Classifying Biomedical Articles using Web Resources : application to KDD Cup 02

Francisco M. Couto
Bruno Martins
Mário J. Silva
Pedro Coutinho

DI–FCUL TR–03–24

July 2003

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

Classifying Biomedical Articles using Web Resources : application to KDD Cup 02

Francisco M. Couto, Bruno Martins and Mário J. Silva
Departamento de Informtica Faculdade de Cincias Universidade de Lisboa
E-mail: {fjmc,bmartins,mjs}@xldb.di.fc.ul.pt

Pedro Coutinho
UMR 6098, Architecture et Fonction des Macromolcules Biologiques
Centre National de la Recherche Scientifique
E-mail: pedro@afmb.cnrs-mrs.fr

July 2003

Abstract

This paper presents a novel approach for text classification on biomedical literature, involving the use of information extracted from related web resources. Our method creates a representation of an article based on information extracted from public online databases, that is afterwards used by traditional statistical text classification algorithms. 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 of searching for additional data on online databases can effectively improve efficiency on text classification systems for biomedical literature.

1 Introduction

As for most fields of scientific research, relevant facts discovered in molecular biology have mainly been published in scientific journals [4]. Extracting knowledge from this large amount of unstructured information is an arduous task, even for human experts. An improvement was the creation and maintenance of structured databases that collect and distribute biological information. Examples are the GenBank or SwissProt databases that deal with biological sequences and describe properties of common biological entities, such as genes and proteins.

In the past few decades, the explosion of available genomic data implicated an exponential growth of these databases, causing a lack of annotations for many recent entries (mostly genomic) [2]. Therefore, a substantial amount of knowledge important to characterize each biological entity is spread through a vast set of heterogeneous sources. The integration of different data sources is a viable approach to correct and complete our knowledge about these biological entities [7, 6]. Motivated by this fact, we introduce in this paper a novel approach for text classification, which involves integrating extracted information from biological web resources into the text classification process.

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 articles or passages from document corpora [19]. Most approaches to text classification are based on Statistical Natural Language Processing [12]. They apply quantitative methods for automated language processing, using probabilistic modeling, information theory, and sometimes linear algebra. Our method extracts information from biological web resources and integrates it in a statistical representation for each article. This representation is then used by a statistical classification algorithm based on a bag-of-words representation, in order to create a text classification model [15].

To evaluate the feasibility of our approach, we implemented the proposed method and submitted an entry to the *KDD2002 Cup challenge: bio-text task* [1]. The task consisted mainly in determining whether an article contained experimental results of interest, and indicating which gene-products (transcript or protein) were associated with the results.

The rest of this paper is structured as follows. Section 2 details the proposed method. In Section 3 we explain the implementation and analyze our results on the KDD Cup task. Finally, in Section 4, we express our main conclusions.

2 Method

Our classification method relies on biological results stored in public databases available on the web. These databases usually associate their data with bibliographic information, which provides a powerful source for document classification.

The motivation for our method derives from the observation that most authors of recently published biomedical articles also submitted their results on these public databases. Since results in a database are stored in a structured form, they can be easily used in an automated system. We named our method WTC (Web Text Classification) and we can describe it as follows:

Input:

- An article with its content and its meta-data (e.g. title, authors, accession number in a bibliographic database).
- Database(s) where relevant information about the articles can be found.
- A controlled vocabulary, i.e. a list of terms that will be used as features in the text classification algorithm. This list should be automatically extracted from the values assigned to database fields.

Output:

- The number of occurrences for each term in the controlled vocabulary.

Procedure:

1. We identify all database accession numbers associated with the article. 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, indicating in which database 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. three letters followed by 9 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 bibliographic information is not available but there is some data describing the information source (e.g. the authors, the date, the laboratory, the technique) we can compare this data with the article's meta-data to figure out if they represent the same information source.

2. We retrieve the content of the database entries, identified by the accession numbers selected in the last step. Depending on the database, we should select the fields that contain relevant information for our objective. This way, we only retrieve the values assigned to these fields.

3. For each term in the given controlled vocabulary, we compute the number of occurrences in the data retrieved from the databases.

By applying WTC to an article, we create a statistical representation of it. This representation is then used by a traditional statistical classification method to build a classification model. The learning process will need a training set of articles where for each article we have to specify its expected classification.

3 Case Study

We experimentally evaluated WTC 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 [18]. 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 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.

