Biomedical Text Mining Applied To Document Retrieval and Semantic Indexing

Anália Lourenço[1], Sónia Carneiro[1], Eugénio C. Ferreira[1], Rafael Carreira[1,2], Luis M. Rocha[3], Daniel Glez-Peña[4], José R. Méndez[4], Florentino Fdez-Riverola[4], Fernando Diaz[5], Isabel Rocha[1] and Miguel Rocha[2]

[1] IBB/CEB, University of Minho, Campus Gualtar, Braga, Portugal
{analia, soniacarneiro, ecferreira, irocha}@deb.uminho.pt
[2] CCTC, University of Minho, Campus Gualtar, Braga, Portugal
{rafaelcc, mrocha}@di.uminho.pt
[3] School of Informatics, Indiana University, Bloomington IN, USA
rocha@indiana.edu
[4] Computer Science Dept., Univ. Vigo, Campus As Lagoas, Ourense, Spain
{dgpena,moncho.mendez,riverola}@uvigo.es
[5] Computer Science Department, University of Valladolid, Segóvia, Spain
fdiaz@infor.uva.es

Abstract. In Biomedical research, the ability to retrieve the adequate information from the ever growing literature is an extremely important asset. This work provides an enhanced and general purpose approach to the process of document retrieval that enables the filtering of PubMed query results. The system is based on semantic indexing providing, for each set of retrieved documents, a network that links documents and relevant terms obtained by the annotation of biological entities (e.g. genes or proteins). This network provides distinct user perspectives and allows navigation over documents with similar terms and is also used to assess document relevance. A network learning procedure, based on previous work from e-mail spam filtering, is proposed, receiving as input a training set of manually classified documents.

Keywords: Biomedical Document Retrieval, Document Relevance, Enhanced Instance Retrieval Network, Named Entity Recognition, Semantic Indexing Document Network.

1 Introduction

In biomedical research, the ability to cross-reference data adequately has become invaluable. Scientific publishing grows at a steady rate and research goals are becoming ever more focused and complex. The urge for automatic curation methods and tools is now greater than ever and the capacity to retrieve the correct set of documents about a particular problem is crucial. An effective biomedical document retrieval system for user-defined queries is particularly important to the expanding body of research on Biomedical Text Mining, that aims at automatically identifying

valuable information (mostly relationships among major biological entities such as genes and proteins). Furthermore, it plays a major role in researchers' daily work as well, since researchers spend much of their time searching for relevant documents to particular problems.

Currently, PubMed is the bibliographic search system with largest life science and biomedical coverage. Between PubMed and the end-users there is the need for an intermediate layer that prevents the user to be flooded with a large set of undesired documents, and thus reducing the time and effort spent in further manual and/or automatic document processing. In other words, PubMed's results should be validated, assessing the relevance of each candidate document based on some given measure. Moreover, documents should be conveniently indexed, allowing intuitive document search, and far more important, sustaining focused searches based on biomedical terminology. Thus, users will not only work over the subset of document that they are actually interested in, but also they will be able to focus further reading and analysis based on mentions to genes, proteins and other biological entities that are meaningful in a given context.

The main contribution of this work is a novel approach to the enhanced retrieval of biomedical documents based on semantic indexing. This approach differs from previous efforts in its goals: we do not focus on a particular query, since the conceptual building of the evaluators holds out regardless of the query. Furthermore, our final retrieval goal relates more directly to the needs of researchers using PubMed, i.e., we aim at delivering a tool that can assist end-users in their daily activities. As such, we addressed the filtering of PubMed's results, but we also provide for an indexing network that displays the documents according to user search perspectives, associating documents with similar contents.

2 Biomedical information

Biomedical information retrieval is mostly supported by bibliographic databases and open-access journals. Currently, PubMed sustains the largest life science and biomedical bibliographic database, containing over 17 million records. Although providing an invaluable service, PubMed search engine is based on user-specified queries, i.e., sets of keywords that the user considers to best describe the query. Achieving an adequate formulation of a query is not straightforward. Users may choose general terms or address broad-scope problems (e.g. a search on "leukemia"). While tracking down eventually relevant documents through such a process, many partially related and irrelevant documents will be retrieved as well.

