Approximate document outlier detection using Random Spectral Projection

Mazin Aouf and Laurence A. F. Park

School of Computing, Engineering and Mathematics,
University of Western Sydney, Australia
{mazin,lapark}@scem.uws.edu.au

Abstract. Outlier detection is an important process for text document collections, but as the collection grows, the detection process becomes a computationally expensive task. Random projection has shown to provide a good fast approximation of sparse data, such as document vectors, for outlier detection. The random samples of Fourier and cosine spectrum have shown to provide good approximations of sparse data when performing document clustering. In this article, we investigate the utility of using these random Fourier and cosine spectral projections for document outlier detection. We show that random samples of the Fourier spectrum for outlier detection provides better accuracy and requires less storage when compared with random projection. We also show that random samples of the cosine spectrum for outlier detection provides similar accuracy and computational time when compared with random projection, but requires much less storage.

1 Introduction

To perform outlier detection is to examine a data set for items that are dissimilar to the majority of the set. Outlier detection has been used for tasks such as computer network intrusion detection, medical fraud detection and credit card fraud detection, and novelty detection. It is also useful in finding clusters in highly imbalanced data sets.

The use of computers to automate tasks and communicate through the Internet, has led to the generation and storage of large amounts of information that must be processed for each outlier detection task. Therefore, we must be able to efficiently and effectively perform outlier detection on a large scale. Most recorded human interaction is in the form of free text (e.g. email, wikis, blogs, social network posts), therefore identifying outliers in large text document collections is a relevant problem that is useful for finding interesting items or suspicious documents that do not belong.

The definition of outlier detection implies that the data must be thoroughly examined to find the small set of outliers. Therefore, outlier detection is a computationally expensive task. It has been recently shown [1] that random projection can be used to project sparse data sets, such as text documents, into a lower dimensional space and approximately preserve the distances between all of the data objects. It has also been shown [2] that random samples of the Fourier and cosine spectrum provide us with a good lower dimensional approximation of sparse data sets for effectively and efficiently identifying clusters. In this article, we will investigate the utility (in terms of accuracy, efficiency and storage required) of these random spectral projections for outlier detection on large text document collections.

The contributions of this article are:

- A description of random spectral projection using the Fourier and cosine transforms (Section 3).
- A comparison of the speed and accuracy of random spectral projection and random projection for outlier detection (Section 4.4).
- An examination of the storage, speed and accuracy of random spectral projection and random projection when performing outlier detection on a large document set (Section 4.6).

The article will proceed as follows: Section 2 describes the current methods used for text document outlier detection. Section 3 describes the theory behind compressive sampling and shows how we will apply this to outlier detection. Section 4 contains the experimental method, results and discussion.

2 Text Document Outlier Detection

In this section we will examine how to perform outlier detection on a collection of text documents. Text document sets exist in high dimensional spaces, therefore, we require a simple method for detecting outliers. We first present the outlier detection method that we will be using and then describe how we will compute the similarity of each document during the outlier detection process.

2.1 Outlier Detection

An outlier is an element of a set that has different qualities in some respect to the majority of the set. Identifying an outlier may be subjective and therefore requires a clear definition in order for detection to take place. In this article, we are focusing on text documents, so an outlier document is one that is written on a different topic to the majority of the document collection. We will be representing each text document as a vector in a high dimensional space, implying that we need an efficient and effective outlier detection method that can be used on high dimensional vector spaces.

A simple outlier detection method examines the similarity of each document to its neighbours [3]. If a document is similar to many documents then it is considered an inlier, if a document is similar to only a few other documents, then it may be an outlier. This simple distance based outlier detection method has order $O(TN^2)$ where N is the number of documents in the collection and T is the dimensionality of the document space. The computational complexity comes from us having to compute the distance of each document from a given document to obtain its outlier likelihood score. An advancement on the distance based method is to also examine the density of the document distributions. If a region in the vector space is densely populated, while another is more spread out, the simple method may wrongly detect a document in the latter region of the vector space as an outlier. Local Outlier Factor (LOF) [4] is an outlier detection method that examines the local data distribution, and computes the outlier likelihood score based on the density around the point. Unfortunately the LOF method requires us to store the identity of each of document's k neighbours and the distances to these neighbours. The computation of the LOF scores requires a scan of the complete set of document vectors to compute the neighbours, having order $O(TN^2)$, then a scan of each point's neighbours to compute the document density, having order $O(Nk)$, and a final scan of the density scores and neighbours, also having order $O(Nk)$, where k is the number of neighbours chosen.

