Evaluating Better Document Representation in Clustering with Varying Complexity

Stephen Bradshaw and Colm O'Riordan
National University of Ireland, Galway, Ireland

Keywords: Clustering and Classification Methods, Mining Text and Semi-structured Data, Context Discovery.

Abstract: Micro blogging has become a very popular activity and the posts made by users can be a valuable source of information. Classifying this content accurately can be a challenging task due to the fact that comments are typically short in nature and on their own may lack context. Reddit[a] is a very popular microblogging site whose popularity has seen a huge and consistent increase over the years. In this paper we propose using alternative but related Reddit threads to build language models that can be used to disambiguate intend mean of terms in a post. A related thread is one which is similar in content, often consisting of the same frequently occurring terms or phrases. We posit that threads of a similar nature use similar language and that the identification of related threads can be used as a source to add context to a post, enabling more accurate classification. In this paper, graphs are used to model the frequency and co-occurrence of terms. The terms of a document are mapped to nodes, and the co-occurrence of two terms are recorded as edge weights. To show the robustness of our approach, we compare the performance in using related Reddit threads to the use of an external ontology; Wordnet. We apply a number of evaluation metrics to the clusters created and show that in every instance, the use of alternative threads to improve document representations is better than the use of Wordnet or standard augmented vector models. We apply this approach to increasingly harder environments to test the robustness of our approach. A tougher environment is one where the classifying algorithm has more than two categories to choose from when selecting the appropriate class.

1 INTRODUCTION

Clustering is the act of grouping items within a dataset into related subsets to gain some insight about the dataset. It can be done as a classifying technique, or as a preprocessing step in conjunction with additional approaches. It is an unsupervised approach that aims to identify inherent patterns in the data and uses those to select a cluster group for each item in the dataset. There are many clustering techniques all with their own strengths and weaknesses. There is no panacea for clustering method that works for all information needs.

Problems that exist with the use of K-means to documents lies in the so called *curse of dimensionality* (CoD). When designing a clustering approach every word in the corpus must be taken into consideration. This leads to a large sparse dataset. In addition, the polysemic nature of words can affect the efficacy of the clustering algorithm. Furthermore in this project the application of clustering is particularly challenging because of the source material. Instead of clustering standard documents, comments are selected for categorising. Comments offer a greater challenge to an automated clustering agent as they are often smaller and less structured than standard documents (Hu and Liu, 2012) (Zamir et al., 1997) (Zheng et al., 2009).

The literature on clustering reflects a realisation that standard approaches to clustering by themselves are suboptimal approaches (Zamir et al., 1997). This is due to two salient issues a) the failure of past approaches to incorporate the connectedness of terms in their analysis and b) issues surrounding the CoD. Techniques to offset these issues include Part of Speech Tagging (POS) (Zheng et al., 2009), Latent Semantic Indexing (Song et al., 2009), the application of nonnegative matrix factorization (Shahnaz et al., 2006) and pointwise mutual information (PMI) (Levy and Goldberg, 2014) to name a few. In short, more recent approaches have focused on better document representation to improve clustering quality. In this paper we outline an approach that uses graphs that attempts to capture the intended meaning of text. These graphs are used to improve the representation of a document. We show that this leads to more efficient clustering of the documents.

[a] *www.Reddit.com*

Bradshaw, S. and O'Riordan, C.
Evaluating Better Document Representation in Clustering with Varying Complexity.
DOI: 10.5220/0006930901940202
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 194-202
ISBN: 978-989-758-330-8

The remainder of this paper is outlined as follows: section 2 talks about related work, the use of external ontologies to boost document representation, clustering, social media as a source as well as a number of other papers that used alternative representational approaches for their document sets. Section 3 outlines the methodology; the source of the data used and the reasons for selecting the various threads for clustering. It talks about how documents can be represented as a vector of terms and how these vectors can be altered to improve clustering efficacy. Section 4 details the result of our experiments, different clustering evaluation approaches are outlined and discussed. Finally section 5 will detail our conclusions and future work.

2 RELATED WORK

There are many clustering approaches that achieve different aims. Examples of clustering approaches are hierarchical clustering and partitioning clustering. Additionally *two step clustering* combines elements of the two. Hierarchal clustering is the act of breaking and joining clusters. It can be done in a top down manner (*divisive*) or a bottom up manner (*agglomerative*). In divisive clustering, one big cluster is created which is further divided into smaller and smaller clusters. This is an inverse approach to agglomerative clustering which initially contains many small clusters that are joined together based on similarity. The final result for both approaches is a tree like structure, which contains the large amalgamated cluster at the top, and a series of smaller clusters at the bottom.