Every document that matches the posted keywords in any of the requested search fields is considered a candidate. However, it is not trivial for the user to pose its query in such a way that the keywords do not bring attention over documents that are not connected to the subject of their interest. For example, if we are interested in searching documents related to "Escherichia coli stringent response", we can impose the co-occurrence of the words all together. In this case, we will certainly miss many relevant documents due to discourse variants (e.g. "stringent response in Escherichia coli"). If we pose a word-free query, i.e., not imposing any word co-occurrence, we will get every document that matches any of our four query words. Probably, the wisest decision would be to re-structure the query, arranging the organism name

"Escherichia coli" and the event/problem "stringent response" as two search terms. Yet, even then we may get a considerable number of partially related or irrelevant documents.

Some initiatives offer related research. In KDD 2002, one of the tasks focused on helping to automate the work of curating biomedical databases by identifying what papers need to be analysed for Drosophila gene expression information. The sub-task 2.3 of the BioCreAtIvE 2004 workshop addressed the automatic extraction and assignment of Gene Ontology (GO) annotations of human proteins, using full-text articles [1]. In turn, the 2004 TREC Genomics the same retrieval task embraced a broader variety of bioinformatics queries [2]. Other works address problems such as: the identification of protein interaction mentions using word proximity networks [3]; the ranking of gene queries for the human genome [4], the construction of content-rich biological networks [5], the association of genes with Gene Ontology codes [6] and the re-ranking of PubMed's results according to their relevance to SwissProt annotation [7]. It is interesting to notice that machine learning techniques are currently combined with Natural Language Processing (NLP) techniques in order to tackle conventional linguistic analysis as well the particular biomedical terminology.

We are also interested in improving retrieval performance. Notwithstanding, our work differs from this line of work as we aim at delivering a rich document indexing network that focusing on relevant documents provides means of navigation through the biological terms that best describe those documents. Users do not end up with a ranked list of documents, but rather a network that can be intuitively traversed.

3 Biomedical document retrieval and semantic indexing

Our workflow for document retrieval and processing encompassed three steps: retrieving documents from PubMed; pre-processing documents, namely PDF to text conversion and basic document structuring; lexicon-based Named Entity Recognition (NER) (bottom part of Figure 1). Any tool that is able to perform such tasks and outputs annotated documents can be used in this stage. The only requirements are a robust NER module (lexicon-based or trained over gold standard corpora) and the tagging of major biological classes (namely, genes, proteins, compounds and organisms). In this work, the @Note Biomedical Text Mining open-source workbench, a tool developed by the authors, is used. @Note supports PubMed search for relevant documents and document retrieval from open-access and subscribed Web-accessible journals. Entrez's eUtils grant access to PubMed and deliver query results. Each PubMed record has a set of external links that the LWP crawling module follows, trying to reach for the full-text document. The original documents in PDF format are converted into plain ASCII files. Plain text documents are tokenised and common English stopwords are filtered. NER is based on a dictionary (obtained by the merge of contents of major biological databases) and expert-specified lookup lists. A term rewriting system encompasses the set of active annotation rules, ranging from simple substitution rules to conditional and evaluated rules. Rules target up to seven-word terms and ignore too short words (less than 3 characters long). Furthermore, @Note sustains a user-friendly environment for the manual curation of document relevance.

