EXPLOITING N-GRAM IMPORTANCE AND WIKIPEDIA BASED ADDITIONAL KNOWLEDGE FOR IMPROVEMENTS IN GAAC BASED DOCUMENT CLUSTERING

Niraj Kumar, Venkata Vinay Babu Vemula, Kannan Srinathan
Center for Security, Theory and Algorithmic Research
International Institute of Information Technology, Gachibowli, Hyderabad, India
niraj_kumar@research.iiit.ac.in, vinaybabu.vv@gmail.com, srinathan@iiit.ac.in

Vasudeva Varma
Search and Information Extraction Lab
International Institute of Information Technology, Gachibowli, Hyderabad, India
vv@iiit.ac.in

Keywords: Document clustering, Group-average agglomerative clustering, Community detection, Similarity measure, N-gram, Wikipedia based additional knowledge.

Abstract: This paper provides a solution to the issue: "How can we use Wikipedia based concepts in document clustering with lesser human involvement, accompanied by effective improvements in result?" In the devised system, we propose a method to exploit the importance of N-grams in a document and use Wikipedia based additional knowledge for GAAC based document clustering. The importance of N-grams in a document depends on several features including, but not limited to: frequency, position of their occurrence in a sentence and the position of the sentence in which they occur, in the document. First, we introduce a new similarity measure, which takes the weighted N-gram importance into account, in the calculation of similarity measure while performing document clustering. As a result, the chances of topical similarity in clustering are improved. Second, we use Wikipedia as an additional knowledge base both, to remove noisy entries from the extracted N-grams and to reduce the information gap between N-grams that are conceptually-related, which do not have a match owing to differences in writing scheme or strategies. Our experimental results on the publicly available text dataset clearly show that our devised system has a significant improvement in performance over bag-of-words based state-of-the-art systems in this area.

1 INTRODUCTION

Document clustering is a knowledge discovery technique which categorizes the document set into meaningful groups. It plays a major role in Information Retrieval and Information Extraction, which requires efficient organization and summarization of a large volume of documents for quickly obtaining the desired information.

Since the past few years, the usage of additional knowledge resources like WordNet and Wikipedia, as external knowledge base has increased. Most of these techniques map the documents to Wikipedia, before applying traditional document clustering approaches (Tan et al., 2006; Kaufman and Rousseeuw, 1999).

(Huang et al., 2009) create a concept-based document representation by mapping terms and phrases within documents to their corresponding articles (or concepts) in Wikipedia, and a similarity measure, that evaluates the semantic relatedness between concept sets for two documents. (Huang et al., 2008) uses Wikipedia based concept-based representation and active learning. Enriching text representation with additional Wikipedia features can lead to more accurate clustering short texts. (Banerjee et al., 2007)

(Hu et al., 2009) addresses two issues: enriching document representation with Wikipedia concepts and category information for document clustering.

Next, it applies agglomerative and partitional clustering algorithm to cluster the documents.

All the techniques discussed above, use phrases (instead of bag-of-words based approach) and the frequency of their occurrence in the document, with additional support of Wikipedia based concept.

On the contrary, we believe that the importance of different N-grams in a document may vary depending on a number of factors including but not limited to: (1) frequency of N-grams in the document, (2) position of their occurrence in the sentence and (3) position of their sentence of occurrence.

We know that keyphrase extraction algorithms generally use such concepts, to extract terms which either have a good coverage of the document or are able to represent the document's theme. Using these observations as the basis, we have developed a new similarity measure, which exploits the weighted importance of N-grams common to two documents, with another similarity measure based on a simple measure of common phrase(s) or N-gram(s) between documents. This novel approach increases the chances of topical similarity in clustering process and thereby, increases the cluster purity and accuracy.

The first issue of concern to us was to use Wikipedia based concepts in document clustering, leading to effective improvements in result along with lesser human involvement in system execution. In order to accomplish this in our system, we concentrate on a highly unsupervised approach to utilize the Wikipedia based knowledge.

