applied sciences

MDPI

Article

A Query Expansion Method Using Multinomial Naive Bayes

Sergio Silva [1,2,3,*], Adrián Seara Vieira [1,2,3], Pedro Celard [1,2,3], Eva Lorenzo Iglesias [1,2,3] and Lourdes Borrajo [1,2,3]

[1] Computer Science Department, Escuela Superior de Ingeniería Informática, Universidade de Vigo, 32004 Ourense, Spain; adrseara@uvigo.es (A.S.V.); pedro.celard.perez@uvigo.es (P.C.); eva@uvigo.es (E.L.I.); lborrajo@uvigo.es (L.B.)
[2] CINBIO-Biomedical Research Centre, Universidade de Vigo, 36310 Vigo, Spain
[3] SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36312 Vigo, Spain
* Correspondence: smachado@alumnos.uvigo.es

Abstract: Information retrieval (IR) aims to obtain relevant information according to a certain user need and involves a great diversity of data such as texts, images, or videos. Query expansion techniques, as part of information retrieval (IR), are used to obtain more items, particularly documents, that are relevant to the user requirements. The user initial query is reformulated, adding meaningful terms with similar significance. In this study, a supervised query expansion technique based on an innovative use of the Multinomial Naive Bayes to extract relevant terms from the first documents retrieved by the initial query is presented. The proposed method was evaluated using MAP and R-prec on the first 5, 10, 15, and 100 retrieved documents. The improved performance of the expanded queries increased the number of relevant retrieved documents in comparison to the baseline method. We achieved more accurate document retrieval results (MAP 0.335, R-prec 0.369, P5 0.579, P10 0.469, P15 0.393, P100 0.175) as compared to the top performers in TREC2017 Precision Medicine Track.

Keywords: query expansion; information retrieval; multinomial naive bayes; relevance feedback

Citation: Silva, S.; Seara Vieira, A.; Celard, P.; Iglesias, E.L.; Borrajo, L. A Query Expansion Method Using Multinomial Naive Bayes. *Appl. Sci.* **2021**, *11*, 10284. https://doi.org/10.3390/app112110284

Academic Editor: Arturo Montejo-Ráez

Received: 15 September 2021
Accepted: 28 October 2021
Published: 2 November 2021

1. Introduction

Information Retrieval (IR) is a field of computer science that processes text documents and retrieves those that are more similar to a user query based on the resemblance of the contents of the documents and the keywords of the query. In a particular way, the task of information retrieval is gaining importance in the field of biomedicine.

The exponentially growing amount of clinical data makes it remarkably difficult to extract relevant information that meets the needs of each individual user [1]. The most well-known techniques make use of keywords to search for specific items that contain them, but language semantics, polysemy, synonymy, and hyponymy make keywords useless in many cases [2]. Therefore, the retrieval process in information retrieval systems must be improved in order to deal with all this complexity and deliver appropriate results that meet what the user is looking for.

One of the most widely-used techniques to improve the retrieval process is query expansion (QE). QE is the process of reformulating a given query in order to retrieve the more suitable documents that meet a user's needs. Over the years, several query expansion techniques have been analyzed, but even recent elaborate architectures are having problems surpassing the performance of classic techniques [3]. Because of this, our work is focused on the extraction of terms that expand the original query in order to improve the relevance of the retrieved documents.

Related Works

So far, several authors have worked on diverse research lines related to query expansion, the improvement of efficiency and performance being the common aim, in order to offer more relevant information to the user and better fulfill their needs.

Zhu et al. [2] evaluated the use of auxiliary collections to address polysemy, synonymy, and hyponymy in clinical text retrieval. These semantic relations complicate the retrieval process as different words can relate to the same topic. In order to deal with this problem, they proposed a pseudo-relevance feedback method that looks for new terms in the auxiliary collections in order to expand the initial query. The authors concluded that the use of all available data, in some cases, is inadequate and may not lead to improvements in the recovery system. In these cases, the authors suggested additional resources and a selection of the collection that is suitable for the query.

Ehman et al. [4] proposed the Normalized Difference Measure metric, a measure that takes into account the relative frequency of documents and terms in order to improve text classification. This metric analyzes all the terms found in the documents and benefits from the inclusion of new relevant terms in a query when used as a classifier in information retrieval systems.

Araújo et al. [5] implemented a pseudo-feedback query expansion method that allows the user to select expansion words from a list of possible relevant terms. The authors used the top three retrieved documents to extract terms based on document and word frequencies, word length, and query length. The obtained results showed an overall improvement in the number of relevant retrieved documents, despite the fact that the results in some cases were not better than the base case due to the low number of relevant documents.

Afun et al. [6] suggested a combination of several query expansion methods such as Ontologies, Association Rules, WordNet, Methathesaurus, Synonym Mapping, Local Co-occurrence, and Latent Semantic Indexing. The authors noted several limitations to the previous techniques (e.g., performance reduction, term relationship loss), emphasizing the importance of choosing the right technique for each specific case.

Agosti et al. [7] reviewed multiple query expansion techniques that had been applied to information retrieval systems used in clinical trials. The authors concluded that it is not possible to build an expansion technique pattern that correctly applies to a huge text corpus. They reported that the use of weighted keyword expansion and query reduction (removal of words that are not relevant) improved the performance of information retrieval in clinical trials.

Xu et al. [8] proposed a supervised query expansion model that could be applied to highly diverse biomedical datasets. The authors performed a term extraction for each query, proceeded to assign labels to each term, and then ranked them to know which were the most relevant. Owing to these three steps, and with the use of rank weights, the authors were able to enhance the queries and improve the performance of biomedical information retrieval.

Azad and Deepak [1] surveyed multiple query expansion techniques, weighting methods, and ranking methodologies for information retrieval. They found that the most frequent queries are composed of one, two, or three words, which increases its ambiguity and makes the retrieval process difficult. This exposed the increasing need for query expansion techniques to enhance the original queries with the use of relevant terms, making it easier for information retrieval systems to obtain more suitable elements.

