Web Search Relevance Ranking

HUGO ZARAGOZA[1], MARC NAJORK[2]
[1]Yahoo! Research, Barcelona, Spain
[2]Microsoft Research, Mountain View, CA, USA

Synonyms

Ranking; Search ranking; Result ranking

Definition

Web search engines return lists of web pages sorted by the page's relevance to the user query. The problem with web search relevance ranking is to estimate relevance of a page to a query. Nowadays, commercial web-page search engines combine hundreds of features to estimate relevance. The specific features and their mode of combination are kept secret to fight spammers and competitors. Nevertheless, the main types of features at use, as well as the methods for their combination, are publicly known and are the subject of scientific investigation.

Historical Background

Information Retrieval (IR) Systems are the predecessors of Web and search engines. These systems were designed to retrieve documents in curated digital collections such as library abstracts, corporate documents, news, etc. Traditionally, IR relevance ranking algorithms were designed to obtain high recall on medium-sized document collections using long detailed queries. Furthermore, textual documents in these collections had little or no structure or hyperlinks. Web search engines incorporated many of the principles and algorithms of Information Retrieval Systems, but had to adapt and extend them to fit their needs.

Early Web Search engines such as Lycos and AltaVista concentrated on the scalability issues of running web search engines using traditional relevance ranking algorithms. Newer search engines, such as Google, exploited web-specific relevance features such as hyperlinks to obtain significant gains in quality. These measures were partly motivated by research in citation analysis carried out in the bibliometrics field.

Foundations

For most queries, there exist thousands of documents containing some or all of the terms in the query. A search engine needs to rank them in some appropriate way so that the first few results shown to the user will be the ones that are most pertinent to the user's need. The interest of a document with respect to the user query is referred to as "document relevance." this quantity is usually unknown and must be estimated from features of the document, the query, the user history or the web in general. Relevance ranking loosely refers to the different features and algorithms used to estimate the relevance of documents and to sort them appropriately.

The most basic retrieval function would be a Boolean query on the presence or absence of terms in documents. Given a query "word1 word2" the Boolean AND query would return all documents containing the terms word1 and word2 at least once. These documents are referred to as the query's "AND result set" and represent the set of potentially relevant documents; all documents not in this set could be considered irrelevant and ignored. This is usually the first step in web search relevance ranking. It greatly reduces the number of documents to be considered for ranking, but it does not rank the documents in the result set. For this, each document needs to be "scored", that is, the document's relevance needs to be estimated as a function of its relevance features. Contemporary search engines use hundreds of features. These features and their combination are kept secret to fight spam and competitors. Nevertheless, the general classes of employed features are publicly known and are the subject of scientific investigation. The main types of relevance features are described in the remainder of this section, roughly in order of importance. Note that some features are query-dependent and some are not. This is an important distinction because query-independent features are constant with respect to the user query and can be pre-computed off-line. Query-dependent features, on the other hand, need to be computed at search time or cached.

Textual Relevance

Modern web search engines include tens or hundreds of features which measure the textual relevance of a page. The most important of these features are matching functions which determine the term similarity to the query. Some of these matching functions depend only on the frequency of occurrence of query terms; others depend on the page structure, term positions, graphical layout, etc. In order to compare the query and the document, it is necessary to carry out some non-trivial preprocessing steps: tokenization (splitting the string into word units), letter case and spelling normalization, etc. Beyond these standard preprocessing steps, modern web search engines carry out more complex query reformulations which allow them to resolve acronyms, detect phrases, etc.

One of the earliest textual relevance features (earliest both in information retrieval systems and later in commercial web search engines) is the vector space model scoring function. This feature was used by early search engines. Since then other scoring models have been developed in Information Retrieval (e.g., Language Models and Probabilistic Relevance Models) [8] and have probably been adopted by web search engines. Although web search engines do not disclose details about their textual relevance features, it is known that they use a wide variety of them, ranging from simple word counts to complex nonlinear functions of the match frequencies in the document and in the collection.

Furthermore, web search engines make use of the relative and absolute position of the matches in the document. In Information Retrieval publications, there have been many different proposals to make use of term position information, but no consensus has been reached yet on the best way to use it. Most known approaches are based on features of the relative distances of the match terms such as the minimum (or average, or maximum) size of the text span containing all (or some, or most) of the term matches. Web search engines have not disclosed how they use position information.

