A Similarity Measure for Clustering and its Applications

Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, Srinivas Mukkamala, and Bernardete M. Ribeiro

Abstract**—**This paper introduces a measure of similarity between two clusterings of the same dataset produced by two different algorithms, or even the same algorithm (K-means, for instance, with different initializations usually produce different results in clustering the same dataset). We then apply the measure to calculate the similarity between pairs of clusterings, with special interest directed at comparing the similarity between various machine clusterings and human clustering of datasets. The similarity measure thus can be used to identify the best (in terms of most similar to human) clustering algorithm for a specific problem at hand. Experimental results pertaining to the text categorization problem of a Portuguese corpus (wherein a translation-into-English approach is used) are presented, as well as results on the well-known benchmark IRIS dataset. The significance and other potential applications of the proposed measure are discussed.

Keywords**—**Clustering Algorithms, Clustering Applications, Similarity Measures, Text Clustering.

I. Introduction and Motivation

OUR study of similarity of clustering was initially motivated by a research on automated text categorization of foreign language texts, as explained below.

As the amount of digital documents has been increasing dramatically over the years as the Internet grows, information management, search, and retrieval, etc., have become practically important problems.

Developing methods to organize large amounts of unstructured text documents into a smaller number of meaningful clusters would be very helpful as document clustering is vital to such tasks as indexing, filtering, automated metadata generation, word sense disambiguation, population of hierarchical catalogues of web resources and, in general, any application requiring document organization [1], [2]. Document clustering is also useful for topics such as Gene Ontology [3] in biomedicine where hierarchical catalogues are needed.

To deal with the large amounts of data, machine learning approaches have been applied to perform Automated Text Clustering (ATC). Given an unlabeled dataset, this ATC system builds clusters of documents that are hopefully similar to clustering (classification, categorization, or labeling) performed by human experts.

To identify a suitable tool and algorithm for clustering that produces the best clustering solutions, it becomes necessary to have a method for comparing the results of different clustering algorithms. Though considerable work has been done in designing clustering algorithms, not much research has been done on formulating a measure for the similarity of two different clustering algorithms.

Thus, the main goal of this paper is to: First, propose an algorithm for performing similarity analysis among different clustering algorithms; second, apply the algorithm to calculate similarity of various pairs of clustering methods applied to a Portuguese corpus and the Iris dataset; finally, to cross-validate the results of similarity analysis with the Euclidean (centroids) distances and Pearson correlation coefficient, using the same datasets. Possible applications are discussed.

II. Clustering Methods

A cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters [4]; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set.

There are several categories of clustering algorithms. In this paper we will be focusing on algorithms that are exclusive in that the clusters may not overlap.

Some of the algorithms are hierarchical and probabilistic. A hierarchical algorithm clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations, it reaches the final clusters wanted. The final category of probabilistic algorithms is focused around model matching using probabilities as opposed to distances to decide clusters. EM or Expectation Maximization is an example of this type of clustering algorithm.

In [5], Pen et al. utilized cluster analysis composed of 2 methods. In Method I, a majority voting committee with 3 results generates the final analysis result. The performance measure of the classification is decided by majority vote of the committee. If more than 2 of the committee members give the same classification result, then the clustering analysis for that observation is successful; otherwise, the analysis fails.

Manuscript submitted May 18, 2008. This work was supported in part by ICASA (Institute for Complex Additive Systems Analysis), a division of New Mexico Tech.

G. J. Torres, R. B. Basnet (corresponding author), A. H. Sung, and S. Mukkamala are with the Department of Computer Science & ICASA, New Mexico Tech, Socorro, NM 87801, USA (phone: +1-575-835-5126; fax: 505-835-5587; e-mail: {silfalco | rbasnet | sung | srinivas}@cs.nmt.edu).

B. M. Ribeiro is a member of the Department of Informatics Engineering at the University of Coimbra, Coimbra, Portugal (phone: +351-239-790087; e-mail: bribeiro@dei.uc.pt).

Kalton et al. [6] did clustering and after letting the algorithm create its own clusters, added a step. After the clustering was completed each member of a class was assigned the value of the cluster's majority population. The authors noted that the approach loses detail, but allowed them to evaluate each clustering algorithm against the "correct" clusters.

