Web Search Engine Metrics (Direct Metrics to Measure User Satisfaction)

Ali Dasdan
eBay, Inc.
2145 Hamilton Avenue
San Jose, CA, USA
ali_dasdan@yahoo.com

Kostas Tsioutsiouliklis
Yahoo!, Inc.
701 First Avenue
Sunnyvale, CA, USA
kostas@yahoo-inc.com

Emre Velipasaoglu
Yahoo!, Inc.
701 First Avenue
Sunnyvale, CA, USA
emrev@yahoo-inc.com

ABSTRACT
Search engines are important resources for finding information on the Web. They are also important for publishers and advertisers to present their content to users. Thus, user satisfaction is key and must be quantified. In this tutorial, we give a practical review of web search metrics from a user satisfaction point of view. We cover metrics for relevance, comprehensiveness, coverage, diversity, discovery freshness, content freshness, and presentation. We will also describe how these metrics can be mapped to proxy metrics for the stages of a generic search engine pipeline. The practitioners can apply these metrics readily and the researchers can get motivation for new problems to work on, especially in formalizing and refining metrics.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Performance evaluation (efficiency and effectiveness)*. A.1 [**Introductory and Survey**].

General Terms
Measurement, Performance.

Keywords
Web metrics, web search, relevance metrics, coverage metrics, comprehensiveness metrics, diversity metrics, discovery freshness metrics, content freshness metrics, presentation metrics.

1. OVERVIEW
The modern Web search engines are ecosystems. Indispensible for finding information and distributing content on the Web, they are mostly free to the users and publishers and the cost is usually born by the advertisers and the search engine companies which are looking for economic advantage. Quality is directly linked to the richness and health of the ecosystem

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Search engines are also complex machines where the information is gathered by crawling the Web and consuming feeds, content is indexed, matched to an incoming query expressing the information need, and the results are presented to the user. User's satisfaction of the service can be measured directly or inferred indirectly. Impact of various stages of the machinery on quality can be estimated by using proxy metrics.

In this tutorial, we focus on various aspects of the Web search engine quality assessment for user satisfaction including relevance, coverage, comprehensiveness, diversity, discovery freshness, content freshness, and presentation metrics which we briefly describe below and cover in more detail during the tutorial.

1.1 Relevance
A search result is relevant to the extent it is related to or answers the input query. In Information Retrieval (IR), relevance is typically measured along the following four axes: explicit vs. implicit, absolute vs. preference, binary vs. multi-valued grade, and the underlying user model with its related metrics.

Explicit judgments are provided by paid panelists in a laboratory environment. Implicit judgments are obtained from search session logs of real users. Since a click does not necessarily imply relevance, various approaches have been developed to infer relevance from clicks.

A judgment is absolute when utility of a query-result pair is judged independently of the other search results for the same query. For preference judgments, a panelist is presented with two search results and is asked to decide which one is more relevant.

Historically, relevance has been measured on a binary scale of relevant vs. irrelevant; however, modern search engines use a graded scale such as highly relevant, relevant, somewhat relevant, and irrelevant.

Finally, there are many IR metrics to map relevance to a number for comparison. One of the most commonly used metrics is Discounted Cumulative Gain in which the contribution of a search result increases with its grade but decreases with its position

1.2 Temporal and geographic relevance
In addition to the four axes of traditional IR relevance discussed above, two more axes are interesting for web search: time-sensitive and geography- or locale-sensitive queries. Each can best be described by an example.

An example of time-sensitive query is "WWW conference" since usually the intent is to get the official page of the upcoming WWW conference rather than that of a previous one. For locale-

sensitivity, an example query is about the show times for a movie, where probably the intent is to find that at nearby theaters.

Search results can be quantified for these axes if queries can be classified according to their time- and locale-sensitivity and editors are asked to grade with the corresponding intent in mind.

1.3 Coverage

This implies the presence of a certain set of search results of interest in the search results page. When taken over all users using different languages all over the world and content from the entire Web, coverage refers to the size, comprehensiveness, and diversity of a search engine's index. Since comprehensiveness and diversity also refer to similar metrics for user satisfaction, they will be discussed in separate sections below.

