

# Supporting the Curation of Biological Databases with Reusable Text Mining

Olivo Miotto<sup>1,2</sup>

Tin Wee Tan<sup>2</sup>

Vladimir Brusic<sup>3,4</sup>

olivo@iss.nus.edu.sg

tinwee@bic.nus.edu.sg

v.brusic@uq.edu.au

<sup>1</sup> Institute of Systems Science, National University of Singapore, 25 Heng Mui Keng Terrace, Singapore 119615

<sup>2</sup> Department of Biochemistry, The Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, Singapore 117597

<sup>3</sup> Australian Centre for Plant Functional Genomics, School of Land and Food Sciences and Institute for Molecular Bioscience, University of Queensland, Brisbane QLD 4072, Australia

<sup>4</sup> Institute for Infocomm Research, Singapore, 21 Heng Mui Keng Terrace, Singapore 119613

## Abstract

Curators of biological databases transfer knowledge from scientific publications, a laborious and expensive manual process. Machine learning algorithms can reduce the workload of curators by filtering relevant biomedical literature, though their widespread adoption will depend on the availability of intuitive tools that can be configured for a variety of tasks. We propose a new method for supporting curators by means of document categorization, and describe the architecture of a curator-oriented tool implementing this method using techniques that require no computational linguistic or programming expertise. To demonstrate the feasibility of this approach, we prototyped an application of this method to support a real curation task: identifying PubMed abstracts that contain allergen cross-reactivity information. We tested the performance of two different classifier algorithms (CART and ANN), applied to both composite and single-word features, using several feature scoring functions. Both classifiers exceeded our performance targets, the ANN classifier yielding the best results. These results show that the method we propose can deliver the level of performance needed to assist database curation.

**Keywords:** biological databases, database curation, text mining, machine learning

## 1 Introduction

Biological research increasingly depends on computational analysis of data. A wide range of biological data repositories have emerged to fulfill this need. Primary data sources, such as GenBank [3], offer vast quantities of data with a very broad content coverage. To cope with the increasingly large volume of published data, these databases often streamline their data entry process by relying on automated submission mechanisms. At the opposite end of the scale, there are also thousands of specialized data repositories, focusing on particular molecules, organisms or diseases, which offer smaller sets of richly annotated records. To ensure data of the highest quality, these databases generally follow a manual data entry and curation (annotation) process [7].

Manual curation is performed by domain experts- knowledgeable scientists who represent valuable and often scarce resources. Their primary source of data is scientific literature, usually peer-reviewed journal articles. Database curators search biomedical research literature for facts of interest, and manually transfer knowledge from published papers to the database. Recently, widespread online publication of journals has dramatically improved the availability of literature [12] and the automation of search operations, both essential for a curator. However, electronic publishing has also caused an

increase in the volume of literature, which is compounded by the continuous rapid expansion of biological knowledge. As a result, the manual curation process remains a time-consuming, expensive process that is prone to omissions and inconsistencies [15]. This knowledge transfer bottleneck slows down the pace of research, and therefore there is considerable interest in technological solutions that minimize the curators' involvement, or replace them altogether. In particular, *text mining* techniques enable various degrees of automation of the analysis of scientific literature, such as the identification of named entities, the classification of documents, and the extraction of relevant facts [4]. Although they are still not capable of fully automated extraction of correct information from texts, these approaches keep improving, and perform useful tasks. Yet, their adoption is very limited, largely because they require knowledge and infrastructure available only to a few database curators, and not at all to average scientists. The use and configuration of text-mining tools must be simplified, hiding complexity from users with limited skills in programming and linguistics.

In this paper, we discuss the text mining needs of biological database curators, and assess the suitability of current text mining techniques for use in intuitive curator-oriented tools. We propose a method for text categorization, which can be trained by a curator to select the most relevant documents, and does not require domain-specific programming or knowledge of linguistics. The key characteristics of a curator-oriented software tool implementing this method are also described here. To demonstrate that current text mining techniques can deliver appropriate performance under these conditions, we tested two machine learning algorithms on a real-life curation problem: the identification of scientific abstracts that describe allergen cross-reactivity. We report results obtained using different feature selection mechanisms (single word vs. composite features), in association with four different scoring functions. Our results show that standard text mining software, without domain-specific adjustments, can deliver an adequate level of performance to support curation tasks.

