Volume 3, No. 8, August 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

Impact of Similarity Measures on Causal Relation Based Feature Selection Method for Clustering Maritime Accident Reports

Santosh Tirunagari[*1], Maria Hanninen[2], Guggilla Abhishek[3], Kaarle Stahlberg[2], and Pentti Kujala[2]

*[1]Department of Information and Computer Science, Aalto University, School of Science, Otaniemi, Espoo, Finland
Santosh.tirunagari@aalto.fi[1]
[2]Department of Applied Mechanics, Aalto University, School of Engineering, Tietotie – 2, Otakaari, Espoo, Finland
(maria.hanninen, kaarle.stahlberg, pentti.kujala) @aalto.fi[2]

[3]Department of Computer Science, Teegala Krishna Reddy Engineering College, Saroor Nagar, Hyderabad, 500079
guggillaabhishek@gmail.com[3]

Abstract: Unsupervised document clustering is an automated process in which documents are analyzed based on their similarity. In this paper, we propose a new feature selection method based on causal relations to classify maritime accident reports in unsupervised manner. We also compare the impact of different similarity measures on proposed feature selection method. Based on the analysis, we conclude that the proposed feature selection method has better performance over the conventional method due to the effect of dimensionality curse. The impact of similarity measures improves with the proposed feature selection method. In the analysis, we have compared Correlation, Cosine, Spearman, Bray-Curtis, Euclidean, City-block, Squared-Euclidean, Standardized Euclidean, and, Chebychev similarity measures. The first two produced the best results, followed by the next two. The rest did not produce good results with the maritime accident reports used in our analysis. Interestingly Chi-Square gave good results with proposed method in our analysis.

INTRODUCTION

Clustering is an unsupervised learning approach that forms groups of documents such that the documents which are similar will fall in the same cluster, while the documents that are different are separated a part to different clusters [5]. The similarity between any two documents is measured by similarity measure or metrics. Many similarity metrics have been proposed and are widely used, including Euclidean, Cosine, Correlational, Spearman, Chebychev, etc. Each similarity metric has its own characteristics which makes it suitable for some problem and less suitable for some others. In this paper, we take document clustering as a problem and analyze the performance of the K-means clustering over a number of popular similarity metrics.

The importance of considering distance measures becomes particularly important in the case of very sparse high-dimensional data, related to the curse of dimensionality. It has been shown that this phenomenon has an impact on various techniques for classification (including k-NN classifiers) and clustering [1]. It also effects information retrieval results [2]. Hence in our experiments, we propose causal relation based feature selection method as the dimensionality reduction method and analyze the performance of the similarity metrics over the proposed and conventional method. In the following, we describe the methods and data used in our experiments as well as the results of the experiments and their interpretation. Even though we aim at a high degree of generalizability, it is to be noted that the results are subject to the data used and the preprocessing through which the texts are transformed into numerical vectors.

The rest of the paper consists of three more sections. Section 2 describes the methods we have used in our experiments that are similarity measures, proposed feature selection method and K-means clustering. Section 3 explains the data and experiments. Section 4 and Section 5 provides the results and conclusions respectively.

METHODS

In the following, we will describe the similarity measures, evaluation methods and our proposed algorithm.

Similarity Metrics

In the implementation of the clustering algorithm, the similarity measures must be predetermined. This measure shows the degree of closeness or separation of the data objects [5]. The performance of the clustering algorithm depends on choosing a good similarity measure over the input data. Universally there is no solid similarity metric for a particular type of clustering. Not all distance or similarity measures is a metric, a measure d is said to be a similarity metric [1] if it satisfies all the four properties namely non-negativity, isolation, symmetry and triangular-inequality.

non-negativity: The distance between any two points must be non negative. $d(i, j) \geq 0$

isolation: The distance between two points is zero if and only if the points are identical. $d(i, j) = 0$ iff $x = y$.