McDonald et al.[3] proposed a technique that extends deep learning architectures where queries and documents are analyzed together in order to obtain their similarities. The authors claimed that even state-of-the-art complex architectures do not improve the performance of classic algorithms such as BM25 and BM25+extra. Their proposed method helps to achieve better performance, offering an improvement of BM25, although unable to surpass BM25+extra in some cases.

Wang et al. [9] implemented a pseudo-relevance feedback technique to expand the queries using terms extracted from the top-ranked documents returned in a first search. The authors used Rochio+BM25 to extract the expansion terms, outperforming baseline models in terms of MAP and precision at different positions.

This paper shows the work developed to expand an initial query from a set of first-retrieved documents using the Multinomial Naive Bayes technique as an autonomous selector of terms. The proposed technique improves performance and helps information retrieval systems to obtain a higher number of relevant documents for specific queries.

It is organized according to the following structure: Section 2 presents a detailed overview of the information retrieval process, techniques used in text preprocessing and representation, the selection of attributes, and the measures used in the evaluation of the information retrieval system. Section 2.3 describes the procedures carried out and the proposed query expansion technique. Section 3 describes the evaluation methodology followed and the results obtained. Finally, in Section 4, the conclusions of the study are presented and future perspectives are discussed.

2. Materials and Methods

The general information retrieval process obtains documents that are relevant to a given query. This involves tasks related to text preprocessing, document indexing, and the execution of an initial query that could then be expanded to obtain better results. Figure 1 shows an overview of the process.

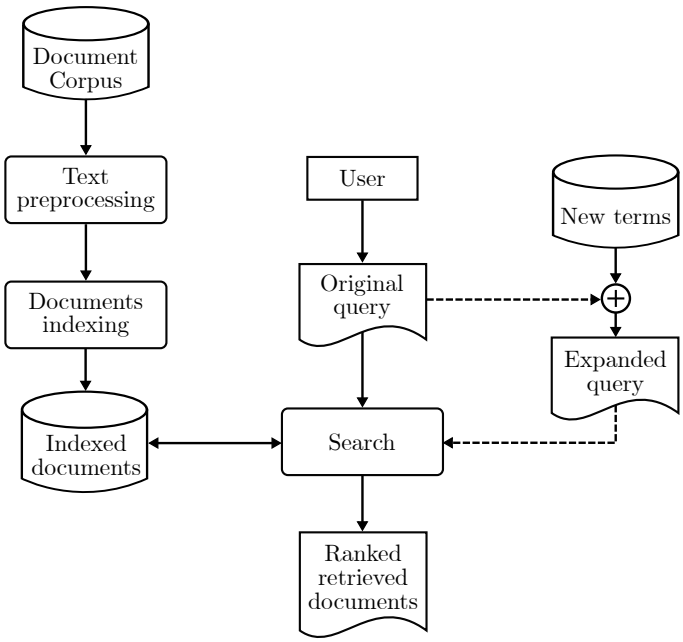

Figure 1. Information retrieval—general process.

2.1. Text Preprocessing and Matching

The information retrieval process involves data preprocessing and matching. The *preprocessing* step includes tasks related to tokenization, stopword removal, sentence detection, stemming, lemmatization, and term weighting.

Tokenization allows for the transformation of a document into words using white spaces, commas, periods, and tab delimiters as separators during the token-building process. According to [10], there are different text delimiters that can lead to a complex process of tokenization. Since most scientific documents are written in English, the recognition and extraction of tokens is carried out considering a specific set of characters. A whole set of special characters is disregarded as they contribute nothing to the knowledge and only function as token separators. Among them are: (,), /, {, }, [,], :, ;.

There are some issues that need to be taken into account, such as the identification of abbreviations, dates, acronyms, and letter capitalization. Case transformation allows for the standardization of the words contained in documents, thus dropping different versions of the same word. Given that the stopword list is presented in lowercase, all the letters are transformed to its lowercase variant.

Stopword removal is based on the elimination and non-consideration of words that are very frequent and offer little significance. The main advantage of this procedure is the reduction of data size, and thus the decrease of computational cost and the improvement of accuracy. There are lists of stopwords available for the English language. However, new terms may be added to these lists depending on the structure of the documents and the needs of particular circumstances.

Stemming is a process of reducing words to their word stem, preserving only the morphological root. Suffixes of words such as plurals and gerunds, among others, are removed. According to the literature, *Porter Stemmer* and *Krovetz Stemmer* are the most frequently used stemmers in information retrieval systems in the English corpora. Porter Stemmer was developed by Martin Porter at Cambridge University in 1980 and was first published in Porter, M.F [11]. It is a process of removing word suffixes, such as gerunds and plurals, and replacing inflectional endings. It consists of rules dealing with a specific suffixes and according to certain conditions. Lemmatization uses dictionaries and a morphological analysis of words in order to reflect the base form of a word, consequently collapsing the inflectional forms.

Document indexing is based on the frequency of the words that each document contains. Words with a high number of repetitions have a higher frequency, while the others have a lower frequency [12]. This is not always desirable behavior as some words such as and, the, and or appear frequently in documents but do not offer relevant information. One of the most widely applied algorithms is Term Frequency-Inverse Document Frequency (TF-IDF) [13], a statistical measure that assesses the importance of a word for a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the document collection, which recognizes the fact that some words are generally more common than others.

The matching step allows for the calculation of the similarity between documents and queries, with an associated weighting of terms. In general, a retrieval system returns a list of ordered documents where the first is the document most similar to the query. Taking this into account, it is possible to reformulate the query and expand it to be more representative of the need of the user; this technique is known as query expansion. According to the literature, query expansion techniques can be classified as: query-specific, corpus-specific, or language-specific [14].

Query-specific terms are based on the extraction of new terms from a subset of documents retrieved by a specific query. It is an approach of relevance feedback systems in which the new terms are obtained from a set of relevant documents. Although this technique is widely used and very effective, it requires users to indicate which documents are relevant.

In the corpus-specific technique, the entire content of a specific full-text database is analyzed to identify terms that are used in similar ways. This can be performed manually (although this requires a lengthy and ad hoc process) or automatically. Traditional automatic thesaurus construction techniques group words based on their patterns of occurrence at the document level [15,16]; that is, words that often occur together in documents are considered similar. These thesauri can be used for automatic expansion or manual consultation.

