Searching Political Data by Strategy

Roberto Cornacchia
Spinque
roberto@spinque.com

Jaap Kamps
Univ. Amsterdam
kamps@uva.nl

Wouter Alink
Spinque
wouter@spinque.com

Arjen P. de Vries
CWI
arjen@acm.org

Abstract

Professional search could benefit significantly from advanced information retrieval techniques. However, current search systems fail in the matching between the conceptual level of the professional information needs and the data processing level of the available search technologies. Search by strategy is a novel paradigm that puts the modelling phase of complex search paths at a central spot. We illustrate some of the key advantages of this paradigm in a realistic use case: searching semantically enriched political data.

1 Introduction

Information professionals can be found everywhere, from industrial research departments finding the latest breakthroughs in technologies, to local librarians helping young readers find their favourite books, and from research journalists doing background checks on breaking news stories, to individual heritage professionals analysing events in years that have long since passed. Their activity is also referred to as *professional search* and is different from incidental or general user-oriented search in the following ways: the person behind the search is willing to invest effort and time in the process and can provide the system with high-quality feedback; she has a good understanding of the search domain and of the search target; it is often necessary to connect several requests together into a multi-step search process; high recall in results tends to be important; documenting the exact search process is sometimes even more important than results themselves (e.g. in e-discovery or evidence based methods).

Information professionals tend to ignore advanced search technologies, often because it is hard to relate them to their own perception of a search task. In practice, expert users have been shown to face complex information needs where search effectiveness can be improved when queries include both natural language and structure [KK10]. However, users tend to have difficulty to use query languages to transform their information needs into a proper semi-structured query [KMRS06]. Hence there is a need for supporting searchers during their whole task or search episode – both when interactively constructing a complex query, and when interactively exploring the result space [dV11].

Search by Strategy is a novel approach where search is adapted to the requirements of professionals – not the other way around. It allows domain specialists to implement their own complex and multi-step search strategies, in a way that is powerful but easy to analyse, justify and document. This high level of control (the user remains in charge) may persuade information professionals to experiment with (and share with their colleagues) new approaches to solving their information seeking tasks.

In: M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg (eds.): Proceedings of the Integrating IR technologies for Professional Search Workshop, Moscow, Russia, 24-March-2013, published at http://ceur-ws.org

2 Search by Strategy

Search by Strategy is an iterative 2-stage search process that separates search strategy definition (the *how*) from actual searching and browsing of data collections (the *what*). A visual, graph-based query environment, the Strategy Editor (see Figure 1), allows search professionals to abstract away technical details and express their domain knowledge as high-level search strategies. Modelling domain-specific search strategies in this framework corresponds to designing graph structures, where edges represent data-flows consisting of any kind of object, including for instance patent documents, medical reports, images, abstracts, paragraphs, speech-segments, companies, locations, keywords. The nodes connected by such edges are pre-defined (though extensible), general-purpose operational blocks, that either provide a data source or modify their input data-flow applying operations such as selections, ranking, retrieval of semantically related objects, data fusion, and others. Search strategies defined in this framework can be inspected, optimised, saved, and eventually published as independent web-based search engines. All the data operations involved in the query process that a strategy describes are automatically translated into a probabilistic relational query language and executed on top of an SQL database engine.

2.1 Arbitrary retrieval units - we need more than just documents

In information retrieval, especially in its most traditional forms related to full-text search, a document is the smallest piece of information being indexed, searched and returned to users. Even with the advent of multimodal and multimedia information retrieval, only few exceptions such as XML retrieval have explored how to search and present sub-document retrieval units. Expert finding and entity search are examples of search processes focusing on different concepts than documents.

This particular facet of search flexibility is at the core of Search by Strategy. Take an example multi-step search strategy: "*find patents by academic researchers who also worked in companies that are active in research field X or that hold significant prior-art patents on topic Y*". It is not unusual for patent specialists to approach such complex tasks by issuing sub-queries to (possibly distinct) systems and combining the respective outputs and inputs manually (i.e., performing a 'human join operation'). In Search by Strategy, the whole process can be described as a strategy, where intermediate retrieval units include patents, people, companies, universities, and each can be inspected independently by activating the respective sub-graph. Also, refining the document concept allows to apply different search options to smaller units. For example, multi-language documents such as patents can be searched using different language options for each document section (e.g. search English abstract and German claims simultaneously).