The second issue is to remove noisy entries, namely N-grams that are improperly structured with respect to Wikipedia text, from identified N-grams, which results in a refinement of the list of N-grams identified from every document. This requires the use of Wikipedia anchor texts and a simpler querying technique (see sec-2.4).

Another issue of high importance is to reduce the information gap between conceptually similar terms or N-grams. This can be accomplished by applying community detection approach (Clauset et al., 2004; Newman and Girvan, 2004) because this technique does not require the number and the size of the community as a user input parameter, and is also known for its usefulness in a large scale network. In our system, using this technique, we prepare the community of Wikipedia anchor texts. Next, we map every refined N-gram to the Wikipedia anchor text community resulting in a reduction in the information gap between N-grams, which are conceptually same but do not match physically. Therefore, this approach results in a more automatic mapping of N-grams to Wikipedia concepts.

Finally, we present a document vector based system, which uses GAAC (Group-average agglomerative clustering) scheme for document clustering that utilizes the facts discussed above.

2 PREPARING DOCUMENT VECTOR MODEL

In this phase, we prepare a document vector model for document clustering. Instead of using bag-of-words based approach, our scheme uses the index of the refined N-gram (N = 1, 2 and 3), their weighted importance in document (See Section 2.2) and the index of Wikipedia anchor text community (See Section 2.4), to represent every refined N-gram in the document (see Figure 1).

We then apply a simple statistical measure to calculate the relative weighted importance of every identified N-gram in each document (See Section 2.2). Following this, we create the community of Wikipedia anchor texts (See Section 2.3). As the final step, we refine the identified N-grams with the help of Wikipedia anchor text by using semantic relatedness measure and some heuristics, and map every refined N-gram with the index of Wikipedia anchor text community (See Section 2.4).

2.1 Identifying N-grams

Here, we used N-gram Filtration Scheme as the background technique (Kumar and Srinathan, 2008). In this technique, we first perform a simple pre-processing task, which removes stop-words, simple noisy entries and fixes the phrase(s) and sentence(s) boundaries. We then, prepare a dictionary of N-grams, using dictionary preparation steps of LZ-78 data compression algorithm. Finally, we apply a pattern filtration algorithm, to perform an early collection of higher length N-grams that satisfies the minimum frequency related criteria. Simultaneously, all occurrences of these N-grams in documents, are replaced with unique upper case alphanumeric combinations.

This replacement prevents the chances of unnecessary recounting of N-grams (that have a higher value of N) by other partially structured N-grams having a lower value of N. From each qualified term, we collect the following features:

F = frequency of N-gram in the document.

P = total weighted frequency in the document, referring to the position in a sentence, either the initial half or the final three-fourths of a sentence).

The weight value due to the position of N-gram, *D*, can be represented as:

$$D = (S - S_f) \quad \text{--- (1)}$$

where

S = total number of sentences in the document,

S_f = sentence number in which the given N-gram occurs first, in the document.

Finally, we apply the modified term weighting scheme to calculate the weight of every N-gram, W, in the document, which can be represented as:

$$W = \left(D + p + \log_2 \left(\left\{ \frac{(F+1)}{(F-p+1)} \right\} \right) \right) \times N \quad \text{--- (2)}$$

Where N= 1, 2 or 3 for 1-gram, 2-gram and 3-gram respectively.

2.2 Calculating Weighted Importance of Every Identified N-gram

The Weighted Importance of an identified N-gram, W_R, which is a simple statistical calculation that is used to identify the relative weighted importance of N-grams in document, can be calculated using:

$$W_R = \frac{(W - W_{avg})}{SD} \quad \text{---- (3)}$$

where

W_{avg} = Average weight of all N-grams (N = 1, 2 and 3) in the given document.

SD = Standard deviation of the weight of all N-grams in given document.

W = Calculated weight of N-gram (See Equation 2)

2.3 Using Wikipedia Anchor Texts

We use the collection of anchor texts in Wikipedia, to remove noisy N-grams and reduce the information gap between conceptually-related N-grams that do not match due to differences in writing schemes.

