DATABASE
The Journal of Biological Databases and Curation

Database, 2018, 1–12
doi: 10.1093/database/bay052
Original article

Original article

A Field Sensor: computing the composition and intent of PubMed queries

Lana Yeganova, Won Kim, Donald C. Comeau, W. John Wilbur and Zhiyong Lu*

National Center for Biotechnology Information (NCBI) / National Library of Medicine (NLM) at the National Institutes of Health (NIH), 8600 Rockville Pike, Bethesda, MD 20894, USA

*Corresponding author: Tel.: +1 (301) 594 7089; Email: luzh@ncbi.nlm.nih.gov

Citation details: Yeganova,L., Kim,W. Comeau,D.C. *et al.* A Field Sensor: computing the composition and intent of PubMed queries. *Database* (2018) Vol. 2018: bay052; doi:10.1093/database/bay052

Received 29 December 2017; Revised 19 April 2018; Accepted 17 May 2018

Abstract

PubMed® is a search engine providing access to a collection of over 27 million biomedical bibliographic records as of 2017. PubMed processes millions of queries a day, and understanding these queries is one of the main building blocks for successful information retrieval. In this work, we present Field Sensor, a domain-specific tool for understanding the composition and predicting the user intent of PubMed queries. Given a query, the Field Sensor infers a field for each token or sequence of tokens in a query in multi-step process that includes syntactic chunking, rule-based tagging and probabilistic field prediction. In this work, the fields of interest are those associated with (meta-)data elements of each PubMed record such as article title, abstract, author name(s), journal title, volume, issue, page and date. We evaluate the accuracy of our algorithm on a human-annotated corpus of 10 000 PubMed queries, as well as a new machine-annotated set of 103 000 PubMed queries. The Field Sensor achieves an accuracy of 93 and 91% on the two corresponding corpora and finds that nearly half of all searches are navigational (e.g. author searches, article title searches etc.) and half are informational (e.g. topical searches). The Field Sensor has been integrated into PubMed since June 2017 to detect informational queries for which results sorted by relevance can be suggested as an alternative to those sorted by the default date sort. In addition, the composition of PubMed queries as computed by the Field Sensor proves to be essential for understanding how users query PubMed.

Introduction

PubMed (www.pubmed.gov) is a search engine developed and maintained by the National Center for Biotechnology Information at NLM. PubMed works on MEDLINE®, a collection of over 27 million biomedical bibliographic records as of 2017, and has witnessed a steady growth of scholarly information over the last decades. PubMed processes on average 3 million queries a day and is recognized as a primary tool for scholars in the biomedical field (1–3). Given the significance of PubMed, improving the

understanding of user queries offers tremendous opportunities for providing better search results.

For general web search engines, the problem of query understanding spans a whole spectrum of studies ranging from identifying high-level query intent (informational, navigational or transactional) (4–8), to identifying finer-grained query information, such as person, age, movie, travel, job domains (9), to understanding query semantics (10–12). Many queries asked on the web target structured or semi-structured web data, such as commercial products, movies etc. Mapping the unstructured language of these queries into a structured representation has been studied extensively and shown to improve retrieval results (13–15). Other approaches used in query understanding include statistical machine learning (7), deep learning (9, 12), mapping to Wikipedia semantic space (10, 11), relying on query logs (11, 16) and click information (17).

Despite the extensive research into general web search, there has been less published research on usage patterns for online biomedical information resources. It is, however, known that there are important differences between the two (18–21). In the biomedical domain there are a few studies aimed at understanding how health information is being searched for and the information needs of domain users such as clinicians, medical researchers or patients (18, 19, 22–25). The two most comprehensive biomedical query log analyses are the study of 1 day of PubMed queries (19) and the study of 1 month of PubMed queries (18). Both analyze statistical properties of query logs such as query length, user sessions, size of the result set and attempt to characterize queries in terms of semantics and the intent. The work described in (18) manually annotates a random set of 10 000 queries from PubMed logs by mapping the segments of queries to sixteen predefined categories of semantic types. The study in (19) attempts to perform semantic analysis of queries by mapping them to the MeSH controlled vocabulary.

One major aspect of queries examined by both the general and biomedical search domains is query intent. As defined by Broder (5), general web queries can be characterized as informational, navigational or transactional (usually not observed in scholarly searches). Extending this definition to PubMed, informational queries, also known as topical searches, such as *colon cancer*, or *familial Mediterranean fever*, are intended to satisfy information needs on a particular topic. Navigational queries, also known as known-item queries (26), such as *Katanaev AND Cell 2005, 120(1): 111–22,* are intended to retrieve a specific publication. In PubMed, navigational queries can be composed of citation elements including author name, title, volume, issue, page and/or date, or are complete citations. Only a small percentage of PubMed queries include explicit fields assigned by a user and are trivial to understand. The vast majority of queries have no field assignments, although the latent structure information is assumed. For these queries, the burden of mapping query segments into fields is shifted to the search system.

The reason why predicting query intent is important is that it frequently drives the behavior of a search engine (27). Informational queries focus on access to free text, which tends to retrieve many documents and a sorting function is crucial for displaying the results. In contrast, navigational queries require syntactic parsers and access to structured citation data, and represent an intent of a user to find a specific document or a website. This distinction is particularly important for searching scholarly citation databases, such as PubMed, where navigational queries constitute a significantly larger portion of all queries, compared with a general search domain. As we demonstrate in this study, navigational queries account for just about half of all searches, while in general search domain they are reported to account for 10% of the queries (6).

Although health-related queries and health information retrieval have drawn attention toward developing new tools and techniques specific for this domain (28), to our knowledge, there are no applications that can infer the intent of PubMed queries algorithmically. Two recent studies consider predicting the intent of academic queries (29, 30). The study (29) reports that in academic search engines navigational queries constitute 7.6% of queries, however, the computation relies on explicit cues, such as ISBN number, DOI or other citation related tags to classify queries. Given that only a small percentage of PubMed queries include explicit fields assigned by a user, more sensitive methods are needed for classifying query intent. The study (30) presents a binary classification approach for predicting the intent of scholarly queries, where authors report an F1 score of 0.677 on a set of 579 manually annotated scholarly queries using their best method (Gradient Boosted Trees). They use features such as number of tokens in a query, the ratio of query terms identified as author names, whether the query has punctuation or not, etc. to drive the training. To address the problem of predicting the intent of biomedical queries, we developed Field Sensor, a web-scale tool that assigns a field to each token or sequence of tokens in a query, by computing a mapping between a query segment and a field, along with the likelihood of that mapping. For example, given the query *sleep apnea, cushing* it identifies that *sleep apnea* is a text, and *cushing* is an author name, and predicts the mapping *sleep apnea [text], cushing [author]*. Based on the field assignments the query intent is inferred as follows: the query is considered informational if it consists of *text* fields only, otherwise we call it navigational.

