Best Practices:

Getting the Most Out of Concept Search

LLM LIQUID LITIGATION MANAGEMENT, INC.

Abstract

Concept search is a powerful tool in the e-discovery arsenal. When properly utilized, concept search can be a valuable way to expand the scope of your review or to narrow its focus. However, concept searches are often misused because of the prevalence of the Boolean paradigm. This paper outlines best practices that will allow you to get the most out of concept searching.

Boolean Logic

Boolean logic was developed in 1854 by George Boole, and applied to internet search engines in the early 1990s. Since then, it has dominated the industry. As a result**,** most people have intimate familiarity with Boolean – otherwise known as keyword – searches. The search query entered is a word, or perhaps a short phrase, which is then located within the set of searchable documents. Operators such as AND, OR, and NOT may be used to create a more complex query consisting of multiple keywords in a manner similar to an algebraic equation. Any hits that are returned will exactly match the search query, regardless of context. For example, if the query is “bomb”, the results “that movie was a bomb” and “how to build a bomb” will both be returned as results. This specificity can be useful, but it is also limiting when you need to distinguish between concepts rather than single words.

Latent Semantic Indexing

Latent semantic indexing, or LSI, was developed in the late 1980s at Bell Laboratories, and was first used to assign documents for review in 1992. A mathematical approach to textual analysis, LSI forms the backbone of the concept search technology.[1] Concept search has become increasingly central to the government/intelligence community and the e-discovery industry; its success is due to the fact that, unlike Boolean searches, context matters. In the example given above, a concept search would be able to distinguish between “that movie was a bomb” and “how to build a bomb” because it would determine that the word “bomb” is being used differently within each phrase.

Concept searches distinguish the relationships between the words in a mathematically quantifiable way, much the same way that the human mind does intuitively. The LSI technology allows the best features of a computer – speed and thoroughness – to be combined with the best abilities of a human reviewer.

Changing the Boolean Paradigm

For the reviewer used to entering keywords and Boolean operators to search, it may be difficult to establish the correct process for concept searching. It is not uncommon for a reviewer unfamiliar with the differences between concept and keyword searches to treat them interchangeably and obtain subpar results because of this misuse. As a result, some reviewers view concept searches as being inferior to keyword searches, and neglect them as a part of their e-discovery toolkit. In reality, keyword and concept searches complement each other when used correctly.[2]

It cannot be emphasized enough that concept searches are not a replacement for keyword searches, and that keyword searches are not a substitute for concept searches. Both types of searches have distinct limitations and cannot offer all of the functionality a reviewer needs when used alone. The most efficient and thorough review process combines both types of searches; it is simply an issue of using the right type of search at the right time, as determined by the review goals.

Because concept searches are context sensitive, the search query is constructed differently than the typical query one would use for a keyword search. Instead of entering a single keyword or a string of keywords joined by Boolean operators, text of substantial length should be submitted as the query for a concept search. Paragraphs (either created by the reviewer, or excerpted from longer texts) or entire documents form the basis for a query, which provides a contextual framework.

Establishing basic best practices for concept searches will assist reviewers in determining when it is most appropriate to use one search method over the other to get the results that they need. Learning how to make concept searches work will improve the efficiency of the review process from start to finish.

When Should Concept Searches be Used?

At times, it may be unclear whether or not a concept search is the best search method to use. In general, if you want to see all instances of the usage of a single word or highly specific phrase – such as a company name, or the name of a witness – you should use a Boolean search. For example, if you needed to see all documents where “wafer” was used, regardless of how it was being used, it would be best to do a keyword search. It would be a bad idea to do this with a concept search, because there are multiple ways in which the word “wafer” could be used, and without providing more detail in the query, the system would not understand exactly what you wanted.

If, on the other hand, you wanted to find all of the documents in which “wafer” was used in the context of computer hardware, rather than in the context of baked goods, you would want to use a concept search (more details on how to format your search query are below). This would allow you to discriminate between the two usages that could exist in your pool of searchable documents. If you wanted to do this with a Boolean search, it would be almost impossible because you would have to enter many different operators in order to exclude or include very specific linguistic cues that distinguish between the two types of “wafer”; it is almost certain that you would not be able to think of all of these cues, and even if you could, it would be prohibitively time consuming and complex to include them all.

Additionally, reviewers should be aware how concept searches can be used in different stages of the discovery process. When used early in the discovery process, concept searches provide a structured way to explore your set of searchable documents. For example, searching for a single well defined concept may lead to the discovery of other highly relevant concepts of that you might not have even been aware of. If you are unsure what terms or ideas might be important to your matter, casting a wide net with a concept search will be informative. This means reviewers can be more confident that they are not excluding a vitally important piece of the puzzle early on that will negatively impact the rest of the review. After the ideas most important to the review have been defined, concept searches allow reviewers to focus their attention only on the specific documents containing that information.