III. The Similarity Measure Algorithm

To measure the 'similarity' of two sets of clusters, we define a simple formula here: Let $C = \{C_1, C_2, \ldots, C_m\}$ and $D = \{D_1, D_2, \ldots, D_n\}$ be the results of two clustering algorithms on the same data set. Assume C and D are "hard" or exclusive clustering algorithms where clusters produced are pair-wise disjoint, i.e., each pattern from the dataset belongs to exactly one cluster. Then the similarity matrix for C and D is an $m \times n$ matrix $S_{C,D}$ (1).

$$S_{C,D} = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1j} & \ldots & S_{1n} \\ & & \cdots & & & \\ S_{i1} & S_{i2} & \ldots & S_{ij} & \ldots & S_{in} \\ & & \cdots & & & \\ S_{m1} & S_{m2} & \ldots & S_{mj} & \ldots & S_{mn} \end{bmatrix} \quad (1)$$

where $S_{ij} = p/q$, which is Jaccard's Similarity Coefficient [7] with p being the size of intersection and q being the size of the union of cluster sets C_i and D_j. The similarity of clustering C and clustering D is then defined as

$$\mathbf{Sim(C, D)} = \Sigma_{i \le m, j \le n} \mathbf{S_{ij}} \;/\; \mathbf{max\,(m, n)} \quad (2)$$

For Example 1, let $C_1 = \{1,2,3,4\}$ $C_2 = \{5,6,7,8\}$ and $D_1 = \{1,2\}$, $D_2 = \{3,4\}$, $D3 = \{5,6\}$, $D4 = \{7,8\}$ thus m=4 and n=2, then the similarity between clustering C and D is given by the following matrix $S_{C,D}$.

TABLE I
SIMILARITY MATRIX ON EXAMPLE 1 DATA

Cluster	D_1	D_2	D_3	D_4
C_1	2/4	2/4	0/6	0/6
C_2	0/6	0/6	2/4	2/4

In cell C_1D_1, $p=|C_1 \cap D_1|=|\{1,2\}|=2$, and $q=|C_1 \cup D_1|=|\{1,2,3,4\}|=4$. Therefore, cell $C_1D_1=p/q=2/4$. Similarly the other cells of the matrix are calculated. Thus, the similarity between cluster set C and cluster set D in this case is Sim(C, D) = (2/4+2/4+0/6+0/6+0/6+0/6+2/4+2/4)/4 = 0.5

For Example 2, let $C_1 = \{1,2,3,4,5,6\}$ $C_2 = \{7,8\}$; $D_1 = \{1,2,3,4\}$, $D_2 = \{5,6,7,8\}$, thus m=2, n=2 and matrix $S_{C,D}$ is:

TABLE II
SIMILARITY MATRIX ON EXAMPLE 2 DATA

Cluster	D_1	D_2
C_1	4/6	2/8
C_2	0/6	2/4

Thus, the similarity Sim(C, D), according to the similarity matrix above, is (4/6+2/8+0/6+2/4)/2= 17/24 or 0.7083

It is easy to show that $0 < \text{Sim(C, D)} \le 1$; and Sim(C, D)=1 for two identical clustering, where the similarity matrix Sim(C, D) is a square matrix; and that this measure is only applicable to clustering a finite set of patterns into a finite number of disjoint (or non-overlapping) clusters.

Also, we can take the square of summation of the matrix values to define similarity Sim(C, D), i.e., let $\mathbf{Sim(C, D)} = (\sum_{i,j} \mathbf{S_{ij}} \,/\, \mathbf{max(m, n)})^2$, this would have the effect of giving a lower value of similarity but without changing its range of (0, 1]. This similarity measure is a reasonable one to use because, if we define the dissimilarity or "distance" between two clusterings C and D as U (C,D) = 1 – Sim (C,D), then it can be proved that U(C,D) is a good distance measure for it satisfies all desirable properties (non-negativity, identity, symmetry, triangle inequality) of a distance metric.

IV. Methodology

Fig. 1 illustrates the steps carried out for similarity measure of clustering a Portuguese corpus. The details of the "translation based text categorization" technique for foreign-language texts are found in [8] and briefly described below. (The Iris dataset that is used in our second set of experiments does not require any preprocessing.)