Language-specific is a technique present in online thesaurus that is not adapted for any specific text collection. Liddy and Myaeng [17] used Longman's Dictionary of Contemporary English, a semantically encoded dictionary. Others such as Voorhees [18] turned to WordNet [19], a network of lexical relationships built by hand. Borrajo et al. [20]

studied the use of dictionaries in the classification of biomedical texts with three different dictionaries (BioCreative [21], NLPBA [22], and an ad hoc subset of the UniProt database called Protein [23]).

In this work, Indri [24] is used as the search engine to perform the matching between a given query and a set of documents. Indri uses a combination of language modeling and inference networks for the information retrieval procedure. It is able to evaluate a query against a previously indexed corpus, returning a collection of the most relevant documents.

Indri uses a Dirichlet likelihood function for query evaluation prior to term weight smoothing. This function takes into account the frequency of words in a document and in the document collection, and a parameter μ, which takes the value of 2500 by default [25]. The score returned by the Dirichlet probability function is given by:

$$\log\left([C(W,D) + \mu * C(W,C)/|C|)/(|D| + \mu)\right] \quad (1)$$

- $C(W,D)$ represents the word count in the document D;
- $C(W,C)$ represents the word count in the document collection;
- $\mu = 2500$ default.

2.2. Corpus

In this work, the Clinical Trial corpus is used, which is composed of a set of clinical documents, topics (descriptions of the user needs), and relevance judgments performed by specialists in the field [26]. Roberts et al. [27] discussed in greater detail how the corpus was created and showed multiple works using it as an experimental corpus. Clinical Trials are available on the TREC official web page http://www.trec-cds.org/2017.html, accesed on 22 July 2021. The database contains 241,006 documents in txt and xml format. For this work, the *txt* format is selected.

Given that the main objective of this work is to present a new technique for query expansion, all the topics available are used. The topics consist of *disease*, *genetic variants*, *demographic*, and potentially other information about the patients.

The relevance judgments file corresponding to the Clinical Trial collection contains the relevant documents to each query, except for the query related to topic 10. In this case, there is no relevant information, and the query associated with this topic is disregarded.

In general, the documents contain a title, a detailed description of what is carried out in the study, information on the patient condition, intervention, and eligibility factors (these may include gender, age, or the respective criteria for inclusion or exclusion from the study). All documents are indexed with all its content, which means that no specific field in the document is selected.

In order to index the corpus, the documents were preprocessed using Porter stemming, and a list of stopwords for the English language were removed https://www.ranks.nl/stopwords, accessed on 15 May 2021. The terms *age*, *condition*, *detailed*, *eligibility*, *exclusion*, *inclusion*, *intervention*, *title*, *criteria*, *description*, *gender*, and *summary* were added to the stopword list because they were terms related to the names of the field labels in the documents; therefore, they were not relevant. All 241,006 documents were indexed for retrieval. Figures 2 and 3 show examples of document and topic structures used in the Corpus Clinical Trial.

TITLE:
Information Presentation Formats
CONDITION:
Meningioma
INTERVENTION:
Check Symptoms
SUMMARY:
Prevention and early detection of medical problems
can greatly reduce health care costs ...
DETAILED DESCRIPTION:
We will present individuals with medically accurate
information about a medical condition and measure ...
ELIGIBILITY:
Gender: All
Age: 18 Years to N/A
...

Figure 2. Clinical Trial—sample document.

```
<topic number="1">
    <disease>Liposarcoma</disease>
    <gene>CDK4 Amplification</gene>
    <demographic>38-year-old male</demographic>
    <other>GERD</other>
</topic>
```

Figure 3. Clinical Trial—sample topic.

2.3. System Architecture

The new query expansion technique presented in this study is based on relevance feedback and is presented in Figure 4.

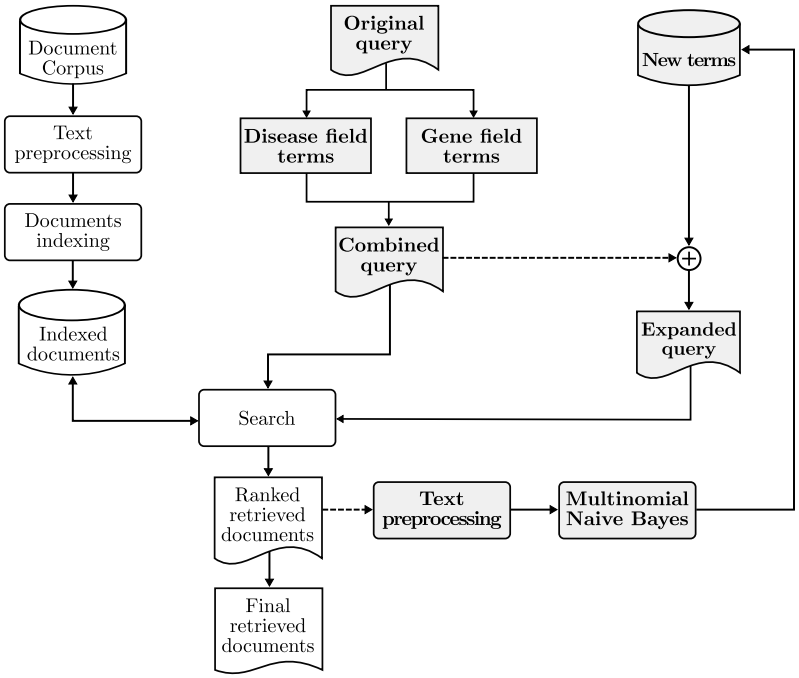

Figure 4. The proposed method.

The main elements of the proposed technique are the Original Query (OQ) found in the corpus, a combination of the terms found in the data fields of the OQ called Combined Query (CQ), and an Expanded Query (EQ) obtained from a combination of the words of

the CQ and new terms from the relevant documents retrieved by the CQ in a first search. From this point, new searches could be performed to further improve the query.

