November 2020

Sensemaking for Broad Topics via Automated Extraction and Recursive Search

Alankar Jain

William M Leszczuk

Sitaram Iyer

Mehrbod Sharifi

Seher Aylin Altiok

Sensemaking for Broad Topics via Automated Extraction and Recursive Search

ABSTRACT

The availability of vast amounts of diverse information related to a broad topic makes it difficult and time-consuming for users to find and digest the right information regarding various low-level topics within the broader space. Current approaches to addressing these challenges include providing curated topical pages, relevant query refinement suggestions, list of subtopics, etc. However, these approaches do not scale and offer inadequate support for sensemaking. This disclosure describes automated techniques that extract information from online information sources by using a query related to a high-level topic to recursively formulate additional queries for subtopics to construct a hierarchical set of topics related to the broad query. The results can be utilized to provide a user interface using the hierarchical topic levels which can make it faster and easier for users to understand and navigate information regarding a high-level topic.

KEYWORDS

- Search engine
- Seed query
- Query refinement
- Query transformation
- Topic deconstruction
- Topic extraction
- Topic hierarchy
- Document structure
- Document parsing
- Sensemaking

BACKGROUND

Users often search the web for information pertaining to a high-level task or goal, such as buying a car, running a marathon, etc. In many cases, these high-level tasks or goals are broad and composed of various sub-components. In such cases, the availability of vast amounts of diverse information regarding each sub-component makes it difficult and time-consuming for

users to find and digest the right information. These issues are compounded by users' unfamiliarity with the topics, which can make it challenging for users to formulate appropriate queries to fulfill their information needs.

Currently, search engines attempt to address these challenges by providing results pages that include curated and/or computed user experiences to help users orient themselves within a broad topic space. For example, a user who searches for "anxiety disorder" can be shown results that include sections for various aspects of the specific disorder, such as "Overview," "Symptoms," "Treatments," "Specialists," etc. However, provision of such experiences require substantial manual curation effort for the search engine provider, thus making it difficult to scale to different topics, locales, and/or languages.

Alternatively, or in addition, the search results can include clickable query refinement suggestions. For example, a user who enters the query "buy a house" can be shown suggestions for additional related queries, such as "get a mortgage to buy a house," "what types of properties are available for purchase?" etc. Further, the user can be presented with a list of steps involved in the task connected to the query. For instance, the results page for the "buy a house" query can provide a sequence of individual steps that are typically involved in purchasing residential real estate. However, these approaches may not be adequate since a user that is unfamiliar with the high-level topic of the query is highly likely to be unfamiliar with the subtopics represented in the query suggestions and/or list of steps, thus requiring further sensemaking effort to develop sufficient understanding of the topic space.

DESCRIPTION

This disclosure describes automated techniques to extract information from online information sources to enhance the information presented in response to a query about a high-

level task (referred to herein as a "seed query") to help users understand the broad topical space connected to the seed query. The information extraction and presentation is performed via the following steps:

1. **Retrieve search result pages:** A search engine is used to retrieve the top N pages corresponding to the query. These pages are likely to include content that provides broad information about the high-level topic.
2. **Extract document structure:** Each retrieved page is parsed based on its structure to extract relevant content. For instance, the content of section headers and/or titles within the page typically provides concise information that can help refine the broader topic. For example, a page pertaining to the query "how to buy a car" can include headers for various related sub-topics such as "research vehicle features," "obtain financing for a vehicle purchase," etc.
3. **Aggregate information across documents:** Noise and/or lack of relevant information within the extracted content of a single page can be overcome by aggregating the extracted information across multiple retrieved pages using appropriate aggregation techniques, such as hierarchical clustering. For instance, such aggregation can produce a list of headers that capture lower-level information related to the broader query, with the list being ranked based on content across the multiple pages retrieved for the query.
4. **Transform aggregated information into queries:** If needed, each header within the ranked list as formulated in the previous step is transformed into a form suitable for a query. The process uses any suitable transformation technique, such as heuristics, to filter unneeded text and/or add relevant contextual information. For example, the extracted header "get pre-approved for financing" within pages retrieved for the seed query "buy a car" can be

transformed into a format suitable for a follow-up query, such as “how to get pre-approved for financing a vehicle purchase.”

5. **Repeat for transformed queries**: For further in-depth exploration, the above steps can be repeated for each of the queries that result from the transformation process performed in the previous step. For instance, in the case of the seed query “buy a car,” the above steps can be repeated for queries connected to lower-level topics, such as “how to get pre-approved for financing a vehicle purchase,” “how to research car features,” etc. Each step can in turn produce queries regarding lower-level subtopics. For instance, the query “how to get pre-approved for financing a vehicle purchase” can lead to a transformed query suggestion such as “how to check your credit report before applying for financing.”

