Classifying Biological Articles using Web Resources

Francisco M. Couto
fcouto@di.fc.ul.pt

Bruno Martins
bmartins@xldb.di.fc.ul.pt

Mário J. Silva
mjs@di.fc.ul.pt

XLDB Research Group, Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal

ABSTRACT

Text classification systems on biomedical literature aim to select relevant articles to a specific issue from large corpora. Most systems with an acceptable accuracy are based on domain knowledge, which is very expensive and does not provide a general solution. This paper presents a novel approach for text classification on biomedical literature, involving the use of information extracted from related web resources. We validated this approach by implementing the proposed method and testing it on the *KDD2002 Cup challenge: bio-text task*. Results show that our approach can effectively improve efficiency on text classification systems for biomedical literature.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Bioinformatics (genome or protein)databases, Feature extraction or construction, Text mining, Web mining*

Keywords

biomedical text classification

1. INTRODUCTION

The classification of biological literature is an important recent research topic, motivated by the large number of biological articles that curators have to read in order to update biological databases, or simply to be aware of progress in a specific area. Text classification applied to biological literature can minimize this effort by automatically selecting only the relevant articles to a given task [3].

Text classification systems are primarily designed to assign categories to documents, in order to support information retrieval, or to provide an aid to human indexers in the assignment task. In the simplest form, binary classification, the system decides the relevant and irrelevant documents (or passages) from large corpora [20]. Most approaches to text classification are based on statistical natural language processing [13]. They apply quantitative methods for automated language processing, using probabilistic modeling, information theory, and sometimes linear algebra. Statistical text classification systems need a training set of documents in order to build a model later used to classify other documents. This training set consists in a representation of each document and its expected classification. After building the model, and given the representation of a new document, the system can then predict its class. Most of the times, when we want to evaluate a model, we create a test set. This set also contains the expected classification for each document, which will later be compared to its predicted classification. The most common form of representation for documents is the bag-of-words. In this approach, features are the set of all words mentioned in the documents, and each document is represented by the number of occurrences of each one of these features in the text.

The classification process requires having appropriate features to describe the instances to be classified (e.g., the meaningful terms occurring in the documents). Since appropriate features are not always available, a usual approach is constructing new ones (e.g., by combining old features in some interesting way [17]). To deal with this problem, most information extraction methods applied to biomedical literature use domain specific knowledge to improve their efficiency in a given domain. The domain knowledge is usually applied to build grammatical rules or training sets that are valid in a specific problem [10]. However, these are very expensive and limited solutions, since they are not normally applicable in other domains.

An alternative approach for generating new features is to use external information sources, such as databases found on the web. For example, Basu et al. formalize movie recommendation as a classification problem, and show that classification performance can be improved using features extracted from the web [1]. This approach cannot always be used (e.g., one is not likely to find large amounts of additional information about very specific problems on the web), but when it is applicable, it is often useful. Cohen proposed a method that produces new features from a collection of web pages [5]. The method reduced the error rate of classifiers in a wide variety of situations. In the field of molecular biology, structured databases that collect and distribute biological information on the Internet are nowadays common. Examples are the GenBank or the SwissProt databases, dealing with biological sequences and describing properties of common biological entities such as genes and proteins. Automatic tools that integrate these data sources, such as ProFAL [7, 6], are a viable approach to correct and complete our knowledge about biological entities [8]. Text classification has thus a perfect application scenario in this problem.

This paper introduces an approach for biological text classification, which involves integrating extracted information from biological web resources into the text classification process. We present a method that, given a collection of articles, extracts related information from biological databases to produce a set of new features,

i.e., a richer representation of each document. If this information is valuable, classification will achieve a higher accuracy than simply using only the text from the articles. To validate the proposed method, we compared its performance with the following types of text classification approaches:

- A standard approach, using the Naïve Bayes statistical classification method with bag-of-words document representation [13].
- State-of-the-art approaches, which use domain based methods.

