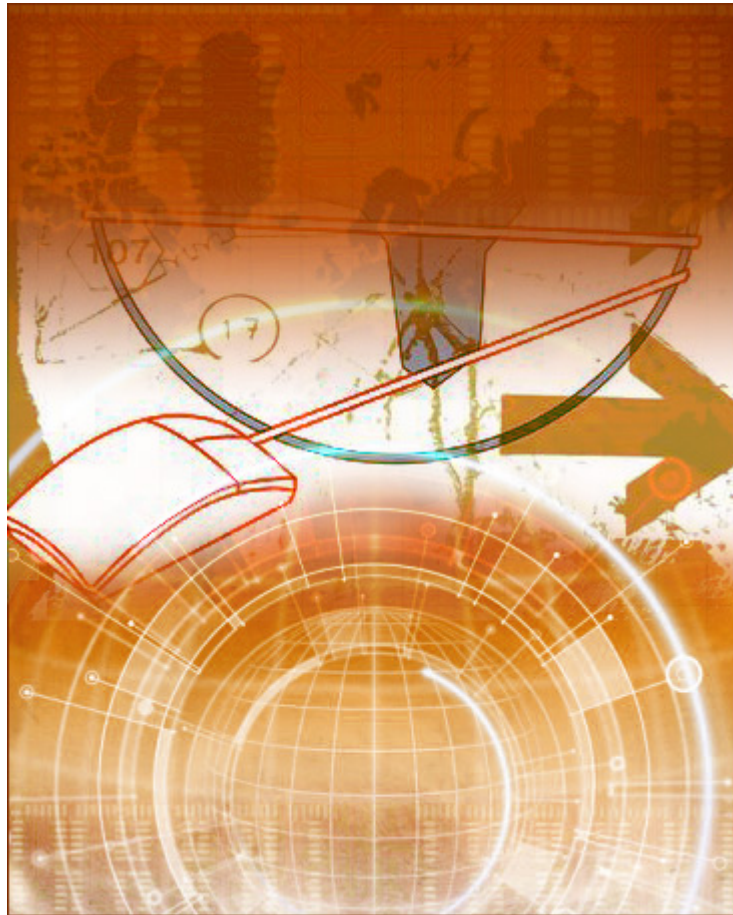


# LSI

## Latent Semantic Indexing - An Introduction



**A Hand Book**  
by Mosaic Services SEO Team

# Table of Contents

## History **3**

Brief Introduction.....	3
Reason for Buying .....	3
Long Term Benefit from this move .....	4

## AdSense and Applied Semantics **5**

History of AdSense ? .....	5
----------------------------	---

## What Is LSI (Latent Semantic Indexing) **7**

LSI Insight.....	7
LSI in Action.....	8

## LSI and search engines **9**

Benefit of Latent Semantic Indexing .....	9
Improved SERPs. ....	9
Stop SPAMS.....	10

## How has Google adopted LSI ? **12**

Proof of LSI.....	12
Inside the Grid.....	12

## Proof of Semantic Rankings **16**

See for yourself.....	16
Cola Example:.....	16
Maid Example: .....	17

# HISTORY

(Google buys Applied Semantics)

## Brief Introduction

Google's tilt towards the semantic indexing of Web pages and establishing relevance through this brilliant statistical method may have been the influence of Applied Semantics, the company known as the innovator of the semantic text processing technology. Google, in an aggressive move which spelled growth and astonished onlookers, took over Applied Semantics in April 2003. The acquisition had several interesting pros and cons and shook the SEO fraternity with expectations and skepticism, at the same time. Let's have some insight on the main concerns.

## Reason for Buying

**"Google and Applied Semantics (previously known as Oingo) started as search engines..."**

The acquisition seemed pretty interesting mainly because of the business focus and the growth path followed by the two companies. The Applied Semantics representative even mentioned the similarity in the work culture between the two companies. To compare, both Google and Applied Semantics (previously known as Oingo) started as search engines with funky names in the late 1990s. While Google maintained the trend, so to say, Applied Semantics diversified business into developing software applications for online advertising, domain name, and enterprise information management markets.