The clustering algorithm has several ways in which it can determine how to divide or join a cluster. Examples of this include single linkage, complete linkage, average linkage and centroid (Sarstedt and Mooi, 2014). The purpose behind these approaches is to establish a good metric for separating items into their respective clusters. Linkage in this instance is the connection between two items in two clusters. Single linkage identifies the nearest neighbours of two clusters. Complete linkage focuses on the distance between the points. Average linkages aims to create a distance between the mean of the items within two clusters, while the centroid approach finds the mean of each cluster and determines the appropriate cluster by measuring the difference from each target point with the mean of each cluster.

The selection of a clustering approach and a separation procedure has a huge bearing on the resulting clusters. Single linkage tends to form one large cluster with the outliers forming smaller additional clusters. Complete linkage is more affected by outliers so it will tend to produce smaller more compact clusters. Centroid and average linkage will produce smaller clusters with a low within-cluster variance (Sarstedt and Mooi, 2014).

Partition clustering differs in approach and outcome to hierarchical clustering. K-means is an example of the most prominent partition clustering approach. Rather than focusing on the distance between clusters, it examines the points within a cluster and endeavours to reduce the variance found within each cluster. It is an iterative approach, that assigns an item to a cluster based on the members that are already contained there. Initially, assignment is randomly selected, but subsequently it calculates the mean of each cluster and uses that to inform the allocation of each subsequent item. Clusters formed in this fashion tend to be more equal in size, so it is important to choose the correct number of required clusters at the beginning to avoid the creation of spurious clusters. For this project, partition clustering was selected as:

- it is computationally less expensive
- we know before the experiment how many clusters are required
- this approach best suits the dataset as the aim is to produce harmonious clusters that are wholly comprised of each individual subreddit
- this approach is less susceptible to the outliers that are inherent in sparse datasets
- Its simple algorithm means that this approach can be applied to large datasets that are represented by sparse vectors

This paper posits that issues associated with CoD can be offset by disambiguating the intended meaning of a term. To achieve this we map the common co-occurrence of terms as they occur in a domain specific environment. While many other approaches have leveraged external sources for improved document representation, typically their sources are not dynamically constructed from active online content. A common source of external information is Wordnet; however there are a number of drawbacks in using this source as it is created by experts and therefore static in nature. Additional drawbacks to using a static source are, the coverage and size may be limited. It is not a domain specific resource the content can quickly become outdated and it tends only to support major languages. Muller et al (Müller and Gurevych, 2009) conducted a study where they compared the similarity in documents using 3 external sources: Wi-

kipedia [1], Wiktionary [2] and Wordnet (Wallace, 2007). Taking the TREC dataset of 2003 they compared the syntactical similarity of the query with the pre-labeled relevant result-set. In 35.5% of the cases the required documents contained multiple query terms, meaning that there are many documents that do not even contain a strong direct syntactical overlap. They showed that through the use of external sources they could improve the coverage of the queries, however Wordnet was found to be less accurate than dynamically created sources such as Wikipedia.

The use of Wordnet for identifying semantic relatedness in text focuses on using synonyms, hypernyms, antonymns etc. Additional semantic factors that have been considered are noun, adjective relations. Zheng et al identify noun phrases in the text and classify them as concepts. They use these concepts to gain a greater understanding of the document, by increasing the importance placed on them through use of a weight which is increased with every re-occurrence (Zheng et al., 2009). The outcome of using graphs to map common co-occurring terms, means that this proposed approach will also capture noun phrase, but will not limit itself to using these when identifying important common terms in a post.

An alternative approach for representing documents can be found in the work of Cai et al, who represent the documents as a matrix and represent it as a second order tensor. This results in the document properties being stored in a more compact format, allowing for processing (Cai et al., 2006).

The strength of the approach proposed in this paper is that it identifies words that co-occur frequently and uses them to disambiguate the intended use of words in the target dataset. The work of Nagarajan et al bears a lot of similarity in approach and intent. As part of their preprocessing steps they use statistical analysis to identify the import terms. Stopwords and words deemed non important are removed and the documents are clustered around the terms deemed important. The advantage to removing non essential terms is that they reduce the complexity involved in computing clusters. They apply a agglomerative clustering approach and correctly identify 90 percent of the corpus. The is an example of K-medoids as specific datapoints are selected as centroids, rather than the hypothetical mean points (Nagarajan and Aruna, 2016).

Hammouda et al use the clustering approach as a preprocessing stage to group items similar in nature. They subsequently identify common phrases in the clusters and use these to label the dataset (Hammouda et al., 2005).