## 2 Background

Curators face major challenges in all stages of the conversion from unstructured scientific literature to structured data *e.g.* database records. Scientific articles are highly specialized and often hard to understand even for experts in the area, making it difficult to identify all the interesting facts at the *knowledge extraction* stage. The variety of writing styles compounds this problem, since facts are not always clearly stated. Analyzing a paper is therefore a lengthy exercise, but waste of effort can be minimized by effective upstream *selection and filtering of documents*. Scientific abstracts are very valuable for evaluating the information content of a paper: they provide more limited information than the full paper text, but are information-rich and typically summarize the main results. Although some repositories base their curation process on scanning all abstracts from selected scientific publications [2], this approach is impractical for both smaller projects and for broad research topics, so pre-selection of articles is highly desirable.

Recently, the use of text mining algorithms has been proposed for streamlining various aspects of the curation process. The phrase *text mining* loosely denotes the analysis of text documents by machine learning and natural language processing (NLP) algorithms. In most cases, the objective is not to discover new knowledge (as a more rigorous definition of text mining [8] would demand), but rather to assist recovery of information from text. The text mining process [6] has four stages, ordered by increasing complexity:

- a) *document categorization* identifies documents relevant to given topics
- b) *named entity tagging* isolates concepts and names important to the problem space
- c) *fact extraction* extracts items of meaningful knowledge
- d) *collection-wide extraction* discovers new knowledge by correlating facts from multiple documents.

*Fact extraction* systems are suitable for automating the annotation of database entries. Recent promising results include the successful annotation of genes and proteins, and extraction of biological interactions [6, 10]. However, even the best state-of-the-art systems are not as accurate as human curators. Automatic maintenance of high-quality databases demands high *precision* (high proportion of true positives), which usually comes at the expense of lower *recall* (capturing a smaller portion of all published knowledge). This trade-off is evident in a study of automated annotation of enzymes [10], which deemed 92% precision and 50% recall as “sufficient for inclusion in a high-quality database”. Indeed, a high precision is necessary if the data in the repository is to be trusted, but 50% recall omits half of all available knowledge, which is an unacceptable trade-off for most human curators.

Current limitations of automated fact extraction suggest that curators are still necessary mediators between published literature and databases. Still, curators can be supported by *document categorization* systems that select and filter documents, reducing workload without compromising the quality of results. Machine-learning classifiers can achieve relatively high recall, at the cost of reduced precision; in other words, they can find a high percentage of all available knowledge, if one can accept somewhat “noisy” results. Since human curators are highly effective as quality filters, it is often acceptable to relax precision requirements to achieve higher recall. In practice, this trade-off depends on the nature of the database and the abundance of documents being reviewed. Supported by text mining systems, curators can efficiently discard irrelevant documents, thus significantly improving annotation speed. Lower work volume and higher speed have effects on quality: while working on our local allergen database, rapid result filtering enabled us to identify errors in existing records, such as incorrect identifiers or non-human experimental results, showing that the advantages of automation extend beyond higher productivity.

The most common approaches to document categorization involve machine-learning classifiers, trained with manually-annotated sets of documents that contain both documents of interest (*positives*) and other documents (*negatives*). Although a variety of successful approaches have been reported, the best results are often obtained as a result of laborious choices of algorithms and document features, to suit the specifics of a particular problem and are, therefore, domain-specific. One prize-winning system, for example, used a combination of sophisticated techniques, and non-obvious document features (figure captions), which are difficult to extract [16]. Such powerful systems are clearly hard to reuse in different contexts, and can only be developed by highly-specialized programmers, often with expertise in *natural language processing* (NLP). Surprisingly, very little research has addressed the need for text mining systems that can be used for a variety of diverse tasks, by curators with limited programming and linguistic expertise.

Cohen and Hersh [4] have stated that current text mining research is biased towards “evaluations based on system output independent of user needs”. They have identified the major challenge in this field: bridging the gap between text mining researchers and database curators, thus “helping biomedical researchers to solve real-world problems that are inhibiting the pace of research”. They highlighted the need for improvement in: a) access to full text articles rather than abstracts, b) identification of the features for analyzing text, c) measurement of true value to users, d) cooperation between end users and text mining researchers. We have identified *usability* and *reusability* of text mining tools as additional areas for improvement. Currently, even the most accurate algorithm cannot benefit database curators, unless text mining experts are available for tools development. Of the thousands of specialized databases currently online, very few (such as BIND [2]) can count on the availability of such experts. Curators therefore need reusable, configurable and customizable tools that serve their needs, without requiring them to become skilled programmers. Limited access and availability of full text articles are still serious barriers to effective text-mining. Even when the full text of a research paper is available, the need for subscription limits the application of automated data mining tools. It is clear that in the foreseeable future, most biomedical discovery from text will be strongly reliant on journal abstracts, which are freely available from large repositories. For example, PubMed [18] contains abstracts of articles published in a large number of biomedical journals

(over 15 million abstracts as of August 2005). Even when high-coverage full-text indexing becomes available, it is likely that searches on abstracts will still be valuable as a preliminary analysis. It is also worth noting that full text articles are currently mostly available in HTML or PDF, formats in which the identification of significant document parts (such as figure captions) is possible but not straightforward. Standard encoding of document structure (such as using XML), clearly benefits the development and use of text mining systems.

