Detecting Digital Newspaper Duplicates with Focus on eliminating OCR errors

Yeshey Peden
ICT Officer
Department of Public Accounts
Ministry of Finance
Thimphu, Bhutan
ypgyeltshen@gmail.com

Richi Nayak
Associate Professor
Higher Degree Research Director
Science and Engineering Faculty
Queensland University of Technology
r.nayak.qut.edu.au

Abstract

With the advancement in digitalization, archived documents such as newspapers have been increasingly converted into electronic documents and become available for user search. Many of these newspaper articles appear in several publication avenues with some variations. Their presence decreases both effectiveness and efficiency of search engines which directly affects user experience. This emphasizes on development of a duplicate detection method, however, digitized newspapers, in particular, have their own unique challenges. One important challenge that is discussed in this paper is the presence of OCR (Optical Character recognition) errors which negatively affects the value of document collection. The frequency of syndicated stories within the newspaper domain poses another challenge during duplicate/near duplicate detection process. This paper introduces a duplicate detection method based on clustering that detects duplicate/near duplicate digitized newspaper articles. We present the experiments and assessments of the results on three different data subsets obtained from the Trove digitized newspaper collection.

Keywords: clustering, duplicate document detection, OCR errors, feature selection

1 Introduction

The Australian Newspaper Digitization Program (ANDP) has initiated the digitization of newspapers archives before the copyright act. ANDP has provided free access to this data, relevant to Australia via the Trove search engine[1]. A significant problem with digitisation of archived newspapers is the presence of identical and nearly identical documents in the resulting collection. These duplicate documents are not only an annoyance to users as search results, but, they also decrease efficiency of search engines (Uyar, 2009). The processing of these duplicate documents and results is not only time consuming, but, their presence also does not add any value to the information presented to users. Duplicate document detection has gained research interest as it assists in search engines in increasing the effectiveness and storage efficiency (Uyar, 2009). While carrying out this task, consideration of the domain/application is very important. For example, in plagiarism detection scenario, even if a sentence or a paragraph of one document is found in another document the two documents could be seen as near-duplicates. This definition of near-duplicate may be looser in other domains as comparison to news domain (Hajishirzi, Yih and Kolcz, 2010).

Newspaper articles exhibit distinct characteristics (Smeaton, Burnett, Crimmins and Quinn's, 1998). Almost all news stories have an element of a continuum and news stories usually evolve over time. A large proportion of articles are related to previously published articles. There exists a dependency between news stories which cannot be ignored when navigating news archives. These dependencies need to be acknowledged while looking for duplicate newspaper articles. If two documents are identified as near identical articles, when in fact one contains new information, this potentially leads to loss of information (Gibson, Wellner and Lubar, 2008). In this paper, we explored document clustering as a technique to identify and generate links between related news stories.

Another significant problem with the use of digitized collections in search engine is OCR (Optical Character recognition) errors. These errors are non-existing words which result due to incorrect letter recognition. OCR helps resolve the growing problem of searching intended information from the digital archives; however, historical newspapers are different from the recent newspapers in image quality, type fonts, ruby characters, noise, and language usage. The direct use of these digitized collections, without any processing for quality improvement, cannot provide satisfactory results (Shima, Terasawa and Kawashima, 2011). Newspapers especially old ones are certain to have OCR errors no matter how well the digitization is done. The reasons include the

[1] http://trove.nla.gov.au/

condition of the source material, the typeface used, the contrast between print and paper and many more. (DLC, 2011).Whereas such challenges do not surface in other web documents,

An abundance amount of work exists on identifying duplicated online web pages (Conrad, Guo and Schriber, 2003; Radlinski, Bennett and Yilmaz, 2011), but, none can be found on digitized newspapers with particular consideration to OCR errors. While identifying exact duplications of plaintext documents has been more straight forward (Gibson, Wellner and Luber, 2008), identifying near duplicates has been challenging (Alonso, Fetterly and Manasse, 2013). The task of defining near duplicates and setting a "resemblance" threshold (a threshold that classifies near identical and non-identical documents) is difficult and varies across the domains. The errors resulted due to OCR software application not only have the potential to affect the clustering solutions but it also makes it difficult to set a threshold and identifying redundancy.While it will be almost impossible to detect exact duplicates, it will be most difficult to decide at what level of threshold to discard the near duplicates. At times, the small differences between the two near duplicate documents might have important information to reveal. Moreover two documents might be near duplicate but a lot of words might not match due to OCR errors. This will result in falsely classifying the documents as not duplicates.