This paper empirically shows that while Wordnet can be used to improve upon document representation, it pales in camparison to dynamically created content. YAGO is another external ontology which is built on the structure of Wordnet; however it additionally incorporates the knowledge found on Wikipedia, giving it substantially more scope. It exploits the hypernym category of Wordnet to inform connections in its facts (Fabian et al., 2007). The paper by (Baralis et al., 2013) demonstrate how this ontology can be used to indicate to an automated agent which sentences are most salient, which they applied to multi document summarisation tasks. Strapparava et al extended the functionality of it by identifying and labelling terms that have emotive connotations (Strapparava et al., 2004).

In a detailed analysis on the state of the art of clustering Shah (Shah and Mahajan, 2012) list a number of factors that result in creating good clusters. These factors are; representation, reduction of dimensionality, reducing the rigidity of cluster definitions so that topics can be represented in two or more clusters, appropriate labelling of clustering for subsequent use, a good estimation of required number of clusters, stability and the use of semantics to properly encapsulate the intended meaning from the text.

Starstedt et al apply a *two-step clustering* method to their dataset, which is comprised of market information. Two-step clustering combines the strengths of hierarchical and partition clusters. The first step employs partitioning elements. The dataset is partitioned into clusters, each one allocated to a leaf on a tree. The second step employs hierarchical methods which order the clusters by importance. As the authors are analysing the factors that affect marketing they can use this ordering step to determine which variables have the highest impact on the resulting clusters. The choice of clustering algorithm differs from the one used in this project as the desired outcome was different. The goal of that project was to identify the impact of a finite number of indicative features, while this projected attempted to identify common themes in a sparse dataset which highlight whether a document is a member of a particular cluster (Sarstedt and Mooi, 2014).

Li et al. (Li et al., 2008) compare the performance of K-means and Dbscan on a storm dataset. The aim of the project was to accurately group instances of storms form their attributes. DBscan showed itself to be more fit to the task as a result of the process it uses to create clusters. While Kmeans created clusters based on the mean value of the attributes, DBscan focused on the density of items around the attributes,

[1] https://www.wikipedia.org
[2] https://www.wiktionary.org

thus producing more accurate cluster results.

In addition to dealing with standard issues related to clustering text, this paper engages the problems associated with dealing with text which is derived from social media. Data from this source tends to be more difficult to process as, it does not contain the same structure as traditional data sources, it can often be short in length and devoid of context. Additionally it is often time sensitive (Hu and Liu, 2012). This last issue is both an advantage and disadvantage when dealing with text. Take for example the case where someone wishes to query a recent event such as the earthquake in Mexico (2017). Prior to the specific event, the terms earthquake and Mexico would not have a link between them, so standard static knowledge sources will not be useful in providing context to the user interested in getting more information on the topic. However following the specfic event, there is a flurry of social media reports on the incidents and it generates conversation threads. Thus the dynamic nature of using Reddit as a source for constructing our external knowledge means that there will be associations built between these two entities. Identifying recent events is known as *event detection* and there have been a number of papers that deal with this using *Twitter* as a datasource (Atefeh and Khreich, 2015) (Sakaki et al., 2010) (Weng and Lee, 2011).

To the best of our knowledge our approach is the first to leverage Reddit as a source of information for extracting semantic relatedness in terms. Much of the work conducted using Reddit has focused on the actions of the users of the site (Gilbert, 2013) (Bergstrom, 2011) (Duggan and Smith, 2013) (Singer et al., 2014) (Potts and Harrison, 2013). A notable exception is the work done by Weninger who have done preliminary work on modeling the threads found in the subreddits. Taking a snapshot of the submissions on the top 25 most popular subreddits over a 24 hour period, they apply Hierarchal Latent Dirichlet Allocation to the conversation threads. They determine that, regardless of the length of a conversation thread, child posts tend to bear a resemblance to the initial parent comment and conclude that this is an indication of a conversational hierarchy. They propose that future use can be made of these topic words with regards to web searches and labeling document clustering (Weninger et al., 2013).

2.1 Methodology

In the following section we outline the steps taken in setting up the experiments. It includes a discussion on how the text is converted to a vector space model and will go into detail on the various ways that the

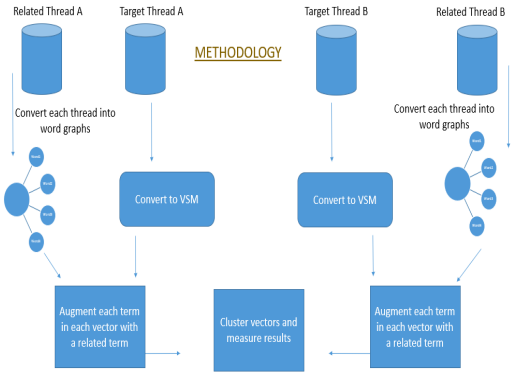


Figure 1: Sample of Methodology Steps Taken.

text was augmented. For the purpose of maintaining a point of comparison, included in this work are the results of applying the clustering algorithm on an unaltered dataset which we will refer to as the standard approach.

