MINING MEDLINE FOR THE VISUALISATION OF A GLOBAL PERSPECTIVE ON BIOMEDICAL KNOWLEDGE

João Pita Costa,
Quintelligence, Ljubljana, Slovenia

Luka Stopar, Flavio Fuart, Marko Grobelnik
Jožef Stefan Institute. Liubliana

Raghu Santanam, Chenlu Sun
Arizona State Universitv. USA

Paul Carlin
South Eastern Health and Social Care Trust, UK

Michaela Black, Jonathan Wallace
Ulster Universitv. UK

1 THE PROJECT

There is an ever increasing number of data sources that potentially could be used to gain new insights into areas such as disease prevention, policy formulation/ evaluation and personalised medicine, but these are not optimised for use within an analytics type user interface. The MIDAS project was funded under a call for 'Big Data supporting Public Health policies' to develop a big data platform that facilitates the utilisation of healthcare data beyond existing isolated systems, making that data amenable to enrichment with open and social data [1]. This aligns closely with a number of themes in Knowledge Discovery in Databases (KDD) in that the platform enables the integration of heterogeneous data sources, providing privacy-preserving analytics, forecasting tools and visualisation modules to deliver actionable information. Policy makers as a result will have the capability to perform data-driven evaluations of the efficiency and effectiveness of proposed policies in terms of expenditure, delivery, wellbeing, and health and socio-economic inequalities, thus improving current policy formulation, delivery risk stratification and evaluation. This H2020 project has a total of 15 partners from 5 EU countries as well as Arizona State University (ASU). The partners are Universities, SMEs and health departments in governmental institutions.

2 THE DATA SET

The day-to-day growth of knowledge available online mandates the need to be assured that information sources are complete and reliable. The state-of-the-art in medical research is aggregated and accessible in MEDLINE [2], through a single searchable platform - PubMed – providing access to references and abstracts on life sciences and biomedical topics. This open data source is frequently used by clinicians and researchers to provide an overview of a certain topic using several filters, tags and advanced search options. MEDLINE incorporates more than 27 million records dating from 1946 to the present day, drawn from more than 5,200 journals worldwide, in about 40 languages. It includes 443218 full-text articles with the key-words string "public health" included. This open dataset includes a comprehensive controlled vocabulary – the Medical Subject Headings (MeSH) – which indexes journal articles and books in the life sciences. It can contribute to scientific literature reviews before carrying out research in a specific topic area. MeSH is composed of 16 major categories that further subdivide from general to more specific in up to 13 hierarchical levels. This rich data structure is annotated by human hand, assisted by semi-automated tools, and therefore is not available in the most recent citations.

3. THE CHALLENGE

The richness of such a complete data source, as MEDLINE is, brings challenges, particularly in the efficient search and choice of appropriate scientific knowledge. Although powerful, PubMed does not provide suitable tools for in-depth analysis and presentation of scientific information. MIDAS aims to enable advanced visualization techniques to support public health policy making, and thus a suitable MIDAS MEDLINE repository had to be developed. It has to allow exploration of a wide range of different visualisation techniques in order to evaluate their applicability to policymaking tasks. Hence, the need for a selection of a powerful, semi-structured text index, that allows free text searches, but also the creation of complex queries based on available metadata, based on Elasticsearch. It combines features provided by no-SQL databases with standard full text indexes, as it is based on the Apache Lucene Index. Kibana is the data visualisation dashboard of choice for Elasticsearch-based indexes used in the context of MIDAS for fast prototyping and support of use-cases. This tool enables one to query large datasets and produces different types of visualization modules that can be later integrated into customised dashboards. The flexibility of such dashboards permits the user to profit from data visualisations that feed on his/her preferences, previously set up as filters to the dataset. Furthermore, the MIDAS MEDLINE tools presented in this paper permit the user to explore the potential of the MEDLINE dataset, based on enhanced UX representations that are easy to understand and to communicate.

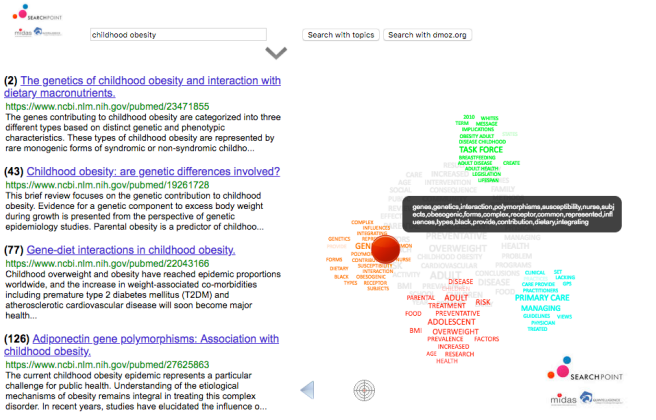

Figure 1. Screenshot of the MEDLINE data visualization tool after a query on "childhood obesity", exhibiting several clustered keywords representing areas of interest that the user can focus on by moving to it the red searchpoint and reindexing the list of related citations positioning in top 5 an article that otherwise would occupy position 126.