As noted previously, concept and keyword searches solve different problems in different ways. Learning to switch between the two depending on your needs rather than relying solely upon one or the other is fundamental.

Concept searches should be used when:

- When trying to distinguish between synonyms
- When you are dealing with a complex concept that cannot be easily expressed in a single word or phrase
- When discovering ideas central to your matter
- When trying to focus review

How Should Concept Search Queries be Constructed?

After you have determined that a concept search is the most appropriate search method to use for your particular goals, the query must be constructed. As mentioned earlier, a common mistake is to assume that concept search queries and keyword search queries are interchangeable. Rather than using a keyword or phrase created with Boolean operators, it is crucial that a longer section of text is used as the search query. For example, an email message containing a discussion of one of the ideas or terms central to the matter. This allows the system to analyze the mathematical – and therefore, linguistic – relationships present in the text. Excerpts of longer documents are also ideal candidates for use as a concept search query.

Avoid the use of a list of bullet points, unrelated sentences, and single words, as these typically do not contain enough conceptually rich content. In order to check that the search query is optimally constructed, ask yourself if a human would be able to perform a search if you gave them the exact same query.

In order to better illustrate the previous point, consider the example of someone who walks into the library looking for a particular book. They cannot remember the title or author of the book, but when they describe the plot and character details to the librarian, the librarian is able to easily locate the book that they were looking for. Regardless of length, your concept search query needs to encapsulate the essence of the idea you are interested in. The query could be five highly descriptive sentences, or a more loosely focused document of several pages as long as it communicates the concept of interest clearly.

When selecting a document or passage of text to use, it is important to carefully clean it so you are providing the best possible material for the analytics system to use. If you are using an email, be sure to remove the headers and footers from the message and only include the content most relevant to your query. When using concept search, the maxim, “Garbage in, garbage out,” should be remembered. You will only get useful results if you start from a solid foundation.

Concept search queries should be constructed:

- From entire documents or longer passages of text
    - There should be enough content to make context clear
    - Length is less important than depth
- Avoiding lists, bullet points, keywords, short phrases, and disjointed sections of text
- From relevant text that has been cleaned of unimportant content

How to Work With the Results

Once a concept search has been performed, the results will appear as a ranked list of documents, with numbers ranging from 0-100. Be aware that these numbers do not represent a percent match; instead, they represent cohesiveness between the query and the result, or in other words, how closely related they are. If no minimum level of cohesiveness is specified by the reviewer, all searchable documents will appear on the list of results, ranked from most to least relevant. In order to prevent this, you will want to set a minimum threshold for what results to consider, between 0-100. Thresholds above 80 are most common. It is best practice to set a lower threshold initially, and then increase it once you have manually determined what results should be included.

Other Uses of the LSI Technology

There are several novel ways in which LSI can be applied to augment the review process that go beyond mere identification of ideas within the searchable documents via a concept search. You can use an excerpt from the complaint to easily identify where the complaint and the discovery documents intersect. Also, LSI is the basis of the keyword expansion functionality, which augments traditional keyword search by suggesting additional search terms. This may give you additional ideas for review that you would not have considered.

Using LSI, you can generate clusters of related documents and modify these clusters based upon your preferences for cohesiveness. In clustering, documents are grouped by concept, and prominent terms are displayed for each cluster; these can then be used as keywords for additional search queries. Different clusters can also be assigned to specific reviewers, or used to quickly identify outliers in the document set. Removing these outliers can streamline and focus the review process.

If you are trying to answer a specific question about your matter (for example, are we guilty of what we have been accused of?), you can create a highly specific, fictional passage of text yourself to use as a query. Think creatively; by deepening your knowledge of how LSI works, you broaden your review horizons.

- Use a passage from the complaint to begin your review
- Keyword expansion
- Generate clusters and view prominent terms
    - Prominent terms can be used as keywords for Boolean search
    - Specific clusters can be assigned to specific reviewers
    - Identify outliers
- Create text to test a question or hypothesis about your matter

In Conclusion

Concept searching is a valuable analytical tool that, when used properly, complements and augments more traditional Boolean searching. In order to get the most out of concept search, you need to know when concept searches are most appropriate given a particular review goal, how to construct the best possible search queries, and how to work with the results. Remember that concept searching operates in a manner similar to the human mind, and if you try to use it while operating within the Boolean paradigm, you will be disappointed with the results. Construct your queries by asking yourself if a human being would understand what you are asking, and remember to clean your queries of irrelevant information.

Developing a thorough understanding of the best practices for concept searching will not only improve your review efficiency but also your review quality.

References

1. Zukas, A., Price, R.J. “Document Categorization Using Latent Semantic Indexing” Content Analyst Company, LLC. 2009. <http://www.contentanalyst.com/images/images/whitepaper_categorization_using_lsi.pdf>

2. The Sedona Conference®, WGI. The Sedona Conference® Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery. The Sedona Conference Journal. Volume 8, 2007.