This paper is a step towards finding the solution to the problems associated with digitized newspaper articles. We propose a solution to deal with unique problems such as eliminating OCR errors and detecting redundant news stories which may negatively affect the search engine performance and the user experience. We propose a methodology using a pre-processing method to eliminate OCR errors and clustering algorithms. Evaluation of the clusters is done using both internal and external measures. The proposed method is based on the assumption that the closest neighbours within a cluster can represent near duplicates or exact duplicates.

The remainder of this paper is set out as follows. Section 2 will describe the approach taken. This is followed by experiments, results and evaluation in section 3 and finally ending with the conclusion.

2 The Proposed Clustering based Method

The proposed document duplication detection method starts with initial pre-processing that includes removal of all non-alphanumeric characters and stop words (i.e. extremely common words which are not valuable) and stemming of the terms to its root form using Porter's stemmer (Porter, 1980). This is followed by identification of the more common words and OCR error using a simple method consisting of frequency count and threshold cut-off. A sample of raw data from Trove dataset displayed in Figure 1 contains non-existing words such as "luetnc", "uecrioi", "gnnitttt" and "oiilitioiiali" which resulted due to incorrect letters.

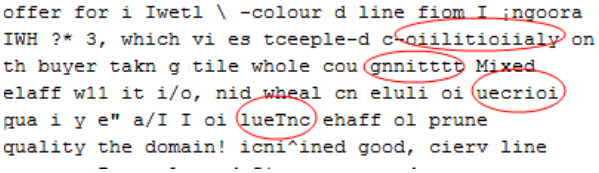

Figure 1: A sample of raw data from Trove dataset prior to cleaning

For each term in the document collection, Term Frequency (TF) is calculated. TF is the number of times the term occurs in the document collection D.

TF (t, D) = f (t, D)

Selection of a threshold cut-off is critical. Setting a higher threshold may lead to elimination of valuable words. Setting the threshold too low may still leave the noise or OCR errors in the collection. It is not possible to get a perfect threshold cut-off that leads to elimination of all the noise while keeping all the important words. The aim is to find a cut-off that might not lead to complete elimination of the OCR errors but, it should remove most of the common ones, and it should also lead to a reduced word space while keeping the important terms. This will reduce the computational time and, additionally, lead to a more accurate clustering solution.

The decision on the threshold cut-off is made by plotting the frequency distribution of the terms. Since the number of terms is too big to be put in one plot, random points are picked for visualization shown in Figure 2. The x-axis of the graph represents the number of times (n) a term occurs within the document collection and the y-axis shows the terms that occurs n times. It is seen that a total number of 842385 terms occurs one time and also the maximum count of times a term occurs is 35,717.

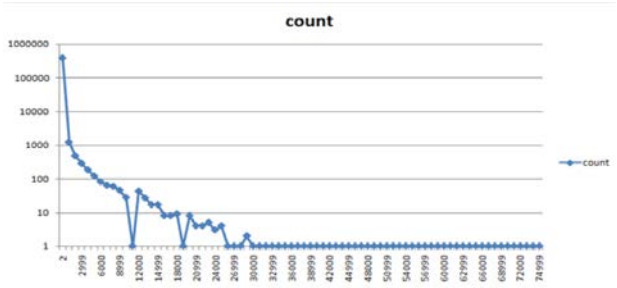

Figure 2: A graph presenting frequency distribution of terms-random points picked due to largeness of data

It can been seen in Figure 3 that the words identified as results of OCR error in Figure 2 actually occur only one time in the entire document collection. Such rarely occurring terms are usually filtered out by setting a correct threshold.

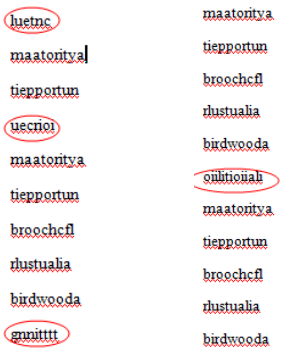

Figure 3: A sample list of stemmed terms occurring one time in the complete document collection

Once pre-processing is completed the document collection is represented in the Vector Space Model (VSM) which represents documents as a feature vector of the words that appear in the documents of the collection (Jin, Chai and Si 2005). Each feature vector is presented with the term-frequency * inverse-document-frequent (tf * idf) weighting. This weighting will consider a term important if it occurs frequently within a document and does not occur so frequently in the collection. This model becomes input to the clustering process.