### 3 Proposed Approach

To help curators streamline their work, we propose a new class of document categorization systems, whose chief characteristics can be summarized as follows:

- Single, intuitive user interface, not requiring significant programming or linguistic abilities
- Ability to connect to major databases (e.g. PubMed) and retrieve documents transparently
- Simple integrated mechanism for rapid annotation of positives and negatives by the curator
- High configurability: curators can specify PubMed queries, provide lists of named entities, specify keywords, word and phrase patterns, etc.
- Ability to make key classification decisions automatically, such as classifier feature choice, classifier parameter, etc., hiding technological complexity from the user
- Ability to learn gradually, incrementally and interactively from the curator's annotations, so that a large initial training corpus is not required

The nine-step workflow of a curator-oriented document categorization systems has been summarized in Figure 1. We distinguish three main stages: during *corpus building* (steps 1-4), the system fetches documents from the source database, and pre-filters them, creating an unclassified corpus; during *curator annotation* (step 5), the curator marks selected documents from the corpus as positives or negatives; and during *unsupervised ranking* (steps 6-9) the system uses the curator's classification to identify the most likely positives from the unclassified documents and present them for annotation. The process of choosing documents from the unclassified pool, presenting them for manual classification, and consequently refining the classifier, is known as *active learning*, and has been successfully used for text classification [17].

In step 1, the tool must be able to handle a large volume of results from broad searches, since it is usually difficult to formulate highly specific queries to support curation. Step 2 extracts the analysis target text from the rest of the document, using a structure-aware approach (such as using the XML standards). In steps 3 and 4, the document text is analyzed for keywords, identifiers, patterns, and entries from ontologies- documents are included in the corpus based on the outcome of this analysis. The tool must therefore allow the curator to specify and maintain lists of keywords and text patterns of interest, to plug in standard vocabularies (such as lists of identifiers and their synonyms), and to specify inclusion criteria for the resulting corpus.

In the annotation step (step 5, see also Figure 4), users are supported by simple mechanisms to aid classification. To facilitate decision-making, the user interface should highlight those items in the text that determined the inclusion of the document in the corpus. By classifying documents, the curator provides new training examples for the next phase, in which the system identifies a set of classification features, and the feature score for each example (step 6). The process of feature selection and scoring should be statistical and fully automated, as should be the training of one or more classifiers (step 7) and the evaluation of these classifier, based on the performance against the classified examples (step 8). In step 9, the best classifier (or combination of classifiers) is used to rank unclassified documents,

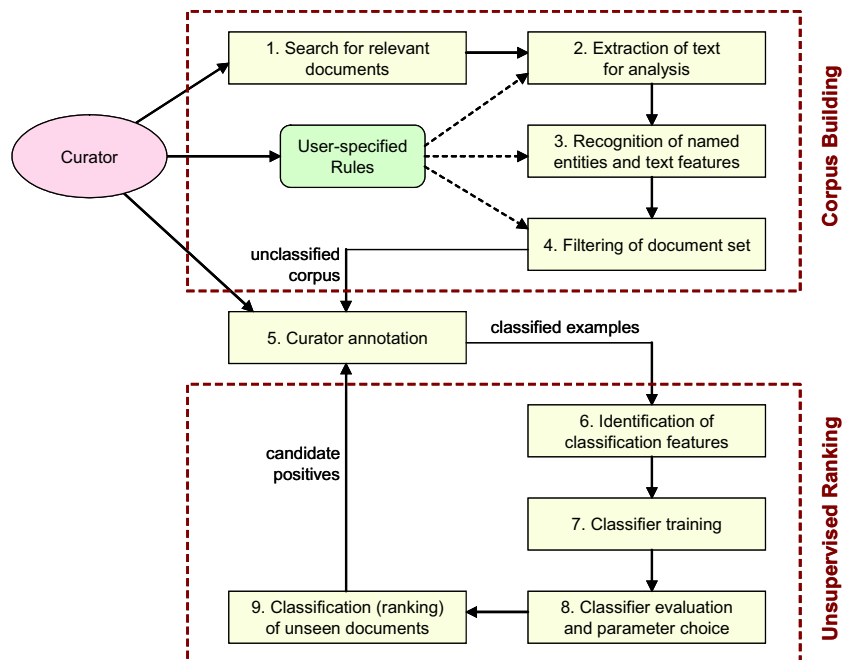


Figure 1: The workflow of a curator-oriented document categorization tool, showing the tasks, and the points at which the database curator is involved.

so that the user is presented with the most likely positive candidates for further annotation. Both corpus building and unsupervised ranking can be executed at user-specified intervals to maintain both the corpus and the trained classifiers up-to-date.