The Field Sensor is a probabilistic field prediction mechanism equipped with two rule-based preprocessing modules. The system starts with the syntactic query chunking module, which splits the query based on logical operators and parentheses. Segments of a query tagged by a user are also identified at this stage and remain unchanged. It is followed by the rule-based query tagging module designed to recognize citation elements of a query originating from volume, issue, page and date fields by considering patterns between numbers and punctuation. And finally, the third module is the probabilistic field prediction module, which given a query segment predicts the field of the segment. The model is based on a Bayesian approach that infers the mapping between a query segment and a field in PubMed based on collection statistics. Our probabilistic field prediction module is related to the probabilistic retrieval model for semi-structured data (PRMS) (13) in the way the mappings between the query words and fields are computed. However, the PRMS is a unigram bag-of-words model, while our model takes into consideration term dependencies and attempts to predict fields for query segments of up to five tokens.

The query term to field mapping predicted by the Field Sensor can be used in multiple different ways. An important functionality of our method is classifying a query as informational or navigational. The Field Sensor has been deployed in PubMed since June 2017 to identify informational queries, for which search results sorted by relevance are suggested to users as an alternative to the default reverse time order (31). An additional application could be to the mixed queries that contain both informational and navigational components. Identifying these fields can leverage the search process by applying different search strategies to informational and navigational components of a query. Furthermore, the Field Sensor is indispensable in query log studies. It allows us to study how biomedical information is being searched for. For queries that do not retrieve any documents, the tool may help us better understand what fields are more likely to be the cause. And finally, since the underlying Field Sensor model is computed based on Medline data, it can be used for any other specialized biomedical citation database, such as bioRxiv (http://biorxiv.org/) or can be retrained for application to other academic search engines.

Materials and methods

In this section, we describe our model behind the Field Sensor, a tool for inferring a field for each token or sequence of tokens in a query. Articles in PubMed are entered into the database in a uniform structured way as: article abstract, article title, author name(s), journal title, volume, issue, page and date. These are the fields we are interested in mapping to. We will label the segments of query that map to the eight fields as *text*, *title*, *author*, *journal*, *volume*, *issue*, *date* and *page*, respectively. Note that *text* field corresponds to vocabulary found in abstracts. While the articles contain additional database fields, e.g. affiliation, we find the outlined eight fields to be the most relevant for our work.

Figure 1 illustrates the overall query processing and field prediction flow in terms of the three core modules: syntactic query chunking, rule-based citation tagging and probabilistic field prediction. Throughout the section, we refer to the example query *Katanaev AND Cell 2005, 120(1): 111–22* to illustrate the functionality of each module by demonstrating how segments of a query are being interpreted at each stage.

Syntactic query chunking

The first step in the process of query understanding is to find segments of a query connected by logical operators (AND, OR), parentheses and brackets. Parentheses and brackets establish the order of operation, but for now we process a query in a linear fashion. Segments of a query tagged by a user are also identified at this stage and remain unchanged. Figure 1 illustrates the query input and output when syntactic chunking is applied in the red box. In this example, the query is partitioned into two segments separated by logical AND. A total of 12.66% of the queries benefit from this module.

Rule-based citation tagging

The rule-base citation tagger is designed to detect citation elements of a query by interpreting punctuation and numerals that are indicative of citation information such as volume, issue, page and date. This is a rule-based approach, which is equipped with a wealth of patterns used for identifying these citation elements. For example, the module:

- recognizes that *page, pp, p* are page indicators, *v* or *vol* are volume indicators;
- interprets patterns indicating the page range, such as 1860–73;
- identifies patterns indicating volume and issue information, e.g. 83(2) or 351: 18.

The blue box in Figure 1 illustrates query output when the citation tagger is applied. A total of 8.8% of queries benefit from this module, and these are queries containing citation information. Most of the queries modified by this module are different from queries that are affected by the

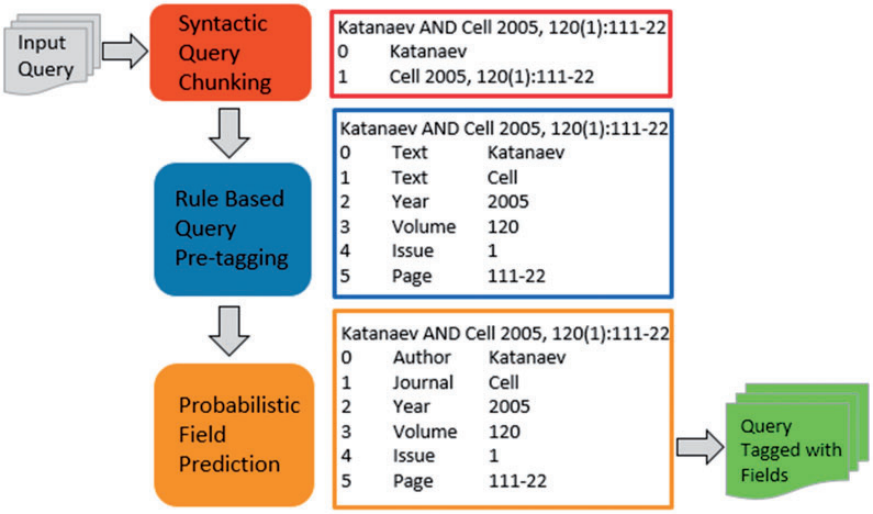

Figure 1. The Field Sensor query processing pipeline. The system consists of three core modules: syntactic query chunking, rule-based citation tagging and probabilistic field prediction. Next to each module we illustrate how query segments are being interpreted by each of these three modules.

first, syntactic query chunking module suggesting that syntactic query chunking mainly targets informational queries, i.e. sequences of text elements linked by logical operators. Of all queries that are modified by either the syntactic query chunking or the rule-based citation tagging, only ~2% are modified by both, indicating that these processes are complementary.