[1] http://kochanski. org/gpk

symmetry: The distance between the points . i and j must be equal to the distance between j and i . that is $d(i, j) = d(j, i)$

triangular-Inequality: Sum of distances from two distinct points must be greater than the distance from third point. That is $d(i, j) \le d(i, h) + d(h, j)$

In this paper, we use a number of similarity measures as given in Table 1. Let x_i and x_j be two row vectors of a data set X with n dimensions.

Table I. Similarity Metrics.

Metrics	Equation
Euclidean	$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{k,i} - x_{k,j})^2}$
Squared Euclidean	$d_{ij} = \sum_{k=1}^{n}(x_{k,i} - x_{k,j})^2$
Standardized Euclidean	$d_{ij}^2 = \sum_{k=1}^{n}(x_{k,i} - x_{k,j})V^{-1}(x_{k,i} - x_{k,j})$
Cosine	$d_{ij} = \left(1 - \frac{x_i . x_j}{\lvert x_i \rvert \times \lvert x_j \rvert}\right)$
Correlational	$d_{ij} = \left(1 - \frac{(x_i - \bar{(x_i)}).(x_j - \bar{(x_j)})}{\lvert (x_i - \bar{(x_i)}) \rvert * \lvert (x_j - \bar{(x_j)}) \rvert}\right)$
Spearman	$d_{ij} = 1 - \frac{(r_i - \bar{r}_i)(r_j - \bar{r}_j)}{\sqrt{(r_i - \bar{r}_i)(r_i - \bar{r}_i)'}\sqrt{(r_j - \bar{r}_j)(r_j - \bar{r}_j)'}}$
Bray Curtis	$d_{ij} = \frac{\sum \lvert x_i - x_j \rvert}{\sum \lvert x_i + x_j \rvert_{ij}} = \frac{\sum \lvert x_i - x_j \rvert}{\sum \lvert x_i + x_j \rvert}$
Chi-square	$d_{ij} = \frac{\sum (x_i - x_j)^2}{\sum (x_i + x_j)}$
Chebychev	$d_{ij} = max_k \lvert x_{k,i} - x_{k,j} \rvert$
Cityblock	$d_{ij} = \sum_{k=1}^{n} \lvert x_{k,i} - x_{k,j} \rvert$

Here, For the standardized Euclidean V is the n -by- n diagonal matrix whose j^{th} diagonal element is $S_{(j)^2}$, where S is the vector of standard deviations. For Spearman r_i and r_j are the coordinate-wise rank vectors of x_i and x_j . $\bar{r}_i, \bar{r}_j = \frac{N+1}{2}$, where N is number of dimensions

K-means Algorithm

K-means is a widely used clustering algorithm because of its simplicity and high speed of execution on larger datasets and is briefly described in [6]. It is a partition based clustering with the main goal of partitioning n objects in to a set of K clusters, where each object belong to exactly one cluster. The number of clusters k is given in advance. Given a set X of n points in a d - dimensional space and an integer k . The algorithm chooses a set of k points $c_1, c_2, \ldots, c_k$ in the d -dimensional space to form clusters $C_1, C_2, \ldots, C_k$ such that the Cost (C) is minimised.

$$Cost(C) = \sum_{i=1}^{k}\sum_{x \in c_i} L_2^2(x - c_i)$$

where x is the data object and c_i is the centroid or mean of the cluster i .

***Term Frequency* (TF):** TF is the total count of the particular word repeated in a document and is calculated as

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of times the term t_i occurs in document d_j and the denominator is the sum of number of times all terms occur in document d_j [4].

***Inverse Document Frequency* (IDF):** IDF is often described as a heuristicis [8], is the total number documents containing the term. It is calculated as the logarithmic value of the quotient when number of documents in the collection is divided by number of documents containing the term.

$$idf_i = \log \frac{\lvert D \rvert}{\lvert d : t_i \in d \rvert}$$

where $\lvert D \rvert$ is the total number of documents in the collection and $\lvert d : t_i \in d \rvert$ is the number of documents where the term t_i appears [4].

