Active Subtopic Detection in Multitopic Data

Benjamin Bergner   Georg Krempl

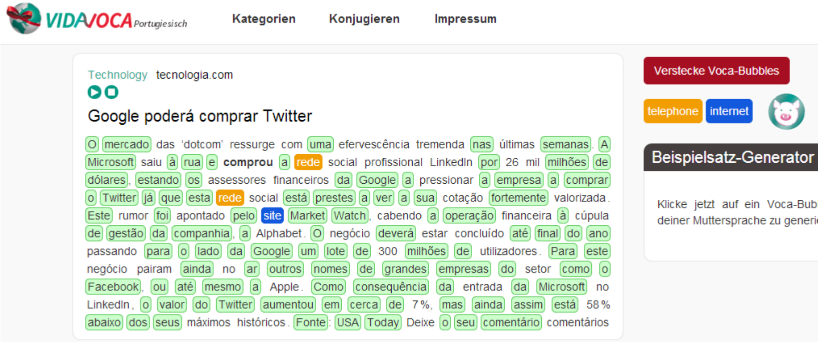
Knowledge Management & Discovery
Faculty of Computer Science
Otto-von-Guericke University
Magdeburg, Germany

Outline

- Introduction
- Related Work: Clustering by Intent (CBI)
- Enhancements: Multitopic CBI
- Evaluation
  - Aims
  - Setup
  - Results
- Summary & Outlook

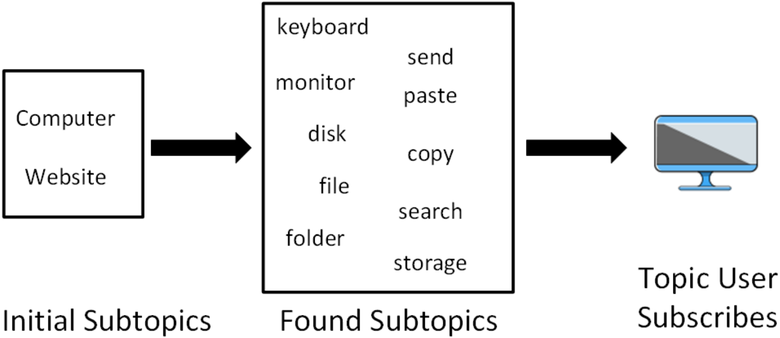
App: Vocabulary Creator for Language Learning

- User reads foreign language texts
- Words of interested topic are highlighted
- How to build comprehensive topic vocabulary sets?



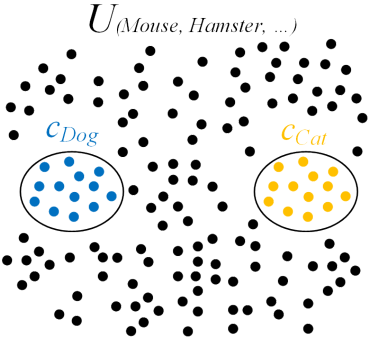
App: Vocabulary Creator for Language Learning

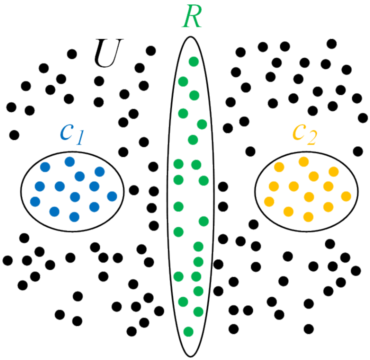
- Input:
  - set of documents
  - an **intention**, e.g. two words from same topic → they act as subtopics
- Desireable Output: More subtopics of same topic (and documents they occur in)
- How: Probabilistic Active Incremental Clustering



Related Work: Clustering by Intent (Forman et. al 2015)

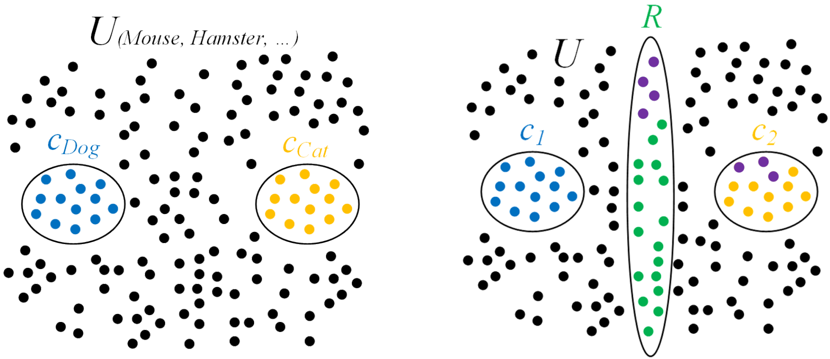
- Dots: represent documents
- $C1$,$C2$: documents that contain predefined subtopics
- Residual Set R: documents most unsure how to cluster
- Get feedback for words in R that best discriminate R from L
- Positive Feedback creates new subtopics and increases $|L|$, repeat
- Forman: Hewlett Packard analyzes support logs, customer surveys with CBI to find meaningful groups based on subtopics