2.2 Modeling Thread Language

Initially 10 threads were chosen at random from Reddit. For each thread in our experiment, we modelled the language as follows.

1. The top thousand submissions from each thread are retrieved. These 1000 threads are selected from the subcategory *hot*, which is a subcategory that displays the trending posts in a particular subreddit. The strength of using posts found in this category is that they have been upvoted by users of the site, thus we can conclude that they are indicative of the theme found in a given subreddit.
2. Threads are made up of a parent comment and responses or child comments. These are concatenated so each document is representative of a parent thread and all of its child comments. If a comment did not contain a child comment, it is omitted from the corpus. The reasoning for this is that a thread that does not generate a response is not deemed interesting by the subredditors, so can be unrelated or poorly phrased.
3. We remove stop words and punctuation, and convert the documents to a vector space model representation. Additionally a TF-IDF weighting scheme is applied to the vectors.
4. Any vector less than 20 words is omitted and from the remaining subset, 1000 vectors were chosen. Some threads do not contain 1000 posts, in which instance every post greater than 20 words in length is selected.

5. Graphs are constructed from content found in a related thread. To represent the semantic relatedness in the text, we adopt a sliding window of two over the vectors. Each term was assigned a node and each co-occurrence is recorded by incrementing the weight between the term and target term. Co-occurrence is deemed if a term falls within a radius of two terms from the target term. The persistent co-occurrence of two terms indicated that these terms are somehow related and can be used to infer a relationship, pertinent to that particular topic.

2.3 Selecting Related Threads

We selected the related Reddit threads based on the nature of the content. So if the initial thread is about rugby then we selected National Rugby League (NRL) as the related thread, python with learnPython etc. Table 1 shows a list of the original thread paired with the corresponding thread used. To verify the relatedness of the threads, similarity was ascertained through the application of cosine similarity on the text of the thread, with the text found in the alternative thread. We found that the selected threads exhibited a high similarity with their related threads. It should be noted that cosine similarity only measures the common presence, absence and frequency of text. It does not reflect any level of the connectedness in terms. For our purpose however, it is sufficient as a rule of thumb measure of accuracy to confirm our intuition.

Table 1: Table List of Threads Used.

Thread	Related Thread
Rugbyunion	NRL
LearnPython	Python
Movies	Fullmoviesonyoutube
Music	popheads
England	London
Quadcopters	Quadcopter
Worldnews	News
Politics	ukpolitics
Boardgames	Risk
Ireland	Dublin

2.4 Creating Cluster Groupings

To test the robustness of our approach we created three sets of thread groupings. Group3 contains five instances of three threads. For the subsequent two groups we added one thread to the groupings. So Group4 contains five sets of four threads and Group5 contains five groupings of five threads. Increasing the number of clusters should increase the difficulty level for the classifier. Groups were selected through a random process and there are instances where a thread appeared in more than one classifier problem. To achieve this we created an array of all of the proposed threads. We generated a random number between one and ten (the number of selected threads). That number corresponded with the index position of a thread. If the thread had not already been included in the grouping we added it, otherwise we continued to generate a random number until each grouping was full. Table 2 contains a full listing on the threads contained in each group.

Table 2: Selected Groupings.

Group3 – Size 3
[Ireland, Rugbyunion, learnPython]
[Boardgames, Quadcopter, Politics]
[Boardgames, Quadcopter, Ireland]
[learnPython, Politics, Worldnews]
[Boardgames, learnPython, Worldnews]
Group4 – Size 4
[Music, Politics, England, Boardgames]
[Music, Ireland, Politics, England]
[Music, Worldnews, learnPython, Ireland]
[Ireland, Quadcopter, England, Boardgames]
[learnPython, Politics, Ireland, Worldnews]
Group5 – Size 5
[England, learnPython, Rugbyunion, Politics, Quadcopter]
[England, Movies, Music, learnPython, Boardgames]
[Ireland, England, Politics, Boardgames, Rugbyunion]
[Ireland, England, Quadcopter, learnPython, Politics]
[Boardgames, England, Politics, Ireland, Quadcopter]

2.5 Graph Augmentation

Augmenting the target thread with information found in the related thread was undertaken in the following way. For each term found in the target thread, we queried the related graph to determine its representation there. If the term is present, we select the highest weighted correlate. This term was then added to the original vector of terms. So while our original document was represented like this $d =< t_1, t_2, ...t_n >$ our augment thread retained the initial terms, but in addition we added one correlate per term. Thus augmented documents can be represented as $d =< (t_1, tr_1), (t_2, tr_2), ..(t_n, tr_n) >$, where tr_i indicates an addition of a correlate taken from the related thread graph.