Our rule-based citation tagger processes a text string in three steps. In the first step, the algorithm looks for matching parentheses, square brackets or quotes. If matching parentheses are found, the string between the parentheses is labeled *parenthetical*. Likewise, the string between the square brackets is labeled *Tag*, and the string preceding the *Tag* is labeled *Tagged*. The string contained between matching quotes is labeled *Quoted* and is intended to not be split down by this module.

Step two consists of assigning more definite labels to tokens that represent citation information. For this purpose, the tokens are considered in order of occurrence and are tested against a known set of terms that are abbreviations for months (e.g. *Jan*, *Dec*), page number indicators (e.g. *p*, *pp*) and volume number indicators (e.g. *v*, *vol*). When such tokens are recognized and occur in an appropriate context we label them *Month*, *PageIndicator* or *VolumeIndicator*. If the token is not recognized as one of these, we examine the individual characters. If the first character is a digit or the second character is a digit and the token contains a hyphen, the label is changed to *Numeric*. Otherwise we test for string being alphabetic and pass it on for processing by the next module.

Step three again examines the tags in the order the tokens occur. If a token is found labeled *PageIndicator* this indicates the following token must have the label *Page*. Likewise, the label *VolumeIndicator* must be followed by a *Volume* label. The label *Numeric* receives special treatment. If the corresponding token is an integer between 1900 and current year the label is changed to *Year*. If the token contains a '–', e.g. *111–22*, it is labeled *Page*. If the *Numeric* labeled token represents an integer in the range 1–31 and follows a token with the label *Month*, the *Numeric* label is changed to *Day*. An integer inside parentheses is considered further. If it is in the range 1900–current year it is labeled *Year*. Otherwise, if it is preceded by a *Numeric* label, it is tested to match one of the volume and issue patterns, e.g. *83(2)*, and is assigned an *Issue* label. Finally, an attempt is made to label *Numeric* tokens that appear next to tokens not labeled *Numeric* as *Volume* or *Issue* if these labels are not already taken. Furthermore, if both labels are assigned they must appear adjacent and in the order *Volume* followed by *Issue*. Not all details are included here, but the foregoing provide the major features.

Probabilistic field prediction

Queries containing no parsing information or field indicators constitute 78.54% of all queries. The probabilistic field prediction module predicts query fields by establishing relationships between query tokens and fields in the PubMed database. Our method assumes that a query has an implicit mapping of each query term or sequence of terms into one of the eight fields, and that the distribution of words in the fields of the database provide the basis for the inference process.

After being processed with the first two modules, the probabilistic field prediction is applied to the segments of

query where no other parsing information is available. The fields that we consider for the calculation are article abstract, article title, author name, journal title, volume, issue, page and date. We assume that a query Q is composed of m terms $Q = (t_1, \ldots, t_m)$ and we want to predict the probability that a term t in a query should be interpreted as originating from a field F_i in a PubMed record.

We begin with an application of Bayes' theorem

$$P(F_i t) = \frac{P(tF_i)P(F_i)}{P(t)}. \quad (1)$$

To obtain an estimate of the left side we estimate each factor on the right side. The factor $P(t|F_i)$ is the probability of observing term t in the field F_i of a PubMed record. We compute $P(t|F_i)$ for each one of the eight fields using a language model (32) as follows:

$$P(t|F_i) = \frac{\text{freq}(t \in F_i)}{\text{freq}(F_i)}. \quad (2)$$

The factor $P(F_i)$ is a prior probability of the field being the source of terms. We obtain the estimates for $P(F_i)$ from the set of 10 000 manually annotated PubMed queries discussed in the next section. Under the assumption that this list of fields is exhaustive and mutually exclusive, a given interpretation of a query will only assign one field to each term. This allows us to compute $P\Big(t) = \sum_{i=1}^{8} p(t|F_i)p(F_i)$. We can then apply (1) to predict the most likely assignment of fields to the terms in a query.

A language model generally computes a probability distribution over sequences of words. Given a sequence of length m it assigns a probability $P(t_1, \ldots, t_m)$ to the whole sequence. The unigram model assumes the independence of terms and computes the probability as $P_{uni}(t_1, t_2) = P(t_1)P(t_2)$. The unigram language model is frequently used in speech recognition, machine translation and POS tagging. A useful extension of a unigram model is a bigram model, which assumes that the probability of observing term t_2 is dependent on the preceding term, and computes the probability of bigram $t_1\ t_2$ as $P_{2-\text{gram}}(t_1, t_2) = P(t_1)P(t_2 t_1)$.

First, we compute the probabilities based on a unigram language model, then we extend the analysis to sequences of word pairs. We use a bigram language model to compute a probability of a term pair $P_{2-\text{gram}}(t_1, t_2) = P(t_1)P(t_2 t_1)$ and compare its value with $P_{\text{uni}}(t_1, t_2) = P(t_1)P(t_2)$. For every field where a pair of tokens is found, if $P_{2-\text{gram}}(t_1, t_2) > P_{\text{uni}}(t_1, t_2)$ for that field, we join two terms t_1 and t_2 into a phrase $t_1\ t_2$ and the highest probability field is predicted for the pair. When two terms are joined into a phrase $t_1\ t_2$, the process iteratively continues to check succeeding pairs $t_2\ t_3$, $t_3\ t_4$, and $t_4\ t_5$ and extends the predicted query segment, until it can no longer be extended or has reached five tokens in length (operational decision). For every segment length, the field with the highest probability is recorded. In our language models we do not use smoothing, because we restrict our approach to finding terms that actually appear in the database from which the models are derived. Then we compute a path through a query, somewhat similar to a decoding stage in a Viterbi algorithm (33). Starting with the first token, we use a greedy method by iteratively moving the pointer to the end of the longest predicted segment. Although this does not yield optimal segmentation boundaries, it provides a reasonable approximation.