The Naïve Bayes is a simple method that can achieve relatively good performance on classification tasks. It is based on the assumption that each feature value assignment is probabilistically independent of all other feature value assignments.

In both cases, we used the experimental data provided by the *KDD2002 Cup challenge: bio-text task* [21] to evaluate our method.

The rest of this paper is structured as follows. Section 2 details the proposed method. In section 3 we describe the experimental data used to evaluate our method. Section 4 presents the results achieved by the proposed method. Section 5 analyzes and discusses the results. Finally, in section 6, we express our main conclusions.

2. METHOD

Our classification method relies on biological results stored in public databases available on the web. It is motivated by the observation that most authors of recently published biomedical articles also submit their results to these public databases. Therefore, the databases usually have their data associated with bibliographic information, which provides a powerful source for document classification. Since this information is stored in a structured form, it can be easily used in an automated system. We named our method WeBTC (Web Biological Text Classification) and we can describe it as follows:

Input:

- A collection of articles with its content and its meta-data (e.g. title, authors, accession number in a bibliographic database).
- A biological database where information about the articles can be found.

Output:

- A statistical representation for the articles, where each article is represented by the number of occurrences of each term found in the database.

Procedure:

1. For each article, we identify all the associated database accession numbers. An accession number is a unique identifier for a database entry. This information can be extracted by three different ways:
   (a) Directly from the article content. Most authors present accession numbers in their articles, referencing the database where their results were submitted. It is not hard to find an accession number in the text, since they have a common format depending on the database (e.g. two letters followed by 6 digits). Moreover, sentences with an accession number usually also reference the database common name.
   (b) When the authors of a published article submit their results to a database, they often submit also the article identification. In this case, we only have to identify the database entries that cite the article, which is only possible if the database stores and makes available the bibliographic information.
   (c) When a database entry has no bibliographic information but mentions its source indirectly (e.g. the authors, the date, the laboratory, the technique) we can match this data against the article's meta-data to infer that the article represents the information source of the database entry.
2. We retrieve the content of the database entries, and identify the number of distinct terms mentioned on them.
3. For each article, we compute the occurrences of each term in its associated database entries.

EXAMPLE 1. *The article available in PubMed with the identifier* 12803610*, contains the following sentence:*

"The sequence of the nramp cDNA was filed at the EMBL/GenBank/DDBJ Databases under the accession number AJ514946."

For this article, WeBTC's step 1a extracts the accession number AJ514946 *whose entry is available in the database* GenBank*. Besides other terms, this* GenBank *entry contains the term "*Hordeum vulgare subsp. vulgare"*, which is the name of the organism. Step 2 identifies this term, and step 3 counts at least one occurrence of the term. Therefore, the WeBTC output will contain a representation of the article where the feature representing the term "*Hordeum vulgare subsp. vulgare" *has at least one occurrence.*

3. EXPERIMENTAL DATA

We experimentally evaluated WeBTC for classifying biomedical articles on the *KDD2002 challenge cup competition: bio-text task*. The task consisted on identifying which biomedical articles contained relevant experimental results, and which were the gene products (transcripts and proteins) involved. This represents one stage of the curation process done in FlyBase [19]. FlyBase is a comprehensive database for information on the genetics and molecular biology of Drosophila (fruit fly). The curators take a set of articles and extract new relevant information reported on them. By new relevant information, we mean experimental results applicable to wild-type (non-mutated) fruit flies, which are not just merely citations of other articles.

The task goal was to implement a system with the following behavior:

Input:

- A collection of articles on Drosophila genetics or molecular biology. For each article, the full content was provided as a raw text file.
- An XML template for each article containing its identifiers and the list of the genes mentioned in it. The gene names follow a standardized nomenclature, and a synonym list for each gene was provided.
- Other collections of data from biological databases publicly available on the web could also be used, to better mimic real conditions.

Output:

- For each article, a Boolean decision on whether or not there are relevant experimental results reported on it.
- For each article assumed to have relevant experimental results, the genes involved and the gene-product type (transcript, protein, or both).
- A ranked list of articles, sorted by the assurance degree of having relevant experimental results. The articles more likely to contain experimental results should be ranked higher than the articles with no experimental results.