Output:

- For each article, provide a boolean decision on whether or not there are relevant experimental results reported on it.
- For each article assumed to have relevant experimental results, indicate the genes involved and discriminate also the gene-product type (transcript, protein, or both).
- Return a ranked list of articles, sorted by the assurance degree in them 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. For each article in the training set the expected output was provided. The output of 283 articles specified that they contained relevant experimental results. For these articles, their output was extended with the result evidences. The FlyBase curators gave a list of fields, which allowed values when mentioned in the article were taken as sufficient evidence for having relevant experimental results. Each evidence was represented by the name of the field, its value, and an identifier for the gene-product involved.

3.1 Implementation

Our approach consisted in using the WTC method for feature extraction with a statistical classification method.

3.1.1 Bag-of-words Representation

The first stage of our approach was to represent each article by a bag-of-words. This is a common technique in statistical text classification systems. Instead of using the full text of each article, we selected only the text assumed to contain relevant information.

We started by retrieving a controlled vocabulary of 315 terms from FlyBase database. These terms were retrieved from the values of evidence fields in the database entries describing experimental results. Since there were general terms that were specialized by other terms, we structured the controlled vocabulary as a hierarchical tree.

We removed the introduction, related work and bibliography sections of each article. The information contained in these sections is very important, but is difficult to deal with it through a bag-of-words representation. We observed that these sections usually cite relevant experimental results from other articles. Since this information is of no interest, it would only cause problems to the classification process. We delineated individual sentences on the rest of the text for each article. The sentences that did not contained any term from the controlled vocabulary were removed. After also removing the stop-words, we obtained a bag-of-words representation of each document.

3.1.2 WTC Representation

We retrieved the meta-data of each article through its PubMed identifier (an interface to the public bibliographic database MEDLINE [13]). This identifier was provided for each article. The controlled vocabulary was the same used in the bag-of-words representation to filter the text. The selected external biological sources were:

- MeSH (Medical Subject Headings) [14], keywords for classifying articles.
- GenBank(GenPept) [3], a repository of gene (protein) structure data.

There was no need to execute the first step of WTC 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 of using the articles meta-data to retrieve accession numbers. For each article, we applied WTC three times one with MeSH, other with GenBank, and at last with GenPept. The results were three different WTC representations of each article. Since the controlled vocabulary was organized as a hierarchical tree, we assumed that when a term occurs, all of its ancestor terms also occur.

3.1.3 Statistical Classification Method

We decided to use Support Vector Machines (SVMs), because it has been reported as one of the top performing methods for text classification and because we had previous experience with it [20, 16]. Given a training set in a vector space, this method provides a model to classify the test set in that vector space. A vector space is composed by a list of instances and a list of features. For each instance, a value is assigned to each feature. Each instance of the training set also has to be classified to execute the learning process. We used the SVM light package with a binary classification, which classify each instance with -1 or $+1$ [9].

To decide if each article had relevant experimental information, we concatenated the bag-of-words representation with the three WTC representations and applied it on the learning process. Our instance space was the collection of articles. The feature space was the words and the terms from the controlled vocabulary. For each article (instance), the assigned value to each word or term (feature) was given by number of occurrences from the bag-of-words an from the WTC representations.

Since the correct classification of the test set was not provided, we divided the training set in the following two sets to evaluate our approach: the pre-training set with 782 articles and the pre-test set with 80 articles. We used the pre-training set in the SVMs learning process to automatically classify the pre-test set. Next, we computed the precision and recall just by comparing the SVMs classification of the pre-test set with its correct classification. This was useful to adjust our approach to achieve a better efficiency. Since an article has relevant experimental results if it has at least one type of evidence, we concluded that executing a learning process for each type of evidence of experimental results was better than simply executing it once. To discriminate the learning process for each type of evidence we had only to modify the classification for each instance, the vector space was always the same.

We created the ranked list of articles based on the maximum SVMs classification value of all types of evidence, since our SVMs automatically classified each article with a value between $[-1, +1]$ for each type of evidence. This maximum value represents the article's strongest evidence of having relevant experimental results.

For the articles automatically classified as having relevant experimental results, we used the GenBank and GenPept WTC representation to identify the gene-products involved. The instance space was the collection of article combined with its list of genes, i.e. each instance represented an article and one of its genes. The GenBank and the GenPept database entries contained the gene-product's name, Therefore, we modified the WTC method so it could represent each article and gene by only retrieving information from GenBank or GenPept entries related to that gene. The synonym list was here very useful to identify the genes in these external databases. The GenBank WTC representation was applied to the gene transcript decision, and the GenPept WTC representation was applied to the gene protein decision. In this sub-task, we also concatenated the bag-of-words representation with the WTC representation. However, for each instance (article and gene) the article's sentences that did not mentioned the gene were removed.

