LongLife: a Platform for Personalized Search for Health and Life Sciences*

Patrick Ernst, Erisa Terolli, and Gerhard Weikum

Max Planck Institute for Informatics, Campus E1 4, 66123, Saarbücken, Germany
{pernst, eterolli, weikum}@mpi-inf.mpg.de

Abstract. This work demonstrates Longlife: a system for semantically enhanced, personalized search of information about health issues and life-science topics. The system supports user-friendly access to entities, categories and free-text phrases in a corpus of 21 million documents, comprising scientific publications, clinical trials, encyclopedic articles, biomedical news and health forum posts. Search results can be personalized for two kinds of users: patients can provide descriptions of their health history, symptoms and therapies in layperson terms (as in health discussion forums), and doctors or researchers can target specific entities and categories (for disorders, symptoms, risk factors, drugs etc. – e.g., when searching on behalf of a patient).

1 Introduction

Motivation: Although individual health and precision medicine are of great importance to society, search engines hardly support information needs by patients or doctors. PubMed search over biomedical publications supports filters on fields and MeSH tags, but this is still far from what semantic search can do in other domains such as business or travel where text is enriched with entity markup and background knowledge graphs. The Semantic Web community has worked on creating Linked-Data resources for genes, diseases and drugs (e.g., Bio2RDF, DrugBank, DisGeNET) (incl. work on Sparql querying, e.g., [4]), but there is no linkage with the textual content that doctors and patients provide across the Internet. Moreover, search over online health communities (e.g., ehealthforum.com/health/health_forums.html), where patients and doctors discuss personal experiences with disorders, symptoms and therapies, is very basic. IR research for health has largely focused on clinical data (see, e.g., [5] and references there).

As an example, consider a user or doctor (on behalf of the patient) querying about "pancreatic cysts and abdominal pain". Search engines over clinical articles or health forums merely return all kinds of pancreas-related posts.

Contribution: LongLife provides access to entities, categories and free-text phrases in a corpus of 21 million documents, comprising scientific publications, clinical trials, encyclopedic articles, biomedical news and health forum posts.

The semantic layer of entities and other annotations is automatically generated by named entity recognition [2] and linking entities to the DeepLife biomedical knowledge base [3] which encompasses a variety of LOD datasets (Bio2RDF, DrugBank etc.) and the UMLS taxonomy. In contrast to most prior works on biomedical entities, our method goes beyond major types like genes, proteins, diseases and drugs, by capturing a much wider range of entities like symptoms/syndromes, therapies and nutrition- or lifestyle-related risk factors.

On top of this semantically enriched corpus, LongLife offers personalized search by incorporating individual user information on a per-query basis. Lay users like patients typically pose keyword queries, but can add free-text self-descriptions of their case histories (e.g., like posts in health forums). LongLife automatically detects health-related entities in such texts, infers relevant biomedical categories and expands the user query into a semantic-search request. This way, it can return answers that are of specific relevance to the user, e.g., experience of similar patients. As a second use case, when doctors search on behalf of patients, entities and categories may be manually added and further patient properties can be specified (e.g., blood pressure and other vital signs). Again, LongLife automatically synthesizes the final query from these inputs, and computes personalized rankings of answers.

2 System Overview

Data and Indexing: LongLife has currently indexed 21,036,802 documents crawled from a diverse corpus that covers the full spectrum of biomedical information on the web: 19,884,225 scientific publications, 111,139 encyclopedic articles, 76,554 news articles, 164,756 clinical trials and 1,048,428 health forum posts. LongLife stores the data based on ElasticSearch v.1.7.6. We index the following parts: title, full text, topical domain (e.g., cancer, diabetes etc.) and all biomedical entities using the UMLS thesaurus as entity repository. For entity recognition, we use the method of [2] based on min-hash sketches for matching candidate phrases to entity names. We disambiguate between multiple entity candidates by considering only the most specific entity according to the UMLS type system and picking the highest ranked entity. Every detected entity is linked to the LOD Cloud leveraging a mapping between UMLS and Bio2RDF.

Query Processing: LongLife has a form-based search interface with auto-completion suggestions for each field. Input can take the form of keywords or multi-word phrases, entities and/or categories, where the latter two are identified by having the user choose from auto-completion suggestions. Similar to health forum posts, users are asked to pose a question composed of a short post title and a post body containing a description of the individual case. This input is then processed as follows:

- The user question is cast into a keyword query.
- The query is expanded with informative entities and their semantic categories identified in the full text of the case description (see below).
- The expanded query is issued to ElasticSearch.

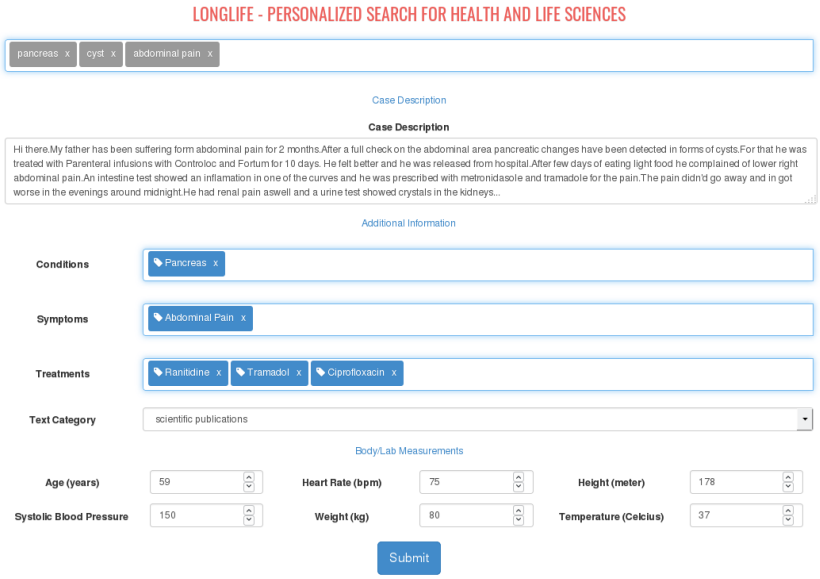

Fig. 1: LongLife Search Interface

- The result ranking is computed by LongLife's customized scoring function that considers the personalized query expansion (see below).