2.3.1. Combined Query (CQ)

The CQ was obtained by using a combination of terms referring to the fields *disease* and *gene*. In general, documents containing terms related to these fields were retrieved. More specifically, it was expected that all documents containing the terms related to the disease and with each of the genes (and their variants, if any) would be retrieved.

In this study, the language modeling tool Lemur (Lemur Project) was used for indexing and query execution. Lemur is a software tool designed to facilitate research in language modeling and IR, using weighting algorithms that provide query analysis methods, document indexing, and query-related document retrieval. This tool was developed by the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts, and by the Institute of Language Technologies (LTI) at Carnegie Mellon University. It is an open source and freely accessible that incorporates Indri as its query language,https://www.lemurproject.org/ software accessed on 12 April 2021.

The Indri query language is based on the Inquery language. It allows for the building of complex queries, and its grammar provides options for term detection, proximity, synonyms, wildcard operations, field restriction, combined beliefs (operators), filter operators, numeric and date field operators, document priors, etc.

Some operations used in this work are:

- *#band(w1 w2 ... wn)* returns documents containing all the terms *w1, w2, ..., wn*;
- *#combine (w1 w2 ... wn)* returns a scored list of documents that contains at least one of the terms;
- *#syn(w1 w2 ... wn)* returns the score of documents containing one of the terms *w1, w2,..., wn*, but considering these as synonyms.

The combined query is written as *#band(Disease GENE variant)*, but if the topic has more than one variant, it is changed to *#band(Disease GENE variant1)*, *# band(Disease GENE variant2)*, etc. For example, for topic 1, we get the combined query *#band(liposarcoma cdk4 amplification)*.

2.3.2. Extraction of New Terms

A set of 29 CQ was executed, saving the recovered documents as a training base consisting of 29 categories, one for each query. Each category is composed of the documents retrieved by the respective query (under category 1 we have the documents retrieved by query 1, and so on).

The training database built upon the retrieved documents was preprocessed the same way as the original documents (stemming, tokenization, stopword removal, case converter, and weighting). To simultaneously carry out the aforementioned operations on the documents, the free WEKA (Waikato Environment for Knowledge Analysis) software was used. It is available on Waikato official web page, https://www.cs.waikato.ac.nz/ml/weka, accessed on 13 May 2021. Weka includes data analysis tools such as textual data preprocessing, filtering, Naive Bayes Multinomial algorithm execution, and data visualization.

2.3.3. Attribute Selection

Attribute selection aims to reduce the number of attributes present in the data. More specifically, in text documents, these attributes refer to words that contain irrelevant information. The application of a classifier is performed on a smaller number of attributes considered the most relevant, which leads to the acquisition of more relevant terms or attributes for each category. The *GainRatio* technique selects attributes that maximize the information gain while minimizing the number of values of an attribute. After calculating the relevance for each attribute, a ranking is generated and the attributes of that ranking are selected, according to a *threshold* value. In this case, this value was 0.

2.3.4. Multinomial Naive Bayes

The expansion of queries is the process of reformulating a given query (Combined Query) in order to improve the performance of the information retrieval system. An evaluation of the initial consultation is carried out, which is expanded with new additional terms in order to be able to retrieve more relevant documents. In general, the expansion of queries may involve the search for synonyms or semantically related words. Moreover, it may employ associated procedures to correct spelling errors, reduce terms to a morphological form, or reweight the terms of the initial consultation, among others. In this study, a Query-specific term approach was adopted using relevant feedback.

A small set of documents was retrieved from an initial consultation, and all of them were considered relevant without any intervention from the user [28]. The content of the retrieved documents was used to obtain the new terms for the CQ expansion. The new query (Expanded Query) was obtained by combining the new terms and the CQ.

The extraction of the new terms is based on the probability that a word belongs to a given category. Once the training base (list of retrieved documents for a query) is organized by categories and the attribute selection is performed, the Multinomial Naive Bayes algorithm is applied.

The Naive Bayes algorithm is widely used in works involving text classification. It is based on probabilistic techniques, assuming the independence of variables. It is assumed that the presence or absence of a given characteristic of a category is not associated with the presence or absence of any other characteristic that is given that category.

The Multinomial Naive Bayes model considers how often the word occurs in documents x_t instead of the binary occurrence. It is calculated as follows, where $|V|$ represents the length of the vocabulary, and $n(C_i)$ is the total number of words in the category C_i:

$$P(d_j|C_i) = \prod_{t=1}^{|V|} P(w_t|C_i)^{x_t} \tag{2}$$

$P(w_t|C_i)^{x_t}$ is the probability of a term w_t occurring in a category C_i, and $n(w_t, C_i)$ is the number of occurences of w_t in the category C_i, as given by:

$$P(w_t|C_i) = \frac{1 + n(w_t, C_i)}{|V| + n(C_i)} \tag{3}$$

Finally, the classification is given by the maximizing function:

$$c^*(d) = argmax_{C_i} P(C_i) \prod_{t=1}^{|V|} P(w_t|C_i)^{x_t} \tag{4}$$

Therefore, the Multinomial Naive Bayes model is a reliable alternative for categorizing documents. In this case, instead of relying on binary values, it uses the frequency of the term. That is, it takes into account the number of times a word or *token* occurs in a document (also called gross frequency) [29]. In particular, the Multinomial Naive Bayes algorithm calculates the probability of a word belonging to a given category.

In this study, for each category (topic), the words w_t that verify the condition $P(w_t|C_i)^{x_t} > 0$ were considered as new terms for the query C_i expansion.

2.4. Expanded Query

In the first stage, the (*CQ*) was generated. It was from here that the expansion of the queries was processed. After the training documents for each query (category) were established, the Naive Bayes Multinomial algorithm was applied.

The CQ was obtained by the terms referring to the fields *disease* and *gene*: *#band(Disease GENE variant)*. Documents containing terms referring to these fields were retrieved at the same time. This initial consultation was performed on the indexed corpus. When the gene had more than one variant, the CQ was written as *#band(Disease GENE variant1)*

#band(Disease GENE variant2). The *band* method uses an *AND* boolean operator, so all documents containing all the terms related to the disease and the genes (and their variants, if any) were retrieved.

