Conceptual Search – ESI, Litigation and the issue of Language

David T. Chaplin
Kroll Ontrack

Across the globe, legal, business and technical practitioners charged with managing information are continually challenged by rapid-fire evolution and growth in the legal and technology fields. In the United States, new compliance requirements, amendments to the Federal Rules of Civil Procedure (FRCP) and corresponding case law, along with technical advances, have made litigation support one of the most exciting professions in the legal arena. In the UK, revisions to the Practice Direction to CPR Rule 31 require parties in civil litigation to consider the impacts associated with electronic documents.

One emerging technology trends—both aiding and complicating the management of electronically stored information (ESI) in litigation in the US, EU and UK alike—is the notion of "conceptual search." This paper focuses on the evolution of conceptual search technology, and predictions of where this science will take legal professionals and technical information managers in coming years and a look at the advantages conceptual search can provide in dealing with the issue of language.

This paper will focus primarily and the latent semantic analysis approach to conceptual search and why this approach is advantageous when searching ESI regardless of the language used in the documents, even to the extent of allowing for cross language searching and accurate searching of documents that contain co-mingle foreign terms with the native language. In order to discuss the language issue the following topics will first be established;

What is conceptual search?
Dominant approaches to conceptual search.
Conceptual Search as a Strategic Litigation Tool

What Is Conceptual Search?

Conceptual search was born out of a need to better locate information in the context of a changing corporate language. Legal teams require access to the information they need to make better, more informed decisions about their cases. Not only is the amount of information growing, there are also significantly more terms being used within the normal corporate lexicon. Abbreviations, acronyms, text and email slang, along with industry and corporate specific terminology, are continually progressing. It is becoming increasingly important that search technology adapts to the changing use of language and the ever-growing amount of information.

Conceptual search is defined as the ability to retrieve relevant information without requiring the occurrence of the search terms in the retrieved documents. Most search technology in use today is traditional keyword search that requires the search term to appear in the retrieved documents. Many of these traditional search engines have mimicked conceptual search through the use of synonym lists and other human-maintained query expansion approaches. True conceptual search retrieves relevant information in a way that does not require the presence of the search terms without the use of query expansion or independently maintained lexicons, taxonomies or synonym lists. This is why conceptual search is distinctly different from keyword search and is the key to why it is able to adapt to changes in language and the use of slang. Conceptual search allows you to locate information about a topic by understanding what words mean in a given context.

The conceptual search engine must measure subtle patterns and relationships that occur in language. The importance of understanding the context of information is amplified when you consider the complexity of language. Effective search requires the search engine to address synonymy (different words with same meaning) and polysemy (same word with different meanings). For example, cellular means something different when the context is biology versus wireless communications. Conceptual search understands these differences and, in effect, smoothes out the idiosyncrasies of speech by analyzing words and how they are used in context. The measurement of how terms are used in context provides the conceptual search engine with the ability to learn new terminology without human intervention.

Dominant Approaches to Conceptual Search

There are two basic approaches to conceptual search: statistical and linguistic. Statistical methods usually learn from text and do not require any pre-built language models. Statistical methods analyze how terms are used within the document collection to be searched. The statistical method determines the underlying structure of the language based on the documents in the collection. Linguistic methods, including natural language processing (NLP) and syntactic approaches, require models of language that are created and maintained by humans. These models are based on insight into the language and content, or from a training set of related text in order to find universal properties of language and to account for the language's development.

There are also two basic methods in producing conceptual search: automatic and manual. Automatic methods allow you to present any source of information to the system without considering structure or syntax. The automatic method allows for the engine to learn as a new language is introduced to the document collection without any human intervention. Manual methods require humans to create and maintain a taxonomy, ontology or synonym list in order to create and maintain relationships. The knowledge is fixed and will have to be altered to account for new vocabulary or relationships.

One may further classify conceptual search by which scientific method of learning is applied. Again, there are two basic approaches: supervised and unsupervised. Supervised learning requires feedback to improve and to initially specify what needs to be learned. Explicit examples need to be supplied to the system for the engine to learn. Unsupervised learning is fully autonomous and can arrive at an optimal solution without requiring user feedback or pre-defined training sets.

Finally, conceptual search technology can be query and non-query based. Methods have been developed that enable conceptual search technologies to automatically cluster or folder documents that are similar in theme. These clusters are labeled and provide the business user with the ability to navigate a large set of information that is organized and appropriately labeled without having to issue a query. The ever-increasing influx of information into critical knowledge management systems requires improved methods automatically organizing and making available documents, without requiring the user to know what search to perform. This approach can also be used to enhance or refine existing corporate taxonomies or to provide a "snapshot" of large document collections.