A choice of clustering algorithm depends upon many factors. Experimental studies in the literature have often portrayed hierarchical clustering to be better than the partitional K-means, in terms of clustering quality but inferior in terms of time complexity (Steinbach, Karypis and Kumar, 2000). In this paper, we apply the repeated bisections partitional clustering approach (Karypis, 2002) to deal with the large dimensions. The method is a K-way clustering solution that is computed by performing a sequence of K-1 repeated bisection on the data instances where K is the desired number of clusters. In addition to having low computational requirements, this bisecting K-means approach has a time complexity which is linear in the number of documents. It is found that as K increases, there is an increment in the optimization of the criterion functions as well.

Using a nearest neighborhood method that uses cosine similarity, for each object (a newspaper article) within a cluster, its ten nearest neighbors is found. If the similarity between the object and its nearest neighbor is within the acceptable threshold, it is considered as identical or near-identical. It is assumed that any near duplicates of that particular document should be its nearest neighbour within the same cluster.

3 Experiments and Evaluation

This section focuses on assessing the effectiveness of the proposed document duplication detection method. We first present the datasets and evaluation measures used in experiments. The results from the pre-processing phase of the experiment will be presented next. This is followed by discussion and analysis of the results obtained from the clustering phase. Finally, we will present and analyze the results of duplicate identification obtained from the cluster evaluation phase.

We use the clustering tool CLUTO to perform clustering (Karypis, 2002). CLUTO is chosen because it is easy to use and is able to operate on very large dataset with respect to number of documents as well as number of dimensions. CLUTO has many clustering algorithms and each of these algorithms has different scalability characteristics. Karypis (2002) has shown that the most scalable method in terms of time and memory is vcluster's repeated-bisecting algorithm that uses cosine similarity function.

3.1 Dataset

The Trove data set consists of a total number of 77,841,027 documents. The documents fall into six categories which are: Literature; detailed lists, results and guides; Advertising; Articles; Family Notices; and Others. The main focus of the study is on the Article category which consists of a total of 58,549,810 documents (as shown in Table 1). The experiment is conducted on three different data subsets selected from three different time period of year 1921 (called as Dataset 1), year 1880 (called as Dataset 2) and year 1862 (called as Dataset 3). The data collection for the period of 1921 is randomly chosen while the years 1862 and 1880 collection are chosen based on queries identified by historians as having duplicates. It is based on the fact that there are duplicates of those particular news articles. For example, running the search query "kipper billy daring attempt" on Trove website is supposed to produce at least three duplicate copies. For each dataset, a two months of data snapshot has been used for clustering experiments.

Category Name	**#Documents**
Literature	12289
Detailed Lists, Results, Guides	7516250
Advertising	11058829
Article	**58549810**
Family Notices	703846
Others	3
Total:	**77841027**

Table 1: Total number of documents under each category

The study is based on the premise that, in archived news-stories in late 1800s and early 1900s, the duplication in news occurs within the two months of the release of the original news story and not beyond that. This assumption has been made based on discussion with historians involved in the project, as well as, based on the facts that the stories used to deliver via telegraphs and printing and delivery process used to take long. Today with partial replacement of newspapers by web news and network television news which gets updated in real time, most duplication would happen on the same day. Even with newspapers, most publication happens daily or weekly. A search completed on Trove search engine for both queries "Kelly expiated his career in crime" and "kipper billy daring attempt" showed that news duplication does not go beyond two months. These two queries were for news that happened in 1860 and 1862 respectively.

In the original data collection, the news articles include some advertisement. There was no heuristic way to identify these ads and distinguish them with the news stories. A simple method is applied to identify and eliminate any ads occurring within the data collection. An initial clustering is conducted on the data collection after

a simple pre-processing of stemming and general stop word removal. Once the clusters are obtained, the best clusters on the top are manually observed. It has been seen that any cluster that has few number of documents and that has very high intra-similarity tends to be ads most of the time. The exact same ads might repeat in several newspapers and appear during extended periods. This might not allow efficient identification of duplicate news articles. Any document within these clusters which is identified as an ad is removed from the collection. This process slightly improves the efficiency of the clustering solution.

