Geodesic Distances for Clustering Linked Text Data

Selma Tekir
Dept. of Computer Engineering
Izmir Institute of Technology
Izmir, Turkey 35430
Email: selmatekir@iyte.edu.tr

Florian Mansmann
Faculty of Computer Science, Box 78
University of Konstanz
Konstanz, Germany 78457
Email: Florian.Mansmann@uni-konstanz.de

Daniel Keim
Faculty of Computer Science, Box 78
University of Konstanz
Konstanz, Germany 78457
Email: Daniel.Keim@uni-konstanz.de

***Abstract*—The quality of a clustering not only depends on the chosen algorithm and its parameters, but also on the definition of the similarity of two respective objects in a dataset. Applications such as clustering of web documents is traditionally built either on textual similarity measures or on link information. Due to the incompatibility of these two information spaces, combining these two information sources in one distance measure is a challenging issue. In this paper, we thus propose a geodesic distance function that combines traditional similarity measures with link information. In particular, we test the effectiveness of geodesic distances as similarity measures under the space assumption of spherical geometry in a 0-sphere. Our proposed distance measure is thus a combination of the cosine distance of the term-document matrix and some curvature values in the geodesic distance formula. To estimate these curvature values, we calculate clustering coefficient values for every document from the link graph of the data set and increase their distinctiveness by means of a heuristic as these clustering coefficient values are rough estimates of the curvatures.**

To evaluate our work, we perform clustering tests with the k-means algorithm on a subset of the English Wikipedia hyperlinked data set with both traditional cosine distance and our proposed geodesic distance. Additionally, taking inspiration from the unified view of the performance functions of k-means and k-harmonic means, min and harmonic average of the cosine and geodesic distances are taken in order to construct alternate distance forms. The effectiveness of our approach is measured by computing micro-precision values of the clusters based on the provided categorical information of each article.

I. Introduction

The principal aim of the information retrieval systems is to retrieve only the relevant documents from the document collection. In order to determine the relevance, similarity/distance measures are utilized on the document representations. The representations are in terms of some low-level features that are directly measurable from data. Document lengths, the frequency of words in documents are examples of such features.

In the classical IR implementation, feature vectors are formed for documents and distance between these vectors is calculated to relate them. The ordinary distance metric is position-independent in the sense that if two data points are shifted by the same amount in one coordinate the distance between them does not change. In other words, it does not take into account the topology of the document space and assumes that the space is flat (zero curvature). A curvature metric can therefore provide additional information about the data points and is dependent on the position in the space.

In order to address the document semantics in a better way; there is a need for a generalization that captures all the geometric structure of space including the notions of distance, angle, volume, and curvature [1]. The formulation of this generalized metric (metric tensor) varies according to peculiarities of the space. Thus, the distance computation is dependent on the inherent space. This paper proposes a geodesic distance metric that extends the classical distance computation with the measurements of curvatures so that the specificities of the document space can be reflected in a better way.

The geodesic distance metric provides a way of combining features, which can be applied to data sets that offer multiple feature spaces. For our experiments, the selected data set contains linked text documents on which link and text based features can be calculated. The text analysis is conducted using the *Vector Space Model* (VSM) [2] of information retrieval. The term weights are calculated based on term frequencies plus some normalization mechanisms such as *inverse document frequencies* (idf). The regular cosine distance that is applied to the tf-idf version of the term-document vectors is used as the basic similarity measure.

In the computation of the geodesic distances, the cosine distance is combined with the curvature measurements. The curvature values are based on the clustering coefficient values from the link graph given the fact that the clustering coefficient values are rough estimates of the curvatures [3].

The importance of the geodesic distance metric lies in the fact that it utilizes a mathematical cost function for combining links with the text similarity measures. There exists link-based ranking approaches as well as retrieval models incorporating link evidence. However, there is a lack of optimal cost functions to combine cosine, link indegrees, PageRank, etc.

The experiments in this paper are conducted on the Wikipedia XML Corpus [4] English subset. Wikipedia seems a good selection because it is a known fact that in contrast to general web links, Wikipedia links are good indicators of relevance. In addition to this; in Wikipedia, outlinks and inlinks are similar in character and both contribute to the semantic analysis of the documents unlike the Web in which indegrees have a dominant role in determining the semantic relatedness [5]. Thus, the clustering coefficient computations that are based on the undirected link graph of the collection are plausible choices as link-based features for the fact that

there is symmetry in the semantic nature of Wikipedia (if A is relevant to B then B is relevant to A, too).

For measuring the effectiveness of the proposed approach, we use data clustering algorithms. An overwhelming theme for different data clustering techniques/algorithms is to convert the objective into an optimization problem and propose an optimization (performance) function accordingly. The proposed optimization function is expected to measure the goodness of the data analysis objective at hand. Thus, dependable performance functions are of vital importance in the field.

