LEGAL SEARCH: foundations, evolution and next challenges. The Wolters Kluwer experience

LA BÚSQUEDA DE INFORMACIÓN LEGAL: fundamentos, evolución y próximos desafíos. La Experiencia de Wolters Kluwer

Ángel Sancho Ferrer[1], Carlos Fernández Hernández[2], Pierre Boulat[3]

Artigo recebido em 04 mar. 2014 e aceito em 09 jun. 2014.

Abstract

In this paper, we explain how legal search is different from other search scenarios and why it should be considered as an area plenty of technical and cognitive challenges. We introduce our experience in this field, based on an extensive analysis of legal users search behavior, identification of patterns, and new search functionalities developed in a multinational environment. The state-of-the art in this area is now based on Natural Language Processing techniques, semantic expansion, advanced queries and documents suggestions, advanced cluster extraction, relevance score ranking and displaying of best results. Furthermore, there are still new challenges to be faced in legal search. New solutions will be based in better integration of the algorithms with the content structures, better understanding of users, specific work environment and iterative prototypes.

Keywords
Legal Search. User behavior. Semantic expansion. Relevance Score Ranking. Advanced query and documents suggestion. Innovation. Iterative processes.

[1] Law degree from Universidad de Barcelona (UB). Senior Applied Research Engineer for Search at Wolters Kluwer. Madrid, Spain. E-mail: asancho@wke.es

[2] Law degree from Universidad Autónoma de Madrid (UAM). R&D Project Manager at Wolters Kluwer. Madrid, Spain. E-mail: cafernandez@wke.es

[3] Law degree from Université de Panthéon Assas (Paris II). R&D Analyst at Wolters Kluwer. Madrid, Spain. E-mail: pboulat@wke.es

Resumen

En este artículo explicamos las diferencias entre la búsqueda de información legal y otros tipos de búsqueda y porqué la consideramos un área llena de desafíos técnicos y cognitivos. Presentamos nuestra experiencia en este campo, basada en un exhaustivo análisis de la conducta de búsqueda de los usuarios, en la identificación de patrones y en el desarrollo de nuevas funcionalidades en un entorno de trabajo multinacional. En la actualidad, las técnicas de búsqueda más avanzadas utilizan el proceso del lenguaje natural, la expansión semántica, sugerencias avanzadas de consultas y documentos, extracción avanzada de fragmentos, presentación de resultados por relevancia y presentación de los mejores resultados. Pero más allá de estos adelantos, siguen identificándose nuevos desafíos para la búsqueda legal. Las nuevas soluciones para los mismos se basarán en una mejor integración de los algoritmos con las estructuras de contenidos, una mejor comprensión de la actividad de los usuarios, en la especialización según el marco de trabajo y en prototipos iterativos.

Palabras clave
Búsqueda de información legal. Comportamiento del usuario. Expansión semántica. Presentación de resultados por relevancia. Sugerencias avanzadas de consultas y documentos. Innovación. Procesos iterativos.

1 Introduction

Search of legal information brings up an important set of particularities and technical challenges compared with other kind of search scenarios, as in the general web, an ecommerce or news site or even an intranet.

First, requests are more complex. Legal cases involve not just finding an object (site, person, book, and item) but a similarity of facts and legal consequences (situations). For this reason, legal search usually cannot be performed with just two or three words: it is usually based on a combination of concepts that can have different meanings depending on the context (open texture).

Second, results must be precise, as the implications of a legal analysis require greater responsibility from legal professionals. In many cases, just a simple document may not be enough: a set of fragments of several documents is needed. Also, different sources have very different authority and there is less redundancy of the information.

Finally, good research needs both legal and technical skills. But the majority of legal professionals are not technically oriented, and cannot be expected from them an extra effort to understand a search engine's internal logic.

To solve as much of these problems as possible, some years ago we started to work to replace the traditional legal search approach, which mainly consisted in transposing traditional print-based research techniques to online search.

In the last ten years state-of-the art developments have reached a high level of efficiency in this field, but still new problems and challenges await to be solved to optimize search process.

In this paper, we explain how legal search is different from other search scenarios and that it must be considered as an area plenty of technical and cognitive challenges. . We introduce our experience in this field, our analysis of legal users search behavior and the state-of-the art in this area. Finally we advance the forthcoming challenges to be solved in legal search.

