Interactive Clustering of Text Collections According to a User-Specified Criterion

Ron Bekkerman, *University of Massachusetts, Amherst*

Hema Raghavan, *University of Massachusetts, Amherst*

James Allan, *University of Massachusetts, Amherst*

Koji Eguchi, *National Institute of Informatics, Tokyo*

Email: ronb@cs.umass.edu

Outline

- Problem statement and motivation
- Underlying technology
  - Combinatorial Markov Random Fields
- Interactive clustering algorithm
- Results and conclusions

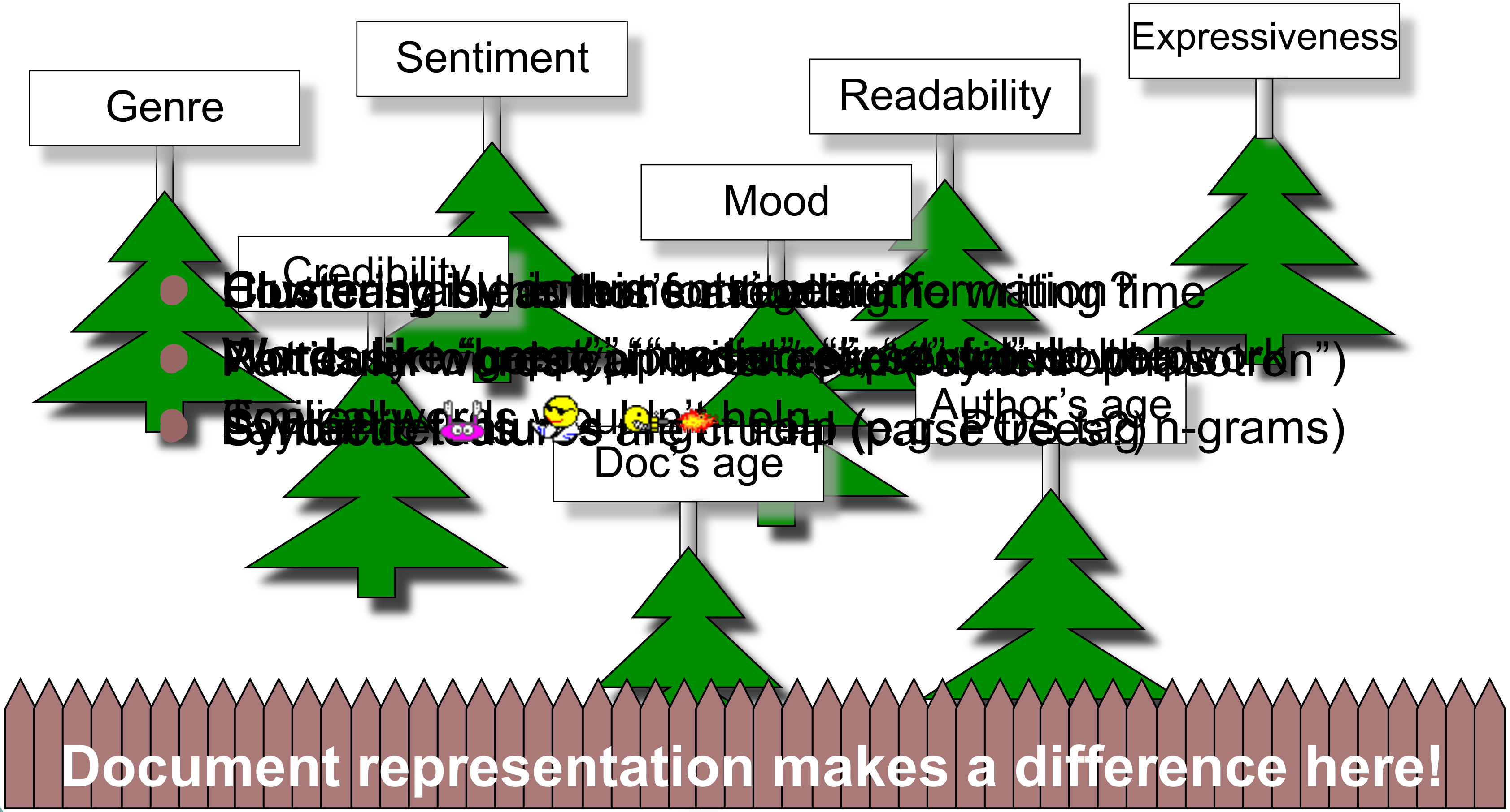
Data analysis

- Given a collection of documents
  - No labels
- Tell me *something* about its structure
  - What something? ☺
- Cluster it!
  - Group documents *by topic*
    - Where topic is some kind of word similarity

Topical clustering

- Example: two news titles:
  - “Seafood Benefits Outweigh Potential Risks”
  - “One Study Calls Fish a Lifesaver, Another Is More Cautious”
- Clustering won’t work here
  - The one based on BOW representation
- But what if you don’t care about the topic?

Non-topical clustering



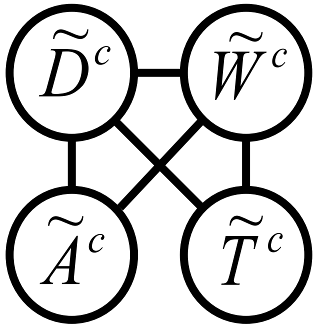
Solution: interactive clustering

- A useful tool for on-the-fly data analysis
- The user comes up with *feature types*
  - Which are *modalities* of the data
- The user comes up with *feature examples*
  - For the modalities where it is possible
- We apply our technology
  - For multi-modal clustering
  - While enriching feature lists

Combinatorial MRF (Comraf)

Bekkerman et al., ECML-2006

- Data modalities are represented with one random variable each
- Which are nodes in a Comraf graph ***G***
- Edges are interactions between the modalities
- Chef's special inference method:
  - Local combinatorial optimization for nodes
  - Iterative Conditional Mode for traversing ***G***

