Electronic Commerce On-Site Search Services: A State of the Art Review

Nan Li[1], Peter Hitchcock[1], James Blustein[1], and Michael Bliemel[2]

[1] Faculty of Computer Science, Dalhousie University,
6050 University Avenue, Halifax Nova Scotia, B3H 1W5, Canada
[2] Faculty of Management, Dalhousie University,
6100 University Avenue, Halifax Nova Scotia, B3H 3J5, Canada
nan_ls@hotmail.com, {hitchcp,Jamie}@cs.dal.ca, m.bliemel@dal.ca

Abstract. Many Electronic Commerce websites have vast product catalogues, which require visitors to use an on-site search function to find and consequently purchase the product they desire. This paper illustrates the importance of successful on-site searches, the main Key Performance Indicators (KPI) for on-site searches, and introduces several popular on-site search algorithms and techniques.

Keywords: On-Site Search, Key Performance Indicators, E-Commerce.

1 Introduction

Over the last decade search engines continue to improve as they grow in popularity. With the rise of powerful search engines like Google, Yahoo! and Bing, why would we still need search services within a website? Why not let the search engines do all the search-related work? Here are the reasons:

A webpage may not be visible in the search result listings. Visitors tend to have little patience. People typically provide two words per search [1] and they do not often look past the first ten results returned by the search engine. Research shows that if a visitor cannot find what they are looking for within three clicks, the chance of the visitor leaving becomes much higher [2]. For a website owner, it does not make sense to let visitors leave the website and visit another link from the search engine's list of results if the item is available from your website. If allowed to leave, the chance that the visitor will return is very low. There are at least 50.6 billion web pages [3] but search engines usually display 10 search results a time. The majority of web pages are therefore invisible when search engines are the primary method of searching.

It is well known that search engines do not crawl web pages on a daily basis, so there are many pages that may have changed since they were last indexed. If search engines are the only method that visitors can use to browse a webpage, they can be mislead by out-of-date or incomplete information.

People are becoming accustomed to searching by inputting a search keyword rather than searching by category. Even if one website has its inventory in well organized categories, visitors may still want to search based on keywords that are in their minds,

R. Sharman, T.S. Raghu, and H.R. Rao (Eds.): WEB 2009, LNBIP 52, pp. 226–234, 2010.

instead of narrowing down their search category by category. Providing both a category and text search box on an e-commerce website has become a basic feature.

Almost all professional e-commerce websites now use a text search box and the inclusion of such a text search box, useful or not, is now the norm. By providing a search box search service, it becomes much easier to log what visitors are searching for on the website. By studying those search keywords, we can gain insight into what visitors are expecting from the website. Based on this, webpage content can be adjusted to better fit the visitors' expectations.

2 Background

Every e-commerce website has one common goal, which is to sell products or services to its visitors. No matter what the product or service is that is being sold, all visitors experience the same searching process as illustrated below in Figure 1, which is subsequently explained in detail.