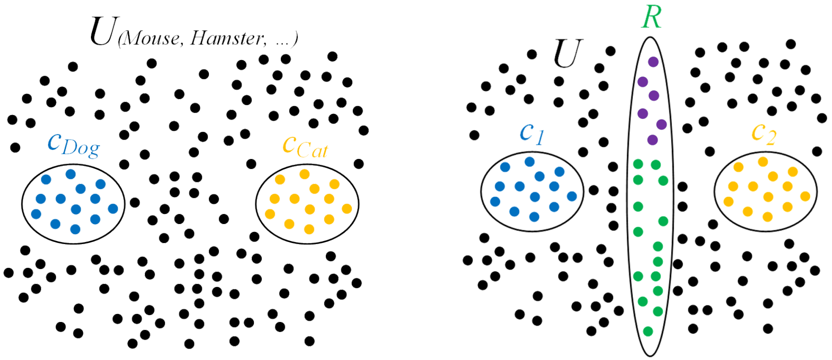
Related Work: Clustering by Intent (Forman et. al 2015)

- Dots: represent documents
- $C1, C2$: documents that contain predefined subtopics
- Residual Set R: documents most unsure how to cluster
- Get feedback for words in R that best discriminate R from L
- Positive Feedback creates new subtopics and increases $|L|$, repeat
- Forman: Hewlett Packard analyzes support logs, customer surveys with CBI to find meaningful groups based on subtopics



Related Work: Clustering by Intent (Forman et. al 2015)

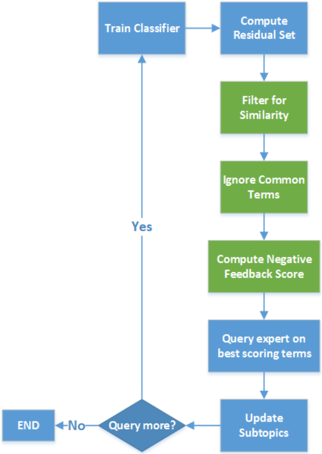
- Dots: represent documents
- $C1$,$C2$: documents that contain predefined subtopics
- Residual Set R: documents most unsure how to cluster
- Get feedback for words in R that best discriminate R from L
- Positive Feedback creates new subtopics and increases $|L|$, repeat
- Forman: Hewlett Packard analyzes support logs, customer surveys with CBI to find meaningful groups based on subtopics



Multitopic Clustering by Intent (MCBI)

Problem and Objective

- In multitopic datasets, also unrelated documents will be considered as residual
- This work aims to extend CBI to make it usable for multitopic datasets: **Multitopic Clustering by Intent (MCBI)**



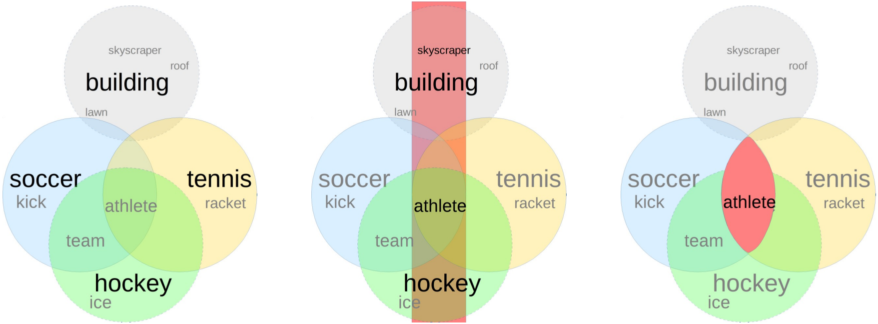
Contributions

- Similarity Set
- Improve usage of negative feedback
- Ignore common words

Remark: Concepts are explained per example, for formulae: see paper

MCBI Contributions: From Residual to Similartiy Set

- Subtopic's word bags with **intent: sports**
- initially given: soccer, tennis
- aim: find hockey, multi-topic environment: building documents
- residual set contains hockey AND building related documents
- word bags of soccer, tennis and hockey overlap when considering similarity



MCBI Contributions: Negative Feedback, Ignore common words

Negative Feedback

- Do not allow negative feedback terms to occur further times
- Candidates that have many relative cooccurences with negative feedback terms will be punished
- they are likely to be estimated negatively, too
- Ex.: Russia/hockey are competing given previously rejected candidate Finland, Russia and Finland occur more often together (also non-sports-relationships) than hockey and Finland (only sports-relationship)

Ignore common words

- Candidate Terms occuring often in L are topic specific, not subtopic specific
- e.g. athlete or stadium occur often in L
- We ignore them to
  - save annotation time
  - prevent them from being added to negative word list which would punish real subtopics

Evaluation: Aim, Setup

Objective

- Optimize interaction with expert
- i.e. find many actively chosen subtopics in few tries

General Setting

- Compare against **random** and **CBI** baseline
  (for comparison between CBI and clustering see Forman et. al 2015)
- Consider average number of positive feedback over iterations