4. THE TOOLS.

MEDLINE Visualiser. When we use indexing services, such as the one described in Section 3 or most Internet search engines, to search for information across a huge amount of text documents, we usually receive the answer as a list sorted by a relevance criteria defined by the search engine. We could try refining the query further, but even by applying this time-consuming procedure, we can never be confident about the quality of the result. For this reason MIDAS is developing an interactive visual tool based on the SearchPoint technology (searchpoint.ijs.si) that helps the information we are looking for emerge more coherently. It summarizes the results of an indexing service and allows users to interactively explore its results based on topics extracted from their textual snippets. After performing the search within MEDLINE, the output presents documents in clustered topics represented as word-clouds, it identifies a handful of grouped keywords representing different topics. In Figure 1 those groups (on the right) are represented by different colors. By moving a pointer closer to one of the clusters, the results are reordered, promoting the documents related to that group to the top of the list. This allows efficient exploration of the medical information retrieved by interactively surfacing relevant information and avoiding the standard answer, which is biased by definition. For instance, when querying the portal against "childhood obesity", related articles are presented as five groups related to different topics that experts discuss when describing the issue. By moving the position of the pointer over the colored groups, the order of results changes accordingly. As shown in Figure 1, the item that was originally at position 126 now occupies the forth position due to its relevance to the subtopic of interest.

MeSH Classifier. The rising importance of automatic annotation based on open bodies of knowledge such as Wikipedia or MEDLINE/MeSH can be very useful to both health professionals and policy-makers. The MIDAS project is providing an automated classifier that is able to suggest the categories of the not yet annotated articles. It learns using the part of the MEDLINE dataset that is annotated with MeSH, and is able to suggest categories to the submitted text snippets from the MeSH thesaurus. We use a nearest centroid classifier [4] constructed from the abstracts from the MEDLINE dataset and their associated MeSH heading major descriptors. Each document is embedded in a vector space as a feature vector of TF-IDF weights. For each category, a centroid is computed by averaging the embeddings of all the documents in that category. For higher levels of the MeSH structure, we also include all the documents from descendant nodes when computing the centroid. To classify a document, the classifier first computes its embedding and then assigns the document to one or more categories whose centroids are most similar to the document's embedding. We measure the similarity as the cosine of the angle between the embeddings. Preliminary analysis shows promising results. For instance when classifying the first paragraph of the Wikipedia page for "childhood obesity", excluding the keyword "childhood obesity" from the text, the classifier returns the following MeSH headings:

Diseases/Pathological Conditions, Signs and Symptoms/Signs and Symptoms/Body Weight/Overweight/Obesity.

This classifier permits the automated annotation of any free text, including both the annotation of articles that do not yet have the MeSH classification available, health records proprietary to a Public Health organization, or even health related news [3].

5. CONCLUSION AND FUTURE WORK

The usage of text data goes beyond static summary reports when the user is able to manipulate dynamic representations of data sources such as MEDLINE, in a visual way that is fit to his/her workflow and topics of interest. It is particularly useful when it can complement the user's own data, to highlight insights that were potentially overseen, or to save time in tasks that are usually exhausting when done through classical methods. MIDAS is further developing the described technology to enable a visual representation of a significant part of the selected documents originated from the search results, permitting a view of the knowledge coverage through multidimensional scaling over a certain research topic. Furthermore, we aim to improve PubMed's precomputed similarity score, utilising a deep learning algorithm – Doc2vec [5] – to create similarity measures between articles in the MEDLINE corpus. On that note we aim to provide better visualisation approaches to navigate the large content of this biomedical open data set during search. In contrast, frequency based algorithms are unable to take advantage of the word context. Our implementation currently builds a matrix of similarity scores for each article in the corpus. Added value is given by the fact that these advanced tools are being developed together with health professionals and policy-makers within the MIDAS project, ensuring the delivery of meaningful technology.

ACKNOWLEDGMENTS

We thank the support of the European Commission on the H2020 MIDAS project (GA nr. 727721).

REFERENCES

[1] Brian Cleland et al (2018). Insights into Antidepressant Prescribing Using Open Health Data, Big Data Research, doi.org/10.1016/j.bdr.2018.02.002

[2] F. B. Rogers, (1963). Medical subject headings. *Bull Med Libr Assoc*. **51**: 114–6.

[3] J. Pita Costa et al (2017). Text mining open datasets to support public health. WITS 2017 Conference Proceedings.

[4] C. Manning et al (2008). Introduction to Information Retrieval, Cambridge University Press, 2008, pp. 269-273.

[5] T. Mikolov and Kai Chen (2013). Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781.