Our survey is an empirical one, both in the analysis and in practice. We've develop our features as we analyze, test and fix. This is not speculative work, but just a practical one. So the bibliographical references included at the end should be considered as a record of our progress together with some helping books we have employed, not as academic references.

This document is therefore structured as follows:

1 How do legal users search?
2 Developments to improve the user's experience.
3 Pending challenges — what is still lying ahead.

2 How do legal users search?

Search engines try to establish a link between the query and some documents of an index that, probably, contain the right answer. This is to say that the query is not just the starting point but a key factor on the quality of the search results — no relevancy or semantic technologies can solve a "bad formulated" query.

In order to create better research systems, the first step is to analyze the information that users enter into the system through keywords in text boxes. Therefore, the better we know

the queries, the better the odds to focus our work correctly in order to solve any search difficulty.

How users build their queries? What type of queries do our customers usually type? Which of them can we possibly solve and which cannot, and what do we need to provide a solution to these difficulties? This is a scarcely explored topic to which we have devoted our recent efforts in Wolters Kluwer, analyzing thousands of Spanish and French logs.

In order to do so during the last two years (2011-2013) we have develop a survey on 5.000 real user's logs on several large pay-per-view databases both from Spain –La Ley Digital (www.laleydigital.es) and Ciss On line (www.cissonline.es) and France –Lamy line (www.lamyline.lamy.fr). Users didn't know that their logs will be going to be analyzed.

We keept anonymized the name and location of the user's but we do identified their field of activity, and so we can determine that those logs came from professional users: mainly lawyers (52%), academia (28%) and public administration (18%), others users reached a 2% of the cases and were ignored.

Our users' requests should be studied through three angles: the external form of the queries, the users' deep interests and the technical solutions needed.

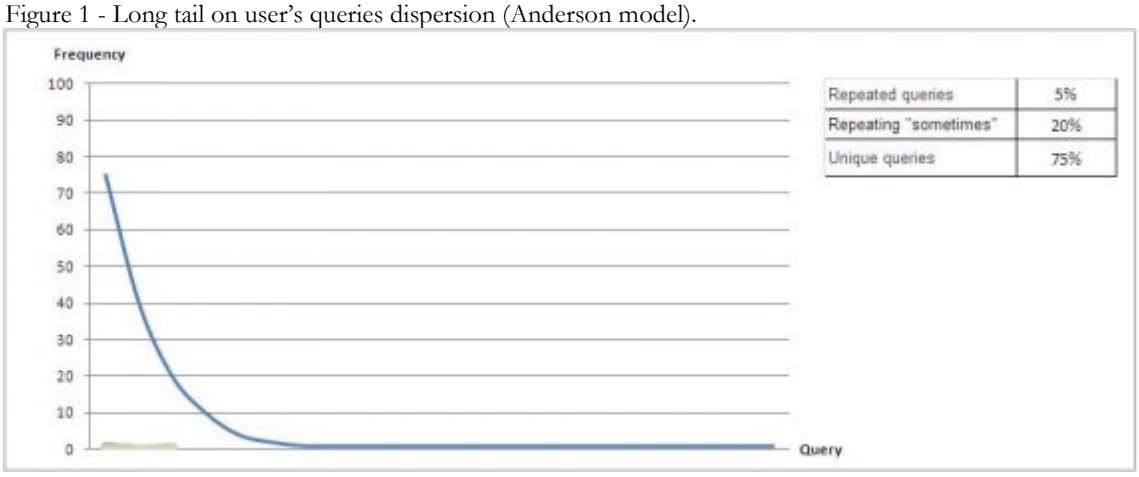
2.1 Legal search from a formal point of view

According to our studies in 80% of cases the information requests that the system has to process is entered in a full text box, with less use of browsing and metadata filtering.

- No Boolean operators are employed
- The average query is typed with two to four concepts (not just words, as many times a set of words expresses a single concept)
- The inclusion of numbers and metadata in the full text box is increasing, as it is reducing the use of the more precise "advanced forms"
- Around 10% of the queries include typos

Indeed, the variety of queries is very large. When analyzing large sets of queries, we found that few of them are repeated, while there are many unique queries (following the classic long tail model defined by Anderson (2004)). (Fig. 1)

Figure 1 - Long tail on user's queries dispersion (Anderson model).