2.2 Semantic search

The shift from a traditional document-oriented view to a more generic interpretation of data is supported by searching semantic annotations that describe properties of and relations among data objects. Such annotations can be imported from external sources or generated by corpus-specific indexing templates and are stored in the index as probabilistic triples. Imagine a system analysing PDF files for their content, outputting the following probabilistic triples to express its uncertainty about the document's language but certainty about its page count:

```
0.9 (this paper, language, English)
1.0 (this paper, pages, 4)
```

When designing a strategy, the building blocks that process such triples take those probabilities into account, affecting all the subsequent results. This feature is crucial for the construction of multi-step strategies. For example: first find patents based on keywords; from those, find inventors by following the predicate `inventor_of`; find companies by following the predicate `works_at`, and so on and so forth.

2.3 Semi-structured search

Data objects may be structured in hierarchies. Taking inspiration from semi-structured information retrieval, we store such hierarchies using pre-post encoding schemas [Gru02] and this makes it possible to perform structural queries from strategy blocks. Consider for example a discussion with multiple speakers in which individual utterances can be grouped into larger (sub-)topics. Relevance to a query of smaller utterances can be propagated easily to ancestor (sub-)topics using hierarchical information. Spinque has recently applied the Search by Strategy approach to a project involving 200 years of Dutch parliament proceedings (ExPoSe: Exploratory Political Search).

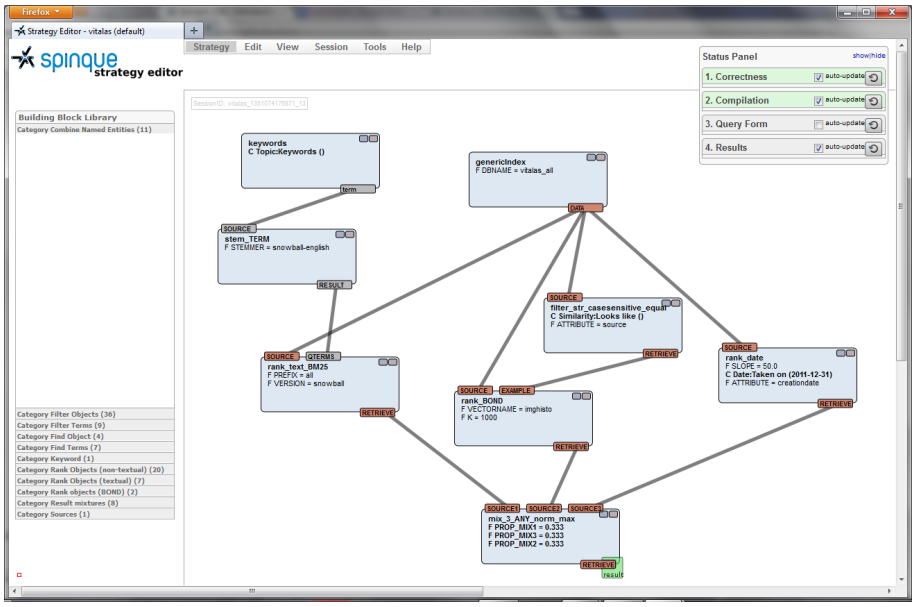

Figure 1: A strategy being designed in the Strategy Editor web application.

2.4 Exploratory search

Exploratory search, mainly implemented in terms of faceted browsing, can be considered for many aspects the opposite of professional search, as it is mostly used by searches who are unfamiliar with the domain at hand and/or with their goal. However, search professionals can benefit from the way exploratory search can be facilitated using the integration of databases and information retrieval. Firstly, fuzzy facets (ranking) are used as a generalisation of Boolean facets (filtering), which yields better search experiences in particular in complex scenarios. Secondly, each applied facet results in a search strategy being implicitly constructed or extended. Therefore, exploratory search can become an alternative way of designing new strategies: by browsing the data rather then by explicitly thinking of a search process.

3 Search political data by Strategy

Parliamentary proceedings and other transcripts of meetings are a common document genre characterised by a complex narrative structure. The essence is not only what is said, but also by who and to whom, and why. The ExPoSe (Exploratory Political Search) project aims at making semantically enriched political data (`http://politicalmashup.nl/`) searchable using novel exploratory search methods. The corpus contains 200 years of the Dutch parliament proceedings, where all topics, speeches and interruptions are hierarchically stored in an open XML standard. Each speech or interruption is labelled with the name and parliamentary role of the speaker. We have developed a modern, easy-to-use search system which allows users to express complex requests on this semantically rich corpus, by using Spinque technology.