We are not investigating the accuracy of the outlier detection method itself, but we are interested in how the choice of random projection type affects the accuracy of the outlier detection. To simplify our experiments so that they focus on the effect of the projection, we will use simple distance based outlier detection.

2.2 Comparing Documents

A text document is a sequence of terms, where the terms describe the content of the document. Each document can be represented as a vector in a vector space, where the vector space has one dimension for each unique term in the document collection. Within this T dimensional space (where T is the number of unique terms in the collection), we can construct a document vector by using the frequency of each term in a document as the corresponding value of each element in the document vector. Doing this, we have:

$$\boldsymbol{d} = \left[f_{d,t_0}\ f_{d,t_1}\ \ldots\ f_{d,t_{T-1}}\right]$$

where f_{d,t_j} is the frequency of term t_j in document d. There has been many similarity functions developed to compare document vectors to queries for information retrieval (vector space methods [5], probabilistic methods [6], language models [7]), but when comparing documents to documents, it has been found that the TF-IDF weighting with cosine similarity is the most appropriate [8]. The TF-IDF weighting we use in this article is of the form:

$$w_{d,t} = w(f_{d,t}) = \log\left(\frac{N}{f_t} + 1\right) f_{d,t}$$

where $w_{d,t}$ is the weight of term t in document d, f_t is the number of documents term t appears in, and N is the number of documents in the collection. We can see that if the term t is common, meaning that f_t is large, then $\frac{N}{f_t}$ will be close to 1 and $\log\left(\frac{N}{f_t} + 1\right)$ will be close to $\log(2) = 0.6931$. If term t is rare, meaning that f_t is small, then $\frac{N}{f_t}$ will be close to N and $\log\left(\frac{N}{f_t} + 1\right)$ will be close to $\log(N+1)$. Therefore, the TF-IDF weighting gives less weight to common terms and more weight to rare terms that define the document. The weighted document vector is given as:

$$\boldsymbol{\delta} = \left[w_{d,t_0}\ w_{d,t_1}\ \ldots\ w_{d,t_{T-1}}\right]$$

The document similarity function is given as:

$$S(d_i, d_j) = \frac{\boldsymbol{\delta}_i \cdot \boldsymbol{\delta}_j}{\|\boldsymbol{\delta}_i\| \|\boldsymbol{\delta}_j\|}$$

where $\boldsymbol{\delta}_i$ is the ith weighted document vector, the inner product $\boldsymbol{\delta}_i \cdot \boldsymbol{\delta}_j = \sum_{k=i}^{T} w_{d_i,t_k} w_{d_j,t_k}$ and the vector norm $\|\boldsymbol{\delta}_i\|$ is $\sqrt{\boldsymbol{\delta}_i \cdot \boldsymbol{\delta}_i}$. We can see that this document similarity function measures the cosine of the angle between the weighted document vectors. If both vectors are the same, the similarity is 1; as the documents become more different, the similarity approaches 0. Note that also the denominator of the similarity function normalises the document lengths, meaning that the weight of a word in a long document will be less than the weight of the same word in a smaller document.

2.3 Random Projection

Outlier detection is dependent on the vector space dimensionality. As the dimensionality of the space increases, so does the time required to compute the outliers. Methods of dimension reduction (such as PCA, NMF and PLSA [9–11]) can be used to map the vector space into a smaller space, where each vector in the smaller space is an approximation of the vectors in the original space. Unfortunately, these methods are computationally expensive and therefore many not be feasible for high dimensional spaces. In this section we will examine a simple method that has been used for dimension reduction as a preprocessing step for outlier detection.