## 4 Materials and Methods

We set up a proof-of-concept system to demonstrate the utility of our method on a specific real-world curation task, and measure its performance. We compared two common generic machine learning classifiers for document selection. Documents were pre-filtered by user queries, and scored using statistically selected features. We compared the use of two types of features, and four scoring functions, to investigate whether any combination therein offered advantages. We did not aim to produce a complete working system; in particular, we did not address the active learning process, which will be the subject of later work.

### 4.1 Curation Task Overview

In a study case for our method, we addressed the curation needs of the ALLERGEN database (<http://research.i2r.a-star.edu.sg/Templar/DB/allergen>). ALLERGEN contains records of human allergen proteins, extracted from literature and enriched with annotations on the biochemical properties of these allergens. We focused on the specific task of identifying information on *allergen cross-reactivity*. Cross-reactive allergens share structural similarities at molecular level, causing the immune system of certain individuals to react to multiple allergens. ALLERGEN stores cross-reactivity information, used for allergen avoidance in patients with severe allergies.

Our document categorization task was to identify all relevant PubMed abstracts that report cross-reactivity. This information generally involves two named allergens, and a statement describing cross-reactive properties. Cross-reactivity statements are not expressed consistently - some abstracts contain

a clear sentence with the words “cross” and “reactivity” (or derivatives), but others imply cross-reactivity indirectly.

The identification of named entities was supported by the WHO/IUIS Allergen Nomenclature [9], a naming standard for allergens. Allergen identifiers consist of a capitalized 3-letter word, followed by one lowercase letter and an integer (e.g. “Mal d 1”). The standard allows some variations (such as “Mala f 1” and “Pru av 3”). The IUIS nomenclature also provides an “official list” of over 600 allergens, which can be used as an ontology. The IUIS nomenclature was not in use before 1994 and is being gradually adopted (Figure 2). We accepted these limitations, with a plan to extend the system to non-standard allergen names later.

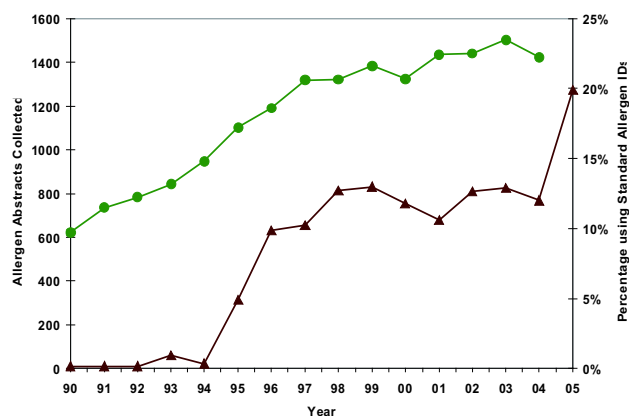


Figure 2: The percentage of abstracts that use IUIS allergen identifiers (triangles) is shown beside the total number of abstract in the corpus (circles) since 1990. Usage of standard identifiers became widespread from 1994, and is currently around 20%. Many abstracts in the corpus have no mention of specific allergens.

## 4.2 Experimental Setup

We built our system using the Aggregator of Biological Knowledge (ABK) [13], an extensible system for retrieving, aggregating and analyzing heterogeneous data from multiple data sources. ABK is a user-oriented platform for biological knowledge discovery comprising four subsystems: a) an *extensible mediator framework* which handles Web-enabled connectivity, b) a *rule-based extractor and resolver* which extracts the information from search results, according to structural rules specified by the user, c) a *recordset manager* which provides data management functionality, and d) a *plug-in tool framework* for analyzing and manipulating stored information. The architecture of ABK is shown in Figure 3.

For our study, the mediator framework delivered user queries to PubMed, retrieving results as XML documents. ABK relies on XML encoding for handling the results, and allows users to specify structural rules for the extraction of data from retrieved documents. We defined rules for extracting the abstract text, title, journal name and year of publication. The resulting records can be viewed in a spreadsheet-like graphical user interface. ABK supports analysis of the record by means of software tools, added as plug-in extensions. We developed a number of reusable text analysis tools, to form a basic literature analysis workbench for conducting our study. Although the ABK system is capable of connecting to a number of data sources, and integrating their knowledge, our study only used documents from PubMed. More details about the ABK system are available in [13].