Business professionals, attorneys and litigation support professionals are spending more and more time searching for information to make better and faster decisions. Conceptual search reduces the number of queries, results sets and redundant hits in the standard process of collecting, reviewing and producing documents in discovery. Ultimately, conceptual searching techniques allow legal teams to retrieve the maximum number of relevant documents, including information that would not ordinarily be found through keyword searches.

Conceptual search also simplifies the process. The legal team enters a phrase or sentence and the technology organizes corresponding documents into groups of topics and sub-topics available for document review. For example, if a reviewer knows that all documents in a particular folder are related to stock options and all documents in another folder are related to going out to lunch or birthday celebrations within an office, the reviewer will be able to move through the documents with the level of speed and precision needed to make the most efficient decision about whether the document is important to the matter at hand.

Further, conceptual search provides an intelligent information access layer that sits between the data and the person conducting the search. The value of this technology is important because it provides:

- ***Contextual location of data***: Relevant information is retrieved based on context, resulting in better and more informed decisions.

- ***Faster identification of data***: A more advanced understanding of the information is achieved, facilitating faster location of relevant information via better, more accurate search results, which provide quicker decision-making ability.

- ***No other technology needed***: The engine does not require a query language, providing a faster path to productivity with no training required.

- ***Automated application of the technology***: An intelligent layer is created that understands your information and continues to learn, providing the ability to automate decisions without human intervention.

- ***The ability to learn more as data volumes increase***: An intelligent layer that "learns" sits between the business professional and the critical business information, providing accurate and relevant search results as language and terminology change and shift.

Simply stated, conceptual search is the key technology that can facilitate better and faster business decisions in a knowledge economy. Conceptual search provides a mechanism to deliver the right information to the right person at the right time.

Concept search has been available for several years as a tool to help legal and business professionals review data that has already been collected. Concept search can also be used before the document review and production begins for strategic analysis, witness identification, early fact assessment and search term formulation.

Conceptual Search as a Strategic Litigation Tool

When a new investigation or lawsuit begins, US and UK lawyers must start the process of trying to answer the who, what, where, when, why and how questions. Sometimes lawyers have a reasonably good understanding of the people, places and things early in the case, but other times they do not. Rarely, however, will the lawyer possess that knowledge to the degree that allows full early case assessment and a full understanding of who the potential witnesses are and what happened in the case.

Concept search can dramatically improve the speed at which the lawyers develop their case theories, increase the accuracy of the analysis, and decrease the expense of the process. Concept search can help attorneys identify people involved in the dispute, sift through mountains of data and provide an objective, machine-generated group of data with similar context and improve the accuracy of the typical "search-term" approach to data analysis.

Starting with even a limited amount of information about the case, the attorney will be able to identify one or two witnesses who may have knowledge of relevant facts. Through the use of concept search technology, names of other potential witnesses may be dropped into search groupings without requiring the use of search terms, without knowing in advance the names of these individuals, without having to account for misspellings or abbreviations and without having to look at the "to" or "from" lines in email headers. Armed with this information at the beginning of a case, attorneys should more quickly focus on the most important witnesses, even those who are not part of the organization, such as customers, suppliers, competitors and potential wrongdoers.

In addition, having earlier witness identification information will help the legal team ensure that they have preserved data for the right group of custodians. Rather than having to start the data identification process by interviewing each person or by preserving "everything," early use of concept search can help the legal team hone in on who is a potentially important witness. The concept search results can then be fine-tuned with custodian interview and analysis to ensure the preservation plan is complete.

In the early investigation phase of a case, lawyers frequently know very little about the facts. The investigation may start with nothing more than an anonymous call to an ethics hotline or an allegation of potential wrongdoing by a single employee. Through the use of concept search, the legal team can analyze the data of the accused wrongdoer and quickly profile the subject matter of the data. As the legal team rapidly culls out the irrelevant information, the potential facts become clearer and other witnesses emerge as possible subjects of the investigation. In incremental steps, the legal team can then collect data of others and run concept search technology against that data for sorting and grouping. This technology will help the team get a picture of the facts more quickly and more cost-effectively than the typical method of having a team of lawyers plod through every email or try to formulate guesses at search terms to zero in on the issues.

For years, lawyers have tried to develop the perfect set of search terms, the unobtainable objective of which is to find all relevant data while excluding all irrelevant data. Taking an overly narrow approach to search terms results in the team missing relevant data, but taking an overly broad approach will leave the legal team with much more data than it needs to review.

Lawyers spend hours and hours making, refining and fighting about search term lists. Typically, lawyers for the producing party want a small, narrow list, but lawyers for the requesting party want a large, broad list. But no human being is capable of developing a search-terms list that factors into account the taxonomy and lexicon of the data, nor can any human anticipate all of the abbreviations, misspellings, or "code" language intended to deceive that are prevalent in the data. Concept search can help. It has been repeatedly shown that two people will use the same term to describe something less than 20% of the time. In information retrieval this phenomenon is called term mismatch. The impact of term mismatch is amplified when you factor in that most search engine queries are short and many are a single term. Conceptual search smoothes out this issue by analyzing the context of the terms in the corpus of text being searched.