Besides match position, web search engines exploit the structure or layout of documents, especially HTML documents. There are a number of ways to do this. One of the simplest is to compute textual similarity with respect to each document element (title, subtitles, paragraphs). More complex solutions integrate matches of different structural elements into a single textual relevance score (see for example [8]).

Another type of textual relevance information is provided by the overall document quality. For example, Web search engines use automatic document classifiers to detect specific document genres such as adult content, commercial sites, etc. Specialized techniques are also used to detect spam pages. Pages may be eliminated or demoted depending on the result of these analyses.

Hyperlink Relevance

The web is a hyperlinked collection of documents (unlike most previously existing digital collections, which had only implicit references). A hyperlink links a span of text in the source page (the “anchor text”) to a target page (the “linked page”). Because of this, one can think of a hyperlink as a reference, an endorsement, or a vote by the source page on the target page. Similarly, one can think of the anchor text as a description or an explanation of the endorsement. One of the innovations introduced by Web Search Engines was leveraging the hyperlink-structure of the web for relevance ranking purposes.

The hyperlink graph structure can be used to determine the importance of a page independently on the textual content of the pages. This idea of using web hyperlinks as endorsements was originally proposed by Marchiori [9] and further explored by Kleinberg [6] and Page et al. [11] (who also introduced the idea of using the anchor text describing the hyperlink to augment the target page).

Exploiting User Behavior

As stated above, hyperlink analysis leverages human intelligence, namely peer endorsement between web page authors. Web search engines can also measure users’ endorsements by observing the search result links that are being clicked on. This concept was first proposed by Boyan et al. [2], and subsequently commercialized by DirectHit [3]. For an up-to-date summary of the state of the art, the reader is referred to [5]. Besides search result clicks, commercial search engines can obtain statistics of page visitations (i.e., *popularity*) from browser toolbars, advertising networks or directly from Internet service providers. This form of quality feedback is query-independent and thus less informative but more abundant.

Performance

There are several aspects of the performance of a web relevance ranking algorithm. There are the standard algorithmic performance measures such as speed, disk and memory requirements, etc. Running time efficiency is crucial for web search ranking algorithms, since billions of documents need to be ranked in response to millions of queries per hour. For this reason most features need to be pre-computed off-line and only their combination is computed at query time. Some features may require specialized data structures to be retrieved especially fast at query time. This is the case for example of term-weights (which are organized in inverted indices [1]), or query-dependent hyperlink features [10].

A more fundamental aspect of the performance of a relevance ranking algorithm is its accuracy or precision: how good is the algorithm at estimating the relevance of pages? This is problematic because relevance is a subjective property, and can only be observed experimentally, asking a human subject. Furthermore, the performance of a ranking algorithm will not depend equally on each page: the best ranked pages are those seen by most users and therefore the most important to determine the quality of the algorithm in practice. Performance evaluation measures used for the development of relevance ranking algorithms take this into account [8].

There exist other, less explicit measures of performance. For example, as users interact with a search engine, the distribution of their clicks on the different ranks give an indication of the quality of the ranking (i.e., rankings leading to many clicks on the first results may be desired). However, this information is private to the search engines, and furthermore it is strongly biased by the order of presentation of results.

Feature Combination

All of the features of a page need to be combined to produce a single relevance score. Early web search engines had only a handful of features, and they were combined linearly, manually tuning their relative weights to maximize the performance obtained on a test set of queries. Modern search engines employ hundreds of features and use statistical methods to tune these features. Although the specific details remain secret, a number of research publications exist on the topic (see for example [13]).

Key Applications

The key application of web search relevance ranking is in the algorithmic search component of web search engines. Similar methods are also employed to bias the ranking of the advertisements displayed in search results. Some of the principles have been applied in other types of search engines such as corporate search (intranet, email, document archives, etc.).

Future Directions

Research continues to improve all of the relevance features discussed here. This research has lead to a continuous improvement of search engine quality. Nevertheless, current relevance features are becoming increasingly hard to improve upon. Considerable research is centered today on discovering new types of features which can significantly improve search quality. Only two of the most promising areas are mentioned here:

Query-understanding: Different types of queries may require very different types of relevance ranking algorithms. For example, a shopping query may require very different types of analysis from a travel or a health query. Work on algorithms that understand the intent of a query and select different relevance ranking methods accordingly could lead to dramatic increases in the quality of the ranking.

