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

![Diagram of initial vs. found subtopics]

**Initial Subtopics**
- keyboard
- monitor
- disk
- file
- folder

**Found Subtopics**
- send
- paste
- copy
- search
- storage

**Topic User Subscribes**
Related Work: Clustering by Intent (Forman et. al 2015)

- Dots: represent documents
- \(C_1, C_2\): 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

\[
U_{(Mouse, Hamster, \ldots)}
\]

\[
C_{Dog}, C_{Cat}
\]

\[
R
\]

\[
C_1, C_2
\]
Related Work: Clustering by Intent (Forman et. al 2015)

- Dots: represent documents
- \( C_1, C_2 \): 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
- $C_1, C_2$: 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 Similarity 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 cooccurrences 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 occurring 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

<table>
<thead>
<tr>
<th>Iteration Setting</th>
<th>Average subtopics found</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count.</td>
</tr>
<tr>
<td># subtopics, NF, IGN</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>5.52</td>
</tr>
<tr>
<td>2, CBI, R = 0.005 · U</td>
<td>0.9</td>
</tr>
<tr>
<td>4, CBI, R = 0.003 · U</td>
<td>1.05</td>
</tr>
<tr>
<td>2, off, −</td>
<td>7.75</td>
</tr>
<tr>
<td>2, on, 0</td>
<td>15.40</td>
</tr>
<tr>
<td>2, on, 0.1</td>
<td>15.55</td>
</tr>
<tr>
<td>2, on, 0.3</td>
<td>15.35</td>
</tr>
<tr>
<td>2, on, 0.5</td>
<td>15.90</td>
</tr>
<tr>
<td>2, on, 0.7</td>
<td>15.60</td>
</tr>
<tr>
<td>2, on, 1.0</td>
<td>10.15</td>
</tr>
<tr>
<td>4, off, −</td>
<td>9.00</td>
</tr>
<tr>
<td>4, on, 0</td>
<td>15.40</td>
</tr>
<tr>
<td>4, on, 0.1</td>
<td>15.90</td>
</tr>
<tr>
<td>4, on, 0.3</td>
<td>15.95</td>
</tr>
<tr>
<td>4, on, 0.5</td>
<td>16.30</td>
</tr>
<tr>
<td>4, on, 0.7</td>
<td>15.90</td>
</tr>
<tr>
<td>4, on, 1.0</td>
<td>9.00</td>
</tr>
</tbody>
</table>

**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 → 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


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

\[ p_{s|u} = Pr(s|\vec{u}) \propto \log Pr(s) + \sum_{i=1}^{|V_u|} \log Pr(u_i|s) \quad (1) \]

\[ Pr(s) = \frac{|L_s|}{|L|} \quad (2) \]

\[ Pr(u_i|s) = \frac{\sum_{u_i \in L_s} u_i + 1}{\sum_{\forall v \in L_s} v_j + |V_L|} \quad (3) \]

uncertainty\(_u = p_{s|u} - p_{s'|u} \quad (4) \]

similarity\(_u = p_{s|u} + p_{s'|u} \quad (5) \]

\[ \text{rejscore}_i \leftarrow \max_{n \in V_{REJ}} \left( \frac{|\text{getDocumentsWithWords}(U \cup L, \{w, n\})|^2}{|\text{getDocumentsWithWords}(U \cup L, n)|} \right) \quad (6) \]

\[ \text{discscore}_i \leftarrow |\text{getDocsWithWords}(U, w)| - |\text{getDocsWithWords}(L, w)| \quad (7) \]