3.2 Evaluation Measures

The quality of the clusters obtained is measured using two measures: internal quality measure and external quality measure. The internal quality measure is based on the examination of the internal similarity (ISIM) that is the average similarity between the objects of each cluster and the output value of the criterion function used. The ISIM has values between 0 and 1. Good clusters have ISIM and output value of the criterion function closer to 1. In contrast a bad cluster has ISIM and output value of the criterion function closer to 0. This is also known as the quantitative measure.

In addition to the internal measures, an alternative qualitative measure is required to evaluate the effectiveness of the clustering algorithm in grouping the similar digitized news articles together. This leads to finding out how successfully the near duplicate articles are detected. An extrinsic method using ground truth is applied. The ground truth for this study is a collection of information retrieved from the Trove search engine based on certain queries. A simple process of manual observation is applied to the results retrieved by the query in order to identify duplicates related to that query. Then the cluster solution is taken for further evaluation.

3.3 Effectiveness of pre-processing

This section discusses the effect of pre-processing phase in removing the OCR errors and reducing word space. An assumption is made that if a terms occurs too many times in the overall document collection, the term is not important and does not provide any valuable information while if a terms occurs very few times such as once, the term is an OCR error term. This conclusion is drawn from manual observations of the data. A cut-off threshold is set to identify too frequent and too infrequent terms.

Experiment on Dataset 1 (Table 2) shows that a number threshold is better suited than a % threshold even though it is not a perfect solution by itself. A % threshold leads to elimination of too many terms which could lead loss of valuable words. It has been seen that increasing the threshold cut-off leads to maximum elimination of OCR errors but it also increases the risk of removing important terms.

Input file	Min & Max Threshold culled	Matrix out put (doc, term, matrix density)
Original data collection	Max:8000, Min:1	115755, 398858, 12249111
Data collection: ads removed	Max:8000, Min:1	115635, 398826, 12482467
Data collection: ads removed	Max:8000, Min:2	115635, 229043, 12178014
Original data collection	MIN:1% ; MAX:70%	115755, 2408, 684543
Original data collection	NONE	115755, 1869496, 13861605

Table 2: Dataset 1(1921) - Comparison of different versions of the data collection with different threshold cut offs

Observation of Table 2 shows that it does reduce the word space of the document collection, but, the reduction is not too great. Moreover comparing the threshold cut offs MIN=1 and MIN=2, the latter proves to be a better choice than the former. The “Min” indicates how many minimum documents the term should appear in, and the “Max” shows how many maximum documents, the term should appear in. Note that this processing is conducted after the standard stop-word removal and stemming so the dataset does not contain many stop-words or rare words. This culling would ensure that errors associated with OCR are removed.

It can be seen from Table 3 that the total number of documents in the 1880 collection is 28717. The original data without any MIN and MAX threshold culled has 1045656 unique terms and the density is 7340510. By eliminating the most frequent and infrequent (OCR errors) terms from the original data using a simple MIN and MAX threshold cut off, the number of unique term has reduced to 130579 and the dimensionality has also been reduced.

Input Data	MIN & MAX Threshold culled	Matrix out put
Original data	Max:8000, Min:2	28717 x 130579 x 6274642
Original data	None	28717 x 1045656 x 7340510

Table 3: Dataset 2 (1880) - Comparison of the original input data with the data from which the frequent and infrequent terms (OCR errors) have been removed

Similar results can be seen for Dataset 3 in Table 4.The identification and elimination of approximately 87.49% of total number of terms in the 1880 data collection (as shown in Table 5) and 86.79% of total number of terms in the 1862 data collection (as shown in Table 6), as infrequent/OCR errors, confirm the concern stated earlier with regard to OCR errors. Figure 4 displays the top ten OCR errors in the three different datasets that are removed by our method.