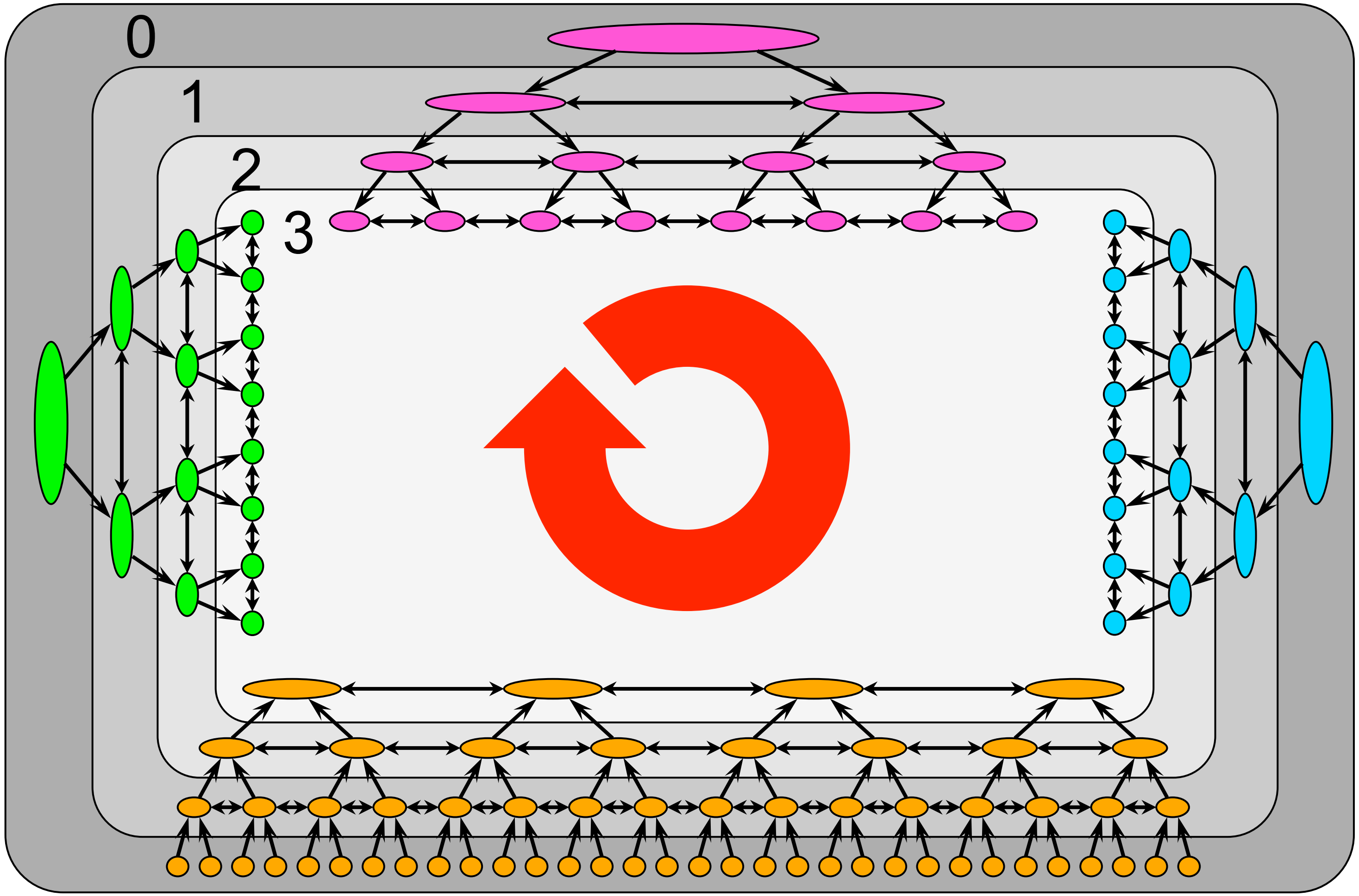


Multi-way distributional clustering

Bekkerman et al., ICML-2005

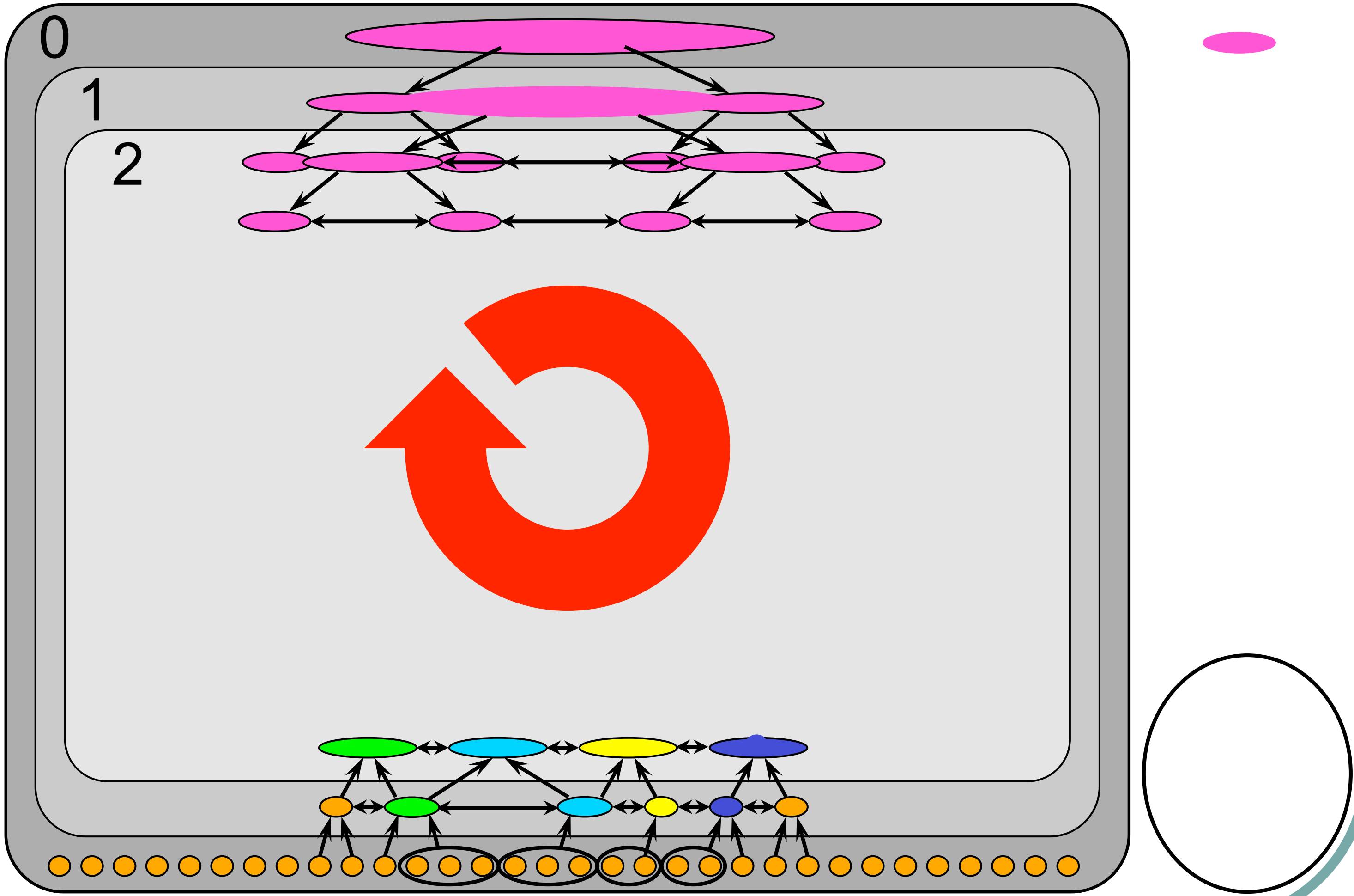
- A specific inference algorithm used in Comrafs for multi-modal clustering
  - *Multi-modal clustering*: simultaneously constructing N clusterings of N data modalities
- Combination of:
  - Top-down clustering for some modalities
  - Bottom-up clustering for the others
  - Local optimization at each iteration

Example: 4-way clustering



Example: interactive clustering

- User labeling
- Doc clustering
- Word clustering
- User correction
- Doc clustering
- Etc…



Clustering by genre

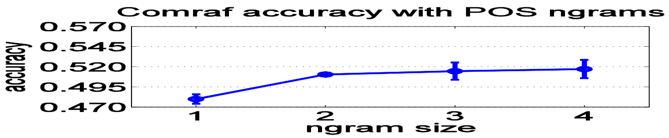
- Feature types:
  - BOW
  - POS tag n-grams (1- 2- 3- and 4-grams)
  - Both BOW and POS n-grams (2-grams)
- No feature examples can be given
- 2-way or 3-way distributional clustering used as it is

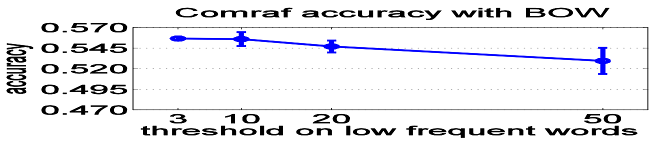
Genre: experimental setup

- Dataset:
  - British National Corpus
  - 21 categories (genres), 32 documents in each
  - Semi-manually POS-tagged (91 tags overall)
  - 64,000 unique words, 6000 POS 2-grams
- Baselines:
  - K-means (WEKA's implementation)
  - LDA (Xuerui Wang's implementation)

