Proc. of the Fifth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2017
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-131-3 doi: 10.15224/ 978-1-63248-131-3-19

Using bipartite graphs projected onto two dimensions for text classification

Stephen Redmond Eleni Rozaki

***Abstract*— In our Big Data world, the amount of text being gathered is ever expanding. For many years, data curators have sought ways to group these documents and identify common topics. As the size of the problem increases, solutions that will scale are needed. The purpose of this work is to present a novel text classifier that can be used for text-mining and interactive information access. The model that is demonstrated can be used to extract hierarchical relations between topics, as well as to conducted unsupervised clustering of documents and keywords. The approach that is taken with this model is the use of a graph-of-words key term extraction and a dimensional projection of the bipartite graph of documents and key terms. This projection makes it possible for terms to be co-clustered in an efficient manner in relation to their documents and the documents in relation to their terms. Furthermore, the key term extraction process that is outlined can be scaled on a large corpus using a distributed processing system such as Apache Spark, and the resultant model can be visually interacted with by users.**

Keywords— ***text mining, classification, clustering, bipartite graph, Apache Spark***

I. Introduction

The amount of text data that is created and needs to be catalogued and searched in the world is ever increasing. The sources of text data are many, from academic collections to newspaper archives, blogs and comments, and even collections of helpdesk tickets. The volume of text that exists means that curators of these data need methods to automate the categorising of the documents. As the size of the problem increases, it is also important to have a means of cataloguing and searching that can efficiently be scaled, in particular through the use of Big Data systems. Large amounts of text data may indicate that a Big Data approach such as MapReduce on Hadoop, which shares the tasks across multiple machines, is an appropriate technology to employ. [1] The current trend is to move to the in-memory execution engine afforded by Apache Spark.[2] One of the reasons that Spark is widely used is because it provides additional data and graph processing options.

There is also a tendency toward using more advanced machine learning techniques in this area. For example, researchers have applied a recurrent convolutional neural network.[3] However, the use of such models can be problematic in regulated industries, as stakeholders, such as regulators or customers, may demand access to the model details. The issues that have been mentioned lead to the question of what approach can be used for the mining of large amounts of text that is simple enough that it could be explained to a stakeholder, while also being visualised by a business user. The purpose of this paper is to present research around developing an approach to text mining that can be applied to any text corpus and scales using big data technologies. The outcome is a method using a graph-of-words key term extraction and a two-dimensional projection of the resulting bipartite graph of documents and key terms.

This paper begins with a discussion of the methodology used to conduct this study. A description of the keyword extraction method is presented followed by the unsupervised clustering method. Then the implementation of the text categorisation model is shown along with an evaluation of the model in real-world practice with the use of public datasets. Finally, the conclusions are presented.

II. Database Selection

In previous research,[4] two suitable corpora were curated and made available for research purposes[1]. One of the data sets was retrieved from the BBC Sport website and contained 737 articles across the topics of athletics, cricket, football, rugby and tennis. The second data set consists of 2,225 articles from the BBC news site on the topics of business, entertainment, politics, sport and technology.

A version of the 20 Newsgroups corpus was obtained from the UCI KDD Archive.[5] The corpus contained about 20,000 separate documents that were comprised of posts to twenty different Usenet newsgroups.

Because of the previous work of researchers,[4] the BBC corpora were already in a good, clean condition and did not need additional reprocessing. The files were put together in a zip file that contained a sub-folder for each of the topics. The 20 Newsgroups set was also packaged as a zip file that contained a sub-folder for each newsgroup. Each document also included additional header records that were not useful for the categorisation process and were removed.

Stephen Redmond
School of Computing
National College of Ireland

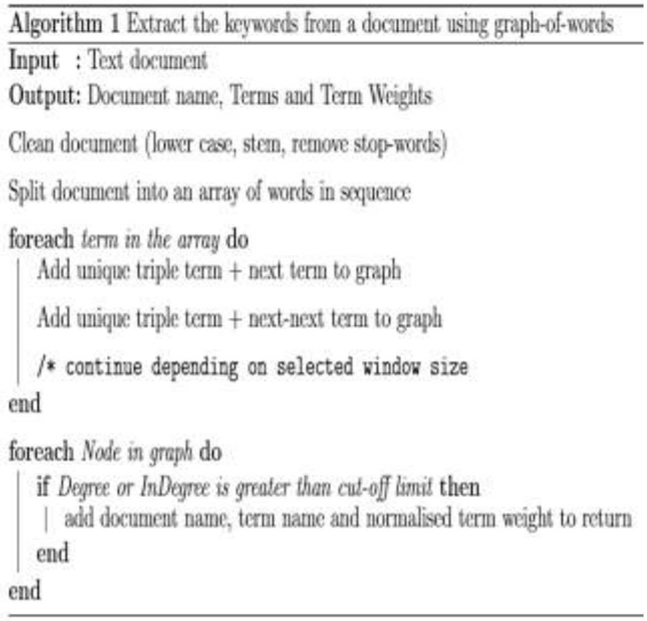
Eleni Rozaki
School of Computing
National College of Ireland