Input File	Min & Max Threshold culled	Matrix out put (doc, term, density)
Orginal Data	Max:8000, Min:2	16696, 103172, 4347401
Orginal Data	None	16696, 782569, 5137437

Table 4: Dataset 3 (1862) - Comparison of the original input data with the data from which the OCR errors have been removed

Total No of unique terms	MIN Threshold	Terms occuring<=MIN Threshold	Terms occuring more than 8000 times	Good usable terms
1045656	2	914886=87.49%	1	130769=12.51%

Table 5: Dataset 2 (1880) - The most frequent, infrequent and good usable terms

Total No of Unique terms	Threshold	Terms occuring<=Threshold	Terms occuring more than 8000 times	Good terms
782569	2	679221=86.79%	1	103347=13.21%

Table 6: Dataset 3 (1862) - The most frequent, infrequent and good usable terms

jrticiefl	pretiid	1t1l
bxohaagw	pretiic	gehmhnh
flezv	flev	vollkommerr
uildltion	1roiu	flez
flg5	flet	icliarfca
iimurbvd	tenipemtui	ptcherd
mfksr	amenlldmenlldmenltmenlld	riwuh
asrest	fler	ttfdrutahd
supuppliosup	sansllst	flep
lrrlgate	flel	sanult
Dataset 1 (1921)	Dataset 2 (1880)	Dataset 3 (1862)

Figure 4: Top 10 OCR errors removed by the defined process in all three datasets respectively

3.4 Internal Evaluation of Cluster Solutions

For Dataset 1, it is seen that the best clusters were produced by a combination of K=2000, method=RBR and criterion function=I2 (as shown in Table 7 and 8). For Dataset 2, Table 9 and 10 shows that combination of K=600, method=RB and criterion function=I1 produced a better solution compared to the rest. For Dataset 3, it is clear from results in Table 11 and 12 that K=400 produces a better clustering solution compared to K=300. Looking at the methods, if based on the output value of the criterion function, RBR methods combined with criterion function I2 outperformed the other combinations. Otherwise if based on the average ISIM value, RB with I1 seems to be performing better than the rest.

Although some combination of the methods outperformed the other, the quality of overall clustering solution is not very high. For a good solution, it is expected that the average ISIM value and the output value of the criterion function at least be 0.5 and above. These values are obtained just around 0.5. An external measure based on ground truth is required to evaluate the solution further and discover the effectiveness of the methodology.

K value	MIN Threshold Cut off	No of Docs Clustered	OUTPUT
100	1	115630 of 115635	I2=2.68e+04
100	2	115630 of 115635	I2=2.74e+04
1000	1	115630 of 115635	I2=4.21e+04
1000	2	115630 of 115635	I2=4.28e+04
2000	1	115630 of 115635	I2=4.79e+04
2000	2	115630 of 115635	I2=4.88e+04

Table 7: Dataset 1, cluto clustering solution for method=rbr; -sim (similarity measure)=cosine; criterion function=I2

Method	Criterion Function	SIM	Output
RB	I2	Cos	I2=3.75e+04
RBR	I2	Cos	I2=4.80e+04
RB	I1	Cos	I1=1.39e+04
RBR	I1	Cos	I1=1.68e+04

Table 8: Dataset 1, comparison of two different methods and two criterion functions in CLUTO using K=2000

Method	Criterion Function	SIM	OUTPUT	AVG no of Docs in each cluster	ISim(AVG)
RB	I2	Cos	I2=1.04e+04	57.428	0.17413
RBR	I2	Cos	I2=1.1202e+04	57.428	0.17938
RB	I1	Cos	I1=4.35e+03	57.428	0.329976
RBR	I1	Cos	I1=4.88e+03	57.428	0.31893

Table 9: Dataset 2, result summary for K=500

Method	Criterion Function	SIM	OUTPUT	K(AVG)	ISim(AVG)
RB	I2	Cos	I2=1.08e+04	47.85667	0.181992
RBR	I2	Cos	I2=1.16e+04	47.85667	0.190278
RB	I1	Cos	I1=4.64e+03	47.85667	0.345382
RBR	I1	Cos	I1=5.20e+03	47.85667	0.327673

Table 10 : Dataset 2, result summary for K=600

Method	Criterion Function	SIM	OUTPUT	AVG no of Docs in each cluster	ISim(AVG)
RB	I2	Cos	I2=6.29e+03	55.65	0.192987
RBR	I2	Cos	I2=6.64e+03	55.65	0.193087
RB	I1	Cos	I1=2.78e+03	55.65	0.34794
RBR	I1	Cos	I1=3.01e+03	55.65	0.335223