2.6 Hypernym Augmentation

The process for augmenting threads with hypernyms is similar to the steps taken in creating the graph aug-

mentation. However instead of querying the graph for a related term, we used Wordnet. Thus the hypernym thread representations appear as follows $d =< (t_1, ht_1), (t_2, ht_2), ...(t_n, ht_n) >$ where ht_i represents the addition of a hypernym of the target term. In many instances there are a number of candidate terms, here we utilise the natural ordering of Wordnet, which ranks terms by most common use. Thus the first term in the list is more likely to be the intended hypernym.

Having conducted the previous steps we clustered the resulting thread representations with K-means.

3 RESULTS

3.1 Metrics

The evaluation of clustering performances is a field of study in itself. One must select an appropriate evaluation metric for the required task, given the clustering approach used and the desired outcome expected. There is no one clustering evaluation method that captures all aspects of a set of clusters (Meilă, 2007) (Milligan, 1996), (Kleinberg, 2003). For this work, four clustering metrics were selected, they were: inertia, homogeneity, majority-representation and adjusted Rand Index. These four metrics were selected as they best reflect the intended outcome of the clustering. We aim to make compact clusters that are reflective of each input thread. Both majority-representation and homogeneity are indicators of the singleness of the predicted clusters, while inertia reflects the concentration of the cluster points. The adjusted rand index is one of the most popular clustering evaluation metrics (Steinley, 2004) (Santos and Embrechts, 2009), that uses inherent and external factors to create a score reflecting the level of quality of a cluster.

3.2 Majority-representation

Majority-representation captures the level of the correct division of categories. We define correct as the allocation of our prelabelled documents into individual categories. We populate a table with all of the predicted labels. The rows represent the dispersion of the class over the clusters; and the columns represent the allocation of that class to that cluster. A class must have the highest row value (highest incident of itself) and be the highest column value (most representative of that cluster).

We assign a designation of one to each categorical cluster that contains a majority representation of one category in that cluster. This metric is useful because it allows us to clearly determine which assigned label matches our predetermined classes. It has been noted that one drawback of K-means is that it does not provide cluster labels; this metric will enable us to clearly determine the most representative cluster for a class.

	a	b	c	d	e	totals
0	1.0	1.0	1.0	n/a	n/a	3.0
1	1.0	1.0	1.0	n/a	n/a	3.0
2	1.0	1.0	1.0	n/a	n/a	3.0
3	1.0	1.0	1.0	n/a	n/a	3.0
4	1.0	1.0	1.0	n/a	n/a	3.0
5	1.0	1.0	1.0	1	n/a	4.0
6	1.0	1.0	1.0	1	n/a	4.0
7	1.0	1.0	1.0	1	n/a	4.0
8	1.0	1.0	1.0	1	n/a	4.0
9	1.0	1.0	1.0	1	n/a	4.0
10	1.0	1.0	1.0	1	1	5.0
11	1.0	1.0	1.0	1	1	5.0
12	1.0	1.0	1.0	1	1	5.0
13	1.0	1.0	1.0	1	1	5.0
14	1.0	1.0	1.0	1	1	5.0
total	15.0	15.0	15.0	10	5	60.0

Table 4: Threads correctly categorised through graph approach

	a	b	c	d	e	totals
0	1.0	0.0	1.0	n/a	n/a	2.0
1	1.0	1.0	0.0	n/a	n/a	2.0
2	1.0	1.0	0.0	n/a	n/a	2.0
3	0.0	1.0	1.0	n/a	n/a	2.0
4	1.0	1.0	0.0	n/a	n/a	2.0
5	0.0	0.0	1.0	1	n/a	2.0
6	0.0	0.0	0.0	1	n/a	1.0
7	1.0	0.0	1.0	0	n/a	2.0
8	0.0	1.0	0.0	1	n/a	2.0
9	0.0	0.0	1.0	0	n/a	1.0
10	0.0	1.0	0.0	1	0	2.0
11	0.0	1.0	1.0	1	1	4.0
12	0.0	1.0	0.0	1	0	2.0
13	0.0	0.0	1.0	0	0	1.0
14	1.0	0.0	0.0	0	1	2.0
total	6.0	8.0	7.0	6	2	29.0