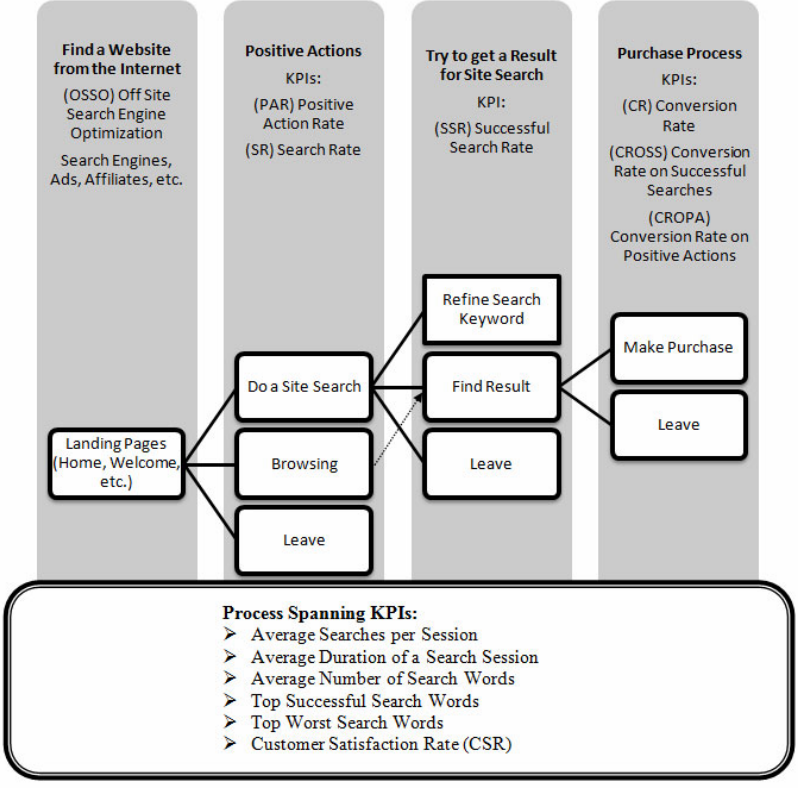

Fig. 1. E-Commerce Website Usage and KPIs from a Visitor's Point of View

The overall process is composed of two major parts: *off-site search optimization* (*OSSO*) and on-site search. For the site owner, the purpose of OSSO is to bring visitors to the website. On-site searches are intended to convert those visitors into customers. This paper concentrates on exploring on-site search approaches and performance measurement metrics.

2.1 Measuring Performance

Key Performance Indicators (*KPI*) are a set of measures or metrics used to define and evaluate certain aspects of e-commerce websites. There are several KPI to help gauge on-site search services. Some KPI are used to analyze how the current on-site search strategies are working, while others are used to design useful on-site search solutions.

Search Rate (*SR*) is the percentage of all the visitors who did at least a search by inputting a search keyword. The formula for SR is:

$$\text{SR} = \text{Count of unique visitors who performed at least one search} / \text{Count of unique visitors} . \quad (1)$$

For the websites which expect to get a high SR, a low SR may indicate several web design problems, such as the search box not being visible enough to the visitors or page content not encouraging visitors to use the search box.

Positive Action Rate (*PAR*) is a measure of how many visitors perform a positive action such as browsing or searching the site instead of leaving immediately after entering the website. The formula for PAR is:

$$\text{PAR} = (\text{count of unique visitors} - \text{count of visitors who leave the website without performing any action}) / \text{count of unique visitors}) . \quad (2)$$

For example, if there were a total of 1,000 unique visitors and 500 of them left the site without doing anything, the PAR would be equal to (1000 – 500) / 1000 = 50%.

Low PAR can be indicative of several types of website issues; one of which may be that the website draws too much inappropriate traffic. Such traffic can be directly caused by a poor marketing strategy. Another shortcoming may be poor website content. The content may be too confusing for the visitors and leave them in a state where they do not know what to do next. The primary use of PAR is to analyze the off-site marketing strategy and landing page performance.

Once visitors have been attracted to a site by OSSO and stay at the site, the next step towards success is to have an intra-site search that will aid them to efficiently find products or services that they will purchase.

Successful Search Rate (*SSR*) is a very important KPI for on-site search. There are a couple of different definitions for SSR. Some website-owners calculate it based on unique visitors who performed searches. It shows the rate of how many visitors actually get a search result page. That formula is:

$$\text{SSR} = \text{Count of unique visitors who get at least one search result} / \text{Count of unique visitors} . \quad (3)$$

For example, assume that there are total of 1,000 unique visitors and 800 of the 1000 did 1,200 searches. Now assume that only 700 of them have ever seen the search result page. The SSR = 700/1000 = 70%.

Another definition of SSR is the rate for all of the search sessions that receive a search result page. That formula is:

$$\text{SSR} = \text{Count of search sessions that get a search result} / \text{Count of search sessions} . \quad (4)$$

As an example, if there are 1,000 unique visitors that make 1,500 searches and 1,125 of the search sessions return a search result, then the SSR is 1125/1500 = 75%.

To simplify the problem and focus on search only, we calculate the SSR according to the second formula. Therefore, we calculate SSR based on search sessions rather than visitors. SSR is the health meter of the current searching strategies. It directly affects the customer satisfaction rate and the very important KPI final conversion rate. Mondosoft [4] reported, "Among the 400 sites in our study, an average of 22% of all searches returned no results."

2.2 Purchase Related KPIs

There are some other website KPI that may also reflect how the on-site search solution is faring, or may detail the effect of the solution on on-site search performance.