Random projection is the act of projecting a vector space to a lower dimensional space using a randomly generated mapping. It has been shown [1] that a random projection of sparse data that approximately preserves the similarity between vectors can computed using the mapping where each element is sampled from a random variable X having the distribution:

x	-1	0	1
$P(X = x)$	1/6	2/3	1/6

To map the document vector space from a T dimensional space to an S dimensional space, we generate an $S \times T$ matrix P_{R} containing values randomly sampled from X. We can then project the weighted document matrix D (containing the weighted document vectors as its columns) using:

$$D_{\mathrm{R}} = P_{\mathrm{R}} D$$

where D_{R} is an $S \times N$ matrix containing the projected document vectors as columns.

3 Random spectral projection

Compressive sampling [12, 13] is a sampling and reconstruction theory that has popularity in the image processing field [14] but has found its way into machine learning [2]. The idea is that if we are able to represent our data in a sparse vector space though the linear transformation Ψ^{-1}, then we are able to spread the information in our data set throughout the dimensions of the vector space using an linear transformation Φ that is maximally incoherent to Ψ. Using this knowledge, we are able to sample at a rate less than given by the Nyquist theorem and still be able to perfectly reconstruct the original signal from the sample. In our case, we have:

$$\boldsymbol{x} = \min \|z\|_1, \text{ s.t. } \boldsymbol{\xi} = \Phi\Psi\boldsymbol{z}$$

where $\boldsymbol{d} = \Psi\boldsymbol{x}$ is our document vector, $\boldsymbol{x}$ is sparse, Φ is the sampling function, $\boldsymbol{\xi}$ is the sample of $\boldsymbol{d}$, and $\|\cdot\|_1$ is the l_1 norm. The coherence of a pair of basis vectors is a measure of how similar they are. Coherence is given as:

$$\mu(\Phi, \Psi) = \sqrt{N} \max_{1 \le i,j \le N} |\langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_j \rangle|$$

where $\boldsymbol{\phi}_i$ and $\boldsymbol{\psi}_j$ are basis vectors of the linear functions Φ and Ψ respectively, and $\mu(\Phi, \Psi) \in [1, \sqrt{N}]$. Therefore if two isometric transformations are maximally incoherent, $\mu(\Phi, \Psi) = 1$.

For outlier detection, we do not need to reconstruct the document vectors, but we do require a method of projecting most of the information in the document collection into a smaller vector space. By performing the projection, we are able to reduce the computation time required, and by preserving most of the data, we will be able to maintain the accuracy of the outlier detection method.

In this article, our data consists of document vectors, which are sparse, therefore, the transformation Ψ^{-1} required to map our document vectors to a sparse set of vectors is simply the identity matrix; we choose $\Psi = I$. Our choice of Φ must be maximally incoherent to $\Psi = I$, therefore, we use $\Phi = \Sigma P_{\text{DFT}}$, where P_{DFT} is the discrete Fourier transform projection, and the matrix Σ selects a sample of s rows from P_{DFT}.

In this section, we will examine the discrete Fourier transform and the discrete cosine transform (a real approximation to the discrete Fourier transform) and how sampling is performed.

3.1 Discrete Fourier Transform

The Discrete Fourier Transform (DFT) is an isometric transformation that decomposes a vector into its various frequency components. The DFT used in this article has the form:

$$W_{d,f} = \sum_{t=0}^{T-1} w_{d,t} \exp\left(\frac{-2i\pi f t}{T}\right)$$

where $i = \sqrt{-1}$ and $W_{d,f}$ is the fth Fourier coefficient of document d. The DFT can also be given as a matrix multiplication:

$$D_{\text{DFT}} = P_{\text{DFT}} D$$

where P_{DFT} contains the elements $\exp(-2i\pi ft/T)$ and D is the matrix containing the set of weighted document vectors as columns. To sample from the transformed vector, we randomly take s coefficients of the DFT. This can be accomplished by randomly selecting s rows of P_{DFT} and the applying the transformation. The sample vectors (the columns of D_{DFT}) now become our representation of the document d to compute outliers from.

The sample vector contains complex elements, therefore we must ensure that the document similarity function $S(d_i, d_j)$ is modified to reflect this. We must ensure that the we take the complex conjugate of one of the document vectors being compared (for use in the inner product), and that we take the real portion of the inner product.