[1] BBC Datasets: http://mlg.ucd.ie/datasets/bbc.html

A. Extracting Keywords

Typical processes used in text analytics [6] such as tokenisation, lower case conversion, stemming using Porter's algorithm and stop word removal are also applied.

Keywords are then extracted using the graph-of-words technique used by other researchers [7]. This method is described in Algorithm 1 above.



```
Algorithm 1 Extract the keywords from a document using graph-of-words
Input  : Text document
Output: Document name, Terms and Term Weights

Clean document (lower case, stem, remove stop-words)

Split document into an array of words in sequence

foreach term in the array do
    Add unique triple term + next term to graph

    Add unique triple term + next-next term to graph

    /* continue depending on selected window size
end

foreach Node in graph do
    if Degree or InDegree is greater than cut-off limit then
        add document name, term name and normalised term weight to return
    end
end
```

The algorithm can easily be executed across multiple nodes on a Big Data platform because it is not dependent on information about other documents in the corpus.

The Term weight was normalised by dividing the Term Weight by the maximum Term Weight in that document.

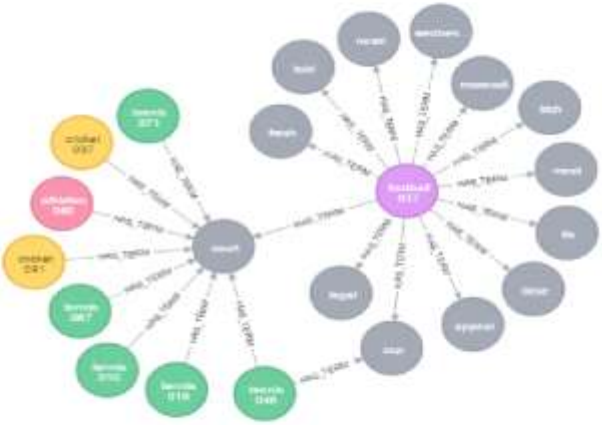


Figure 1. Example of bipartite document and term data

Figure 1 shows an example of the bipartite relationship between keywords and documents. Terms only connect to documents and documents only connect to terms.

III. Clustering with a bipartite graph and weighted centroids

In the previous section, keyword extraction was discussed. In this section, the process of storing the data in a bipartite graph and performing co-clustering is explained. In the bipartite graph, each document is one type of node, and each term is another type of node The edges connect the documents to the terms using the normalised Term Weight extracted from the graph-of-words process. The unique approach of this research project was to assign each node a position (x,y) in a two-dimensional space. For the clustering task, the initial positions of either the nodes or documents were unknown, so they were specified using a random number. Several iterations are performed in which each node is moved relative to its connected nodes. This means that each of the term nodes is moved to the weighted centroid of its connected documents. Next, each of the document nodes was moved to the weighted centroid of its connected terms. The process of moving first terms and then documents was iterated several times, and the terms and documents begin to co-cluster.

If the process was iterated too many times, then the terms and documents begin to converge to a point, which is not useful. A decision must be made regarding the best stage at which the process should cease. By measuring the entropy, or information gain, between two iterations with the use of a Kulback-Leibler divergence, a value can be obtained to compare against a specified parameter to use as a stop condition.

Figure 2 shows the cluster output of the process after eight iterations. The term nodes are coloured light grey while the document nodes are coloured post hoc using known document types.

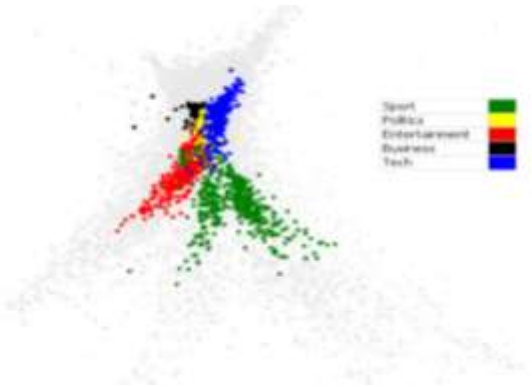


Figure 2. Subject clusters

The figure shows that clusters do emerge, but it could be difficult to extract the groups automatically using a clustering algorithm. However, the nodes can be presented to a user in a visualisation tool, such as QlikView, in order to allow the user to discover and assign the clusters interactively. It does, however, lead to the idea of the classification method.