Applied Semantics had already proved their leadership in managing and processing content, especially in the field of semantic text processing, for which they are considered the innovators. The foundation for this expertise lies in the CIRCA technology patented by them. This technology is designed to recognize, organize, and dig out information from Web sites and information stores on the Internet or any computer in a way that is almost similar to the way the human mind and brain scans for related information. Naturally, this was expected to give results that are closer to relevance and importance. In short, Applied Semantics first showed the way to apply Artificial Intelligence more closely into the search engines periphery.



## Long Term Benefit from this move

Google expected to gain mainly from this enterprise that promised immense potentiality in the world of advertising and indexing and ratings for the search engine. Undoubtedly, Google started to make use of the accomplished Applied Semantics engineering team and went on to launch AdSense, which is a Cost per Click (CPC) service that seemed to have brought huge revenue for Google. Note that AdSense was another key application of the CIRCA technology. It helped Web publishers in understanding a particular theme on a page and deliver advertisements that are highly relevant and targeted for that Web page.

Another interesting edge of this takeover was Applied Semantics' relation with Google and its major rival, Overture. This relationship was actually getting redefined. Overture was into a contract with Applied Semantics in which the latter fed the traffic to Overture's paid listings. This was again one of the best services offered by Applied Semantics and formed an USP for Overture. The contract was supposed to continue till August 2003; however, Google's acquisition of Applied Semantics jeopardized the whole idea. Until the takeover, Applied Semantics' technology behind the Overture listings were considered better and could have beaten Google technique anytime.

---

# ADSENSE AND APPLIED SEMANTICS

Before we delve deeper into the SEO scenario as it stands today, we shall talk about two things - AdSense and Applied Semantics and Google's acquisition of Applied Semantics. This would help us understand the backdrop of the new developments in SEO for Google. If you are wondering why we are so much concerned about Google – we can't help it because it's the most reliable, popular, and yielding search engine in the world today and tests the capability of an SEO expert qualitatively.

## History of AdSense ?

**"... AdSense is a content categorization tool developed by Applied Semantics, a California-based company"**

AdSense is a content categorization tool developed by Applied Semantics, a California-based company that established itself with its innovative semantic text processing technology. Applied Semantics was started in 1998 by Adam Weissman and Gil Elbaz, two Caltech graduates. It was known as Oingo back then. The target plan of the initiators was to make computers more and more close to human-thought, or in other words, make them as intelligent as the human brain.

The duo strove with the support of a team of accomplished linguists and software engineers and succeeded in building a new architecture, CIRCA, that was close to their vision. They patented the CIRCA technology that is able to serve as the common platform for all Applied Semantics' products. Gradually, they designed another product, AdSense, which was a semantic text processing tool and was expected to take contextual ad designing to a new high.

AdSense was claimed to be a better product with superior technology than Google's search engine related programs. Applied Semantics representatives claimed that AdSense technology is able to:

- ➔ process and artificially comprehend the actual content of a Web page and does not simply read from the URL or Web log statistics as done by Google
- ➔ identify and extract the main theme of a page objectively. In comparison, Google

studies just the user trends and patterns that are usually subjective, and hence, inconsistent.

- detect confusing or unclear terms, which helps it to rate similar words or synonyms, plurals, and other forms of a word, unlike Google.
- filter information to a higher level and provide crisp and relevant content.

These were not simply words to boast off. The results were being reflected in the paid listings being trafficked by Applied Semantics. No doubt Google was concerned, more so, when archrival Overture entered into a partnership with Applied Semantics for their own listings.

Apart from these, Applied Semantics had a unique technology and commercial base in DomainPark that involved providing revenue-generating content to several Web sites by blending the paid listings from providers and developing manual directories or categories.

Google was certainly planning to beat its competitor- by hook or by crook. Acquisition was, perhaps, the best way out.

---

# WHAT IS LSI (LATENT SEMANTIC INDEXING)

Latent Semantic Indexing or LSI has changed the world of search engine optimization. One fine morning, SEO experts found that most of their best ranking sites on Google were in jeopardy. Google has simply updated its crawler-program to accommodate LSI and has moved towards a more relevant rating list!