Finally, an expanded query (EQ) was built as *#combine*(t_1 t_2 ... t_n n_1 n_2 ... n_n), employing the boolean operation *OR*. The terms t_1, t_2, ..., t_n are the words contained in the *disease* and *gene* fields of the combined query, and n_1, n_2, ..., n_n are the new terms extracted by the described process. The EQ was, again, performed over the full indexed corpus.

3. Results and Discussion

After the execution of the queries, the measures were extracted using the *trec_eval* tool. It receives the recovered documents and the qrels file as parameters. This tool has been officially developed for its use in many of the tasks organized by the Text REtrieval Conference (TREC). For each query, the values of MAP, R-prec, and P@n were recorded for $n \in \{5, 10, 15, 100\}$. This procedure was exactly the same for both CQ and EQ.

Among the most frequently used measures in information retrieval are MAP, R-prec, and P@n. The Mean Average Precision (MAP) is the mean of the average precision scores for each query:

$$MAP = \frac{\sum_{q=1}^{Q} Ave(P)}{|Q|} \tag{5}$$

The average precision ($Ave(P)$) emphasizes the assignment of a higher ranking to relevant documents. It is the average of the precision of each of the relevant documents in the ranked sequence:

$$Ave(P) = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{\text{number of relevant documents}} \tag{6}$$

- k is the rank in the sequence of retrieved documents;
- n is the number of retrieved documents;
- $rel(k)$ is a binary function that assumes the value of 1 if the item at rank k is a relevant document, and zero if otherwise;
- $P(k)$ is the precision at cut-off k on the list.

R-prec is the precision after R documents have been retrieved, where R is the number of relevant documents for the topic. P@n is the accuracy of the first n documents recovered.

After the evaluation of the results in terms of the average values of the aforementioned measures, there is a clear improvement obtained by the expanded query. As can be seen in Figure 5, there is an increase of approximately 30% in the value of the MAP measure, from 0.261 to 0.335, with the use of query expansion. Regarding the R-prec measure, there is a general improvement of 12%. In relation to P@5 and P@10, the improvement is still significant at about 12% and 13%, which means that even with an increase in the considered number of the first retrieved documents, the system remains robust. In the measure P@15, the improvement was about 12%, while in P@100, it was about 24%.

In Table 1, the *MAP*, *R-prec*, *P*@5, *P*@10, *P*@15, and *P*@100 values obtained for each combined and expanded query are recorded. This shows a general improvement in all queries resulting from the expansion of the initial consultation.

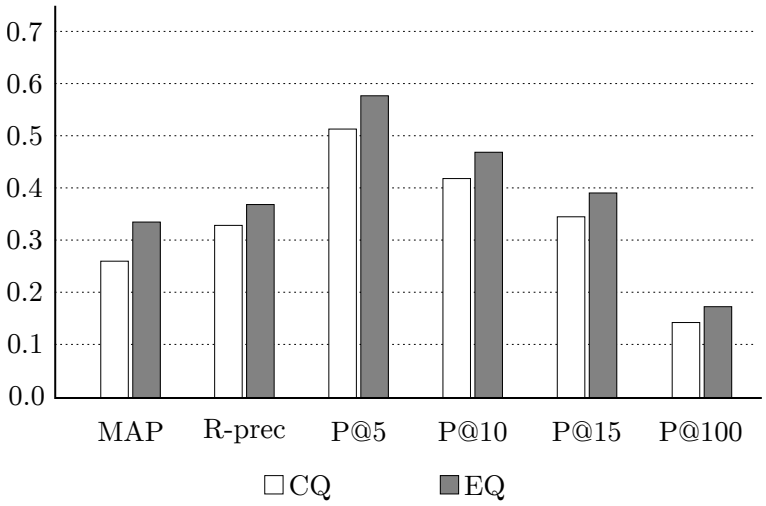

Figure 5. Mean values of measures.

Table 1. Measures of the evaluation of the combined and expanded queries.