3.2 Discrete Cosine Transform

Use of the Fourier transform requires us to work in the complex domain, in which tools may not be readily available to many readers. Therefore, we will also examine the discrete cosine transform (DCT), which is a close approximation to the DFT, but its coefficients are real. The DCT used in this article has the form:

$$W_{d,c} = \begin{cases} \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} w_{d,t} \cos\left[\left(t + \frac{1}{2}\right) \frac{c\pi}{T}\right] & \text{if } c = 0 \\ \sqrt{\frac{2}{T}} \sum_{t=0}^{T-1} w_{d,t} \cos\left[\left(t + \frac{1}{2}\right) \frac{c\pi}{T}\right] & \text{if } c \neq 0 \end{cases}$$

where $W_{d,c}$ is the cth cosine coefficient, The DCT can be represented as a matrix multiplication, in the same fashion as the DFT. We also randomly sample s coefficients from the DCT spectrum to use as our document vector representation in the s dimensional space.

4 Experiments

We will now compare the effects, in terms of accuracy, time and storage required, for outlier detection when using random projection and the DFT and DCT random projections.

4.1 Experimental Environment

Our initial experimental environment consisted of a document set with ten artificially inserted documents from another document collection. For this we used two document sets from the SMART collection[1]. The first document set we used contained all documents from the CRAN document set (aerodynamics documents) and the first ten documents from the MED document set (medical abstracts), giving us 1408 documents and 4589 unique terms. We called this document set CRAN+10MED. The second document set we used contained all of the documents from the CISI document set (information science abstracts) and the first ten documents from the CRAN document set, giving us 1470 documents containing 5676 unique terms. We called this collection CISI+10CRAN. Note that each of the documents sets were parsed and converted into matrices using the

[1] Available from: `ftp://ftp.cs.cornell.edu/pub/smart`

textIR indexing software[2]. This software removes a predefined set of stopwords and performs stemming using the Lovins stemmer.

4.2 Evaluation

Rather than examine the number of outliers predicted correctly, we instead examined the likelihood of a document being an outlier, allowing us to obtain a finer scale for accuracy measurement. Our experiments involved ranking all points in terms of their distance based outlier likelihood score. The ranked lists were then examined and the ranks of the true outliers were found. A method that provides all of the outliers higher in the ranked list is evaluated as being more accurate than a method where the outliers are found in the lower ranks of the ranked list.

To evaluate each outlier detection method's outlier ranked list, we used Average Precision. Average precision uses r, the rank of each true outlier in the outlier ranked list, where r_i is the rank of the ith outlier provided by the method under evaluation, and r_i is ordered from highest rank to lowest rank. For example, if $r_3 = 10$, it means that the outlier detection method ranked an outlier as the 10th most likely outlier, and two other outliers were ranked somewhere from rank 1 to 9. Average precision is defined as:

$$\mathrm{AP}(r) = \frac{1}{O}\sum_{i=1}^{O}\frac{i}{r_i}$$

where O is the number of outliers. If there were three outliers and an outlier detection method ranked the outliers in positions $r = \{1, 3, 6\}$, the Average Precision would be $(1/1 + 2/3 + 3/6)/3 = 0.72$. We can see that if all of the outliers were ranked above all non-outliers, the Average Precision would be 1.

4.3 Procedure

The outlier likelihood score computation requires the parameter k (the number of neighbours to consider). We computed outlier likelihood scores using neighbour distance of $k = 10, 20, 50, 100, 200$ and 500. Each of the sampling methods requires the parameter s (the number of coefficients to sample). We computed outlier likelihood scores using sample sizes $s = 32, 64, 128, 256, 512, 1024$ and 2048. Each trial of the experiment involves random sampling, therefore we repeated each trial ten times to take into account the variability. The mean speed and average precision of the ten trials is reported as the expected value.

