

Web Site Optimization

By Dmitry Yakovlev,

DataArt (www.dataart.com)



What is site optimization?

Site optimization is a number of actions aimed at enhancing the efficiency of a Web site performance. A Web site's typical functions include business representation and awareness (PR), sales (B2C), and potential customer leads (B2B). Consequently, the measure of the site's effectiveness corresponds to its visibility, sales volume and the number of generated leads respectively.

Often, site optimization is understood as promotion in search engines for a number of keyword phrases. However, simply getting a higher ranking doesn't solve a business task as it's only the first step on the way to a sale. After a user enters the site from a search engine link, you should provide him with a quick access to information sought. Convenient navigation, good content organization, easy search, and other usability issues play an important role. The next step is properly placing contact information. In other words, a Web site should provide several ways to establish a contact with company's representatives and to allow for feedback.

A full-fledged site optimization should include all three components above:

1. Optimization for search engines
2. Effective content organization and clear navigation
3. Intuitive feedback mechanism

Metrics: How to measure the effect of optimization?

Improving content and navigation should be based on some measurable characteristics (metrics) which allow to assess effect of the changes on the final result. For example, if the site's primary function is generating business leads, the main metrics would be the number of site visitors who initiated the contact. Analyzing user navigation patterns helps to clarify the reasons why some of them leave the site quickly. It's clear that only a fraction of the contact forms filled out on the site will result in sales, that's why it's difficult to assess the payback based on efforts spent on the site development. However, in most cases these efforts are worth it. For instance, deploying a system for registration and classification of the filled out contact forms and its integration with a corporate CRM system allowed to make conclusion that the site brings contacts requiring a relatively low-cost presales investment.

Visitor tracking

There are two approaches to tracking site visitors – to use counters or to analyze web-server logs. The first one is easy to implement, provides basic statistics and is acceptable for most of the small sites. The major drawback of this approach is loss of information. The counter requests can be blocked by a browser or a corporate proxy-server. Besides, counters can't log crawlers.

Web-server logs analysis provides more complete and detailed information – errors statistics, traffic, images requests and file downloads; cookies contents, crawlers' activity, etc. Practice shows that the best results are achieved by using the combination of both methods. The counter provides general information in real time while web-server logs analysis tools deliver the detailed statistics.

Analysis and Reporting Tools

One of the popular topics for discussions in SEO forums is the choice of a logs analysis system. There is no single solution as every site is unique and its detailed analysis requires very fine tuning of the tools. However, from time to time new analysis tools appear that are claimed to be perfect. One of the most

effective and flexible analysis tools is an OLAP cube, as well as a simple SQL client application to run an arbitrary query against the logs database. Apart from the option of creating various kinds of reports, the analysis system should be able to build and deliver a number of reports generated automatically on a regular basis (daily, weekly, etc). This ensures the continuity of the key metrics monitoring process.

It is also important to track information on the position of your site in search engines and those of your competitors. The analysis of the competitor's site allows evaluating the level of their expertise in optimization techniques and often brings surprising discoveries and new ideas.

Crawlers Identification Problem

One of the problems rarely mentioned by developers of log-analyzers is identifying a crawler. Search bots (chiefly Google bot and MSN crawler) can generate more than half of the overall traffic. So the problem is overstating the overall statistics. A less obvious, but more important problem is a significant distortion of the navigation paths due to bots' random walking through the site. This makes off-the-shelf log-analyzers useless for the comprehensive analysis of the navigation paths.

Case study

The second part of this article talks about DataArt's expertise in building Web analysis systems by example of the company's site, <http://www.dataart.com>. The site is hosted on a remote server, performs a representative function and works as one of the channels for business lead generation. We describe the information gathering and analysis processes; outline consequent changes that should be undertaken and explain their implementation process.

Statistics Collection

Visitor statistics are gathered with the help of a counter and through the logs analysis. A specially developed application downloads logs from the server on a daily basis via an FTP-protocol, saves them locally, imports into an MSSQL database and runs a post-processing procedure. During the import, the application extracts characteristics of the user agents; splits referrers into domains, paths and query strings; detects the most popular search engines and extract phrases; splits IP addresses into octets and saves sessions IDs into a separate field. After that, an SQL server executes a final data processing procedure to extend the data with client's geographic location and time spent on every page, plus identify crawlers.

