Web Search Literacy Skills

Ioannis Karatassis
Institute for Information Systems
University of Duisburg-Essen
47048 Duisburg, Germany
karatassis@is.inf.uni-due.de

Norbert Fuhr
Institute for Information Systems
University of Duisburg-Essen
47048 Duisburg, Germany
norbert.fuhr@uni-due.de

February 22, 2020

Contents

1 Theoretical Skills

1.1 Indexing

S1 indexing_bad_content: Documents containing illegal, inappropriate, or content of low quality may find their way into the index.
S2 indexing_cost: Indexing a website is free of charge for webmasters.
S3 indexing_crawler: Search robots, crawler, or simply robots, systematically browse the web. They visit web pages and deliver their content to the search system for the purpose of web-indexing.
S4 indexing_fields: Information found in documents, e.g., title, url, and content, are usually stored in fields inside the index. The HTML structure of HTML pages is exploited, too.
S5 indexing_new_documents: During indexing of a web page, links are extracted and used to find other web pages and documents. Supported document types will then be analyzed and indexed, too.
S6 indexing_old_content: Searchers might be presented old information on a SERP as search engines retrieve information from their index.
S7 indexing_register: Webmasters may submit their websites to search engines.
S8 indexing_reindex: Indexed documents are revisited by crawlers in order to obtain and index the newest version of the document.
S9 indexing_reindex_interval: The time interval at which a web page is re-indexed varies and depends on several factors, e.g., update frequency and type of website or document.
S10 indexing_veracity_assessment: Indexed contents do not undergo veracity assessment.
S11 indexing_web: Search engines cover only a portion and not the whole world wide web. Some documents can therefore not be retrieved through search engines although they exist online.

Table 1: Category: Indexing

1.2 Searchability

S12 searchability_availability: In general, content that is publicly available may be subject of indexing. Content that requires subscription or registration is usually not indexed and thus not searchable.
S13 searchability_index_exclusion: Webmasters may instruct search robots how to crawl and index pages on their websites via a file named robots.txt or through meta tags. Entire websites and subpages can be excluded from indexing by doing this.
S14 searchability_nolink: Unindexed web pages that are not linked by other indexed web pages cannot be retrieved through search engines without further intervention, e.g., through their webmasters.
S15 searchability_only_index: After hitting the search button, search engines search their databases instead of the world wide web or the Internet to retrieve search results. Documents that are not indexed can thus not be searched at all.
S16 searchability_supported_documents: Besides HTML, only a few other document types are supported, e.g. TXT, PDF, and PPT. Images, videos, and other supported media types located on web pages are usually stored in the index, too, and thus searchable.

Table 2: Category: Searchability

1.3 Web Search

S17 search_content: When retrieving search results, the document's content (body) is matched against the search query per default.
S18 search_fields: Apart from the content, fields stored in the index are matched against search queries by search engines to find relevant documents.
S19 search_logical_and: By default, all searches are AND searches. Search engines find web pages that contain all given keywords. The way how queries are processed differs between search engines though.
S20 search_logical_not: The logical NOT operator is supported in popular search engines. They can be instructed to exclude search results containing the specified keywords.
S21 search_logical_or: The logical OR operator is supported in popular search engines. They can be instructed to retrieve search results containing at least one of the specified terms.
S22 search_match_media_documents: Contents of media documents, e.g., images and videos, are not matched against queries. Instead, search engines use meta-information and web pages for comparison.
S23 search_phrase: Search engines can be instructed to search for phrases.
S24 search_qualitative_search: Search engines are not suitable for qualitative search as indexed documents do not undergo quality control assessment.
S25 search_quantitative_search: Search engines are suitable for quantitative search as they aim at answering every search request.
S26 search_restriction: Fields or properties of search results can be restricted through the filtering tool and query operators, but not through simple keywords.
S27 search_terms: Search engines compare every term of a query with every term within the index to find relevant search results.

Table 3: Category: Web Search

1.4 Linguistic Functions

S28 linguistic_case: Search queries are usually case-insensitive. `Wooden House` and `wooden house` yield same search results.
S29 linguistic_punctuation: Search engines discard most punctuation. Which characters are discarded and which are kept varies from one search engine to another.
S30 linguistic_spelling: When spelling errors are detected, search engines might auto correct them and use the corrected query for retrieval, or simply inform the user.
S31 linguistic_stemming: During indexing and searching, stemming on terms is usually employed. Houses is reduced to house in order to find more potentially relevant documents.
S32 linguistic_stop_words: Stop words are words, which are partially or completely ignored by search engines during indexing and query processing. They are deemed irrelevant for searching purposes because they have very little meaning or occur less frequently in the language.

