An Open-Access Platform for Transparent and Reproducible Structured Searching

Tony Russell-Rose
UXLabs Ltd
Centre Park, Warrington, UK
tgr@uxlabs.co.uk

Jon Chamberlain
School of Computer Science and Electronic Engineering,
University of Essex
Colchester, Essex, UK
jchamb@essex.ac.uk

ABSTRACT

Knowledge workers such as patent agents, recruiters and legal researchers undertake work tasks in which search forms a core part of their duties. In these instances, the search task often involves formulation of complex queries expressed as Boolean strings. However, creating effective Boolean queries remains an ongoing challenge, often compromised by errors and inefficiencies. In this paper, we demonstrate a new approach to structured searching in which concepts are expressed as objects on a two-dimensional canvas. Interactive query suggestions are provided via an NLP services API, and support is offered for optimising, translating and sharing search strategies as executable artefacts. This eliminates many sources of error, makes the query semantics more transparent, and offers an open-access platform for sharing reproducible search strategies and best practices.

CCS CONCEPTS

• **Information systems** → **Query representation**; **Search interfaces**; *Expert search*; • **Computing methodologies** → *Representation of Boolean functions.*

KEYWORDS

query formulation, advanced search, Boolean, search visualisation, professional search

ACM Reference Format:
Tony Russell-Rose and Jon Chamberlain. 2019. An Open-Access Platform for Transparent and Reproducible Structured Searching. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331394

1 INTRODUCTION

Many knowledge workers rely on search applications in the course of their professional duties. Patent agents, for example, depend on accurate prior art search as the foundation of their due diligence process [15]. Similarly, recruitment professionals rely on Boolean search as the basis of the candidate sourcing process [8], and media monitoring professionals routinely manage thousands of Boolean expressions on behalf their client briefs [16].

The traditional solution is to formulate complex Boolean expressions consisting of keywords, operators and search commands, such as that shown in Figure 1. However, the use of Boolean strings to articulate complex information needs suffers from a number of shortcomings [8]. First, they are poor at communicating structure: without some sort of physical cue (such as indentation), parentheses and other operators can become lost among other alphanumeric characters. Second, they scale poorly: as queries grow in size, readability becomes progressively degraded. Third, they are error-prone: even if syntax checking is provided, it is still possible to place parentheses incorrectly, changing the semantics of the whole expression.

(cv OR "cirriculum vitae" OR resume OR "resumé") (filetype:doc OR filetype:pdf OR filetype:txt) (inurl:profile OR inurl:cv OR inurl:resume OR initile:profile OR intitle:cv OR initile:resume) ("project manager" OR "it project manager" OR "program* manager" OR "data migration manager" OR "data migration project manager") (leinster OR munster OR ulster OR connaught OR dublin) -template -sample -example -tutorial -builder -"writing tips" -apply -advert -consultancy

Figure 1: An example from the Boolean Search Strings Repository

To mitigate these issues, many professionals rely on previous examples of best practice. Recruitment professionals, for example, draw on repositories such as the Boolean Search Strings Repository[1] and the Boolean String Bank.[2] However, these repositories store content as unstructured text, and as such their true value as a source of reproducible templates may never be fully realized.[3]

2dSearch[4] offers an alternative approach in which information needs are expressed by combining objects on a two-dimensional canvas. Query suggestions are provided via an NLP services API, and support is offered for optimising, translating and sharing search strategies as executable artefacts. This eliminates many sources of error, makes their semantics more transparent, and offers an open-access platform for sharing reproducible search templates and best practices.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331394

[1] https://booleanstrings.ning.com/forum/topics/boolean-search-strings-repository, accessed 10 Oct 2018.
[2] https://scoperac.com/booleanstringbank, accessed 10 Oct 2018.
[3] http://booleanblackbelt.com/2016/01/the-most-powerful-boolean-search-operator, accessed 10 Oct 2018.
[4] https://2dsearch.com, accessed 24 Feb 2019.

2 RELATED WORK

The application of data visualisation to search query formulation can offer significant benefits, such as fewer zero-hit queries, improved query comprehension and better support for exploration of an unfamiliar database [4]. An early example is that of Anick et al. [1], who developed a two-dimensional graphical representation of a user's natural language query that supported reformulation via direct manipulation. Jones [5] developed a query interface to the New Zealand Digital Library which uses Venn diagrams and query result previews.