	MAP		**R-prec**		**P@5**		**P@10**		**P@15**		**P@100**	
Query	**CQ**	**EQ**	**CQ**	**EQ**	**CQ**	**EQ**	**CQ**	**EQ**	**CQ**	**EQ**	**CQ**	**EQ**
1	0.293	**0.408**	0.353	0.353	1.000	1.000	0.500	**0.600**	0.333	**0.400**	0.050	**0.080**
2	0.243	**0.345**	0.394	**0.402**	**0.800**	0.600	0.700	0.700	0.733	0.733	0.440	**0.450**
3	0.405	**0.615**	0.417	**0.625**	1.000	1.000	**0.900**	0.800	0.667	**0.733**	0.100	**0.200**
4	0.410	**0.454**	0.491	**0.509**	0.600	**0.800**	0.600	**0.700**	**0.600**	0.533	0.360	**0.380**
5	**0.205**	0.173	0.194	0.194	**0.200**	0.000	0.100	**0.200**	0.133	**0.200**	**0.210**	0.170
6	0.404	**0.411**	**0.444**	0.370	1.000	1.000	**0.700**	0.500	**0.600**	0.533	0.160	**0.170**
7	0.369	**0.571**	0.390	**0.546**	0.600	**1.000**	0.600	**0.900**	0.667	**0.867**	0.480	**0.680**
8	0.450	**0.504**	**0.541**	0.525	0.600	**0.800**	0.600	**0.700**	0.667	0.667	0.420	0.420
9	0.314	**0.520**	0.339	**0.532**	0.600	**0.800**	0.600	**0.800**	0.467	**0.600**	0.300	**0.460**
11	0.319	**0.402**	0.316	**0.368**	0.600	**0.800**	0.400	**0.600**	0.333	**0.400**	**0.150**	0.140
12	0.118	**0.266**	0.231	**0.256**	0.600	**0.800**	0.300	**0.700**	0.200	**0.467**	0.100	**0.190**
13	0.090	**0.161**	**0.324**	0.206	0.200	0.200	0.200	0.200	**0.267**	0.200	0.110	**0.120**
14	**0.579**	0.563	0.714	0.714	0.800	0.800	0.500	0.500	0.333	0.333	0.050	**0.060**
15	0.250	**0.253**	0.250	0.250	0.200	0.200	0.100	0.100	0.067	0.067	0.010	0.010
16	0.300	**0.327**	0.400	0.400	0.400	0.400	0.200	0.200	0.133	0.133	0.020	**0.040**
17	0.150	**0.191**	0.303	0.303	0.400	0.400	0.400	0.400	0.400	**0.533**	0.090	0.090
18	0.000	**0.044**	0.000	**0.079**	0.000	**0.200**	0.000	**0.200**	0.000	**0.133**	0.020	**0.050**
19	0.044	**0.307**	0.044	**0.304**	0.200	**0.400**	0.100	**0.400**	0.067	**0.267**	0.010	**0.130**
20	0.200	**0.234**	0.200	0.200	0.200	0.200	0.100	0.100	0.067	0.067	0.040	0.040
21	0.088	**0.246**	0.209	**0.269**	0.400	**0.600**	**0.500**	0.300	**0.400**	0.333	0.190	**0.240**
22	0.059	**0.087**	0.118	**0.147**	**0.600**	0.400	**0.600**	0.400	**0.400**	0.333	0.120	**0.160**
23	**0.243**	0.236	**0.367**	0.333	**0.400**	0.200	**0.500**	0.400	**0.400**	0.333	**0.190**	0.170
24	0.333	**0.580**	0.333	**0.556**	1.000	1.000	0.600	**0.900**	0.400	**0.667**	0.060	**0.110**
25	0.265	**0.459**	0.375	**0.475**	0.600	**0.800**	0.600	**0.900**	0.467	**0.733**	0.210	**0.250**
26	0.213	**0.225**	0.200	0.200	0.200	0.200	0.100	0.100	0.067	0.067	0.010	0.010
27	0.250	**0.313**	0.393	**0.429**	0.400	**0.800**	0.500	0.500	0.400	0.400	0.080	**0.110**
28	0.250	0.250	0.500	0.500	0.200	0.200	0.100	0.100	0.067	0.067	0.010	0.010
29	0.277	**0.385**	0.250	**0.375**	0.400	**0.600**	0.200	**0.300**	0.133	**0.200**	0.010	**0.020**
30	**0.263**	0.198	**0.600**	0.290	0.600	0.600	0.400	0.400	0.400	0.400	0.110	**0.120**
ALL	0.261	**0.335**	0.330	**0.369**	0.514	**0.579**	0.418	**0.469**	0.347	**0.393**	0.142	**0.175**

The bold font refers to the highest value for each measure.

We can draw a comparison between the obtained results and the outcome of the participants of the TREC2017 Precision Medicine Track. As reported in [27], Table 8 shows the best, median, and worst results per topic from over 133 runs at P@5, P@10, and P@15 for the TREC 2017 Precision Medicine Track using clinical trials. Our method clearly outperforms most of median results in all precision ranges. More specifically, it achieves better precision than the median in 25 out of 29 topics for P@5, 26 out of 29 topics for P@10, and 26 out of 29 topics for P@15. In order to better analyze the results, Figures 6–8 show a comparison of the mean precision of the participants of the TREC2017 Precision Medicine Track and that of the proposed method in the top 5, 10, and 15 retrieved documents.

Firstly, Figure 6 shows how our proposed method obtained better or equal results as the mean participants of the TREC2017 Precision Medicine Track, except for the fifth query, which means that none of the first five retrieved documents were relevant due to the high difficulty of the query.

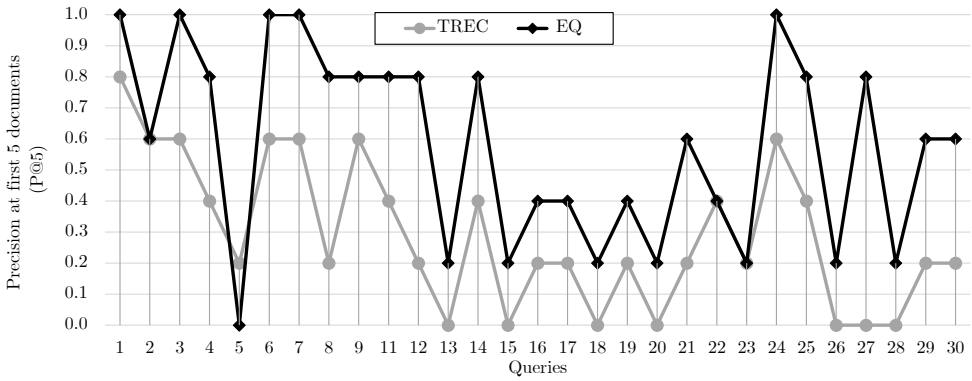

Figure 6. Comparison of the mean precision of the participants of the TREC2017 Precision Medicine Track (TREC) and the proposed method (EQ) in the top five retrieved documents.

Secondly, it can be seen in Figure 7 how the proposed method obtained the same results as the other teams at query number 5. Unlike the precision in five documents, the proposed method always offers better or equal results at P@10.

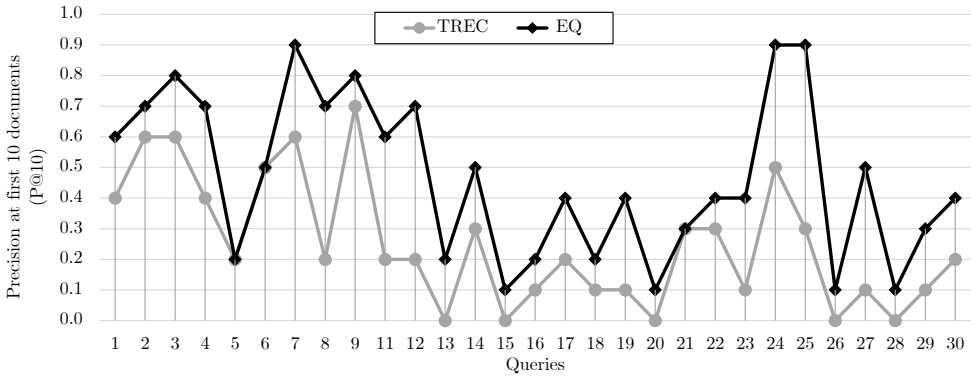

Figure 7. Comparison of the mean precision of the participants of the TREC2017 Precision Medicine Track (TREC) and that of the proposed method (EQ) in the top ten retrieved documents.

