Improving Newsgroup Clustering by Filtering Author-Specific Words

Yuval Marom and Ingrid Zukerman
School of Computer Science and Software Engineering
Monash University

1 Project Aim

Aim: to use clustering to identify topics in electronic discussions and provide descriptions of these topics to interested users.

Target application: a help-desk system, but we have used newsgroups as a test-bed. Newsgroups are useful because
- they provide a good approximation to help-desk systems,
- they are readily available on the Internet in large quantities and diversity, and
- they obviate the need for manual tagging of topics, and thus enable automatic evaluation.

Problem: when people contribute frequently to a newsgroup, their idiosyncratic words dominate the clustering process, as shown in the example on the right.

cluster 1
saved
gif
transparent
advertisements
tom187@earthlink.net
crop
unsolicited
...

cluster 3
photography
light
convert
similar
nelson
colour
view
...

cluster 2
sketches
http://www.sover.net...
comic
smith
tony
demo
realistic
...

Example posting from "edjhann@hotmail.com":

```
To make the type visible make it white
put a Stroke on it from the Layer Styles
dialog. Or some variation of that.

--
Comic book sketches and artwork:
http://www.sover.net/~hannigan/edjh.html
```

Example posting from "vizrosplugins@yahoo.com":

```
The fastest software company is Borland. When I
called them to buy JBuilder 5, I was told the
current version was 6. But what I got from mail
is 7. A month later, I learned 9 was scheduled to
release.

Tony G. Smith
Vizros - Realistic 3D page curl plug-ins and more
Demo at http://www.vizros.com/gallery.html
```

Overall word-usage by "edjhann@hotmail.com" in 85 postings:

word	frequency	proportion
sketches	84	0.988*
http://www.sover.net...	84	0.988*
comic	84	0.988*
artwork	84	0.988*
layer	6	0.070
styles	4	0.047
...	...	...

* words with a significantly high proportion are filtered

2 Dominant Authors

In the example shown here, we are clustering email threads from the newsgroup **comp.graphics.app.photoshop**, using K-Means. The example shows three clusters and their most characteristic words.

The words that are most characteristic of cluster 2 appear in the signatures of two dominant contributors. That is, the clustering algorithm has created this cluster based on the authors, rather than the topics of discussion. The characteristic words are irrelevant to the topic of discussion.

Examples of postings made by these authors are shown below, as are the authors' highest overall word-usages in the newsgroup. The words that characterize cluster 2 are highlighted in the example postings.

Overall word-usage by "vizrosplugins@yahoo.com" in 47 postings:

word	frequency	proportion
tony	42	0.894 *
vizros	36	0.766 *
demo	35	0.745 *
smith	33	0.702 *
realistic	30	0.638 *
software	3	0.064
borland	1	0.021
...	...	...

* words with a significantly high proportion are filtered

3 Filtering Mechanism

We have built a filtering mechanism that removes undesirable influences of dominant authors in a newsgroup. The mechanism works as follows:

1. Build a "profile" for each person posting to the newsgroup. This profile is a distribution of *word posting frequencies* -- the number of postings where a word is used.
2. Consider each word in each posting:
   a) calculate a *word-usage proportion* (word posting frequency divided by the person's total number of postings)
   b) if the proportion is significantly higher than a threshold, filter the word from that posting.

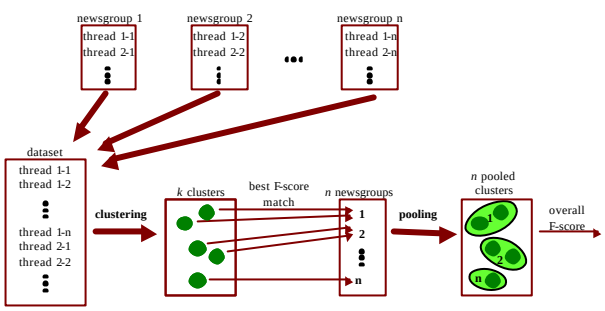
4 Evaluation

Objective: to examine the effect of the filtering mechanism on clustering performance, with respect to
- the topical similarity between the newsgroups, and
- the number of clusters.

