Improving the Decision Value of Hierarchical Text Clustering Using Term Overlap Detection

Nilupulee Nathawitharana
La Trobe University
18585515@students.latrobe.edu.au

Damminda Alahakoon
La Trobe University
D.Alahakoon@latrobe.edu.au

Sumith Matharage
sumith.matharage@gmail.com

Abstract

Humans are used to expressing themselves with written language and language provides a medium with which we can describe our experiences in detail incorporating individuality. Even though documents provide a rich source of information, it becomes very difficult to identify, extract, summarize and search when vast amounts of documents are collected especially over time. Document clustering is a technique that has been widely used to group documents based on similarity of content represented by the words used. Once key groups are identified further drill down into sub-groupings is facilitated by the use of hierarchical clustering. Clustering and hierarchical clustering are very useful when applied to numerical and categorical data and cluster accuracy and purity measures exist to evaluate the outcomes of a clustering exercise. Although the same measures have been applied to text clustering, text clusters are based on words or terms which can be repeated across documents associated with different topics. Therefore text data cannot be considered as a direct 'coding' of a particular experience or situation in contrast to numerical and categorical data and term overlap is a very common characteristic in text clustering. In this paper we propose a new technique and methodology for term overlap capture from text documents, highlighting the different situations such overlap could signify and discuss why such understanding is important for obtaining value from text clustering. Experiments were conducted using a widely used text document collection where the proposed methodology allowed exploring the term diversity for a given document collection and obtain clusters with minimum term overlap.

Keywords: Term overlap, Growing Self Organizing Map, Hierarchical clustering, Text document clustering

1 Introduction

With the introduction of World Wide Web, an enormous amount of text data has become available. Also, the usage of electronic media including electronic format of text data has been increased in every aspect of human life due to easy storage and accessibility. Text mining has recently emerged as a popular research field to find hidden patterns in electronic text data. This research field has received attention of researchers due to availability of massive volumes of text and sophisticated text mining techniques are introduced to analyse text data in different domains (Khadjeh Nassirtoussi et al. 2014; Lu et al. 2013; Minanović et al. 2014). Text classification (grouping of text data into predefined classes) and text clustering (grouping of text data into similar groups without known class labels) are two main areas which fall under the umbrella of Text Mining. However, text clustering techniques are preferred over text classification techniques in most of the cases, due to lack of categorical information in most text data sets (Gunasinghe et al. 2012).

Document clustering is a widely used, useful technique to explore hidden patterns in a collection of text documents. During the document clustering process, documents are grouped into a number of clusters, in which documents belonging to the same cluster demonstrate a high degree of homogeneity and documents belonging to two different clusters demonstrate a

high degree of heterogeneity. Since documents need to be converted into a numerical format prior to clustering, they are represented as a vector of weighted terms, where terms can comprise of individual words or word sequences and the clustering process is carried out based on the detailed term similarities of the document vectors (Kohonen et al. 2000).

Humans group documents into different topics or categories based on the content relevancy to those topics. Due to semantic or syntactic necessity, words or word sequences can be repeated in documents grouped under different topics so it is not possible to expect a one to one mapping between the terms of the documents and the topics. For example, the word 'scheme' can be used in explaining housing schemes, pension schemes, recycling schemes, and colour schemes which are associated with different themes according to human categorization. Because of above reason traditional cluster evaluation measures become less important when clustering text documents and we cannot expect the documents which belong to the same human categorized group to be clustered together.

In general, documents which are associated with similar topics are expected to have significant term overlap and a low term overlap is expected in the documents which belong to very different topics. Despite this general belief there might be situations where documents describing vastly different topics share a considerable percentage of terms, as well as documents describing similar topics share a lesser number of terms. Since term frequencies are used for text document clustering, if same terms with similar frequencies exist within two documents there is very high possibility they will be clustered together regardless of the human categorized group they actually belong to. This highlights the importance of analysing term overlap since the term overlap between the documents have considerable impact in results we obtain by text documents clustering. In addition overlap calculation is highly beneficial to identify the term diversity of a given document set.

Different types of clustering algorithms exists which include hierarchical clustering, partitional clustering, fuzzy clustering, and artificial neural networks (Jain et al. 1999). Several most popular clustering algorithms are K-Means, Self-Organizing Map and SOM based clustering algorithms. Out of the many clustering techniques, Self-Organizing Map (SOM) (Kohonen 1998) based techniques are widely used by the data analysts (Kaski et al. 1998; Kohonen et al. 2000). The Growing Self Organizing Map (GSOM) (Alahakoon et al. 2000) is a prominent member of the SOM family and it has shown great potential in different types of clustering tasks. The GSOM algorithm has been used across diverse disciplines due to its capabilities and several applications of the GSOM can be found in Hsu et al. (2003), Amarasiri et al. (2005), Matharage et al. (2009), Ahmad et al. (2010), Matharage et al. (2011), and Gunasinghe et al. (2012).

In this paper a new methodology is proposed utilizing the unique feature of the GSOM algorithm to analyse a document set with the possibility of obtaining clusters with low term overlap using hierarchical clustering.

The rest of the paper is organized as follows. Section 2 provides a detailed explanation on hierarchical clustering as a term overlap detection technique. Section 3 describes the proposed methodology. Section 4 provides the experimental results and the section 5 conclude the paper.

2 Hierarchical Clustering as a Term Overlap Detection Technique

2.1 Hierarchical Clustering of Text

Hierarchies are a commonly used data structure in data mining and knowledge discovery (Schkolnick 1977). Hierarchical structure facilitates visualizations of data in different granularity levels which let the data analyst to obtain more detailed relationships by initially exploring the abstract relationships. Some of the previous research work have revealed that documents in a collection naturally lead themselves to a hierarchical structure which makes it a suitable data structure to analyse textual data (Merkl 1998).

2.2 The SOM and the GSOM Algorithms

The Self-Organizing Map (SOM) algorithm was introduced by Kohonen (1982) and is capable in mapping high dimensional data distribution to a lower dimensional output, mostly to a two dimensional output space. Since the SOM algorithm is capable in visualizing patterns in input data it is widely used as a clustering technique. Some of the application of the SOM algorithm for text clustering can be found in Kaski et al. (1998), Merkl and Rauber (1999), and Kohonen et al. (2000). However there are certain limitations associated with the SOM algorithm and the main limitation is size of the output network need to be predefined. When clustering a dataset, output map size cannot be defined correctly prior to the clustering as there is no knowledge about the inputting data. In that case there might be a need of running the SOM algorithm for different map sizes and select an appropriate map size. Some of the other limitations include lack of hierarchical clustering capabilities and possibility of learning only from stationary data.

The Growing Self-Organizing Map (GSOM) algorithm (Alahakoon et al. 2000) which has the ability to grow dynamically has overcome the predefined map size limitation in the SOM. Similar to most of the SOM based algorithms, the GSOM also consists with two activation modes namely training mode and testing mode. Actual network growth and smoothing out of weights occur in training mode, while the final calibration of the network for known inputs take place in the testing mode. Several key characteristics and advantages of the GSOM algorithm include structurally unconstrained learning, exploratory dynamics, efficient computing, and visualization (Matharage 2012) which makes it a highly suitable algorithm for text clustering. Matharage (2012) has investigated the usability of the GSOM algorithm as a text clustering algorithm and concluded that when compared to the SOM algorithm, the GSOM algorithm is more efficient, more accurate and has better topology preservation capabilities. The details of the GSOM algorithm is provided in Alahakoon et al. (2000).

2.3 The GSOM with Spread Factor (SF) for Hierarchical Text Clustering

2.3.1 The GSOM with Spread Factor (SF)

In the GSOM, hierarchical clustering is facilitated by the parameter Spread Factor (SF) which controls the growth of the network. When a low SF value is used, the GSOM output becomes a more abstract map whereas more detailed map is given when a high value for SF is used. This characteristic of the GSOM algorithm is very advantageous in data mining since it is possible to obtain an abstract map initially and then further explore the map using a higher SF. Some of the examples of using hierarchical clustering facilitated by the GSOM includes Alahakoon et al. (2000) for the zoo data set (Bache and Lichman 2013) and in Haiying et al. (2004) for the sleep apnea data set (Goldberger et al. 2000).

