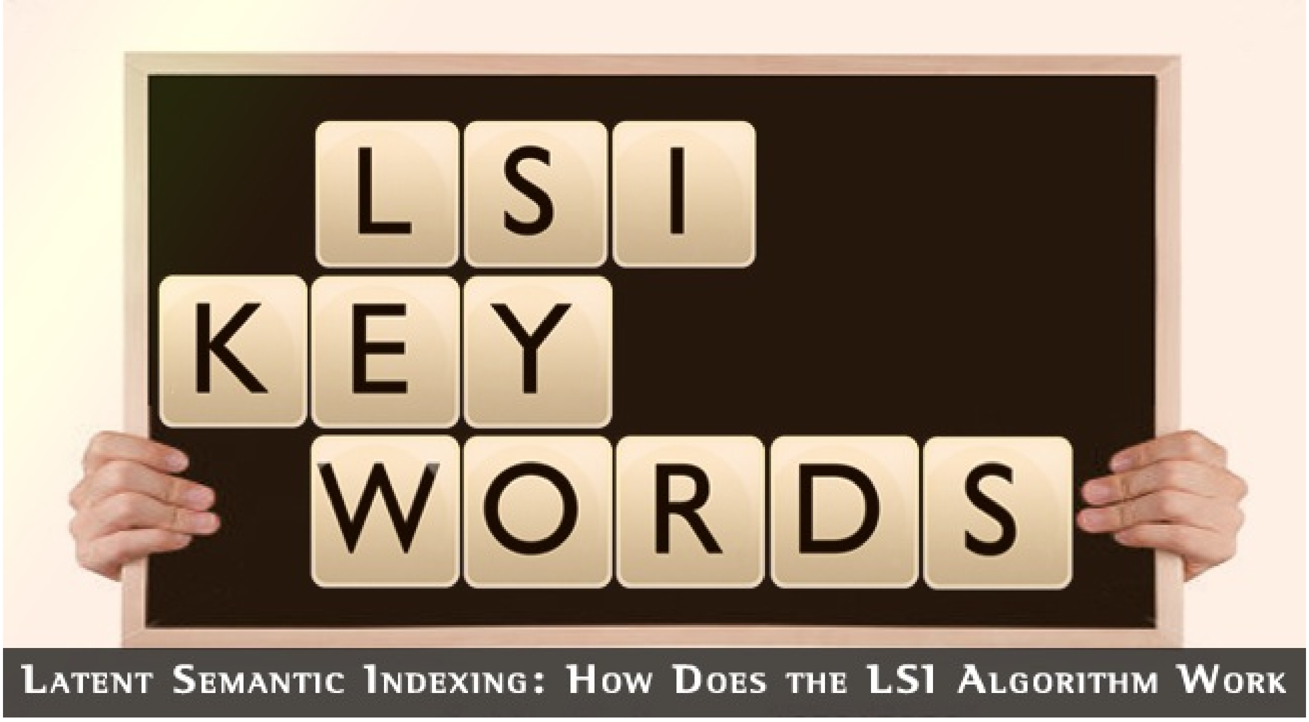
Latent Semantic Indexing: How Does the LSI Algorithm Work?

www.techlila.com http://www.techlila.com/latent-semantic-indexing/

Google uses LSI to assess the meaning of the written content on your blog or website. Latent semantic indexing is a misnomer for 'latent semantic analysis,' a statistical analytical technique that can use character strings to determine the semantics of text – what that the text actually means.



Here we shall discuss some aspects of LSI that make you think differently about keywords and how you write your content. Keep in mind that Google is larger in the search traffic it gets that all the other search engines combined. Also, some of these others use Google data. That is why we focus on it.