Dataset

- Over **4000 wikipedia articles** drawn from most common nouns → heterogenous/multi-topic dataset
- Stop word removal, lemmatization, tf-idf → ca. **6500 unique words**
- Focusing on one closed category for testing: **countries**
- build a list of all countries, languages, denonyms for **auto-evaluation**

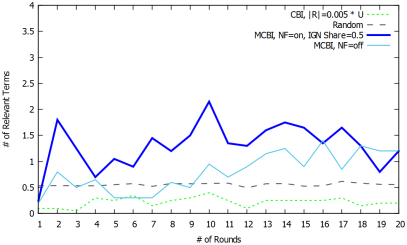
Evaluation Results

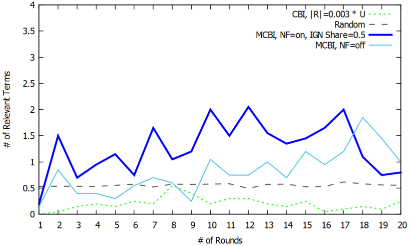
Iteration Setting # subtopics, NF, IGN	Average subtopics found Count.	Lang. & Denonym	Total
Random	**5.52**	**5.60**	**11.12**
2, CBI, R $= 0.005 \cdot \mid$ U $\mid$	**0.9**	**3.9**	**4.8**
4, CBI, R $= 0.003 \cdot \mid$ U $\mid$	**1.05**	**2.95**	**4.0**
2, *off*, –	7.75	8.20	15.95
2, *on*, 0	15.40	8.60	24.00
2, *on*, 0.1	15.55	8.9	24.45
2, *on*, 0.3	15.35	8.9	24.25
2, on, 0.5	**15.90**	**10.30**	**26.20**
2, *on*, 0.7	15.60	10.15	25.75
2, *on*, 1.0	10.15	7.05	17.20
4, *off*, –	9.00	7.10	16.10
4, *on*, 0	15.40	7.90	23.30
4, *on*, 0.1	15.90	8.00	23.90
4, *on*, 0.3	15.95	8.45	24.40
4, on, 0.5	**16.30**	**9.05**	**25.35**
4, *on*, 0.7	15.90	9.35	25.25
4, *on*, 1.0	9.00	6.90	15.90

Table: Detailed evaluation results for random, *CBI* and *MCBI*.

Evaluation Results

- Average number of positive feedback over
- 20 iterations with differing starting values, 20 rounds per iteration, 20 queries per round $\rightarrow$ 400 queries per iteration
- *MCBI* with *NF* = *on* and *IGNShare* = 0.5 performs almost always better than MCBI without those settings





Summary & Outlook

Summary

- MCBI extends CBI to multi-topic environments with residual sets within same topic $\rightarrow$ **similarity sets**
- Active feedback querying from user on candidates
- Incorporation of positive and **negative** user feedback
- **Ignoring** common words
- Evaluation on a wikipedia corpus for most common nouns
- Promising results compared to random and CBI

Outlook

- More extensive experimental evaluation needed
- Parameter tuning
- Tests for applications such recommender systems, information retrieval

Bibliography

- Forman, G.; Nachlieli, H. & Keshet, R. Bifet, A.; May, M.; Zadrozny, B.; Gavalda, R.; Pedreschi, D.; Bonchi, F.; Cardoso, J. & Spiliopoulou, M. (Eds.) Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III Clustering by Intent: A Semi-Supervised Method to Discover Relevant Clusters Incrementally Springer International Publishing, 2015, 20-36
- Jurafsky, Daniel, and James H. Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall.
- Settles, Burr. 2010. Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison

Thank you

Thank you for your attention

Interested? Ask Questions

Appendix: Further Applications

Content-based Recommender Systems

- Key terms in already visited pages act as subtopics
- Find new subtopics
- Check terms from unseen items
- Recommend items that have maximum of found subtopics
- Advantage: Find surprising results

Information Retrieval

- Recognize relations between documents
- Relate those that share many of found subtopics to same category

Appendix: Formulae

$$p_{s|u} = \mathcal{Pr}(s|\vec{u}) \propto \log \mathcal{Pr}(s) + \sum_{i=1}^{|V_u|} \log \mathcal{Pr}(u_i|s) \tag{1}$$

$$\mathcal{Pr}(s) = \frac{|L_s|}{|L|} \tag{2}$$

$$\mathcal{Pr}(u_i|s) = \frac{\sum_{u_i \in L_s} u_i + 1}{\sum_{v \in L_s} v_j + |V_L|} \tag{3}$$

$$uncertainty_u = p_{s|u} - p_{s'|u} \tag{4}$$

$$similarity_u = p_{s|u} + p_{s'|u} \tag{5}$$

$$rejscore_i \leftarrow \max_{n \in V_{REJ}} \left(\frac{|getDocumentsWithWords(U \cup L, \{w, n\})|^2}{|getDocumentsWithWords(U \cup L, n)|} \right) \tag{6}$$

$$discscore_i \leftarrow |getDocsWithWords(U, w)| - |getDocsWithWords(L, w)| \tag{7}$$