

LATENT SEMANTIC INDEXING

Introduction to LSI

LSI (Latent Semantic Indexing) is a filter that Google uses to determine the relevance of a website by quickly comparing its content to that of existing websites within the Google network.

If website A does not contain the expert verbiage that is commonly used on Website B,C,D,E,F,G... within the subject matter then website A will not be found within the top results.

LSI is an algorithm that closely resembles the thought processes that an actual "human" would go through in order to determine if the results of their query are relevant to what they were searching for.

In other words the search engines are the closest they have ever been to being able to quickly determine relevance based on what an actual human would find relevant by comparing the structure and "words" of a page and website and then comparing them to those of websites that are already considered relevant.

For example, if you search for WWII in Google, it will show you World War II related content. How would Google know that WWII means World War II? WWII is semantic (means the same thing) to World War II based on a high number of results.

Semantic does not always means Synonyms, they can be slangs, abbreviations, made up words, etc.

Here is another example, if you were to search the internet in the keyword "golf", the majority of websites you will find with high keyword density "golf" will also have a high density of other words such as:

- sports
- golf
- country
- golf's
- golfer's
- club
- golfer

You can see this for yourself if you go to Google and type "~golf". The mark "~" in front of the keyword means you want to see semantic results. Most people don't know about the Google "~" function to find "search engine determined" synonyms. The key word here is "search engine

determined" not thesaurus determined. Look at the top site for the keyword "golf" and you will see that they don't simply insert the word "golf" all over their home page. What you will find is "expert verbiage", which is words other than the word "golf", yet related to golf.

If you cover all LSI keywords for golf, than the LSI algorithm will determine that you completely cover the theme of golf. Practice this technique throughout all the web pages on your website and you will automatically deserve the #1 position in Google for the appropriate keyword.

Co-Occurrence Algorithm

Co-occurrence is the percentage of websites that contain either the main theme keyword (or key phrase) and a secondary keyword (synonym) as well.

On December 28th 2006 Google filed a new patent application titled "[Detecting spam documents in a phrase based information retrieval system](#)". Google engineer, Anna Lynn Patterson is the inventor of this patent. Here is a quote directly from the patent:

"An information retrieval system uses phrases to index, retrieve, organize and describe documents. Phrases are identified that predict the presence of other phrases in documents. Documents are then indexed according to their included phrases. A spam document is identified based on the number of related phrases included in a document."

Google has filed a patent for an algorithm that will index and rate the "relevance" of web pages to determine if there is also an occurrence of phrases related to the subject matter of a web page.

William Slaw ski of SEObyTheSea has written a very detailed informative article about [Phrase Based Information Retrieval and Spam](#).

To give you a break down you must not only use LSI keywords, but you have to make logical sense on how you use them or they simply won't count towards your ranking.

Here is an example:

The phrase "President of" predicts "President of the United States", "President of Mexico", "President of AT&T", etc.

All of these latter phrases are phrase extensions of the phrase "President of" since they begin with "President of" and are super-sequences thereof.

It's useful because it can predict one of those other phrases. But, if it doesn't predict at least one other phrase that isn't an extension of it, it may be seen as an incomplete phrase.

"President of the United" is an incomplete phrase because the only other phrase that it predicts is "President of the United States" which is an extension of the phrase.

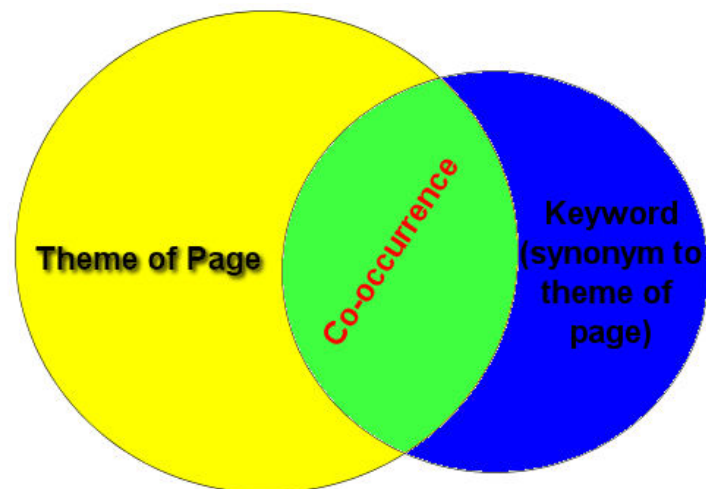
This incomplete phrase list might be kept to help searchers. When a search query is received, it can be compared against the incomplete phrase list.