4.4 Results

Analysis of the results showed us that similar trends were displayed for each value of k, therefore we will only present the results for $k = 100$ but use the entire set of results in later significance tests. Figures 1 and 2 present the mean AP versus computation time plots for the CISI+10CRAN and CRAN+10MED

[2] Available from: `http://staff.scm.uws.edu.au/~lapark/textIR/`

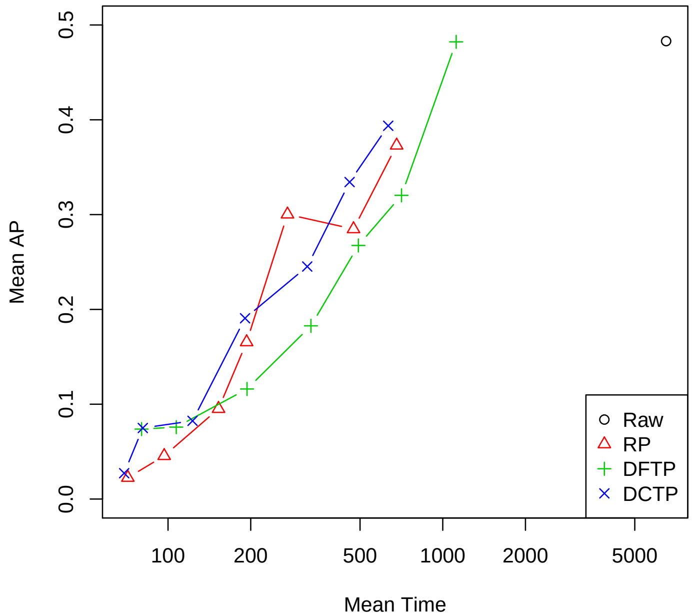

Fig. 1. The mean AP and computation time for outlier detection in the original space (Raw), and for random projection (RP), DFT projection (DFTP), and DCT projection (DCTP) for $k = 100$ on the CISI+10CRAN data set. The results for each projection method have been recorded for projections into 32, 64, 128, 256, 512, 1024 and 2048 dimensions. Note that the time axis is presented in a log scale.

document sets respectively. It should be noted that the plots show that as s increases, so does the computation time. For the majority of the cases, we see that the mean AP increases as s increases. The Raw result, is the AP obtained when no projection is used. If we examine the vertical alignment of each of the plots, we see that the RP and DCTP times closely match, meaning that for a given s, we expect them to complete in the same time. If we examine the times for each DFTP point, we see that they almost match the times of the RP and DCTP methods for the previous s values. Therefore the DFTP method is slower than the DCT and RP; this is likely to be due to the DFTP method producing complex values.

We can compare the mean AP for a given s by examining the horizontal alignment of the points, It can be seen that the DCTP mean AP is greater than RP for Figure 1, but less for Figure 2. We can also see that the mean AP of DFTP for a given s is greater than RP for both plots.

4.5 Significance of results

We have generated 10 results for each s,k pair for each method. To test the hypothesis that the DFT and DCT spectral random projection provides a greater

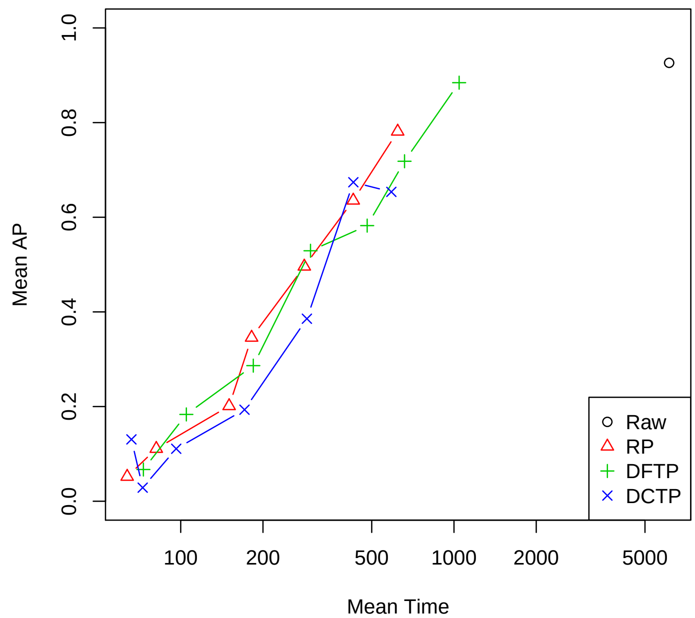

Fig. 2. The mean AP and computation time for outlier detection in the original space (Raw), and for random projection (RP), DFT projection (DFTP), and DCT projection (DCTP) for $k = 100$ on the CRAN+10MED data set. The results for each projection method have been recorded for projections into 32, 64, 128, 256, 512, 1024 and 2048 dimensions. Note that the time axis is presented in a log scale.