IV. Implementing a text classifier

In this section, information is provided about using a similar positioning technique to create a text classifier. For a classifier, the initial starting positions of the documents could be established based on their known class. Furthermore, only a single repositioning of the terms needed to be performed followed by a single repositioning of the documents relative to their terms. Figure 3 shows an example of the classifier model built on the five class BBC Sport corpus. Once the documents were put into their final position, a kd-tree is constructed using the two-dimensional position values of the document nodes[8]. When a new document is received to be classified, the keywords and weights are extracted using the same graph-of-words technique used to build the classifier.

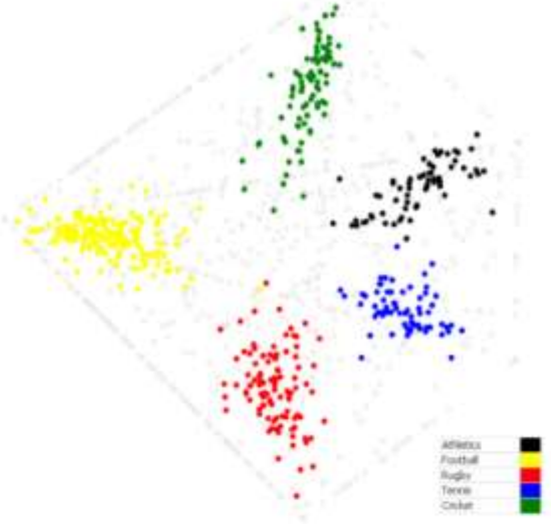


Figure 3: Display of the five class, BBC Sport corpus.

The keywords are mapped to the term nodes in the graph in order to establish the position for each. Then, the weighted centroid is calculated. The centroid is compared against the kd-tree, using the k-Nearest Neighbour algorithm, to locate the documents that are closest to the centroid. If the user requires an absolute response, then the document category that represents the most neighbours is returned. [9][10] However, if a user wanted a fuzzy request, the percentage for each category is returned.

A. Results of the Text Classifiers

The text classifier was implemented against the BBC News dataset and the BBC Sport dataset. RapidMiner was used to implement both a kNN and SVM (Support Vector Machine-linear kernel) model against the same datasets.

This allowed for the results for the datasets to be compared against each other. There were varying results across the three tests for the different corpora.

TABLE I RESULTS OF CLASSIFICATION USING THREE METHODS

Corpus	Classifier F1 score		
	KNN	***SVM***	***Graph***
BBC Sport	0.9777	0.9509	0.9598
BBC News	0.9556	0.9719	0.9206

The results that are shown in Table I show a high degree of accuracy. In addition, the classifiers supervised the clustering results in order to identify the groups and make decisions about the relative topics of the viewers.

B. Results of topic groups classification

Table II shows that the effort to classify across twenty-newsgroup set using each label as a classifier was not successful. The lack of success occurred even though the SVM performed adequately. However, the lack of success was not unexpected when considering the challenge.

TABLE I I RESULTS OF CLASSIFICATION ON 20 NEWSGROUPS DATASET

Corpus	Classifier F1 score		
	KNN	***SVM***	***Graph***
20 Newsgroups	0.1175	0.7788	0.3280

It must be recognised that many of the newsgroups cover similar topics, which means that they have similar keywords. Visually examining the results allowed a user to see this interaction and make decisions about grouping topics. The ability to visually inspect the model is demonstrated in Figure 4.

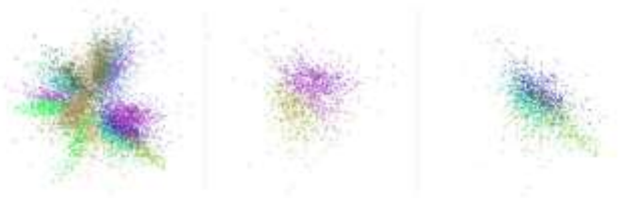


Figure 4: Visually examining the model to identify groups.

