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

June 29, 2019

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.
indexing_bad_content: Documents containing illegal or inappropriate content may get into the index.
indexing_fields: Properties of documents, e.g., title, url, and content, are usually stored in fields inside the index.
indexing_cost: Indexing a website is free of charge.
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.
indexing_old_content: Searchers might be presented old information as search engines display indexed documents that may have been updated after they have been indexed.
indexing_supported_documents: Besides HTML, only a few other file types are actually processed during indexing, e.g. TXT, PDF, and PPT. Images, videos, and other supported document types located on web pages are usually stored in the index, too.
indexing_register: Webmasters may submit their websites to search engines.
indexing_reindex: Indexed documents are revisited by crawlers in order to obtain and index the newest version of the document.
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.
indexing_veracity_assessment: Indexed contents do not undergo veracity assessment.
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: Web Search Literacy Skills. Category: Indexing

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.
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.
searchability_nolink: Unindexed web pages that are not linked by other indexed web pages can not be retrieved through search engines without further intervention, e.g., through webmasters.
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.

Table 2: Web Search Literacy Skills. Category: Searchability

search_content: When retrieving search results, document content is matched against search queries per default.
search_fields: Search engines match the content of fields against search queries to find relevant documents.
search_fulltext: Search engines compare every term of a query against every term within the index to find relevant search results.
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.
search_logical_not: The logical not is supported in popular search engines. They can be instructed to exclude search results containing the specified keywords.
search_logical_or: The logical or is supported in popular search engines. They can be instructed to return search results containing at least one of the specified terms.
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.
search_qualitative_search: Search engines are not suitable for qualitative search as indexed documents do not undergo quality control assessment.
search_quantitative_search: Search engines are suitable for quantitative search as they aim at answering every search request.
search_restriction: Fields or properties of search results can be restricted through the filter tool and query operators.
search_title: When retrieving search results, document title is match against search queries per default.

Table 3: Web Search Literacy Skills. Category: Web Search

linguistic_case: Search queries are case-insensitive. `Wooden House` and `wooden house` yield the same search results.
linguistic_spelling: When spelling errors are detected, search engines might correct them and use the corrected query for retrieval or simply inform the user.
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.
linguistic_stop_words: Stop words are some 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 frequently in the language.
linguistic_punctuation: Search engines discard most punctuation. Which characters are discarded and which are kept varies from one search engine to another.

Table 4: Web Search Literacy Skills. Category: Linguistic Functions

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.
query_define: The define: operator finds definitions of terms.
query_filetype: The filetype: operator restricts search results to those of a certain file type.
query_grouping: Terms or phrases and their operators can be enclosed by parentheses to specify the order in which they are interpreted.
query_intitle: The intitle: operator finds documents that include a specific word as part of the indexed title tag.
query_not: NOT can be used in front of a term to exclude search results that contain that specific term.
query_not_short: The minus sign placed in front of a term is interpreted as NOT by some search engines.
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.
query_or: OR can be used between two keywords to instruct search engines to retrieve websites containing one or both of the keywords.
query_or_short: Some search engines interpret the pipe character as the OR operator.
query_phrase: Quotation marks in queries serve match exact phrases. Search engines find only web pages that have the specified terms together as a phrase.
query_site: The site: operator restricts search results to a particular domain.
query_syntax_operators: Correct syntax ensures keywords are identified as operators. Logical operators must be uppercase, e.g., OR, while field and other operators lowercase followed by a colon, e.g., intitle:gamification.

Table 5: Web Search Literacy Skills. Category: Query Language

ranking: Search engines use a variety of features for ranking search results.
ranking_personalized: Popular search engines use information about the individual to tailor search results. The order in which search results are displayed is affected as a result and potentially the result set itself.
ranking_personalized_different_users: Two users issuing the same query may receive different search results due to personalization.
ranking_personalized_history: Web history of searchers may be used as part of personalized search.
ranking_personalized_language: User language may be used as part of personalized search.
ranking_personalized_location: Location of individuals may be used as part of personalized search.
ranking_personalized_same_user The same user issuing the same query on multiple devices may receive different search results due to personalization.
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.
ranking_term_ordering: Ordering of query terms matters when it comes to ranking. `Wooden house` and `house wooden` might yield different search results.

Table 6: Web Search Literacy Skills. Category: Ranking

tactic_bibble: To look for a bibliography already prepared, before launching oneself into the effect of preparing one; more generally, to check to see if the search work one plans has already been done in a usable form by someone else. [1]
tactic_filter: To refine search results by means of the filter tool or query operators.
tactic_hubspoke: Following links from a landmark site such as a directory or a search result, then returning to it to follow other paths, is a tactic specific to the Internet, and facilitated by browser features such as history and tabs. [2]
tactic_phrase: To use a phrase search to maximize the ranking of terms comprised of several words. [2]
tactic_select: To break complex search queries down into subproblems and work on one problem at a time. [1]
tactic_suggestion: To use any form of context-related suggestions provided by the search engine to further the search.
tactic_sub: To move downward hierarchically to a more specific (subordinate) term. [1]
tactic_type: To select the appropriate type of search results with regard to search intention.

Table 7: Web Search Literacy Skills. 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.