Nitsche and Nürnberger [7] developed a system based on a radial user interface that supports interactive visual refinement of vague queries. A further example is Boolify[5], which provides a drag and drop interface to Google. More recently, de Vries et al [3] developed a system which utilizes a visual canvas and building blocks to allow users to graphically configure a search engine. Similarly, Scells and Zuccon [14] developed a platform to edit and explore Boolean queries using a tree visualization, based on a common representation [13]. *2dSearch* differs from the prior art in offering:

- A visual framework which eliminates many errors associated with traditional command-line query builders;
- Search results that update in real-time, and individual blocks with hit counts that can be enabled/disabled on demand;
- Queries that are analysed and validated, with common errors (e.g. duplication, orphaned lines, redundant bracketing) detected and corrections offered;
- A platform-agnostic representation and support for multiple databases which mitigates inefficient 'translation' of search syntaxes across databases;
- Interactive query suggestions that avoid the problems of phrase boundary detection and 'query drift' that undermine traditional query expansion techniques;
- Outputs that may be shared as executable artefacts or exported as traditional Boolean strings.

3 A NEW APPROACH

3.1 Building search strategies

At the heart of *2dSearch* is a graphical editor which allows the user to formulate search strategies using a visual framework in which concepts are expressed as objects on a two-dimensional canvas. Concepts can be simple keywords or attribute:value pairs representing controlled vocabulary terms (e.g. Mesh terms) or database-specific search operators (e.g. field codes and other commands). They can be combined using Boolean (and other) operators to form higher-level groups and then iteratively nested to create expressions of arbitrary complexity.

The application consists of two panes (see Figure 2): a query canvas and a search results pane (which can be resized or detached). The canvas itself can be resized or zoomed, and includes an 'overview' pane which allows the user to navigate to elements outside the current viewport. A sliding menu is offered on the left, providing file I/O and other options. This is complemented by a navigation bar which provides support for document-level functions such as naming and sharing search strategies.

[5]https://www.kidzsearch.com/boolify/, accessed 23 Oct 2018

Although *2dSearch* supports the creation of complete strategies from a blank canvas, its function and value are most readily understood by reference to an example (i.e. text-based) search strategy, such as that shown in Figure 1. A trained professional may be able to mentally 'parse' the expression and interpret the general approach, but without associated documentation it is difficult to understand exactly what the searcher intended let alone optimise, debug or re-use strategies expressed in this form.

However, when opened using *2dSearch*, its structure becomes much more apparent (see Figure 2). It can be seen that the overall expression consists of a conjunction of five OR clauses (the pale blue blocks), with a number of field tags (dark blue) and various negated terms (white on black). By displaying them as nested groups, *2dSearch* offers support for abstraction in which lower-level details can be progressively hidden, and meaningful names given to sub groups so that they can be more effectively reused.

To edit the expression, the user can move terms from one block to another, or create new groups simply by combining terms. They can also cut, copy, delete, and lasso multiple objects. If they want to understand the effect of one block in isolation, they can use its hit count (Figure 3) or execute it individually. Conversely, if they want to remove one element from consideration, they can temporarily disable it. In each case, the effect of each change is displayed in real time in the adjacent search results pane.

Crafting syntactically correct search expressions can be an error prone and tedious process [2]. Line numbers, parentheses, square brackets, punctuation, whitespace characters and Boolean operators all have the potential for errors. However, by using a visual representation, the task of generating syntactically correct expressions can instead be delegated to system-level functions.

3.2 Optimising search strategies

2dSearch functions as a meta-search engine, so is agnostic of any particular search technology or platform. To execute a given query and retrieve results, the semantics of the canvas content must be mapped to the syntax of the underlying database. This is achieved via an abstraction layer or set of 'adapters' for search platforms such as Bing, Google, PubMed, Google Scholar, Epistemonikos, TRIP Database, etc. These are user selectable in the interface via a drop-down control.

It is common for healthcare information professionals to want to search more than one database, particularly when undertaking a systematic literature review [9]. In practice, this requires a process of 'translation' of the search strategy to match the syntax of the target database and the search operators it supports. For a relatively simple query this may not be a major undertaking, particularly if such operators form a relatively small proportion of the overall search strategy. However, the user still has to understand which elements are platform-specific, identify the closest equivalent in the other database and manually edit their query, all of which is laborious and time consuming [2].

