Information Retrieval for Precision Oncology

Jurica Ševa[1], Julian Götze[2], Mario Lamping[3], Damian Tobias Rieke[3,4], Reinhold Schäfer[5], Ulf Leser[1,*]

Abstract: Diagnosis and treatment decisions in cancer increasingly depend on a detailed analysis of the mutational status of a patient's genome. This analysis relies on previously published information regarding the association of variations to disease progression and possible interventions. Clinicians to a large degree use biomedical search engines to obtain such information; however, the vast majority of search results in the common search engines focuses on basic science and is clinically irrelevant. We developed the *Variant-Information Search Tool*, a search engine designed for the targeted search of clinically relevant publications given a mutation profile. VIST indexes all PubMed abstracts, applies advanced text mining to identify mentions of genes and variants and uses machine-learning based scoring to judge the relevancy of documents. Its functionality is available through a fast and intuitive web interface. We also performed a comparative evaluation, showing that VIST's ranking is superior to that of PubMed or vector space models.

1 Introduction

Precision oncology denotes treatment schemes in cancer in which medical decisions depend on the individual molecular status of a patient [Ga13]. The most relevant molecular information is the set of variations (mutations) each individual carries. When faced with the variant profile of a patient, clinicians critically depend on accurate, up-to-date, and detailed information regarding the clinical relevance of the present variations. Finding such information is highly laborious and time-consuming, often taking hours or even days [Do17]. We demonstrate Variant-Information Search Tool (VIST), a search engine specifically developed to aid clinicians in precision oncology in their search for clinically relevant information for a (set of) variations or mutated genes. The core of VIST is its ranking function which, given a (set of) variation or a (set of) gene and a cancer entity, ranks those documents of its corpus highest which contain clinically relevant information. The main difficulty when developing a ranking function for such a novel and quickly emerging field are (a) the lack of gold standard data and (b) the complexity of the concept "clinical relevance", encompassing, among other, information about gene-mutation-drug associations, frequencies of variations within populations, mode of action of drugs and molecular functions.VIST copes with this

[1]Department for Computer Science, Humboldt-Universität zu Berlin; [2]University Hospital Tübingen; [3]Charité – Universitätsmedizin Berlin; [4]Berlin Institute of Health (BIH); [5]Deutsches Krebsforschungszentrum; [*]Corresponding author: leser@informatik.hu-berlin.de.

doi:10.18420/btw2019-39

complexity by using: (1) advanced information extraction to pre-filter documents based on the genes and variations they mention, and (2) machine learning (ML) document classifiers trained on a silver-standard corpus of clinically related documents. VIST furthermore offers several metadata filters (journal, year of publication), highlights key phrases (i.e., the clinically most important sentences) and mentions of query entities when displaying documents, links out to external databases, and allows mixing of entity and classical keyword search. VIST was developed inn close interaction with medical experts and is freely available at https://triage.informatik.hu-berlin.de:8080/. It is the first search engine directly targeting clinical relevance of documents which required the integration of ML methods into the ranking. This discerns it technically from other biomedical IR systems, such as GeneView [Th12] or DigSee [Ki13]. The algorithmic problem of finding clinically relevant documents for variation data was also studied in the recent TREC Precision Medicine evaluation [Ro17]. However, the precise task was different from what we target in VIST, as also general medical data and comorbidities of patients were included, which would be very sensitive to implement in a public search engine like VIST.

2 VIST

VIST is a document retrieval system which ranks PubMed abstracts according to their clinical relevance for (a set of) queried variation(s) and/or gene(s) and a cancer entity. When inserted into the index, documents undergo a comprehensive processing pipeline including textual preprocessing, metadata extraction, named entity recognition, classification regarding cancer-relatedness, cancer type, and clinical relevance, and keyphrase detection [Še18]. We detect gene mentions using GNormPlus, variations using tmVar, and drugs using tmChem. All documents and annotations are indexed using Solr. To rank documents against a query, we pre-rank all documents according to two scores: one for their relatedness to cancer in general, and one for clinical relevance. In both cases, we use supervised document classification trained on CIViC [Gr17] and OncoKB [Ch17] with tf-idf weighting. Results are computed by first retrieving all documents containing any of the given variants / genes and filtering for cancer type. Remaining documents are ranked according to a linear combination of "keyword score"(cosine similarity to query), "cancer score"(confidence of the cancer-relatedness classifier), and "clinic score"(confidence of the classifier for clinical relevance).