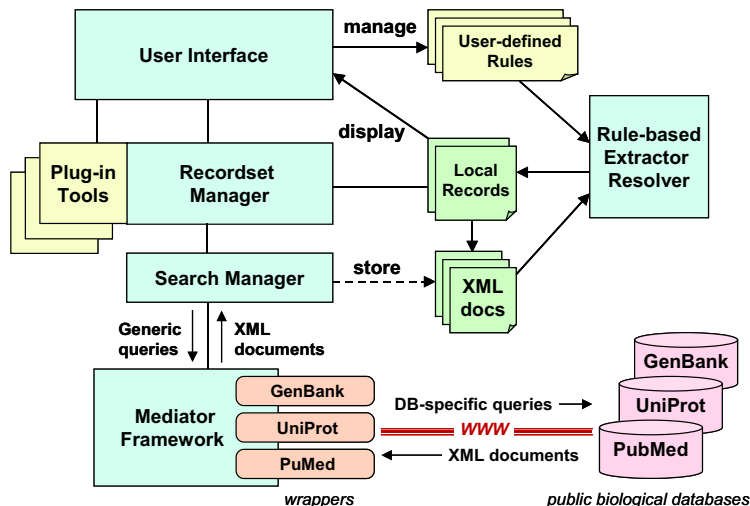


Figure 3: Architecture of the ABK system.

### 4.3 Corpus Collection and Annotation

We used ABK to collect 26,997 PubMed abstracts containing the word “allergen”, and automatically extract their abstract text. Named entities were easily identified by an ABK plug-in (the *Text Analyzer Tool*) which performed generic text analysis tasks, such as identifying sentences, and matching regular expressions and keywords from user-supplied lists. The tool was configured to find keywords such as “cross” and “reactive”. It also found identifiers, both from the IUIS “official list”, and by matching the IUIS nomenclature pattern with a regular expression. This basic analysis uncovered 71 identifiers used in literature but not included in the IUIS official list, showing it lags behind current usage. We finally filtered those documents that contained at least two different named allergen identifiers, forming a corpus of 584 abstracts.

To create training and test sets, the corpus was manually annotated by a curator to separate positives and negatives (a positive is defined as an abstract that contains information on cross-reactivity between allergens). The annotation process was supported by the *Corpus Annotator Tool*, an ABK plug-in (Figure 4). This tool displays the abstract text, highlighting the named entity features discovered by prior steps. Highlighting helps focus the curator’s attention to key terms, and speeds up annotation. Annotation of each abstract is a straightforward task: a button click determines if the abstract is a positive or negative, while a double click on the text selects *key sentences*—those sentences which capture the cross-reactivity information. Annotation of the full corpus by an expert took approximately 4 hours, identifying 73 positives and 511 negatives. Only 39 positives captured cross-reactivity information in a single key sentence, while 28 required two key sentences, and the remaining 6 contained three or four key sentences. A higher number of key sentences indicate that the abstract is vaguely worded, which was sometimes hard to interpret even for the curator. Six positives did not contain the words “cross” and/or “reactive” (or their derivatives).

We collected statistics on the position of key sentences within the abstracts, with the intent of investigating if this information can be used bias score features. Key sentences were significantly more likely to be found in the last third of the abstract than in the rest (Figure 5).

### 4.4 Feature Selection and Scoring

To train classifiers, we compared two types of features: *single-word* and *composite* features, the latter consisting of group of words that co-occur frequently in sentences. This comparison aimed at testing whether the widely used bag-of-words approach [11] has inherent weaknesses. The presence of a phrase

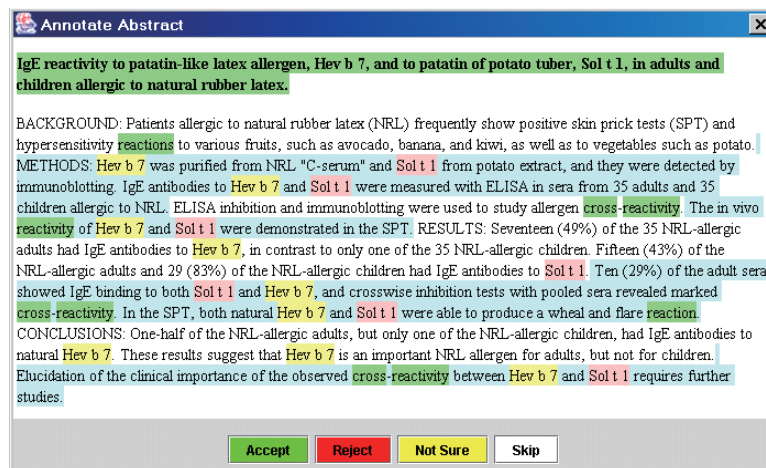


Figure 4: Screenshot of the ABK Corpus Annotator Tool. Features highlighted include: IUIS allergen identifiers (yellow), other allergen identifiers (pink), cross-reactivity keywords (green), and sentences containing at least two identifiers (light blue).

