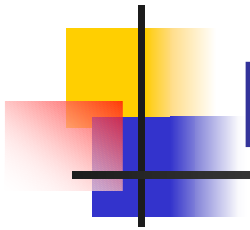
Classifying Biological Articles using Web Resources

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

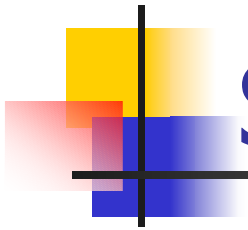
Outline

- Introduction
- Method
- Experimental Data
- Results
- Conclusions



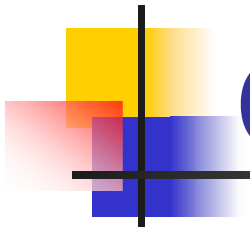
Problem

- MEDLINE
  - "Beginning in 2002 over **2,000** completed references are added daily each Tuesday through Saturday, January through October"
  - "over **460,000** added last year"
  - Reading 10 articles per day, takes **112** years to read those articles
- Curators have to read the literature to update their databases



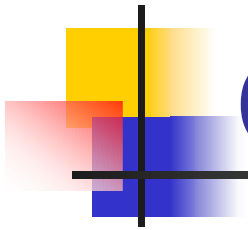
Solution

- Text mining tools:
  - Text classification systems
  - Select only the relevant articles
- Classification needs appropriate features:
  - Derived from domain knowledge
  - Supplied by experts
  - Expensive an limited solutions



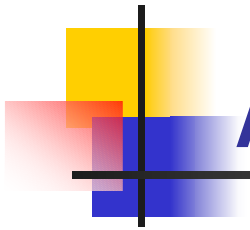
Our Approach

- Use of public web resources for generating new features
- Why to use it?
  - Similar approaches reduced the error rate of classifiers in other situations
- Is it feasible?
  - Besides literature a large amount of additional information about molecular biology is available on the web
  - E.g. GenBank, SwissProt ....



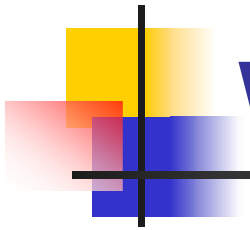
Outline

- Introduction
- Method
- Experimental Data
- Results
- Conclusions



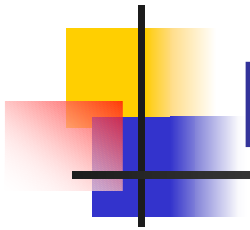
Approach

- Assumption:
  - A large amount of articles have their results published in public databases
    - e.g. GenBank, SwissProt
- Hypotheses:
  - The use of this information can improve the performance of a standard classification system when applied to biological literature



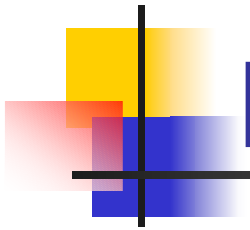
WeBTC (Web Biological Text Classification)

- Input:
  - A collection of biological articles
  - A biological database
- Output:
  - A statistical representation for each article



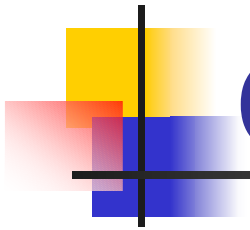
Procedure

- Identify all the database accession numbers related to the given articles:
  - Directly from the article content
  - Cited in the database
  - Article's meta-data matching
- Identify the distinct terms in the selected database entries
- For each article, compute the occurrences of each term in its database entries (bag-of-words)



Example

- The article in PubMed (pmid=12803610), contains the sentence:
  - ”The sequence of the nramp cDNA was filed at the EMBL/GenBank/DDBJ Databases under the accession number AJ514946.”
- The GenBank “AJ514946” entry contains the term (organism name):
  - "Hordeum vulgare subsp. vulgare",
- In the WeBTC output:
  - The organism name will represent a feature
  - Having at least one occurrence.



Outline

- Introduction
- Method
- Experimental Data
- Results
- Conclusions

Experimental Data

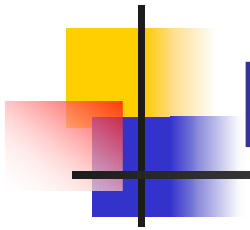
KDD Cup 2002: bio-text task

- Text classification for FlyBase
- Input:
  - A collection of articles about Drosophila (862 for training and 213 for testing)
  - The genes mentioned and a synonym list
  - To better mimic real conditions other external information could also be used
- Output:
  - Select the articles with relevant experimental results
  - For which genes (transcript, protein, or both)
  - A ranked list of articles

Articles' Representations

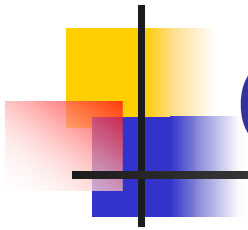
- Application of WeBTC to the following databases:
  - MeSH, GenBank and GenPept
- Integration of the three representations into a single one:
  - WeBTC representation
- Bag-of-words representation of each article using Bow
- KDD Cup:
  - Filtering of the GenBank and GenPept entries to identify which genes had relevant results