Table 11 : Dataset 3, result summary for K=300

Method	Criterion Function	SIM	OUTPUT	AVG no of Docs in each cluster	ISim(AVG)
RB	I2	Cos	I2=6.63e+03	41.7375	0.2023825
RBR	I2	Cos	I2=7.00e+03	41.7375	0.207855
RB	I1	Cos	I1=3.05e+03	41.7375	0.345963
RBR	I1	Cos	I1=3.32e+03	41.7375	0.337478

Table 12: Dataset 3, result summary for K=400

3.5 External Evaluation of Cluster Solutions

The three datasets are further evaluated using an extrinsic method. The quality of cluster is evaluated by finding out how successfully the near duplicate articles can be detected using the cluster outputs. Analysing the results obtained using Dataset 1 (as shown in Table 13), it is established that when the cosine value is equal to 1, two documents are exact duplicates and when the cosine value is below 1 and above 0.4, the documents could be near duplicate.

Document ID	Cosine Value
83949238	1
70768975	0.602139
66634151	0.589655
16882523	0.58184
51105364	0.578736
27953167	0.568314
79392563	0.496901
80494852	0.484702
92886510	0.481333
20456292	0.474761

Table 13: Dataset 1 - Nearest neighbors of document id 83949238

Human observation shows that all 10 nearest neighbours obtained for that document can be called near duplicates but the low cosine values resulted because of the OCR errors, and due to the fact that the same information could have been presented using different words.

We have not yet used a semantic model to incorporate the semantic words in the matching process. With Dataset 2, based on results from Table 14, it is understood that when cosine value is 1 or 0.9999, the document is an exact duplicate and the documents are near duplicate if the cosine value is 0.8 and above. The documents are more of an update on the news if the threshold is 0.4 and above. It is seen that that better observation can be made from dataset 2 due to presence of less OCR errors.

Document ID	Cosine value	Comments
13476046	1	exact duplicate
13483531	0.957245	near duplicate
13477812	0.945398	near duplicate
78918195	0.90304	near duplicate
813385	0.842917	near duplicate
65379447	0.54443	contains information about Ned Kelli's execution but it is more of an update of the news with more details
65960937	0.50208	contains information about Ned Kelli's execution but it is more of an update of the news with more details
77590206	0.492915	contains information about Ned Kelli's execution but it is more of an update of the news with more details.
89686423	0.479583	contains information about Ned Kelli's execution but it is more of an update of the news with more details
2984082	0.44524	there is near duplicate information but because of the use of different terms, due to OCR errors and due to presence of other news, cosine similarity value is low

Table 14: Dataset 2 -Nearest neighbors of document id 13476046

With Dataset 3, from the results in Table 15 it is found to be difficult to set a cosine value threshold for identifying near/exact duplicates. For a particular document, human observation proved that some of its nearest neighbours obtained are near duplicate information. But with the presence of OCR errors, the cosine value is calculated lower than it should be.

Document ID	Cosine Value	Comments
4604263	1.0	
13225704	0.72897	Near duplicate. Could have had higher cosine value if other news stories were not merged in it.
59790774	0.665777	Near duplicate but cosine value is lower because of too many OCR errors
13225448	0.319914	Near duplicate but it just contains a shortsummary news. No details given
18687137	0.294292	Near duplicate. Could have had higher cosine value if other news stories were not merged in it
4604821	0.291291	Related news but not near duplicate
18687088	0.249018	Contains a brief summary of the news. No details and other news stories are merged in it
60509989	0.221332	Contains a brief mention of the kipper billy news(update) but because of the OCR errors and also it contains other news stories merged in it
79976917	0.106073	Similar news but not related
90253966	0.090444	Completely different news

Table 15: Dataset 3 - Nearest neighbours of document id 4604263

While the methodology has been successful, to some extent, in detecting near/exact digitized news documents, it is clear that the pre-processing phase could have been improved further. The pre-processing phase did not completely eliminate the OCR errors and no consideration was given to incorporate semantic models to deal with synonyms and hyponyms. These flaws had impact on the clustering solution and therefore clear observation could not be made at the end. Anyway this project was a first step forward and, with more sophisticated pre-processing, improved results can be obtained.

