Business from technology

Precision and recall in classifying scientific literature: comparing topic modeling to Kernel-based spectral clustering

GTM Conference 2013, Atlanta, USA

Arho Suominen[1], Stephen Carley[2], Hannes Toivanen[1] and Alan Porter[2,3]

[1]VTT Technical Research Centre of Finland

[2]Technology Policy and Assessment Center, Georgia Tech

[3]Search Technology, Inc.

Background

- The availability of machine learning methods that can be applied to text, such as Topic modeling (Blei & Lafferty 2009) and string kernels (Karatzoglou & Feinerer 2007), have shown promise as partitioning (dimension reduction) methods for textual information.
- Studies show that the text-based clustering methods can differentiate between document groups with high accuracy (Karatzoglou & Feinerer 2010; Wei X. & Croft 2006).
- The increase in algorithm development and subsequent software tools that enable the use of the algorithms have made using the practical.

Background

- Knowledge Discovery in Databases (KDD) process stages
  - Selection,
  - Pre-processing,
  - Transformation,
  - Data Mining (discovery of previously unknown properties of data) and
  - Interpretation/Evaluation
- Machine learning (prediction based on known properties of training data)
- String kernels in machine learning and data mining are kernel functions that operate with strings.

Our study

- We extend the earlier work of Yau et al., who worked on a sample of scientific publications which where partitioned with Topic modeling.
- We extend the study by using kernel based spectral clustering, which would enable us to analyze data without any feature extraction prior to analysis. (Lodhi et al. 2002)
  - In practice this would enable analysis without building a document-term-matrix co-occurrence matrix prior to analysis.
- We partition the same data and compare our results with earlier K-means and Topic modeling based results by Yau et al.

Data

- we generate a document collection from seven different scientific areas from the ISI Web of Science; MEMS, Solar Cells, Tissue engineering, RNAi or RNA inference, Graphene, Genetic Algorithm and stochastic or non-linear programming.
- We note that some of the areas are related with others and some are not.
  - For example, Genetic Algorithm correlates with stochastic programming; RNAi is correlated with Tissue Engineering; and Graphene is almost independent of all the others.

Thus we can look at the results with different conditions of relatedness.

- The sample was also limited by at least one author being affiliated with the Georgia Institute of Technology, or Emory University in the case of RNAi.
  - This was done to enable the easy use of expert opinion if needed

Data

- In practice, the data was gathered based on the seven different keywords representing the technologies appearing in the topics field, together with variants of the universities name in the author affiliations field.
- This resulted in a dataset of 1254 documents.

Category(Keyword)	Document #
MEMS	345
Solar cell or photovoltaic	178
Tissue Engineering	217
RNAi or RNA inference	127
Graphene	180
Genetic Algorithm	114
(stochastic or non-linear) programming	99

Source: ISI Web of Science

Pre-processing

- We excluded records with no abstract available reducing the set to 1206 records.
- We used several levels of pre-processing.
  1. No pre-processing – excluding Rights Reserved; ©, and years removed
  2. Using Abstract and titles, the Data that was pre-processed in R to lower case, remove punctuations, removing general stopwords, removing selected scientific stopwords, and by stemming
  3. *1) (i) Merging the Keywords (author's) and Keywords Plus fields (ii) Cleaning the merged field, Applying a Acronym Eliminator to this field, 2) Applying the Acronym Eliminator to the Abstract field, 3) A Chemical Compounds thesaurus (e.g. the chemical acronym 'C2H2' converts to the full word 'acetylene').*
  4. *Process in 2. with added processing in R with putting* to lower case, remove punctuations, removing general stopwords, removing selected scientific stopwords, and by stemming

Method

- The traditional approach to text classification is to map the document to a high dimensionality feature vector
  - Losing the word order and focusing on retaining the frequency of terms in the document.
  - Pre-processing includes the removal of non-informative words and replacing words with their stems.
- Lodhi et al. proposed, in 2002, an approach based on symbol sequences and the use of kernels
  - No need for domain knowledge
  - Document is considered as a long sequence
  - “The more substrings two documents have in common, the more similar they are considered” (Lodhi et al. 2002)

Method

- In this study, we used spectral clustering with the kernel based approach.
  - Spectral clustering embeds datapoints into a subspace of a normalized affinity matrix.
  - String kernel is used to define the affinities between documents.
- The spectral clustering approach we use is based on a algorithm by Ng et al. (2001).
- The function specc() in the R package *kernlab* was used to implement the study.

Methods

Calculated measures:

$$Precision = \frac{tp}{tp+fp}$$

$$Recall = \frac{tp}{tp+fn}$$

$$F-score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Results

Data	Average (St.Dev.)		
	Precision	Recall	F-score
K-means for a dtm (Yau et al.)	0,70 (0,35)	0,66 (0,30)	0,66 (0,30)
LDA for a dtm (Yau et al.)	0,92 (0,04)	0,88 (0,05)	0,90 (0,04)
Specc with kernel no pre-processing*	0,63 (0,17)	0,61 (0,10)	0,61 (0,09)
Specc with kernel pre-processing 1.	0,73 (0,27)	0,73 (0,20)	0,72 (0,22)
Specc with kernel pre-processing 2.	0,67 (0,24)	0,64 (0,16)	0,64 (0,18)
Specc with kernel pre-processing 3.	0,63 (0,27)	0,64 (0,16)	0,60 (0,18)

* Added to the presentation, wasn't included in the abstract

Discussion

- Increased pre-processing makes results worse.
  - With no pre-processing the results were worse than with a minimal processing approach.
  - However, Lodhi et al. (2002) argument on direct analysis through kernel based string clustering might need more work.

  → The impact of pre-processing is interesting. Could the minimal pre-processing approach create a practical point for analysis as it has taken out terms that are in almost every document while keeping variation in the whole corpus?
- Topic modeling by Hierarchical Dirichlet Process yields significantly better results than the string kernel specc.
- With a minimal pre-processing approach kernel based spectral clustering produced better results than a document-term-matrix based K-means analysis.

Discussion

- Noted based on review comments:
  - Analysis is tricky without building a training set.
    - → Yes, definitely tricky. The good results of Topic modeling and specc are even surprising.
  - What about 100 million documents
    - → Without parallel computing difficult, even though the new statistical tools (algorithms) available are faster.
    - → The approximately 1000 documents were easily analyzed, with a regular laptop, in minutes. Sets ranging about 10,000 will take several hours. Full-text will run months.
  - What about legal or scientific jargon?
    - → Minimal pre-processing yielded the best results.

Limitations and Future work

- Limitations:
  - The mock sample created is probably not the gold standard.
  - Not a practical approach for large sets without parallel computing.
  - Precision, recall and the F-score are only metrics – the actual partitioning results can be a better representation of the situation.
- Future work
  - Testing other approaches to cluster with kernels.
  - Validating the needed pre-processing scheme.
  - Extending the work through larger sets of data with parallel computing.

arho.suominen@vtt.fi

QUESTIONS?