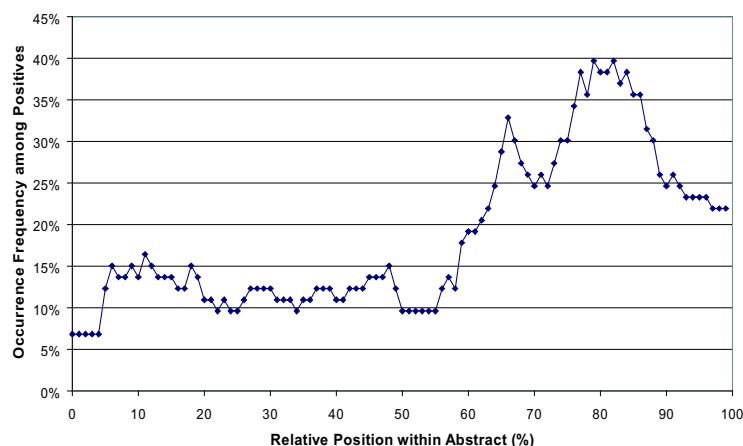


Figure 5: Key sentence occurrences in different parts of abstracts. Each abstract was divided into 100 bins, and a value of 1 was assigned to each bin that overlapped with a key sentence.

such as “high blood pressure” in an abstract is clearly more informative than the presence of each constituent word. The identification of such word combinations usually demands linguistic analysis. However, we can find sets of frequently co-occurring words, known as *frequent itemsets* [1], using commonly available statistical algorithms without linguistic analysis; frequent itemsets can be used as composite features [5].

The Text Analyzer Tool split each abstract sentence into words, discarding stop words and words beginning with digits. The remaining words were stemmed by the Porter stemmer algorithm [14], reducing term variants (e.g. “analysis”, “analyses”, and “analyze”) to their common roots. For each abstract sentence, the *Sentence Transaction Tool* (an ABK plug-in) produced a transaction record, consisting of all stemmed words, without repetitions. Separate positive and negative transaction files were produced; only key sentence transactions were included for positives. The transactions were analyzed by the Apriori algorithm [1]. In positive examples, 1547 itemsets (chiefly combinations of the most frequent words) had at least 5% statistical support; negatives were more heterogeneous (623 itemsets with support of 1% or above).

To make the classifiers more generic, the combined feature lists was reduced by excluding 58 words specific to major sources of human allergens (such as “dog”, “dust” and “cockroach”), based on the assumption that they were irrelevant towards cross-reactivity classification. This exclusion is problem-specific and cannot be automated; however, it requires domain knowledge rather than technical knowledge, and can be easily carried out by a curator, given a sufficiently intuitive user interface.

To select the most informative features, we generated feature score vectors using the *Abstract Statistics Tool* plug-in. Each vector was tagged with the abstract class (negative or positive) and a value for each frequent itemset: 1 if the itemset could be found in the abstract, 0 otherwise. The feature vectors were used to measure the information gain of each itemset and, as a result, we selected the top 64 features (following the rule-of-thumb of using approximately ten training examples for each feature). We used the same process to identify single-word features, by configuring Apriori to identify frequent itemsets of length one.

The selected features were used to produce data files for training and testing the classifiers. Each record consisted of a vector containing a score for each feature, and a class identifier. We experimented with four different score functions:

1. **PRESENCE**. Score is 1 if the feature is found in the abstract, 0 otherwise.
2. **COUNT**. Score is the number of times the feature is found in the abstract.
3. **POSITION**. Same as COUNT, but score is doubled for occurrences in the last 35% of the abstract (based on results shown in Figure 3).
4. **COLOCATION**. Same as COUNT, but score is doubled in sentences that contain one allergen identifier, and quadrupled in those with two or more.

## 4.5 Document Classification

We used the resulting data files to train and test two types of classifiers, which are representative of highly diverse approaches to machine learning:

1. **Artificial Neural Network (ANN)**. We used an ANN from the commercial Neuroshell 2 suite (<http://www.wardsystems.com/products.asp?p=neuroshell2>), choosing a Probabilistic Neural Network (PNN) architecture, using a genetic algorithm for determining appropriate feature smoothing factors.