Support for query translation is provided via a 'Messages' tab on the results pane (Figure 4). For example, if the user tries to execute via Bing a query string containing operators specific to Google, an alert is shown listing the unknown operators. *2dSearch* also identifies redundant structure (e.g. spurious brackets or duplicate

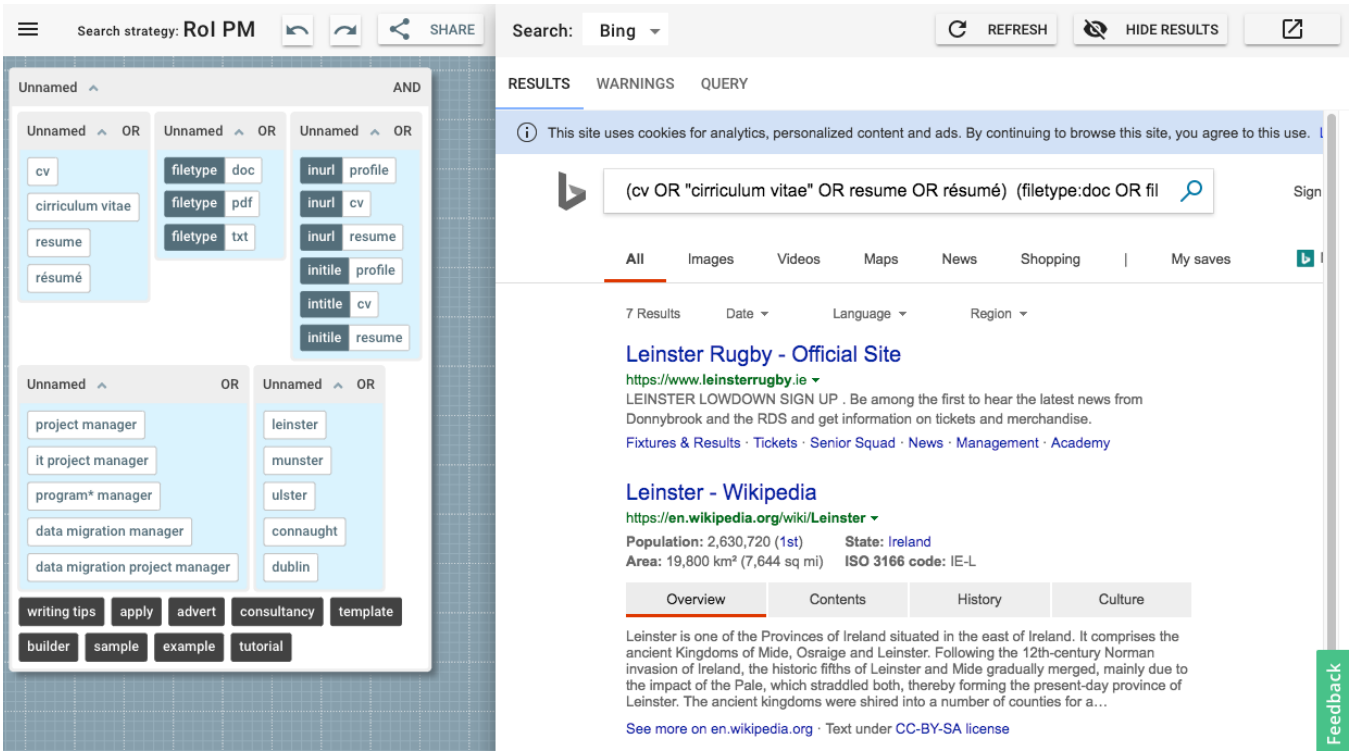

Figure 2: The *2dSearch* app showing the two-dimensional canvas (left) and the search results pane (right).

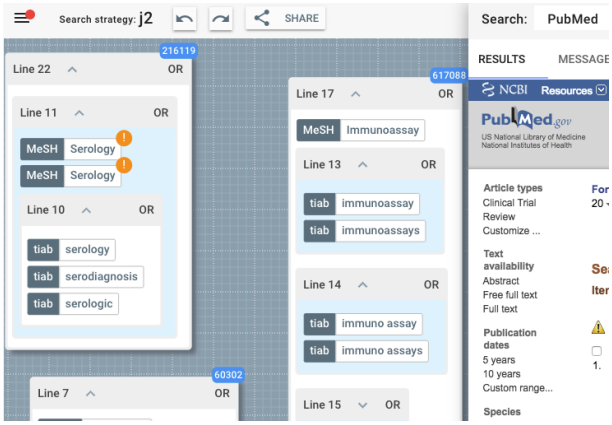

Figure 3: Support for error checking and hit counts

elements) and supports comparison of canonical representations. Query suggestions are provided via an NLP services API (Figure 5) which utilises various Python libraries for word embedding, keyword extraction, etc. and SPARQL endpoints for linked open data ontology lookup [11].