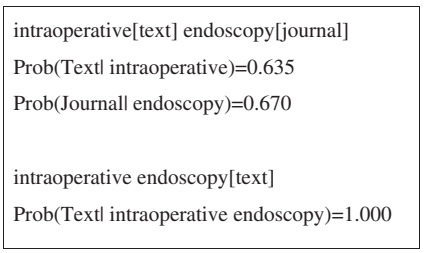

Figure 2. Example of probabilistic field assignments using unigram and bigram language models.

In Figure 2, we work out an example to demonstrate how the probabilities are computed for a sample query *intraoperative endoscopy*. The prediction based on the unigram language model assigns the highest probability field to single tokens and results in predicting *endoscopy* as a journal name. However, as we apply the bigram language model and compute the probability of *intraoperative endoscopy*, we correctly predict the phrase *intraoperative endoscopy* as a *text* field. In this example, *text* turns out to be the only field where that bigram is found.

Given the correlation between words in article abstracts and article titles, the disambiguation of whether a user intends to perform a keyword search or an article title query can be crucial to returning the appropriate search results. After the standard field prediction, the Field Sensor includes an additional check to verify that a query segment is a full title or is a significant portion of a title. We use the *title* field designation to match a full title or a significant portion of a title in a query. The *text* field designation is for the text word query segments representing the subject of interest, occurring either in a title or an abstract. A strong correlation is also present between journal titles and text terms, particularly for short journal names such as *cancer*, *blood* and *circulation*. Many journal names match frequent terms and present a problem. In addition, the language model-based probability calculation may favor the field with smaller vocabulary size, and *journal* field is an

example of such. When a term can be interpreted both as a text term and a journal name, the model is more likely to predict it as a journal name. We post process the initial field predictions for a single token predicted as *journal* to check if the query contains additional citation information such as volume, issue, date or page and/or author name. Otherwise, if a likelihood of a *journal* field prediction is below 0.8, we tag the term as *text*.

Database indexing, data preparation and implementation

An essential component of the Field Sensor is the database indexing step. PubMed data is pre-processed separately for each one of the eight fields of interest. In each field, we collect single term probabilities for every term, as well as joint and conditional probabilities for pairs of terms. These values are stored separately for each field in a readily accessible format to facilitate quick access to the values. One important detail in this process is tokenization, which defines the rules for handling spaces and nonalphanumeric characters when splitting a text into tokens. It is crucial for the database tokenization to be consistent with the query tokenization for optimal retrieval of database entries.

The Field Sensor is implemented in C++ as a general tool for understanding the composition of health and biomedical queries. The current implementation has an average throughput of ~800 queries per second on a single thread, which meets the maximal search traffic of PubMed. The Field Sensor has been integrated and deployed in PubMed to distinguish informational queries from navigational. The PubMed usage details are highlighted in (31).

Evaluation data

Gold standard of 10K manually annotated PubMed queries

We use the publicly available dataset of 10K manually annotated queries described in (18). That study conducted semantic analysis of queries, where query contents were labeled with 1 of the 16 semantic categories: *Body part, Cell component, Tissue, Chemical/drug, Device, Disorder, Gene/protein, Living being, Research procedure, Medical procedure, Biological process, Title, Author name, Journal name, Citation* and *Abbreviation*. Seven annotators with expertise in various areas of biomedicine and/or information science were recruited to annotate the query set.

To use this manually annotated data for the Field Sensor evaluation, we slightly modify the category definitions to fit our setup. The four semantic classes of *Title, Author name, Journal name* and *Citation* are used as defined. The remaining twelve categories are conflated into one class *text*, as we do not intend to distinguish between these categories and identify them as textual elements. Note that the *Citation* category includes volume, issue, page and date, which are all annotated as *Citation*. Hence for this set we evaluate how well we distinguish the four citation elements from the remaining categories, but do not measure the performance for each class separately. Figure 3 presents the composition of the 10K set in terms of the fields. As we mentioned in the section 'Methods', the composition of the 10K set is also used to estimate the factor $P(F_i)$ representing a prior probability of the field in (1). We obtain the estimates for $P(F_i)$ from this set for the five fields: *Text*, *Title, Author name, Journal name, Citation*; and since the *Citation* category includes volume, issue, page and date, we uniformly distribute the probability of the *Citation* category between these four fields.

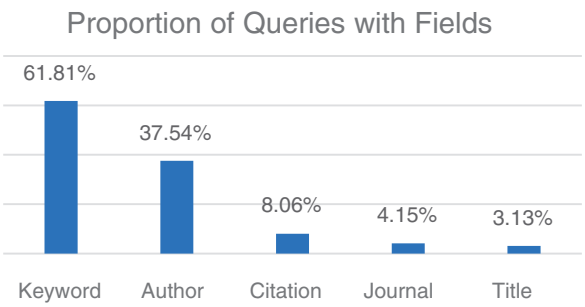

Figure 3. Distribution of fields computed for the 10K queries annotated with five fields: Keyword, Author, Citation, Journal Name and Title.

Silver standard of 103K machine annotated PubMed queries

A reliable way of automatically creating a high quality annotated query set is by establishing a unique mapping between a query and a document. Here we describe a collection of 103K machine annotated queries. The set is obtained automatically and has not undergone manual annotation, and for that reason we refer to it as a Silver Standard. Using this automated process enables us to reliably annotate an arbitrarily large number of queries.

In constructing this set, we consider how the tokens of a query can be mapped to the fields of a PubMed record. In this mapping, we employ the following fields of the PubMed article: title, author names, journal title, volume, issue, page and date. Abstracts are not considered. As we compute a mapping, we evaluate what fraction of the query has found a match to information in the PubMed document and compute a score to reflect the level of

match. Based on the score, we compute a probability of the answer being correct. A probabilistic analysis on a large number of queries has been performed and a scoring function has been calibrated based on the analysis. Details of this method are currently being summarized in a separate work. However, we do have a resulting high quality annotated dataset that can be used to evaluate the Field Sensor.