Table 3: Results of threads accuracy standard approach

	a	b	c	d	e	totals
0	1.0	1.0	1.0	n/a	n/a	3.0
1	1.0	1.0	0.0	n/a	n/a	2.0
2	1.0	1.0	0.0	n/a	n/a	2.0
3	0.0	1.0	1.0	n/a	n/a	2.0
4	1.0	1.0	1.0	n/a	n/a	3.0
5	1.0	0.0	1.0	1	n/a	3.0
6	1.0	0.0	0.0	1	n/a	2.0
7	1.0	1.0	1.0	1	n/a	4.0
8	0.0	1.0	0.0	1	n/a	2.0
9	1.0	0.0	1.0	0	n/a	2.0
10	1.0	1.0	0.0	0	0	2.0
11	1.0	1.0	1.0	1	1	5.0
12	0.0	0.0	1.0	1	1	3.0
13	0.0	1.0	0.0	1	0	2.0
14	1.0	0.0	1.0	0	1	3.0
total	11.0	10.0	9.0	7	3	40.0

Table 5: Results from categorised through hypernym approach

Figure 2: Tables show whether a category was correctly clustered.

We can see from Tables 3 - 5, that the standard approach performs the worst, only dominating in under half of the categories while also being the strongest representation of itself. The hypernym approach achieves second best results in this area, as it showed itself to be correctly separated in two out of every three instances. The graph approach achieved best results as it always made a distinct cluster consisting of distinct allocation of each member of its corpus.

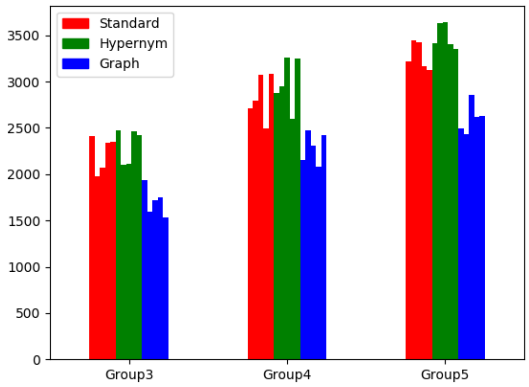


Figure 3: Combined Inertia Scores.

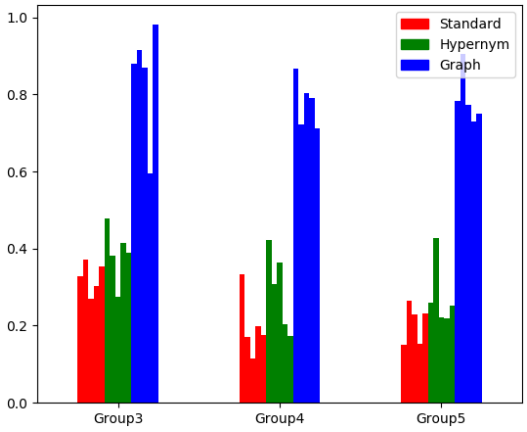


Figure 4: Combined Homogeneity Scores.

3.3 Inertia

Inertia is a standard evaluation technique for any clustering approach. It reflects the sum of the distance of each point in a cluster to the centroid of that cluster. Figure 3 shows the evaluation score for inertia in the clusters. The x-axis shows a label indicating the group being evaluated. The y-axis indicates the sum of the combined inertia scores. In our experiments, we found that the addition of hypernyms to a cluster on average improves results, but it also adds a level of noise to the cluster. This is reflected in earlier work, where we showed that the addition of synonyms and hypernyms improve performance on the mean of instances but also add in additional noise to the final score *citation withheld*. The Graph approach performed best in every cluster iteration. A low inertia score reflects tighter more compact cluster representations. It represents a better document clustering.

3.4 Homogeneity

A homogeneity score reflects the singleness of a class. It looks at the presence of the predominant class, and measures the purity of the clusters. Singleness/purity reflect the make-up of the cluster. Clusters that contain a high level of one class, with few representations of any other will factor highly on this scale. While clusters that consist of documents from many classes will receive a low score. It differs from majority-representation as homogeneity scoring takes the average of the representation and produces a normalised score. It does not consider the distribution of the class - so a document set with a hundred documents that produces a hundred clusters will have a perfect homogeneity score. Majority-representation requires the majority representation of a class to be assigned to one cluster, as well as being the most represented class in that cluster. Figure 4 shows the homogeneity scoring from our experiments. Homogeneity is normalised so that values can be between 0 and 1. Higher scores reflect more harmonious classes. In this instance the graph approach shows itself to be almost twice as compacted as the other representations. the hypernym approach is consistently second and the standard approach performs worst.