Each link in Wikipedia is associated with an anchor text, which can be regarded as a descriptor of its target article. They have great semantic value i.e. they provide alternative names, morphological variations and related phrases for target articles. They encode polysemy, because the same anchor text may link to different articles depending on the context in which it is found, requiring an additional technique to group similar concepts.

Wikipedia Anchor Text Community Detection: In the community detection scheme, we deploy the well-organized anchor text link structure of Wikipedia, which is the basis for the anchor text graph. In this scheme, we consider every anchor text as a graph node, saving time and calculation overhead, during the calculation of the shortest path for the graph of anchor texts, before applying the edge betweenness strategy (Clauset et al., 2004; Newman and Girvan, 2004) to calculate the anchor text community.

We used the faster version of community detection algorithm (Clauset et al., 2004) that has been optimized for large networks. This algorithm iteratively removes edges from the network and splits it into communities. Graph theoretic measure of edge betweenness is used to identify removed edges. Finally, we organize all the Wikipedia anchor text communities that contain Community_IDs and related Wikipedia anchor texts, for faster information access.

2.4 Refining N-grams and Adding Additional Knowledge

In this stage, we consider all N-grams (N = 1, 2 and 3) that have been identified using the N-gram filtration technique (Kumar and Srinathan, 2008) (See Section 2.1) and apply a simple refinement process.

2.4.1 Refinement of Identified N-grams

In this stage, we consider all identified N-grams (N = 1, 2 and 3) (See Section 2.1) and pass it as a query to our Wikipedia anchor text collection. We use Jaccard coefficient to achieve this. We consider any anchor text, B that yields the best match for N-gram, A, if the highest value of $J(A,B)$ occurs for this N-gram-anchor text pair and is greater than or equal to the threshold value, ∂.

$$J(A,B) = \begin{cases} (|A \cap B| / |A|) \geq \partial \\ (|A \cap B| / |B|) \geq \partial \end{cases} \quad \text{--- (4)}$$

where

$J(A,B)$ = Jaccard Coefficient between A and B

A = Identified N-gram

B = Wikipedia anchor text

$A \cap B$ = count of the words that are common to A and B.

In the entire evaluation, in order to get at least a meaningful match between A and B, we put ∂ = 0.6.

2.4.2 Proposed Document Vector Model

Our proposed document vector model contains details about three features of each refined N-gram: (1) the index of the refined N-gram, (2) its weighted importance in the document (See Section 2.2) and (3) the index of the anchor text community, as obtained above and written as community_ID. The addition of "community_ID" with each refined N-gram reduces the information gap that exists between different N-grams. our final document vector model can be represented as follows:

Our Proposed Document Vector Model

Document_ID:[N-gram_ID,Weight,Community_ID] ……..
Document_ID:[N-gram_ID,Weight,Community_ID] ……..

Note. This model contains three elements: the N-gram_ID, its weighted importance (calculated using Equation 3; see Section 2.2) and the community ID of anchor texts.

Figure 1: Our proposed Document vector model

3 CLUSTERING

GAAC based method has been widely studied for the following reasons: (1) it doesn't depend on the user's input parameter to set the number of clusters, (2) it reduces the chances of chaining effect, and (3) it is sensitive to noise.

In this section, we introduce a new similarity measure and use a GAAC (Group-average agglomerative clustering) based scheme to cluster the documents. In Section 3.1, we discuss the proposed similarity measure and in Section 3.2, we discuss the clustering scheme.

3.1 Proposed Similarity Measure

We introduce a new similarity measure, which includes both the weighted importance of an N-gram (N = 1, 2, and 3) in the document and the commonness of the identified N-grams. The weighted importance of an N-gram is the statistical measure of the relative weighted importance of the N-gram in the document (See Section 2.2). The commonness is a simple feature that uses common N-grams for the calculation of similarity measure. We then, calculate the final similarity measure, by calculating the harmonic mean of both, the weighted importance of an N-gram (N = 1, 2, and 3) in the document and the commonness of the identified N-grams.