This matching approach was developed to provide a more effective solution to single citation queries. Single citation queries are those navigational queries that can be mapped to a unique PubMed article. Single citation queries are usually long queries and contain a full title and/or some combination of journal name, author, and volume, issue, page and date. Note that a *text* field is not represented in this dataset. A textual segment of a query is allowed to match part of a title, however, a matching between a query segment and an abstract is not available in this analysis. The reason is 2-fold. The first reason is efficiency—matching over abstracts with this algorithm is significantly more time consuming than matching over titles. Second, users seldom employ terms not present in the title in constructing a single citation query.

For this evaluation, we processed ∼3 million queries collected from a single day on 12 October 2016 and reduced it to a set of 102 971 queries that map to a unique PubMed document with a very high probability. We will refer to the set as the 103K Silver set. On manual review of 500 randomly selected queries from this set, we found the query parses to be 99% accurate.

Experiments and results

We evaluate the performance of the Field Sensor on the gold standard of 10K queries and the silver standard of 103K queries. The two test sets exhibit complementary properties. The 10K set does not distinguish between the four citation elements of *volume*, *issue*, *page* and *date*, which are all conflated into a single *citation* field. The 103K set, on the other hand, is enriched in citation annotations for each of these four, which allows us to evaluate the performance of the Field Sensor on these fields. Compared with the 10K set, the 103K set is also enriched in the *title* field. This is explained by the fact that these are the fields that help establish a unique mapping between a query and a PubMed article. Another aspect of the 103K dataset is that *text* annotations are not present because the abstract field was not included for matching.

Results and analysis on the 10K set

The Gold standard set of 10K queries, contains 9490 annotated queries and 510 queries for which there was no reasonable annotation found (these queries are removed from consideration). We will refer to the set as 10K_GS. The 9490 annotated queries contain 29 426 tokens. Some segments of gold standard queries are not annotated, which reduces the number of annotated tokens to 25 195.

Table 1. Token-based and query-based analysis of the Field Sensor

	Token-based comparison			Query-based comparison		
	P	*R*	*F*	*P*	*R*	*F*
Author	0.980	0.967	0.974	0.969	0.969	0.969
Text	0.932	0.957	0.944	0.957	0.980	0.968
Citation	0.953	0.918	0.935	0.964	0.935	0.949
Journal	0.882	0.926	0.904	0.796	0.891	0.841
Title	0.882	0.789	0.833	0.767	0.741	0.753

Precision/Recall/*F*-measure are computed for the five fields of interest: Author name, Key Word, Citation, Journal name and article Title.

To evaluate the performance of the Field Sensor we applied it to the 10K set. The analysis of the Field Sensor is reported on a query level (on 9490 queries) and token level (based on 25 195 annotated tokens). We refer to the set of predicted annotations as 10K_FS and compare them to the manual annotations 10K_GS. For a query level analysis, a sequence of fields predicted is correct if it matches the gold standard annotation. Otherwise, if at least one of the fields does not match the gold standard annotation, the prediction is considered incorrect. With a token-level analysis, the predicted field is compared with the gold standard field on a token basis. Of 9490 the Field Sensor correctly annotated 8798 queries, which constitutes 93.28% overall accuracy of the tool. Table 1 presents Precision, Recall and *F*-measure for the five fields computed on a token and a query level.

Misspellings

When computing the field predictions for the 10K set, we observed that 431 queries contained tokens not found in PubMed. Most of these non-found terms are misspellings such as '*surgical mask in operation theater*' or '*growthcartilage*'. The PubMed search is equipped with the autocorrect feature that would potentially be of help in these misspelled cases, but the spell checker is outside the scope of this study.

We analyzed the distribution of fields of the misspelled words by comparing them to the manually assigned tags in the 10K_GS. The analysis shows that in 60% of the cases, a misspelling is a text element, in 33% it is an author name, in 4% it is a title token. Journal name and citation misspellings each constitute 1.3%. Based on these statistics the Field Sensor labels the misspelled tokens as *text*.

The recall/precision and *F*-measure figures presented in Table 1 are computed based on that assumption.

Of 9490 queries, 692 contain at least one incorrectly annotated token, and the subsequent section presents a detailed analysis of the errors observed. The analysis revealed that 187 of the 692 queries (27%) contained a misspelling, i.e. a token not found in PubMed. The prediction of the Field Sensor on the misspelled token does not in any way demonstrate the performance of the tool. What is more, the presence of misspelled tokens hinders the evaluation of the Field Sensor. For example, when an author's last name is misspelled, it results in author initials following the last name not being linked to the last name. Or a misspelled title token results in the full title not being recognized. For error analysis, we do not consider the 187 queries that contain misspelled tokens, and closely examine the remaining 505 queries where an error has occurred for a reason other than misspelling.

Detailed error analysis on 10K set

We identify four sources of error that contribute the majority of differences detected on the 505 queries. The most differences are between text words and author names (32.26%), followed by text words and article titles (24.71%), text words and journal names (24.15%) and text words and citations (13.2%). These four classes cover ∼95% of the errors and for them we provide a detailed error analysis. The remaining cases are quite minor and affect only thirty queries.

Distinguishing between text words and author names

These types of differences occur in author names that are also frequently used common English words. For example, *sweet* is most frequently interpreted as *text*, but when searched for in PubMed as a last name *sweet[author]* retrieves 3923 PubMed documents. Such errors are also observed in eponymously named diseases, such as *Alzheimer's disease* where the last name is used within the disease name. In the 10K set, for 110 queries a text words is predicted as an author name, and in 61 queries an author name is predicted as text. Examples of queries in this category are presented in Figure 4.

Distinguishing text words and titles

Given the correlation between words in article abstracts and titles, the disambiguation of whether a user intends to perform a keyword search or an article title query can be crucial to returning the anticipated search results. The number of instances where a text is predicted as title is 66 queries. The number of instances where a title is predicted as text is 65 queries.

QUERY=summer and Kirkpatrick
10K_GS: summer[author] and Kirkpatrick[author]
10K_FS: summer[text] and Kirkpatrick[author]
QUERY=buller day stress
10K_GS: buller[author] day[author] stress[text]
10K_FS: buller[author] day[text] stress[text]
QUERY=musk and glucose
10K_GS: musk[text] and glucose[text]
10K_FS: musk[author] and glucose[text]
QUERY= diabetes + gravidarum + ketonen
10K_GS: diabetes[text] gravidarum[text] ketonen[text]
10K_FS: diabetes[text] gravidarum[text] ketonen[author]
QUERY=Tagetes marigold
10K_GS: Tagetes[text] marigold[text]
10K_FS: Tagetes[text] marigold[author]

Figure 4. Distinguishing between Text Words and Author names.