Source: Prepared by the authors.

2.2 From a cognitive perspective

While trying to understand users' intents behind the query, we realized that there is a lot of relevant information which is not provided in the text box.

What is the problem? No information request in the users' mind can be expressed with just a handful of words, so there is always context and expectations that go beyond the text box. Let's imagine that those few words would go to an expert librarian mailbox: most of the time she would answer with another mail asking for more information. Even if query suggestions are becoming more and more accurate (that may help to reduce the problem), new functionalities are needed to answer many nuances that make user's need more explicit.

We have found three broad categories of "information needs" or "user intents".

2.2.1 Queries aimed to retrieve a document

The object of these queries is mainly (a) a whole legislative document, (b) a single part of a legislative document, (c) a case, (d) an authored document, (e) a collective agreement, (f) a form or model and (g) administrative materials.

These queries (that appears up to a 15% of our studies) are often composed through an extensive use of metadata (legislation range, Court names, official numbers, case names,

dates, etc.), or a combination of metadata and concepts, as this is the most common way of identify a document.

Of course, as we'll explain later, suggestions is the first resource to improve the search experience in these cases.

2.2.2 Queries aimed to frame an issue or learning about it

These queries range 40% to 60% of the cases, depending on the type of product – higher for the Advisor's products, lower for the legal products.

They are usually composed of just one or two concepts (a concept can include several words), as they are not intended to go in deep into the issue, but just to get a first or general approach to it.

This terse description of topics usually provide a large set of results, making the user analyze a number of documents.

2.2.3 Queries aimed to investigate on particular aspects of a subject

These queries range 30% up to 50% of the cases, depending on the type of product –lower for the Advisor's products, higher for the legal products.

These queries employ are usually longer, between 3 to 6 concepts. This create shorter results list and so the need of recall is higher. Also queries' suggestions can avoid zero results or false negatives because the chosen words are not the ones corresponding with the documents included in the product.

2.3 From the point of view of the technologies needed

How difficult is each query for the search engine? That is, what requests can be solved by any standard technology out-of-the box, and which ones need tuning (parameters, indexing process) or new components (semantic dictionaries, relevancy, suggestions), or maybe are even beyond the actual state of the art.

Easy queries. This happens when the query terms match the vocabulary of the collection and it includes a number high enough to retrieve very few results. Then, there is no need to use linguistic processing in the query or relevancy analysis in the result list. The more expert the human is, the better his or her guesses of keywords.

Queries that can only be solved by means of a set of specific search functionalities and tuning. When the size of the content and the complexity of the domain increases, it is needed a better processing of the query terms and the lists of results. As we'll explain the next section, recall is improved with things like semantics and precision through better relevancy sorting.

Queries that currently cannot be solved. Even with the latest technical advances, there are queries that lead to no good results. Let's suppose that a human expert would receive that kind of complex query: What would he or she do? We'll come back to this idea in the last section.

3 Latest developments to improve the user's experience

Early techniques in information retrieval tried to reproduce in an electronic display the search methods traditionally employed in paper works (e.g. indexes and taxonomies), but also forcing users to employ the language that the machine could understand (boolean operators). This way, first search screens appeared plenty of metadata boxes and boolean options, hard to use and far away from the average user knowledge. Because of these requirements, search process was difficult, and results tend to be poor.

As said before, despite of all the sophisticated options offered to search, users tend to use a simple full text option with no boolean operators as a more natural way to express their information needs in a database.

And to deal with that mainstream pattern, new search technologies should be put in place, encoding the expertise of the best researchers in the formulation of the query and sorting the results.

3.1 First phase: Natural Language Processing and relevance score ranking (1998-2008)

3.1.1 Natural Language Processing

Thus a first challenge was how to process queries including no operators.

The immediate problem is what pieces of information provided by the user in his query are going to be searchable and how? That drives to solve problems as: (a) what operators should be chosen by default --AND instead of OR? (b) what Stop Words should we employ, considering the legal language where expressions such as "ever", "never", "in some cases", "with", "without", "no"… can play as query conditionings? (c) What

stemmers to be applied in each language so the meaning is retained? (d) What tokenization rules, considering legal numeric formats (official laws numbers, official court cases numbers, others official documents formats)?