Genre: results

Doc representation	K-means	LDA	Comraf
BOW	9.1%	55.4±0.1%	55.7±0.2%
POS 2-grams	23.2%	44.7±0.2%	51.0±0.2%
BOW + POS 2-grams			**58.5±0.6%**





Clustering by sentiment

- Application of interactive clustering of movie reviews
  - *Harry Potter and the Goblet of Fire*
  - 1613 reviews downloaded from IMDB.com
  - With original user ratings
- Ratings are translated to 4 categories:
  - Strongly disliked, somewhat disliked, somewhat liked, strongly liked

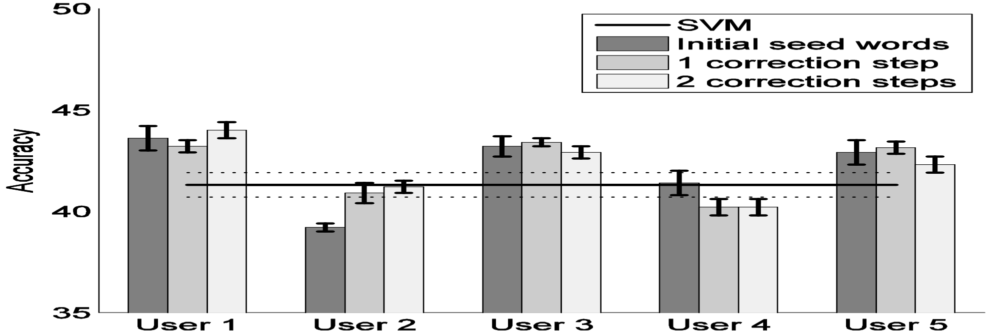
Sentiment: experimental setup

- 5 users were given a list of 563 words
  - Marked between 26 and 58 words from it
  - Revised word clusters after each iteration
- Oracle:
  - For each category C, 25 words were chosen
    - From a list of 4295 “sentimental” words
    - Their distribution over categories had a peak at C
- Baseline (besides k-means and LDA):
  - SVM trained on 22,546 reviews
    - To 46 popular Hollywood movies of 2005

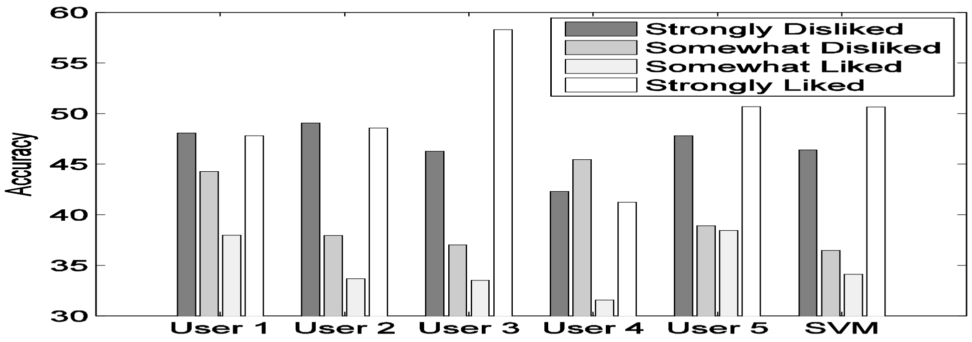
Sentiment: accuracy

Doc represent.	K-means	LDA	SVM	Comraf
BOW	28.2	37.0±0.2	39.1±0.3	40.3±0.8
Sentiment list	29.0	40.2±0.5	41.3±0.6	43.0±0.9
Interactive clustering (Oracle)				**47.1±0.2**

Sentiment: accuracy by user



Sentiment: accuracy by category



Conclusion

- One of the first works on non-topical clustering
- General approach proposed
  - Based on Comraf paradigm
- Scores are high for genre
  - Low for sentiment
    - Which shows how difficult the problem is
- Thank you!