3.5 The Adjusted Rand index (ARI)

The Rand index, first proposed in 1971 (Rand, 1971) is an intrinsic evaluation approach that does a pairwise evaluation of four factors when assessing k clusters: true positive, true negative, false positive and false negative. It iterates over the clusters and evaluates each cluster point with that of each additional cluster point. If both cluster points are the same then it increments the value of true positive. If they both are different then it increments true negative. The final score is calculated by summing the true positive and true negatives and dividing them by the sum of all of the pairings in the cluster set. The adjusted Rand index (Hubert and Arabie, 1985) can be used in the instances where the true labelling of a document set is known. It is an extension to the Rand Index that attempts to offset the element of chance assignment of pairs of documents that could occur when the corpus is large, by factoring the actual index against the proposed index. In addition the ARI extends the range of the evaluation score over a [-1,1] width. This addition allows for negative scores being applied and reduces the congregation of results around one (Meilă, 2007) (Rand, 1971).

Figure 4 shows our final clustering evaluation indices. The results of the experiments show that in each cluster grouping, graph approach rates notably higher on the Rand Index Scale. The hypernym is superior to the standard approach and the performance of these two approaches reduces as the complexity of

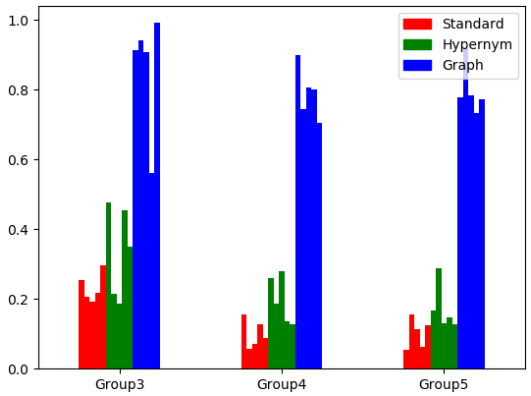


Figure 5: Rand Index Measure.

the clustering increases. There is no corresponding dip in the graph approach.

3.6 Clustering Complexity

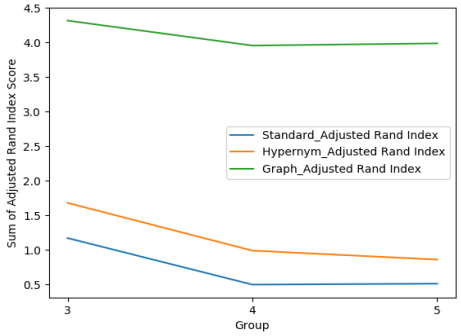
Figure 6 show the performance of each approach as the complexity of the task increases. Predictably, the error rate for group3 is lower in all metrics and increases as we move to group4 and group5. In all metrics, the graph approach shows itself to have superior results, that higher in MR, homogeneity and ARI and a lower rate in the overall inertia rate.

4 CONCLUSIONS

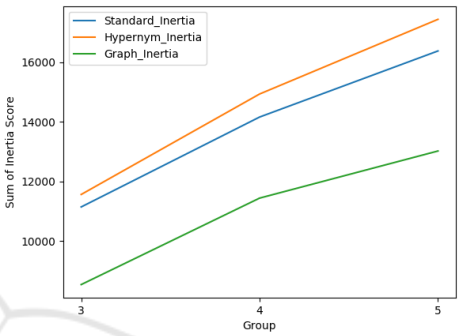
In this paper we have outlined an approach that utilises Reddit as an external source for creating better representation of documents. Improvement was measured using three standard cluster evaluation indices and one of our own creation. We compared the performance of using these graphs against another establish procedure of using Wordnet and show that better results are obtained when the source is dynamic. Future work will involve dynamically sourcing the related threads so that the system can automatically extract a related source and to test its efficacy when used with other supervised classifying approaches. This work has future applications of opinion mining where improvement representation of comments can be used for determining attitudes toward various items in the news, recommender systems and query expansion.

ACKNOWLEDGEMENTS

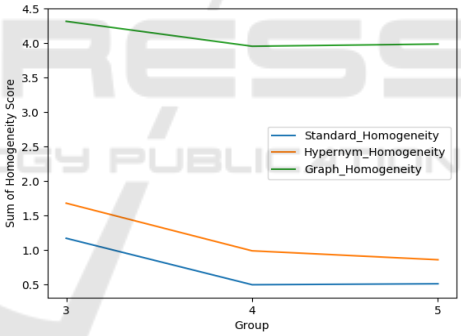
The authors acknowledge the support of Irelands Higher Education Authority through the IT Investment Fund and ComputerDISC in NUI Galway.