In developing a search strategy, information professionals will often create alternative versions to understand the contribution of individual search blocks and to find the best combination. Support for enabling/disabling search blocks, real-time synchronisation, and automated term suggestions make it possible to iteratively refine a strategy without needing to maintain multiple versions.

3.3 Working with existing search strategies

Many information professionals routinely save their search strategies (which can be the results of hours of work) as text documents within their local file system [10]. Various attempts have been made to provide a central repository for such artefacts (e.g. MedTerm Assist [12]), but the practice has not yet been widely adopted. Similarly, some health information specialists use the PRESS forum for peer-reviewing search strategies [6], but this is a private forum which uses Microsoft Word as the medium of exchange. Even if shared as PDF, content copied from such files can lose vital non-print characters, introduce spurious line breaks, and be altered by auto-correct and other unwanted transformations. By contrast, searches in *2dSearch* can freely distributed as hyperlinks generated using the 'Share' button (see Figure 3). In this way, 2dSearch offers an open access, online platform for saving, sharing and executing search strategies as validated, reproducible artefacts.

4 SUMMARY AND FURTHER WORK

2dSearch is a platform for search strategy formulation in which information needs are expressed by manipulating objects on a two-dimensional canvas. Transforming logical structure into physical structure mitigates many of the shortcomings of Boolean strings. This eliminates syntax errors, makes the query semantics more

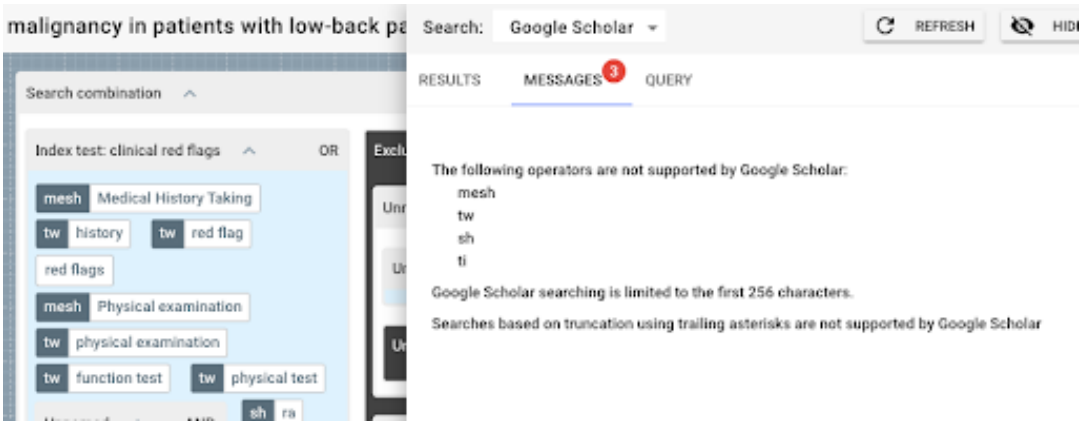

Figure 4: Support for query optimisation and translation.