3.2 Results

In this section, we compare our results with the results from the other 31 submissions. These results 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 [5], the article decision and the gene-product decision by the standard F-measure [12]. The overall score was obtained by the sum of these three scores. The results were normalized to range from 0% to 100% representing the efficiency of the systems.

The graphic in figure 1 shows the results for the three sub-tasks and the overall score. The *Best* values represent the highest score that was always obtained by the same team. The *1Q* values represent the score limit of the first quartile [11], i.e. in this case 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 *Our* values represent our submission scores. Our overall score was in the first quartile as 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%.

The ClearForest and Celera team developed the winning system of the KDD Cup task [17]. The system was implemented through a rule-based general In-

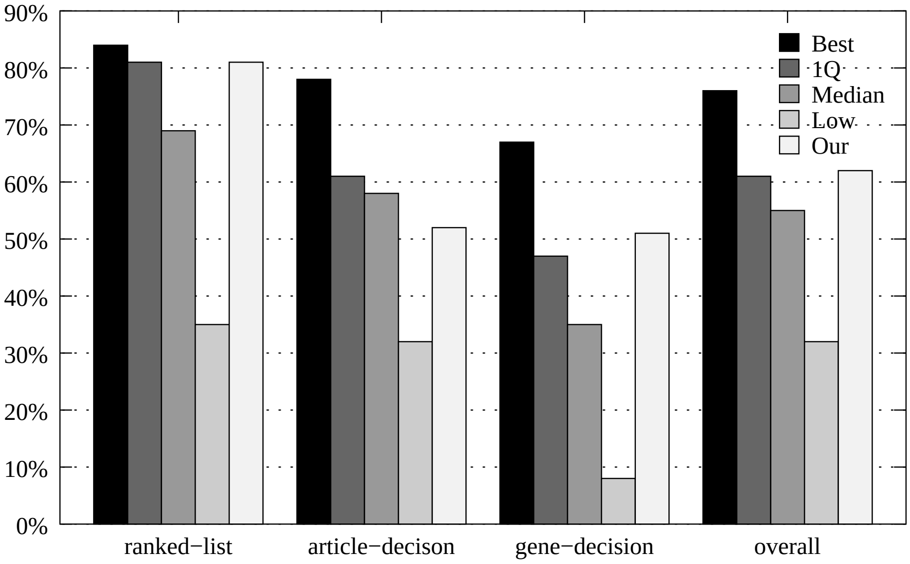

Figure 1: KDD Cup Scoring Results

formation Extraction language. The rules were sequences of pattern matching terms, and were built based on knowledge domain and on purpose for the task.

A team from Singapore obtained a honorable mention by developing a system based on feature extracting with a Naïve Bayes Classifier [10]. 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 [8]. Their system was also based on feature selection and on statistical classification methods. Their 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.

4 Conclusions

This paper introduced a novel approach for text classification. It involves integrating extracted information from biological web resources with common statistical text classification methods. The performance of our approach was evaluated through the *KDD2002 Cup challenge: bio-text task*. Our system was classified on the first quartile. The main problem of our system was a low recall in one of the sub-tasks. The reason for this problem was the small number of external biological sources used. Due to the short period available to implement the system, we were not able to cover other resources, which certainly would increase our recall. Nevertheless, the evaluation indicates that WTC can provide an effective alternative to the essential domain knowledge provided by human curators. This deserves further study, as we hope it could match the performance of methods that are based on domain knowledge introduced and maintained manually.

References

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

[2] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Longman Higher Education, 1999.

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

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

[5] 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.

[6] F. Couto, M. Silva, and P. Coutinho. Rebil : Relating biological information through literature. In *Intelligent Systems for Molecular Biology*. Intelligent Systems for Molecular Biology, poster, 2003.

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

[8] 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.

[9] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter 11 : Making Large–Scale SVM Learning Practical. MIT Press, 1999.

[10] 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.

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

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

[13] MEDLINE. PubMed database at the National Library of Medicine. www.ncbi.nlm.nih.gov.

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

[15] T. Mitchel. *Machine Learning*. McGraw-Hill, 1997.

[16] M. S. N. Maria. Theme-based retrieval of web news. *3rd International Workshop on the Web and Databases*, May 2000.

[17] 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.

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

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

[20] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.