COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 131 (2016) 63–77

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

BIOMedical Search Engine Framework: Lightweight and customized implementation of domain-specific biomedical search engines

Alberto G. Jácome [a], *Florentino Fdez-Riverola* [a], *Anália Lourenço* [a,b,*]

[a] *ESEI—Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, Universidad de Vigo, 32004 Ourense, Spain*

[b] *Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*

ARTICLE INFO

Article history:
Received 26 July 2015
Received in revised form
10 February 2016
Accepted 29 March 2016

Keywords:
Search engine framework
Biomedical literature
Vertical engine
Text mining
Web application

ABSTRACT

Background and objectives: Text mining and semantic analysis approaches can be applied to the construction of biomedical domain-specific search engines and provide an attractive alternative to create personalized and enhanced search experiences. Therefore, this work introduces the new open-source BIOMedical Search Engine Framework for the fast and lightweight development of domain-specific search engines. The rationale behind this framework is to incorporate core features typically available in search engine frameworks with flexible and extensible technologies to retrieve biomedical documents, annotate meaningful domain concepts, and develop highly customized Web search interfaces.

Methods: The BIOMedical Search Engine Framework integrates taggers for major biomedical concepts, such as diseases, drugs, genes, proteins, compounds and organisms, and enables the use of domain-specific controlled vocabulary. Technologies from the Typesafe Reactive Platform, the AngularJS JavaScript framework and the Bootstrap HTML/CSS framework support the customization of the domain-oriented search application. Moreover, the RESTful API of the BIOMedical Search Engine Framework allows the integration of the search engine into existing systems or a complete web interface personalization.

Results: The construction of the Smart Drug Search is described as proof-of-concept of the BIOMedical Search Engine Framework. This public search engine catalogs scientific literature about antimicrobial resistance, microbial virulence and topics alike. The keyword-based queries of the users are transformed into concepts and search results are presented and ranked accordingly. The semantic graph view portraits all the concepts found in the results, and the researcher may look into the relevance of different concepts, the strength of direct relations, and non-trivial, indirect relations. The number of occurrences of the concept shows its importance to the query, and the frequency of concept co-occurrence is indicative of biological relations meaningful to that particular scope of research. Conversely, indirect concept associations, i.e. concepts related by other intermediary concepts, can be useful to integrate information from different studies and look into non-trivial relations.

Conclusions: The BIOMedical Search Engine Framework supports the development of domain-specific search engines. The key strengths of the framework are modularity and extensibility

* *Corresponding author*. ESEI—Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, Universidad de Vigo, 32004 Ourense, Spain.

E-mail addresses: agjacome@esei.uvigo.es, riverola@uvigo.es, analia@uvigo.es (A. Lourenço).

http://dx.doi.org/10.1016/j.cmpb.2016.03.030

in terms of software design, the use of open-source consolidated Web technologies, and the ability to integrate any number of biomedical text mining tools and information resources. Currently, the Smart Drug Search keeps over 1,186,000 documents, containing more than 11,854,000 annotations for 77,200 different concepts. The Smart Drug Search is publicly accessible at http://sing.ei.uvigo.es/sds/. The BIOMedical Search Engine Framework is freely available for non-commercial use at https://github.com/agjacome/biomsef.

1. Introduction

In Life Sciences, the wealth of knowledge in journal publications is of significant importance for researchers in making scientific discoveries. However, the acquisition of such information is becoming increasingly difficult due to the large volume and heterogeneity of articles, and the intense rate of publication in these fields.

The PubMed search engine is quite powerful in that its keyword-based Boolean query interface is practical and easy to use, and it lays on the millions of abstracts available in the Medline database [1]. However, it can be difficult for researchers to explore and search such huge volume of data in an efficient manner. Queries about general topics are likely to return a large number of potentially relevant documents (hundreds or thousands of documents) that require further manual revision. Conversely, queries about very specific topics demand that the researcher knows in advance the most relevant keywords of the topic (which often vary over time).

It is much more intuitive for users to query about concepts relevant to the domain and with which they are familiar with. Within this context, domain-specific literature search is currently of high demand by biomedical researchers [2]. Efforts are being invested in the development of alternative ways to query documents from the Medline database and offer topic-driven search facilities. Alternatives are quite varied in purpose and nature, but most of them take advantage of text mining technologies and domain-specific semantics to offer a more focused and contextualized search experience [3,4].

Today, major biological databases, such as UniProt protein resources, BioCyc pathway knowledge bases and BRENDA enzyme information system enable concept-driven searches of the curated documents (abstracts or full-texts) [5–7]. Furthermore, the annual report of the Nucleic Acids Research journal about biomedical databases and Web services presents a wide variety of specialized literature search engines, created to meet the information needs of particular research communities [8].

Table 1 presents a summary of recent projects that developed domain-specific literature search systems. The purpose is not to provide a complete list of available domain-specific systems but rather to describe the technical standpoint of these applications, most notably the support provided by domain-specific semantics, the use of text mining technologies, and the visual artifacts used to enable concept-driven knowledge discovery. Text mining methods and tools are used to incorporate semantic functionality such as the automatic suggestion of synonyms for user-submitted query terms, the assessment of document relevance, the description of document contents, or the inspection of documents in which concepts are related in specific ways. Controlled vocabularies and ontologies, such as DrugBank [18], UMLS [19], MESH [1], and SNOMEDCT [20], help define the knowledge domain to be captured by the search engine and offer the means to convert keyword-based queries in domain concepts, and navigate within domain semantics [21].

For example, the Protein Interaction information Extraction (PIE) system implements a machine learning approach to provide a PubMed-like search interface that prioritizes documents mentioning protein–protein interactions [9] while Alkemio Web tool uses a naïve Bayesian classifier to predict the relatedness of chemicals to query topics [11]. In turn, the PolySearch2 supports the discovery of associations between human diseases, genes, drugs, metabolites and toxins in MEDLINE abstracts, PubMed Central full-text articles, and text-rich biological databases [13].

Typically, all search engines present results as a ranked list. Semantic annotations are visually highlighted and some semantic facets (based on concept types) may exist to narrow down the list. Graph- or tree-based visualizations are provided as advanced means of navigation, namely, to find indirect associations between concepts of interest.

From a technical standpoint, one obvious observation is that most projects undertake the construction of the domain-specific search engine from scratch, and that all these new implementations include modules for document retrieval, text mining, document indexing and scoring, and query execution. Typically, document retrieval relies on MEDLINE web services and text mining relies on open solutions for linguistic processing and biomedical entity recognition. Moreover, document/concept scoring is often based on well-known algorithms, such as the term frequency inverse document frequency (TF-IDF) algorithm [22].

Understandably, domain requirements and goals of analysis determine the practical choice of methods and tools used, but many technologies and tools have the potential of being applied across domains. In particular, the core infrastructural components linking main operations and resources in most biomedical semantic search engines could be generalized with the help of a web development framework.

There are general purpose and open-source search engine development frameworks, but they are hardly used in bioinformatics applications. These frameworks are typically equipped to deal with distribution and scalability concerns, i.e. enabling the construction of large-scale engines, rather than delivering generalized semantic functionality, i.e. enabling the construction of domain-specific engines (Table 2). So, the learning curve associated with using these frameworks together with the need to program additional biomedical semantic processing and analysis modules discourages bioinformatics application.