STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVII**, Number 3, 2012

ANALYZING AND TUNING USER QUERIES TO SEARCH ENGINES

DARIUS BUFNEA

Abstract. There are certain situations when a web site's visitor using a search engine as a referrer will not be correctly redirected to the desired product or information page, although such a page exists within the web site. This paper presents a solution for a web site to locally further analyze and tune user queries to search engines in order to lead the user to the page describing the product he is interested in. This can improve the visibility of products within the website and in the same time increase its conversion rate. Misdirecting the user and failing in satisfying his interest will be reflected in the revenue amount of a website; on the contrary, satisfied visitors can become potential clients and on a long time scale can improve the website's ranking within a search engine.

1. Introduction

As web technologies developed and their usage on business oriented application as e-commerce have increased, the research community and in the same time private enterprise studies have led to research for solutions to maximize the impact of the web interaction with an online consumer. The impact of the web interaction with an online consumer can be quantified by using certain variables such as: time spent by the visitor within the website, number of products description visualized, conversion rate, number of ordered products or number of displayed commercial ads.

In this paper we present a solution for analyzing and tuning user queries to search engines in logic implemented at web application's tier.

Web sites and web applications in Internet depend on a large scale on visitors sent by search engines. For a regular web site, the percent of these visitors can be as high as 70 to 80 percent from the total number of visitors. Our presented method enhances a search engine's effort to redirect the

Received by the editors: September 17, 2012.

2010 *Mathematics Subject Classification.* 68U35, 68M11.

1998 *CR Categories and Descriptors.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*.

Key words and phrases. Web Referrer, Search Query, User Experience.

visitor to the desired page, i.e. the page that contains information about a product searched by the visitor. This method should not be confounded with Search Engine Optimization (SEO) techniques. Our method rather complements these approaches. There are certain situations when a search engine might fail in redirecting the user to the correct web page, even if such a page exists within the website. A missed search redirected user may leave the website and give up in further searching for the product, even if the web site is offering a search feature. The short term impact is the loss of a possible customer. On a long time run, such a visitor repeated behavior may lead to web site's position demotion in a search engine's results for certain keywords. It is well known that Google for example permanently adjusts their order of return search results based on previous users' experience and feedback.

2. Previous Work

From the mid 90s as web usage increased, numerous different techniques were proposed in order to maximize the interaction between a web site and an online user visiting the site. These techniques are based on different approaches such as data mining, pattern discovery or artificial intelligence. All these techniques aim the same goal: getting the user more closely to the product or information he is interested in. Some studies analyze market basket data ([1], [2]), an approach useful for determining and generating relevant web content for similar consumer-desired products based on previous collected data. Another direction of research ([3], [4]) is pattern discovery across a user's visit path within a website, approach that can be used to create dynamically adaptive web sites based on users' online behavior patterns.

In order to maximize the interaction with its visitors, a website with a relative constant audience can rely on user based personalization techniques via cookies or server side saved configuration settings.

More recently, semantic web approaches, although designed to increase interaction between websites and web applications, can also improve a product visibility within a website for a human person trying to reach that product's specification page. For example, GoodRelations [5] might help a referrer in locating pages of other web sites describing similar products.

3. Contributions

In this paper we focus on those situations when the visitor is sent to the web site from a search engine, i.e. the search engine is a referrer for the web site. There are situations when a search engine will fail to redirect visitors to the most search query related page within the web site. A typical scenario is the visitor landing on the root page of the web site, simply because news about

a product he's interested in is posted on that page. It's very common for web site's root page to have a higher page rank than an in-site page containing a product's description.

Our idea is based on the assumption that business logic running at the web server can extract from the user's query more semantic information than the search engine does. This semantic information is used to redirect user to a more search query related page than the one user has landed on. This approach is similar in many aspects with Search Engine Optimization (SEO) techniques, but while SEO is used to improve a web site visibility in search engines, our method is used to improve a product visibility within the website itself.

