

## **LES 2004 Winter Meeting:**

*Licensing Converging Technologies - Bridging the Gap*

### **PRESENTATION PAPER**

---

## **Latent Semantic Analysis: New Intellectual Property Data Search Technology For Non-Obvious Licensing Opportunities**

---

by

**Andy Gibbs, CEO**

PatentCafe.com, Inc.  
441 Colusa Avenue • Yuba City, CA 95991  
t: 530.671.0200 • f: 530.671.0201  
<http://www.patentcafe.com>

---

#### **BACKGROUND**

Generally, the methodology of licensing patents incorporates to varying degrees a sequence of processes that includes identification of licensing targets, developing a position that allows patent assertion, and ultimately the execution of a licensing agreement. Focusing on patents as a means to identify licensing targets, the process may look like the flow in CHART 1.

One of the many methods of identifying potential licensees uses patent searching or patent data mining. Patent searches can result in a pool of patents that are clustered around certain technology groups or patent classifications, thereby carrying along the corresponding list of companies (assignees) conducting business in those targeted technology areas.

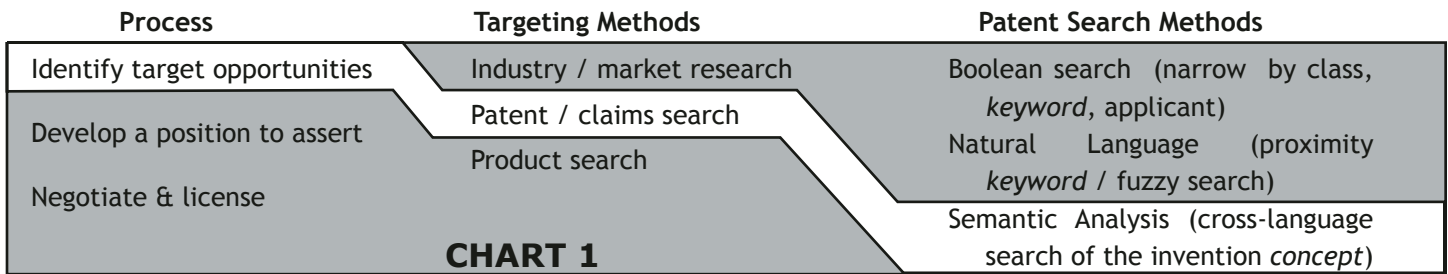
The method of searching patents varies depending on the search technology overlying the patent database. These search technologies include highly structured command line (ie: Dialog), keyword searching (word, word combinations, Boolean match/no match), natural language (more forgiving than Boolean - linguistic variants of the root search words, word stem, phrases), and semantic analysis (neural net indexing that learns the concepts represented by words). There are significant differences in the results that may be obtained through the use of these search technologies, especially when the objective is to identify licensee targets rather than to simply conduct pure patent research. While the use and limitations of command line, Boolean and natural language search technologies are fairly well known, it's the objective of

---

this paper is to discuss the newer concept-based semantic search technology that is enjoying rapid adoption to search large volumes of unstructured data.

words with the query.

LSA works by mapping relationships between each word



*[Semantic search technology is increasingly the technology of choice for National Security, and was implemented as the litigation discovery tool used to search and relevancy rank more than 30 million pages of Enron E-mail against concept queries.]*

### “OFF TARGET” PRECISION

Keyword searching has the advantage of being precise. That is, the researcher will definitely see results that match their search queries. There are many assumptions the researcher makes when relying on keyword searching, however. These assumptions include: (a) the researcher knows every possible keyword or jargon used by every industry that may be a rich environment of targeted licensees, and (B.) that the patent data being searched is complete and accurate (contains the words that it is supposed to contain - definitely not always the case).

The danger associated with relying on keyword matching to identify licensing targets is that while the researcher can see the results they expect, they cannot possibly know what possibly important patents they search query is missing.

