	90.5	7004571	86.7	6120133	**79.9**	**6780339**	61.0	5554213		
6705702	90.4	6595621	90.4	6595621	86.4	6648464	79.8	6902274	59.9	5260009		
6854829	90.4	6830320	90.4	6830320	86.4	6439710	79.6	6273559	59.7	6175422		
6966627	90.4	6431704	**90.4**	**6431704**	86.3	5861900	79.6	6059871	59.6	6659598		
6244687	90.3	6565190	90.3	6565190	86.1	6966643	79.2	6695440	55.2	6827769		
7004571	90.2	5724079	90.2	5724079	85.9	6336708	79.1	6084621	52.4	6169605		
6595621	90.1	6776474	90.1	6776474	85.3	5843219	79.0	5693129	**52.0**	**5622611**		
6830320	90.1	6761437	90.1	6761437	85.2	6074052	78.9	5302197	51.2	5594652		
6431704	90.0	6312117	90.0	6312117	85.1	6450098	78.9	5098475	51.0	6001482		
6565190	89.9	6354689	89.9	6354689	84.8	6618066	78.9	6585369	49.8	6713389		
5724079	89.9	6361156	89.9	6361156	84.7	6569230	78.9	6089697	49.3	6503831		
6776474	89.9	6905191	89.9	6905191	84.7	5580372	78.8	6412928	47.2	6467897		
6761437	89.8	5557307	89.8	5557307	84.6	6375304	78.8	6441774	**47.2**	**6852560**		
6312117	89.8	5650808	89.8	5650808	84.5	6425331	78.7	6412909	**47.1**	**6802456**		
6354689	89.8	5521622	89.8	5521622	84.4	6665091	78.6	5764263	**46.9**	**6750076**		
6361156	89.7	5563639	89.7	5563639	84.3	6475271	78.5	6344819	46.3	6761758		
6905191	89.7	6902252	89.7	6902252	84.3	6273536	78.5	6120141	45.8	6811595		
5557307	89.7	5992962	89.7	5992962	84.3	6783581	78.4	5902390	45.1	6835833		
5650808	89.5	6491374	89.5	6491374	84.0	6752494	78.2	6655773	**44.2**	**6822231**		
5521622	89.4	6299287	89.4	6299287	84.0	6162289	78.1	5989325	**44.2**	**6627882**		
5563639	89.2	6755506	89.2	6755506	83.8	6890069	78.0	6676244	44.2	6197482		
6902252	89.2	6557971	89.2	6557971	83.8	5026427	**78.0**	**6507002**	**44.1**	**6723985**		
5992962	89.1	5600353	89.1	5600353	83.8	6422698	77.9	6409330	42.9	6300045		
6491374	89.0	5412411	89.0	5412411	83.7	5601023	77.8	5746815	**42.8**	**6787766**		
6299287	89.0	6505911	89.0	6505911	83.6	6575566	77.8	5114477	42.6	7062848		
6755506	89.0	6497510	89.0	6497510	83.5	5514207	77.8	6069190	42.6	6860928		
6557971	89.0	6557988	89.0	6557988	83.5	5948150	77.8	6247804	42.4	7087752		
5600353	88.9	6619794	88.9	6619794	83.5	6394575	77.7	5644350	**42.0**	**6633031**		
5412411	88.8	5659342	88.8	5659342	83.5	6030440	77.6	5605566	41.8	6821462		
6505911	88.8	5598192	88.8	5598192	83.3	6964700	77.5	6502925	**41.3**	**6768107**		
6497510	88.7	6378980	88.7	6378980	82.9	6637876	77.4	6783222	41.2	6447093		
6557988	88.7	6902331	88.7	6902331	82.9	6637868	77.3	5936027	41.2	6808659		