LSI is a methodology involving statistical probability and correlation that helps deducing the semantic distance between words. It's obviously a complex methodology but can be easily applied to understand the relation between certain words in a paragraph or in a document. This methodology is being used while indexing a page in the search engine's database.

## LSI Insight

**“LSI is a methodology involving statistical probability and correlation that helps deducing the semantic distance between words...”**

Delving deeper, LSI is concerned not only with studying a document for keywords and listing it in the database, but also with studying a collection of documents and recognizing and identifying the words that are common between these documents. This way it can conclude on the semantic relation between the words being used in these documents. The process then finds out which other documents include or makes use of these semantically close words. The resultant documents are indexed to be related or closely relevant to a context, according to latent semantic indexing.

LSI regards the documents with certain proportion of words being used frequently to be semantically close. If there are fewer words common among documents, they are supposed to be semantically distant. Therefore, LSI introduces interdependence of measure and it rates the relevance of any document on a scale of 0 to 1. Unlike regular keyword searches, LSI can acknowledge the measure of how close is a document to another or how relevant is a credential to a particular context.

## LSI in Action

Let's consider an example here. In a document that discusses Stephen Covey and his preaching, words such as 'effective', 'habits', 'interdependence', 'independence', 'synergic', 'paradigm', 'continuum', 'public victory', 'private victory', 'circle of influence' and so on would be found frequently. Once the search engine indexing tool that uses the LSI technique recognises these commonly-used words from a given set of documents, it can find other documents or Web pages on the net that contains the same set of keywords in almost similar frequency and index them in the database beside the relevant context (Stephen Covey and his preaching) that it leads to.

Now compare this simple method with a human brain's approach to search a context. If you are given a set of document and asked to locate the one's that discuss a particular context, what will you do? Anyone will try to find out the things in common in the sample context and use the observation to compare the rest of the documents to classify them. This intelligence has been added to the lifeless crawler-software or computer technology through the LSI technology.

Quite obviously, the LSI algorithm doesn't understand anything about the meaning of a word in a document. It just reads through the pattern of the usage of particular words and calculates the correlation of their occurrence and hence their correlation with a particular context. Let's get into the practical side of it, that is, how it is applied to a search engine technique.

---

# LSI AND SEARCH ENGINES

## Benefit of Latent Semantic Indexing

**“LSI fits in wonderfully and enhances the search engine’s power to converge with artificial intelligence...”**

Now, what do we gain through the use of this LSI technology?

To answer briefly, we can say that with LSI, search engines took a step forward to give us an ideal search result. Now you would ask - what is an ideal search result? Then, answer this! What do you look for when you type in a keyword or a context in Google’s search text box?

With the number of Web pages increasing voluminously on the Web, we would like to rely doubtlessly on the search engine and want to use them as a librarian with huge capacity to recall, ability to give the most precise and relevant results and that too, with wonderful sense of ordering. More technically, the ideal search engine should be able to cater to this trinity of recall, precision, and order. And this is where LSI fits in wonderfully and enhances the search engine’s power to converge with artificial intelligence.

Let’s have a look at the dumb computer problems that LSI can well take care of.

As we said earlier, a conventional search engine based on keyword searching may not give you the best results. This is simply because the search engine programs cannot differentiate between:

- ➔ Similar words with different meanings, ex: Monitor workflow or monitor
- ➔ Words that are similar in meaning but spelled differently, ex: disease and maladies
- ➔ Singular and plural forms of words, ex: button and buttons
- ➔ Words with similar roots, such as differed, differs, and different
- ➔ Other grammatically different words, such as roast, roasted

## Improved SERPs.

The LSI, because it focuses on a bunch of keywords, so to say, and not a single keyword,



and through its studied pattern of the relationship between semantically close and distant words in a collection of documents, it do not get confused between singular and plurals, or synonyms. It simply goes on to find the context developed by a bunch of keywords. So that, when you search for Tiger Woods, it doesn't go on to look for Web pages that has used the keywords "tiger' and 'woods' but lists a collection of pages that discusses Golf. This is what is called **relevance feedback**.

Usually, you will find that your search results are reduced with the increase in the number of keywords you search for. This is because a search engines functions better when they study, index, and recall for shorter and a simpler set of keywords. LSI goes the other way round and first focuses on knowing and analyzing a document exhaustively before indexing or categorizing it. Therefore, a latent semantic search engine allows a user to do an iterative search and provides useful feedback to frame a better search, if needed.