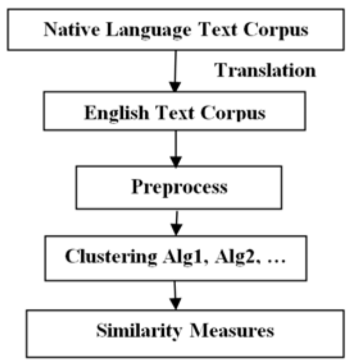

Fig. 1 Methodology for calculating similarity measure of clustering the Portuguese dataset

A. The Datasets

The Portuguese CETEMPublico corpus consisting of 1.5 million extracts, more than 225 million tokens, is excerpts of Portuguese newspaper Publico [9]. There are total of 9 different categories which are shown in Table III. The "nd" category which is short for "not defined" was excluded from our experiments. 1000 randomly chosen documents with at least 75 tokens were extracted for each category and then translated into English using the Google translation service [10].

Iris dataset, one of the most popular datasets in pattern recognition literature, was used as benchmark dataset. The dataset can be downloaded from Machine Learning Repository at University of California, Irvine [11]. The dataset is summarized in Table IV.

TABLE III
GENRES COVERED IN THE CETEMPUBLICO CORPUS

ID	Category	Description	Samples
1	clt	Culture	1000
2	clt-soc	Culture-Society	1000
3	com	Technology	1000
4	des	Sports	1000
5	eco	Economics	1000
6	opi	Opinions	1000
7	pol	Politics	1000
8	soc	Society	1000
Total:			**8000**

TABLE IV
IRIS DATASET

ID	Category	Samples
1	Iris Setosa	50
2	Iris Versicolour	50
3	Iris Virginica	50
Total		**150**

B. *Preprocessing Portuguese Dataset*

Tokenization was carried out by using suitable delimiters such as white-space and punctuation marks. Stop words or functional words such as article, prepositions, etc. that are not useful in the text categorization process were removed during preprocessing. Stemming was used to extract the root form of each word in the document. Since stem word as features performs better than single words and noun-phrase [12], we applied the popular and publicly available Porter Stemmer algorithm to stem translated English words [13].

Though there are various term weighting schemes such as BINARY, TF, LOGTF, LOGTFIDF, IDF, TF-CHI, TF-RF [14], [15], we used the traditional but popular weighting scheme, TF.IDF which is one of the best performance wise.

C. *Clustering Algorithms*

In experimenting with our clustering similarity algorithm the following clustering algorithms were studied:

A. Repeated Bisection
B. Direct
C. Agglomerative
D. Graph
E. K-means
F. K-medoids
G. EM

For the first four algorithms (A - D), gCLUTO [16], a cross-platform graphical application for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters, was used. gCLUTO is built on top of the CLUTO clustering library.

For K-means (E) and K-medians (F) the Matlab Fuzzy Clustering and Data Analysis Toolbox [17] was utilized.

Finally, for Expectation Maximization (G) the WEKA (Waikato Environment for Knowledge Analysis) [18] tool was used.

D. *Clustering Similarity Analysis*

After applying the clustering algorithms on Portuguese and Iris datasets, clustering similarities were calculated using the proposed algorithm. The results were then verified by calculating centroid Euclidean distance and Pearson correlation.

Euclidean distance:

$$d = \sqrt{\sum (X - Y)^2} \tag{3}$$

Pearson correlation coefficient:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} \tag{4}$$

V. EXPERIMENTAL RESULTS

Pair-wise similarity matrix for all the clustering algorithms mentioned in section IV.C and with the human-labeled actual categories (H) was generated using the Similarity Algorithm we've proposed and cross verified with results from Euclidean distance and Pearson correlation.

Due to space limitation only the final similarity matrix between Repeated Bisection and the rest of the algorithms including human-labeled actual categories are shown and cluster is abbreviated as (**Cl.**) in the result tables. The significant values in the result tables have been **emphasized**: value closest to 1 in similarity matrix, smallest value for Euclidian distance i.e. smallest distance between cluster centroids, and closest value to +1 for Pearson correlation i.e. best positive relationship between centroids.

A summary of the similarity among A, B, C, D, and H on the Portuguese Dataset is as shown in Table V. Algorithms E, F, and G were not applied to Portuguese dataset due to their implementation limitation as they ran out of memory on a machine with 4 GB RAM. Repeated Bisection and Agglomerative gave 78% (highest) similar clusters. Repeated Bisection was also most similar to Human-labeled actual categories with 66% similarity compared to the rest of the algorithms.