For some queries, it may not be completely clear whether a user is searching with keywords or article title. For example, a PubMed search with a query '*schizophrenia and multiple sclerosis*' returns 518 articles when interpreted as a keyword query and exactly 1 result [PMID: 3059470] when interpreted as a title. Moreover, examining the differences between the Field Sensor results and manual annotations we observed that annotators were not always consistent in distinguishing between the text and title fields. Examples of queries in this category are presented in Figure 5.

Distinguishing text words and a journal name

A strong correlation is also present between journal titles and text words, particularly for single token journal names such as *cancer, diabetes, circulation, blood, drugs* etc. Many journal names match frequent text words and present a challenge. The number of queries where a text word is predicted as a journal title/part of a journal title is 86 queries, and the number of queries where a journal title is predicted as text is 37.

Distinguishing text words and citation information

The number of instances where a text word is predicted as citation is 24, and the number of instances where a citation element is predicted as a text word is 46. This error can occur on numbers that are legitimately part of the text, but are interpreted as a citation element. For example, in a $QUERY = KLN\ 47$ we predict *47* to be a citation element. Another source of differences in this group are queries containing terms like *review*, where the term is interpreted as

QUERY=rsv pneumonia in postmortum infants
10K_GS: *rsv pneumonia in postmortum infants [title]*
10K_FS: *rsv pneumonia[text] in postmortum infants [text]*
QUERY=does aprotinin prevent stroke
10K_GS: *does aprotinin prevent stroke [title]*
10K_FS: *does[text] aprotinin[text] prevent[text] stroke [text]*
QUERY=self care theory and accidental patient falls
10K_GS: *self care theory and accidental patient falls[text]*
10K_FS: *self care theory and accidental patient falls[title]*

Figure 5. Distinguishing between Text Words and Titles.

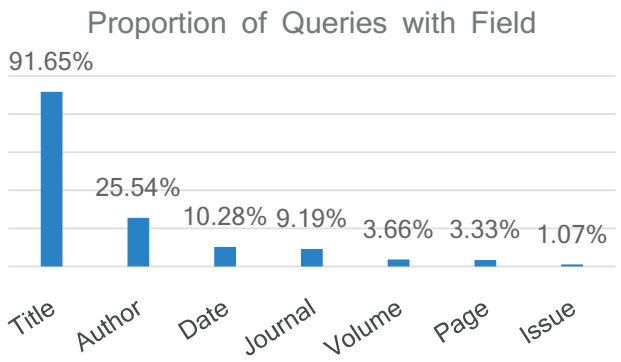

Figure 6. Distribution of fields computed for the 103K queries annotated with seven citation fields: Title, Author name, Date, Page, Volume, Issue and Journal Name.

citation information in the 10K_GS, while we interpret it as text. Another example in this group is a *QUERY= textbook of pediatrics and Nelson*. We interpret textbook as a text element, while in 10K_GS it is interpreted as citation information.

Results and analysis on the 103K set

The 103K Silver Standard dataset contains 102 971 queries and the total number of annotated terms in the set is 828 078. Figure 6 presents the distribution of fields in this dataset. These navigational queries are heavily enriched in article titles which appear in whole or in part in over 91% of queries. Author names are present in 25.5% of queries, followed by date in 10.23% of queries and journal name in 9.17% of queries. Even in citation queries, volume, issue and page are still quite minor.

Of the 102 971 queries, the Field Sensor completely agrees with the Silver standard annotations on 93 716 queries, which constitutes 91.01% overall accuracy of the tool. In terms of tokens, for 98.23% of tokens (813 404) we agree with the Silver Standard annotation, and disagree on the remaining 1.77%.

Table 2. Token-based and query-based analysis of the Field Sensor on the 103K set

	Token-based comparison			Query-based comparison		
	P	*R*	*F*	*P*	*R*	*F*
Title	0.990	0.993	0.991	0.986	0.987	0.986
Author	0.972	0.940	0.956	0.987	0.942	0.964
Date	0.910	0.955	0.932	0.978	0.955	0.967
Page	0.969	0.898	0.932	0.980	0.902	0.939
Volume	0.928	0.838	0.881	0.947	0.841	0.891
Issue	0.983	0.686	0.808	0.994	0.685	0.811
Journal	0.742	0.725	0.733	0.821	0.637	0.717

Precision/Recall/*F*-measure are computed for the seven fields of citation queries: Title, Author name, Date, Page, Volume, Issue and Journal Name.

We further compute the Precision, Recall and *F*-score on the token and query level for each one of the seven fields. The token level computation evaluates the fraction of correctly identified tokens within a field. The query level computation evaluates the fraction of queries with all field spans correctly identified. Table 2 presents the Precision. Recall and *F*-measure for the seven fields computed at the token and query level.

For the three most frequent fields in the 103K dataset (Title, Author names and Date) the performance of the Field Sensor is quite outstanding. Currently our efforts are directed toward improving the journal name recognition, however, since journal names appear in 9.19% of queries in this citation-rich set, the overall impact of this field is modest.

The utility of the field sensor: predicting query intent and query composition

PubMed logs record user interactions with PubMed such as search, retrieval and click through information. In this section, we demonstrate the application of the Field Sensor on the PubMed query logs. We predict query intent, examine the composition and lengths of PubMed queries, and distill the citation information search patterns as examples of the Field Sensor usage. The analysis is performed on a random day of PubMed queries logged in the system on Wednesday 12 October 2016 which contains 3 054 498 anonymized queries.

Using the Field Sensor, we can predict the query intent with high accuracy and web-scale speed (processing about 800 queries per second on a single thread). The web-scale processing makes it feasible to apply the Field Sensor to an arbitrarily large set of queries and classify them as information or navigational based on the Field Sensor predictions. Applied to 1 day of queries, we predict 47.68% of queries to be informational and 52.31% navigational. For

reference, the accuracy achieved on the 10K dataset is 95.24% on this binary classification task, where 53.08% of queries are predicted to be informational and 46.91% navigational. The gold standard split is 53.69% informational and 46.31% navigational. Note that the performance of our method on the binary classification task is higher than that on the task of predicting all eight fields.