LSI is more close to human-generated taxonomies and categorization and takes a long step in structuring unstructured data. Hence, it is more **archive-friendly**. It allows archivists to efficiently label and index the LSI-generated categories. LSI does half the job and every document need not be indexed from scratch.

LSI helps in pointing out any part of content that is relevant but not covered in a document by comparing the data or content words on a given topic. This can find use in several contexts, one of them being a kind of automated grading system, where an assignment is compared to a sample of given quality.

LSI can investigate the semantic relationships within a text to decide on the relevance and consistency in the component parts. Adopting this into an application would enhance readability and comprehension. Naturally, these properties can be used effectively in instructional design and techniques.

## Stop SPAMS

However, the final and more relevant use of LSI is perhaps its power to filter information and prevent spamming or distribution of unsolicited electronic mail. By adapting and adjusting a latent semantic algorithm on your mailbox and feeding the details of known spam messages into it, junk mail can be prevented more effectively than with the current

system of keyword based approaches.

LSI is an extremely methodical technique that needs high amount of monotonous precision, one that a computer can do efficiently. As obvious, the technique involves a purely mechanical search based on an extensive evaluation of a set of words and comparing their presence in a much larger set of documents. The process can be automated because the computer does not need to understand either the search query or the meaning of the words.

---

# HOW HAS GOOGLE ADOPTED LSI ?

## Proof of LSI

**“A normal way of representing this is the grid or the matrix form; this is the reason why experts call the LSI method as ‘thinking inside the grid’...”**

Semantic text processing essentially understands linguists. Think of a statement; say, I am optimizing a paragraph for search engine. At least three to four words (I, am, a, for) in the statement are excesses, in the sense that they don't contribute actively towards the meaning of the sentence. They simply add value to the sentence grammatically. In this way, natural language contains numerous redundant and unnecessary words, from the point of view of search engines or semantic meanings. Functional words, conjunctions, prepositions, auxiliary verbs, and several other forms of words just add meaning to a sentence but do not add much content. Ironically, these are the most frequently used words in English.

In the very the first step in LSI, these words are picked up and ignored. The document is then left with words that may have some semantic meaning. We can discard:

- ➔ Articles, prepositions, and conjunctions
- ➔ Common verbs and pronouns
- ➔ Common adjectives (big, late, high)
- ➔ Frilly words (therefore, thus, however, albeit, etc.)
- ➔ Any word that appear uniquely in every document or in a particular document

## Inside the Grid

Now, our document has a much-reduced collection of words on which we can apply our statistical methodology. We can now start to index this collection of words in the document. A normal way of representing this is the grid or the matrix form; this is the reason why experts call the LSI method as ‘thinking inside the grid’. The grid or matrix contains the documents listed along the horizontal axis and the words contained in the documents along the vertical axis.

For the conventional keyword search, we just put a cross (X) in the column for any

document where a particular word (listed on the row) appears or just leave the column blank if the word does not appear. The grid then shows like this:

Document name/ Keywords contained	Elevation	Topography	Height	Tiger
GIS mapping	X	X	X	
Topology	X	X	X	
Rainfall harvesting	X	X		
Poetries of William Blake				X

Obviously, a grid may contain a cross or a blank. There is no midway and this way we can have an analysis of our document on keyword search. Note that we have left out any word or may have included it under any other column head if the form of the word varies, say it is 'topologies' that appear somewhere in the document and not 'topology'. If instead of looking for the presence of each keyword in a document we take into account how many times a word appears in the given document, the grid may appear something like this:

Document name/ Keywords contained	Elevation	Topography	Height	Tiger
GIS mapping	5	8	6	1
Topology	6	6	3	0
Rainfall harvesting	2	3	7	0
Poetries of William Blake	0	0	0	5

These figures give certain mathematical meaning. We can calculate the mean, median, and mode of the occurrence of certain words in the document and the correlation between them. This gives us a detailed analysis on our document collection. In case of LSI, we do exactly this. After removing unnecessary words from the documents, we generate the term-document matrix. A graphical representation of this matrix would give you the term-space and will have as many dimension as the number of content-wise meaningful words. This is because, to graphically represent the matrix, you will need as many axes to the

graph as there are content words.

