Improving the integrity of Internet search through the use of extrinsic business information

John Nagle
SiteTruth
999 Woodland Avenue
Menlo Park, CA 94025
650-326-9109

nagle@sitetruth.com

ABSTRACT

We have developed a search system resistant to “search engine optimization”. We perform an automated “due diligence” on web sites which offer products and services for sale. By mining multiple databases to obtain information about the legitimacy of the seller, legitimate businesses can be moved upward in search results, while less legitimate businesses can be moved downward.

Categories and Subject Descriptors

J.7 [Computer Applications, Computers in Other Systems]

General Terms

Security, Human Factors, Legal Aspects

Keywords

Web, spam, search, integrity

1. INTRODUCTION

Search engines began as a form of straightforward indexing. Today, they reflect an ongoing battle between “search engine optimization” companies trying to promote products and services, and search companies trying to produce search results resistant to attempts to game the search system.

We have developed a search system resistant to such gaming. We perform an automated “due diligence” on web sites which offer products and services for sale. By mining multiple databases to obtain information about the legitimacy of the seller, legitimate businesses can be moved upward in search results, while less legitimate businesses can be moved downward. This approach is resistant to most “search engine optimization” techniques. The goal is to create a user view of the World Wide Web with more real businesses and far fewer marginal web sites, thus reducing user frustration and making web search more satisfying and effective.

2. RATING BASIS

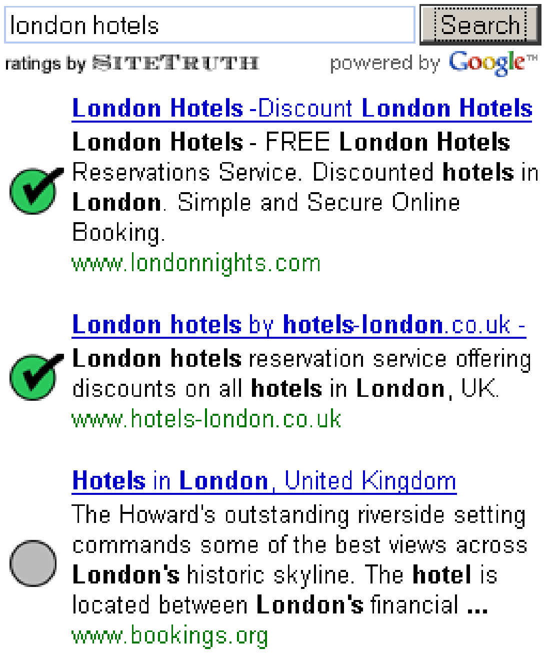
The basic question SiteTruth answers is “If I needed to find the business behind this web site, could I do so?” Ratings are based not on popularity, but on legitimacy as a business.

Under California law, operating an online business which accepts credit cards, without clearly disclosing the actual name and address of the business, is a criminal offense.[1] This applies to all web sites that accept transactions from Californians, which covers most of the English-speaking world. We, as a California-based operation, take this as a basis for site integrity checking – if we can't determine the name and address of the business, it will receive a very low ranking.

We thus attempt to find the name and address of each business with a web site. We look at SSL certificates[2], seals of approval from organizations such as the Better Business Bureau, connections to payment sites which identify their merchants, and the contents of the site itself[3]. We make little use of WHOIS information due to its low data quality.

3. THE USER'S VIEW

Figure 1: Results for a search phrase known to attract substantial “search engine optimization” efforts.



london hotels

✓ **London Hotels** -Discount **London Hotels**
London Hotels - FREE **London Hotels** Reservations Service. Discounted **hotels** in **London**. Simple and Secure Online Booking.
www.londonnights.com

✓ **London hotels** by **hotels-london**.co.uk -
London hotels reservation service offering discounts on all **hotels** in **London**, UK.
www.hotels-london.co.uk

○ **Hotels** in **London**, United Kingdom
The Howard's outstanding riverside setting commands some of the best views across **London's** historic skyline. The **hotel** is located between **London's** financial **...**
www.bookings.org

Here we are annotating a search from a popular search engine. The integration between the search engine and SiteTruth's system is being performed in the user's browser, using JavaScript and XML queries to the SiteTruth server. The SiteTruth system immediately returns results for sites previously rated. Unknown sites are processed with a quick “site crawl”, looking at key pages for the items mentioned above. This typically takes between two and twenty seconds, and as each site is rated, the icons in the annotated search result are updated. This approach allows us to offer real-time coverage of the entire World Wide Web without previously examining every site in existence.