Conversion Rate (*CR*) is the percentage of the all the visitors who actually buy something. For example, if the total number of unique visitors is 1,000 and 40 of them finally make a purchase, then the final CR = 40/1000 = 4%. Getting a high CR is the goal for all e-commerce websites. One of the motivations to increase SSR is to produce a higher CR. Aside from SSR, several other factors can affect a visitor's purchase decision. Examples of these alternative factors include price, shipping charges, and availability. A high conversion rate indicates a good SSR, but a poor conversion rate does not necessarily reflect a low SSR.

Conversion Rate on Successful Searches (*CROSS*) is the KPI that shows how many of the visitors who get at least one search result actually makes a purchase. If the total number of unique visitors is 1,000 and 400 of them gets at least one search result then forty of them actually purchase goods. The CROSS = 40 / 400 = 10%. This KPI focuses on the visitors who performed at least one successful search. A low CROSS indicates that there may be some pricing, shipping or availability issues that are stopping the visitors from buying the product after finding it through the search function.

Conversion Rate on Positive Actions (*CROPA*) shows how many visitors who performed a positive action actually buy something. If the total number of unique visitors is 1,000 and 500 of them performed a positive action and forty of them actually buy something, then CROPA = 40 / 500 = 8%. Some site owners prefer this KPI since it focuses attention on the performance of on-site features only by filtering out inappropriate traffic.

2.3 Spanning the Entire Process

Average Searches Per Session (*ASPS*) refers to the number of searches a typical visitor will usually do on a website. This KPI relates to visitor search behavior. Based on

the Mondosoft website search survey [4], the average searches per session is between two and four. That research indicates that visitors are likely to refine their search keywords four times during a search process, at the most. If the visitor still cannot find a positive search result within the maximum four searches, the chance that the visitor will leave the website is high.

All website owners must study this KPI and learn the patience limits of visitors in order to survive. This KPI can also be a reference when filtering out suspected unwanted visitors. If one visitor searches too many unusual terms or too many times, it may be an indication that the visitor may be software and not really a potential customer.

Average Duration of a Search Session measures how much time users spend on each search, on average. A search session represents the entire process of a visitor trying to find a product by searching using a keyword. A visitor may have many search sessions. This KPI shows the time between search sessions. By studying this KPI, we can get an idea if the content on the search result page is appropriate. For example, if most of the search result pages provide many search results, you may expect people to finish the reading the entire list within 2–3 minutes. However, if the KPI shows that duration of the search is only 1 minute, then it either means that the visitor did not find what they were looking or they simply gave up reading. The challenge is how to get the right number for this KPI. One must consider that visitors may do something not related to search between search sessions, such as loading another static information page between searches on the same site or grabbing a cup of coffee.

Average Number of Search Words answers the question "Are most of the visitors searching using multiple words or single words?" This KPI is very important in determining the right search strategies. Searches that use multiple words strongly suggest a search approach that understands natural language.

Top Successful Search Words is the certain number of search keywords that visitors used most frequently to search in the website to get a search result. This measure indicates what most of the visitors are expecting from the website. Those search words may present some products or certain information. With the help of this KPI, a website owner will know what products or information should be promoted throughout the website. This KPI also indicates whether the searching strategy applied to this kind of search is working properly.

Top Worst Search Words are the most commonly used search terms that do not match anything in the index. The inability to get results from a search results in a poor situation for both the visitor and the website. However, the top "worst" search words can provide very important and valuable information. If a search keyword constantly returns no results but different visitors keep using it in their searches, it may indicate several things:

- The keyword may refer to a product that visitors are expecting to be on this website and considerations regarding adding the item to the website's inventory should be made.
- Visitors may be searching for an item that is in the inventory but the visitors are not using the correct term that refers to the item. It shows a need for a change to the site's searching strategy as changing a visitor's search habits is an unreasonable solution.

Customer Satisfaction Rate (*CSR*) indicates if customers are satisfied with the website and how much they appreciate the services that are provided by the website. CSR is difficult to measure. Some websites provide a survey, asking customers if they were satisfied with the services offered and if not, what factors caused the dissatisfaction. By analyzing the survey data, website owners can get some measure of the CSR. However, for those who do not have, or have but do not analyze, this kind of survey, it may be difficult to know customer satisfaction levels. Some other KPI can help in this regard. SSR in particular has a strong affect on CSR. A low SSR leads to low CSR —no one appreciates a poor search service.

