Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup

Alexander S. Yeh, Lynette Hirschman, and Alexander A Morgan

Introduction

- The idea behind ‘challenge cup’ was to present teams with real or realistic training and test data to “create measurable forward progress in [the] field”
- They mined papers from the Flybase database

(FlyBase is a comprehensive database for information on the genetics and molecular biology of Drosophila. It includes data from the Drosophila Genome Projects and data curated from the literature. FlyBase is a joint project with the Berkeley Drosophila Genome Project.)

Methods: Contest Set-up

- Given a set of papers (full text) on genetics or molecular biology and, for each paper, a list of the genes mentioned in the paper
- Determine whether the paper meets FlyBase gene expression curation criteria, and for each gene, indicate whether the full paper has experimental evidence for gene products (mRNA and/or protein)

What then needed to Return

- A ranked list of the papers in order of probability of the need for curation, the presence of experimental evidence needs a higher ranking.

(curated: articles from the literature that have been reviewed by the curation staff at RGD who have read the article and extracted the specific information of interest to RGD which was subsequently loaded into the database.)

- A yes/no decision on whether on curate each paper

- The each gene in each paper an individual decision about whether the paper has evidence for gene products

Data Training

- Data consisted of 862 ‘cleaned’ full text papers
- Genes renamed to standard convention
- Matched to flybase standards

Results

Sub Task:	Best:	1st	Med	Low
Ranked List:	84%	81%	69%	35%
Y/N Paper:	78%	61%	58%	32%
Y/N Products:	67%	47%	35%	8%

Winning strategy

- Manually constructed rules that were matched against patterns deemed of ‘interest’

- All teams moved away “bag of words” approach common in test classification did more with domain experts

Lessons Learned

- PDF form not suitable for Processing, furthermore HTML had its own challenge. Too many linked file mapping.
- Many times to properly “mine” requires a significant biology understanding and understanding of flybase conventions

Lessons Learned

- The more hightec automated weighted techniques produced far less ‘correct’ answers than those programs written manually to the specific constraints of the task.
- It was important to know both what and ‘where’ to look for features and patterns.

Third sub task most difficult

- Associations and indicators varied for each gene. Different patterns.
- A way to combat the structure may be use more extensive linguistic structure indicators. Similarities and better relationship structures would help the system more with reliability

Points of Note

- Training of the test data in not practical in normal circumstances.
- Nor is requiring golden html.
- *Either transcripts or proteins* papers failure shows text mining over reliance on simple associations.
- What about the things that should be left in that are mined out?