For example, if the search query is "President of the United," the search system can automatically suggest to the user "President of the United States" as the search query.

Each good phrase is used with sufficient frequency and independence to represent meaningful concepts or ideas expressed in the corpus.

Keep in mind that case sensitive also plays a role, if your content says president of the united states and the LSI and Co-occurrence phrase is President of the United States, your phrase will not be valued and will be rejected.

Chart example:



Theme of Page can be any specific keyword or general phrase. In a niche retail market, if you specialize in Nissan 350Z, then Nissan 350Z would be the Keyword. The theme would be automotive.

Co-occurrence is the percentage (%) of web pages that contain both the Theme of the page (keyword) AND the keyword (synonym).

Keywords (LSI Words) If you have a page whose main keyword (theme) is "350Z". Your LSI words may be "Sports Car"

Spam Filters

How does Google know you're stuffing keywords into your content? From the foregoing, the number of the related phrases present in a given document will be known. A normal, non-spam document will generally have a relatively limited number of related phrases, typically on the order of between **8 and 20**, depending on the document collection. By contrast, a spam document will have an excessive number of related phrases, for example on the order of between 100 and 1000 related phrases. Thus, the present invention takes advantage of this discovery by identifying as spam documents those documents that have a statistically significant deviation in the number of related phrases relative to an expected number of related phrases for documents in the document collection.