As a simple example, consider a user submitting a vanity query to a search engine, i.e., a query for his or her name. If the same user is expecting to get his or her homepage as a response but finds out that either the homepage is not returned or not ranked high enough, the user will not be satisfied with the search engine.

One way to measure coverage is by samples of URLs of interest.

1.4 Comprehensiveness

A result set returned for an input query can be made up of all relevant results but the set may still be not comprehensive enough. Here, comprehensiveness refers to the usefulness of the search results. In turn, a page can be useful if it is informative and/or novel, i.e., if the user actually learns something about the answer to the input query.

As an example, consider that a user issues the input query "WWW2010" and gets the following pages at different and not necessarily consecutive ranks: the W3C page on WWW conferences, the Wikipedia page on the WWW conferences, and the official WWW2010 page. Although all these pages are relevant, the user can easily find out that they are ordered in the increasing order of comprehensiveness. Note that comprehensiveness also depends on the query because, for example, the Wikipedia page is probably the most comprehensive for the input query "WWW conference".

One way to measure the relative comprehensiveness of a page is the dwell time distribution, where the dwell time is the time spent on the page by users. Pages abandoned too quickly may not present much satisfactory information.

1.5 Diversity

This metric matters especially for ambiguous queries. If an input query has many facets, then its search results page is diverse when it contains results for each facet. The number of search results for each facet may be proportional to the importance of each facet, determined by the amount of content or previous user clicks, or even previous search history of the user.

Even if the input query is unambiguous, diversity is still important because search results can be constructed from a variety of sources and can have a variety of characteristics. For example, if the search results are news or blog entries, then the sources matter as they may reflect different political views. Even for an arbitrary query, it is possible to have content coming from small or large sites, user generated content or content from well-known publishers, etc.

Diversity can be measured in many ways. For the more structural setting for diversity like the site or page characteristics, traditional IR metrics such as recall apply with minimal modifications. The more difficult part is to automatically determine the facets of the input query and measuring if each facet is covered in the search results. When the facets of a query and its results can be identified, most IR metrics can be modified to take into account the information overlap between the results.

1.6 Discovery freshness

This implies the coverage of very new content in the search. This metric is especially effective for news and blog content but can be used for any content that changes fast. Beyond coverage, it can also measure the latency of search engines in acquiring and presenting new content.

Consider the following example. A user just heard some news on TV and wants to get more information. She can either go directly to a news source or search for the information using a search engine. It is typically more convenient to do the latter because search engines not only aggregate news from many sources but they also rank the results to help users focus on the top results. The user will not be satisfied if the search engine does not return results related to the news that she was looking for.

One way to measure this metric is by checking for the coverage of news or blog content by a search engine.

1.7 Content freshness

Discovery freshness is for the content that a search engine is expected to discover. In contrast, content freshness is for the content that a search engine already has. The goal is to see how fresh the content has been kept.

Content freshness for a web page matters for three different reasons. First, it matters for the accessibility status of the page, for example, is the page still accessible or has it become a hard or soft 404 error? Second, it matters for the content of the page; for example, when a search engine cannot find the relevant page for a query because of some new content that has been added recently to the page but not yet been discovered by the search engine. Third, content freshness matters for the links on the page because such links lead to new content to be discovered by the crawler.

One way to evaluate this metric is by measuring the staleness of the content accessed by users.

1.8 Presentation

This metric covers many aspects: query assistance (spelling correction, disambiguation, similar queries, key phrases from web pages, interactivity), search results page layout (organic and sponsored results, navigation features, cached content links), speed (from query submission to the fully loaded search results page), vertical inclusion (news, blogs, photos), structured content inclusion (e.g., using SearchMonkey from Yahoo!), title and abstract for search results, paid-inclusion content in organic results, and finally the number of sponsored results and their placement in the search results page.

2. REFERENCES

See http://dasdan.net/ali/www2010.php