2. **Decision Tree (CART)**. We included a decision tree classifier, using the CART 5.0 package (<http://www.salford-systems.com/cart.php>). A cost of 4.0 was assigned to misclassified positives.

Classifier performance was assessed in terms of recall (R) and precision (P), using a test set consisting of 30% of the examples, randomly chosen by the classifier. As we previously stated, our main objective is to pre-select documents before manual curation, and the intervention of a human curator allows the precision requirements to be relaxed, privileging higher recall. We set performance targets to  $R > 75\%$  and  $P > 40\%$ , which was deemed to be a reasonable trade-off, when accounting for the time necessary for a curator to visually discard false positives.

Table 1: Classifier performances. Results were obtained using *composite features* (A), and *single-word features* (B). Recall and precision against a random test dataset are shown for each scoring function. The best performance figures are circled.

(A) Classifier Performance using **Composite Features**

		PRESENCE	COUNT	POSITION	COLOCATION
ANN	Recall	78.9%	79.0%	63.2%	78.9%
	Precision	68.2%	68.2%	60.0%	50.0%
CART	Recall	80.0%	76.5%	64.7%	88.2%
	Precision	41.4%	50.0%	44.0%	44.1%

(B) Classifier Performance using **Single Word Features**

		PRESENCE	COUNT	POSITION	COLOCATION
ANN	Recall	78.9%	63.2%	63.2%	68.4%
	Precision	93.8%	92.3%	63.2%	81.3%
CART	Recall	76.5%	76.5%	88.2%	64.7%
	Precision	44.8%	44.8%	46.9%	44.0%

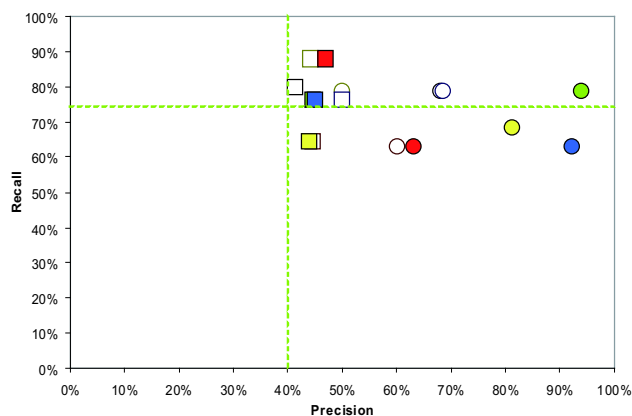


Figure 6: Plot comparing classifier performance figures as reported in Table 1. ANN classifiers are represented by circles, and CART ones by squares. Solid markers show use of single-word features, and unfilled markers denote composite features. The dotted lines show the predetermined performance targets for our classifiers.

## 5 Results and Discussion

We observed that CART builds its decision tree almost solely on features derived from positives, while the ANN classifier also recognizes patterns in negatives. These differences account for several of the variations in classifier performances, which are shown in Table 1.

The most important result is that both types of classifiers exceeded our performance criteria when used with both single-word and composite features, without using any special scoring functions. Figure 6 clearly shows that ANN classifiers are considerably more precise. However, lowering the precision threshold (and forcing the human curator to manually discard more false positives) permits the use of CART classifiers, which boost the recall by about 10%. The means that an additional 10% of knowledge is incorporated in the database: there is a clear trade-off between human effort and database coverage.

CART's lower precision is largely due to its dependence on recognizing positives. On the other hand, ANN is precise precision when using single-word features, many of which were derived from negatives. Although systematic reliance on negative features can actually decrease performance (e.g. when classifying diverse documents), we believe our corpus was representative of allergen-related PubMed abstracts.

The role of the different scoring function was varied, and in some case they impacted classification negatively. The COUNT function presented no performance advantage over PRESENCE. Interestingly, both POSITION and COLOCATION boosted the performance of CART classifiers, but brought no benefit to ANN- probably because these functions are primarily designed to boost recognition of positives. The high impact of POSITION when using single-word features with CART indicates that the presence of certain words in the last third of the abstract is a stronger indicator that the presence of positive-related phrases.

Overall, the performance of ANN classifiers is appropriate to support curation tasks. Although ANN performed well without applying functions, the highest recall figures were obtained applying the COLOCATION function to the CART classifier. This indicates that combinations of classifiers could yield even higher performance, a hypothesis that will be worth exploring further.

Finally, we investigated whether our results were biased by our pre-filtering technique (*i.e.* selection based on IUIS identifiers), by applying the three top-performing ANN classifiers to the full set of 26,997 retrieved PubMed abstracts, which includes positives that do not use the IUIS standard. We then compared the classifiers' prediction against a list of abstracts that do not use the IUIS nomenclature, previously identified during manual curation of the ALLERGEN database. We found that ANN classifiers could identify 91% of these abstracts, indicating that pre-filtering bias was not a significant issue.