mean AP for a given s,k pair when compared to random projection, we used bootstrap sampling to generate the distribution of the increase in the mean AP of the DFTP and DCTP outlier detection over RP outlier detection. We can test for a significant increase in mean AP for each of the methods, using the bootstrap distribution. The resulting distributions were approximately Normal for each method on each document set. We obtain p values of 0.060 and 0.0001 for the DFT projection on the CISI+10CRAN and CRAN+10MED collections respectively. This shows that the DFT random spectral projection provides a significant increase in mean AP over random projection. For DCT projection, we obtained p values of 0.398 and 0.995 on the CISI+10CRAN and CRAN+10MED collections respectively. This implies that we have no evidence of the DCT projection providing a significant increase in mean AP over random projection.

From these results, we can see that using DFTP produced more accurate results than when using RP for both document sets. But we must ask what caused the change in accuracy of the DCTP method across document sets? If we examine the Raw scores for each document set, we see 0.48 for the CISI+10CRAN set and 0.92 for the CRAN+10MED set, meaning that the outliers had more separation in the CRAN+10MED set, but were harder to find in the CISI+10CRAN set. From this, we can make the hypothesis that the separation of the RP and

Method	Memory		Time	
	Function	Projection	Projection	Outlier
Raw	NA	NA	NA	3.4 years
RP	321.4 MiB	222.3 MiB	16.6 sec	2.6 days
DCTP	NA	222.3 MiB	38.2 min	2.4 days
DFTP	NA	444.7 MiB	38.6 min	4.4 days

Table 1. The memory used for the projection function (Function) and the projected matrix (Projection), and the time taken to perform the projection (Projection) and compute the outliers (Outlier) when using no projection (Raw), random projection (RP), DCT projection (DCTP), and DFT projection (DFTP) on the ZIFF+10FR document collection. Note that the Raw completion time of 3.4 years was extrapolated.

DCTP methods grow as the difficulty in detecting the outliers increases. Testing this hypothesis will be left for future work.

4.6 Large Document Collection

To examine the benefit of using the DFTP and DCTP methods on large documents sets, we combined the 56920 articles from the TREC[3] Disk 2 Ziff Publishing collection (computing articles) and the first 10 newspaper articles from the TREC Disk 2 Financial Review collection (finance articles). This formed a document collection with 56930 documents and 82295 unique terms; we call this document set ZIFF+10FR. Using a sparse matrix format, we were able to store the term frequency matrix in 83.93 megabytes.

We then ran the outlier detection method on the weighted term frequency matrix, and on random projections of the matrix using RP, DFTP and DCTP. Each method projected the matrix to $s = 512$ sample features and we computed the $k = 5$, 10, 20, 50, 100, 200 and 500 neighbour distances. The memory consumption and timing results are presented in Table 1, while the accuracy results are in Table 2.

We can see that even though the random projection and both spectral random projections required more storage than the sparse raw data, the computation time for the projected data was much faster, allowing us to obtain results. DCTP required the least storage and computation time, while RP required the most storage (due to it needing storage for its projection matrix), and DFTP required the most time (due to it using complex numbers). The accuracy results show us that the outlier detection task was difficult (shown by the low AP scores). We can see that for each k the DFTP AP is the greatest (for some values of k it is at least 10 times greater than the RP and DCTP AP). We can also see that the DCTP AP is either greater than or equal to the RP AP.

[3] `http://trec.nist.gov/`

Method	AP						
	5	10	20	50	100	200	500
RP	0.0001	0.0016	0.0029	0.0027	0.0022	0.0017	0.0011
DCTP	0.0001	0.0016	0.0036	0.0034	0.0025	0.0019	0.0013
DFTP	0.0001	0.0021	0.0107	0.0209	0.0303	0.0290	0.0196

Table 2. Accuracy results for outlier detection on each projection method using the ZIFF+10FR document collection. The average precision (AP) for k = 5, 10, 20, 50, 100, 200 and 500 is shown.