The ability to distinguish informational and navigational queries allows us to better understand how users query PubMed. For example, we have observed a noticeable growth in query size compared with the average query size of 3.54 and median length of 3 reported in earlier studies from 2009 (18, 19). Based on the 12 October 2016 data, an average number of tokens per query is 5.18 and the median is 3. For computing these averages, we tokenized queries by defining the tokens as space separated sequences of characters, and excluded from the analysis noisy queries containing more than 100 tokens. To understand the reasons behind the growth in length we compute the average query length at six points in time. These points represent 1 day of PubMed logs collected on the same date of 20 January for six consecutive years 2012–17. As depicted in Figure 7, we observe that informational queries remain on average at about the same size, however, the length and proportion of navigational queries follows a growing trend. The average size of navigational queries has increased from 5.3 in 2012 to 7.0 in 2017. Compared with the 10K set, the proportion of navigational queries has also increased. This may reflect that search systems are becoming better at parsing long queries and users are becoming comfortable copy-pasting article title or the whole citation.

The availability of predicted field data allows us to better understand the complexity of the information-seeking process of users searching for citation information. We analyze query fields to uncover what are the most frequent ways of accessing citation information. The analysis of fields is conducted on the navigational queries from

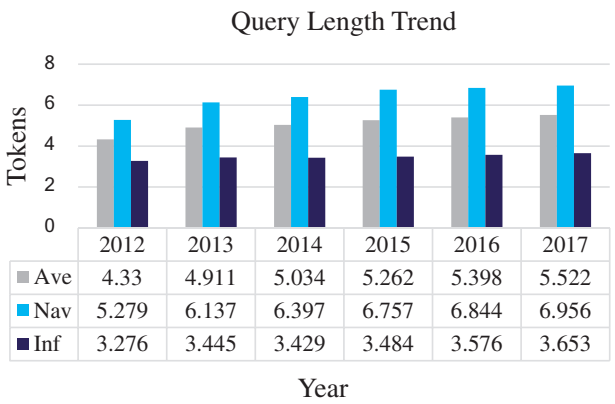

Figure 7. The average query length computed over query logs collected on 20 January for six consecutive years ranging from 2012 to 2017.

12 October 2016. To obtain the patterns, we group sequences of tokens from the same field into one entity. For example, queries consisting of two authors author1, author2 will be counted as an author query. Figure 8 presents the 13 most frequent patterns that account for >80% of the navigational queries. About 28% of navigational queries are author name queries. These are queries consisting of one or more author names. The next largest category are title queries that account for ∼24% of all navigational queries, and together with author queries comprise more than half of all navigational queries. Other popular search patterns are author name followed by text and text followed by author name. These two categories together comprise ∼12.5% of navigational queries. Queries consisting of PMIDs only also prove to be a popular way for accessing an article, and contribute 4.5% to navigational searches. The abundance of long title queries, with the average length of a title query being 11.57 tokens, explains the length of navigational queries. Figure 8 presents the distribution of query sizes in each of the categories outlined.

Conclusions and discussion

We here present the Field Sensor, a novel probabilistic tool for computing query composition and predicting query intent. The tool labels each segment of a query with a PubMed record field (text, title, author, journal, volume, issue, page and date). We evaluate the tool on a manually annotated dataset of 10K queries as well as a machine annotated dataset of 103K queries and demonstrate its superior performance. The software achieves a production-grade speed for PubMed.

The essential functionality of the Field Sensor is to infer the query intent. As a part of the PubMed search system, it is used to detect informational queries and refer a user to relevance ranked search results. Applied to a random day of PubMed queries, the Field Sensor predicts 48% to be informational and 52% navigational. To our knowledge, the Field Sensor is the first web scale tool for inferring intent and computing the composition of biomedical queries. In addition, field predictions allow us to study on a large scale how biomedical information is being searched for in PubMed. The underlying Field Sensor model is trained on Medline data, and the Field Sensor can be adapted to queries in another domain by being retrained on the corresponding domain data.

In the future, we plan to study an alternative method for computing the probabilities. In the current setting, the probabilistic field prediction is based on a language model, where the likelihood of a term in a field is computed as the frequency of the term in the field divided by the size of

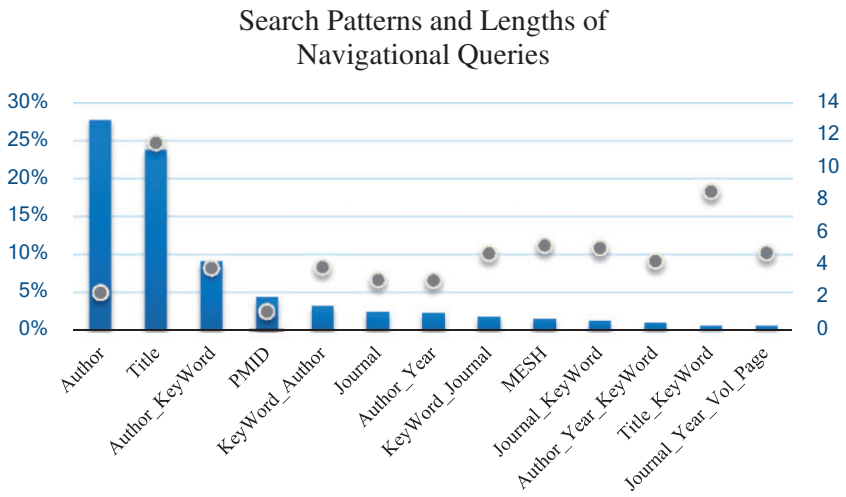

Figure 8. The 13 most frequent patterns for accessing citation information in PubMed that account for >80% of the navigational queries. Percentage of queries in each pattern are reflected in the Bar chart against the primary *Y* axis, and the average length of queries within that pattern are reflected with the scatter plot measured against the secondary *Y* axis.

the field over the whole database as measured by term numbers. Another plausible approach to probability computation is to compute the likelihood of a term as a ratio of the number of documents where that term appears and the size of the PubMed collection (~27 million documents). We believe this may neutralize some of the problems we are observing with journal names, but could introduce other errors. In the future, we also plan to pay closer attention to the misspelled tokens. They present a source of errors, not only because the misspelled token is not accurately labeled, but also because it hinders the analysis of the remainder of a query. While treating misspellings as *text* was reasonable, the first logical step to improve processing of those queries would be spelling correction. We are planning to incorporate a spell-checker in the next generation of the Field Sensor.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Conflict of interest. None declared.

