The Evolution of Semantic Search on the Web:
A LexisNexis® White Paper
In January 2009, a J.P. Morgan survey sought to identify which factors would cause Web users to switch from their preferred search engine to a new one. The results provided an interesting glimpse into what is most important to those of us who rely on the Internet for our daily research needs.

According to the survey, the top consideration—expressed by 45 percent of respondents—was for results that better match the search terms entered into the engine. This factor was more important than speed of response, ease of navigation and all other considerations.

This fundamental goal—the ability to discover more precise search results—is at the heart of a new generation of search technology that is the subject of this white paper.

The New World of Semantic Search

“Semantic Search” is a process used to improve online searching for individuals conducting research on the Internet—that is, when there is no particular document you’re seeking but when you’re trying to locate a number of documents that will help you locate the information you want. Rather than use ranking algorithms (as many popular internet search engines do), to predict relevancy, Semantic Search uses the science of meaning in language (“semantics”) to produce highly relevant search results. The idea is to deliver the information queried by a researcher, rather than forcing them to review a laundry list of loosely related keyword results.

Aside from convenience for those of us who have come to rely on online search engines as research tools, why is this important? First, consider that approximately 80 percent of the information stored and maintained by corporations is in unstructured form. Second, according to IDC, a global marketing intelligence firm, this means that knowledge workers spend an estimated 48 percent of their time searching and analyzing information and that the search related activity costs organizations more than $28,000 per worker each year.

These costs—in combination with the increasing ineffectiveness of keyword search as an exhaustive tool and the massive influx of information in disparate locations—have caused the need for semantic search to explode onto the search landscape in recent months. It now appears that 2009 will be remembered as the year that knowledge workers began to seriously value semantics and the science of meaning in combination with straight keyword search.
Common Misconceptions

The unfortunate side of this new market momentum is that many search engines now claim to be “semantic” when in fact they are not. In addition, search companies that drove, with limited success, the knowledge management initiative of the last 15 years are re-writing their marketing literature in order to position their technologies as being semantic. It’s almost like the shift in the market has all of a sudden made them relevant again.

Unfortunately, the products have infrequently lived up to the marketing in that few, if any, are truly semantic in nature. With this newly cluttered landscape, it is not only important to understand exactly what semantic search is, but it is also important to understand what semantic search is not.

For years, the state of the art in search involved linking keyword queries to taxonomies, lexicons and thesauri as a means for providing a form of structured semantic assistance to searching. This technique for recognizing relationships in data has been around for years with every new list creator boasting of a new breakthrough in learning. The reality is that all of these structured systems utilize manually created lexicons in an attempt to understand meaning and improve search results. In short, they are not learning systems, but static systems.

Unfortunately, it is simply not possible to manually update the static lists at the same rate that information changes. The result is that the lists grow old and searches become less and less effective. So, it’s important that these common misconceptions are no longer allowed to proliferate under the guise of semantic searching. True semantic search is NOT:

- Only powered by pre-built lexicons;
- Only powered by pre-built synonym lists; nor
- Only powered by pre-built thesauri.

Most importantly, while all of these things can be useful features in a semantic search solution, the learning in true semantic search is not static—it is very dynamic. A true semantic system must continually learn.

Learning Systems Make the Difference

Real semantic search involves some sort of machine learning, a scientific discipline concerned with the design and development of algorithms that allow computers to learn based on data. A major focus of “machine-learning research” is automatically learning to recognize complex patterns and make intelligent decisions based on data. Machine learning methods include: Bayesian Inference, Latent Semantic Analysis, neural networks or support vector machines among others.

Products using these methods have the ability to learn from information either by consuming example documents (unsupervised), by user training (supervised), or both. Other Web engines use forms of Natural Language Processing (NLP) to derive patterns in information as a means for creating more dynamic lists or thesauri. Regardless of the method, machine learning is a key component of a true semantic search engine.
This is important because search engines are generally judged on recall and precision. That is, their ability to produce recall (i.e., the largest number of related documents) and precision (i.e., the most relevant documents). As the amount of data being searched has grown, keyword search has increasingly fallen short in producing both recall and precision.

Anyone who has searched the Web and received millions of results for their query knows this to be true. Recall on the Web is overwhelming; results can easily be in the thousands or more. Using a keyword, or better yet multiple keywords, in an attempt to create precision in a given query results in eliminating too many relevant results. In short, it has become a lose-lose scenario where you have a choice between too many or too few results. This simply may not be good enough anymore.

The semantic advantage lays in a search engine’s ability to match on the meaning of words in a user’s query or in a source document. In other words, semantic search engines are able to go beyond keyword matching and match on concepts.

In fact, the more powerful semantic search engines will sometimes produce relevant search results that do not contain any of the query words, bringing an element of serendipity to search. In addition, many of the linguistic challenges that typically wreak havoc on keyword engines like polysemy (words with multiple meanings) and synonomy (multiple words with the same meaning) are handled intelligently and naturally by semantic engines.