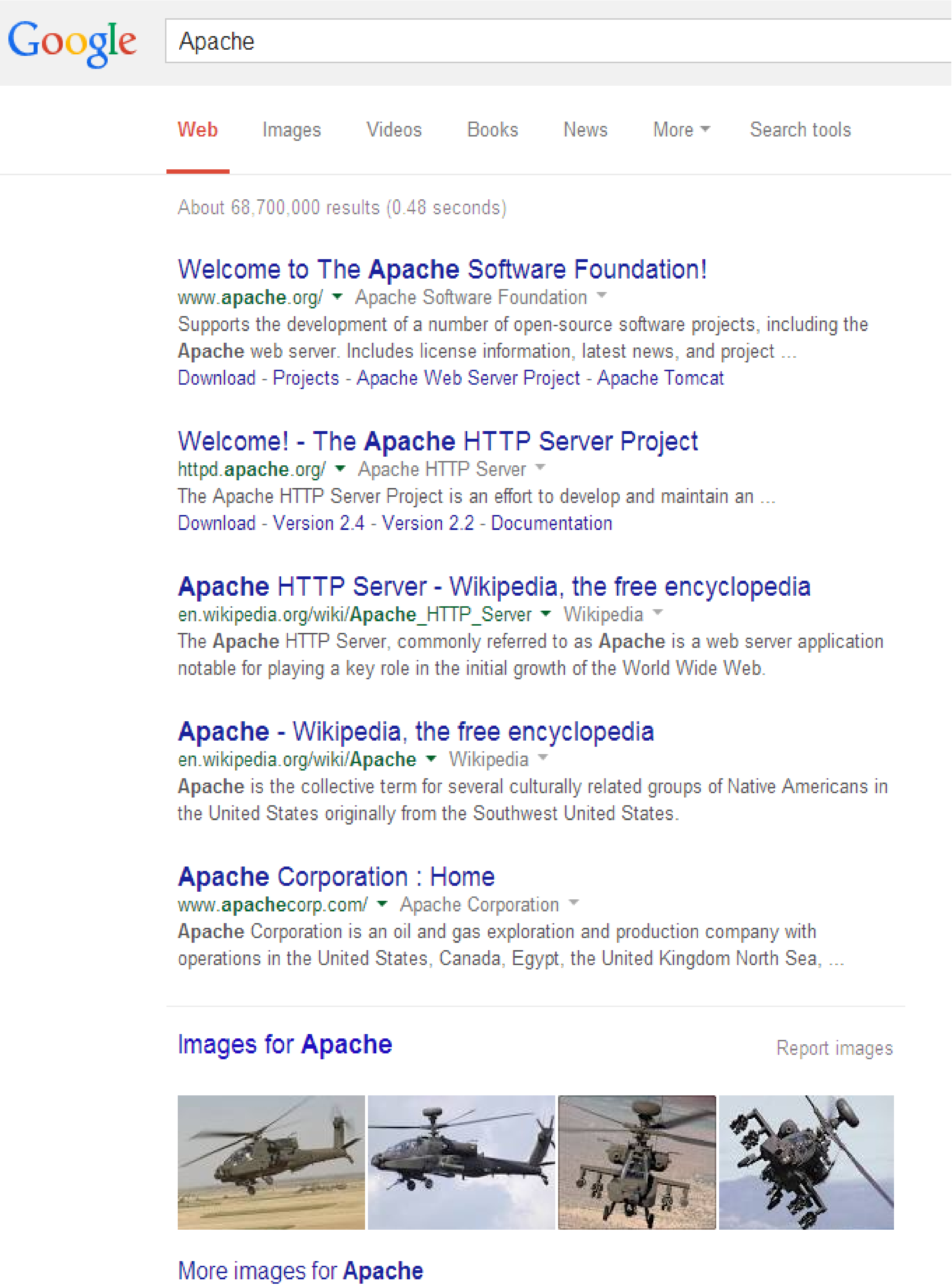
What is Latent Semantic Indexing?

Google's objective is to provide its clients with the best information it can when they carry out a search. Google must therefore fully understand exactly what information searchers are looking for when they use keywords for information, and also how well each indexed web page provides that information.

Google's latent semantic indexing (LSI) algorithm was developed to use the character strings in a document to establish its semantic relevance to the search term (keyword) used. In other words, to help establish the true meaning of the text on a blog post or web page.