Going by this application of the theory, if we try to analyse a real-life document collection and note down the occurrence of each content word, we will get numerous relevant content words. If these are recorded in the matrix, as above, and plotted on a graph, the result in the term space will also have numerous dimensions. This is true for each document in our collection. Each document is considered as a vector with the content words as their component. The documents with several common words will have vectors that are near to each other and hence, will be concluded to be semantically close. Documents with fewer common words will have vectors that are far apart and hence, are semantically distant.

It is mathematically possible to describe this space, although it is difficult to visualize such a space. However, if you try to visualize this multi-dimensional space, you can gain another interesting insight into LSI. Try looking at a branch of a tree full of green leaves. Since, there are leaves propping out at every possible direction, you will always fail to see all the leaves. That is, from whichever angle you try to look at the branch, few leaves will be hidden behind few others so that you can never see all the leaves at one go.

This idea can be contemplated as 'loss in information' and is a similar idea that you can use to visualize your n-dimensional term space. From whichever angle you look from, some vectors in your n-dimensional term space always overlaps others and the boundaries blur or collapse. In other words, similar keywords or content words loses their distinct identity and get squeezed together. Hence, the difference between singular and plurals, or synonyms or similar meaning words tend to attain a null value.

This idea can be contemplated as 'loss in information' and is a similar idea that you can use to visualize your n-dimensional term space. From whichever angle you look from, some vectors in your n-dimensional term space always overlaps others and the boundaries blur or collapse. In other words, similar keywords or content words loses their distinct identity and get squeezed together. Hence, the difference between singular and plurals, or synonyms or similar meaning words tend to attain a null value.

One thing to note here is that, although loss of information is deemed as a bad idea, it is

converted into a blessing when it comes to LSI. This technique of using or exploiting the feature of natural language, namely, similar-meaning words occur together, cuts off noise or unnecessary information. In the final lap, we can remove the hash from the hay.

Everyday, Google is taking a step to convert its whole search mechanism into an LSI-enabled one. Although, LSI is not adapted uniformly and in entirety, and not all searches will return a semantic word set now, the transition is visible in the search results. Conducting a search for 'phone' will show results in which the keyword 'phone' is contained and highlighted. However, if you add the tilde (~) before your keyword and search, ('~phone') your result will show the Web site for Nokia and the word 'Nokia' is now highlighted. From its new method of indexing, Google has determined that Nokia is relevant to phone.

---

# PROOF OF SEMANTIC RANKINGS

For serious Net surfers, Google's relevancy- and context-based search has caught their imagination. It's simple and gives you results without typing in exact or actual keywords or a combination of keywords. It can read your thoughts and present you a list of readings on the context you are searching. By now you know the baseline of LSI and context-based search. We'll show you a few proof of semantic rankings on Google's search engine.

## See for yourself

**"Notice the difference in the search results with the tilde and without it, in all of the various cases..."**


A very simple differentiation is still being maintained between a normal keyword search and a context-based search on Google. If you are going in for the LSI results, start your search word with a tilde (~) sign.

As proof of semantic rankings on Google's search engine, we have taken up random words or contexts and typed them with the tilde in the search box. We present few examples below followed by a brief analysis of the results that appear.

## Cola Example:

**Tip:** Save time by hitting the return key instead of clicking on "search"

[News results for ~cola](#) - [View today's top stories](#)

 [Coca-Cola deal hikes Patrick's finances for gov campaign](#) - [Boston Herald](#) - 7 hours ago  
[Can & Able: CPM, Coca-Cola water down differences](#) - [Economic Times](#) - 15 hours ago  
[Miller Appoints Coca-Cola Veteran CMO](#) - [Adweek](#) - 20 hours ago


[Pepsi World Flash Check](#)  
Pepsi's official web site.  
[www.pepsi.com/](#) - 9k - [Cached](#) - [Similar pages](#)

[Pepsi World](#)  
Copyright 2005 © PepsiCo, Inc. | [Terms and Conditions](#) | [Company Info](#).  
[www.pepsi.com/home.php](#) - 5k - 6 Apr 2005 - [Cached](#) - [Similar pages](#)  
[ [More results from www.pepsi.com](#) ]

Here is a search query on "cola". Notice the difference in the search results with the tilde and without it, in all of the various cases.