We may be years away from the time that courts and litigations on either side of the Atlantic Ocean use concept search in lieu of search terms to identify relevant data. But concept search can be used today to fine-tune the search term approach to data identification that litigants are comfortable using.

Concept search can be used to group the data before search terms are developed. The grouped data could be reviewed by the producing party and used to develop search terms to propose to the requesting party. The requesting party, on the other hand, could apply concept search technology to a set of production data that had been identified solely by the use of search terms. Analyzing the grouped data, the requesting party could then provide the producing party with additional search terms to apply against the main data collection. Approaching term-based data productions iteratively has always been the most accurate approach. Including concept search technology in this iterative approach makes the process even better.

The issue of Language

Latent Semantic Analysis (LSA) is a statistical approach to information retrieval that is designed to analyze how terms are used in context and measures the correlation between all the terms in the corpus of text being searched. This means that each term is in fact a token and hence the language of the term is irrelevant. What is relevant is how that term / token is used in context with all the terms / tokens in the corpus of text. LSA by its very approach to text analysis and retrieval is language independent and has the ability to learn the relationships between terms in an automatic an unsupervised indexing scheme. Proper parsing and tokenizing of the language (especially in the case of double byte languages such as Chinese and Japanese) is required and the need exists for a well thought out stop word list telling the search engine what terms should not be indexed due to the noise they would create.

LSA does not have any need to analyze parts of speech or sentence structure which natural language processing requires and in so doing makes the statistical approach a better information retrieval solution. When multiple languages are being processed or when cross lingual or multi-lingual documents are present the ability to understand relationships between terms is critical. LSA with its ability to measure the correlation between terms assists information retrieval in environments containing documents with acronyms, abbreviations, slang from the integration of chat like communications within corporate emails, multi-lingual text and documents introducing new and expanding terminology. In litigation events these conditions exist and they present challenges in processing the ESI and properly preparing for the litigation.

Information retrieval challenges in litigation within the European Union are amplified by the numerous languages present in the union with twenty-seven independent states sharing common business interests. Conditions exist that heighten the probability of many search challenges due to language. The critical nature of processing and making available for search ESI in a litigation requires careful consideration of the tools that will be utilized.

The Future for Conceptual Search

One thing is clear, the use of conceptual search and document clustering technologies have been utilized in the litigation process before case law and legal opinions have called for the utilization of advanced search solutions. In the United States this has definitely been the case and the same environment exists around the world. The volume of email and other ESI is a consistent problem regardless of where the litigation is taking place. Legal practitioners will always react to the issues of e-discovery in different ways but will always be a segment that will attempt to get ahead of the problems by using new technologies including advanced search tools.

While many issues in discovery are the same in the US, UK and EU the application of advanced search will need to accommodate differences in the collection and review process, regulations and data protection concerns and the growing likelihood that the litigation will require processing data from different countries encompassing many languages. For instance, the EU countries have data protection laws (Council Directive 95/46/EC, 1995 O.J. (L. 281)31 (EC)) that are drastically different from the US in regards to what is considered personal data and broadly defining processing as including collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction. While search is search and discovery is still discovery, how and when you are able to utilize the advanced tools will differ based upon these regulations. The US does not have the same data protection hurdles to discovery with the courts not receptive to most data protection arguments.

In the UK, The Practice Direction to CPR Rule 31 states that parties should consider electronic documents in a litigation. Lawyers are using technology as a mean to explore and better manage electronic disclosure. Savvy lawyers are positioning themselves as expert in electronic disclosure by exploring advanced discovery tools in the period of time prior to the existence of case law. Conceptual search and document clustering technologies are beginning to be implemented as the UK legal community embraces the challenges associated with managing electronic evidence. As in the US the need to focus data collection and reduce the data lawyers need to review in the initial stages of a case is critical. The reduction of data in the early stages is effective in decreasing the time and cost of processing and reviewing the information. Over time, the integration of advanced tools deeper in the litigation process will improve the discovery task as lawyers learn how to apply technology to a problem that is created by technology.

US, UK and EU litigators and business professionals alike are increasingly relying upon technology, like conceptual search, to do their jobs. As more business and legal professionals collect and exchange ESI for multilingual business, litigation, and regulatory purposes, search technology will continue to improve. No matter the global location, one tenet rings true -- the days of searching through file cabinets to locate information are gone. Instead, search technology has and will continue to become an integral part of the corporate and legal business culture in locating, preserving and exchanging electronically stored information.