2.3.2 Automatic Cluster Identification in the GSOM

In general clusters in a GSOM output can be identified by human inspection. When cluster identification is carried out by humans, clusters are identified as a group of hit nodes which are separated by dummy nodes or non-hit nodes. When the GSOM output is obtained using a high SF value cluster identification via visual inspection becomes easy as the output is a more detailed map with increased branching out. If a low SF value is used, the output will be an abstract map where two groups of hit nodes might not be separated by non-hit nodes. In such situation it might not be possible to identify clusters accurately by human inspection as cluster boundaries are unclear. Ahmad et al. (2010) emphasises the requirement of having an automatic cluster identification mechanism when cluster boundaries are not clear.

The K-Means and Davies-Bouldin (DB) index based method (Davies and Bouldin 1979) and Data Skeleton Model (DSM) based method (Alahakoon et al. 2001) has been used for the automatic cluster identification in the GSOM outputs (Ahmad et al. 2010; Amarasiri et al. 2003). When using the K-Means and DB index based method, value for K needs to be pre-determined and value range 2 to $\sqrt{N}$ is suggested by Vesanto and Alhoniemi (2000) where N is the total number of hit nodes in the map. The K value which minimizes the DB-Index value

is selected as the final number of clusters. In Data Skeleton Modelling the path of the spread (POS) during node growth is manipulated to identify and separate clusters from the GSOM output.

3 Methodology

Figure 1 illustrates the proposed methodology for using GSOM based hierarchical clustering to capture term overlap in a document corpus. Each stage of the methodology is explained in following subsections.

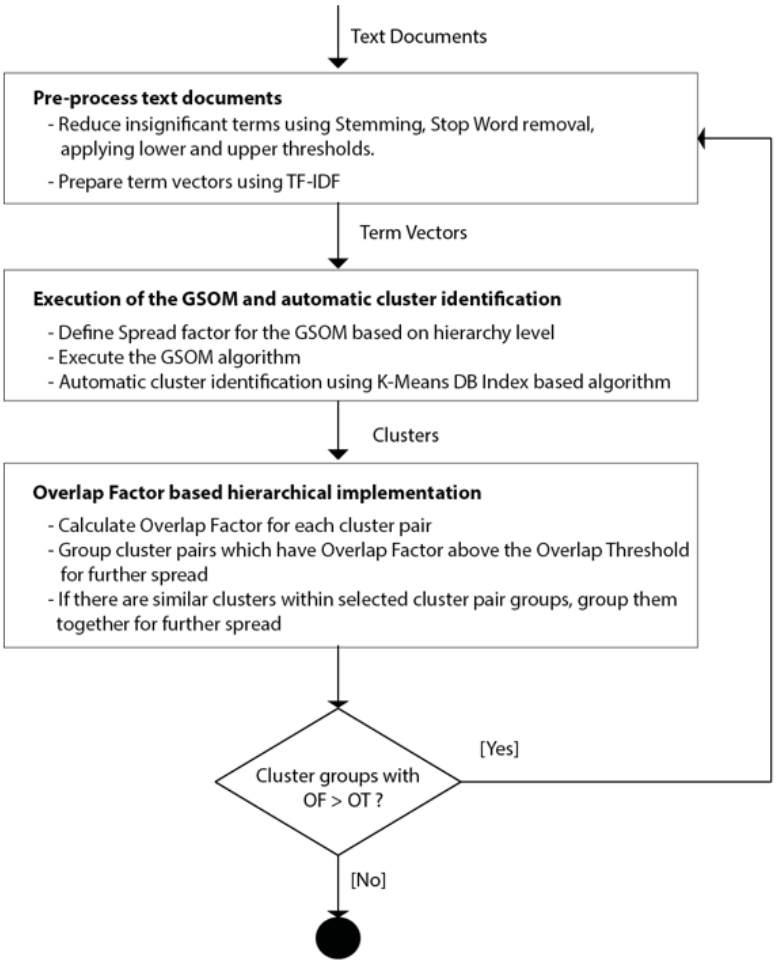

Figure 1: Proposed methodology for using GSOM based hierarchical clustering to capture term overlap in a document corpus

3.1 Pre-Process Text Documents

Input documents are pre-processed to reduce the insignificant terms. Stemming and stop word removal are executed and lower and upper thresholds are applied where the lower threshold helps to remove terms which occur very infrequently and the upper threshold helps to remove terms which occur very frequently.

After pre-processing text documents need to be represented as term vectors prior to the clustering process (Kohonen et al. 2000). Vector Space Model (VSM) (Salton et al. 1975) is a popular document representation technique where documents are represented as term vectors, in which the dimensionality is equivalent to the total number of terms extracted from the complete document set. The importance of each term is represented by a weight value and three popular term weighting techniques associated with the VSM are Binary representation, Term Frequency (TF), and Term Frequency – Inverse Document Frequency (TF-IDF). Binary representation indicates existence of a term in a document where weight 1 is given if term exists

and 0 is given when term does not exist. The equations for TF and TF-IDF term weighting techniques are presented by (1) and (2) respectively.

$$TF_{t,d} = \frac{n_{t,d}}{N_d} \quad (1)$$

$TF_{t,d}$ – term frequency of term t in document d, $n_{t,d}$ – number of occurrences of term t in document d, N_d – total number of terms in document d

$$TF - IDF_{t,d} = TF_{t,d} \times \log\frac{|D|}{DF_t} \quad (2)$$

$TF - IDF_{t,d}$ – term frequency – inverse document frequency of term t in document d, $TF_{t,d}$ – term frequency of term t in document d, |D| – total number of documents, DF_t – number of documents which contain term t.

In the proposed methodology term vectors are prepared (coded) using the TF-IDF weighting technique.

3.2 Execution of the GSOM and Automatic Cluster Identification

Term vectors are directed to the GSOM algorithm and SF value is chosen based on the level of the hierarchy. Lower SF value is selected for the first level of the hierarchy with the aim of obtaining an abstract map. Clusters in the output map are obtained using the K-Means and DB Index based automatic cluster identification mechanism.

3.3 Overlap Factor Based Hierarchical Implementation

Clusters obtained are directed to the term overlap calculation phase which uses two novel parameters introduced in the methodology, namely Overlap Factor (OF) and Overlap Threshold (OT). OF is introduced based on the Jaccard index (Jaccard 1908). Jaccard index which is also known as Jaccard similarity coefficient measures the similarity between finite sample sets (Jaccard 1908). The calculation of Jaccard index is presented by (3).

$$J(A, B) = \frac{(A \cap B)}{(A \cup B)} \quad (3)$$

where, A and B are finite sample sets

OF measures the term overlap between the cluster pairs which is an effective indicator of cluster purity. When calculating the OF only the most frequent terms which are considered as representative terms for clusters are taken into consideration. One major concern regarding the OF calculation is how many terms should be selected from each cluster. Selecting a limited number of highly frequent terms will not provide satisfactory results as the overlap in terms tend to be low. Therefore it is important for the analyst to decide how many terms are needed for the term overlap calculation and the best approach for such selection.

If the sizes of resultant clusters obtained via proposed methodology varies vastly then it is more appropriate to select fixed number of terms rather than a percentage of top most terms for each cluster. The reason is that if a particular cluster size is small and only few documents are mapped to the cluster, the total term count will be comparatively low than the other clusters. Even though total term count is low, most of the terms exist within the cluster might be very frequent within the cluster and if we select only the top most percentage only few terms out of frequent terms will be selected as representative terms for relevant cluster which will have negative impact for OF calculation.