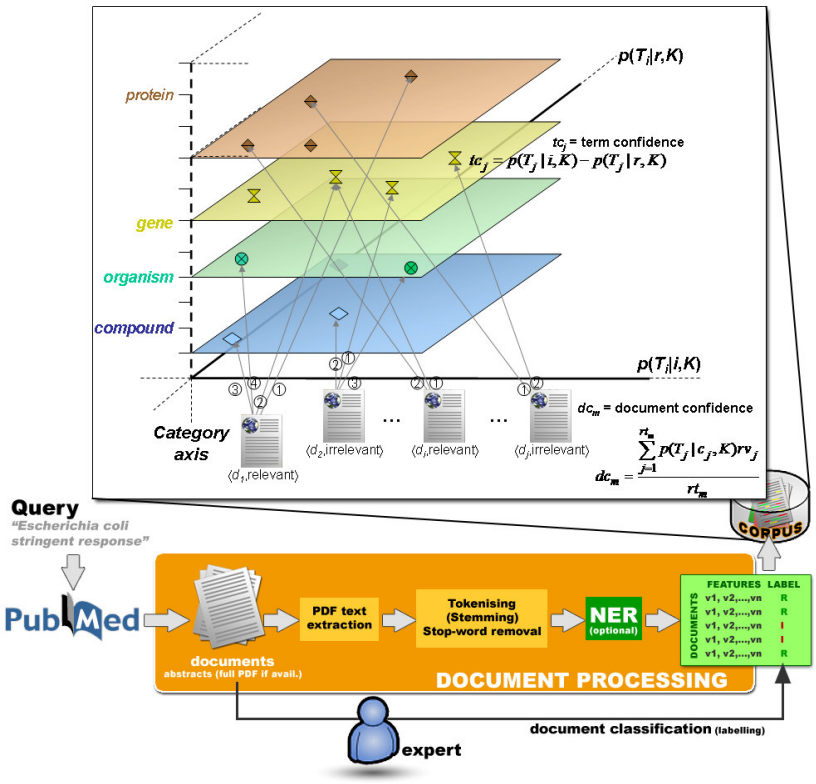

Fig. 1. Document retrieval and semantic indexing workflow.

Taking as input the pre-processed and annotated set of documents, we are interested in selecting the most relevant terms of major biological classes (genes, proteins, compounds and organisms) for each document. Without any further information, the only way of doing this is to base it on the frequency of each word in the document. But, if we have available a collection of classified documents (a corpus), we can use information about the underlying distribution of the corpus in relation to the target concept (relevant or irrelevant) to assess the relevance of each term inside a specific document. In this context, the relevance measure of a term should be able to identify highly predictive terms. The relevance of each term of the document is defined as:

$$r(T_j,d)=\left\{\frac{\left|p(i\mid K)\,p(T_j\mid i,K)-p(r\mid K)\,p(T_j\mid r,K)\right|}{p(T_j\mid K)}\right\}p(T_j\mid d) \qquad (1)$$

The relevance measure $r(T_j,d)$ tries to conjugate the local and global relevance of the term T_j. The first factor in $r(T_j,d)$ depends on the whole corpus K and expresses the utility of term T_j in order to discriminate among irrelevant or relevant documents and therefore it evaluates the global relevance of T_j. The second factor in $r(T_j,d)$ only depends on the specific document which is being processed and, hence, it can be

viewed as a measure of the local relevance of T_j. As a consequence of this definition, the relevance of a term T_j which appears in two different documents only depends on the local relevance (since the first factor of Exp. (1) will be the same). Moreover, the relative relevance of two terms T_j and T_k, which appear in a specific document d, not only depends on the local information, but also depends on the global information which will be probably different for both terms. This is particularly important because we are interested in ordering (by relative relevance) different terms in a specific document in order to select the most relevant ones. Finally, this formulation can be used to select the most relevant terms in two ways: (*i*) a fixed number of terms ordered with respect to $r(T_j, d)$ or (*ii*) a variable number of terms depending on a fixed percentage of the whole sum of individual relevance values (if the terms of a document d are ordered descending by $|r(T_j, d)|$ and R is the sum of $|r(T_j, d)|$ over all the terms T_j belonging to d, then given a percentage α, the first k_α terms, whose partial sum of relevance values exceeds the quantity of αR, will be selected as the most relevant terms).

Based on the previous formulation for selecting relevant terms of each document in a corpus K, we present here our EIRN (*Enhanced Instance Retrieval Network*) model for efficient and flexible document indexing and retrieval. Our EIRN memory structure is borrowed from the previous successful SPAMHUNTING system [8], an instance-based reasoning model that outperforms classical machine learning techniques as well as other successful lazy learner approaches in the domain of anti-spam filtering.