**Tip:** Save time by hitting the return key instead of clicking on "search"

[News results for cola](#) - [View today's top stories](#)

 [Coca-Cola deal hikes Patrick's finances for gov campaign](#) - Boston Herald - 7 hours ago  
[Can & Able: CPM, Coca-Cola water down differences](#) - Economic Times - 15 hours ago  
[Miller Appoints Coca-Cola Veteran CMO](#) - Adweek - 20 hours ago

[Welcome to Coca-Cola](#)  
[www.cocacola.com/](http://www.cocacola.com/) - 3k - [Cached](#) - [Similar pages](#)

[Welcome to Coca-Cola](#)  
The site for Coke. Get wallpaper, music, read the FAQ and find out about the company and its products.  
[www.coca-cola.com/](http://www.coca-cola.com/) - 3k - [Cached](#) - [Similar pages](#)

The listings are utterly different. Whereas in the normal keyword search, Google has presented generously the page that has used the keywords optimally, in the search with the tilde, the search engine has tried to filter in according to the context


## Maid Example:

Same is for the search with 'maid' as the search keyword. Artificial intelligence is able to give you a listing of relevant Web pages.

It deduced that you are looking for people who can housekeep for you. A normal keyword search on 'maid' would have brought you irrelevant results with the keyword used in various other contexts. Check it out yourself.

**Tip:** Save time by hitting the return key instead of clicking on "search"

[News results for maid](#) - [View today's top stories](#)

 [Jacko maid says Victorian boy groped](#) - Melbourne Herald Sun - 52 minutes ago  
[Defence counsel refutes prosecution's submission that maid is ...](#) - Channel New  
[Maid: Jackson Showered With Boy](#) - CBS News - 22 hours ago


[Molly Maid Quality Maid Service](#)  
A site to find household help, be a **maid** or own a franchise.  
[www.mollymaid.com/](http://www.mollymaid.com/) - 10k - [Cached](#) - [Similar pages](#)

[Maid Service From Maid Brigade](#)  
Professional **maid** service from the **maid** service company with over 20 years of experience.  
[www.maidbrigade.com/](http://www.maidbrigade.com/) - 8k - [Cached](#) - [Similar pages](#)

[Maid of the Mist](#)  
Riverborne journey to the foot of the falls. Passengers taken on from both Canadian

Here's another example of semantic or context-based search. Type in the words 'Tiger' and 'Golf' in the Search text box on Google home page. Here's the result you get.

News results for **Tiger golf** - [View today's top stories](#)

 [Tiger finds himself up the creek](#) - Australian - 47 minutes ago  
[Beware, this Tiger's ready to pounce](#) - MSNBC - 12 hours ago  
[Golf Masters starts after 5 1/2-hour delay](#) - Xinhua - 14 hours ago

[Official Website for Tiger Woods](#)  
Official site, offering news, biographical information and statistics, audio and video clips, photos, and merchandise.  
[www.tigerwoods.com/](http://www.tigerwoods.com/) - 5k - [Cached](#) - [Similar pages](#)

[Tiger Golf - MiniClip.com](#)  
GET NEW FREE GAMES BY EMAIL. Send 2 a Friend. Email me this as an attachment (816k). Put this on my website for free.  
[www.miniclip.com/tigergolf.htm](http://www.miniclip.com/tigergolf.htm) - 14k - [Cached](#) - [Similar pages](#)

Google don't try at all to give you a list of Web pages that had Tiger and golf appearing all over their body text, anchor, or URL. Instead, it detects the semantically close words and concludes on the context of your search.

You can have the proof for yourself. Try the search with different other keywords you can think of and compare the results.

---

**G-68, EAST OF KAILASH,  
NEW DELHI-110065 INDIA**

**+91-11-51623530**

**+91-11-51623531**

**[www.sem.mosaic-service.com](http://www.sem.mosaic-service.com)**