The equation for OF calculation for cluster A and B is provided in (4).

$$\text{OF(A, B)} = \frac{(T_A \cap T_B)}{(T_A \cup T_B)} \quad (4)$$

where, T_A and T_B are representative terms for cluster A and B respectively and $0 \leq OF \leq 1$

OT is the threshold value which helps to decide whether further analysis is required for the obtained clusters in the GSOM output map and the value range for OT is $0 \leq OT \leq 1$. During this

phase OF value is calculated for each cluster pair and the cluster pairs which have OF greater than the OT are selected and grouped for further analysis. Each cluster in selected cluster pair group is then compared with other selected cluster pair groups and if same cluster is found in two cluster pair groups then they are grouped together. Afterwards clusters within groups are merged and selected for further spreading.

For each analysis, value for OT is predefined and this value can be consistent with hierarchical analysis or can be adjusted to different values in different hierarchy levels. When higher value is selected for OT only few clusters are selected for further analysis whereas large number of clusters are selected if low OT is selected. By varying the value of the OT, the data analyst can control the number of clusters which are chosen for further analysis.

According to the methodology if cluster groups are selected for further analysis, similar steps to the first iteration will be followed. During the execution of the GSOM, appropriate SF value can be selected based on the level of the hierarchy. This process executes iteratively until there is no further analysis required in the obtained clusters as OF calculated for cluster pairs is less than the predefined OT. This approach allows hierarchical clustering of a text document collection based on term overlap and obtain set of document clusters with minimum term overlap. Algorithmic form of the process is presented by Algorithm 1.

Input : *I* - Input text document collection
Initialise *SF* to a lower value
Initialise Overlap Threshold *OT*
Cluster(*I*) function
repeat
    Select *D* = TermSet(*I*) based on thresholds
    I - Input text document collection
    Generate weight matrix *W* using TF-IDF as the term weighting technique
    Feed *W* to GSOM *G* with relevant *SF* value
    Let's take *C* = Clusters(*G*)
    for iteration $i \leftarrow 1$ to *N* **do**
        N - number of clusters
        J = DocumentSet(C_i) - assign the document set mapped into the cluster C_i as the input collection
        Select *F* = TopTermSet(*J*) - assign the top most *n%* or *n* number of terms with highest frequency into *F* where TopTermSet(*J*) ∈ TermSet(*I*)
    end for
    for each cluster pair $x(C_i,C_j) \leftarrow 1$ to *P* **do**
        P - number of distinct cluster pairs
        Calculate *OF* = OverlapFactor(F_{C_i}, F_{C_j})
        if (*OF* >*OT*) then
            OT - Overlap Threshold
            $G = x(C_i,C_j)$ - assign the overlapping cluster pair to group *G*
        end if
    end for
    for each cluster group in *G*, G_i **do**
        for each cluster group in *G*, G_j where ($G_i \neq G_j$) **do**
            if $C_i \in G_j$ **then**
                Check whether any of the clusters in the group G_i is similar to any of the clusters in the group G_j
                Set G_i = merge(G_i, G_j) - merge the clusters in the groups G_i and G_j and Remove G_j from *G*
            end if
        end for
    end for

for each cluster in G **do**
I = DocumentSet(G_i) - assign the document set mapped into the cluster G_i as the input collection
Cluster(I) - recursively calls the clustering function. *SF* value should be increased based on the level of the hierarchy
end for
until *Overlap Factor value for each obtained cluster pair is lower than the pre-defined Overlap Threshold*

Algorithm 1: The GSOM and Term Overlap based Hierarchical Text Clustering Algorithm

4 Experimental Results

4.1 Dataset

A benchmark dataset known as 'BankSearch dataset' (Sinka and Corne 2005) which was used in many document clustering experiments (Borgelt and Nürnberger 2004; D'hondt et al. 2010) is used for experiments. The BankSearch dataset was formed using web pages and contains 11 document categories with each category containing 1000 documents. The dataset includes distant categories which are associated with different themes and close categories which are associated with the same theme according to human categorization. Distant and close groups exist within the dataset facilitate performing autonomous clustering experiments for close and distant categories separately. In addition one category which is 'Sport' is added as parent for 'Soccer' and 'Motor Sport' categories to facilitate hierarchical clustering.

Dataset ID	Dataset Category	Associated Theme
A	Commercial banks	Banking and finance
B	Building societies	Banking and finance
C	Insurance agencies	Banking and finance
D	Java	Programming languages
E	C/C++	Programming languages
F	Visual Basic	Programming languages
G	Astronomy	Science
H	Biology	Science
I	Soccer	Sport
J	Motor sport	Sport
K	Sport	Sport

Table 1: Data set categories and associated themes in BankSearch data set (Sinka and Corne 2005)

4.2 Investigate the Effect of Overlap Factor for Hierarchical Clustering with the GSOM

The main objective of the experiment was to identify the impact of term overlap for hierarchical analysis of text documents. The experiments were conducted for two distant categories, two close categories and mixture of distant and close categories which were selected from the above mentioned dataset (Sinka and Corne 2005). During the experiments how documents are structured hierarchically based on term similarity was investigated.

Steps in the proposed methodology were followed to obtain the output clusters and document pre-processing was carried out as the first step. First web documents were converted to plain text documents removing the tags used for web page design. Afterwards stop words and special characters were removed and Porter's stemming algorithm (Porter 1980) was applied to the

document set. Lower threshold of 2% and upper threshold of 50% were selected for thresholding and only the terms which had frequency in between were selected. After extracting the terms, term vectors were prepared using TF-IDF. For the first level of the hierarchy we have used SF value of 0.1 and the term vectors were directed to the GSOM algorithm. A learning rate of 0.05 with 50 training iterations and 100 smoothing iterations were used as the GSOM parameters. The nodes generated by the GSOM algorithm were separated into clusters by using the K-Means and DB Index based algorithm. During cluster separation we have executed K-Means algorithm for k = 2 to k = $\sqrt{N}$ where N was the total number of hit nodes. In addition we observed that for text clustering more accurate K-Means results were obtained when most hit nodes were used as cluster centroids instead of random initialization. Because of that when running the K-Means and DB Index based automatic cluster separation algorithm weight vectors of the most hit nodes were used as the cluster centroids.

Following sections provide details of the experiment results for selected document categories. Under each experiment GSOM output maps are presented where nodes with identical colour represent the same cluster and grey nodes represent the non-hit nodes. Clusters in the output maps were labelled based on the most frequent terms and few of most frequent terms are presented for each cluster. Top 35% of most frequent terms were used from each cluster for overlap calculation in each hierarchy level and calculated overlap values are rounded to 2 decimal places and presented in tables. In figures and tables term 'C' is used to represent clusters, 'SC' is used to represent sub clusters, and 'SSC' is used to represent further sub clusters obtained for sub clusters.

4.2.1 Experiment Results Obtained for Two Distant Categories

For the experiment categories 'Commercial banks' (A) and 'Soccer' (I) were selected which are associated with two different themes, 'Banking and finance' and 'Sport' respectively. The total document count used for the experiment was 2000. The main objective in this experiment was to identify how term overlap plays a role in document clustering when the input documents belong to vastly different themes in human categorization perspectives.

Figure 2 illustrates the GSOM output obtained for the first level of the hierarchy. Seven clusters were identified in the GSOM. From the representative terms of each cluster, only one cluster mainly represented the 'Commercial banks' category and other six clusters mainly represented the 'Soccer' category. The overlap for each cluster pair was calculated using (4) and the resulting overlap is shown in Table 2. The predefined OT was 0.4 and according to the overlap matrix cluster groups C2 and C3, C2 and C4, and C2 and C7 had OF greater than the predefined OT. As mentioned by the proposed algorithm if the selected cluster groups have common clusters they should be merged together for further analysis. Since all above cluster groups included C2 cluster, we have merged them all for further analysis. In addition it was noted that all four clusters selected for further analysis were mainly representing the 'Soccer' category. All the documents mapped to the selected four clusters were pre-processed following a similar approach used in the first hierarchy level and clustered using the GSOM algorithm with SF 0.3. Cluster separated GSOM output for the second level of the hierarchy is presented in Figure 3.