3 Evaluation

VIST was extensively evaluated to assess and optimize its performance. First, we used a prototype version of VIST to curate a new corpus of clinically (ir-)relevant documents for performing evaluation, resulting in 188 individual scores., of which 119 are used for evaluation; the others were removed due to inconsistent ratings. We assessed the accuracy of the clinical-relevance classifier on this data set using cross-validation. Finally, we used

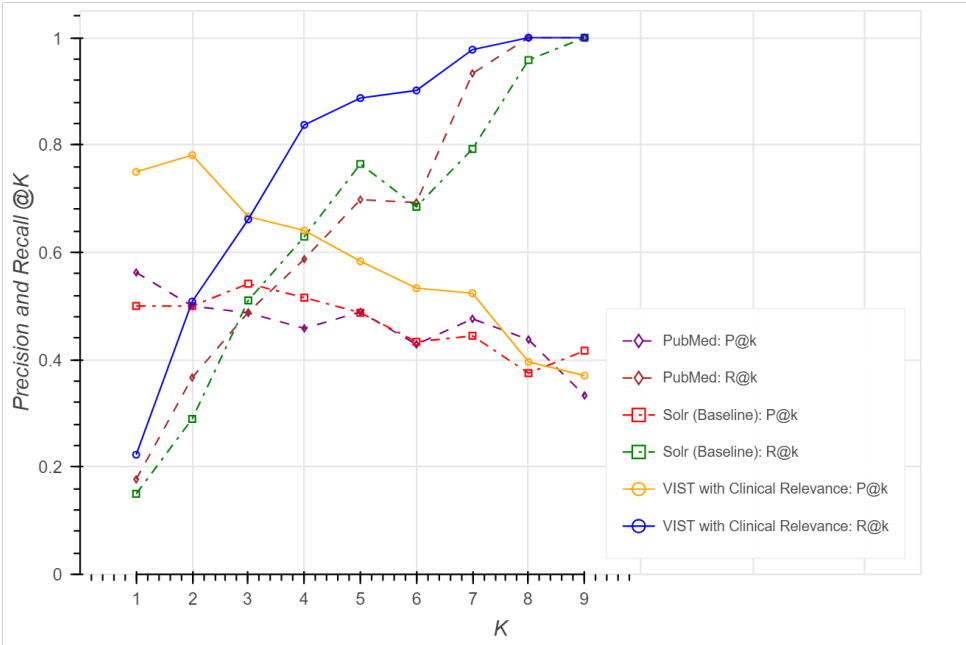

Fig. 1: Precision/Recall @ k averaged over all queries. k denotes the k'th document in the result set that is also contained in the test set.

this set to compare the ranking performance of VIST with that of PubMed and a pure VSM. Due to lack of space, we here only report on the third evaluation. In this comparison we cannot compare absolute ranks of results as VIST performs result filtering, leading to different result sizes. Instead, we use two metrics which are robust to different result set sizes. Regarding the ratio of the average rank of all relevant documents from our test data to the average rank of all irrelevant documents, VIST performs best in 11 out of the 16 queries and very close to the best in 5 out of 16 queries. Second, Figure 1 shows average precision@k and recall@k for all three systems; therein, k denotes the k'th document in the ranked result that is also contained in the test set. Clearly, VIST outperforms VSM-ranking and PubMed in both regards.

4 Web Interface and Demonstration

The VIST web interface allows users to define queries and inspect matching documents. Additionally, it offers entity highlighting, various document filters, and a help page. The query shown below is taken from the evaluation queries, also available in the user interface as example query.