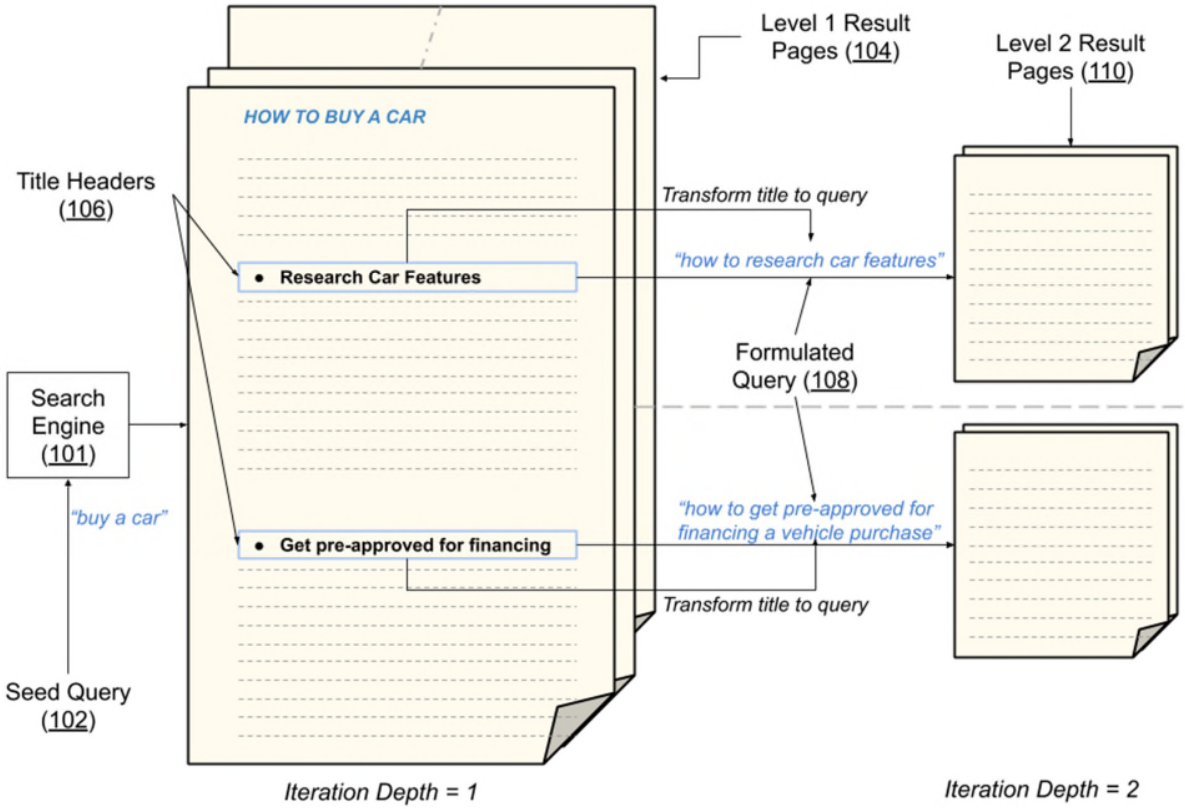

Fig. 1: Automated extraction of headers for transformation into queries for subtopics

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. A user invokes a search engine (101) for a seed query (102, “buy a car”) regarding the broad high-level topic of buying a car. The pages of the results for the query (104) are parsed to extract title headers (106) or other content that indicates subtopics within the pages.

The extracted headers are aggregated across all research pages, and each is suitably transformed to formulate a corresponding query (108) pertaining to a subtopic related to the seed query. For instance, Fig. 1 shows queries pertaining to the subtopics of researching car features and obtaining financing within the high-level topic of buying a car. The process can continue recursively with the formulated queries as the seed for the next level.

The described techniques can be implemented within any application or service that provides search functionality. The number of pages (N) and the depth of iteration for query processing can be chosen, e.g., determined dynamically, based on the topic of the seed query and/or the content of the initial set of pages that are retrieved based on the seed query. The outcomes of the process described above can be used to structure search results that are presented using any suitable user interface (UI) techniques, including but not limited to: query refinement suggestions, “how to” queries, information snippets (e.g., structured answers), a tree representing topical hierarchy, etc.

The described techniques can be applied to any high-level query that pertains to a task or topic that includes multiple subtopics, possibly with multiple hierarchical levels. By automatically retrieving and parsing search results at each level, and formulating and executing queries pertaining to subtopics, the described techniques can automatically construct a journey that the user can take to fulfill their information needs related to the high-level topic. As the techniques are automated, manual curation is not necessary, and thus, the techniques can scale to

serve a large number of queries, across different topics, locales, and/or languages. Provision of a user interface using the hierarchical topic levels can make it faster and easier for users to understand and navigate information regarding a high-level topic, thus improving the user experience.

CONCLUSION

This disclosure describes automated techniques that extract information from online information sources by using a seed query related to a high-level topic to recursively formulate additional queries for subtopics to construct a hierarchical set of topics related to the broad query. The results can be utilized to provide a user interface using the hierarchical topic levels which can make it faster and easier for users to understand and navigate information regarding a high-level topic.