Lastly, Figure 8 proves how the results keep getting better than the mean results of the other teams in most cases, only being unable to reach it at query number 9. This query only has two definitely relevant documents and 60 partially relevant ones, making the retrieval process deeply difficult as the number of first-analyzed documents is increased. Even in this case, the obtained precision (0.60) is very close to the mean precision of the teams (0.667).

Furthermore, of the top overall systems in Table 6 [27], the proposed method surpassed the best team run [30] at P@5 (0.5448), P@10 (0.4448), and P@15 (0.3885), where we attained 0.579, 0.469, and 0.393, respectively. If we analyze the standard deviation of the best team results excluding the best team and duplicates, the values we get are 0.0175 (P@5), 0.0178 (P@10), and 0.0195 (P@15), while the proposed method improves them to 0.034 (P@5), 0.024 (P@10), and 0.005 (P@15). This means that our method obtains a significant improvement when analyzing the first five and ten documents, only managing to match the best team results at P@15. To better visualize this analysis, Figure 9 shows a comparison of the mean improvement of the teams and their respective previous team to the improvement of the proposed method and the best team results.

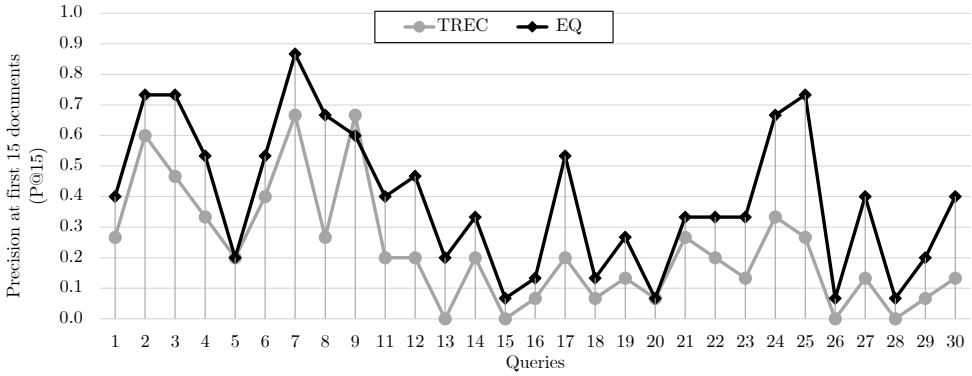

Figure 8. Comparison of the mean precision of the participants of the TREC2017 Precision Medicine Track (TREC) and the proposed method (EQ) in the top 15 retrieved documents.

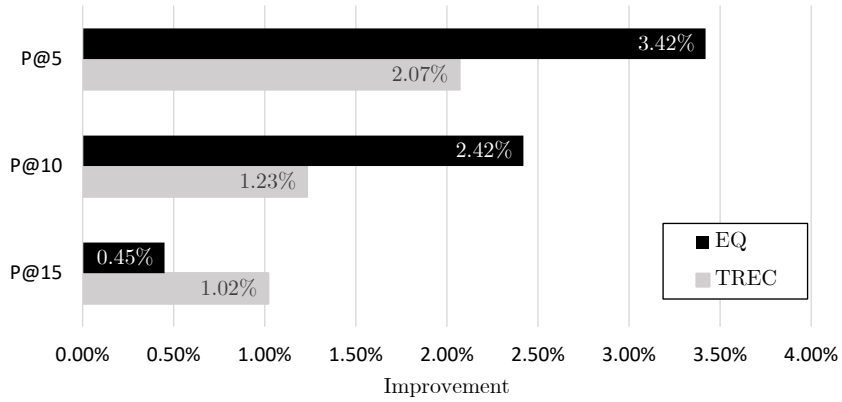

Figure 9. Mean improvement of the best teams at the TREC2017 Precision Medicine Track (TREC) over their respective previous team as compared to the improvement of the proposed method (EQ) over the best team results.

4. Conclusions

In this study, a new unsupervised query expansion technique using Multinomial Naive Bayes was presented. An expanded query was obtained by the combination of terms found in the original query and the new terms retrieved by the Multinomial Naive Bayes method. The extraction of the vocabulary from the documents retrieved by the combined query, and the selection of terms that were likely to belong to a category both proved to be effective in recovering more relevant documents.

More specifically, the application of this Pseudo-Feedback technique proved to be satisfactory considering the MAP and Precision results. The first 5, 10, 15, and 100 documents retrieved were considered in the evaluation process. Even when the first 100 documents

were taken into account, the results improved, which shows that it is possible to increase their quantity and continue to improve the retrieval process quality.

The proposed query expansion technique allows for the improvement of a lightly defined query made by the user in order to obtain better results. This will help users to find relevant documents that fulfill their needs, and easily filter documents found in large and specialized collections of documents, such as the medical corpora, where technical lexicon vocabulary make it difficult to find relevant content through a short query composed of keywords.

Inspired by the success of this new method, more techniques could be researched. We plan to review some topic modeling techniques that could offer more terms inspired in the topic covered by the first retrieved documents, which would allow for a more complex and wider query expansion. In addition, new ways of combining the terms found in the original query and the expansion terms are being studied.

Author Contributions: Conceptualization, S.S.; methodology, S.S.; software, S.S.; validation, L.B., E.L.I. and A.S.V.; formal analysis, S.S.; investigation, S.S. and A.S.V.; resources, S.S.; data curation, S.S. and A.S.V.; writing—original draft preparation, S.S. and P.C.; writing—review and editing, S.S., P.C. and L.B.; visualization, L.B.; supervision, E.L.I. and L.B.; project administration, E.L.I. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Clinical Trial corpus used in this work is available on the TREC official web page http://www.trec-cds.org/2017.html.