In the competition, each output item was considered a sub-task that was evaluated separately. The collection of articles was divided in two sets: the training set with 862 articles and the test set, with 213 articles. The expected output for each article in the training set was provided. Only 283 articles of the training set reported relevant experimental results. The output of these articles was extended with result evidences.

3.1 Implementation

Our implementation of WeBTC in this specific case-study started with the retrieval of the meta-data of each article through its PubMed identifier (an interface to the public bibliographic database MEDLINE [15]). This identifier was provided for each article. We selected the following external biological databases to use with WeBTC:

- MeSH (Medical Subject Headings), a collection of keywords for classifying articles [16].
- GenBank (GenPept), a repository of gene (protein) structure data [2].

There was no need to execute the first step of WeBTC to associate each article with the MeSH terms, since PubMed already manually classifies each article with a set of MeSH terms. We retrieved the GenBank and GenPept accession numbers in the articles' text and through the citations. However, in our evaluation we did not implement the third approach, involving the use of the articles' meta-data to retrieve accession numbers. Then, we executed WeBTC three times: with MeSH, GenBank, and GenPept. The result was three different statistical representations of each article. We combined their features to integrate these representations into a single one, which we named the WeBTC representation.

For each article, we created its bag-of-words representation from its text using Bow, a toolkit for statistical language modeling, text retrieval, classification and clustering [14]. We used the stemming algorithm available in Bow to increase the features quality both in WeBTC and bag-of-words representations. Given the statistical representations of each article, Bow built the models using the Naïve Bayes statistical classification method.

4. RESULTS

4.1 WeBTC vs. Standard Approach

We built a model from the WeBTC representations and another model from the bag-of-words representations. We also implemented a combined model that only considers an article relevant if both models agree in doing so.

Table 1 presents the results obtained by the three models that predicted the classification of the 213 articles in the test set. TruePositives are the number of articles that a model correctly predicted to be relevant. TrueNegatives are the number of articles that a model correctly predicted to be irrelevant. FalsePositives are the number of articles that a model incorrectly predicted to be relevant. FalseNegatives are the number of articles that a model incorrectly predicted to be irrelevant.

	Bag-of-words	WeBTC	Combined
TruePositives	41	19	15
FalsePositives	19	2	0
TrueNegatives	103	120	122
FalseNegatives	50	72	76

Table 1: Results of the three models.

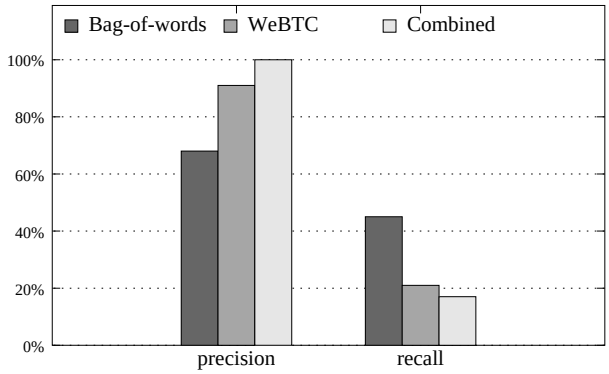

Figure 1: Precision and recall of the three models

Figure 1 compares the precision and recall obtained by the three models. Precision and recall are standard measures of text mining systems that can be directly calculated through these values. Precision is the number of TruePositives divided by the number of all Positives. Recall is the number of TruePositives divided by the number of TruePositives and FalseNegatives. Results show that WeBTC achieved s significantly higher precision. The combined model enabled us to achieve 100% precision. However, the use of WeBTC also implied a reduction on recall.