Table 4: Category: Linguistic Functions

1.5 Ranking

S33 ranking_features: Search engines use a variety of features for ranking search results. Some of them can be affected through the way a query is constructed.
S34 ranking_personalized: Popular search engines use information about the individual to tailor their search results. The order in which search results are displayed is affected as a result and probably the result set itself.
S35 ranking_personalized_different_users: Two users issuing the same query may receive different search results due to personalization.
S36 ranking_personalized_history: The web history of searchers may be used as part of personalized search.
S37 ranking_personalized_language: The user's browser language may be used as part of personalized search.
S38 ranking_personalized_location: The location of searchers may be used as part of personalized search.
S39 ranking_personalized_same_user: The same user issuing the same query on multiple devices may receive different search results due to personalization.
S40 ranking_term_frequency: Term frequency, i.e., the number of occurrences of a query term inside a document, usually affects the document's position inside the result list.
S41 ranking_term_ordering: The ordering of query terms matters when it comes to ranking. `Wooden house` and `house wooden` are actually two different queries and therefore might yield different search results.

Table 5: Category: Ranking

2 Practical Skills

2.1 Query Language

S42 query_and: Spaces between keywords in queries are interpreted as a logical and. However, the behavior differs between search engines. Some search engines provide the AND operator for better control.
S43 query_define: The define: operator instructs search engines to show definitions of terms.
S44 query_filetype: The filetype: operator restricts search results to those of a certain file type, e.g., PDF.
S45 query_grouping: Terms or phrases and their operators can be enclosed by parentheses to specify the order in which they are interpreted, and to construct more complex queries.
S46 query_intitle: The intitle: operator finds documents that include a specific word as part of the indexed title tag.
S47 query_not: NOT can be used in front of a term or phrase to exclude search results that contain that specific unit.
S48 query_not_short: The minus sign placed in front of a term or phrase is interpreted as short version for NOT by some search engines.
S49 query_number_rage: Two numbers, separated by two periods, e.g., 10..20, represent the number range operator. Search results contain numbers in the specified range.
S50 query_or: OR can be used between two terms or phrases to instruct search engines to retrieve websites containing one or both of them.
S51 query_or_short: Some search engines interpret the pipe character as the OR operator.
S52 query_phrase: Quotation marks in queries serve to match exact phrases. Search engines find only documents that have the specified terms together as a phrase.
S53 query_site: The site: operator restricts search results to a particular domain.
S54 query_syntax_operators: Correct syntax ensures keywords are identified as operators. Logical operators, i.e., AND, OR, and NOT, must be in uppercase while field and other types of operators in lowercase and followed by a colon, e.g., intitle:gamification.

Table 6: Category: Query Language

2.2 Search Tactics

S55 tactic_bibble: To look for a bibliography already prepared, before launching oneself into the effect of preparing one. More generally, to check if the search work one plans has already been done in a usable form by someone else. [1]
S56 tactic_filter: To refine search results by means of the filtering tool or query operators.
S57 tactic_hubspoke: Follow links from a search engine result page in a hub and spoke pattern, perhaps using separate windows/tabs, or the browser back button. Adapted from [2].
S58 tactic_phrase: To use a phrase search to maximize the ranking of terms comprised of several words. [2]
S59 tactic_select: To break complex search queries down into subproblems and work on one problem at a time. [1]
S60 tactic_sub: To move downward hierarchically to a more specific (subordinate) term. [1]
S61 tactic_support: To use any form of context-related suggestions provided by the search engine to further the search.
S62 tactic_type: To select the appropriate type of search results with regard to search intention.

Table 7: Category: Search Tactics

References

[1] Marcia J. Bates. Information search tactics. *Journal of the American Society for Information Science*, 30(4):205–214, 1979. ISSN 1097-4571. doi: 10.1002/asi.4630300406.

[2] Alastair G. Smith. Internet search tactics. *Online Information Review*, 36(1): 7–20, 2012.