

April 6, 2005

## **Google United – Google Patent Examined**

**By Jim Hedger, StepForth News Editor, [StepForth Placement Inc.](#)**

Thoughts on Google's patent... "[Information retrieval based on historical data](#)"

Google's newest patent application is lengthy. It is interesting in some places and enigmatic in others. Less colourful than most end user license agreements, the patent covers an enormous range of ranking analysis techniques Google wants to ensure are kept under their control. Some of the ideas and concepts covered in the document are almost certainly worked into the current algorithm running Google. Some are being worked in as this article is being written. Some may never see the blue-light of electrons but are pretty good ideas so it might have been considered wise to patent them. Google's not saying which is which. While not exactly War and Peace, it's a pretty complex document that gives readers a glimpse inside the minds of Google engineers. What it doesn't give is a 100% clear overview of how Google operates now and how the various ideas covered in the patent application will be integrated into Google's algorithms. One interesting section seems to confirm what SEOs have been saying for almost a year, Google does have a "sandbox" where it stores new links or sites for about a month before evaluation.

Google is in the midst of sweeping changes to the way it operates as a search engine. As a matter of fact, it isn't really a search engine in the fine sense of the word anymore. It isn't really a portal either. It is more of an institution, the ultimate private-public partnership. Calling itself a media-company, Google is now a multi-faceted information and multi-media delivery system that is accessed primarily through its well-known interface found at [www.google.com](http://www.google.com).

Google is known for its from-the-hip style of innovation. While the face is familiar, the brains behind it are growing and changing rapidly. Four major factors (technology, revenue, user demand and competition) influence and drive these changes. Where Microsoft dithers and .dll's over its software for years before introduction, Google encourages its staff to spend up to 20% of their time tripping their way up the stairs of invention. Sometimes they produce ideas that didn't work out as they expected, as was the case with Orkut, and sometimes they produce spectacular results as with Google News. The sum total of what works and what doesn't work has served to inform Google what its users want in a search engine. After all, where the users go, the advertising dollars must follow. Such is the way of the Internet.

In its recent SEC filing, the first it has produced since going public in August 2004, Google said it was going to spend a lot of money to continue outpacing its rivals. This year they figure they will spend about \$500 million to develop or enhance newer technologies. In 2004 and 2003, Google spent \$319 million and \$177 million respectively. The increase in innovation-spending corresponds with a doubling of Google's staff headcount which has jumped from 1628 employees in 2003 to 3021 by the end of 2004.

Over the past five years Google has produced a number of features that have proven popular enough to be included among its public-search offerings. On their front page, these features include Image Search, Google Groups, Google News, Froogle, Google Local, and Google Desktop. There are dozens of other features which can be accessed by clicking on the "more" button near the upper right of the screen. We

believe that Google is working to tie all these features together to present its users with search options that are, for want of a better phrase, more relevant than those offered by its competitors. As the Internet and technologies available for users advances, different types of files become searchable and therefore relevant to users. Take Google Video as an example. Now Google (and some of its competitors) can find and read text from closed captioning scripts. As well quotes from recent episodes of virtually any TV show are searchable and can be served back to users along side the clip where the quote originated. Now, imagine a merging of video, textual, graphical and audio files in organic search results. This is, in our opinion, the true intent of the ideas contained in the patent document.

The patent document relates primarily to sorting and cataloging organic search results. As we know them today, organic search results at Google are influenced by a number of factors, many of which involve an evaluation of incoming links. Google needs to ensure its users and advertisers that it is capable of taking action against the darker facets of the search engine optimization sector. Recent stories in the mainstream press have left many with the impression that dark-art SEO and link-spamming is the surest way to get top placements. Google engineers take pride in their work and the popularity of their organic search results is the bedrock on which their profitable business models are built. They can't afford to allow link-spam and deceptive SEO techniques to dominate their organic listings, especially as these listings are about to address and catalog a much more robust and complicated Internet.

Over the past ten months, SEOs have complained and questioned the phenomena known as the Google Sandbox. The sandbox theory explains the time-lag between link-acquisition for a site and link-recognition and reward by Google. A few key sections of the patent document fill in the blanks for SEOs on what Google is examining when a finely crafted link-building campaign falls into the sandbox. The biggest influencer is links and Google is finding new and improved ways to evaluate them.

Google's core algorithm is based on measuring links coming into a page. Because of this, link-building is part of any good search engine optimization campaign. In the span of a month, incoming links to one or more pages of a website might jump by hundreds or thousands. Some of those links might be useful in Google's eyes and some might be useless. The question is, how does it sort which is which?

Google collects a lot of data when it examines a page and the links directed on to or off of that page. When Google mentions they are using "historic data" to determine the value of links directed to your page, they are referring to a number of factors. It knows how long the page has been online, or at least when it first became aware of said page. It also knows how long pages linked to have been online. It knows how often links get clicked and also knows which computer, (and in many cases, exactly who) is clicking the link and where that clicking is coming from.

For an example, check out the following sections from the patent document:

- 1. A method for scoring a document, comprising: identifying a document; obtaining one or more types of history data associated with the document; and generating a score for the document based on the one or more types of history data.*
- 2. The method of claim 1, wherein the one or more types of history data includes information relating to an inception date; and wherein the generating a score includes: determining an inception date corresponding to the document, and scoring the document based, at least in part, on the inception date corresponding to the document.*
- 3. The method of claim 2, wherein the document includes a plurality of documents; and wherein the scoring the document includes: determining an age of each of the documents based on the inception*

*dates corresponding to the documents, determining an average age of the documents based on the ages of the documents, and scoring the documents based, at least in part, on a difference between the ages of the documents and the average age.*

- 4. The method of claim 2, wherein the generating a score for the document includes scoring the document based, at least in part, on an elapsed time measured from the inception date corresponding to the document.*

By the time a reader gets to item 63, the document has covered dozens of page, site, link and URL related factors that may or may not be included in the current working algorithm.

Here is a quick breakdown of "history factors" we think are relevant to Google's algorithm today. Please note, each item might refer to a specific page and at the same time, also refer to all other pages associated with it.

- How long a domain or URL has been registered.
- Has ownership of a domain changed after previous registrations expired?
- Has the physical location of the registrant changed?
- How lengthy is the URL itself? Was it registered to game the index?
- How many pages are included in the website? (A one document or page website is not considered a highly relevant source of information.)
- Freshness and age of document.
- Use of anchor text (both on site and in links directed to site).
- "Trust Factors" regarding sites or pages outbound links refer to, and inbound links are found on.
- The "discovery date" of a particular link and the history of changes involving that link.
- Rate of growth for new links. A sudden burst of growth likely indicates some form of link-spam.
- Variations in anchor text used to phrase links directed to a page being evaluated. If the same anchor text is used in every inbound link, are they phrased that way for branding purposes or spamming purposes?
- Number of searches for keyword phrase associated with the anchor text used in links.
- Number of times Google users click on Google results by entering keyword phrases used in anchor text of incoming links. Does the page being evaluated receive visitors for that keyword phrase on Google's search engine?
- How do users actually behave while on the page, site or document being evaluated?

There is a lot more to find in this document. Thus far, the more we explain, the more questions we have. One thing we are very sure about, the intent of the ideas covered in the patents extends beyond the search tool we know now. We expect to publish a white paper on our analysis of the patent and its implications early next week. Until then, we advise our clients to stay the course. We have long preached a very conservative approach to Google based on relevant link building (which can be slow going but very effective), highly stratified content that is relevant only to the topic addressed by the site, and clear paths based on multiple keyword phrases for spiders to follow.