TABLE V
FINAL SIMILARITY AMONG A, B, C, D, AND H ON PORTUGUESE CORPUS

Algorithms	A	B	C	D	H
A	1	0.6634	**0.7813**	0.5963	**0.6634**
B	**0.6634**	1	0.6009	0.5812	0.5967
C	**0.7813**	0.6009	1	0.5919	0.6511
D	**0.5963**	0.5812	0.5919	1	0.5856

The similarity among all 7 (A - G) algorithms and actual categories on Iris dataset is shown in Table VI. Repeated Bisection and Direct algorithms resulted 100% similar clusters while they both gave clusters 95% similar to actual categories. As expected K-means and K-medoids resulted 90% similar clusters, but resulted the clusters that are least similar to human-labeled actual categories.

A. Repeated Bisection (A) vs. Human Categorization (H)

1) Results on Portuguese Dataset

A_0 - A_7 are clusters given by Repeated Bisection Algorithm and H_0 - H_7 are human-labeled actual categories. The Sim(A, H) is 0.6634.

TABLE VII
SIMILARITY MATRIX BETWEEN A AND H

Cl.	H_0 soc	H_1 eco	H_2 clt-soc	H_3 des	H_4 pol	H_5 clt	H_6 com	H_7 opi
A_0	0.0358	**0.5544**	0.0310	0.0104	0.0229	0.0192	0.0514	0.0305
A_1	0.0100	0.0110	0.0991	0.0025	0.0030	0.0136	**0.5737**	0.0100
A_2	0.0987	0.0088	0.1123	0.0056	0.0139	0.0315	0.0197	**0.1217**
A_3	0.0105	0.0033	0.0110	**0.7675**	0.0043	0.0129	0.0361	0.0183
A_4	0.0379	0.0258	0.0105	0.0048	**0.5387**	0.0177	0.0013	0.1372
A_5	**0.2781**	0.0481	0.0539	0.0058	0.0245	0.0235	0.0043	0.1779
A_6	0.0472	0.0037	**0.3282**	0.0062	0.0150	0.0444	0.0100	0.0150
A_7	0.0639	0.0010	0.0173	0.0045	0.0163	**0.5353**	0.0116	0.0433

TABLE VIII
CENTROID EUCLIDIAN DISTANCE BETWEEN A AND H

Cl.	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7
A_0	0.0435	0.0400	0.0408	0.0502	**0.0128**	0.0423	0.0440	0.0398
A_1	0.0406	0.0336	**0.0108**	0.0498	0.0461	0.0418	0.0455	0.0395
A_2	0.0405	**0.0335**	0.0472	0.0503	0.0504	0.0341	0.0438	0.0342
A_3	0.0405	0.0423	0.0460	**0.0069**	0.0498	0.0419	0.0437	0.0398
A_4	0.0360	0.0367	0.0443	0.0440	0.0428	0.0273	**0.0113**	0.0309
A_5	0.0327	0.0292	0.0397	0.0429	0.0395	0.0239	0.0321	**0.0170**
A_6	0.0349	**0.0205**	0.0427	0.0473	0.0474	0.0390	0.0413	0.0333
A_7	**0.0109**	0.0328	0.0417	0.0432	0.0467	0.0341	0.0385	0.0312

TABLE IX
PEARSON CORRELATION BETWEEN A AND H

Cl.	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7
C_0	0.4199	0.5278	0.5595	0.3436	**0.9593**	0.4722	0.4487	0.5171
C_1	0.4533	0.6422	**0.9672**	0.3128	0.4502	0.4416	0.3643	0.4783
C_2	0.4753	**0.6583**	0.3864	0.3172	0.3585	0.6476	0.4271	0.6333
C_3	0.4254	0.3940	0.3790	**0.9864**	0.3372	0.4133	0.3856	0.4409
C_4	0.4843	0.4882	0.3716	0.3938	0.4738	0.7206	**0.9543**	0.6168
C_5	0.5540	0.6611	0.4793	0.4057	0.5438	0.7754	0.6166	**0.8795**
C_6	0.5399	**0.8505**	0.4339	0.3208	0.3691	0.4520	0.4144	0.5766
C_7	**0.9479**	0.5524	0.4062	0.3782	0.3326	0.5237	0.4279	0.5685

2) Results on Iris Dataset

Repeated Bisection (A) did show a slight deviation from a perfect match to human categorization. Clusters A_1 and H_2 showed a distance of 0.1073 and A_2 and H_1 0.0643. Clusters centroid A_0 showed a perfect match with the human labeled centroid H_0. The similarity values of the Pearson correlation coefficient support this as they range from 0.99992-1. This supports the 95% similarity result obtained from our similarity algorithm.