TF-IDF is the product of TF and IDF

$$tf - idf_{ij} = tf_{i,j} \times idf_i$$

Precision: precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall: Recall is given by

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-measure: F-measure is the harmnic mean of Precision and Recall

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Here, tp is true positives i.e correct result, fp is false positives i.e unexpected result, fn is false negatives i.e missing result.

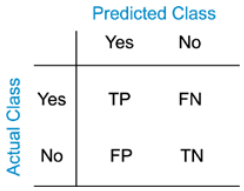

Figure 1: Confusion matrix showing tp, tn, fn, and, fp.

Normalized Mutual Information (NMI) : Bad clusterings have purity values close to 0, a perfect clustering has a purity of 1. High purity is easy to achieve when the number of clusters is large, purity is 1 if each document gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters.
A measure that allows us to make this tradeoff is normalized mutual information or NMI :

$$\mathrm{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2}$$

where I is mutual information and H is the entropy.

Causal Relations as Proposed method: NLTK-Python is used as a tool here. First the raw text is tagged with parts of speech using NLTK POS TAGGER, later this part of the code is used to extract the causal relations from the documents.

```
chunker = nltk.RegexpParser(r'''
NP:
{<DT|IN|JJ>?<NN|PRP><V.*><.*>+<NN|PRP><V.*>?}
''')
```

Where DT is determiner, NN is the noun forms, V is all kind of verbs, PRP is prepositions, JJ is adjectives. A accident report typically consists of 60 pages. The causal relations extracted from a report is maximum of 25 to 30 sentences. Hence a 60 page report is transformed to a half page text document, with all the important terms which caused the accident. Some of the example causal relations extracted from a 60 pages report, are as follows:

Cause 1: "It is probable that he had been drinking alcohol , which would have contributed to his fatigue."

Cause 2: "He was aware that Karin Schepers would sail that evening , but did not take the opportunity to rest during the day."

Cause 3: "In this accident , the master's alcohol consumption , possibly exacerbated by fatigue , resulted in him behaving in a manner that any junior watchkeeper would have found difficult , and which placed the safety of the vessel and crew, and the environment at risk."

Cause 4: "Once it was confirmed that the vessel was aground, all the crew should have been alerted by the sounding of the general alarm."

Cause 5: "if both vessels had then maintained their courses, Admiral Blake's closest point of approach (CPA) to Boxford would have been about cables on the container ship s starboard side."

Cause 6: "It is also possible that the inexperienced cadet only reported the light to the master when he could see it clearly."

These causal relations provide significant terms which caused the accident. These significant terms are enough for classifying the maritime accident reports. Hence the dimensionality of the TFIDF matrix is also reduced, improving the performance and time complexity of the K-means clustering algorithm.

DATA AND EXPERIMENTS

The document collection used in the experiment is 'MAIB accident reports' [2]. There are 11 categories of accidents. We concentrated on only 4 types of accidents with 135 documents. In the paper we focus these set documents as shown in the Table 2.

Table II. Document Collection.

Accident Type	Documents
Collisions	55
Groundings	44
Machinery	21
Fire	15
Total	**135**

As the preprocessing step, the document collection is parsed from pdf to text files. In the next step puntuation symbols are removed and the text is converted to lower case. we proceed with removing the stopwords and applying Porter's stemming [3]. After this process all the unique terms are collected. TF and IDF are calculated for unique terms. TFIDF matrix is obtained which is given as a input to K-means algorithm. This method is reffered to conventional method. Here, at this stage, we introduce our causal relations based feature selection method. We use NLTK-python code shown in Section 2, to extract causal relations. These causal relations are used to collect the unique terms and generate TFIDF matrix.

The number of unique terms in the conventional method were 5809. Whereas in applying proposed feature selection method, only 1076 unique terms were collected. The proposed method is not only able to reduce the dimensions but also extracts significant terms for unsupervised classification.