Based on the Case Retrieval Networks (CRN) indexing properties [9], our model defines two measurements: (*i*) *Term Confidence* and (*ii*) *Document Confidence* for maintaining as much information as possible about existing data (terms and documents). Figure 1 depicts an example of our EIRN model to document retrieval. The EIRN network used in this work is characterized by a two-dimensional space, where the terms (cells) are connected and organized according to the probability of representing irrelevant and relevant documents. Each cell in the network is associated with a *term confidence* (*tc*) which represents a measure of how much we can trust it to classify a given document. The value of *tc* for a given term T_j is given by Eq. (2).

$$tc_j = p(T_j \mid i, K) - p(T_j \mid r, K) \tag{2}$$

where $p(T_j \mid i, K)$ and $p(T_j \mid r, K)$ stand for the probability of the term T_j belonging to irrelevant and relevant documents, respectively.

The basic learning process in the EIRN network consists in topology modification and term confidence adaptations. Based on a corpus K of training documents, learning in an EIRN network is carried out by presenting all training documents to the network in a sequential fashion. For each training instance presentation, the network performs a so-called *learning cycle*, which may result in term confidence adaptation and topology modification. Figure 1 clarifies this situation showing those cells with closest values for $p(T_j \mid r, K)$ and $p(T_j \mid i, K)$ parameters located in nearby points.

In the first step of each learning cycle, the relevant terms (*rt*) of the actual input document d_m, are linked with the terms present in the network, adding new terms to the model if necessary. Each new connection is weighted up with a relevant value (rv_j) which represents the importance of this term to the actual document. The value

of rv_j depends on the relevant terms (rt_m) of the input document d_m and the current term T_j. rv_j is calculated using

$$rv_j = \frac{w_k}{2^{j-1}} \quad with \quad w_k = \frac{2^{rt_m - 1}}{2^{rt_m} - 1} \tag{3}$$

The second step consists of the adaptation of the term confidence affected in the previous step and the calculation of the actual *document confidence* (dc_m). The parameter *dc* represents a measure of document coherence by means of its relevant terms and aids in the identification of rare document contents. The value of *dc* for a given pair $\langle d_m, c_j \rangle$ is calculated by:

$$dc_m = \frac{\sum_{j=1}^{rt_m} p(T_j \mid c_j, K) rv_j}{rt_m} \tag{4}$$

where c_j represents the actual class of the document d_m, rt_m stands for the number of relevant terms for d_m, $p(T_j \mid c_j, K)$ represents the probability of the term T_j belonging to a document with the same class as document d_m and rv_j is calculated using Eq. (3).

Every time a given document needs to be classified, the EIRN network obtains a set *M*' composed of the documents most similar to the target document *d*'. In this sense, we can conceive the EIRN memory structure as a dynamic *k-nearest neighbor* mechanism able to retrieve a different number of neighbors depending on the terms selected from the unclassified document, *d*'. This is done by selecting the relevant terms of the new document as described previously and projecting them into the network term space (see Figure 1). To perform this selection stage, the system encompasses two sequential steps: (*i*) calculating the distance between *d*' and the set of documents that shares the greatest number of common terms (*cf*') and (*ii*) selecting those documents with a similarity value greater than the mean average value.

In order to calculate the similarity between two documents, given a set of shared relevant terms, we use a weighted distance metric that takes into account the relevance of each common term. The underlying idea is to weight those terms that are more relevant to the target document *d*', using the position occupied by each of them in the arrows coming from the target document to the memory structure in Figure 1. The value of the distance between the target document *d*' and a given document d_m is:

$$D(d', d_m) = \sum_{j=1}^{cf'} d(d'_j, d_{mj}) rv_j \tag{5}$$

where *cf*' is the number of common terms between *M*' and *d*', rv_j represents the importance of each term to the target document *d*' and measures the distance between the position assigned to the common term *j* in the two documents, calculated as the difference between the situation of this term in the arrows coming from the target document *d*' and a given document d_m to the memory structure in Figure 1.