The LSI algorithm considers all the constituent terms used in the text of a document to establish its true meaning in relation to the keywords employed. That is why it is important to be specific when searching for information on Google. If you use the word 'Apache' as a search term in Google, your first six results include the web server application, the Native American tribe, the apache helicopters and an oil and gas exploration company using that name.

You cannot assume that Google will return the same definition for the search term as you are thinking about. You must not only use your main keyword, but define its definition in relation to your use of it.



Apache

About 68,700,000 results (0.48 seconds)

Welcome to The **Apache** Software Foundation!
www.**apache**.org/ Apache Software Foundation
Supports the development of a number of open-source software projects, including the **Apache** web server. Includes license information, latest news, and project ...
Download - Projects - Apache Web Server Project - Apache Tomcat

Welcome! - The **Apache** HTTP Server Project
httpd.**apache**.org/ Apache HTTP Server
The Apache HTTP Server Project is an effort to develop and maintain an ...
Download - Version 2.4 - Version 2.2 - Documentation

Apache HTTP Server - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Apache**_HTTP_Server Wikipedia
The **Apache** HTTP Server, commonly referred to as **Apache** is a web server application notable for playing a key role in the initial growth of the World Wide Web.

Apache - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Apache** Wikipedia
Apache is the collective term for several culturally related groups of Native Americans in the United States originally from the Southwest United States.

Apache Corporation : Home
www.**apache**corp.com/ Apache Corporation
Apache Corporation is an oil and gas exploration and production company with operations in the United States, Canada, Egypt, the United Kingdom North Sea, ...

Images for **Apache**

More images for **Apache**

In the same way, if you use the concept of LSI in your text, then you can increase the potential for your page or blog post to be listed for this search term in the context in which you are using it. That is true, even if the text on your webpage or blog post does not mention the keyword ‘apache’ – or any other keyword you are using. Here is why.

LSI Involves Character String Analysis

The algorithm does not use a dictionary as we know it, but involves a complex statistical and mathematical analysis of the individual characters and character strings used in the text that makes up a particular web page. It is immaterial whether that is a blog page, a post, a full page on a website or even a post comment or forum string. LSI is used to evaluate the content of any individual file published on the web.

Because of this, Google has introduced a very powerful means of distinguishing between language, vocabulary and semantics. How do these differ? Here’s how:

Language: The word usually refers to the forms of written or spoken words specific to a certain country or community. Thus, the text string ‘pain’ means ‘bread’ in French, but not in English. Google cannot tell the difference unless it understands the language being used.

Vocabulary: This word refers to the words used within a specific language. It can also refer to the range of words known by a specific person within a language. ‘Bad language’ refers to use of vocabulary and not language. ‘Box’ is a word in English vocabulary, but that word has many meanings, leading us to:

Semantics: The same word can be used to mean many things. Thus the word ‘box’ can mean a container, to fight, to recite the points of a compass or a type of bush. Semantics refers to the meaning of words in a specific context. Bad language can express meaning very well – so may form effective semantics!

Syntax: The way in which words are combined in a sentence. Thus, the words ‘the dog bit the man’ could be expressed using different syntax as ‘*the man bit the dog.*‘ Before LSI Google could not distinguish the difference – it just detected the words.

Google’s LSI algorithm takes each of these factors into account. If you think deeply enough, you will be able to understand why keyword repetition is pointless, and why a low keyword density (KD) can provide you with better ranking results than a high KD.

The algorithm analyzes the meaning of the words in your webpage or blog using semantics and syntax, and matches these with the perceived meaning of the keywords used by the person carrying out the search.

Keyword or Meaning?

What is more important to your readers? Keywords or meaning? Let’s say you are seeking information online using a search engine. What would you rather find: a page offering lots of repetitions of your keyword, but very little else, or one that provides you with what you are looking for, even if you can’t find your search term in the text?

You know the answer to that, and so does Google. The problem was, that in the early days of its existence, the Google search engine was able to find, index and rank only those web pages that contained the search term (keyword/phrase) used by those seeking information.

