Document Clustering using Character N-grams: A Comparative Evaluation with Term-based and Word-based Clustering

Yingbo Miao, Vlado Kešelj, Evangelos Milios
Faculty of Computer Science, Dalhousie University
{ymiao, vlado, eem}@cs.dal.ca

ABSTRACT

We propose a novel method for document clustering using character N-grams. In the traditional vector-space model, the documents are represented as vectors, in which each dimension corresponds to a word. We propose a document representation based on the most frequent character N-grams, with window size of up to 10 characters. We derive a new distance measure, which produces uniformly better results when compared to the word-based and term-based methods. The result becomes more significant in the light of the robustness of the N-gram method with no language-dependent preprocessing. Experiments on the performance of a clustering algorithm on a variety of test document corpora demonstrate that the N-gram representation with n=3 outperforms both word and term representations. The comparison between word and term representations depends on the data set and the selected dimensionality.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Clustering, I.2.7 [Natural Language Processing]: Text analysis, I.5.3 [Clustering]: Similarity measures, I.5.4 [Applications]: Text processing

General Terms: Algorithms, Experimentation, Performance.

Keywords: Text clustering, text mining, N-gram text representation.

1. INTRODUCTION

Motivated by some recent positive results in using character N-grams in building author profiles and their use in automated authorship attribution [3], we explore the idea of using frequent character N-grams as vector features in document clustering. We built a *vector space* for the documents and found that the results are better than using words, especially for low dimensionality. We also experiment with multi-word terms as the vectors' features. Our hypothesis is based on the assumption that a multi-word term representation is a more compact representation of meaning in a domain than words, and therefore has the potential of providing better clustering performance at lower dimensionality. We perform automatic term extraction based on the *C/NC* value method [2] to retrieve the terms using a combination of linguistic and statistical criteria.

Since the document clustering we use (k-means) requires a *distance* or *dissimilarity* measure rather than similarity, we use the following scaled sine distance measure, where d_1 and d_2 are the document vectors using TFIDF weights:

$$distance(d_1, d_2) = \frac{\pi}{180} \cdot \sqrt{1 - cosine(d_1, d_2)^2}$$

However, without feature selection, we can get very high dimensionality, which is one of the major challenges of document clustering. Several approaches to dimensionality reduction are available; e.g., reduction to a set of the most frequent words or terms.

Character N-grams. A character (byte) N-gram is a substring of N characters (bytes) of a longer string [1]. The N-gram models are based on systematically collecting and counting the N-grams using a "sliding window" of length N over a document. Generally, character N-grams produce better results than byte N-grams. Moreover, dimensionality and sparsity with character N-gram representation are lower than with byte N-gram representation. Our experiments were performed on character Tri-grams.

2. DOCUMENT REPRESENTATION AND CLUSTERING

For each feature (N-grams, terms and words) two different feature selection approaches were experimented with. The first feature selection approach is specific to each feature type (document frequency for N-grams, C-value for terms and collection frequency for words), and is explained in this section. The second feature selection approach used is generic, i.e. the same for all three feature types, and consists of a term-frequency variance measure. Let f_j be the frequency of a feature (word/term/N-gram) t in document d_j, and n_0 be the total number of documents in the collection. The features are ranked by a quantity that is proportional to the term-frequency variance (Eq. 8 of [4]):

$$\sum_{j=1}^{n_0} f_j^2 - \frac{1}{n_0}\left[\sum_{j=1}^{n_0} f_j\right]^2 \qquad (1)$$

Experiments with both feature selection approaches have been carried out.

Document Clustering using N-grams. In our approach, we attempted to follow the idea proposed in [3]. We retrieve the most frequent N-grams and their frequencies from each document as its profile. To explore a wider range of distance measures, we use the distance measure

$$\sum_{1 \le j \le L} \frac{(v_1(j) - v_2(j))^2}{\left(\frac{(v_1(j)+v_2(j))}{2}\right)^{\alpha}} \qquad (2)$$

depending on parameter α. From our experimental results, we find that $\alpha = 1$ produces the best clustering quality. N-gram size has impact on clustering quality as well as vector dimensionality and

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

sparsity. Dimensionality and sparsity increase with increasing N-gram size. In other words, the larger the N-gram size, the smaller is the number of common N-grams among documents. The best clustering quality is given by Quad-gram representation or 5-gram. Tri-gram representation, however, produces comparable entropy and accuracy with a more practical dimensionality.

An equivalent operation to stop-word removal in the context of N-grams is ignoring the N-grams that appear in all or almost all documents. This may be regarded as analogous to removing N-grams with high document frequency (DF), and we performed those experiments. They showed that having an upper bound on N-gram DF does not improve performance, and if the bound is made tighter, the performance started to decrease. The N-gram profiles are significantly different from word vectors as they are much more dense: If the N-gram size is small (e.g., less then 5), then most N-grams have high document frequency. If the N-gram size is large, then most N-grams have lower document frequency, and removing ”stop” N-grams would not make significant changes. One of the advantages of the N-gram approach is independence of the specific language, while stop-word removal is language-specific.