3.1. **Technical Fundamentals.** A significant number of requests for resources sent to a web server are accompanied by a `referer`[1] HTTP header defined in RFC 2616. In fact, all requests, not directly typed in a web browser by the user, are accompanied by such a header. This header is added by the web browser to requests made by following links from one page to another, even if pages are hosted on different domains. The header is also present in requests for resources hosted on the same domain as the referrer, for e.g. in request for images needed for correctly rendering a web page.

The `referer` HTTP header is also present in requests for web pages made via a search engine. Although, search engines typically are not hot linking to other websites - rather they redirect the visitor using HTTP redirects - the URL of the search page within the search engine is preserved as a referrer in the request made for the visited landing page (i.e. the web pages where the user is redirected by the search engine). The referrer URL within the search engine always includes the user's search query as a value to a specific search engine dependent attribute. For example, a typical `referer` header for a request made via Google will include a `q` attribute whose value is set to the user's query.

With the knowledge about the user's query, we can locally make better assumption about its semantic than a search engine does. The query might be used to identify if the search engine correctly redirects the user to the most query related page and, if this is not the case, locally redirect him to a more content related page to his query.

3.2. **The search query module.** For a better understanding of our approach we'll give some real world examples about the inappropriate behavior of a search engine in our context.

[1]The name of the header, i.e. `referer` is deliberately misspelled in the RFC 2616 because of historical considerations. Correct spelling should be referrer.

a) A news site containing on its main page the latest news is indexed by a search engine. Subsequently in time, a visitor is landing on the news site's main page from the search engine, his query being related to news posted on the main page at crawling time. If the news that might catch the visitor's attention is at visiting time demoted from the main page, the visitor will not find it and might bounce and go on to the next site within the search result.

b) Someone is looking for flights from London to New York using a search engine. The search engine can easily redirect a user to a page on the flight operator's web site containing the schedule of all departures from New York to London. Even if the flight operator's site is offering a very accessible search form or a "return flights" button, some users will abandon their search experience (i.e. visit) to the flight operator's web site, simply because they don't see a departure hour from London to New York. Instead, such a user will prefer to jump to the next site in the search result, often a competitor's web site.

c) Some person is looking for a notebook with a built in 3G wireless modem. Once again, a search engine can easily redirect the user to an online store's web site offering for sale a notebook with wireless 801.11 connectivity (i.e. Wi-Fi), a classical dial up PSTN modem and 3 GB of RAM. The users will immediately catch the missed landing page, and might prefer visiting the next site within the search result, even if the above online store is really offering a product the user is interested in.

In all these scenarios a web module run by a web site can make new assumptions about the user's query semantic, further refining the search query.

The main task of this module should be:

- Identify from the referrer request header (in case there is one) if the visit is coming from a search engine, and if this is the case what is the user's query value;
- Parse the user's query and match it to products stored in the database backend server. The matching process should take into consideration different product characteristics such as product name, manufacturer, description, specifications and so on. In the matching process different similarity matching algorithms should be tested. Although we implement such an algorithm in our test module (discussed in different writing up paper) future research is to be done. This algorithm is based on the vector space model and is used to map user's queries to product's attributes within the backend database server. Speed and accuracy of the matching process are critical for our technique's success.
- Redirect the user to the match product's description page through a transparent mechanism. Such a redirection can be implemented using HTTP 3xx headers or by using a client side script to perform the jump

from the missed landing page to the desire one. Preliminary researches we have performed are in favor of the client side scripting approach, mainly in order to avoid bounce rate increasing.

The search query module should be run for every page request coming from a search engine result index. Some solutions for running this module are either call it from within a master page or implement it as a server side technology independent filter that intercept the user's request before it reaches the final resource (i.e. the missed landing page).