The example shown above is from our live system, at “www.sitetruth.com”. We also offer a version of the Open Directory of web sites in which each commercial site has been rated by our criteria, with the higher-rated sites appearing first and the lower-rated sites far down the list.

Table 1. Rating symbols

Symbol	Interpretation
✓	A trusted third party has verified the legal existence of the web site operator.
?	The web site contains information identifying the owner, but that information has not been verified.
X	The ownership of the web site cannot be determined.
●	A non-commercial web site, one not engaged in commerce.
◌	Web site evaluation is in progress. Appears briefly during rating.

This simple system minimizes user confusion. We rate only sites sites engaged in commerce, and thus the gray circle indicates a neutral ranking. Some commercial sites which are not obviously selling online also receive the neutral ranking.

4. EFFECTS ON INTERNET COMMERCE

As a test case, we use a set of sites drawn from the 112,000+ sites in the Open Directory. Our current experience with this test set is that about 60% of business web sites get a green check mark, about 10% get a yellow question mark, and the rest get a red X. About half of those red Xs are justified; the business is not identifying itself. Thus, we are currently classifying about 85% of business sites correctly, and are working to increase this percentage. Misclassification usually occurs because the site does give a business name and address, but in a form we did not recognize. Generally, we will pick up any valid mailing address acceptable to the United States Postal Service, including addresses outside the United States. We attempt to validate such unverified addresses against incorporation records and business directories. Our coverage for the English-speaking world is reasonably good. To assist sites in improving their ratings, we offer “web master tools” which allow site operators to easily find out why their site did not receive a high rating. Usually, a few simple changes to the site will increase the rating of a valid business. Less valid businesses may have problems. Which, of course, is the point. Our goal is to have a system which cannot be “gamed” without committing a felony.

We are trying to avoid the mistake made with “high assurance” certificates – inadequate provision for small business. We thus are working on arrangements with off-site payment service providers through which they identify their merchants on “checkout” pages, and we then accept that as valid. We will do this for any payment provider which will execute a Relying Party Agreement under which they take some modest financial responsibility for validating the identity of the merchant. We also make some special arrangements for major auction sites with established reputation systems.

Simply validating the identity of a business is a strong initial step towards improving the integrity of Internet search. This is sufficient to reject a huge number of marginal web sites and businesses as being unidentifiable. Beyond this, once the business behind a site has been identified, it can be evaluated as a business, using the usual financial rating services. That is beyond the subject of this brief paper.

Adding business rating to a major search system will affect the business ecosystem of the web. There will be complaints. It will cause a major shakeup of the “search engine optimization” industry. We will thus deploy and publicize the web-based tool for checking integrity data well in advance of using that data to affect search results. After a few months of minor turmoil, the situation should settle, with higher standards for business identification on the Web.

5. CONCLUSIONS

It appears to be possible to make a significant improvement in the quality of search results by using easily obtainable extrinsic information to validate the legitimacy of businesses. This rewards legitimate businesses for complying with laws and rules regarding disclosure. Users of the search engine are then less likely to be led to disreputable companies. The result should be a more satisfying experience for web users.

6. REFERENCES

[1] California Business and Professions Code, section 17538.

[2] CA/Browser Forum (http://www.cabforum.org/) *Overview of “High Assurance” SSL certificates.*

[3] Tatsuhiko Kagehiro, Masashi Koga, Hiroshi Sako, Hiromichi Fujisawa, "Address-Block Extraction by Bayesian Rule," pp. 582-585, 17th International Conference on Pattern Recognition (ICPR'04) - Volume 2, 2004. *Address extraction from unstructured text.*