## 6 Conclusion and Future Work

In this paper, we have proposed a method for generic and reusable text mining techniques in support of biological database curation, and demonstrated with experimental results the feasibility of this approach. Our method enabled seamless retrieval of documents from simple queries, and extraction of the desired text. Relevant parts of the text were identified by patterns and controlled vocabularies, to aid text analysis and visually assist the annotation task. Annotation was rapid and intuitive, and we showed that the downstream processed of feature selection, feature scoring and classification can be automated. Most importantly, we have shown that this approach can deliver around 80% recall and as high as 94% precision on a real-world database curation task using a standard ANN implementation. The exact choice of features may be more problem-dependent, and our results indicate that combining the results of classifiers using different types of features may be advantageous, though further investigation is required. We obtained consistently good results using compound word features with no special scoring function.

Several areas require further investigation. In the near future, we will be investigating the active learning aspects of the system, to allow for gradual training that does not require a large annotated corpus. We will also investigate automatic selection and tuning of classifier parameters, as well as other types of classifiers. The use of other PubMed fields, such as the MESH terms annotations, will be explored, as will the use of combined results from different classifiers. Finally, user mechanisms for harnessing more complex ontologies (rather than simple word lists) deserve more in-depth research. We believe that these are opportunities rather than obstacles, and look forward to a new class of tools that will truly impact the work of biomedical database curators.

## References

- [1] Agrawal, R., Imielinski, T., and Swami, A.N., Mining Association Rules between Sets of Items in Large Databases, *Proc. of the ACM Intl. Conf. on Management of Data (SIGMOD 93)*, 207–216, 1993.
- [2] Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D’Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M.J., Dumontier, M.R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J.P., Parker, B., Pintilie, G., Pirone, R., Salama, J.J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B.F., and Hogue, C.W., The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Res.*, 33:D418–24, 2005.
- [3] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L., GenBank, *Nucleic Acids Res.*, 33:D34–38, 2005.
- [4] Cohen, A.M. and Hersh, W.A., A survey of current work in biomedical text mining, *Brief. Bioinform.*, 6(1):57–71, 2005.
- [5] Deshpande, M. and Karypis, G., Using conjunction of attribute values for classification, *ACM Intl. Conf. on Information and Knowledge Management (CIKM 2002)*, 356–364, 2002.
- [6] de Bruijn, B. and Martin, J., Getting to the (c)ore of knowledge: mining biomedical literature, *Int J Med Inform.*, 67(1-3):7–18, 2002.
- [7] Fredman, D., Siegfried, M., Yuan, Y.P., Bork, P., Lehv aslaiho, H., and Brookes, A.J., HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources, *Nucleic Acids Res.*, 30:387–391, 2002.
- [8] Hearst, M., Untangling Text Data Mining, *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
- [9] Hoffman, D., Lowenstein, H., Marsh, D.G., Platts-Mills, T.A.E., and Thomas, W., Allergen Nomenclature, *Bull. of the World Health Organization*, 72(5):796–806, 1994.
- [10] Hofmann, O., Schomburg, D., Concept-based annotation of enzyme classes, *Bioinformatics*, 21(9):2059–2066, 2005.
- [11] Joachims, T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *European Conf. on Machine Learning (ECML-98)*, 137–142, 1998.

- [12] Markovitz, B.P., Biomedicine's electronic publishing paradigm shift: copyright policy and PubMed Central, *J. American Medical Informatics Assoc.*, 7(3):222–229, 2000.
- [13] Miotto, O., Tan, T.W., and Brusica, V., Extraction by Example: Induction of Structural Rules for the Analysis of Molecular Sequence Data from Heterogeneous Sources, *Proc. 6th Intl. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL'05), Lecture Notes in Computer Science* 3578, 2005.
- [14] Porter, M.F., An algorithm for suffix stripping, *Program*, 14(3):130–137, 1980.
- [15] Rebholz-Schuhmann, D., Kirsch, H., and Couto, F., Facts from text– Is text mining ready to deliver? *PLoS Biol.*, 3(2):e65, 2005.
- [16] Regev, Y., Finkelstein-Landau, M., and Feldman, R., Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002, *ACM SIGKDD Explorations Newsletter*, 4(2):90–92, 2003.
- [17] Tong, S. and Koller, D., Support vector machine active learning with applications to text classification, *Journal of Machine Learning Research*, 2:45–66, 2001.
- [18] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L., and Yaschenko, E., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 33:D39–45, 2005.