RESULTS

The TFIDF matrices obtained using conventional method and proposed method are given as an input to the K-means clustering algorithm with different similarity metrics and output is generated in the form of document ids with their corresponding cluster ids. As the document categories are

[2]http://www.maib.gov.uk/publications/investigation_reports/reports_by_incident.cfm

already known, F-measure, NMI, precision and recall are calculated. As there are 4 categories of the documents we gave 4 clusters as the input to K-means algorithm. Due to random assigning of centroids in K-means algorithm, the results obtained are different, every time the algorithm is run. Hence to maintain average evaluation, we ran algorithm for 50 times and average of the results were considered.

From the Figure 2, the conventional method produced a highest F-measure of 0.5 and lowest of 0.1, the highest NMI of 0.28 and lowest NMI of 0.012. The first four similarity metrics Cosine, Correlational, Bray-Curtis and Spearman produced good results over 0.4 F-measure and over 0.25 NMI.

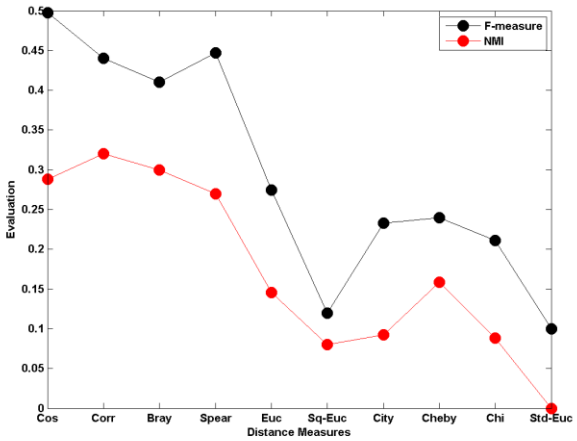

Figure 2: NMI and F-measure for different similarity measures on conventional method.

From Figure 3, it is observed that the Precision and Recall for the first four similarity metrics were found good when compared to the rest. The Precision for the first four similarity metrics was found over 0.5 and Recall was over 0.35.

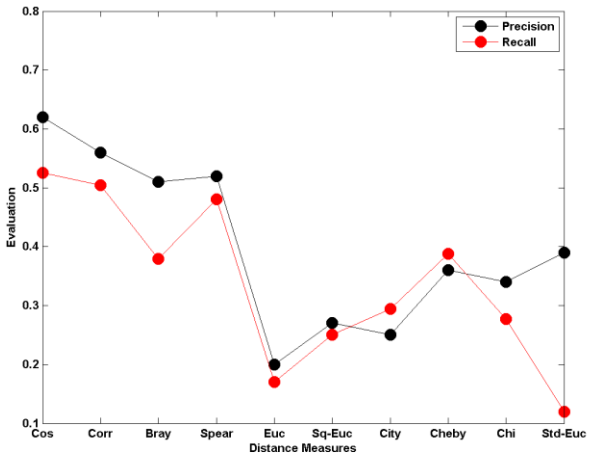

Figure 3: Precision and Recall for different similarity measures on conventional method

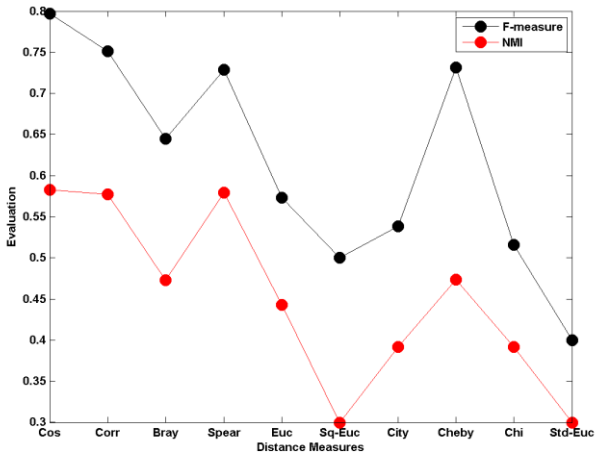

Figure 4: NMI and F-measure for different similarity measures on proposed method.