Personalized Query Expansion: We expand the initial keyword query with biomedical entities extracted from the medical case description. Since UMLS covers a broad spectrum of entities, we constrain, by default, the entity set to symptoms, diseases, medical findings and pharmacological substances. Each entity is assigned an weight computed as the squared Pointwise Mutual Information PMI^2 to the document's domain. PMI^2 between entities a and b is $log\frac{p(a,b)^2}{p(a)p(b)}$ [1]. The domain is the health topic that the document belongs to (e.g., cancer, diabetes, etc.). It is mostly derived from document meta-data, e.g., keywords field of PubMed articles or the names of sub-forums in health communities.

Optionally, we further expand the query with the semantic types/categories of entities obtained from DeepLife [3]. The selected categories do not only encode typing information derived from UMLS, but also reflect relational facts harvested from a large text collection. For example, for *Ibuprofen* we retrieve the categories *anti-inflammatory agent* (type) and also *treatment of fever* (fact) among others.

Answer Scoring: Longlife uses a linear combination of TF-IDF-style scores. We define a query $Q = (T, E, C)$ where T is the set of user's question keywords, E is the set of extracted entities from the case description and C is the set of semantic categories for E. For document $D = (D_t, D_e, D_c)$,

$$score(D,Q) = \lambda_T \sum_{t \in T} idf(t) \frac{tf(t,D_t)}{\sqrt{D_T}} + \lambda_E \sum_{e \in E} PMI^2(d,e) idf(e) \frac{tf(e,D_e)}{\sqrt{D_E}} + \lambda_C \sum_{c \in C} idf(c) \frac{tf(c,D_c)}{\sqrt{D_C}}$$

where d is the domain and $\sqrt{D_{\{T,E,C\}}}$ are normalization factors. We tuned $\lambda_T = 1.0$, $\lambda_E = 0.6$, $\lambda_C = 0.1$ via grid search with relevance labels from crowdsourcing.

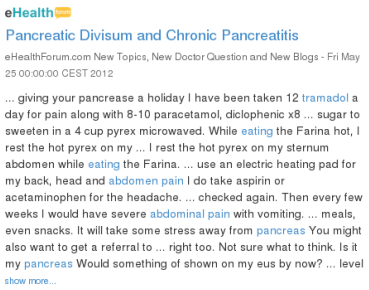

Fig 2: Health Forum Top Result

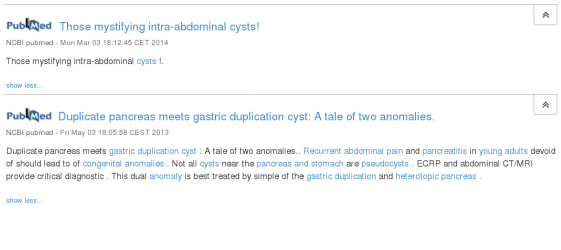

Fig 3: Top Two Results from Scientific Articles

3 Demo Scenarios

LongLife supports both lay users and professionals to discover relevant documents for their specific queries within the entire corpus or the sub-corpus of their choice (e.g., scientific articles only or forum posts only). Figure 1 shows a screenshot of the input functionality of our system. We illustrate the benefits of LongLife by the following two use-case scenarios.

Lay User Scenario: Consider the patient with the case in Figure 1 searching health forums for other users with similar experience. All she has to do is pose the question and provide the description. LongLife automatically converts these inputs into well-crafted query by inferring entities and categories and expanding the query. The top results for this example search is shown in Figure 2.

Professional Scenario: Doctors and researchers are interested in clinical trials and publications. LongLife provides an advanced search box for such experts, where users can specify entities and categories of interest, via convenient autocompletion. Another important feature is to specify vital parameters and lab values of a patient, such as height, weight, age, heart rate and blood pressure. These measurements are automatically mapped into medical entities such as obesity, hypo/hypertension, tachycardia etc., and harnessed for result ranking. Top results of scientific articles for the search example of Figure 1 are shown in Figure 3.

References

1. F. Role et al.: Handling the impact of low frequency events on co-occurrence based measures of word similarity. KDIR 2011
2. A. Siu et al.: Fast entity recognition in biomedical text. Workshop on Data Mining for Healthcare at KDD 2013
3. P. Ernst et al.: DeepLife: An Entity-aware Search, Analytics and Exploration Platform for Health and Life Sciences. ACL 2016
4. A. Hasnain et al.: BioFed: Federated Query Processing over Life Sciences Linked Open Data. Journal of Biomedical Semantics 2017
5. G. Zuccon, B. Koopman: Tutorial on Health Search: From Consumers to Clinicians. WSDM 2019. https://github.com/ielab/health-search-tutorial/tree/wsdm2019