*The first mover in a given market traditionally enjoys a limited time of exclusive exploitation and profitability.*

Keyword searching is especially problematic when considering the phenomenon of parallel invention - the simultaneous discovery of the same or similar inventions occurring in dissimilar markets or industry segments without the benefit of the transmission of ideas between these segments. The use of entirely different jargon used by the respective industries, and the human assignment of completely different patent classification numbers can create a large, critically important body of patent data that remains obscured by the precision of keyword searching.

### WHAT IS “SEMANTIC”?

Latent Semantic Analysis, (LSA) is emerging as the advanced search technology that overcomes the objectionable limitations of keyword searching. LSA is a concept-based search technology that results in an associative, almost “human-like” understanding of documents. Most simply, LSA technology such as Semetric™ can retrieve relevant documents even when those documents do not share any

and every other word in large sets of documents. It’s a technology ideally suited for patent collections of 20 million documents ... or more. This association of words based on the context in which they are used creates a form of numeric “meaning” and “understanding” for the software. Not only are **conceptual** relationships measured between words, but also the theme of an entire document is considered to return the **most relevant** documents available.

The indexing of documents in a traditional database results in the creation of geometric vector representation of each document within a searchable Concept Space.

The results from a query are conceptual. Because LSA operates on concepts, not keywords, searches are not constrained by the individual words that users choose in the query.

Once these attributes are all brought together - the picture becomes clear:

A concept-based semantic analysis search on a patent database will (a) find every document that matches the **concept** of the search query, (b) return the results in a relevancy-ranked order, with the most relevant at the top of the search results list.

### JUST OVER 10 YEARS OLD

Although keyword searching has been around about as long as computers (and Boolean since 1854 when **George Boole invented** binary algebra - now called Boolean logic), Latent Semantic Indexing (LSI) was created in the early 1990s. Over the past decade, advancements to LSI have been swift and effective. Latent Semantic Analysis is a highly refined filtering system reaching back to LSI.

The newness of the technology however, is not the reason that LSA has not been widely implemented. Highly functional advancements of LSI that afford practical application of the technology (rather than academic or theoretical application) are even younger at about 5 years. Since large scale database development can take years to create, those patent databases operating 5 years

ago are still trying to realize the financial return on their investment. Just as General Motors is compelled to build automotive parts on expensive investment tooling before re-designing and re-tooling for next generation designs, older patent databases must ultimately scrap their current technology, and essentially must replace it entirely if semantic tools are to be employed.

Consequently, semantic search tools will initially be found in new databases that either replaced older patent databases, or were developed as entirely new products without the necessity of waiting until the "old tooling" was fully exploited and depreciated.

Rapid advancements in semantic search technology continue to widen the lead - the newest semantic search technology already combines concept searching plus Boolean - delivering the best attributes of old and new. As with any technology improvement, there are distinct advantages to early adopters.

### SEMANTIC ANALYSIS TO IDENTIFY TARGETS

Successful (profitable) licensing depends on the ability to identify the most lucrative technology or market growth segments, and subsequently the potential licensees conducting business within those segments.

Parallel invention theory already shows us that the myopic searching of keywords can miss patent activity in unrelated industry segments, and correspondingly, miss identifying a rich environment of potential licensees.

Since LSA searches the concept of the invention, it readily identifies un-obvious technology, market or industry segments that indeed may be lucrative targets.

The value of these new-found opportunities is further enhanced by the fact that, when compared to competitors still using keyword search technology, the early adopters find themselves alone in exploiting the new market areas.

### CLAIMS DESCRIBE THE INVENTION

Keywords can be applied to the claims section of patents. The hope of course, is that such a search will identify patents and applicants that are linked to infringing products.

Not until you can search the meaning of the claims will the researcher truly identify the broad range of "un-tapped" markets and applicants.