Acknowledgments: The SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from the University of Vigo for hosting its IT infrastructure. We also appreciate the support provided by Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding of ED431C2018/55-GRC Competitive Reference Group. Pedro Celard is supported by a pre-doctoral fellowship from Xunta de Galicia (ED481A 2021/286). The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Azad, H.K.; Deepak, A. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.* **2019**, *56*, 1698–1735. [CrossRef]
2. Zhu, D.; Wu, S.; Carterette, B.; Liu, H. Using large clinical corpora for query expansion in text-based cohort identification. *J. Biomed. Inform.* **2014**, *49*, 275–281. [CrossRef] [PubMed]
3. McDonald, R.; Brokos, G.I.; Androutsopoulos, I. Deep relevance ranking using enhanced document-query interactions. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, Brussels, Belgium, 31 October–4 November 2018; pp. 1849–1860.
4. Rehman, A.; Javed, K.; Babri, H.A. Feature selection based on a normalized difference measure for text classification. *Inf. Process. Manag.* **2017**, *53*, 473–489. [CrossRef]
5. Araújo, G.; Mourão, A.; Magalhães, J. NOVASearch at Precision Medicine 2017. In Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017) Proceedings, Gaithersburg, MD, USA, 15–17 November 2017.
6. Afuan, L.; Ashari, A.; Suyanto, Y. A Study: Query Expansion Methods in Information Retrieval. *J. Phys. Conf. Ser.* **2019**, *1367*, 012001.
7. Agosti, M.; Di Nunzio, G.M.; Marchesin, S. An analysis of query reformulation techniques for precision medicine. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 973–976.
8. Xu, B.; Lin, H.; Yang, L.; Xu, K.; Zhang, Y.; Zhang, D.; Yang, Z.; Wang, J.; Lin, Y.; Yin, F. A supervised term ranking model for diversity enhanced biomedical information retrieval. *BMC Bioinform.* **2019**, *20*, 1–11. [CrossRef] [PubMed]
9. Wang, J.; Pan, M.; He, T.; Huang, X.; Wang, X.; Tu, X. A Pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. *Inf. Process. Manag.* **2020**, *57*. [CrossRef]

10. Junior, J.R.C. *Desenvolvimento de uma Metodologia para Mineração de Textos*; Pontificia Universidad Catolica de Rio de Janeiro: Rio de janeiro, Brasil, 2007.
11. Porter, M.F. An algorithm for suffix stripping. *Program* **1980**, *14*, 130–137. [CrossRef]
12. Zipf, G.K. *Human Behaviour and the Principle of Least-Effort: An Introduction to Human Ecology*; Martino Fine Books: Eastford, CT, USA, 1949.
13. Baeza-Yates, R.A.; Ribeiro-Neto, B. *Modern Information Retrieval*; Addison-Wesley Longman: Reading, MA, USA, 1999.
14. Gauch, S.; Wang, J.; Rachakonda, S.M. A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Trans. Inf. Syst. (TOIS)* **1999**, *17*, 250–269. [CrossRef]
15. Crouch, C.J.; Yang, B. Experiments in automatic statistical thesaurus construction. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 21–24 June 1992; pp. 77–88.
16. Qiu, Y.; Frei, H.P. Concept based query expansion. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, USA, 27 June–1 July 1993; pp. 160–169.
17. Liddy, E.D.; Myaeng, S.H. DR-LINK's linguistic-conceptual approach to document detection. In Proceedings of the 1st Text Retrieval Conf. (TREC-1), Gaithersburg, MD, USA, 4–6 November 1992. [CrossRef]
18. Voorhees, E.M. *Query Expansion Using Lexical-Semantic Relations*; SIGIR '94; Springer: London, UK, 1994; pp. 61–69.
19. Miller, G.A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K.J. Introduction to WordNet: An on-line lexical database. *Int. J. Lexicogr.* **1990**, *3*, 235–244. [CrossRef]
20. Borrajo, L.; Romero, R.; Iglesias, E.L.; Marey, C.R. Improving imbalanced scientific text classification using sampling strategies and dictionaries. *J. Integr. Bioinform.* **2011**, *8*, 90–104. [CrossRef]
21. Hirschman, L.; Yeh, A.; Blaschke, C.; Valencia, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinform.* **2005**, *6*, S1. [CrossRef] [PubMed]
22. Zhou, G. Recognizing names in biomedical texts using hidden markov model and SVM plus sigmoid. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004; pp. 1–7.
23. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]
24. Strohman, T.; Metzler, D.; Turtle, H.; Croft, W.B. Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, Atlanta, GA, USA, 19–20 May 2005; Volume 2, pp. 2–6.
25. Turtle, H.; Flood, J. Query evaluation: strategies and optimizations. *Inf. Process. Manag.* **1995**, *31*, 831–850. [CrossRef]
26. Hiemstra, D.; van Leeuwen, D. Creating a Dutch information retrieval test corpus. In *Computational Linguistics in the Netherlands 2001*; Brill Rodopi: Leiden, The Netherlands, 2002; pp. 133–147.
27. Roberts, K.; Demner-Fushman, D.; Voorhees, E.M.; Hersh, W.R.; Bedrick, S.; Lazar, A.J.; Pant, S. Overview of the TREC 2017 precision medicine track. In Proceedings of the Text Retrieval Conference (TREC) NIH Public Access, Gaithersburg, MD, USA, 15–17 November 2017; Volume 26.
28. Mitra, M.; Singhal, A.; Buckley, C. Improving automatic query expansion. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; pp. 206–214.
29. Raschka, S. Naive Bayes and Text Classification I-Introduction and Theory. *arXiv* **2014**, arXiv:1410.5329.
30. Mahmood, A.A.; Li, G.; Rao, S.; McGarvey, P.B.; Wu, C.H.; Madhavan, S.; Vijay-Shanker, K. *UD_GU_BioTM at TREC 2017: Precision Medicine Track*; TREC: Gaithersburg, MD, USA, 2017.