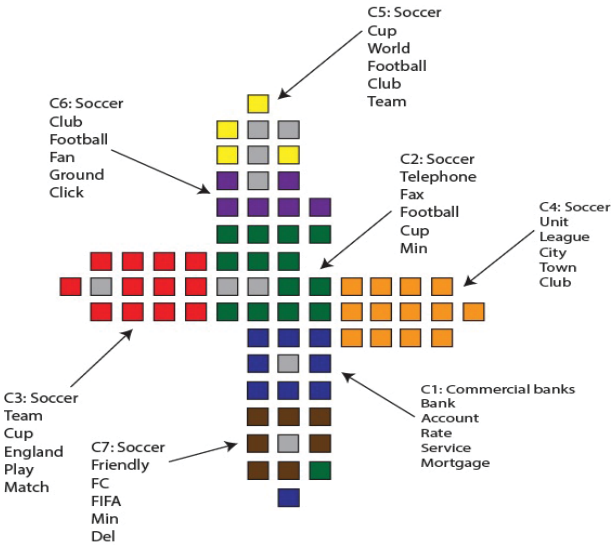

Figure 2: GSOM hierarchical clustering with SF 0.1 for two distant categories from the dataset (First level of the hierarchy)

ID	C1	C2	C3	C4	C5	C6	C7
C1		0.30	0.23	0.28	0.15	0.25	0.25
C2	0.30		0.41	0.41	0.25	0.39	0.42
C3	0.23	0.41		0.36	0.24	0.33	0.38
C4	0.28	0.41	0.36		0.29	0.38	0.39
C5	0.15	0.25	0.24	0.29		0.28	0.26
C6	0.25	0.39	0.33	0.38	0.28		0.31
C7	0.25	0.42	0.38	0.39	0.26	0.31	

Table 2: Overlap Matrix of each cluster pair for the 1st level of the hierarchy for distant categories

We have observed that the merged four clusters were divided into six sub clusters in the next level of clustering. The merged four clusters mainly represented the 'Soccer' category and the resultant sub clusters obtained in the second level of the hierarchy were also mainly representing the 'Soccer' category. To investigate the impact of hierarchical clustering for term overlap we have calculated the OF for cluster pairs and relevant overlap matrix is provided in Table 3.

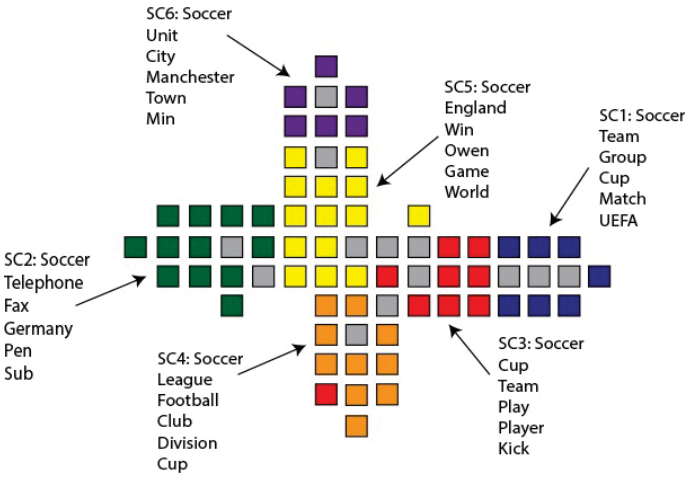

Figure 3: GSOM hierarchical clustering with SF 0.3 for merged C2, C3, C4, and C7 clusters obtained for distant categories in the first level of the hierarchy

ID	SC1	SC2	SC3	SC4	SC5	SC6
SC1		0.26	0.36	0.23	0.39	0.16
SC2	0.26		0.36	0.33	0.27	0.27
SC3	0.36	0.36		0.39	0.37	0.18
SC4	0.23	0.33	0.39		0.24	0.18
SC5	0.39	0.27	0.37	0.24		0.24
SC6	0.16	0.27	0.18	0.18	0.24	

Table 3: Overlap Matrix of each cluster pair for the 2nd level of the hierarchy for merged C2, C3, C4, and C7 clusters obtained for distant categories

According to the overlap matrix overlap was minimized between the cluster pairs in this second hierarchy level. An interesting observation was that number of highest important terms repeated across the clusters was minimized in the second hierarchy level. In the first hierarchy level more important terms were repeated across the clusters including 'Cup', 'Football', 'Club', and 'Team'. However in the second level of the hierarchy number was reduced to two and only 'Cup' and 'Team' were repeated, confirming higher cluster purity for distant document groups. Figure 4 presents the GSOM hierarchical clustering overview diagram for the selected distant categories including the distribution of the categories for each cluster.

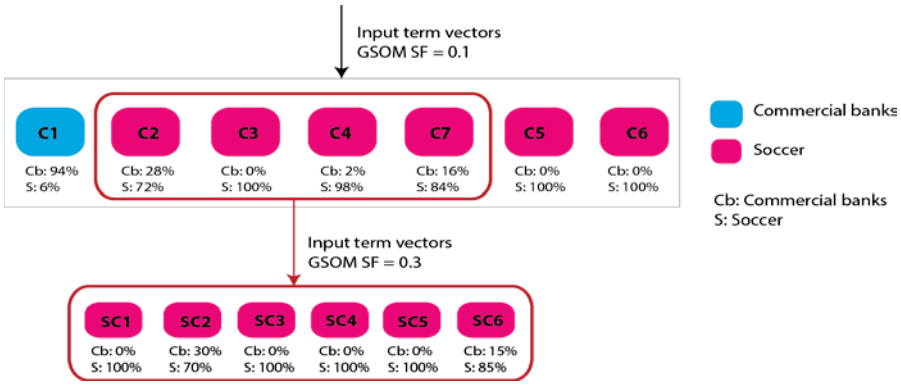

Figure 4: GSOM hierarchical clustering overview diagram for two distant categories from the dataset

To further verify the accuracy of the methodology we have executed experiments for same two distant categories changing the two parameter values which are number of terms selected for

the overlap calculation and OT. Rather than 35% of most frequent terms, most frequent 200 terms were used from each cluster for overlap calculation and OT 0.3 was used for the analysis. In the first hierarchy level when clustered with SF 0.1 seven clusters were obtained where one cluster represented 'Commercial banks' category and other six clusters mainly represented the 'Soccer' category. Two clusters representing 'Soccer' has been selected for further analysis and when clustered with SF 0.3 for the second level of the hierarchy seven sub clusters were obtained where one sub cluster represented 'Commercial banks' and all others were 'Soccer'. The overlap was minimised in the second level of the hierarchy where no further analysis was required. Detailed results of the experiments were not provided due to space limitations.

4.2.2 Experiment Results Obtained for Two Close Categories

Close categories used for the experiment were 'Commercial banks' (A) and 'Insurance agencies' (C) and the total document count used for the experiment was 2000. The main objective was to identify the possibility of obtaining clusters with low term overlap for documents which are associated with similar themes according to human categorization.

Figure 5 illustrates the GSOM output obtained for the first level of the hierarchy after automatic cluster separation. According to the results four clusters mainly represented 'Commercial banks' and the remaining two clusters representing the 'Insurance agencies.' The result overlap matrix (calculated as previously) is provided in Table 4.

Our pre-defined Overlap Threshold was 0.5 and according to the overlap matrix presented in Table 4 the cluster groups C1 and C3, C1 and C5, C2 and C6, C3 and C5 had OF more than the pre-defined OT. The selected cluster groups which have common clusters are merged together for further analysis. So Clusters C1, C3, and C5 and the clusters C2 and C6 were merged as separate groups.