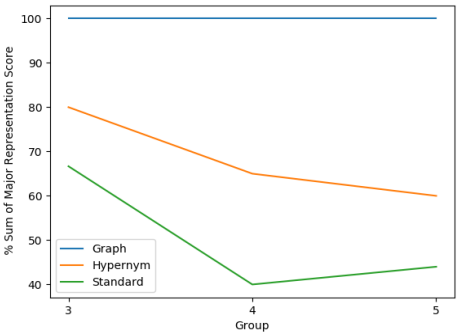
(a) Sum Total of ARI Per Grouping



(b) Sum Total of Inertia Per Grouping



(c) Sum Total of Homogeneity Per Grouping



(d) Sum Total of M-R Per Grouping

Figure 6: Results Per Grouping.

REFERENCES

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

Baralis, E., Cagliero, L., Jabeen, S., Fiori, A., and Shah, S. (2013). Multi-document summarization based on the yago ontology. *Expert Systems with Applications*, 40(17):6976–6984.

Bergstrom, K. (2011). Dont feed the troll: Shutting down debate about community expectations on reddit. com. *First Monday*, 16(8).

Cai, D., He, X., and Han, J. (2006). Tensor space model for document analysis. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 625–626. ACM.

Duggan, M. and Smith, A. (2013). 6% of online adults are reddit users. *Pew Internet & American Life Project*, 3:1–10.

Fabian, M., Gjergji, K., Gerhard, W., et al. (2007). Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*, pages 697–706.

Gilbert, E. (2013). Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803–808. ACM.

Hammouda, K. M., Matute, D. N., and Kamel, M. S. (2005). Corephrase: Keyphrase extraction for document clustering. In *MLDM*, volume 2005, pages 265–274. Springer.

Hu, X. and Liu, H. (2012). Text analytics in social media. *Mining text data*, pages 385–414.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Kleinberg, J. M. (2003). An impossibility theorem for clustering. In *Advances in neural information processing systems*, pages 463–470.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Li, X., Ramachandran, R., Movva, S., Graves, S., Plale, B., and Vijayakumar, N. (2008). Storm clustering for data-driven weather forecasting. In *24th Conference on International Institute of Professional Studies (IIPS). University of Alabama in Huntsville*.

Meilă, M. (2007). Comparing clusteringsan information based distance. *Journal of multivariate analysis*, 98(5):873–895.

Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In *Clustering and classification*, pages 341–375. World Scientific.

Müller, C. and Gurevych, I. (2009). A study on the semantic relatedness of query and document terms in information retrieval. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1338–1347. Association for Computational Linguistics.

Nagarajan, R. and Aruna, P. (2016). Construction of keyword extraction using statistical approaches and document clustering by agglomerative method. *International Journal of Engineering Research and Applications*, 6(1):73–78.

Potts, L. and Harrison, A. (2013). Interfaces as rhetorical constructions: reddit and 4chan during the boston marathon bombings. In *Proceedings of the 31st ACM international conference on Design of communication*, pages 143–150. ACM.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International Conference on Artificial Neural Networks*, pages 175–184. Springer.

Sarstedt, M. and Mooi, E. (2014). *Factor Analysis*, pages 235–272. Springer Berlin Heidelberg, Berlin, Heidelberg.

Shah, N. and Mahajan, S. (2012). Semantic based document clustering: A detailed review. *International Journal of Computer Applications*, 52(5).

Shahnaz, F., Berry, M. W., Pauca, V. P., and Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386.

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., and Strohmaier, M. (2014). Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd International Conference on World Wide Web*, pages 517–522. ACM.

Song, W., Li, C. H., and Park, S. C. (2009). Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Systems with Applications*, 36(5):9095–9104.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386.

Strapparava, C., Valitutti, A., et al. (2004). Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

Wallace, M. (2007). *Jawbone Java WordNet API*.

Weng, J. and Lee, B.-S. (2011). Event detection in twitter. *ICWSM*, 11:401–408.

Weninger, T., Zhu, X. A., and Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 579–583. ACM.

Zamir, O., Etzioni, O., Madani, O., and Karp, R. M. (1997). Fast and intuitive clustering of web documents. In *KDD*, volume 97, pages 287–290.

Zheng, H.-T., Kang, B.-Y., and Kim, H.-G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences*, 179(13):2249–2262.