If your blog post or web page did not contain the phrase keyword used within the search term several times, then it would not be listed and made visible to anybody using that phrase – in future we refer to keywords as being one word or several. So guess what!

Keyword Stuffing and Software

It became common practice to stuff web pages full of keywords. The more the merrier, and Google would rank them high in the results pages for that keyword. Entrepreneurs made their fortunes by designing software (apps to you younger people) that would take one page of text and generate hundreds of others, changing nothing but the keyword used for that page.

Many people that were using the search engine to find for what was them very important information, were being presented with page upon page of useless drivel that gave them nothing but adverts and the same stuff repeated over and over again.

Even webmasters complained to Google about how such pages could be listed above those that were genuinely offering information. The reason was, of course, keyword repetition. The way the ranking algorithm was set up was that the more keywords, the higher the ranking. This had to stop.

Google spotted this, and so decided to do something about it, but what? It began by using the Adsense algorithm that used semantics to establish the best type of ads for any specific web page. It developed this concept further, using latent semantic analysis to create the latent semantic indexing algorithm. More on this shortly.

LSI uses the concept of latent semantic analysis to survey all the vocabulary, syntax and semantics on a page to establish its true meaning. By means of LSI, Google can compare the search term used by its customer with indexed web pages and establish which best matches that search term/keyword by analyzing all the vocabulary on the page, not just keywords. How does it do that?

The Problem of Ambiguity in Language and Vocabulary

Let’s say that you are writing a book about the invention and use of locks throughout history. You need some information online about the topic, so you enter the search term ‘locks and their history.’ Or maybe, ‘the history of locks.’

The first question to consider would be “what type of locks?” Are you writing about security locks – locks and keys, or about canal locks? Or maybe you are even referring to locks of hair? To most people, only the first two options would be likely – canal locks or those that need keys to open.

To Google, however, all are the same. The keyword is ‘locks’ or even ‘history of locks’ but how does the machine know what you are talking about? It cannot ask you – all it can do is to take the character string that makes up the words, and search for it within its indexed blog posts and web pages.

The Effect of the Applied Semantics Acquisition

The answer came after Google purchased a Santa Monica company in 2003 known as Applied Semantics. This firm was working on algorithms that applied semantics to the understanding of the true meaning of written text. Google purchased the company and then applied its principles to its Adsense program.

This is the program mentioned earlier, where Google places relevant PPC adverts on your web pages. Applied Semantics principles were used to establish the best type of advert for your page, based upon the true focus of its content.

Google continued the development of this mathematical analytical technique and finally came up with what it referred to as Latent Semantic Indexing. Using LSI, it is possible for Google to index and then rank your page for its meaning and total content rather than upon only its use of keywords.

How Does the LSI Algorithm Work: Keywords and Semantics?

Google will look at other vocabulary on your page then carry out a statistical analysis of the context and syntax of such vocabulary. If a Google user searches for 'history of security locks' then Google will take other vocabulary of its indexed pages into account. If your page contains words such as 'keys,' 'levers,' and 'doors' then it will associate this vocabulary with security locks.

After taking other ranking factors into account, it will then list your page in the search results pages for that keyword (history of security locks) ranked according to the benefits that Google believes it to be offering to the searcher. Before LSI, the searcher would also be given pages focusing on canal locks and even on hair.

Avoid Keyword Stuffing

The whole point being made here is that there is no longer a need for the excessive repetition of keywords. Since Google introduced LSI, all you need do is to make sure that you use as many synonyms and related terms as you can to the keyword you are chasing.

But do not get us wrong – keywords still count. You should still use relevant keywords, but Google is using the LSI concept to determine what website content is really about: what it is really saying. It is catching out pages written specifically to get listed for individual keywords, but that have little useful content other than meaningless repetitions of the keyword.

You can still use them, although even 1.5% KD or less will be quite sufficient. Use keywords sparingly and also use other terms meaning the same thing in the context of your niche. That's the answer to your question – How Does the LSI Algorithm Work – and why latent semantic indexing is a term you need not remember as long as you understand and apply the concepts Google is using when it employs LSI in its indexing and ranking algorithms.