References

1. Falagas,M., Pitsouni,E., Malietzis,G. *et al.* (2008) Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J.*, **22**, 338–342.
2. Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, **2011**, baq036.
3. Wildgaard,L.E. and Lund,H. (2016) Advancing PubMed? A comparison of 3rd-party PubMed/MEDLINE tools. *Library Hi Tech*, **34**, 669–684.
4. Ashkan,A., Clarke,C.L., Agichtein,E. *et al.* (2009) Classifying and characterizing query intent. In: *The Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*.
5. Broder,A. (2002) A taxonomy of web search. In: *The Proceedings of ACM SIGIR*.
6. Jansen,B.J., Booth,D.L. and Spink,A. Determining the user intent of web search engine queries. In: *WWW 2007*.
7. Mendoza,M. and Zamora,J. (2009) Identifying the intent of a user query using support vector machines. In: *Proceedings of the 16th International Symposium on String Processing and Information Retrieval, 2009*. Springer-Verlag Berlin, Heidelberg, Saariselkä, Finland, pp. 131–142.
8. Figueroa,A. and Atkinson,J. (2016) Ensembling classifiers for detecting user intentions behind web queries. *IEEE Internet Comput.*, **20**.
9. Hashemi,H.B., Asiaee,A. and Kraft,R. (2016) Query intent detection using convolutional neural networks. In: *International Conference on Web Search and Data Mining, Workshop on Query Understanding, 2016*.
10. Hu,J., Wang,G., Lochovsky,F. *et al.* (2009) Understanding user's query intent with wikipedia. In: *Proceedings of the 18th International Conference on World Wide Web, 2009*. ACM New York, New York, NY, USA, Madrid, Spain.
11. Ren,X., Wang,Y., Yu,X. *et al.* (2014) Heterogeneous graph--based intent learning with queries, web pages and wikipedia concepts. In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM'14)*. ACM New York, New York, NY, USA, pp. 23–32.
12. Kale,A., Taula,T., Hewavitharana,S. *et al.* (2017) *Towards semantic query segmentation. In: SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17)*, Tokyo, Japan.
13. Kim,J., Xue,X. and Croft,W.B. (2009) A probabilistic retrieval model for semistructured data. In: *European Conference on Information Retrieval*.
14. Nikolaev,F., Kotov,A. and Zhiltsov,N. (2016) Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in*

Information Retrieval. ACM New York, New York, NY, USA, Pisa, Italy, pp. 435–444.
15. Sarkas,N., Paparizos,S. and Tsaparas,P. (2010) Structured annotation of web queries. *ACM SIGMOD International Conference on Management of Data (SIGMOD'10)*. ACM New York, New York, NY, USA, Indianapolis, IN, USA.
16. Radlinski,F., Szummer,M. and Craswell,N. Inferring query intent from reformulations and clicks. *WWW 2010*.
17. Pitler,E. and Church,K. (2009) Using word-sense disambiguation methods to classify web queries by intent. In: *2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1428–1436.
18. Islamaj Dogan,R., Murray,C., Névéol,A. *et al*. (2009) Understanding PubMed user search behavior through log analysis. *Database*, **2009**, bap018.
19. Herskovic,J., Tanaka,L.Y., Hersh,W. *et al*. (2007) A day in the life of PubMed: analysis of a typical day's query log. *J. Am. Med. Inform. Assoc.*, **14**, 212–220.
20. Wilbur,W.J., Kim,W. and Xie,N. (2006) Spelling correction in the PubMed search engine. *Inf. Retr.*, **9**, 543–564.
21. Hersh,W. and Voorhees,E. (2009) TREC genomics special issue overview. *Inf. Retr.*, **12**, 1–15.
22. Bampoulidis,A., Lupu,M., Palotti,J. *et al*. (2016) Interactive exploration of healthcare queries. In: *14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, Bucharest, Romania.
23. Tsikrika,T., Muller,H. and Kahn,C.E.J. (2012) Log analysis to understand medical professionals' image searching behaviour. In: *Proceedings of the 24th European Medical Informatics Conference (MIE2012)*. European Federation for Medical Informatics and IOS Press, Pisa, Italy, pp. 1020–1024.
24. White,R. and Horvitz,E. (2014) From health search to healthcare: explorations of intention and utilization via query logs and user surveys. *J. Am. Med. Inf. Assoc.*, **21**, 49–55.
25. Zhang,Y. (2014) Searching for specific health-related information in MedlinePlus: behavioral patterns and user experience. *J. Assoc. Inf. Sci. Technol.*, **65**, 53–68.
26. Ogilvie,P. and Callan,J. Combining document representations for known-item search. In: *SIGIR 2003*.
27. Bernstam,E., Herskovic,J. and Hersh,W. (2009) Handbook of Research on Web Log Analysis.
28. Hersh,W. (2009) Information Retrieval: A Health and Biomedical Perspective.
29. Li,X., Schijvenaars,B.J.A. and De Rijke,M. (2017) Investigating queries and search failures in academic search. *Inf. Process. Manag.*, **53**, 666–683.
30. Khabsa,M., Wu,Z. and Giles,C.L. (2016) Towards better understanding of academic search. In: *JCDL 2016, The 16th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, Newark, NJ, USA.
31. Fiorini,N., Lipman,D.J. and Lu,Z. (2017) Cutting edge: towards PubMed 2.0. *eLife*, **6**:e28801.
32. Manning,C., Raghavan,P. and Schütze,H. (2009) *An Introduction to Information Retrieval*. Cambridge University Press.
33. Viterbi,A.J. (1967) Error bounds for convolutional codes and asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.