Note. To check if an N-gram is common to two documents D_1 and D_2, it is sufficient to check the values of the community_Ids of the N-gram in D_1 and D_2 (see Figure 1). If they are same, then we consider that the given N-gram is common to both the documents, D_1 and D_2. This scheme is used in all the similarity measure processes.

3.1.1 Similarity Measure based on Weighted Importance of N-grams

This scheme uses weighted importance of N-grams in the document and considers only N-grams common to two documents, having a positive weighted importance, for similarity measure. The weighted importance based similarity, W_I between documents D_1 and D_2 can be calculated using:

$$W_I = \frac{(SD1 + SD2)}{(2 \times CN)} \quad \text{---- (5)}$$

where

SD1 = sum of positive weighted importance of all N-grams in document D_1, that are also common to D_2

SD2 = sum of positive weighted importance of all N-grams common to D1 and D2

CN = count of the total number of common N-grams between documents D_1 and D_2, that have a weighted importance greater than zero.

3.1.2 Similarity Measure based on Commonness of N-grams

Similarity measure score based on the commonness of N-grams, W_C, between documents D_1 and D_2 can be calculated as follows:

$$W_C = \frac{\#Common\_N-grams(D_1, D_2)}{Min(N-grams(D_1), N-grams(D_2))} \quad \text{--- (6)}$$

where

$\#Common\_N-grams(D_1, D_2)$ = the count of total number of common, unique N-grams between documents, D_1 and D_2.

$N-grams(D_1)$ = total number of distinct N-grams in the document, D_1

$N-grams(D_2)$ = total number of distinct N-grams in the document, D_2

"*Min*" = the smaller value among $N-grams(D_1)$ and $N-grams(D_2)$

3.1.3 Calculating Final Similarity Measure

To calculate the final similarity measure between documents D_1 and D_2, we calculate the H.M. (harmonic mean) of the two similarity measures discussed above, W_I and W_C, i.e.,

$$W = \frac{(2 \times W_I \times W_C)}{(W_I + W_C)} \quad \text{--- (7)}$$

Where 'W' = Final similarity measure for documents, D_1 and D_2.

3.2 Document Clustering Using GAAC

We use the Group-average agglomerative clustering (GAAC) algorithm for computing the clusters. In this scheme, GAAC initially treats each document as a separate singleton cluster, and then iteratively merge the most similar cluster-pairs in each step. Here, the similarity between two clusters is equal to the average similarity between their documents. To calculate the similarity between two documents, we use the similarity criteria defined in Section 3.1(See Equation 7 above). We continue merging, till the average similarity of clusters being merged is greater than or equal to "0.5".

4 EXPERIMENTS

Similar to (Hu et al., 2009; Huang et al., 2009; Huang et al., 2008), we have also employed the bag-of-words based system as our experimental benchmark. The details are given below:

Wikipedia Data: Periodic data dumps can be downloaded from http://download.wikipedia.org.

Dataset Used: A 20 Newsgroups dataset consisting of 2000 messages from 20 Usenet groups can be downloaded from http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.

Experimental Settings: After obtaining the data set, we prepare all possible combinations of 2 to 15 different topics, by merging the document sets of related topics together, and randomly select six combinations from each set. The main aim of this experiment is to test the behaviour of the system, for different cluster numbers.

Preparation of Experimental Corpus set: The format of the corpus_set-ID is **[C_K]**, where K is the number of document sets/classes. (See Table 3)

Baseline Algorithm Used for Comparison: We compared our devised system using BOW based approach, by using two different baseline techniques: (1) Partitional Clustering algorithms (Tan et al., 2006; Kaufman and Rousseeuw, 1999) like ***K-means*** and ***Bi-Secting K-means,*** and (2) ***GAAC*** (Tan et al., 2006). Since *Bi-Secting K-Means* and *K-Means* may produce different clustering results each time the experiment is conducted, due to random initialization, we calculate the average result, by repeating the experiment five times, on every dataset.