Figure 4, shows the improvement in scale for NMI and F-measure. The highest F-measure is found at 0.8 and lowest at 0.41. The highest NMI is found at 0.69 and lowest at 0.3. This shows that the proposed method is performing well with all the similarity metrics. The similarity metric Chi-Square performance was also improved with the proposed method. Due to more dimensions the similarity measures fail to locate the distances, which in turn fail the K-means clustering algorithm to place the data points into proper clusters.

Figure 5, shows the Precision and Recall. First four similarity metrics hold good when compared to the rest. The Precision for the first four similarity metrics was over 0.65 and Recall was over 0.45. This is better when compared to the conventional method. Also the similarity metric Chi-Square performance was improved with the proposed method.

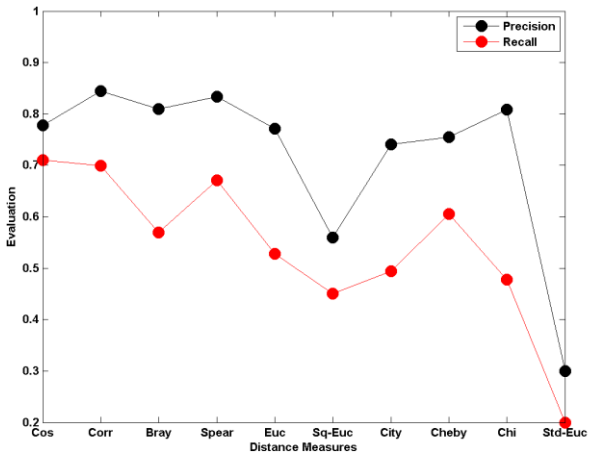

Figure 5: Precision and Recall for different similarity measures on proposed method

CONCLUSIONS

The main aim behind these experiments was to compare and analyze the performances of the similarity metrics for clustering text documents over conventional method and causal relation based feature selection method. We find that correlational similarity and cosine similarity metrics gave us good results and then followed by Spearman and

Bray-curtis for both the methods. The distance metrics like Euclidean, Cityblock, Squared-Euclidean, Standardized Euclidean and Chebychev did not produce good results with both the methods. Interestingly Chi-Square method produced good results with the proposed method. It is also shown that when the dimensions are reduced with causal relation based feature selection method the similarity metrics performance is improved. Based on these experiments in future this work would be carried out on finding normalization methods which can improve the performance of the other distance metrics. This work can be carried out for cross language information retrieval with larger document collection.

ACKNOWLEDGMENT

The study was conducted as a part of CAFE project, financed by the European Union - European Regional Development Fund - Regional Council of Päijät-Häme, the City of Kotka, Kotka-Hamina regional development company Cursor Ltd., Kotka Maritime Research Association Merikotka and the following members of the Kotka Maritime Research Centre Corporate Group: Port of Hamina Kotka, Port of Helsinki, Aker Arctic Technology Inc. and Arctia Shipping Ltd.

REFERENCES

[1] Radovanovi, Milo; Nanopoulos, Alexandros; Ivanovi, Mirjana (2010),"Hubs in space: Popular nearest neighbors in high-dimensional data", Journal of Machine Learning Research11: 2487–2531

[2] Radovanovi, Milo; Nanopoulos, Alexandros; Ivanovi, Mirjana (2010),"On the existence of obstinate results in vector space models", Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 186–193.

[3] M. F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130137.

[4] Taghva Kazem and Veni Rushikesh, Effects of Similarity Metrics on Document Clustering, Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations, IEEE Computer Society 2010 222–226.

[5] Anna Huang, "similarity Measures for Text Document Clustering", NZC-SRSC 2008, April 2008, Christ Church, New Zealand.

[6] An efficient k-means clustering algorithm: Analysis and implementation, T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002), 881–892.

[7] Indexing by Latent Semantic Analysis, Scott C. Deerwester and Susan T. Dumais and Thomas K. Landauer and George W. Furnas and Richard A. Harshman, JASIS, 41 (1990), 391–407.

[8] Stephen Robertson, (2004) "Understanding inverse document frequency: on theoretical arguments for IDF", Journal of Documentation, Vol. 60 Issn: 5, pp. 503–520