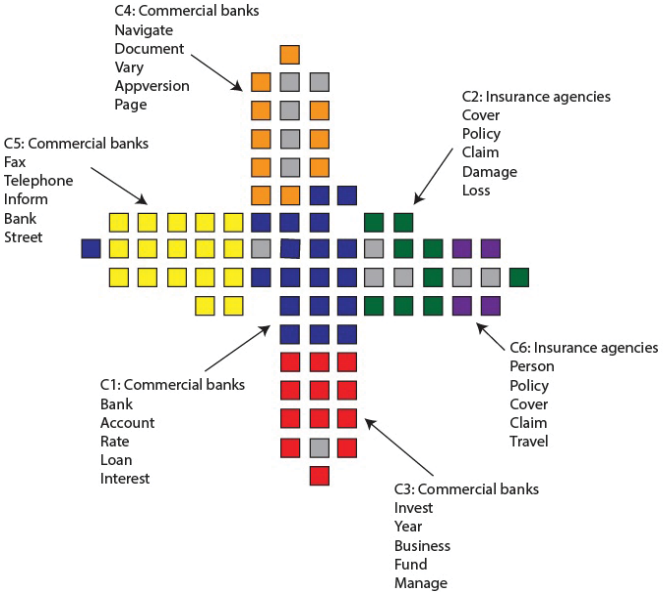

Figure 5: GSOM hierarchical clustering with SF 0.1 for two close categories from the dataset (First level of the hierarchy)

ID	C1	C2	C3	C4	C5	C6
C1		0.45	0.51	0.42	0.52	0.43
C2	0.45		0.40	0.34	0.45	0.59
C3	0.51	0.40		0.36	0.52	0.38
C4	0.42	0.34	0.36		0.43	0.32
C5	0.52	0.45	0.52	0.43		0.42
C6	0.43	0.59	0.38	0.32	0.42	

Table 4: Overlap Matrix of each cluster pair for the 1st level of the hierarchy for close categories

According to Figure 5 C1, C3, and C5 clusters grouped together for further analysis mainly represented the 'Commercial banks' category and the clusters C2 and C6 represented the 'Insurance agencies' and were further analysed separately. All the documents mapped into a group were processed using a similar approach followed in the first hierarchy level. Figure 6 illustrates the resultant map obtained by further analysis of C1, C3, and C5 clusters and Figure 7 illustrates the resultant map obtained by further analysis of C2 and C6 clusters with SF 0.3.

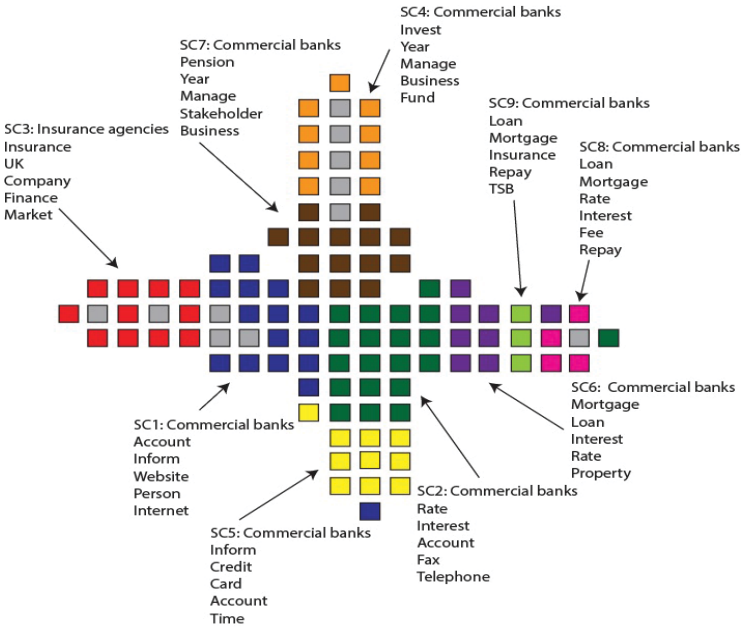

Figure 6: GSOM hierarchical clustering with SF 0.3 for merged C1, C3, and C5 clusters obtained for close categories in the first level of the hierarchy

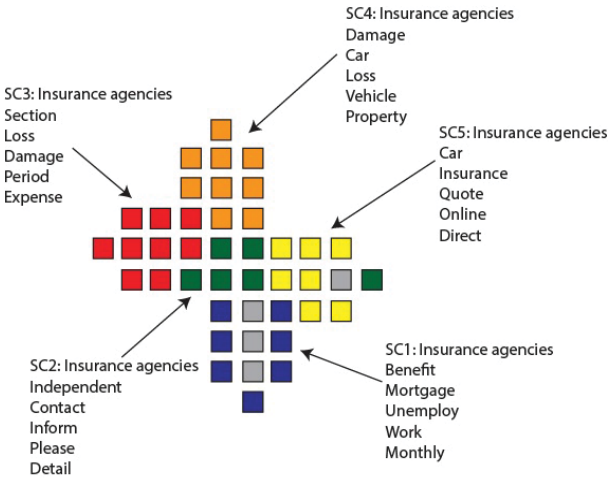

Figure 7: GSOM hierarchical clustering with SF 0.3 for merged C2 and C6 clusters obtained for close categories in the first level of the hierarchy

In Figure 6, merged C1, C3, and C5 cluster group was separated to 9 sub clusters. Out of the 9 sub clusters 8 represented the 'Commercial banks' category and 1 represented the 'Insurance agencies'. The initial clusters which were further spread in this second level of the hierarchy were mainly representing 'Commercial banks' category. However in this second hierarchy level, we observe a cluster which was well separated from other clusters and represent the 'Insurance agencies' category which confirm that few 'Insurance agencies' documents included in the clusters C1, C3, and C5 were separated in this second level of the hierarchy.

These results prove the ability of the GSOM algorithm to well separate text documents belonging to different categories using the terms when high SF value is used. We have calculated the OF for each sub cluster pair and relevant overlap matrix is given in Table 5. According to Table 5 some clusters still have high term overlap which exceeds the predefined OT. In addition Figure 6 illustrate that several most important terms are repeated across sub clusters so clusters are not well separated based on terms. In order to verify whether it is possible to obtain higher cluster purity we can merge sub clusters with OF greater than the OT by following the steps in the proposed methodology and further analyse them with higher SF value than 0.3.

ID	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9
SC1		0.48	0.47	0.37	0.63	0.44	0.41	0.42	0.41
SC2	0.48		0.46	0.45	0.47	0.54	0.50	0.51	0.46
SC3	0.47	0.46		0.44	0.45	0.42	0.46	0.40	0.43
SC4	0.37	0.45	0.44		0.37	0.41	0.56	0.37	0.39
SC5	0.63	0.47	0.45	0.37		0.44	0.41	0.43	0.41
SC6	0.44	0.54	0.42	0.41	0.44		0.44	0.63	0.54
SC7	0.41	0.50	0.46	0.56	0.41	0.44		0.40	0.43
SC8	0.42	0.51	0.40	0.37	0.43	0.63	0.40		0.56
SC9	0.41	0.46	0.43	0.39	0.41	0.54	0.43	0.56	

Table 5: Overlap Matrix of each cluster pair for the 2nd level of the hierarchy for merged C1, C3, and C5 clusters obtained for close categories

According to Figure 7, C2 and C6 clusters were separated to five sub clusters. The clusters C2 and C6 mainly represented the 'Insurance agencies' category and the resultant sub clusters were also representing the same category. Overlap matrix for further analysis of merged C2

and C6 clusters is given in Table 6 and according to Table 6 we can say that none of the sub clusters exceeded the OT such that no further analysis was required. Figure 8 presents the GSOM hierarchical clustering overview diagram for the selected close categories including the distribution of the categories for each cluster.