3.1 The '*knettergek*' case

Let us focus on an interesting case from 2007. During one of the parliamentary sessions, the leader of a large populist party addressed a minister with an offensive expression – "Zij is *knettergek* geworden" ("She has gone *completely crazy*") – in response to her opinions. This caused a major discussion about the decline of standards in parliament and the case was closely followed by the national press. Journalists writing about this case are a good example of search professionals. They may want to investigate whether the term '*knettergek*' had ever been used in parliament before. If so, who else used it? In what context and addressing whom?

Let us see how to answer these questions using the Spinque-powered search system of the ExPoSe project. We start by performing standard text ranking on all the parliamentary proceedings, using the keyword '`knettergek`'. Figure 2 shows a small sample of the result set. Relevant fragments of the proceedings are returned, with the possibility of navigating their inherent hierarchical structure (**semi-structured search**) to inspect larger or smaller utterances. The blue arrow in the figure depicts that the second result is a reply to a previous speech,

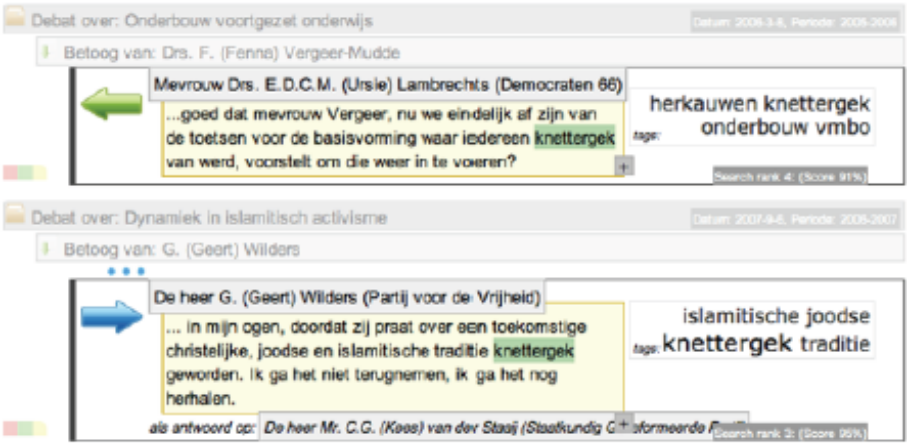

Figure 2: Who uses offensive language in the Dutch parliament?

which we can click and inspect. To follow this semantic link corresponds to implementing the **semantic search** approach outlined in Section 2. Identifying how often the term '*knettergek*' is used per political party, person, or year is a clear use case for **exploratory search** in the form of (fuzzy) faceted search. Notice that this implicitly corresponds to the construction of a more complex search strategy under the hood. Once relevant speech portions are identified, we may request a list of people or political parties as result, which illustrates the importance of allowing **arbitrary retrieval units** as discussed in Section 2.

With the system retrieving focused information from the proceedings, we were able to find out quickly that the term '*knettergek*' has actually been used in parliament by several people from several political parties, across several years. However, it was likewise easy to establish that only once this utterance had been used as an direct characterisation of a member of the government. One could imagine further questions, for example about the relationship between the use of the term '*knettergek*' and the details of the person saying it, such as age or place of birth/residence. To join two separate corpora with parliament proceedings and bio information is a simple operation when designing search strategies.

4 Summary

We have briefly outlined the Search by Strategy approach in the context of professional search and presented a use case where journalists can easily formulate complex requests on top of semantically enriched political data. The main novelty of this approach is that it emphasises flexible search modelling as the key for empowering search professionals to use sophisticated technology. Search professionals "think in terms of strategies" already: they need tools to put their own thoughts in practice. Particular attention is being put in making this flexible framework more and more user-friendly, with a steadily growing 'block library' for the implementation of more best-practice and experimental IR tasks. Indexing is also an active area of development, moving more and more towards self-organising database storage, fast on-the-fly indexing and better query performance.

Acknowledgments We thank Maarten Marx for the discussions on the system and the specific use case.

References

[dV11] Arjen P. de Vries. What to do when one size does not fit all? In *Proceedings of the fourth workshop on Exploiting semantic annotations in information retrieval*, ESAIR '11, pages 1–2, New York, NY, USA, 2011. ACM.

[Gru02] Torsten Grust. Accelerating XPath location steps. SIGMOD'02, pages 109–120, New York, NY, USA, 2002. ACM.

[KK10] Marijn Koolen and Jaap Kamps. Searching cultural heritage data: does structure help expert searchers? RIAO '10, pages 152–155, Paris, France, 2010. CDI.

[KMRS06] Jaap Kamps, Maarten Marx, Maarten de Rijke, and Börkur Sigurbjörnsson. Articulating information needs in XML query languages. *ACM Trans. Inf. Syst.*, 24(4):407–436, October 2006.