TRUMIT: A Tool to Support Large-Scale Mining of Text Association Rules

Robert Neumayer
e-mail: neumayer@idi.ntnu.no

NTNU
Norwegian University of Science and Technology

George Tsatsaronis
Web: http://www.idi.ntnu.no/~gbt/
e-mail: george.tsatsaronis@biotec.tu-dresden.de

biotec
Biotechnology Center TU Dresden
TECHNISCHE UNIVERSITÄT DRESDEN
TRANSINSIGHT

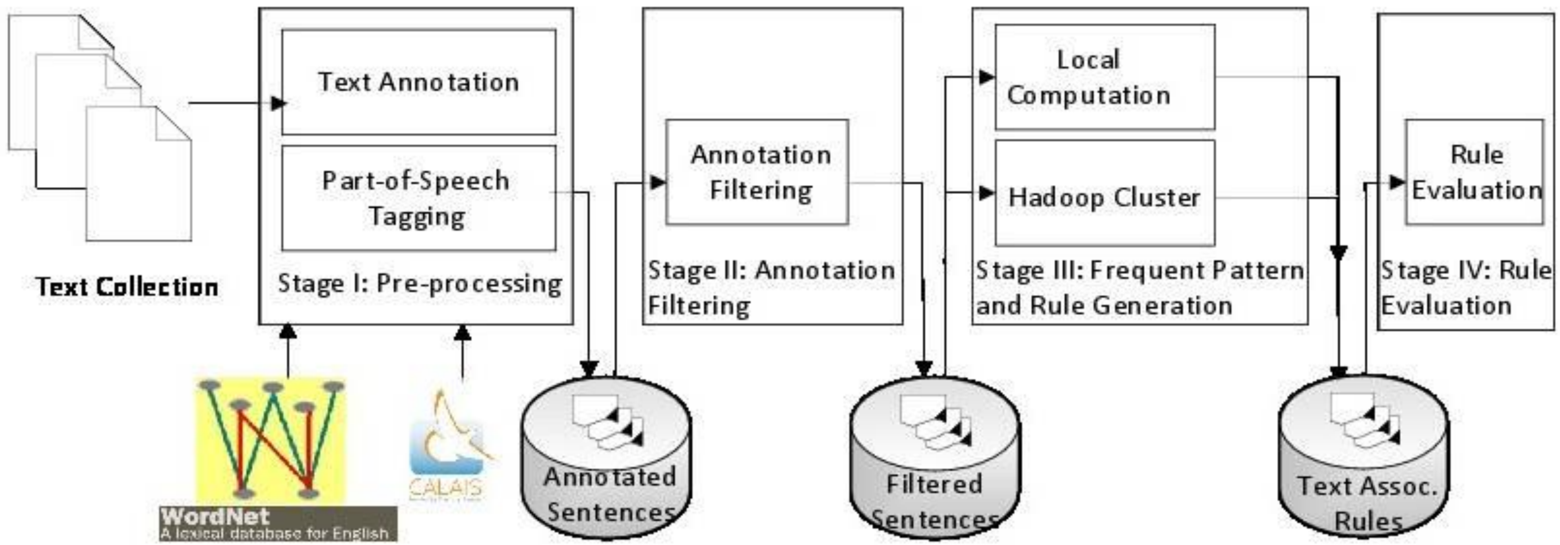
Kjetil Nørvåg
e-mail: noervaag@idi.ntnu.no

NTNU
Norwegian University of Science and Technology

System Description

- Efficient mining of association rules from text
- Term annotation using several difference annotators
- Association rule extraction between terms or categories of terms
- Novel unsupervised evaluation measures for weighting and ranking the importance of text rules
- Demonstration in two large text collections: ***Wikileaks*** and ***TREC-7***

Architecture



Methodology

- $D = \{d_1, d_2, \ldots, d_n\}$ the set of transactions, where D the set of all text sentences
- $A = \{a_1, a_2, \ldots, a_m\}$ the set of all annotations produced from applying all annotators of TRUMIT. Examples are: *Person, Company, Date, etc.*
- Examples of term rules: *Google => 1998, or Bill Gates => Microsoft*
- Examples of category rules: *Company => Date, Person => Company*
- Efficient implementation using: *fp-growth* on *Apache Mahout*

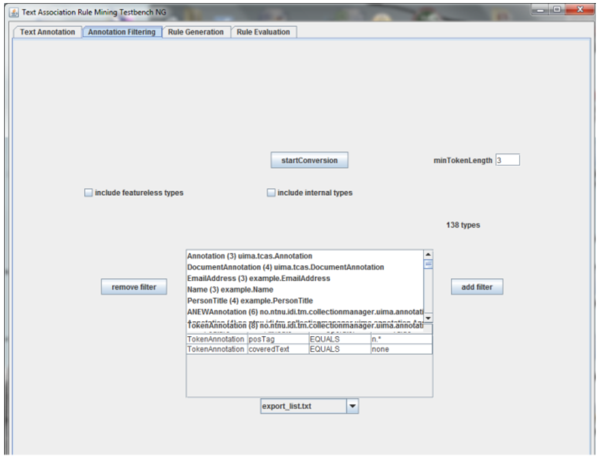
Annotators

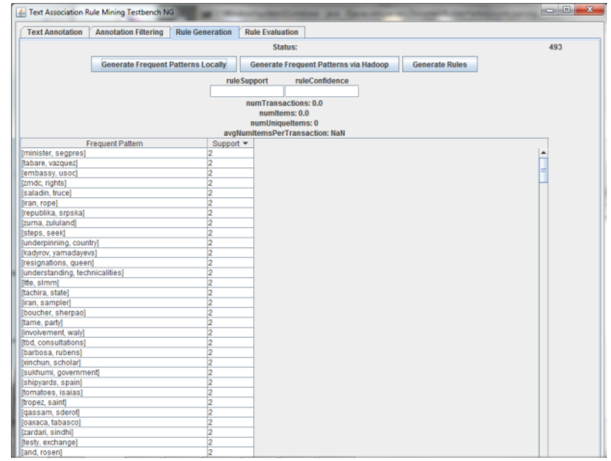
- Annotators' plugins through *Apache UIMA*

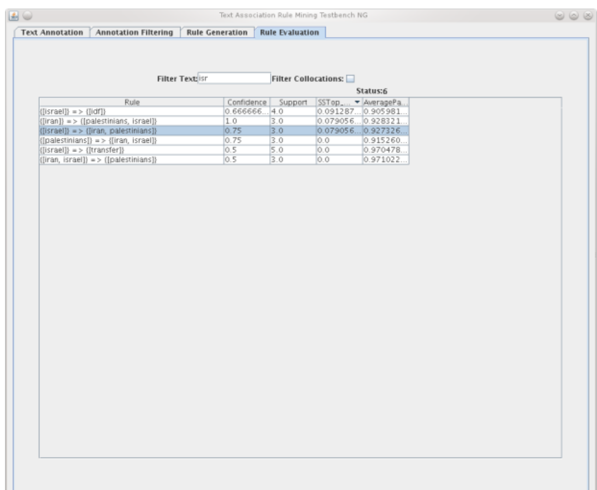
- Currently eight annotators are used:
  - Language annotator
  - Open Calais annotator
  - POS annotator
  - Stanford NER annotator
  - WordNet domain annotator
  - Maui keyword annotator
  - Lexical emotion annotator
  - Wikipedia miner annotator

- Any other annotator can be incorporated very easily

Demo







PKDD 2011, September 5-9, Athens, Greece. “TRUMIT: A Tool to Support Large-Scale Mining of Text Association Rules", Neumayer, Tsatsaronis, and Nørvåg.