Analyzing search logs, in about 75% of them a new obvious pattern revealed the next challenges to deal with: queries are composed by words, but many times a set of words is a unity with its own meaning. For instance: 'Value Added Tax' should be searched as a literal string of words, as it is a concept with an own meaning. A document including separately those words would not be a good result.

3.1.2 Semantic expansion

But just to identify concepts is not enough to encode the expert's knowledge. A concept can be expressed in several ways (for instance, 'Value Added Tax' is very often commonly expressed as 'VAT'), and so, a powerful to-steps way to improve search is, first, to identify concepts in the search and, second, to add to the query of that concept as much synonyms as possible.

It should also include typos, because with just a misspelling technology for query correction it would not retrieve documents with misspelled words in it.

3.1.3 Semantic dictionaries

To identify what sets of words have a meaning just when they are together, and what other words have that same meaning, a dictionary of concepts and expressions (synonym ring) must be curated.

Our experience shows us that the best source of that dictionary is a human build one. This is because no automatic process can, at the same time, add synonyms and avoid the risk of introducing ambiguous synonyms as a human expert does.

3.1.4 Relevance score ranking

In large databases, a big problem arises: any simple query can provide a high number of results, but, of course, not all of them are similarly relevant for a legal user. The "best first" search design principle is needed because 80% of queries just look at the first page, or even the first three results.

How a document can be identified as more relevant? Some clues come from the operators and metadata used by experts to filter and sort the lists. Some content, as the title, contains more meaning, as using operators of proximity and order of the query terms. But also the

idea of "authority" of a document, which under a legal point of view changes per content type and with criteria as ranges of legal documents, court preeminence or certain geographical areas or territorial.

That is, relevancy is not a simple statistical out-of-the-box term frequency calculation, but implies the definition of the applicable legal criteria to codify them into an actionable set of values and weights to be employed by the search engine. This also reveals the importance of the document's analysis, aimed to enrich them with a useful set of metadata.

3.2 Second phase: Advanced query suggestions. Best results (2008-2013)

Once reached a satisfactory level of search quality in that first phase (circa fall 2008), new problems become easily identifiable:

1 A good query is the foundation of an efficient search, thus the more we help users to build good queries, the best results they will get.
2 Any query can display a high number of documents in the results list, in different content types. Even if all of them are properly ordered through an accurate relevance algorithm, is that kind of list really useful for the users?
3 How can we check the quality of our search technologies? How to measure search quality?

3.2.1 Advanced queries' suggestion

Improving relevance algorithms has limits, because it cannot solve a poor quality query. So to improve the search experience we must work in query building components that act before the search engine receives the request. This is like the advice of an expert librarian, which improves the formulation of the information need, with well-chosen terms and in enough detail to lead to good and few results.

This is to say that to offer good suggestions of queries we need to be able to evaluate those that provide these good results. So we need to map a set of queries with an index of documents, in order to eliminate those that are not good enough in that collection.

That set of queries can come from a variety of sources but traditional ones (Thesaurus, semantic dictionary…) are oriented to browsing or query enrichment, and so the most rich and problem oriented are the queries from other users.

Processing those queries needs several steps, to checking them against each index of contents, to eliminate duplicated ones and enrich the set with expressions resulting from a semantic dictionary or a multilevel thesaurus.

It is also of critical importance to establish the score of each suggestion at index and search time.

This score is an expression of the probability of usefulness of each suggestion for a certain query. Is not as simple as the number of times it has been asked. This process will have the side effect of completing the clean-up of the index building with the elimination of queries which do not provide "good" results in each destination product, not just the zero results ones.

The score **at index time** defines a value for the suggestion independent of the query. A prior probability of what the user is looking for or will lead to more useful results.

The score **at search time** takes into account the concrete user's query. For complex information requests, with multi-term queries, factors are not only if the suggestion starts with a term, but if the terms are in order and closer.

3.2.2 Suggestions of documents

An interesting finding came at that point: some queries were obviously searching for concrete laws, so why not to present it directly next to the query, thus avoiding the result list step? Suggestions of documents pose additional challenges, as mixing content types or having a very reliable measure of its authoritativeness.

3.2.3 Best Results (iReport or Best of)

In professional databases results are classified in tabs indicating sources or content types (i.e. laws, judicial opinions, authored comments, forms). Large results lists forces professionals to invest many time in the analysis and filtering of those candidates trying to find the most convenient documents to build an argument.