4.2 WeBTC vs. State-of-the-art Approaches

Since the combined model achieved a better performance, we applied it in our submission to the KDD Cup. The results of 32 state-of-the-art systems were provided by the KDD Cup organization committee, which applied a scoring method to evaluate each of the sub-tasks. They scored the ranked list by the ROC curve [4], the article decision and the gene-product decision by the standard F-measure [13]. The overall score was obtained by the sum of these three scores, normalized to a 0% to 100% range representing the efficiency of the systems.

Figure 2 shows the results for the three sub-tasks and the overall score. The *Best* values represent the highest score, which in this case was always obtained by the same team. The *1Q* values represent the score limit of the first quartile [12], i.e. in this contest it represents the ninth highest score. The *Median* values represent the arithmetic average of all scores. The *Low* values represent the lowest score obtained. The *WeBTC* values represent our submission scores. Our overall score was in the first quartile in two sub-tasks. The exception was in the article decision sub-task, where our score was even lower than the median. In this sub-task, we achieved a precision of 81% but a recall of only 38%.

5. DISCUSSION

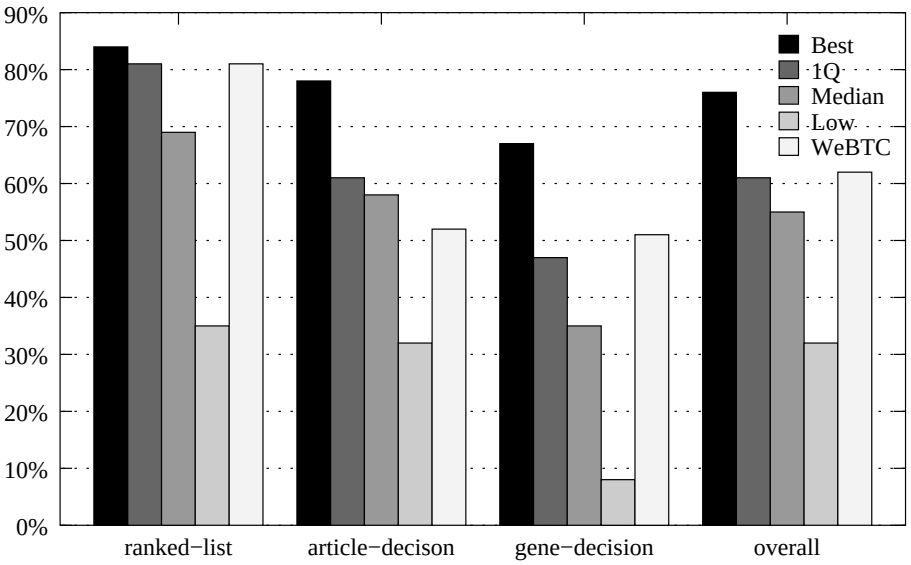

Figure 2: KDD Cup Scoring Results

The main problem of our approach was the low recall. This happened because we were not able to retrieve information for all articles due to the small number of external biological sources used. If we had more time to give our predictions, then we certainly would retrieve more information since the databases used would be more complete and we could also cover other resources. If we presented results obtained with information retrieved after the KDD Cup deadline, we certainly would have a larger recall but that would not show the effectiveness of our approach, because database curators in real situations also have a deadline to classify the articles. Thus, to present results of WeBTC with a higher recall maintaining its precision we have to cover in due time a broader range of resources. On the other hand, for the articles with information available, WeBTC provided a very accurate prediction, reaching 100% precision. The high levels of precision are very useful for database curators, since they do not have to manually verify predictions of relevant articles.

The ClearForest and Celera team developed the winning system of the KDD Cup task [18]. Their system was implemented through a rule-based general Information Extraction language. The rules were built specifically for the task with basis on domain knowledge, and were essentially sequences of terms to use in pattern matching. A team from Singapore obtained an honorable mention by developing a system based on feature extracting with a Naïve Bayes Classifier [11]. However, their feature extraction was based on a set of keywords manually extracted from the training texts and on manual selection of positive examples. Another honorable mention was given to a team from UK [9]. Their system was also based on feature selection and on statistical classification methods, but feature selection was also based on relevant keywords supplied by local domain experts.