Personalization: In principle it is possible to exploit user information to "personalize" web search engine results. Different results would be relevant to a query issued by a layperson than a topic expert, for example. There are many different ways to personalize results: with respect to the user search history, with respect to the user community, with respect to questionnaires or external sources of knowledge about the user, etc. Many scientific papers have been written on this topic, but the problem remains unsolved. Commercial web search engines have mainly shied away from personalized algorithms. Google has proposed several forms of personalized search to its users, but this feature has not had much success. Nevertheless, the search continues for the right way to personalize relevance ranking.

Experimental Results

Evaluation of web relevance ranking is difficult and very costly, since it involves human judges labeling a collection of queries and results as to their relevance. The most careful evaluations of web relevance features are carried out by web search engine companies, but they are not disclosed to the public. There have been very many partial evaluations of search engines published, but they are always controversial due to their small scale, their experimental biases and their indirect access to the search engine features.

A number of experimental benchmarks have been constructed for public scientific competitions. Although they are small and partial they can be used for experimentation (see below).

Data Sets

Commercial search engines dispose of very large data sets comprising very many documents (e.g., hundreds of millions), queries (e.g., tens of thousands), and human relevance evaluations (e.g., hundreds of thousands). These data sets are routinely used to develop and improve features for relevance ranking. See [10,13] for examples of this.

Publicly available data sets for experimentation are very small compared to those used by commercial search engines. Nevertheless, they may be used to investigate some features and their combination. The most important datasets for web relevance ranking experiments are those developed in the Web-track of the TREC competition organized by NIST [4].

URL to Code

Due to the extraordinary cost of developing and maintaining a full-scale web search engine, there are no publically available systems so far. The Nutch project (http://lucene.apache.org/nutch/) is aiming to build an open-source web-scale search engine based on the Lucene search engine. Other retrieval engines capable of crawling and indexing up to several millions of documents include INDRI (http://www.lemurproject.org/indri/), MG4J (http://mg4j.dsi.unimi.it/) and TERRIER (http://ir.dcs.gla.ac.uk/terrier/).

Cross-references

- Anchor Text
- BM25
- Document Links and Hyperlinks
- Field-Based Information Retrieval Models
- Information Retrieval
- Language Models
- Relevance
- Relevance Feedback
- Text Categorization
- Text Indexing and Retrieval
- Vector-Space Model
- WEB Information Retrieval Models
- Web Page Quality Metrics
- Web Search Relevance Feedback
- Web Spam Detection

Recommended Reading

1. Baeza-Yates R. and Ribeiro-Neto B. Modern Information Retrieval. Addison Wesley, Reading, MA, 1999.
2. Boyan J., Freitag D., and Joachims T. A machine learning architecture for optimizing web search engines. In Proc. AAAI Workshop on Internet Based Information Systems, 1996.
3. Culliss G. The Direct Hit Popularity Engine Technology. A White Paper, DirectHit, 2000. Available online at https://www.uni-koblenz.de/FB4/Institutes/ICV/AGKrause/Teachings/SS07/DirectHit.pdf. Accessed on 27 Nov 2007.
4. Hawking D. and Craswell N. Very large scale retrieval and Web search. In TREC: Experiment and Evaluation in Information Retrieval, E. Voorhees and D. Harman (eds.). MIT Press, Cambridge, MA, 2005.
5. Joachims T. and Radlinski F. Search engines that learn from implicit feedback. IEEE Comp., 40(8):34–40, 2007.
6. Kleinberg J. Authoritative sources in a hyperlinked environment. Technical Report RJ 10076, IBM, 1997.
7. Langville A.N. and Meyer C.D. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, Princeton, NJ, 2006.
8. Manning C.D., Raghavan P., and Schütze H. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK, 2008.
9. Marchiori M. The quest for correct information on the Web: hyper search engines. In Proc. 6th Int. World Wide Web Conference, 1997.
10. Najork M. Comparing the effectiveness of HITS and SALSA. In Proc. Conf. on Information and Knowledge Management, 2007, pp. 157–164.
11. Page L., Brin S., Motwani R., and Winograd T. The PageRank citation ranking: bringing order to the Web. Technical Report, Stanford Digital Library Technologies Project.
12. Richardson M., Prakash A., and Brill E. Beyond PageRank: machine learning for static ranking. In Proc. 15th Int. World Wide Web Conference, pp. 707–715.2006,
13. Taylor M., Zaragoza H., Craswell N., Robertson S., and Burges C. Optimisation methods for ranking functions with multiple parameters. In Proc. Conf. on Information and Knowledge Management, 2006, pp. 585–593.