ID	SC1	SC2	SC3	SC4	SC5
SC1		0.43	0.35	0.36	0.34
SC2	0.43		0.42	0.43	0.42
SC3	0.35	0.42		0.43	0.28
SC4	0.36	0.43	0.43		0.32
SC5	0.34	0.42	0.28	0.32	

Table 6: Overlap Matrix of each cluster pair for the 2nd level of the hierarchy for merged C2 and C6 clusters obtained for close categories

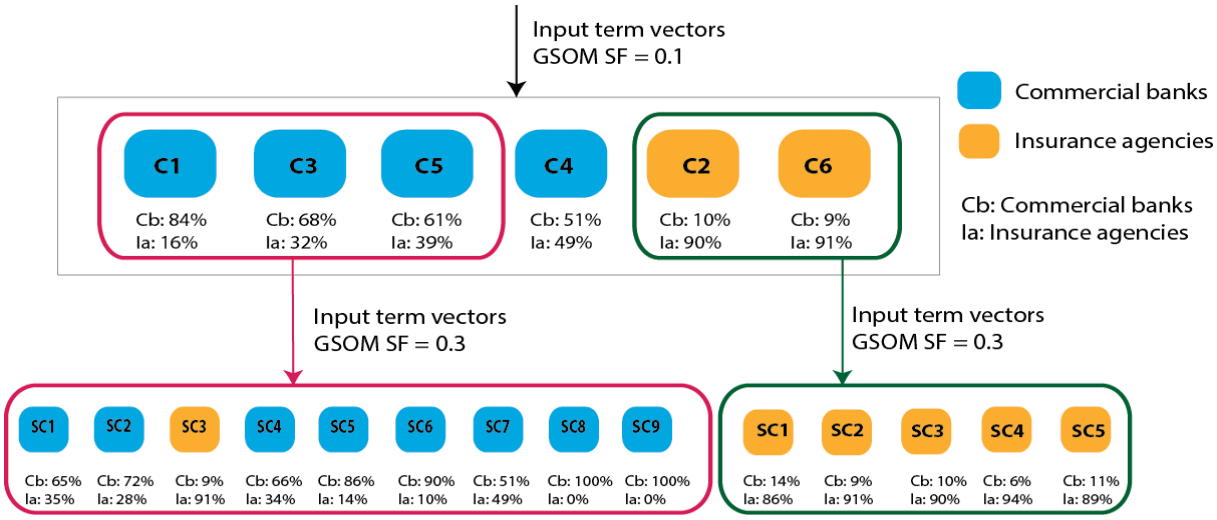

Figure 8: GSOM hierarchical clustering overview diagram for two close categories from the dataset

4.2.3 Experiment Results Obtained For Mixture of Close and Distant Categories

For this experiment we have selected 'Commercial banks' (A), 'Insurance agencies' (C), and 'Soccer' (I). 'Commercial banks' and 'Insurance agencies' are associated with the same theme which is 'Banking and finance'. However 'Soccer' is associated with a different theme which is 'Sport'. During this experiment requirement for running more iterations was expected as close and distant categories tend to separate in the first level of the hierarchy and further drill down might be required for clusters representing close categories to obtain minimal term overlap. The total document count used for the experiment was 3000 and the steps in the proposed methodology were followed.

The first hierarchy level generated 8 clusters after automatic cluster separation was executed. Figure 9 illustrates the cluster separation in the GSOM output. According to the results four clusters mainly represented the 'Soccer' category, one cluster represented the 'Commercial banks' category and three clusters represented the 'Insurance agencies' category. The resulting overlap matrix is provided in Table 7. OT value used for the experiment was 0.5 and according to the overlap matrix presented in Table 7 the cluster groups C1 and C2, C2 and C5 had OF more than the pre-defined overlap threshold value. So the clusters C1, C2, and C5 were merged for further analysis. Two clusters selected for further analysis mainly represent 'Insurance agencies' whereas the other cluster represent 'Commercial banks'. This indicates that it is likely to obtain higher term overlap between the clusters associated with similar themes. Documents associated with 'Soccer' theme are distributed among four clusters which prove that the GSOM

algorithm has well separated the 'Soccer' related documents in the first hierarchy level. In addition the OF calculated between the clusters labelled as 'Soccer' didn't exceed the pre-defined OT which verified that those clusters were well separated based on terms even though they belonged to the same theme. When looking into the cluster separation and overlap matrix we can make the conclusion that GSOM algorithm has well separated the distant category from the close categories in the first hierarchy level.

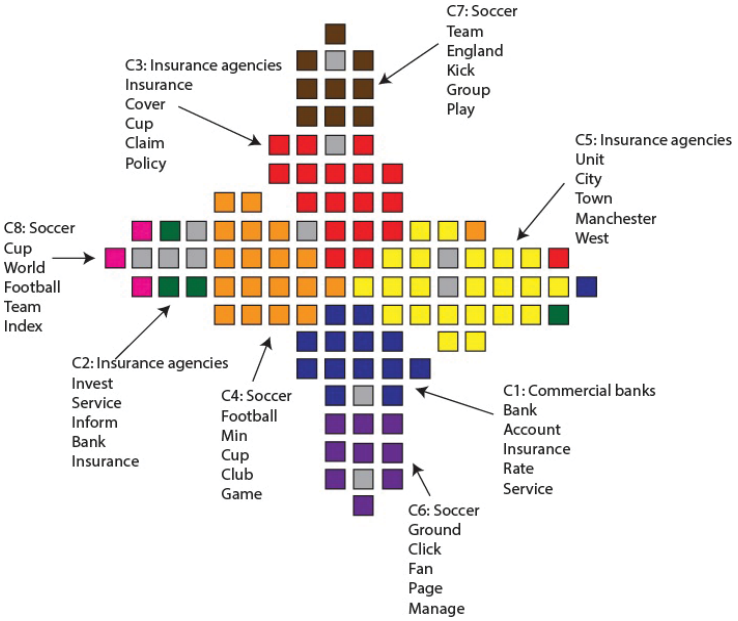

Figure 9: GSOM hierarchical clustering with SF 0.1 for mixture of close and distant categories from the dataset (First level of the hierarchy)

ID	C1	C2	C3	C4	C5	C6	C7	C8
C1		0.58	0.49	0.28	0.47	0.39	0.21	0.12
C2	0.58		0.46	0.33	0.52	0.44	0.22	0.14
C3	0.49	0.46		0.38	0.46	0.36	0.34	0.17
C4	0.28	0.33	0.38		0.44	0.37	0.38	0.25
C5	0.47	0.52	0.46	0.44		0.46	0.28	0.20
C6	0.39	0.44	0.36	0.37	0.46		0.23	0.17
C7	0.21	0.22	0.34	0.38	0.28	0.23		0.18
C8	0.12	0.14	0.17	0.25	0.20	0.17	0.18	

Table 7: Overlap Matrix of each cluster pair for the 1st level of the hierarchy for mixture of close and distant categories

For the second hierarchy level we have used the SF 0.3 for the GSOM algorithm. The merged three clusters were separated into ten sub clusters which is illustrated in Figure 10. Five sub clusters mainly represented 'Commercial banks' category, four sub clusters 'Insurance agencies' and one sub cluster was 'Soccer'. The merged three clusters used for further analysis were mainly representing 'Commercial banks' and 'Insurance agencies' categories. However in this second hierarchy level we obtain a sub cluster which represent 'Soccer' category and it verifies that some of the 'Soccer' related documents assigned to the three clusters selected for further analysis were separated in the second hierarchy level. We observed that several sub clusters had OF greater than the OT and most frequent terms were still repeated among number of sub clusters. This repetition of most frequent terms was significantly visible for sub clusters labelled as 'Commercial banks'. This observation verified that the sub clusters labelled as 'Commercial banks' consist of similar terms and in the term overlap matrix all of those

clusters had OF greater than the OT which confirmed the observation. The overlap matrix for the second hierarchy level is presented in Table 8 and according to Table 8 sub cluster groups SC1 and SC3, SC1 and SC6, SC1 and SC7, SC2 and SC6, SC2 and SC7, SC5 and SC7 showed OF greater than the pre-defined OT. Following the merging rule in our proposed algorithm we have merged the sub clusters SC1, SC2, SC3, SC5, SC6, and SC7 for further analysis. Five sub clusters selected for further analysis mainly represented the 'Commercial banks' category and the other sub cluster mainly represented the 'Insurance agencies' category.