Starting a New Search. The initial query is of the format *Q: [keyword(s), gene(s), variant(s)]*. At least one item has to be specified. Matching abstracts are presented in a ranked order based on VIST relevance score. For each document, its title, PMID, publication year and ranking score are shown. The basic interface is shown in Figure 2.

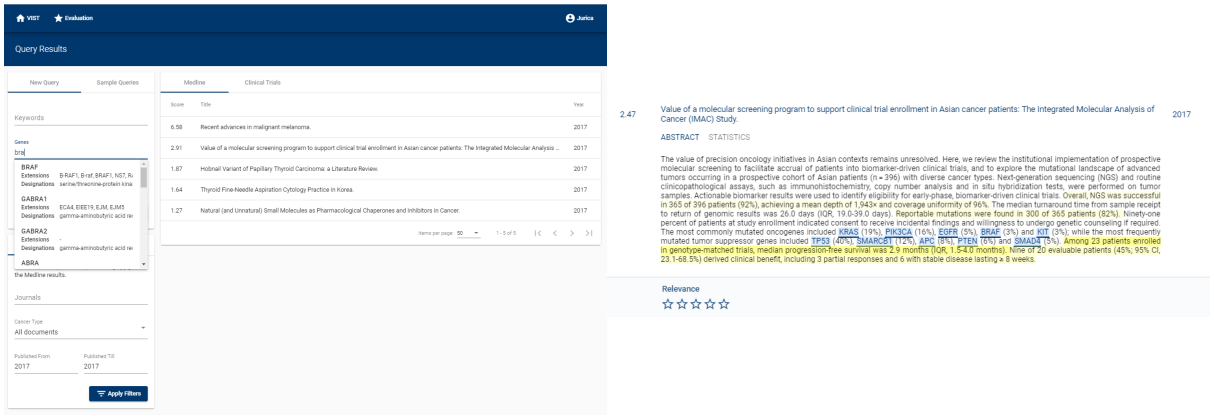

Fig. 2: VIST web interface: Left: Search interface and result overview. Right: Detailed search result with entity and keyphrase highlighting.

Filtering and highlighting of retrieved documents. Enabled as soon as a search yields a non-empty result. VIST allows narrowing returned results by (a) journals, (b) year of publication, and (c) cancer type.

Viewing Document Details. Key sentences and annotated entities are visually highlighted. Key sentences are represented with yellow background with varying transparency levels corresponding to confidence of the detection method. Found genes and drugs are linked to relevant databases (NCBI Genes and DrugBank, respectively). The interface also shows MeSH keywords and a link to the original publication.

References

[Ch17] Chakravarty, Debyani et al.: OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology, 1(1):1–16, 2017.

[Do17] Doig, Kenneth D. et al.: PathOS: a decision support system for reporting high throughput sequencing of cancers in clinical diagnostic laboratories. Genome Medicine, 9(1):38, 2017.

[Ga13] Garraway, Levi A et al.: Precision Oncology: An Overview. Journal of Clinical Oncology, 31(15):1803–1805, 2013.

[Gr17] Griffith, Malachi et al.: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nature Genetics, 49(2):170–174, 2017.

[Ki13] Kim, Jeongkyun et al.: DigSee: disease gene search engine with evidence sentences (version cancer). Nucleic Acids Research, 41(W1):W510–W517, 2013.

[Ro17] Roberts, Kirk et al.: Overview of the TREC 2017 Precision Medicine Track. TREC, pp. 1–12, 2017.

[Še18] Ševa, Jurica et al.: Identifying Key Sentences for Precision Oncology Using Semi-Supervised Learning. In: Proceedings of the BioNLP 2018 workshop. pp. 35–46, 2018.

[Th12] Thomas, Philippe et al.: GeneView: a comprehensive semantic search engine for PubMed. Nucleic Acids Research, 40(W1):W585–W591, 2012.