Latent Semantic Analysis is the first technology capable of searching a large scale patent data based on the text of the CLAIMS of a patent. LSA allows the researcher to use the entire text of a sample patent claim as the search query. LSA does much better in returning relevant results if it is not constrained to searching just a few keywords (since keywords cannot adequately describe a concept).

	A	B	C	D	E	F
1	Document	Filing Date	Assignee Info	Domestic Classification		
2	US 20020040925	5/11/01		Main: 235 Sub: 379		
3	US 20010042785	5/8/01		Main: 235 Sub: 379		
4	US 20030222136	5/31/02		Main: 235 Sub: 379		
5	US 20010042784	3/2/99		Main: 235 Sub: 379		
6	US 20020104878	1/4/02		Main: 235 Sub: 379		
7	US 20030024979	7/26/02		Main: 235 Sub: 379		
8	US 20030111526	10/30/02		Main: 235 Sub: 379		
9	US 06554184	5/5/00		Main: 235 Sub: 379 Main: 235 Sub: 380		
10	US 06467684	3/2/99		Main: 235 Sub: 379 Main: 235 Sub: 380		
11	US 06267292	7/22/99		Main: 235 Sub: 379 Main: 235 Sub: 380		
12	US 06021943	8/27/97		Main: 235 Sub: 379 Main: 235 Sub: 380		
13	US 20030080185	10/26/01		Main: 235 Sub: 380		
14	US 20020185529	6/8/01		Main: 235 Sub: 380		
15	US 20030168510	11/1/01		Main: 235 Sub: 380		
16	US 20040011864	9/11/02		Main: 235 Sub: 380		
17	US 05770844	10/25/96		Main: 235 Sub: 380 Main: 235 Sub: 375		
18	US RE36365	9/4/96		Main: 235 Sub: 380 Main: 235 Sub: 375		
19	US 05477038	10/25/93		Main: 235 Sub: 380 Main: 235 Sub: 379		
20	US RE39255	4/2/01		Main: 235 Sub: 380 Main: 235 Sub: 379		
21	US RE37122	4/17/98		Main: 235 Sub: 380 Main: 235 Sub: 379		
22	US 06189787	10/27/99		Main: 235 Sub: 380 Main: 235 Sub: 382		
23	US 06592030	7/20/00		Main: 235 Sub: 380 Main: 705 Sub: 35		
24	US 06105865	7/17/98		Main: 235 Sub: 380 Main: 705 Sub: 35		
25	US 20030029914	10/8/02		Main: 235 Sub: 383		
26	US 05471669	3/17/94		Main: 235 Sub: 383 Main: 235 Sub: 487		
27	US 20020008146	11/20/98		Main: 235 Sub: 472.01		
28	US 04725719	7/21/86		Main: 235 Sub: 487 Main: 235 Sub: 381		
29	<b>US 05287268</b>	<b>11/16/92</b>		<b>Main: 364 Sub: 405 Main: 364 Sub: 406 N</b>		
30	<b>US 05479510</b>	<b>11/15/94</b>		<b>Main: 380 Sub: 024 Main: 380 Sub: 022 N</b>		
31	<b>US 05671364</b>	<b>6/5/95</b>		<b>Main: 395 Sub: 239 Main: 395 Sub: 201 N</b>		
32	US 06415271	3/26/99		Main: 705 Sub: 039 Main: 705 Sub: 068		
33	US 05937396	12/4/96		Main: 705 Sub: 043 Main: 705 Sub: 039		
34	US 20020065666	11/30/00		Main: 705 Sub: 1		
35	US 20030093293	9/30/02		Main: 705 Sub: 1		
36	US 04941090	1/27/89		Main: 705 Sub: 14 Main: 705 Sub: 16 M		
37	US 06105001	8/15/97		Main: 705 Sub: 14 Main: 705 Sub: 17 M		
38	US RE36116	2/15/96		Main: 705 Sub: 16 Main: 705 Sub: 14 M		
39	US 20020152124	4/10/01		Main: 705 Sub: 16 Main: 705 Sub: 39		
40	US 20020174016	4/2/02		Main: 705 Sub: 16 Main: 705 Sub: 41		
41	US 20020178063	5/25/01		Main: 705 Sub: 17		
42	US 06014636	5/6/97		Main: 705 Sub: 17 Main: 705 Sub: 16 M		
43	US 20010023409	4/18/01		Main: 705 Sub: 17 Main: 705 Sub: 39 M		
44	US 20030149629	2/6/02		Main: 705 Sub: 17 Main: 705 Sub: 43 M		
45	US 20040006489	7/3/02		Main: 705 Sub: 2		
46	US 20020104880	1/18/01		Main: 705 Sub: 2		