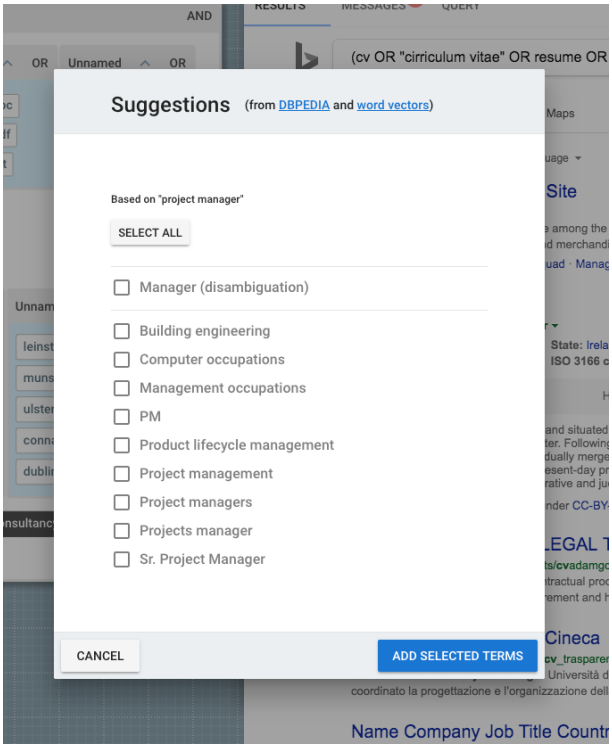

Figure 5: Interactive query suggestions

transparent and offers an open-access platform for sharing reproducible search templates and best practices.

Adopting a database-agnostic approach presents challenges, but it also offers the prospect of a *universal* framework in which information needs can be articulated in a *generic* manner and the task of translating to an underlying database can be delegated to platform-specific adapters. This could have profound implications for the way in which professional search skills are taught, learnt and applied.

REFERENCES

[1] P. G. Anick, J. D. Brennan, R. A. Flynn, D. R. Hanssen, B. Alvey, and J. M. Robbins. 1990. A Direct Manipulation Interface for Boolean Information Retrieval via Natural Language Query. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '90)*. ACM, New York, NY, USA, 135–150.

[2] W. M. Bramer, M. L. Rethlefsen, J. Kleijnen, and O. H. Franco. 2017. Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study. *Systematic Reviews* 6, 245 (2017).

[3] A. P. de Vries, W. Alink, and R. Cornacchia. 2010. Search by strategy. In *Proceedings of the third workshop on Exploiting semantic annotations in information retrieval*. ACM, 27–28.

[4] J. H. Goldberg and U. N. Gajendar. 2008. Graphical condition builder for facilitating database queries. *U.S. Patent No. 7,383,513. 3* (2008).

[5] S. Jones. 1998. Graphical Query Specification and Dynamic Result Previews for a Digital Library. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (UIST '98)*. ACM, New York, NY, USA, 143–151.

[6] J. McGowan, M. Sampson, D. M. Salzwedel, E. Cogo, V. Foerster, and C. Lefebvre. 2016. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *Journal of Clinical Epidemiology* 75 (2016), 40 – 46.

[7] M. Nitsche and A. Nürnberger. 2006. QUEST: Querying Complex Information by Direct Manipulation. *In: Yamamoto S. (eds) Human Interface and the Management of Information. Information and Interaction Design. HIMI 2013. Lecture Notes in Computer Science* 8016 (2006).

[8] T. Russell-Rose and J. Chamberlain. 2016. Searching for talent: The information retrieval challenges of recruitment professionals. *Business Information Review* 33, 1 (2016), 40–48.

[9] T. Russell-Rose and J. Chamberlain. 2017. Expert Search Strategies: The Information Retrieval Practices of Healthcare Information Professional. *JMIR Med Inform* 5, 4 (2017), e33.

[10] T. Russell-Rose, J. Chamberlain, and L. Azzopardi. 2018. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[11] T. Russell-Rose and P. Gooch. 2018. 2dSearch: A Visual Approach to Search Strategy Formulation. In *Proceedings of DESIRES: Design of Experimental Search & Information REtrieval Systems (DESIRES 2018)*.

[12] A. A. Saleh, M. A. Ratajeski, and J. LaDue. 2014. Development of a Web-based repository for sharing biomedical terminology from systematic review searches: a case study. *Medical reference services quarterly* 33, 2 (2014), 167–178.

[13] H. Scells, D. Locke, and G. Zuccon. 2018. An Information Retrieval Experiment Framework for Domain Specific Applications. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1281–1284.

[14] H. Scells and G. Zuccon. 2018. searchrefiner: A Query Visualisation and Understanding Tool for Systematic Reviews. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM.

[15] J. I. Tait. 2014. An introduction to professional search. In *Professional search in the modern world*. Springer, 1–5.

[16] J. Wing Pazer. 2013. The importance of the Boolean search query in social media monitoring tools. *DragonSearch white paper* (2013). https://www.dragon360.com/wp-content/uploads/2013/08/social-media-monitoring-tools-boolean-search-query.pdf (retrieved 22-Mar-2018).