Clustering Short Texts using Wikipedia

Somnath Banerjee
Hewlett-Packard Labs
Bangalore, India
somnath.banerjee@hp.com

Krishnan Ramanathan
Hewlett-Packard Labs
Bangalore, India
krishnan.ramanathan@hp.com

Ajay Gupta
Hewlett-Packard Labs
Bangalore, India
ajay.gupta@hp.com

ABSTRACT
Subscribers to the popular news or blog feeds (RSS/Atom) often face the problem of information overload as these feed sources usually deliver large number of items periodically. One solution to this problem could be clustering similar items in the feed reader to make the information more manageable for a user. Clustering items at the feed reader end is a challenging task as usually only a small part of the actual article is received through the feed. In this paper, we propose a method of improving the accuracy of clustering short texts by enriching their representation with additional features from Wikipedia. Empirical results indicate that this enriched representation of text items can substantially improve the clustering accuracy when compared to the conventional bag of words representation.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Clustering

General Terms
Algorithms, Experimentation

Keywords
Clustering, Wikipedia, Feed Reader

1. INTRODUCTION
In recent years, dynamic content like news and blog posts are being delivered in the form of RSS or Atom feeds that can be read using a feed reader. A feed reader periodically downloads the updated contents from the subscribed sources and provides an interface to the users to read and manage the feed items.

Although, in this paradigm, users actively subscribe to the preferred feed sources, there is still the problem of information overload. Many popular feed sources usually send a large number of items everyday. For example, Google News (http://news.google.com/) sends 350+ news items per day through its feed. Users often subscribe to multiple feed sources and that increases the problem.

One way to deal with the information overload problem is to cluster similar items. Often different feed sources send articles (items) on the same topic. This is especially true for the news domain where several different sources cover the same news stories. The same news source also often delivers multiple articles describing different aspects or developments of a story. By clustering similar items, a feed reader can provide a better interface to the users. It can also help in filtering duplicate or very similar items and make the information more manageable for a user. The Google News homepage is one example of such an interface that clusters the news articles belonging to the same topic and presents only one article for a topic. This enables the display of manageable amount of information in their news homepage. Interested users can view the other stories of a topic by following a link.

Feed sources usually send a few lines of text for a feed item along with a link to the full content of that item. Therefore, a feed reader usually has access to very short length of text for each feed item. A clustering method implemented within a feed reader thus should be able to work with only short texts. This makes the clustering task challenging. In this paper, we propose a method of improving the clustering accuracy using Wikipedia by enriching the representation of the short texts to be clustered. In our method, the conventional bag of words representation of text items is augmented with the titles of select Wikipedia articles. It has been observed previously that additional features from WordNet can improve clustering results [2]. Here we show how Wikipedia can be used as the additional feature source for text clustering, especially when very few lines of text are available for each item. To evaluate our method, we did a comparative study by representing a collection of short news articles using the bag of words method and using the proposed method. We then ran several different clustering algorithms using each representation. Results indicate that for most clustering algorithms, significantly higher accuracy is obtained with our proposed method of representation.

2. FEATURE GENERATION FROM WIKIPEDIA
Recently, Gabrilovich [1] has demonstrated the value of using Wikipedia as an additional source of features for text categorization and determining the semantic relatedness between texts. Here, we have used a similar technique to generate additional features of text items for clustering algorithms.

We downloaded[1] the English Wikipedia dump of November 26, 2006. After removing templates, articles describing Wikipedia features, and articles containing less than 50 non "stop words" we had 1,174,107 articles. We then created a Lucene (http://lucene.apache.org/) index of these Wikipedia articles.

Given the text of a feed item, we create two query strings from the text (more on this in next section). We then use the query strings to retrieve the top matching Wikipedia articles from the Lucene index. The titles of the retrieved Wikipedia articles now serve as additional features of the feed item for clustering. The titles of the Wikipedia articles are referred here as Wikipedia concepts.