TABLE X
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND H

Cl.	H_0	H_1	H_2
A_0	**0.0000**	3.2082	**4.7545**
A_1	4.6557	1.5174	0.1073
A_2	3.1561	**0.0643**	1.6746

TABLE XI
CORRELATION BETWEEN A AND H

Cl.	H_0	H_1	H_2
A_0	**1.0000**	0.7623	0.6166
A_1	0.6227	0.9809	**0.9999**
A_2	0.7695	**0.9999**	0.9770

B. Repeated Bisection (A) vs. Direct (B)

1) Results on Portuguese Dataset

A_0 - A_7 are clusters given by Repeated Bisection and B_0 - B_7 are clusters given by Direct algorithms. The Sim(A, B) is 0.7813.

TABLE XII
SIMILARITY MATRIX BETWEEN A AND B

Cl.	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7
A_0	**0.8303**	0.0004	0.0000	0.0005	0.0058	0.0605	0.0074	0.0000
A_1	0.0032	**0.8965**	0.0004	0.0000	0.0009	0.0144	0.0038	0.0000
A_2	0.0029	0.0030	0.0081	**0.5000**	0.0219	0.0187	0.0315	0.0110
A_3	0.0017	0.0000	**0.9402**	0.0000	0.0051	0.0031	0.0062	0.0004
A_4	0.0012	0.0000	0.0008	0.0009	**0.5139**	0.0148	0.2877	0.0004
A_5	0.0045	0.0014	0.0022	0.0243	0.0806	**0.5542**	0.0469	0.0056
A_6	0.0151	0.0342	0.0057	**0.2234**	0.0018	0.0345	0.0969	0.0596
A_7	0.0004	0.0004	0.0009	0.0022	0.0059	0.0072	0.0250	**0.8179**

2) Results on Iris Dataset

Clusters A_0 - A_2 are clusters given by Repeated Bisection and B_0 - B_2 are clusters given by Direct clustering algorithm. The similarity between Repeated Bisection and Direct is 1, suggesting 100% similarity between the clusters given by A and B, which infers that Repeated Bisection and Direct algorithms gave clusters with 100% similarity.

TABLE XIII
SIMILARITY MATRIX BETWEEN A AND B

Cl.	B_0	B_1	B_2	B_0	B_1	B_2
	----Fractional values----			----Decimal Values----		
A_0	**50/50**	0/105	0/95	**1.0000**	0.0000	0.0000
A_1	0/105	**55/55**	0/100	0.0000	**1.0000**	0.0000
A_2	0/95	0/100	**45/45**	0.0000	0.0000	**1.0000**

The comparison of the results of the Repeated Bisection and Direct clustering algorithms showed perfect matches with each other once again supporting our 95% similarity comparison result.

TABLE VI
FINAL SIMILARITY AMONG A - G CLUSTERING ALGORITHMS ON IRIS DATASET

Toolbox	gCLUTO				Matlab Fuzzy-clustering toolbox		Weka	Actual
Clustering Algorithm	Repeated Bisection	Direct	Agglomerative	Graph	K-Means	K-Medoid	EM	
Repeated Bisection	**1**	**1**	0.93603	0.74252	0.67408	0.66896	0.84922	**0.95303**
Direct	**1**	**1**	0.93603	0.74252	0.67408	0.66896	0.84922	**0.95303**
Agglomerative	0.93603	0.93603	**1**	0.72289	0.67092	0.66682	0.86789	**0.94505**
Graph	**0.74252**	**0.74252**	0.72289	**1**	0.50520	0.66682	0.67306	0.72250
K-Means	0.67408	0.67408	0.67092	0.50520	**1**	**0.90343**	0.66639	0.67178
K-Medoid	0.66896	0.66896	0.66682	0.66682	**0.90343**	**1**	0.66370	0.66740
EM	0.84922	0.84922	0.86789	0.67306	0.66639	0.66370	**1**	**0.88041**