Experimental Data



Models

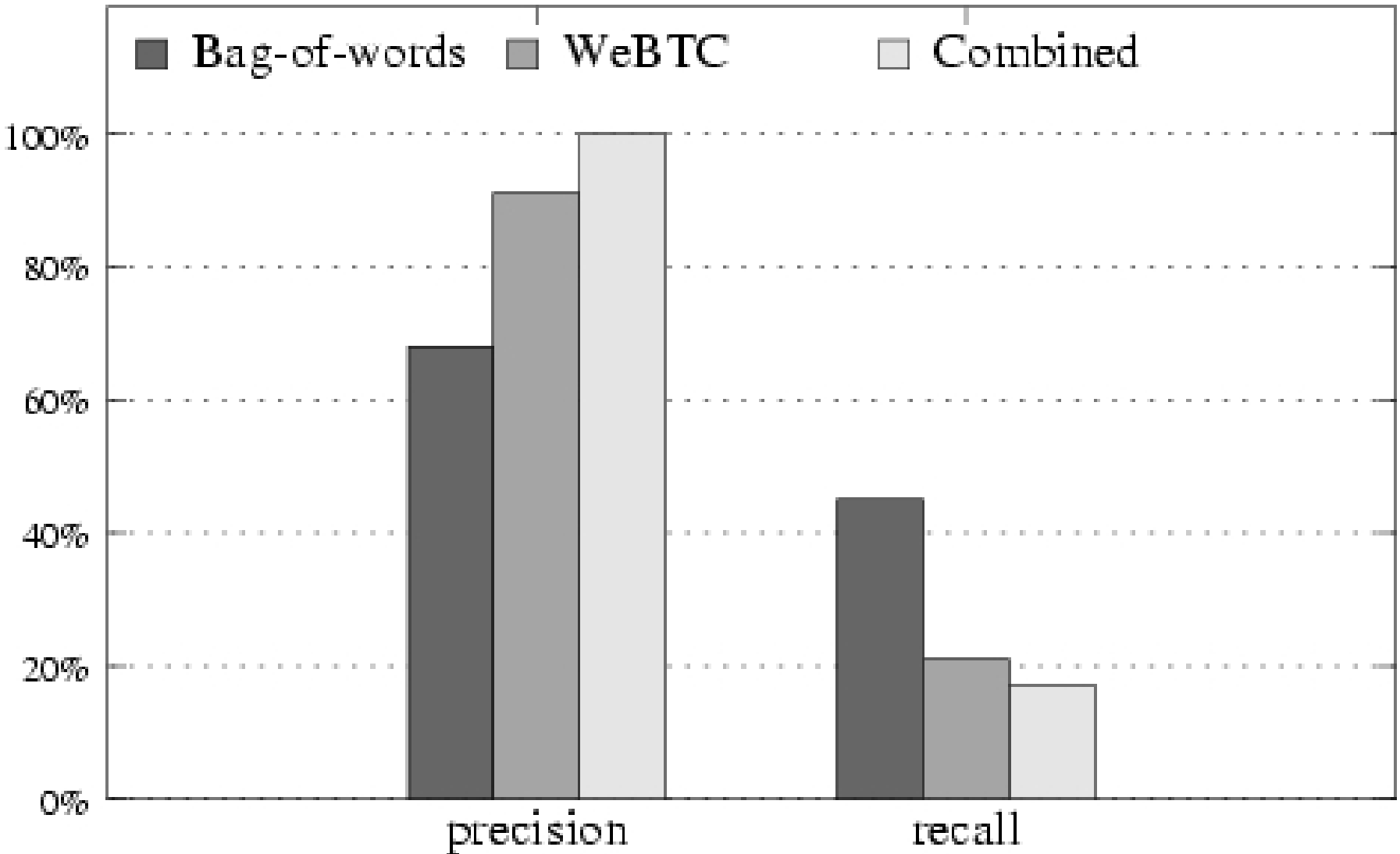
- Naïve Bayes statistical classification method
- Three Models:
  - From the WeBTC representation
  - From the bag-of-words representation
  - Combined, that considers an article relevant iff both models agree in doing so
- KDD Cup:
  - A model for each kind of evidence