3 On Site Search Approaches

In this part, we introduce major text search approaches. Website owners should pick the right searching strategy based on the websites' searching requirements and data characteristics. For most websites, one searching strategy will not fulfill all of the searching requirements. Implementing and applying two or even more searching strategies on one website may be needed.

Exact Search is the strategy that returns exact matches for the searched keyword only. In SQL, the syntax for exact search is select * from table where column = 'keyword'. This gives visitors the most accurate results. There is no additional technique challenge in this kind of search except improving searching speed. Creating proper indexes for the data will improve searching speed. Exact search is the best solution when visitors know precisely what they are searching for, and under what name it is indexed. Unfortunately, the system cannot find exact matches to visitors' search text most of the time and will return no search results if no other search strategies are used to follow up.

Partial Search is also known as wildcard or pattern search. Based on the wildcard's position, it can return different results. In SQL, the syntax for partial searches can be one of the following:

```
select * from table where column like 'keyword%'
select * from table where column like '%keyword'
select * from table where column like '%keyword%'
```

This type of search returns the results that contain the search keyword if the keyword shows up in any part of the searched data.

The wildcard can be a single character, such as "?". If searching for "n?t" the results will include words such as "not" and "net". The wildcard can be multiple characters, using "*". For example, a search for "scien*", the results can include "science" and "scientific" [5]. A leading wildcard in the search terms may cost more time than a term that has a wildcard in the end. The reason is that the searcher cannot ignore any section for the leading wildcard search, and it must go over the entire index until it finds the entries for the term. On the other hand, if we know what the results start with, we can jump to matching sections directly. Proper indexing can increase partial search speed as well.

Both of these SQL-based approaches require care to implement. The most straightforward approaches leave gaping security holes [6]. Other approaches, which are less risky and more forgiving of users' mistakes, depend on comparing relaxed versions of the user's search terms against a similarly relaxed version of the database index.

Phonetic Search is defined as strategic indexing and searching through words based on their pronunciation. Most of the phonetic algorithms are restricted to words in English. If those algorithms are applied to a non-English language, the results will surely be a mess [7, 8]. The following are some techniques for phonetic search.

The most common phonetic search algorithms are Soundex [8] (which was patented over 70 years ago) and some of its variants: Metaphone [9], Double Metaphone [10], Phonix [11], and Editex [8]. In brief, they all work by creating indexes in which codes represent letters and combinations of letters (such as "ph"). In this way, many words with similar pronunciations are identical in the index. The entries in the index point to the actual words.

Grammatical Searches are algorithms that try to match two strings using their grammatical characteristics. Like phonetic searching, they are also specific to a single natural language, e.g. English. Different languages require a different design algorithm to fulfill this search strategy. To implement grammatical search algorithms, it may be necessary to work out a library for mapping purposes. Stemming and synonym searches are two of the major grammatical search strategies.

Stemming Search is a strategy that uses the etymological root of the search keyword instead of the search keyword itself to search through the data and return matches that contain the same root. The stemming search matches all different types of a words including its verb, noun and adjective. For example if the keyword is "mice", it will match both "mouse" and "mice". If the keyword is "running", it matches "run", "running", "ran" and "runs". It was first published by Lovins [12]. Since then, several types of algorithms appeared. Among them is an algorithm written by Porter [13] which was widely used. His "Surfix Stripping" algorithm set up a rule that was formed through several steps.

All stemming algorithms are attempts to reach a balance between two types of error: overstemming and understemming. Overstemming means stem two words to the same root that should not. On the other hand, understemming means stem to different roots for two words that should have the same root [14].

Synonym-based search is more complex than stemming but can yield better results. Sometimes synonyms search and stemming search are quite confusing. For example, if the search keyword is "cars", the word "car" will also be in the search result for both algorithms. However, "automobile" will be in the result for synonyms search but not stemming search. Another example is a search of "intelligent". A synonym search might return matches that include the words "smart" and "clever". To implement synonyms search, a set of rules is not enough. We may need another dictionary for those matches that cannot be covered.