TABLE XIV
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND B

Cl.	B_0	B_1	B_2
A_0	**0.0000**	4.6557	3.1561
A_1	4.6557	**0.0000**	1.5721
A_2	3.1561	1.5721	**0.0000**

TABLE XV
PEARSON CORRELATION BETWEEN A AND B

Cl.	B_0	B_1	B_2
A_0	**1.0000**	0.6227	0.7695
A_1	0.6227	**1.0000**	0.9786
A_2	0.7695	0.9786	**1.0000**

C. Repeated Bisection (A) vs. Agglomerative (C)

1) Results on Portuguese Dataset

A_0 - A_7 are clusters given by the Repeated Bisection and C_0 - C_7 are clusters given by Agglomerative algorithm. The Sim(A, C) is 0.6078.

TABLE XVI
SIMILARITY MATRIX BETWEEN A AND C

Cl.	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
A_0	0.0376	0.0057	0.0348	0.4578	0.1046	0.0115	0.0291	0.0263
A_1	**0.5085**	0.0037	0.0155	0.0353	0.0197	0.0143	0.0572	0.0231
A_2	0.0111	0.0122	0.0372	0.0130	**0.1867**	0.0330	0.0786	0.0343
A_3	0.0366	**0.6403**	0.0253	0.0058	0.0169	0.0218	0.0211	0.0337
A_4	0.0091	0.0111	**0.3647**	0.0246	0.0735	0.0196	0.0740	0.0347
A_5	0.0321	0.0086	**0.1704**	0.0298	0.1365	0.0142	0.0715	0.0744
A_6	0.0203	0.0053	0.0330	0.0201	0.0372	0.0256	0.0449	**0.2781**
A_7	0.0226	0.0102	0.0313	0.0114	0.0223	**0.3342**	0.1618	0.0641

2) Results on Iris Dataset

Clusters A_0 - A_2 are clusters given by Repeated Bisection and C_0 - C_2 are clusters given by Agglomerative clustering algorithm. The Sim(A, C) is 0.9360. Observe that clusters A_0 and C_0 are 100% similar; however clusters A_1 and C_1, and A_2 and C_2 are 87% and 86% similar respectively, which brought the average similarity down to 93% compared to Repeated Bisection and Direct.

TABLE XVII
SIMILARITY MATRIX BETWEEN A AND C

Cl.	C_0	C_1	C_2	C_0	C_1	C_2
A_0	50/50	0/98	0/102	1.0000	0.0000	0.0000
A_1	0/105	48/55	7/100	0.0000	0.8727	0.0700
A_2	0/95	0/93	45/52	0.0000	0.0000	0.8653

The comparison of the results of the Repeated Bisection and Agglomerative clustering algorithm showed near perfect matches with each other supporting our 94% similarity comparison result.

TABLE XVIII
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND C

Cl.	C_0	C_1	C_2
A_0	0.0000	4.7460	3.2698
A_1	4.6557	0.0946	1.4382
A_2	3.1561	1.6562	0.1407

TABLE XIX
PEARSON CORRELATION BETWEEN A AND C

Cl.	C_0	C_1	C_2
A_0	1.0000	0.6110	0.7616
A_1	0.6227	0.9998	0.9812
A_2	0.7695	0.9755	0.9998

D. Repeated Bisection (A) vs. Graph (D)

1) Results on Portuguese Dataset

A_0 - A_7 are clusters given by Repeated Bisection algorithm and D_0 - D_7 are clusters given by Graph algorithm. The Sim(A, D) is 0.5963.