6619794	88.7	6812953	88.7	6812953	82.8	6439703	77.3	6164756	41.0	6019455		
6604818	88.7	6629752	88.7	6629752	82.8	6899753	**77.3**	**6474794**	40.6	7063924		
5659342	88.6	6350028	88.6	6350028	82.7	6578955	77.3	5746817	40.0	7022385		
5598192	88.6	5903290	88.6	5903290	82.7	6733111	77.2	5892524	39.8	7066976		
6378980	88.6	5943072	88.6	5943072	82.6	6476096	77.2	5568173	39.5	6911412		
6902331	88.6	6976748	88.6	6976748	82.6	6382776	77.2	6742868	39.0	6660680		
6812953	88.5	5966154	88.5	5966154	82.5	6056812	77.2	5320668	38.9	6991754		
6871934	88.5	6145961	88.5	6145961	82.4	6231654	77.1	5105209	38.7	6818276		
6629752	88.5	6375307	88.5	6375307	82.4	6899426	77.0	6033055	38.7	6686205		
6350028	88.4	5680162	88.4	5680162	82.3	6048389	76.9	6641651	38.6	6649413		
5903290	88.4	6398337	88.4	6398337	82.3	6913353	76.8	5808639	38.3	7087341		
5943072	88.4	5108503	88.4	5108503	82.3	6659583	76.7	6086185	38.1	6864201		
4630076	88.3	4975117	88.3	4975117	82.2	6737109	76.7	5863320	38.1	7034091		
6976748	88.3	4380772	88.3	4380772	82.0	4421559	76.7	6187082	37.7	6753108		
5966154	88.3	6257698	88.3	6257698	82.0	6503311	76.6	4849770	37.3	6346290		
6145961	88.3	6739684	88.3	6739684	82.0	6395079	76.6	5805178	36.4	6967183		
6375307	88.3	6923529	88.3	6923529	81.8	6805736	76.6	6336694	35.1	5776359		

Legend and Keyword Results Summary correlating to additional keywords appended to a Boolean query string:

top 100 Heuristic only / no Boolean	92 of top 100 remained after adding keyword #1 & #2	100 of previous top 100 remained after adding keyword #3	8 of previous top 100 remained after adding keyword #4	37 of previous top 100 remained after adding keyword #5	12 of previous top 100 remained after adding keyword #6	42 of previous top 100 remained after adding keyword #7
dropped between columns	<u>8% eliminated</u> with 2 keywords	0% eliminated with 3 keywords	<u>92% eliminated</u> with 4 keywords	<u>73% eliminated</u> with 5 keywords	<u>88% eliminated</u> with 6 keywords	<u>58% eliminated</u> with 7 keywords
dropped but relevant	**4** relevant patents dropped with first 2 key-words	**4** relevant patents missed / dropped between A&B	**11** relevant patents missed / dropped be-tween B&C	**9** relevant patents missed / dropped be-tween C&D	**13** relevant patents missed / dropped be-tween D&E	total of **37** relevant pat-ents missed / dropped

Conclusion

We've acknowledged the various problems with the Boolean search process in general, and more specifically related these problems to the inordinately high legal and financial risks associated with patent documents.

The demands to perform a patent search that attempts to identify <u>all</u> of the relevant documents within the scope of available resources (time, budget, computing time, a given patent data quality) keep researchers reliant on the time-honored practice of crafting a lengthy, complex Boolean search string. But it's been shown that such restrictions, although they produce relevant patents in a final results list, more dangerously drop an increasing number of relevant patents <u>that should have been included</u> in the final search report.

The application of heuristics, namely a synonym lookup table, or more preferably a Semantic / artificial intelligence **expert system** that allows the researcher to use a less restrictive Boolean query, results in the Best-First search results list that positions the most relevant documents at the top. Researchers are then able to manage very large search results lists without filtering the list to a more manageable quantity by using more keywords.

The results of applying heuristics to Boolean patent searching are faster time to identify the most relevant patents, but more importantly, the identification of the largest number of relevant patents that will serve as acceptable prior art.