**Semantic Search: Seeing Through the Hype**

While semantic search engines offer many new benefits to searchers, most products on the market today exhibit frustrating limitations that ultimately take away from the overall value. The most common limitations associated with semantic search are:

- Lack of transparency: Most semantic search engines on the market today are “Black Boxes”—that is they are not able to show precisely how they generate a search result;

- Lack of control: In addition to being black boxes, semantic search engines provide the searcher little ability to enhance or affect the search result by “engaging” with the query;

- Inability to scale indexes: Semantic search engines can only search the documents they index and don’t allow the user to extend the intelligence to other content sources. In short, the semantic provider must store and index all data to be searched. It’s just not reasonable to expect any provider to re-index all current and future indexes of all content as the availability of content continues to expand exponentially; and

- Inability to scale intelligence: Semantic indexes are generally larger than keyword indexes, typically live in memory and cannot scale to learn from tens of millions of documents.

The existence of these limitations has created an opportunity for a semantic search solution that finally answers the needs of power searchers without attempting to replace the searcher themselves (or their expertise in Boolean searching) as a critical part of the search process.
Semantic Search on TotalPatent™ from LexisNexis®: A conversation between searcher and technology

Before embarking on yet another semantic search solution, LexisNexis chose to set a higher bar. The solution had to rise above the challenges and limitations that have hindered the other semantic search solutions on the market today—it had to be different. It had to learn from ALL patent documents—every single one of them. It had to scale to provide semantic search capability beyond just LexisNexis indexes and into client databases and even the open Web. Most of all, the LexisNexis semantic search solution had to answer our customers’ requests for complete transparency with control of each query remaining in the hands of the users; it could not be a black box.

To that end, LexisNexis teamed up with PureDiscovery to finally bring to market the first fully transparent, highly scaleable, federated semantic search platform for patents and related technical literature.

The new TotalPatent semantic search solution accomplishes four breakthrough objectives in online search:

1. It is transparent: No black box here. Each query is enhanced by the machine intelligence and shown to the user for their complete understanding and engagement.

2. It gives control to the user: Not only is the semantic search transparent, but users are given the ability to add, delete, increase or decrease the importance of all query words (concepts) in a unique visual query interface.

3. It is fully federated: While we maintain the largest full text patent and non-patent literature databases in the world, we didn’t stop there. Our semantic search platform can federate semantic searches to virtually any index, whether it resides internally or on the web.

4. It utilizes the most scalable semantic technology in the world: We started building our semantic intelligence with the USPTO patent index of over 7 million documents, we have since added 2.5 million Elsevier journal articles, and we are just getting started. Our technology knows no current limit to how much it can learn. Unlike other semantic technologies that must sample smaller sets of documents for learning, our semantic index includes the learning from over 10 million full text documents, and it’s growing every day.

At the center of the TotalPatent service and what makes it all possible is the creation of the BrainSpace, an artificial intelligence layer of a scale that has never been achieved before. We started with over 10 million full text patent and non-patent literature documents and clustered them into dozens of clusters. Each of these clusters was then transformed into a “brain”. It is this collection of “brains” that together form the BrainSpace. There have been many attempts in the past to incorporate artificial intelligence into patent search, but never before has there been a product that constructed an intelligence layer on this scale.
Next came an equally important challenge—how to integrate the user into the experience. What makes the TotalPatent semantic search experience so unique is that LexisNexis has re-invented semantic search to occur NOT as a transaction, but rather as a conversation between the machine learning and the user of the system. TotalPatent semantic search engages the user in a search experience that encourages interaction with the machine intelligence. To make this possible, LexisNexis had to create the interaction as a visual experience. We accomplished this by transforming each query into what is called a QueryCloud™, a visual representation of the newly generated semantic query. This “visual sandbox” that allow users to interact with the BrainSpace.

As you can see, each user query is transformed into a tag cloud that shows the top 20 most relevant extracted and inferred words and phrases (concepts) in varying font sizes, visually depicting the relevance of each word. By increasing or decreasing the font size of terms and phrases (relevance), underlining words and phrases (required) and even adding and subtracting words and phrases, the user is completely engaging the artificial intelligence to focus in on highly relevant results.

With LexisNexis semantic search technology, we’ve leveraged the intelligence from millions of full-text technical publications, organized that intelligence in a breakthrough fashion, and then made it accessible to the experience and expertise of the searcher. The result is a unique and valuable tool for patent and non-patent literature.

LexisNexis has opened up a new paradigm in the possibilities and power of searching by presenting the user with real learning to help them perform their jobs better and faster. Welcome to the new world of semantic search.

**About PureDiscovery**

PureDiscovery (www.purediscovery.com) is the creator of KnowledgeGraph, a semantic software platform that transforms an organization’s documents into a working collective intelligence. PureDiscovery’s technology semantically connects people and knowledge in ways that were simply not possible before.