Figure 10: GSOM hierarchical clustering with SF 0.3 for merged C1, C2, and C5 clusters obtained for mixture of close and distant categories in the first level of the hierarchy

ID	SC1	SC2	SC3	SC4	SC5	SC6	SC7	SC8	SC9	SC10
SC1		0.49	0.52	0.42	0.44	0.51	0.52	0.40	0.34	0.42
SC2	0.49		0.43	0.43	0.49	0.55	0.53	0.46	0.28	0.41
SC3	0.52	0.43		0.42	0.37	0.39	0.47	0.41	0.35	0.46
SC4	0.42	0.43	0.42		0.40	0.43	0.40	0.36	0.30	0.42
SC5	0.44	0.49	0.37	0.40		0.41	0.51	0.49	0.29	0.42
SC6	0.51	0.55	0.39	0.43	0.41		0.46	0.37	0.25	0.38
SC7	0.52	0.53	0.47	0.40	0.51	0.46		0.49	0.30	0.41
SC8	0.40	0.46	0.41	0.36	0.49	0.37	0.49		0.28	0.39
SC9	0.34	0.28	0.35	0.30	0.29	0.25	0.30	0.28		0.32
SC10	0.42	0.41	0.46	0.42	0.42	0.38	0.41	0.39	0.32	

Table 8: Overlap Matrix of each cluster pair for the 2nd level of the hierarchy for merged C1, C2, and C5 clusters obtained for mixture of close and distant categories

Figure 11 illustrates the output map using SF 0.5 and according to Figure 11 we observed that merged six clusters were separated to six sub clusters. Four sub clusters mainly represented 'Commercial banks', one sub cluster 'Insurance agencies' and the other sub cluster 'Soccer'. Similar to the results for the second hierarchy level we observed that one sub cluster represent 'Soccer' category even though the merged clusters directed for further analysis mainly represented 'Commercial banks' and 'Insurance agencies' categories. This observation proves that few documents belonging to 'Soccer' category were included to the clusters used for further analysis and they were separated in the third hierarchy level.

Overlap matrix for third hierarchy level is shown in Table 9. Within the clusters further analysed in this third hierarchy level there was one cluster mainly representing 'Insurance agencies'. It can be seen that same most important terms presented for relevant cluster in Figure 10 were repeated for the sub cluster labelled as 'Insurance agencies' obtained in the third hierarchy level. However when comparing relevant two clusters labelled as 'Insurance agencies' in the second and third hierarchy levels, we can say that the cluster purity is higher for the cluster we obtained in the third hierarchy level. Since the term overlap calculated in the third hierarchy level didn't exceed the OT which shows that relevant cluster does not have significant term overlap with any other sub cluster in the third level of the hierarchy. According to Figure 11 we notice that there were three sub clusters which still had OF greater than the OT. Those clusters mainly represent 'Commercial banks'. However, one sub cluster which represented 'Commercial banks' category was separated well as it did not have significant term overlap with other sub clusters. We can further analyse the three sub clusters which had OF greater than the OT to verify whether it is possible to obtain high cluster purity by minimising the term overlap between the resultant clusters. Figure 12 presents the GSOM hierarchical clustering overview diagram for the selected close and distant categories including the distribution of the categories for each cluster.

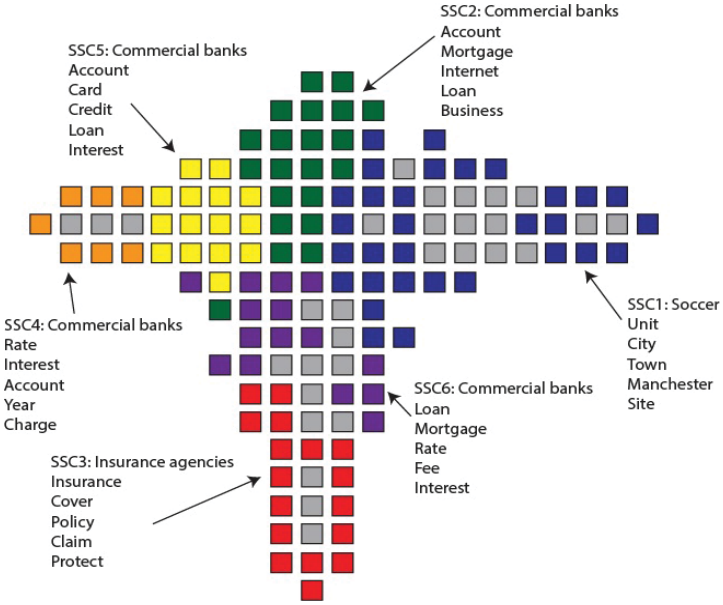

Figure 11: GSOM hierarchical clustering with SF 0.5 for merged SC1, SC2, SC3, SC5, SC6, and SC7 sub clusters obtained for mixture of close and distant categories in the second level of the hierarchy

ID	SSC1	SSC2	SSC3	SSC4	SSC5	SSC6
SSC1		0.42	0.37	0.31	0.37	0.35
SSC2	0.42		0.46	0.49	0.56	0.59
SSC3	0.37	0.46		0.37	0.48	0.47
SSC4	0.31	0.49	0.37		0.47	0.49
SSC5	0.37	0.56	0.48	0.47		0.49
SSC6	0.35	0.59	0.47	0.49	0.49	

Table 9: Overlap Matrix of each cluster pair for the 3rd level of the hierarchy for merged SC1, SC2, SC3, SC5, SC6, and SC7 sub clusters obtained for mixture of close and distant categories

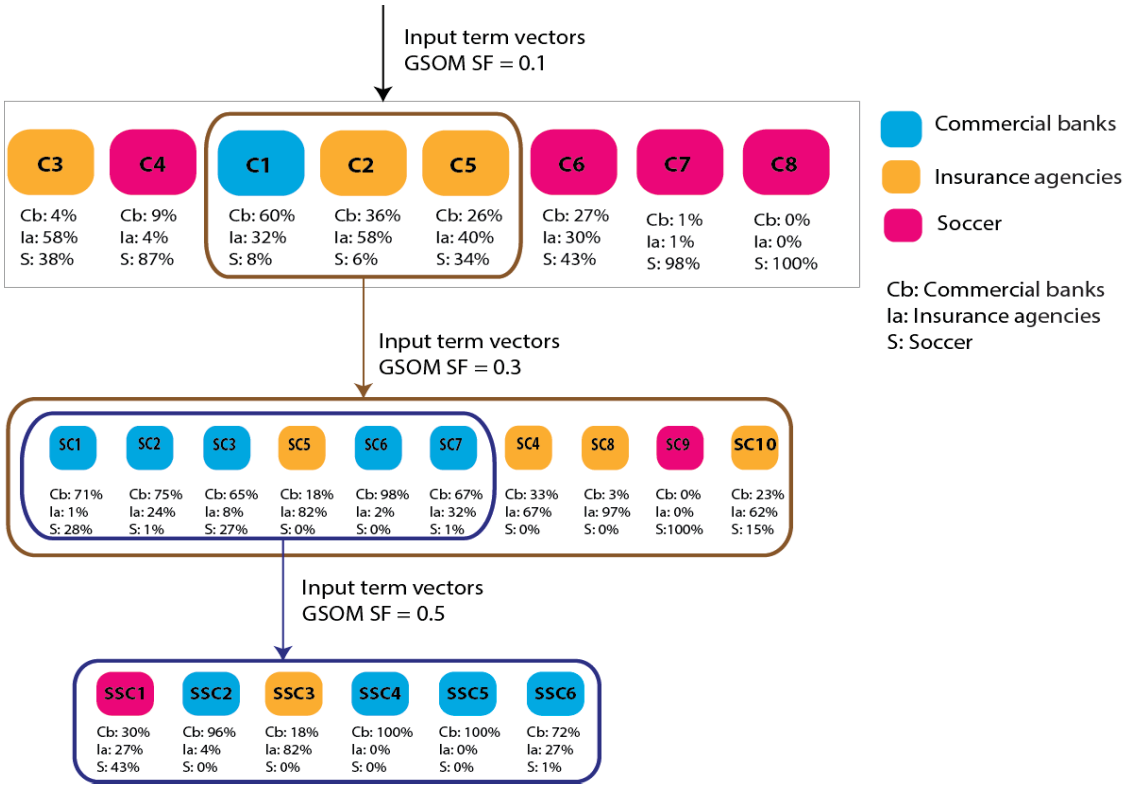

Figure 12: GSOM hierarchical clustering overview diagram for mixture of close and distant categories from the dataset

5 Conclusion

In this paper we have introduced a novel methodology to analyse text documents hierarchically based on the terms in the documents. According to our experiment results we can conclude that the GSOM hierarchical clustering can be used together with term overlap between the clusters to conduct an exploratory analysis of a given document set. The experiment results showed that the proposed methodology will lead to obtaining clusters with minimal term overlap when executed for distant categories. For close categories it might be required to execute the steps several times iteratively to obtain clusters with minimal term overlap. When the document set consists with a mixture of distant and close categories, distant categories tend to be separated within first few levels of the hierarchy showing minimum term overlap whereas close categories might require more iterations to obtain minimum term overlap between the clusters. However it was noticed that the analysis output we obtained by following the novel methodology is highly dependent on the repetitive terms within the given document set regardless of the themes they are actually associated with.