4.7 Discussion

This work shows the potential of the DFT and DCT projections when compared to random projection, but is only one component in high dimensional outlier detection. There have been other methods proposed in the literature to increase the speed of distance based outlier detection, but they were not used in this analysis. We kept this analysis as simple as possible, to examine the effect of the sample projection only.

The findings from this work can be combined with existing methods of approximation to increase speed. For example, neighbour distances can be computed on a random sample of documents rather than the whole set [15]. This approximation can also be applied to increase the speed of a DFT or DCT projection. LOF [4] can be used in place of the distance based outlier detection method. Methods of increasing the speed of distance based outlier detection method are given in [3]. These may be directly applied to the projected document space to further reduce the computation time of the DFTP and DCTP methods.

Rather than using random projection to compute the distances, it can be used to select a candidate set of neighbours [16]. The distances to all candidates may then be computed in the original space. The analysis in this article has shown that the DCT or DFT can also be used in place of random projection in this case.

5 Conclusion

Outlier detection is an important process that allows us to automatically identify objects that are different from the majority of the data. When examining text, we can use outlier detection to identify documents that are interesting or out of place. Random projection is used in outlier detection to obtain a lower dimensional approximation of the data, in order to speed up the detection process.

In this article, we presented the concept of spectral random projection, using the discrete Fourier transform (DFT) and the discrete cosine transform (DCT), and we examined its utility for outlier detection on a text document collection.

We showed that using the DFT projection provides significantly greater accuracy than when using random projection and requires less storage, but requires additional time. We also showed that the DCT projection provides similar accuracy and computation time as random projection, but it requires much less storage (60% less for the large document set).

References

1. Deegalla, S., Bostrom, H.: Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In: Proceedings of the 5th International Conference on Machine Learning and Applications. ICMLA '06, Washington, DC, USA, IEEE Computer Society (2006) 245–250
2. Park, L.A.F.: Fast approximate text document clustering using compressive sampling. In Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M., eds.: Machine Learning and Knowledge Discovery in Databases. Volume 6912 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2011) 565–580
3. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. The VLDB Journal **8**(3-4) (February 2000) 237–253
4. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. SIGMOD '00, New York, NY, USA, ACM (2000) 93–104
5. Zobel, J., Moffat, A.: Exploring the similarity space. ACM SIGIR Forum **32**(1) (Spring 1998) 18–34
6. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments, part 2. Information Processing and Management **36**(6) (2000) 809—840
7. Song, F., Croft, W.B.: A general language model for information retrieval. In: CIKM '99: Proceedings of the eighth international conference on Information and knowledge management, ACM Press (1999) 316–321
8. Park, L.A.F., Ramamohanarao, K., Leckie, C.A., Bezdek, J.C.: Adapting spectral co-clustering to documents and words using latent semantic analysis. In: AI 2009: Advances in Artificial Intelligence. Volume 5866 of Lecture Notes in Computer Science., Springer Berlin / Heidelberg (December 2009) 301–311
9. Ding, C., He, X., Simon, H.D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proc. SIAM Int'l Conf. Data Mining (SDM'05). (April 2005) 606 – 610
10. Park, L.A.F., Ramamohanarao, K.: An analysis of latent semantic term self-correlation. ACM Transactions on Information Systems **27**(2) (2009) 1–35
11. Park, L.A.F., Ramamohanarao, K.: Efficient storage and retrieval of probabilistic latent semantic information for information retrieval. The International Journal on Very Large Data Bases **18**(1) (January 2009) 141–156
12. Candes, E., Wakin, M.: An introduction to compressive sampling. Signal Processing Magazine, IEEE **25**(2) (3 2008) 21–30
13. Candes, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid, Spain, European Mathematical Society (2006)
14. Goyal, V., Fletcher, A., Rangan, S.: Compressive sampling and lossy compression. Signal Processing Magazine, IEEE **25**(2) (3 2008) 48–56
15. Wu, M., Jermaine, C.: Outlier detection by sampling with accuracy guarantees. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '06, New York, NY, USA, ACM (2006) 767–772
16. de Vries, T., Chawla, S., Houle, M.: Density-preserving projections for large-scale local anomaly detection. Knowledge and Information Systems **32** (2012) 25–52 10.1007/s10115-011-0430-4.