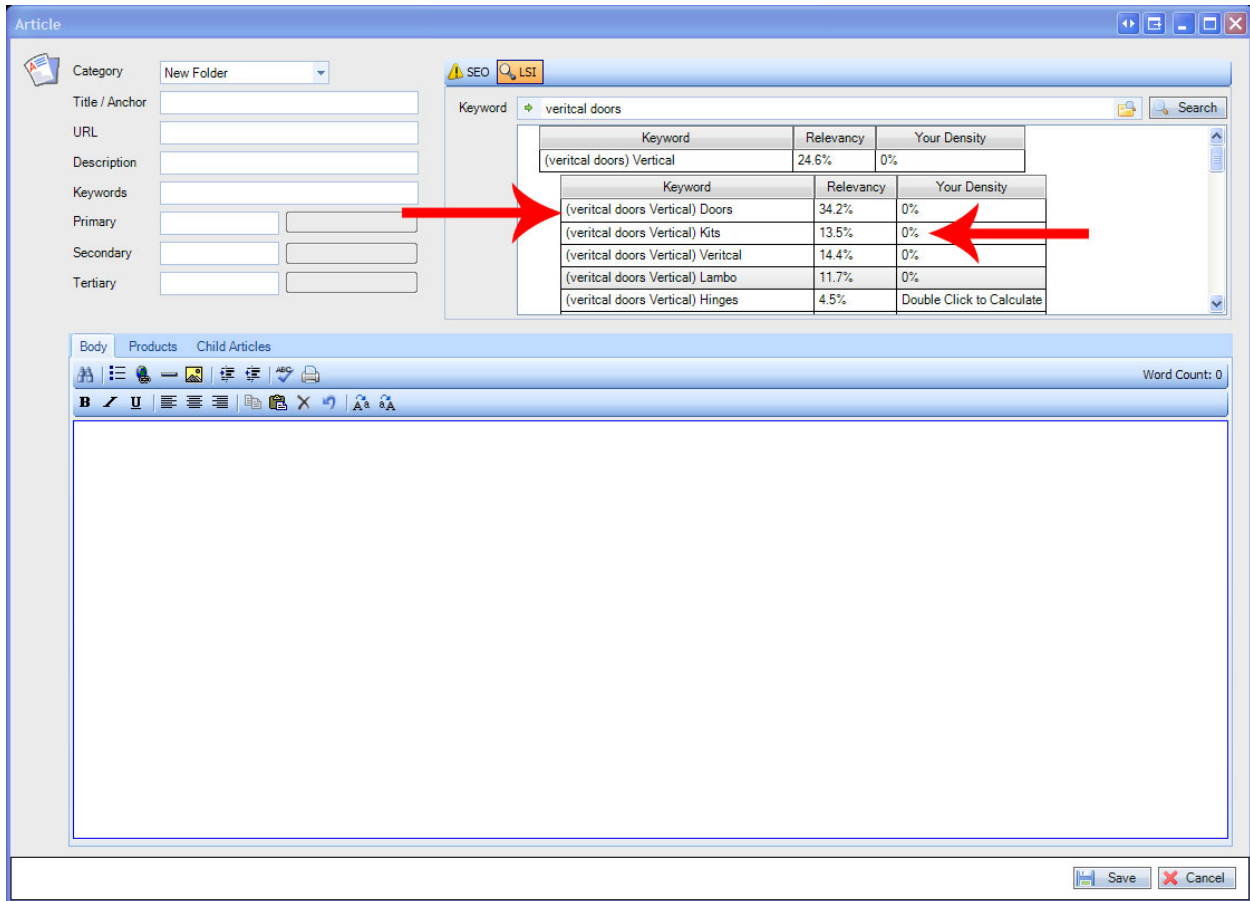
Shopping Cart Elite LSI Tool

Using the LSI tool for your advantage is easy. The goal is to enter 8-20 LSI keywords into your content so your content will be more relevant. As indicated in the screen shot below, simply enter your primary, secondary, tertiary keywords separately or enter them in a phrase inside the LSI tool.

Keyword	Relevancy	Your Density
(vertical doors) Vertical	24.6%	Double Click to Calculate
(vertical doors) Custom	7.8%	Double Click to Calculate
(vertical doors) Body Kits	3.0%	Double Click to Calculate
(vertical doors) Inch	3.6%	Double Click to Calculate
(vertical doors) Horizontal	4.2%	Double Click to Calculate
(vertical doors) Exit	3.0%	Double Click to Calculate
(vertical doors) Security	3.6%	Double Click to Calculate

Click search and Shopping Cart Elite will automatically grab all the data from Google tilde searches and perform calculations against your article with the LSI Keywords. You will get a list of LSI keywords, density within your article and relevancy between your keyword and the LSI keyword.

The screen shot below shows you the Keyword that is being used to find an LSI keyword within the parenthesis “(Keyword)”. The LSI keyword that is relevant for the keyword inside the parenthesis “(Keyword)” is outside of the parenthesis “(Keyword) LSI Keyword”.



Keyword – The keyword used in the screen shot is “Vertical Doors”

LSI Keyword – One of the LSI Keywords for “Vertical Doors” is Vertical (24.6% relevancy). Usually the 1st level that shows up is treated as a general keyword, by double clicking on the phrase you will get into the 2nd level of LSI keywords. Notice after we click on “(Vertical Doors) Vertical” the next level came up. This is because we performed a search for “Vertical Doors Vertical” when we clicked on it. The LSI keywords for Vertical Doors Vertical are:

- Doors – 34.2%
- Kits – 13.5%
- Lambo – 11.7%

Hinges – 4.5%

You can double click on any first level LSI keyword/phrase and it will take you to the second level, third level, etc. As the levels go down they show you what the LSI keyword/phrase is for the top level LSI keyword/phrase. Going down 1 level is mandatory, going down 2-3 levels is recommended.

The screenshot shows the 'Article' editor interface. On the left, there are fields for Category (New Folder), Title / Anchor, URL, Description, Keywords, Primary, Secondary, and Tertiary. On the right, the LSI analysis is displayed. The main keyword is 'vertical doors'. Below it, there are three levels of sub-keywords. A red arrow points to the 'Conversion' sub-keyword.

Keyword	Relevancy	Your Density
(vertical doors Vertical) Lambo	11.7%	0%
(vertical doors Vertical Lambo) Upgrade Kits	10.0%	0%
(vertical doors Vertical Lambo Upgrade Kits) Conversion	23.0%	Double Click to Calculate
(vertical doors Vertical Lambo Upgrade Kits) Power Upgrade Kit	8.2%	Double Click to Calculate
(vertical doors Vertical Lambo Upgrade Kits) Degree	6.6%	Double Click to Calculate

Relevancy – Relevancy shows you how relevant this keyword is to the main keyword as well as the additional keywords that might have been added as you go down the levels. If relevancy is above 10% you must put that keyword/phrase into your content. Any keyword/phrase that is above 3% is recommended to be a part of your content.

Density – Density shows you the density percent of the LSI keyword within your article content. If you double click on the phrase “density” next to each keyword, it will show you the percent. You should keep the percent under .7% or usually 1 keyword.

The article you read about LSI is to explain the concept of how it works. To practice it in Shopping Cart Elite all you need to do is insert 8-20 LSI words into your content that is relevant to your article. Inserting only one word or phrase will do the job. Don't forget that case

sensitive and co-occurrence phrases count. Don't just stuff the LSI keywords into your content; make logical sentences out of them.