Fuzzy Search can be defined to include all of the string similarity matching algorithms that include phonetic and grammatical matching. We however, limit the term to algorithms that focus on correcting the misspelling (typo) problems in the search keyword only. This focus should be a strategy applied only after the failures of phonetic, stemming and synonyms searching or where those three strategies are not applicable. Levenshtein string distance [15] is one of the most well-known fuzzy

search algorithms. It is also referred to as edit distance. It calculates the number of deletions, insertions or substitutions to translate a string (source string) to another string (target string). This calculated number is called the distance between these two strings. If the distance is 0, means the two strings are identical. The larger the number is, the more difference there is between the two strings. This algorithm can be used to return all of the strings that are close to the search keyword. However, since it is based on the string distance only, the majority of the time, it may only return words that are totally unrelated. That is why there are many hybrid solutions that implement string distance and other ranking algorithms to get more accurate results. Editex, referred to above in the section on phonetic algorithms, is one of them.

All of the algorithms described here in section 6 can be implemented by all of the programming languages, such as Java, C++, VB, or SQL server. They are all designed to solve the issue of accuracy in the returned search results to lead to a higher successful search rate (SSR). This in turn can lead to increased customer satisfaction rates with the site (CSR) as well as higher conversion rate on successful searches (CROSS).

4 Recommendations

Many electronic commerce websites continue to frustrate potential customers by providing a search box which often does not provide usable results. These poor results are the responses to queries that are not exact matches to the names of products contained within their databases. Most visitors are users that have become accustomed to finding millions of results with search engines, and often do not realize that on-site search services are far less sophisticated than those of commercial search engines. It is imperative for electronic commerce website operators to improve their visitors' experience with on-site search services if they wish to remain competitive.

In this paper we showed what key performance metrics can be used to assess how well on-site search services are functioning. As product catalogues and customer expectations continue to evolve, it is important to measure and monitor how well on-site search services are performing. We recommend that website operators consider using some or many of the search approaches outlined in this paper to reduce the number of failed searches that visitors are experiencing.

Acknowledgements. The authors wish to thank the anonymous reviewers for their valuable feedback that served to make this a better article, as well the first author's employer who wished to remain anonymous. Publication of this research was also supported through a research grant from the Social Science and Humanities Research Council of Canada (SSHRC).

References

1. Pu, H.-T., Chuang, S.-L., Yang, C.: Subject categorization of query terms for exploring Web users' search interests. J. Am. Soc. Inf. Sci. Technol. 53(8), 617–630 (2002)
2. Hobo. The 3 Click Rule of website design, http://www.hobo-web.co.uk/tips/31.htm

3. Netcraft, Web Server Survey (January 2009), http://news.netcraft.com/archives/2009/01/16/january_2009_web_server_survey.html
4. Mondosoft. Web Site Usability Metrics; Search Behavior and Search Trends (May 2004), http://www.mondosearch.com/upload/pdf/whitepaper/wp-searchbehavior.pdf
5. Microsoft. Wildcard Characters. MSDN Visual Studio Developer Center, http://msdn.microsoft.com/en-us/library/069b0htd(VS.71).aspx
6. Schwartz, R.: Processing Web Forms Carefully. Linux Magazine 4(9), 56–59 (2002)
7. Repici, D.J.: Understanding Classic SoundEx Algorithms (2007), http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm
8. Zobel, J., Dart, P.: Phonetic string matching: lessons from information retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Zurich (1996)
9. Nelson, A.: Implement Phonetic ("Sounds-like") Name Searches with Double Metaphone Part VI: Other Methods & Additional Resources (2007), http://www.codeproject.com/KB/string/dmetaphone6.aspx
10. Atkinson, K.: Lawrence Philips' Metaphone Algorithm (2001), http://aspell.net/metaphone/
11. Zobel, J., Dart, P.: Finding approximate matches in large lexicons. Softw. Pract. Exper. 25(3), 331–345 (1995)
12. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics 11(1&2), 22–31 (1968)
13. Porter, M.F.: An algorithm for suffix stripping, in Readings in information retrieval, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco (1997)
14. Paice, C.D.: Method for evaluation of stemming algorithms based on error counting. Journal of the American Society for Information Science 47(8), 632–649 (1996)
15. Gilleland, M.: Levenshtein Distance. In: Three Flavors, http://www.merriampark.com/ld.htm