Geography

To obtain geographical information from an IP address, one can use an existing database or service, priced from \$50 to \$600 depending on the level of detail. There are also some free databases, providing information only on a country level. Usually, these are reduced versions of complete commercial databases (for example www.MaxMind.com or www.IP2Location.com). The geographical information allows to determine which regions are most interested in your services, so you can make arrangements with local resellers or establish your local representative offices.

Bots Identification

The most reliable method for crawlers' identification is by IP address. This method can be used for the most popular search engines, though their addresses may change occasionally. Unfortunately, such method is not applicable to distributed robots (utilities for sites downloading, personal search systems, pilot bots), which can come to the site from virtually any IP address.

Some crawlers can be detected on the basis of the user agent information if you keep the corresponding list. Some bots disguise themselves as popular browsers (IE, Mozilla) or actually are those browsers (for example, when IE downloads a site to make it available offline). The latter case can be resolved using adaptive methods which analyze the behavior of a remote client. If many pages are requested from single IP in a very short period of time (for example 50 pages in one minute) most probably this is a robot. Apart from

that, you can consider several indirect signs, like an empty referrer field or robots.txt file retrieved. Such a multi-level crawler identification scheme turned out to be highly effective in practice.

Feedback Mechanism

The feedback mechanism is usually implemented by using web forms and placing company e-mails and phone numbers on a contact page. In most cases it's more convenient for visitors to initiate a contact via a feedback form as it provides a number of predefined fields that can be filled out quickly.

Designing an effective feedback form is a rather complicated task which requires a detailed analysis of navigation paths and a lot of experiments with the interface. The number and arrangement of the fields, their color scheme, plus presence of various elements and even the name of the Web form – all affect the number of submitted forms. For example, a gaudy design, an abundant number of links to other site sections, and too many pictures distract the visitor from filling out the form and increase the number of exits from the page. The forms and fields which are too small, or a superfluous number of form fields also reduce the efficiency of the feedback mechanism.

To avoid the distortion of information, such as misprints in e-mails, it's useful to perform data validation on the client's side. Additional data validation on the server side allows to cut off a major amount of spam generated by robots. Make sure that the validation procedure on the server side is not stricter than the one on the client side, otherwise some forms may be lost.

As more users have been installing Windows XP SP2, more new problems with pop-ups and cookies have evolved. The first problem is that pop-up windows are simply blocked by the browser without notifications. The second problem is that the site that relies on cookies will not work for users whose browser doesn't accept cookies. According to our estimates, the number of browsers that don't accept cookie has grown from 2% to 27% within the first quarter of 2005 and has been continuing to grow.

It is useful to attach some technical information to each submitted form. This information includes the visitor's IP, geographic location, a site the visitor came from and a detailed path of his browsing through the site. The navigation path allows to make assumptions about the visitor's intentions and the kind of information about the company he already possesses. The visitor's geographic location helps to decide which company representative should follow up with the contact. Additionally, this technical information allows to filter spam and to save valuable time of the sales manager.

Analysis and reports

Once the logs are transferred to an SQL server and processed they are ready for analysis. Separating log analysis from logs processing gives the analyst more freedom in using the most appropriate tools and provides a possibility of custom tools development. As mentioned before, the most flexible and universal tool is an OLAP cube. OLAP cube can be explored via Excel, a web-interface or third-party components. The advantages of this approach are hierarchical views, multidimensional tables, the ability to execute arbitrary queries, fast data retrieval and a convenient interface. The screenshot below demonstrates how an OLAP cube can be used to deduce geographical distribution of traffic generated by search engines for the last three months. Data mining took less than two minutes.

Web Logs (OLAP) Modify Shared Page ▼

www.DataArt.Com Logs (PivotTable)

Is Bot (Blank)

				Country ▼			
				Russian Federation	United Kingdom	United States	Grand Total
Year ▼	Quarter	Month	Search Engine ▼	Hits	Hits	Hits	Hits
2004			Google	74.31%	55.64%	56.96%	58.04%
2005	Quarter 1	January	MSN	46.67%	96.23%	77.27%	78.21%
			Rambler	21.48%	0.94%	11.49%	10.50%
			Yahoo	7.41%		0.06%	0.54%
			Yandex	3.70%	2.83%	11.17%	9.36%
			Total	20.74%			1.39%
			Total	100.00%	100.00%	100.00%	100.00%
		February		43.27%	42.41%	44.73%	44.33%
		March		26.46%	35.55%	35.87%	35.40%
		Total		100.00%	100.00%	100.00%	100.00%
Grand Total				100.00%	100.00%	100.00%	100.00%

Apart from analytical reports that require active participation of an analyst, there is a number of standard reports that can be generated automatically, such as daily reports on the number of visitors, traffic generated from target search engines, number of page loads, number of initiated contacts, etc. The average and total time spent on a particular page by a visitor are very useful metrics demonstrating the interest in a particular part of the site. Pages which trigger extended periods of visitors' time and attract many visitors constitute so-called "zones of interest" and should be designed with great care. Pages with a smaller number of visitors but longer periods of time spent on them often contribute more to the overall time spent on the site than the pages with a higher number of hits.