Evaluation Metrics: Cluster quality is evaluated using two metrics: Purity (Zhao and Karypis, 2001) and F-measure standard evaluation metrics (Steinbach et al., 2000). Standard deviation of F-measure score and cluster purity is also calculated, in order to measure the variation in results.

Results: In Table 3, we present: (1) cluster purity score and (2) F-measure score with standard deviation, using all the four systems. A bold font value is used to represent higher score. Since, *K-means* has a higher accuracy than *Bi-Secting K-means* and *GAAC*, when using a bag-of-words based approach, we compared our result with *K-means*. We also observed that *GAAC* (with bag-of-words based approach) is better in cluster purity score than the other two partitioning algorithms, *Bi-Secting K-means* and *K-means*.

From the results obtained with the 20-Newsgroup dataset (See Table 3), it is clear that:

- Our devised scheme performs better than the bag-of-words based approach, and shows an average improvement of 9%, in F-measure score over *K-means*, based on BOW approach.
- It also shows improvements of more than 10% in F-measure score over *K-means*, based on BOW approach, for corpus_set_ID: C_2, C_7, C_{13}, C_{14} and C_{15}.
- Our approach shows an average improvement of 16% in purity, over *K-means* based approach.

5 CONCLUSION

In this paper, we introduce a new similarity measure, based on the weighted importance of N-grams in documents, in addition to other similarity measures that are based on common N-grams in documents. This new approach improves the topical similarity in the cluster, which results in an improvement in purity and accuracy of clusters. We reduce the information gap between the identified N-grams and remove noisy N-grams, by exploiting Wikipedia anchor texts and their well-organized link structures, before applying a GAAC based clustering scheme and our similarity measure to cluster the documents. Our experimental results on the publicly available text dataset clearly show that our devised system performs significantly better than bag-of-words based state-of-the-art systems in this area.

Table 3: Cluster Purity and F-score value for 20-Newsgroup Dataset

Corpus Set-ID	Bi-Secting K-means (BOW)		K-means (BOW)		GAAC (BOW)		Our Approach	
	Purity (SD)	F-score (SD)	Purity (SD)	F-score (SD)	Purity (SD)	F-score (SD)	Purity (SD)	F-score (SD)
C_2	0.65 (0.05)	59.07 (1.27)	0.63 (0.15)	61.01 (1.91)	**0.79** (0.05)	58.02 (1.79)	0.78 (0.04)	**68.76** (1.04)
C_3	0.53 (0.06)	63.213 (2.82)	0.52 (0.09)	63.11 (1.25)	0.68 (0.06)	56.89 (3.79)	**0.71** (0.06)	**66.08** (1.11)
C_4	0.60 (0.17)	63.125 (2.01)	0.54 (0.16)	65.00 (1.94)	0.56 (0.04)	55.25 (5.00)	**0.73** (0.08)	**68.97** (1.51)
C_5	0.62 (0.06)	48.01 (4.37)	0.50 (0.11)	55.00 (5.80)	0.58 (0.05)	53.00 (3.32)	**0.62** (0.09)	**58.31** (1.27)
C_6	0.52 (0.10)	51.25 (1.25)	0.53 (0.07)	49.167 (3.33)	0.57 (0.09)	45.08 (3.86)	**0.65** (0.06)	**55.83** (2.01)
C_7	0.55 (0.08)	49.642 (1.89)	0.56 (0.07)	54.762 (5.13)	0.63 (0.14)	48.00 (3.03)	**0.70** (0.05)	**59.79** (2.53)
C_8	0.56 (0.11)	45..02 (4.23)	0.54 (0.10)	53.437 (2.06)	0.58 (0.10)	47.17 (2.06)	**0.65** (0.05)	**57.18** (2.07)
C_9	0.56 (0.07)	43.888 (2.13)	0.54 (0.05)	52.222 (1.69)	0.57 (0.05)	41.01 (2.47)	**0.63** (0.06)	**56.76** (1.41)
C_{10}	0.53 (0.05)	44.341 (5.26)	0.57 (0.08)	53.51 (2.46)	0.56 (0.07)	41.73 (3.11)	**0.59** (0.07)	**57.89** (2.43)
C_{11}	0.51 (0.17)	45.43 (4.01)	0.55 (0.04)	51.818 (3.19)	0.53 (0.06)	45.00 (2.90)	**0.54** (0.08)	**55.23** (1.73)
C_{12}	**0.51 (**0.08)	45.00 (3.91)	0.50 (0.08)	50.00 (2.69)	0.52 (0.12)	44.17 (4.51)	**0.51** (0.06)	**54.93 (2.33)**
C_{13}	0.50 (0.09)	46.87 (5.00)	0.49 (0.12)	49.01 (3.09)	**0.52** (0.06)	45.00 (3.31)	**0.52** (0.09)	**55.91** (2.79)
C_{14}	0.51 (0.16)	47.00 (5.13)	0.51 (0.20)	51.071 (4.12)	0.51 (0.17)	46.09 (3.00)	**0.55** (0.09)	**56.19** (2.13)
C_{15}	0.50 (0.10)	41..666 (4.73)	0.49 (0.21)	49.667 (4.00)	0.48 (0.13)	43.34 (3.91)	**0.52** (0.10)	**55.91** (1.79)
S_{avg}	0.55 (0.10)	50.57 (3.43)	0.53 (0.11)	54.20 (3.05)	0.58 (0.09)	47.84 (3.29)	**0.62** (0.07)	**59.12** (1.87)