TABLE XX
SIMILARITY MATRIX BETWEEN A AND D

Cl.	D_0	D_1	D_2	D_3	D_4	D_5	D_6	D_7
A_0	0.0182	0.0022	0.0151	0.0819	0.0437	**0.4552**	0.0121	0.0385
A_1	0.0297	0.0358	0.0069	**0.5432**	0.0160	0.0157	0.0056	0.0208
A_2	0.0240	0.0094	0.0263	0.0229	0.0325	0.0188	0.0088	**0.1887**
A_3	0.0099	0.0535	0.0160	0.0119	0.0156	0.0087	**0.6552**	0.0102
A_4	0.0325	0.0069	0.1788	0.0048	0.0270	0.0648	0.0211	**0.2567**
A_5	0.0276	0.0090	0.0077	0.0165	**0.2869**	0.0588	0.0175	0.1499
A_6	**0.2042**	0.0101	0.1659	0.0175	0.0711	0.0129	0.0063	0.0274
A_7	0.0674	**0.3143**	0.0637	0.0352	0.0299	0.0168	0.0348	0.0703

2) Results on Iris Dataset

Clusters A_0 - A_2 are clusters given by Repeated Bisection and D_0 - D_3 are clusters given by Graph clustering algorithms. The Sim(A, D) is 0.7425. Notice that Graph algorithm suggested 4 clusters instead of requested 3; which significantly brought the overall similarity to 74% though there are two cluster sets that are as high as 100% similar.

TABLE XXI
SIMILARITY MATRIX BETWEEN A AND D

Cl.	D_0	D_1	D_2	D_3
A_0	0.0000	0.0000	0.0000	**1.0000**
A_1	0.3454	**0.6363**	0.0100	0.0000
A_2	0.0000	0.0000	0.0000	**0.9782**

The comparison of the results of the Repeated Bisection and Graph clustering algorithm showed differences with each other supporting our lower 72% similarity comparison result.

TABLE XXII
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND D

Cl.	D_0	D_1	D_2
A_0	4.9639	4.4668	**3.2108**
A_1	0.4564	**0.2868**	1.5085
A_2	1.8694	1.4103	**0.0694**

TABLE XXIII
PEARSON CORRELATION BETWEEN A AND D

Cl.	D_0	D_1	D_2
A_0	0.6044	0.6318	**0.7674**
A_1	0.9996	**0.9998**	0.9793
A_2	0.9731	0.9811	**0.9999**

E. Repeated Bisection (A) vs. K-means (E)

The comparison of the results of the Repeated Bisection and K-means clustering algorithm showed significant difference with each other once again supporting our 67% similarity comparison result.

TABLE XXIV
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND E

Cl.	E_0	E_1	E_2
A_0	0.6700	4.0561	**0.2635**
A_1	4.4354	**0.6187**	4.5921
A_2	2.8878	**0.9544**	3.1245

TABLE XXV
PEARSON CORRELATION BETWEEN A AND E

Cl.	E_0	E_1	E_2
A_0	0.9902	0.6858	**0.9997**
A_1	0.7238	**0.9964**	0.6074
A_2	0.8499	**0.9924**	0.7567

F. Repeated Bisection (A) vs. K-medoids (F)

The comparison of the results of the Repeated Bisection and K-medoids clustering algorithm showed significant difference with each other once again supporting our 66% similarity comparison result.

TABLE XXVI
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND F

Cl.	F_0	F_1	F_2
A_0	**0.3340**	4.0372	0.5154
A_1	4.6003	**0.6400**	4.5303
A_2	3.1419	**0.9325**	2.9908

TABLE XXVII
PEARSON CORRELATION BETWEEN A AND F

Cl.	F_0	F_1	F_2
A_0	**0.9996**	0.6862	0.9956
A_1	0.6016	**0.9964**	0.6925
A_2	0.7520	**0.9924**	0.8255

G. Repeated Bisection (A) vs. EM (G)

The comparison of the results of the Repeated Bisection and EM clustering algorithm showed significant difference with each supporting the 84% similarity comparison result, as it was not quite as bad as K-means, nor as good as algorithms A, B, or C.

TABLE XXVIII
CENTROID EUCLIDEAN DISTANCE BETWEEN A AND G

Cl.	G_0	G_1	G_2
A_0	3.3668	4.9992	**0.0000**
A_1	1.3707	**0.4098**	4.6557
A_2	**0.2328**	1.9494	3.1561

TABLE XXIX
PEARSON CORRELATION BETWEEN A AND G

Cl.	G_0	G_1	G_2
A_0	0.7365	0.6174	**1.0000**
A_1	0.9878	**0.9999**	0.6227
A_2	**0.9986**	0.9773	0.7695

VI. CONCLUSION

The similarity measure that we proposed has experimentally demonstrated consistently similar results to popular measures of Euclidian distance (between cluster centroids) and Pearson correlation. The measure provides the benefit of allowing the aggregated comparison between differing algorithms to allow users to identify the best available clustering algorithm for their applications. Our results show that Repeated Bisection and Direct hierarchical clustering algorithms consistently produced clusters that are most similar to expert labeled categories for both smaller data sets with fewer features (Iris) and large dataset (translated Portuguese-English corpus) with a much larger number of features. Though there remains much room for additional research, our preliminary results indicate that Repeated Bisection and Direct algorithms can be used for clustering both small and large scale datasets, for example, foreign language text document clustering.