Approach:
1. Merge threads from separate newsgroups into a single dataset.
2. Test the ability of the clustering mechanism to separate these threads back into the correct newsgroups.

Issues:
- Clustering is an unsupervised learning mechanism, therefore in order to evaluate clustering performance, we need to determine which clusters match which newsgroups.
- The number of clusters is not always equal to the number of newsgroups.

Solution:
1. Calculate the F-score (details in the box below) for each cluster-newsgroup match.
2. Choose the match with the best F-score.
3. Pool clusters that match the same newsgroup.
4. Calculate an overall F-score as a measure of how well the pooled clusters match the newsgroups.



F-score calculation:

Precision: $P_{ij} = \frac{|\text{cluster } i \cap \text{newsgroup } j|}{|\text{cluster } i|}$

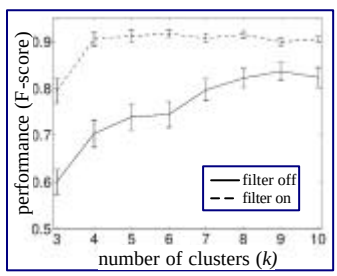
Recall: $R_{ij} = \frac{|\text{cluster } i \cap \text{newsgroup } j|}{|\text{newsgroup } j|}$

F-score: $\frac{2 P_{ij} R_{ij}}{P_{ij} + R_{ij}}$

Dataset 1

Newsgroups:
lp.hp
comp.text.tex
comp.graphics.apps.photoshop

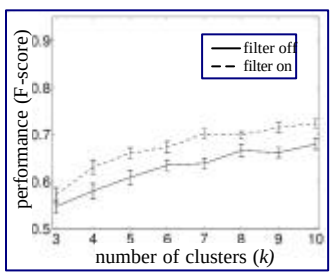
Results:
- Performance is much poorer without filtering, suggesting that author-specific words create undesirable overlaps between the newsgroups.
- These overlaps are resolved as the value of *k* increases, because more clusters enable the detection of finer differences between the threads.



Dataset 2

Newsgroups:
talk.politics.mideast
talk.politics.guns
talk.religion.misc

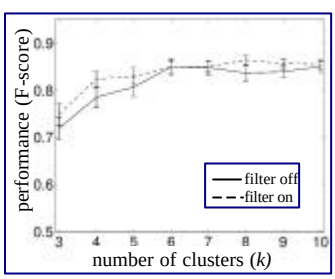
Results:
- These newsgroups discuss fairly similar topics, so there is a large topical overlap between the threads. Therefore, separating these newsgroups is difficult, yielding a poorer performance.
- However, filtering consistently improves performance, which means that there are also undesirable overlaps created by author-specific words.



Dataset 3

Newsgroups:
talk.politics.mideast
rec.sport.hockey
sci.space

Results:
- These newsgroups discuss very different topics, so the threads are different enough for the clustering to perform similarly well with and without filtering.
- Nonetheless, filtering has an effect for lower values of *k*, suggesting that some overlap is created by author-specific words.



5 Conclusion

- The results show that our filtering mechanism generally improves clustering performance, where the magnitude of its effect depends on the topical similarity between the newsgroups, and the number of clusters.
- We have experimented with newsgroups of varying degrees of topical similarity. The least related newsgroups provide a benchmark for clustering performance, while the more related ones exemplify our target help-desk application.
- Author-specific words used by dominant authors can have a detrimental discriminative influence on the clustering of newsgroup threads, and they can also lead to the extraction of uninformative and confusing topic descriptions. Thus, a filtering mechanism such as ours has both quantitative and qualitative benefits.

Acknowledgments

This research was supported in part by Linkage Grant LP0347470 from the Australian Research Council and by an endowment from Hewlett Packard.

MONASH University