One advantage of the novel methodology is that it allows investigating the diversity of the document content for a given document set. In addition it allows making decisions on what documents are more likely representing a similar or close document category based on term overlap.

Organizations collect large volumes of documents which contain information necessary for decision making at various levels. Automating the separation of documents into useful groups is highly beneficial since this will cut down lengthy manual processing which becomes almost impossible due to large volumes and high frequency of document collection. Traditionally documents are grouped in to pre-defined categories but a document may belong to several such categories or better fit into a new category (no existing category) based on the content. The methodology proposed in this paper enables exploration of a document corpus by content and hierarchically separating into categories and sub categories based on term overlap.

References

Ahmad, N., Alahakoon, D., and Chau, R. 2010. "Cluster Identification and Separation in the Growing Self-Organizing Map: Application in Protein Sequence Classification," *Neural Computing & Applications* (19:4), pp. 531-542.

Alahakoon, D., Halgamuge, S. K., and Srinivasan, B. 2000. "Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery," *IEEE Transactions on Neural Networks* (11:3), pp. 601-614.

Alahakoon, D., Halgamuge, S. K., and Srinivasan, B. 2001. "Mining a Growing Feature Map by Data Skeleton Modelling," *Data Mining and Computational Intelligence* (68), pp. 217-250.

Amarasiri, R., Alahakoon, D., and Smith, K. A. 2005. "Hdgsom: A Modified Growing Self-Organizing Map for High Dimensional Data Clustering," *HIS'04: Fourth International Conference on Hybrid Intelligent Systems, Proceedings*, pp. 216-221.

Amarasiri, R., Wickramasinghe, L., and Alahakoon, L. D. 2003. "Enhanced Cluster Visualization Using the Data Skeleton Model," in *Intelligent Systems Design and Applications. Springer*, pp. 539-548.

Bache, K., and Lichman, M. 2013. *{Uci} Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences.

Borgelt, C., and Nürnberger, A. 2004. "Experiments in Document Clustering Using Cluster Specific Term Weights," *Proc. Workshop Machine Learning and Interaction for Text-based Information Retrieval (TIR 2004)*, pp. 55-68.

D'hondt, J., Vertommen, J., Verhaegen, P.-A., Cattrysse, D., and Duflou, J. R. 2010. "Pairwise-Adaptive Dissimilarity Measure for Document Clustering," *Information Sciences* (180:12), pp. 2341-2358.

Davies, D. L., and Bouldin, D. W. 1979. "Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1:2), pp. 224-227.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., and Stanley, H. E. 2000. "Physiobank, Physiotoolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation* (101:23), pp. E215-220.

Gunasinghe, U., Matharage, S., and Alahakoon, D. 2012. "A Sequence Based Dynamic Som Model for Text Clustering," *IEEE International Joint Conference on Neural Networks (IJCNN)*.

Haiying, W., Azuaje, F., and Black, N. 2004. "An Integrative and Interactive Framework for Improving Biomedical Pattern Discovery and Visualization," *IEEE Transactions on Information Technology in Biomedicine*, (8:1), pp. 16-27.

Hsu, A. L., Tang, S. L., and Halgamuge, S. K. 2003. "An Unsupervised Hierarchical Dynamic Self-Organizing Approach to Cancer Class Discovery and Marker Gene Identification in Microarray Data," *Bioinformatics* (19:16), pp. 2131-2140.

Jaccard, P. 1908. "Nouvelles Recherches Sur La Distribution Florale," *Bulletin de la Sociète Vaudense des Sciences Naturelles* (44), pp. 223-270.

Jain, A. K., Murty, M. N., and Flynn, P. J. 1999. "Data Clustering: A Review," *ACM computing surveys (CSUR)* (31:3), pp. 264-323.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. 1998. "Websom - Self-Organizing Maps of Document Collections," *Neurocomputing* (21:1-3), pp. 101-117.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. 2014. "Text Mining for Market Prediction: A Systematic Review," *Expert Systems with Applications* (41:16), pp. 7653-7670.

Kohonen, T. 1982. "Self-Organized Formation of Topologically Correct Feature Maps," *Biological Cybernetics* (43:1), pp. 59-69.

Kohonen, T. 1998. "The Self-Organizing Map," *Neurocomputing* (21:1-3), pp. 1-6.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., and Saarela, A. 2000. "Self Organization of a Massive Document Collection," *IEEE Transactions on Neural Networks* (11:3), pp. 574-585.

Lu, Y., Zhang, P., Liu, J., Li, J., and Deng, S. 2013. "Health-Related Hot Topic Detection in Online Communities Using Text Clustering," *PLoS ONE* (8:2).

Matharage, S., Alahakoon, O., Alahakoon, D., Kapurubandara, S., Nayyar, R., Mukherji, M., Jagadish, U., Yim, S., and Alahakoon, I. 2011. "Analysing Stillbirth Data Using Dynamic Self Organizing Maps," *IEEE International Workshop on Database and Expert Systems Applications (DEXA), 2011 22nd*, pp. 86-90.

Matharage, S., Gunasinghe, U., and Alahakoon, D. 2009. "Growing Self Organizing Map with an Imposed Binary Search Tree for Discovering Temporal Input Patterns," *2009 IEEE International Conference on Industrial and Information Systems*, pp. 222-226.

Matharage, S. S. 2012. *An Autonomous Incremental Learning Model for Efficient Mining of Text Data*. in: Faculty of Information Technology. Clayton School of Information Technology. Monash University.

Merkl, D. 1998. "Text Classification with Self-Organizing Maps: Some Lessons Learned," *Neurocomputing* (21:1–3), pp. 61-77.

Merkl, D., and Rauber, A. 1999. "Uncovering Associations between Documents," *Proc. International Joint Conference on Artificial Intelligence (IJCAI99)*.

Minanović, A., Gabelica, H., and Krstić, Ž. 2014. "Big Data and Sentiment Analysis Using Knime: Online Reviews Vs. Social Media," *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings*, pp. 1464-1468.

Porter, M. F. 1980. "An Algorithm for Suffix Stripping," *Program-Automated Library and Information Systems* (14:3), pp. 130-137.

Salton, G., Wong, A., and Yang, C. S. 1975. "A Vector Space Model for Automatic Indexing," *Commun. ACM* (18:11), pp. 613-620.

Schkolnick, M. 1977. "A Clustering Algorithm for Hierarchical Structures," *ACM Trans. Database Syst.* (2:1), pp. 27-44.

Sinka, M. P., and Corne, D. W. 2005. "The Banksearch Web Document Dataset: Investigating Unsupervised Clustering and Category Similarity," *Journal of Network and Computer Applications* (28:2), pp. 129-146.

Vesanto, J., and Alhoniemi, E. 2000. "Clustering of the Self-Organizing Map," *IEEE Transactions on Neural Networks*, (11:3), pp. 586-600.