Concept-based patent search for "method of transferring cash between remote persons" returned some expected and unexpected patent classifications, notably one in the robotics industry segment. the researcher would then need to create a claims chart to complete the comparison of the inventions.

364: Electrical Computers And Data Processing Systems  
 235: Cash Register  
 380: Secret Communication Electronic, Funds Transfer  
 705: Accounting Computer Controlled, Monitored  
**395: Robot, Arm Movement Control**

**CHART 2.**

In real world application, searching the concept of the patent claims has resulted in licensors identifying entirely new market segments. As an example:

- A multinational oil company's IP licensing manager tasked with licensing the company's technology exhausted the traditional targeting methods: search patent classifications, search industry keywords [such as "mining", "water treatment", "petroleum"]. Within minutes of searching on the entire text of the claims of one of the patent earmarked for licensing, the manager found an entirely new body of product technology that potentially infringed the oil pipe patents - miniature medical needles. Both inventions claimed similar methods of reducing interior-wall fluid friction reduction methods - one for oil, the other for blood. The oil pipe patents enjoyed a priority date advantage.

---

*This is a great example of how a parallel field of invention was identified and exploited using the stick licensing approach.*

- A manufacturer of self-heating, plug-in air fresheners identified a the un-obvious food industry with it's active self-heating cans as a lucrative licensing environment. There were no uniformly overlapping industry keywords or patent classifications that would have allowed identification of the food market using keyword search technology.

Of course, after identifying new market areas, it's the licensing professional's job to then perform a commercial analysis of market size and sales volumes, for each target licensee size and sales volumes, for each target licensee market size, and assess the retaliatory legal response, or financial investment likely in order to see the licensing through to conclusion.

#### METHODOLOGY

Using a semantic search engine to mind patent data, a search query such as the following may be used: "method of transferring cash between remote persons". In actuality, a longer, more comprehensive query would be desired.

The search results included US issued patents and published applications, and were ranked according the how relevant they were to the search query (the most relevant patents were on top of the list).

This data set thus described the universe of patents that claim or describe an invention similar to the concept of the query.

Three data fields (patent filing date, patent number, US classification) from top 100 were exported, although 10,000 could have been. The exported data set was then opened in Excel and sorted according to patent classification.

TO identify the most active patent assignees, exporting and sorting the assignees would have produced data that could have been charted - identifying the licensees based on their level if investment in the technology areas.

However, it's interesting to note that within the first 100 patents, one relating to robotics appeared. Should this identify a previously ignored potential licensing industry area in which to license, then drilling further into the classification would have followed.

It's important to understand that the Class 395 was

identified in minutes, rather than through serendipity some later date. A concerted effort to enter the robotics industry with the subject technology could yield significant licensing revenue that would have never been considered otherwise.

#### SUMMARY

While each technology or methodology applied to patents will usually yield predictable results, LSA stands unique in being able to identify new industry or technology segments containing potential licensing targets previously not considered.

Whether traditional keyword technology remains the core patent searching method employed by the researcher, today's competitive licensing environment and shareholders' expectation of earnings growth nearly require the augmentation of traditional search methods with more advanced semantic search technology.