The first image in the panel shows the entire model with overlapping topics. A user might imagine that the groups talk.religion.misc, alt.atheism and soc.religion.christian would be a natural grouping. The second panel shows that the group soc.religion.christian does not overlap with the other two groups. Instead, the third panel shows that soc.religion.christian overlaps with the three talk.politics groups. Additionally, the two sport groups of rec.sport.baseball and rec.sport.hockey do not overlap very well.

Table III Grouping of Newsgroups based on Topic Similarity

Science	**Computer**	**Politics**
sci.crypt	comp.graphics	talk.politics.guns
sci.electronics	comp.sys.ibm.pc.hardware	talk.politics.mideast
sci.med	comp.os.ms-windows.misc	talk.politics.misc
sci.space	comp.sys.mac.hardware	soc.religion.christian
	comp.windows.x	
Misc	**Motors**	**Baseball**
misc.forsale	rec.autos rec.motorcycles	rec.sport.baseball
Religion		**Hockey**
alt.atheism		rec.sport.hockey

Table IV Results of classification using three methods

Corpus	**Classifier F1 score**		
	KNN	***SVM***	***Graph***
8 Topic Groups	0.2575	0.8445	0.6381

Based on visually exploring the model, a new set of topic groups emerged, which are shown in Table III . When using these groupings, the classification improved in both the SVM and the bipartite graph method. The improvement was due to the improved separability of the topics.

By using the visualisation to explore the topics, the user was able to make such discoveries and improve the overall results. It was the ability of the user to visually interact with the model and to discover the relationships between topics that was important.

C. Conclusion

The purpose of this paper was to demonstrate the use of an approach to text mining that could be applied to any corpus and scaled using big data technologies. The outcome was a novel classifier using a graph-of-words keyword extraction method and a bipartite co-clustered graph represented on a two-dimensional plane that could be visualised and explained to business stakeholders or regulators. The production of the graph could be scaled using Apache Spark. The text classifier produced satisfactory results on some of the datasets used. However, even in areas where the results were not as strong. a user could visually examine the model to discover where there were topics that overlapped and could potentially be merged. This method is a valuable addition to the body of knowledge regarding text mining because of the use of a classifier that was both visually interactive and explainable. From a real-world standpoint, it is important for companies to be able to explain the model that is being used. In fact, it may become even more important for companies to explain the models they use because of new regulations, such as the EU General Data Protection Regulations.

Table V Contingency Table of Results from BBC News Classifier

Predicted	bus.	entert.	politics	sport	tech	**Total**
Actual						
bus.	140	1	1	0	5	**147**
entert.	15	89	17	0	0	**121**
politics	6	3	123	1	1	**134**
sport	0	0	3	155	2	**160**
tech	2	0	2	0	110	**114**
Total	**163**	**93**	**146**	**156**	**118**	**676**

References

[1] Lin, J. and Dyer, C. (2010). Data-intensive text processing with mapreduce, Synthesis Lectures on Human Language Technologies 3(1): 1–177.

[2] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S. and Stoica, I. (2012). Proceedings of the 9th

USENIX conference on Networked Systems Design and Implementation, USENIX Association, pp. 2–2.

[3] Lai, S., Xu, L., Liu, K. and Zhao, J. (2015). Recurrent convolutional neural networks for text classification., AAAI, Vol. 333, pp. 2267–2273.

[4] Greene, D. and Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering, Proc. 23rd International Conference on Machine learning (ICML'06), ACM Press, pp. 377–384.

[5] Mitchell, T. M. (1997). Machine learning. 1997, Burr Ridge, IL: McGraw Hill 45(37): 870–877.

[6] Rousseau, F. and Vazirgiannis, M. (2015). Main core retention on graph-of-words for single-document keyword extraction, European Conference on Information Retrieval, Springer, pp. 382–393.

[7] Uysal, A. K. and Gunal, S. (2014). The impact of preprocessing on text classification, Information Processing Management 50(1): 104 – 112.

[8] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching, Communications of the ACM 18(9): 509–517.

[9] Iswarya, P. , Radha V. (2015). Ensemble learning approach in Improved K Nearest Neighbor algorithm for Text Categorization, IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15.

[10] Taeho Jo (2017). Using K Nearest Neighbors for Text Segmentation with Feature Similarity, 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE).

Stephen Redmond is a research student in the area of data analytics at the National College of Ireland. His current research interests include text mining techniques, decision support systems and advanced analytics techniques.

Eleni Rozaki (S-22, M-13) is currently working as an associate lecturer in the School of Computing at the National College of Ireland. Her Ph.D. degree is in the area of data mining and business analytics at Cardiff University, United Kingdom. Her current research interests include, decision support systems and text and data mining techniques.