This is why it is worth to go a step beyond and take the risk of trying to identify just the very best results of a list and display them in a different way, as a new document, ready to be used, printed or emailed.

The algorithms analyze the value of each content type for that customer, what is a minimum threshold of quality (something that cannot be based on the unreliable simple

TF*IDF score), and the best fragments, so the results are presented like a short structured report.

4 Next challenges

Despite all those developments, still there are pending scenarios to optimize. Some of them are not actual complains but hidden. This is what make us think that in the next years the professional search experience will change again.

4.1 Bad queries

Of course, an obvious pattern to analyze are the zero results queries, when no document in the collection has matched the terms of the query, and also the less obvious one of the results with bad quality, despite all the terms are, of course, in the documents.

But also there are false positives when there are too many good results, and so the user has still too many documents to read.

In most of those cases, an expert librarian would be able to retrieve useful information, reducing or changing the result set by adding, changing or eliminating words or filters.

So a first important area to investigate is how users reformulate to guess a better query.

4.2 Good answers

Another important area to analyze is what can be "an answer".

The actual paradigm is a large result set structured by content type (tabs) and with facets to help in the filtering. This was an important advance to organize large amounts of documents with different authority and create a first form of dialog with the content set.

But, in some cases, a small set of documents can be enough. This is why authored content (yearbooks, manuals, forms) on paper still solve some kinds of information needs.

Even, as we explained before, just one document can be enough, because this was the level of the request.

Also the snippets (or keyword in context or query-oriented summaries) is an important tool too as "in many cases, an information need can be satisfied by viewing the document

surrogate alone" (HEARST, 2009, *online*) if the extracted sentence is readable and contains all the words of the query.

Sometimes the answer is just a fact, and the question is a natural language one, like it happens in Google Graph, Wolfram Alpha, Siri or Watson.

5 Conclusions

We don't envision legal search in the medium term to be transformed into a computational engine, but it will be enriched with internal inference engines and external dialogs that go beyond the reactive "photocopy and highlight" limited capabilities of the actual search engines.

We think that professional searches are still not solved and so in three to five years will be something different. The main forces that will allow the development of these developments need the following:

1 **Search technologies that are better integrated with the content structures**. So a better exploitation of the core assets of Wolters Kluwer can be done (moving to the target of "computable knowledge").
2 **Better understanding of users**. New customer insights can be obtained by mining the query logs assets both with big data techniques and human evaluation (how an expert would have given a better result than the search engine).
3 **Specific work environment**. Not only is needed a better integration between different departments (analysis area, development teams, architecture, subject matter experts …) but also a legal search need team with mixed profiles, composed by software engineers and legal experts. These multidisciplinary teams took an integrated approach to the solution of the problems, exploring a wide range of points of view, to deliver knowledgeable and experienced answers. This cross-learning allows experts and technicians to suggest new useful ideas, as well as a quick process to discard others. In a multidisciplinary team, any expertise can separately solve the problem, because the research continuously requires a variety of influences from the other experts and knowledge fields.
4 **Iterative prototypes**. This method allows to learn what works from the UX and technology perspectives and to improve it at each iteration, instead of just using process oriented methodologies where each area completes its tasks (requirements, components) and delivers them as an input to the next phase in a well-defined and closed process. Because as said Scott Berkun (2009, *online*) "requirements are not a

> product […] The design/engineering team know first-hand all the stupid things requirements often include since they've been forced to experience them before. Want to avoid these stupid mistakes? Get their input early."

Search is, still, a huge unexplored topic, which can partly be solved by applying big data techniques to analyze users' logs, in combination with cognitive models, including strengthening the relevance and natural language algorithms, content tuning, as well as new interfaces in order to adapt search engine to the natural workflow.

6 Bibliography

ANDERSON, C. The Long Tail. **Wired**, 12 oct. 2004. Available at: <http://wrd.cm/1o3NZrH>. Accessed on: 25th mar. 2014.

BERKUN, S. Why requirements stink. **Scott Berkun**, 25th mar. 2009.Available at: <http://bit.ly/1kKm8aj>. Accessed on: 25th mar. 2014

HEARST, M. **Search user interfaces**. Cambridge: Cambridge University Press, 2009. Available at: <http://searchuserinterfaces.com**>.** Accesed on: 25th mar. 2014.