An important component of any large site is a search subsystem that allows a visitor to find information by-passing the main navigation. Actually, the search subsystem extends the site navigation and therefore must deliver highly relevant results. One of the most effective implementations of the search system is queering GoogleAPI service. When using this approach you have to bear in mind that your site should be regularly indexed by GoogleBot. Otherwise, the results will correspond to an outdated content. For a site with such a search subsystem, one of the most important metrics is a number of pages retrieved by GoogleBot. Knowing what visitors search over the site helps to optimize the content structure and the navigation system by making the most popular pages easily accessible.

Standard reports are easily generated by running one or several SQL queries. There is a number of software products that can be used for generating these reports in a variety of formats (Crystal Reports and MSSQL Reporting Services, etc.). DataArt uses MS SQL Reporting Services as it provides an integrated development environment, web-interface access and a scheduled delivery of reports to specified e-mail addresses. Employees responsible for running the web site start their daily duties with reviewing reports for the previous day. These reports usually include:

1. Total number of visitors (number of sessions, distinct IP addresses, pages retrieved by users and crawlers).
2. A list of the most popular pages
3. Full paths of visitors who searched the site
4. Full paths of visitors from particular countries that visited pages with web forms
5. Referrers from search engines and phrases they searched for

Metrics and Organization of Site Optimization Activities

One of the primary functions of the site is generating business leads. Consequently, the most appropriate metrics is the number of visitors who initiated a contact. In practice, such approach can't be used directly as a significant part of submitted forms is spam. The problem is solved by marking up the forms as either useful or spam. The number of reasonable requests is considered as a measure of the site performance. Another measure is the number of visitors from a particular geographical location directed by a search engine query based on target keywords or phrases. The correlation between the number of visitors from a target group and the number of initiated contacts is also a good measure which provides information for usability improvement.

The second important function of the site is a proper business presentation, where the site serves as an online "business card". The measure of the site's popularity is the total number of site visitors. However, you can distinguish narrower groups by geographical regions. There is an opinion that you don't need to take into account "accidental" visitors (those who don't become clients). It's not true, as the total amount of visitors increases the overall brand recognition. One of the techniques to attract more site visitors is tweaking content so that it appeals to a wider audience, and promoting it in search engines for popular keywords. If the target audience and the keywords were determined correctly, the site will receive a significant flow of visitors. Any relevant content is useful – photographs, whitepapers, images, reference data, clipart, etc. It's preferable that the content is original and is created by the company's employees or business consultants.

Here is a list of metrics we keep an eye on:

1. Number of initiated contacts
2. Number and volume of initiated sales
3. Number of visitors on the pages with Web forms
4. Number of target visitors
5. Correlation between the number of target visitors and the number of filled forms
6. Total number of visitors
7. Site placement in search engine results
8. Number of pages requested by crawlers of major search engines

These metrics are calculated over short and long-term periods of time to evaluate two kinds of changes: local, caused by a particular content alteration, and global, caused by a more profound phenomena. In addition, we track and analyze positions of competitors' sites in the most popular search engines.

The effects caused by these changes might be revealed with a significant time delay. This is especially true for search engine optimization. Moreover, these effects are often unpredictable. That's why even small changes to the site should be documented. For example, changing the color of a small banner can drastically affect the number of clicks. In another example a single phrase located at the top of the page can influence the position of the site in search engine results by specific keywords and thus attract a larger targeted traffic.

Management issues can affect the decision making process if the site is maintained by several people. Our practice shows that discussing the site changes in large groups delays the final decision. That's why it's more reasonable to have decisions made in small groups of two to four people based on site statistics versus personal opinions.

Conclusion

The successful site development requires coordinated efforts of the whole team to make it not simply good, but the best in the category. Designers, authors, developers, analysts, managers – all contribute to making the site better.