For common, very often used queries, this module can also cache the match product in order to increase and improve the web page serving speed. Other important aspects that should be taken into consideration by such a module are: words order in user's query, frequently typos or words abbreviation often used by visitors in their queries.

3.3. **Experiment.** In order to assert the method presented in this paper, we made an experiment over a commercial in production web server. This enterprise web server was also used for implementing and testing our method. Because of confidential agreements we cannot disclosure the name of the web server, but we have the approval for using any obtain results for research purpose in scientific papers.

The obtained results support the idea presented in this paper, but actual values are extremely site depended. They were obtained for a very search engine friendly web site (i.e. a web site that presents its content semantic in an easy understandable way by the search engine). Actual percent of these values might be different from site to site being affected by some parameters such as: site niche, the degree of site search engine optimization and advertising campaigns. The experiment was conducted over one month by a web application server module, but the measurements were also confirmed using standard tools such as Google Analytics [7].

In this experiment we were interested in:

- Percent of visitors that have landed on our web site via a search engine using a query string that can be matched to a product. For previous given examples, a product can be a notebook, a flight from London to New York or a news title. The matching process is important in order to further determine if the user's landing page is a good one, i.e. it describes the product the user is interested in, or not.
- Percent of web visitors that have landed via a search engine on a wrong page (i.e. page not describing the product they are interested in);
- Percent of visitors that have landed on a wrong page and they subsequently used website's available search form or navigate further to find information about their product of interest;

– Percent of visitors that have landed on a wrong page and they subsequently abandoned their visit on the web site (i.e. they bounce).

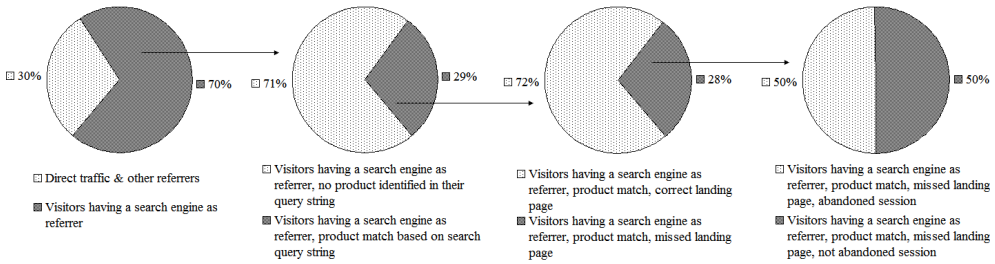

FIGURE 1. Visits, product match and users' behavior for a test web site

We are especially interested in the percent of visitors that have landed on a wrong page and they subsequently abandoned their visit on the web site, those visitors becoming potentially lost customers. Our method increases product match ratio based on user's query and subsequently increases conversion rate by correctly redirecting user's browser to the desired page.

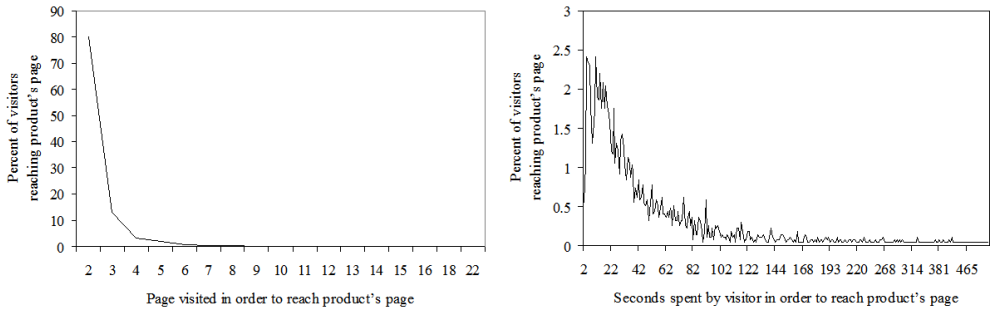

FIGURE 2. Page visited in order to reach products page (left) Seconds spent by visitor in order to reach products page (right)

Furthermore, we are also targeting visitors that have landed on a wrong page and subsequently they are continuing their session looking for information about a specific product. Figure 2 depicts this behavior, showing the number of visited pages and the time spent in searching for a specific product by percent of visitors - percent from the total amount of missed landing users

that are not abandoning their session. Our method helps in reducing user's effort in reaching product information by decreasing the total number of pages that he has to visits or the amount of time he spent searching that information.