Note. Purity refers to *Cluster Purity*; F-score refers to *F-measure Score* (See Section 4.3.3); SD = Standard Deviation; S_{avg} = Average Score

REFERENCES

Banerjee, S., Ramanathan, K., Gupta, A., 2007. Clustering Short Texts using Wikipedia; *SIGIR'07*, July 23–27, Amsterdam, The Netherlands.

Clauset, A., Newman, M., Moore, C., 2004. Finding community structure in verylarge networks. *Physical Review* E, 70:066111, 2004.

Hammouda, K., Matute, D., Kamel, M., 2005. CorePhrase: Keyphrase Extraction for Document Clustering; In *IAPR: 4th International Conference on Machine Learning and Data Mining*.

Han, J., Kim, T., Choi, J., 2007. Web Document Clustering by Using Automatic Keyphrase Extraction; Proceedings of the *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*.

Hu, X., Zhang, X., Lu, C., Park, E., Zhou, X., 2009. Exploiting Wikipedia as External Knowledge for Document Clustering; *KDD'09*.

Huang, A., Milne, D., Frank, E., Witten, I. 2008. Clustering Documents with Active Learning Using Wikipedia. *ICDM 2008*.

Huang, A., Milne, D., Frank, E., Witten, I., 2009. Clustering documents using a wikipedia-based concept representation. In Proc *13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Kaufman, L., and Rousseeuw, P., 1999. Finding Groups in data: *An introduction to cluster analysis*, 1999, John Wiley & Sons.

Kumar, N., Srinathan, K., 2008. Automatic Keyphrase Extraction from Scientific Documents Using N-gram Filtration Technique. In the Proceedings of *ACM DocEng*.

Newman,M., Girvan,M., 2004. Finding and evaluating community structure in networks. *Physical review* E, 69:026113, 2004.

Steinbach, M., Karypis, G., and Kumar, V., 2000. A Comparison of document clustering techniques. Technical Report. Department of Computer Science and Engineering,University of Minnesota.

Tan,P., Steinbach,M.,Kumar,V., 2006. *Introduction to Data Mining*; Addison-Wesley; ISBN-10: 0321321367.

Zhao, Y., Karypis, G., 2001. Criterion functions for document clustering: experiments and analysis, Technical Report. Department of Computer Science, University of Minnesota.