Given the distance between two documents, the similarity is obtained by the following expression, where the document coherence is used to consider those texts which are most consistent with the corpus:

$$S(d', d_m) = \frac{1}{D(d', d_m)} dc_m \tag{6}$$

Every time the aforementioned document retrieval stage is executed by selecting those documents with higher values for the similarity with the target document *d*', the system assigns a class label to the new document *d*' based on a proportional weighting voting algorithm. Each document in *M*' returns one vote and by means of recounting the existing votes, a final classification is provided by the system.

4 Experiments

A case study concerning the behavior of the bacterium *Escherichia coli* under stress conditions is used to validate our EIRN model. The query *Escherichia coli amino acid starvation* was posed to PubMed, aiming at documents related to amino acid starvation, i.e., the condition that initiates the overall response to stress. Amino acid starvation triggers stringent response, while other conditions of starvation (e.g. nitrogen starvation) initiate other stress responses. Thus, any paper that addresses another starvation condition, but refers to amino acid starvation might be included in the results as well. Out of 258 documents retrieved from PubMed, an expert curator labeled 76% as irrelevant and 24% as relevant.

For the experiments, we have used a 10-fold cross-stratified validation scheme for improving the quality of the achieved results [10]. With respect to the representation of each document, our EIRN network was created using all the terms, capturing the maximum quantity of information (α=100%).

Figure 2 shows the percentage of correct classifications (%TP+TN), percentage of false positives (%FP) and percentage of false negatives (%FN) belonging to the two analyzed queries. The proposed model drastically reduces the number of FN errors (relevant documents not detected) in both queries when the NER process is applied. Moreover, the system is able to correctly classify a higher number of documents.

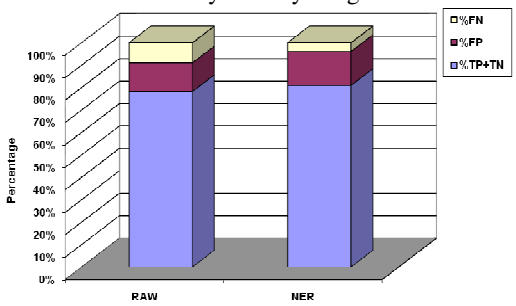

Fig. 2. Percentage of correct classifications, false positive and false negative errors.

Table 1 shows basic performance measures. The first column shows the accuracy of the classifier. The use of NER slightly improves the accuracy and the recall, thus its use increases the proportion of well classified documents within the relevant documents. On the other hand, the proportion of well classified documents within the irrelevant documents (measured by the specificity) is approximately the same. Regarding the predictive behaviour of the classifier, the use of NER barely changes the value of the precision of the classifier, but it improves significantly the negative predictive value.

Table 1. Different performance results of the classifier: accuracy, recall (or sensitivity), specificity, precision (or positive predictive value) and negative predictive value with 10-fold cross-validation

	Accuracy	**Recall**	**Specificity**	**PPV**	**NPV**
RAW	0.78	0.63	0.83	0.54	0.88
NER	0.81	0.84	0.80	0.57	0.94

In order to show the effect of R:I ratio on the predictive values, Figures 3a and 3b show the extrapolated values of precision and the estimated values of the negative prediction value, when the probability of relevant/irrelevant documents varies in the available corpus.

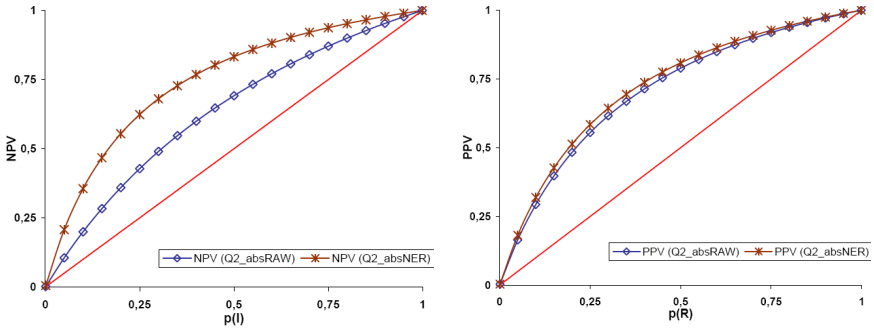

Fig. 3. Model behaviour analysis for different scenarios of R:I query results. (left) positive predictive value (precision) of the classifier (right) negative predictive value of the classifier.