	Module off	Module on
Bounce rate	44.74%	41.75%
Visits coming from search engines	70.22%	70.47%
Visitis having a search engine as referrer and a product was identified in user's query string	20.37%	20.43%
Visits having correct landing pages	14.66%	20.43%
Visits having missed landing pages	5.71%	-
Visitors that have landed on a wrong page and they subsequently continued visiting our website	2.86%	-
Bounce visitors because of missed landing pages	2.85%	-
Conversion rate	2.23%	2.64%

FIGURE 3. Web site behavior with our module set to off / on

Figure 3 presents the results obtained with our module set to off for one month and to on for another month. When set to off, in fact it was running in a special state performing no redirects, but only measurements and collecting the above statistical data. When set to on, in case of an initially missed landing page, it consequently redirects the user to the page describing the product identified in user's search query. A major benefit of running the proposed module is reducing of the total bounce rate. This is done by eliminating visitors that bounce because they land on wrong pages. In fact, even if in a missed landing pages situation, each session will count at least two requests, one for the initially missed lading page and one for the page describing the product visitor is interested in. By leading visitors more precisely by their interest, an increase in the conversion rate can also be observed, the potentially lost customers being transformed in potentially gain customers.

Note: All percent values in figure 3 are relative to the total amount of site traffic. We counted only visitors using key words that we successfully match against a product from the web site's offer. We were not interested on visitors using generic key words such as company's name or web site's name as the search query.

4. Conclusions and future work

In this paper we have advanced a method for analyzing and tuning user queries to search engines. Our method helps visitors in reaching the product

information they are interested in and in the same time it makes a product more visible within the website. We are currently focusing our efforts in evaluating different similarity algorithms based on the vector space model in order to improve the matching process between search queries and products description. The speed of the matching process is extremely important, because any delay in serving the desired page to a client may be reflected in the Page Rank of a website, site speed being one of the criteria that Google uses in computing a website Page Rank.

References

[1] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, Shalom Tsur, *Dynamic itemset counting and implication rules for market basket data*, in Proceedings of the 1997 ACM SIGMOD international conference on Management of data 1997 (SIGMOD '97), New York, NY, USA, pp. 255-264.

[2] Alexandros Nanopoulos, Yannis Manolopoulos, *Efficient similarity search for market basket data*, The VLDB Journal 11, 2 (October 2002), pp. 138-152.

[3] Alina Campan, Darius Bufnea, *Automatic Support for Improving Interaction with a Web Site*, in Studia Universitatis Babeş-Bolyai, Informatica, Vol. XLV(2), pp. 95-103, 2000.

[4] Poonam Goyal, Navneet Goyal, Ashish Gupta, T. S. Rahul, *Designing self-adaptive websites using online hotlink assignment algorithm*, in Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia 2009 (MoMM '09), ACM, New York, NY, USA, pp. 579-583.

[5] Martin Hepp, *GoodRelations: An Ontology for Describing Products and Services Offers on the Web*, in Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008), Acitrezza, Italy, September 29 - October 3, 2008, Springer LNCS, Vol. 5268, pp. 332-347.

[6] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee, *Hypertext Transfer Protocol – HTTP/1.1*, RFC 2616, June 1999.

[7] *Google Analytics*, Google Inc., <http://www.google.com/analytics/>.

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogălniceanu St., 400084 Cluj-Napoca, Romania

E-mail address: bufny@cs.ubbcluj.ro