All the approaches described above use domain knowledge as a crucial component of their systems. The main conclusion retained from the KDD Cup was that statistical text classification systems reasoning without considering domain knowledge achieved poor results. Our approach attempts to obtain domain-specific knowledge through information automatically extracted from external biological sources available on the web.

6. CONCLUSIONS

This paper introduced a novel approach for text classification, which involves the integration of extracted information from biological web resources with common statistical text classification methods.

In our case-study, WeBTC was able to significantly increase the precision (reaching 100%) of a standard classification method. Its low levels of recall are due to the small number of articles for which information in the external databases was found. If more information had been retrieved, WeBTC would certainly achieved higher levels of recall while maintaining its remarkable levels of precision.

The performance of WeBTC was also evaluated in the *KDD2002 Cup challenge: bio-text task* versus state-of-the-art systems. The evaluation indicated that WeBTC provided an effective alternative to enhance the performance of standard classification methods.

WeBTC deserves further study. We hope to match the performance of state-of-the-art methods, which are based on domain knowledge introduced and maintained manually.

7. ADDITIONAL AUTHORS

Additional authors: Pedro Coutinho (Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique), email: pedro@afmb.cnrs-mrs.fr.

8. REFERENCES

[1] C. Basu, H. Hirsh, and W. W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *AAAI/IAAI*, pages 714–720, 1998.

[2] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, B. Rapp, and D. Wheeler. GenBank. *Nucleic Acids Research*, 30:17–20, 2002.

[3] C. Blaschke, R. Hoffmann, J. Oliveros, and A. Valencia. Extracting information automatically from biological literature. *Comparative and Functional Genomics*, 2:310–313, 2001.

[4] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[5] W. W. Cohen. Automatically extracting features for concept learning from the web. In *Proc. 17th International Conf. on Machine Learning*, pages 159–166. Morgan Kaufmann, San Francisco, CA, 2000.

[6] F. Couto, M. Silva, and P. Coutinho. Improving information extraction through biological correlation. In *Data Mining and Text Mining for Bioinformatics Workshop co-located with 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Dubrovnik-Cavtat, Croatia, September 2003.

[7] F. Couto, M. Silva, and P. Coutinho. ProFAL: Protein functional annotation through literature. In *VIII Conference on Software Engineering and Databases (JISBD)*, Alicante, Spain, November 2003.

[8] M. Gerstein. Integrative database analysis in structural genomics. *Nature Structural Biology*, Structural genomics supplement:960–963, November 2000.

[9] M. Ghanem, Y. Guo, H. Lodhi, and Y. Zhang. Automatic scientific text classification using local patterns: KDD CUP 2002 (task 1). *SIGKDD Explorations*, 4:95–96, 2002.

[10] L. Hirschman, J. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561, 2002.

[11] S. Keerthi, C. Ong, K. Siah, and et al. A machine learning approach for the curation of biomedical literature - KDD Cup 2002 (task 1). *SIGKDD Explorations*, 4:93–94, 2002.

[12] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics*, chapter Quartiles, pages 35–37. Princeton, NJ: Van Nostrand, 1962.

[13] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[14] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[15] MEDLINE. PubMed database at the National Library of Medicine. `www.ncbi.nih.gov/PubMed`.

[16] MeSH: Medical Subject Headings. `www.nlm.nih.gov/mesh/meshhome.html`.

[17] G. Pagallo and D. Hassler. Boolean feature discovery in empirical learning. *Machine Learning*, 1990.

[18] Y. Regev, M. Finkelstein-Landau, R. Feldman, and et al. Rule-based extraction of experimental evidence in the biomedical domain - the KDD Cup (task 1). *SIGKDD Explorations*, 4:90–92, 2002.

[19] G. Rubin. Around the genomes: The drosophila genome project. *Genome Research*, 6:71–79, 1996.

[20] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[21] A. Yeh, L. Hirschman, and A. Morgan. Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles. *SIGKDD Explorations*, 4:87–89, 2002.