Outline

- Introduction
- Method
- Experimental Data
- Results
- Conclusions

WeBTC vs. Standard



WeBTC vs. Standard

- Bag-of-words

41	19
50	103

- WeBTC

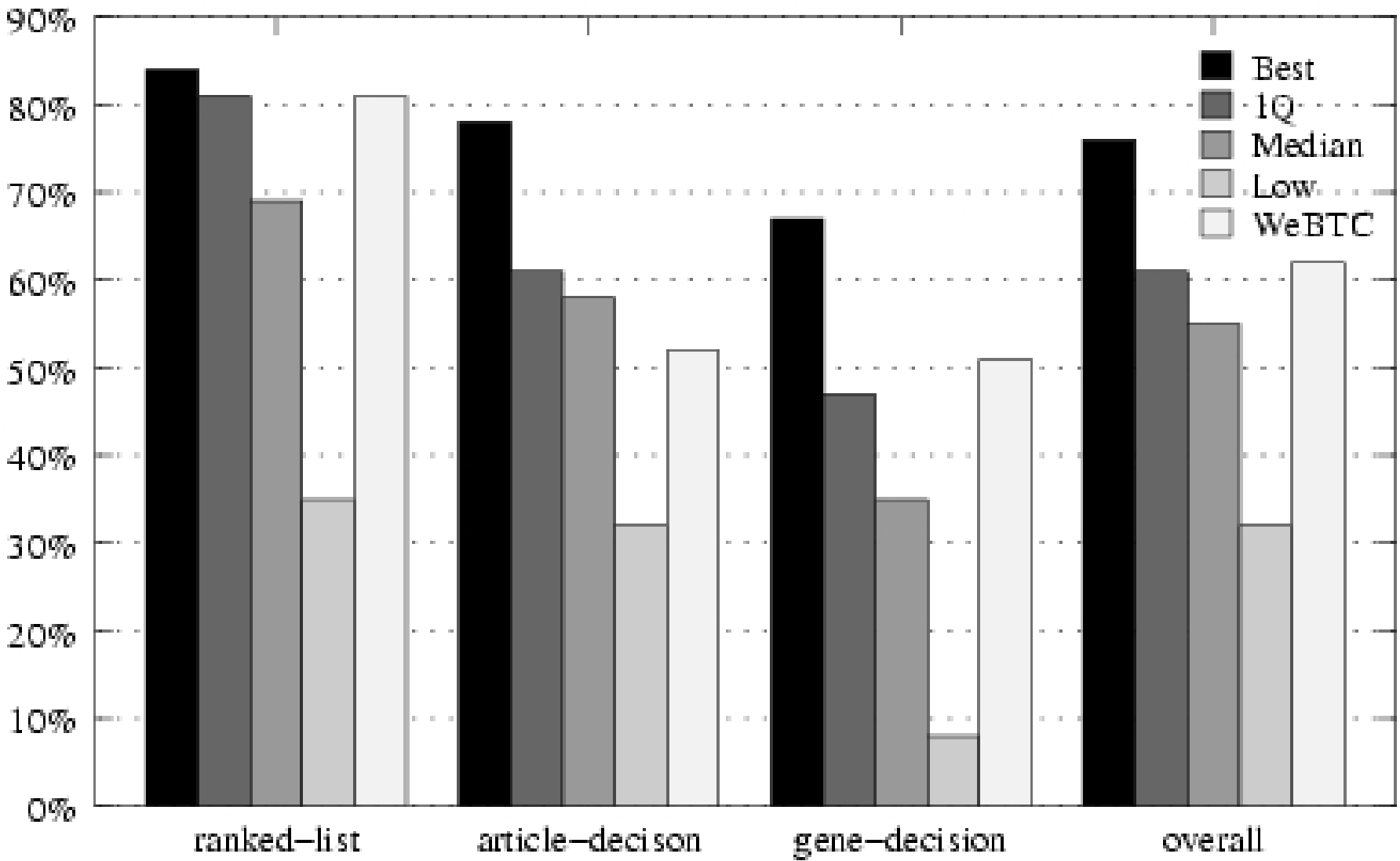
19	2
72	120

- Combined

15	0
76	122

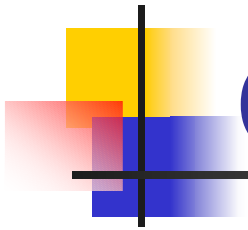
Rows:(MethodPositive, MethodNegative) - Columns:(RealPositive,RealNegative)

WeBTC vs. State-of-art



Main Problems

- Low recall in article-decision:
  - precision of 81% but recall of 38%
- Why?
  - Small number of databases used
  - Recent articles not yet in the databases
  - Meta-data match not implemented
  - Pessimistic threshold:
    - good ranked list
    - positives: training=33% vs. test=43%

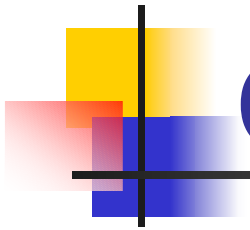


Other Approaches

- Winning systems used domain knowledge supplied by experts
- Weak results for statistical text classification systems without domain knowledge
- Our approach automatically obtained the domain knowledge from the external databases

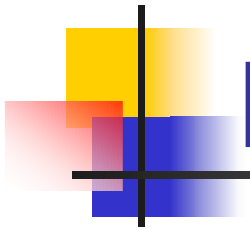
Outline

- Introduction
- Method
- Experimental Data
- Results
- Conclusions



Conclusions

- WeBTC a text classification method that explores public biological web resources
- Improved precision (reaching 100%)
- Recall can be improved by covering a large range of web resources
- KDD Cup:
  - Better performance than standard classification systems
  - Effective alternative to automatically obtain the domain knowledge



Future Work

- Match the performance of state-of-art methods based on domain knowledge introduced manually.
- Explore other biological databases
- Use of ontologies
- Integrated in the ReBIL project:
  - http://xldb.fc.ul.pt/rebil/
- BioCreative 2004