Document Clustering using Terms. Terms are extracted automatically based on their *C Value* (a frequency-based weight that accounts for nested terms). In order to reduce the dimensionality, the n terms with highest *C Value* are chosen as the vector features. TFIDF is used as the weighting scheme and scaled sine is used as the distance measure. In this paper, K-means is the clustering method used.

Document Clustering using Words. In word based clustering, we remove the stop-words and apply Porter's stemming algorithm. We select the words that have the highest collection frequency (number of occurrences in the whole collection).

We apply the TFIDF weighting scheme to compute the elements of the vector representation of documents and the scaled sine distance measure between documents in the K-means clustering algorithm.

3. EXPERIMENTAL RESULTS

Evaluation Methodology. Two evaluation measures are used to evaluate the clustering quality: *entropy* and *accuracy*. Two data sets are used in this paper: Reuters-21578 and CISI-CRAN-MED. After removing from Reuters the articles having less or more than one topic, there are 8,654 documents left which belong to 65 topics. CISI-CRAN-MED data set has 3,893 documents, which belong to three topics.

N-gram Feature Selection Based on Document Frequency. We performed experiments on feature selection based on document frequency. Two values of document frequency, minimum document frequency df_{min} and maximum document frequency df_{max}, are changed in our experiments. For any N-gram to be included in the vector space definition, it should appear in at least df_{min} documents at most df_{max} documents. Experiments show the importance of setting a minimal document frequency bound df_{min} and even with small values of df_{min} we can achieve significant dimensionality reduction. Generally, with increasing df_{min}, the clustering quality becomes better at first. However, the quality becomes worse if we keep increasing the df_{min}.

Comparison of Word, Term, and N-gram Representations under feature-specific feature selection. We performed experiments on comparing clustering performance using Tri-gram, term and word representation with different dimensionalities, and using the two feature selection approaches. To get lower dimensionality for Tri-grams, firstly we chose the Tri-grams with maximum $df =$ (Total number of documents) / 2; secondly, we sorted Tri-grams by decreasing df and chose the first n Tri-grams. For term-based representation, we use *Filter 1*, *C-Value* and choose the n terms with highest *C-Value*. For word-based representation, we perform stop-word elimination and word stemming (Porter's stemming algorithm). The dimensionality was reduced by using the n most frequent words over the collection. The results are the means of 30 runs of the clustering algorithm to reduce the impact of the initial random centroids.

When compared to word and term representation, tri-gram representation produces the best clustering results on both Reuters and CISI-CRAN-MED data sets. A standard T-test confirms that clustering performance using Tri-grams is significantly better than using terms on both of the Reuters and CISI-CRAN-MED data sets. Moreover, term representation generally gives significantly better results than word representation on the Reuters data set. On the other hand, there are not significant differences between the results of term and word representation when dimensionality is 1,000 or 2,000 on the CISI-CRAN-MED data set. When the dimensionality is 3,000 or 4,000, we see that word representation produces better clustering quality than term representation on the CISI-CRAN-MED data set.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel approach for document clustering. Using a character N-gram-based representation gives best clustering performance with the lowest dimensionality. A possible justification for this result is that the N-gram method has less sparse-data problems, it finds common patterns between words with the same roots but different morphological forms (e.g., finance and financial), without treating them as equal, which happens with word stemming. It also detects phrasal patterns with N-grams that bridge two words. To reduce dimensionality with words, we have to ignore a large set of words, while frequent N-grams collect frequencies of multiple words, and hypothetically retain more information. In addition, we find that feature selection based on document frequency can be applied to document clustering with character N-gram representation. More detail about the experiments summarized in this short paper is available as a technical report [5].

Acknowledgments. The research presented here was funded in part by grants from the Natural Sciences and Engineering Research Council of Canada and IT Interactive Services, Inc.

5. REFERENCES

[1] W. B. Cavnar. Using an n-gram-based document representation with a vector processing retrieval model. In *TREC-3*, pages 269–278, 1994.

[2] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, August 2000.

[3] V. Kešelj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *PACLING'03*, pages 255–264, August 2003.

[4] J. Kogan, C. Nicholas, and V. Volkovich. Text mining with information-theoretical clustering. *Computing in Science and Engineering*, May 2003 (accepted).

[5] Y. Miao, V. Kešelj, and E. Milios. Document clustering using character n-grams: A comparative evaluation with term-based and word-based clustering. Technical report, http://www.cs.dal.ca/research/techreports/2005/, Faculty of Computer Science, Dalhousie University, 2005.