4 Conclusion

This paper presented a clustering based duplicate detection method for digitized newspaper stories. Three different data subsets have been employed to test the method. Correction of OCR errors has been found a significant issue dealing with digitized collection of the archived documents. Employing a simple method of threshold cut-off to eliminate the OCR errors has not been found to be a perfect solution. While for each dataset, some combination of the methods outperformed the other, the overall clustering solution was not found to be optimal. This would have been due to the reason that there has not been 100% removal of OCR errors from the digitized news article collection. The risk of removing important terms prevented the removal of 100% of the errors. It is found that increasing the number of clusters leads to better quality clusters but it has also been established that increasing the value of the cluster number beyond a certain point leads to clusters containing very few documents which does not serve the purpose of keeping all similar news documents together.

The clusters were further evaluated by using an extrinsic nearest neighborhood method that was used to find the 10 nearest neighbor for each document/object in a cluster.

In future, to make the duplicate detection more robust and efficient, a query dependant method can be explored and adapted. More time can be spent on the pre-processing phase of the methodology to find a way to eliminate the OCR (Optical Character recognition) errors without removing any valuable terms. The simple threshold cut off method can be improvised into a more sophisticated method. Likewise for dimensionality reduction, the same simple method was used. Other methods such as the PCA

(Principal Component Analysis) (Indhumathi and Sathiyabama, 2010) and SVD (Singular Value Decomposition) can be explored and experimented with as well.

It would also be important to look into other clustering methods. The methodology developed in this study adapted the bisecting partitional approach which does not consider a new incoming document. Methods such as adaptive K-means clustering allow clusters to grow without depending on the initial selection of cluster representation. It would be interesting not only to explore the different software, tools and methods but it will also be interesting to expand the study to understand the evolution of news stories and enhance user search experience.

5 Acknowledgement

We would like to acknowledge CRC Smart Services and Prof Kerry Raymond for facilitating data acquisition from National Library of Australia. We acknowledge Prof Paul Turnbull, Dr Sangeetha Kutty, Sharon Pingi and Behzad for constructive discussion and assistance in processing the data. Yeashey will like to thank AusAID scholarship and AusAID student contact officers in QUT for their full support.

6 References

Alonso, O., Fetterly, D. and Manasse, M. (2013): Duplicate news story detection revisited. In *Information Retrieval Technology*: 203-214, Springer Berlin Heidelberg.

Conrad, J. G., Guo, X. S., and Schriber, C. P. (2003, November): Online duplicate document detection: signature reliability in a dynamic retrieval environment. In *Proceedings of the twelfth international conference on Information and knowledge management*: 443-452, ACM.

DLC (2011, May). A brief history of tactile writing systems for readers with blindness and visual impairments.http://www.dlconsulting.com/digitization/the-unique-challenges-of-newspaper-digitization.

Gibson, J., Wellner, B., and Lubar, S. (2008): Identification of Duplicate News Stories in Web Pages. In *Workshop Programme*: 26.

Hajishirzi, H., Yih, W. T., and Kolcz, A. (2010, July): Adaptive near-duplicate detection via similarity learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*: 419-426, ACM.

Indhumathi, R., and Sathiyabama, S. (2010): Reducing and Clustering high Dimensional Data through Principal Component Analysis. *International Journal of Computer Applications*, 11:8.

Jin, R., Chai, J. Y., and Si, L. (2005, August): Learn to weight terms in information retrieval using category information. In *Proceedings of the 22nd international conference on Machine learning*: 353-360, ACM.

Karypis, G. (2002): CLUTO-software for clustering high-dimensional datasets. http://gloros.dtc.umn.edu/gkhome/cluto/cluto/download.

Porter, M. F. (1980): An algorithm for suffix stripping. *Program: electronic library and information systems*, *14*(3): 130-137.

Radlinski, F., Bennett, P. N., and Yilmaz, E. (2011, February): Detecting duplicate web documents using click through data. In *Proceedings of the fourth ACM international conference on Web search and data mining*: 147-156, ACM.

Shima, T., Terasawa, K., and Kawashima, T. (2011, September): Image processing for historical newspaper archives. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*: 127-132, ACM.

Smeaton, A. F., Burnett, M., Crimmins, F., and Quinn, G. (1998, March): An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts. In *BCS-IRSG Annual Colloquium on IR Research*.

Steinbach, M., Karypis, G., and Kumar, V. (2000, August): A comparison of document clustering techniques. In KDD workshop on text mining, Vol. 400: 525-526.

Uyar, E. (2009): *Near-duplicate news detection using named entities* (Doctoral dissertation, BILKENT UNIVERSITY).