Consequently, and in order to avoid the effect of the R:I ratio and give a more robust performance measure of the classifier, Table 3 shows the *f-score*$_\beta$ (for three different weights of β), the kappa coefficient and the diagnostic odds ratio. The kappa and DOR measures show that the use of NER improves the performance of the classifier, since kappa coefficient and DOR grows significantly.

Table 2. The f-score values for different balanced weights, kappa coefficient and diagnostic odds ratio with 10-fold cross-validation

	F-score			**Kappa**	**DOR**
	β=0.5	**β=1.0**	**β=2.0**		
RAW	0.56	0.58	0.61	0.44	8.38
NER	0.61	0.68	0.77	0.55	20.93

To measure the contribution of each biological class in our EIRN structure, Table 3 shows the individual value of the Cohen's Kappa coefficient for classification (using abstract with NER) as well as the total amount of terms stored in our EIRN model for each biological class. As we can see from Table 3, "compounds" is the biological class with highest impact on the model (better Kappa coefficient). Our model is able to correctly classify (using abstracts with NER) and efficiently index relevant documents with a percentage of terms below the 50% of the total amount.

Table 3. Contribution of biological classes in the EIRN indexing structure

EIRN terms	Kappa	Query
20848	0.45	(C)*ompounds*
15926	0.41	(G)*enes*
14290	0.38	(P)*roteins*
13321	0.02	(O)*rganisms*
36774	0.49	(C+G)
51064	0.51	(C+G+P)
64385	0.55	(C+G+P+O)

5 Conclusions

This work proposes a novel approach to the retrieval of biomedical documents based on Text Mining oriented semantic indexing. The approach does not focus on a particular query, since the conceptual building of the evaluators holds out regardless of the query. Furthermore, our final retrieval goal relates more directly to the needs of researchers using PubMed, i.e. we aim at delivering a tool that can assist end-users in their daily activities. We address the filtering of PubMed's results, but we also provide for an indexing network that displays the documents according to user search perspectives, associating documents with similar contents and allowing term-specific views. A network learning procedure, based on previous work on e-mail spam filtering, is applied, receiving as input a training set of manually classified documents. The resulting network provides distinct user perspectives and allows navigation over documents with similar terms and can be used to assess document relevance.

References

1. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005;6 Suppl 1:S1.
2. Hersh W, Bhupatiraju RT, Ross L, Johnson P, Cohen AM, Kraemer DF. TREC 2004 Genomics Track Overview. In Proc. 13th Text Retrieval Conference (TREC), p. 13-31.
3. Abi-Haidar A, Kaur J, Maguitman A, Radivojac P, Retchsteiner A, Verspoor K, et al. Uncovering Protein-Protein Interactions in the Bibliome. *Genome Biology* 2008;247-55.
4. Sehgal AK, Srinivasan P. Retrieval with gene queries. *BMC Bioinformatics* 2006 Apr 21;7.
5. Wang P, Morgan AA, Zhang Q, Sette A, Peters B. Automating document classification for the Immune Epitope Database. *BMC Bioinformatics* 2007 Jul 26;8.
6. Raychaudhuri S, Chang JT, Sutphin PD, Altman RB. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research* 2002 Jan;12(1):203-14.
7. Mostafa J, Lam W. Automatic classification using supervised learning in a medical document filtering application. Information Processing Management 2000;36(3):415-44.
8. Méndez JR, Glez-Peña D, Fdez-Riverola F, Díaz F, Corchado JM. Managing irrelevant knowledge in CBR models for unsolicited e-mail classification. *Expert Systems with Applications* 2008.
9. Lenz M, Auriol E, Manago M. Diagnosis and Decision Support. *Lecture Notes in Artificial Intelligence* 1998;1400:51-90.
10. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proc. 14th International Joint Conference on Artificial Intelligence, p. 1137-43.