With the intriguing initial results, our future work will include expansion and verification of the proposed algorithm through the use of larger datasets along with various feature sizes, sample sizes, and expansion with the numbers of categories to work with.

The techniques can be extended to various real-world problems such as classification and clustering of malware, email analysis (finding social graph among the users based on email contents, for instance) in digital forensics. Since unsupervised clustering algorithms do not give accuracy; the proposed algorithm can be applied to find the best clustering algorithm for many real-life applications where clustering techniques are applied. The approach should enable users to experimentally compare various clustering algorithms and choose the one that best serves the problem.

ACKNOWLEDGMENT

The authors would like to thank Messrs. D. Chen, R. Lopes, R. Koduru, S. Mungala, K. Bondili, and M. Batta for their assistance in translating the Portuguese documents.

REFERENCES

[1] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, 2002, vol 34, No. 1, pp. 1-47.

[2] C. J. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, L. M.Hage, and W. E. Hammond, “Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse,” American Medical Informatics Association Annual Fall Symposium (formerly SCAMC), 1997, pp. 101-5.

[3] K. Seki and J. Mostafa, “An Application of Text Categorization Methods to Gene Ontology Annotation,” Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp. 138-145.

[4] M. Matteucci. (2008). A Tutorial on Clustering Algorithms. Available: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/.

[5] Y. Pen, G. Kou, Y. Shi, and Z. Chen, “Improving Clustering Analysis for Credit Card Accounts Classification,” LNCS 3516, 2005, pp. 548-553.

[6] A. Kalton, K. Wagstaff, and J. Yoo, “Generalized Clustering, Supervised Learning, and Data Assignment,” Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, ACM Press, 2001.

[7] T, Kardi. (2008). Similarity Measurement. Available: http://people.revoledu.com/kardi\/tutorial/Similarity/.

[8] M. K. Sankarapani, R. B. Basnet, S. Mukkamala, A. H. Sung, and B. Ribeiro, “Translation Based Arabic Text Categorization,” Proceedings of Second International Conference on Information Systems Technology and Management, Dubai, March 2008.

[9] Linguateca. (2007). Linguateca. Available: http://www.linguateca.pt /Repositorio/.

[10] Google. (2008). Google Translate. Available: http://translate .google.com/translate_t.
[11] A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository: Iris Data Set. Avaialable: http://www.ics.uci.edu/~mlearn/MLRepository.html.
[12] C. Liao, S. Alpha, and P. Dixon, "Feature Preparation in Text Categorization," Procedings of the Autralasian Data Mining Workshop, Canberra, Australia, 2003.
[13] M. F. Porter, "An Algorithm for Suffix Stripping, Readings in Information Retrieval," Morgan Kaufmann Publishers Inc, 1997.
[14] M. Lan, S.-Y Sung, H.-B. Low, and C.-L. Tan, "A Comparative Study on Term Weighting Schemes for Text Categorization," IJCNN, 2005, vol. 1, pp. 542-545.
[15] C. Liao, S. Alpha, and P. Dixon, "Feature Preparation in Text Categorization," Procedings of the Autralasian Data Mining Workshop, Canberra, Australia, 2003.
[16] G. Karypis. (2008). gCLUTO – Graphical Clustering Toolkit | Karypis Lab. Available: http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview.
[17] J. Abonyi and B. Balasko, B. (2008). Fuzzy Clustering and Data Analysis Toolbox. Available: http://www.fmt.vein.hu/softcomp/fclusttoolbox/.
[18] University of Waikato. (2008). Weka 3 –Data Mining with Open Source Machine Learning Sofware in Java. Available: http://cs.waikato.ac.nz/~ml/weka/.