[1] Available at http://download.wikimedia.org/enwiki/

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

3. EXPERIMENTAL SETUP

We used Google News to generate the labeled dataset. In our experience, we observed that the clustering algorithm that Google uses is quite accurate. Thus it provides an easily available labeled dataset for clustering. As mentioned earlier, Google News homepage contains short articles on different news topics and for each article there is a link pointing to the other articles on the same topic. Figure 1 shows one such article and the associated link (circled red) to the other articles of the same news topic. Visiting any such links pointing to the other articles will display a page that contains 30 news articles on the corresponding news topic. If there are more than 30 news articles for a topic then there will be a link to the 'Next' page that contains another 30 (or less) articles on the same topic.

Figure 1. A news clip from Google News homepage

We took a snapshot of the Google News homepage on February 16, 2007. That page had 26 articles on different news topics and so 26 links pointing to the other articles of the corresponding topics. We crawled those 26 links and the 'Next' page from each link and then extracted the news articles by parsing the crawled pages. This way we gathered 1557 news articles belonging to the 26 different topics (i.e., approximately 60 articles per topic). Each news article here consists of a title and one line of description. Note that the original 26 links of the Google News homepage served as cluster identifiers.

News articles of the above dataset were represented in two different ways and six different clustering algorithms were run using each of the representation. The first representation is the simple bag of words representation. That is, each article is represented by a vector of terms appearing in the article [3]. The weight of each term is the frequency of the term in the article. Since the title of a news article is more representative than the description, we got better results by giving greater importance to the title. In our experiments, we obtained best results by doubling the weights of the terms appearing in the title of a given article. We refer to this representation method as *Baseline*.

In the second method, the term frequency vector of the above method is augmented with selected Wikipedia concepts. For a given news article we used the title and the description as two separate query strings. Using each query we retrieved top 10 matching Wikipedia concepts from the Lucene index; i.e. we retrieved total 20 Wikipedia concepts. This list of 20 Wikipedia concepts was then represented by a vector where the weight of a concept is the frequency of the concept in the list. In this case also we observed that giving greater importance to the title of a news article is beneficial. We obtained better results by doubling the weights of the concepts retrieved by the title query string. A new representation of the given article was generated by augmenting this vector to the term frequency vector constructed by the first method. We refer to this representation method as *Wiki_Method*.

For our experiment we used the freely available clustering package SenseClusters[2]. SenseClusters uses CLUTO[3] as the clustering engine and it provides six different clustering algorithms; rb, rbr, direct, agglo, graph and bagglo (the details of these algorithms are available in the CLUTO manual). We ran all the six clustering algorithms by representing the news articles using the *Baseline* and *Wiki_Method*. We used the pk3 cluster stopping measure to automatically determine the number of clusters in the dataset. All other clustering parameters were left at defaults (SenseClusters documentation contains the details of the pk3 measure and other clustering parameters).

3.1 Results

Table 1 shows the clustering accuracy achieved by the *Baseline* and *Wiki_Method* with the different clustering algorithms. Except the case of 'rbr' clustering algorithm, the *Wiki_Method* achieved better accuracy than the *Baseline*. It also achieved the overall best accuracy of 89.56%.

Table 1 Clustering accuracy (in percentage)

	rb	rbr	direct	agglo	graph	bagglo
Baseline	63.20	79.38	67.05	22.03	81.62	23.57
Wiki_Method	85.42	63.65	82.66	83.88	**89.56**	43.67

4. CONCLUSIONS

We have proposed a method of improving the accuracy of clustering short text items using Wikipedia as an additional knowledge source. Our experiment shows that this method can substantially improve clustering accuracy. The results obtained here also corroborate the recent findings that world knowledge can help in the different information retrieval tasks.

Future work includes testing the method for the incremental clustering problem as that is the more realistic scenario for a feed reader. Also, as we observed that additional Wikipedia concepts did not help all the different clustering algorithms, more understanding is required on how and when these additional features should be used.

5. REFERENCES

[1] E. Gabrilovich. *Feature Generation for Textual Information Retrieval Using World Knowledge*. PhD Thesis, Department of Computer Science, Technion – Israel Institute of Technology, Haifa, Israel, 2006

[2] A. Hotho, S. Staab, and G. Stumme. *Ontologies Improve Text Document Clustering*, In the Proc of the Third IEEE International Conference on Data Mining (ICDM'03), Melbourne, Florida, USA, 2003

[3] G. Salton, editor. *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1988

[2] http://senseclusters.sourceforge.net/

[3] http://glaros.dtc.umn.edu/gkhome/views/cluto/