We are given the Wikipedia categorical information as part of the data set. The most common text clustering algorithm *k-means* [6] is used for the tests. The rationale for selecting *k-means* is two-fold. First, as we already know the number of categories to look for in the data set, we easily set the k, the main argument of the algorithm. Second, there exists an abstract framework for integrating multiple feature spaces for the k-means algorithm. The second property can be attributed to the simple but powerful nature of the k-means performance function. For practical reasons the algorithms are run on some subsets of the whole data collection.

The rest of the paper is organized as follows: In Section II, we provide some related work to define the context and give an overview of the state-of-the-art. In Section III, we discuss the geometric meaning of the geodesic distance in comparison to the existing similarity/dissimilarity measures and give the calculation scheme. In Section IV, we provide detailed information about the experimental setup-data set, algorithms, parameters, evaluation metrics and depict the experimental results. Finally, we conclude this paper and raise some issues for future work in Section V.

II. Related Work

Ma et al [7] claim that they are the first researchers that use geodesic distance in text mining related research areas. Their work deals with the query-based sentence retrieval and compares geodesic with cosine distance in this context. The method constructs a graph of all sentences including the candidate (query) ones. In order to define the local neighborhood, a threshold variable epsilon is introduced. If the distance between two sentences are below the threshold value then a direct link is established between them. The geodesic distances are computed over the links by utilizing shortest-path algorithms on the sentence graph. Resulting rankings and correct ratio plots for given queries according to both geodesic distance and cosine are provided. The results show that for the particular values of the parameter epsilon the correct ratio values of the geodesic distance are superior to cosine's, for some other range it degenerates the cosine angle distance.

In hyperbolic IR [8], which is non-Euclidean, a geometric meaning is introduced to the positions in space. The query vector is assumed to be at the center of the hyperbolic sphere and the other documents are evaluated according to their hyperbolic distances to the query vector at the center of the sphere. In short, if a non-Euclidean aspect is to be introduced to a metric space model, the points should be specialized. Another important point is that change of hyperbolic distance according to the radius of the hyperbolic sphere as a parameter introduces equivalent ranking as traditional similarity measures plus weighting schemes.

Xiao et al [9] associate with the geodesic a cost based on length and sectional curvature. The sectional curvature is determined by the degree to which the geodesic bends away from the Euclidean chord. Hence for a geodesic in space, the sectional curvature can be estimated easily if the Euclidean and geodesic distances are known. Put it another way if the Euclidean distance and sectional curvature values are known, geodesic distances can be easily computed. Lou [3] states that *clustering coefficient values* are rough estimates of the sectional curvatures. Getting the sectional curvature values from the link graph, taking cosine from the term-document matrix, the geodesic distance computation can be easily adapted to the text documents with links.

In PageRank [10], global link structure of the document set is utilized to calculate the ranks of the documents. The *authority* concept is introduced in HITS [11] to determine the importance of the documents. An outlink from a source document to a target one means that the source gives some authority to the target. Additionally, it is also critical from whom you get authority. Therefore, there are a set of hub documents from which having inlinks is more valuable. This hub-authority pattern is the key idea and applied in a local context after filtering out documents by text-based queries.

Language models provide mechanisms to utilize link evidences along with the text content scores. The experimental results of Kamps and Koolen [5] show that local degree priors are better than the global degree priors and weighted local/global priors are even more helpful. Thus, the proposed approach is plausible as it presents a compromise between global and local by evaluating local connectivity on the global link graph.

Strehl et al [12] provide a framework for evaluating the impact of similarity measures on clustering web pages. In this work, the fundamental similarity measures are discussed along with their geometric interpretation. The clustering algorithms that are better suited to term-document matrix based text data are determined and the existing similarity/distance measures' performance with these algorithms are stated.

Oikonomakou and Vazirgiannis [13] review web document clustering approaches. They classify the existing algorithms according to characteristics or features that are used. The processed features in the web context are text and/or link-based features. Thus; text-based, link-based, and hybrid approaches exist for web document clustering. This work proposes a hybrid approach for this purpose.

In their work which combines the link-based measures with the content-based classifiers, Calado et al [14] state that the effectiveness of the combination approach may depend on the importance given to each of the sources of evidence to be combined. More weight should be given to those that provide more reliable information. They recognize finding the ideal weights for each of the evidences to be combined as the funda-

mental problem. They set their objective as pursuing methods to automatically determine such weights and alternative ways to combine link-based and content-based evidences.

Yang [15] points out that there is no unanimity in the research findings related to link analysis and/or fusion methods. Some claim that combining results of various retrieval methods is beneficial to retrieval performance, others' results state that fusion in general seemed to decrease retrieval performance. The main question according to Yang is finding out the reason of the general failure of the fusion may be due to the characteristics of test collections, failings of link analysis, inadequacies of fusion formula, or combinations of all or any of the above. He believes the future fusion efforts should focus on discovering the fusion formula that can best realize the fusion potential of combining diverse retrieval methods.

In this work, we attempt to use geodesic distances to better address the semantics of linked text documents. In other words, geodesic distance formula is proposed as a way of combining text-based and linked-based features. This paper is an extended version of the SSCI CIDM 2011 paper [16]. Our extensions include the evaluation of k-harmonic means algorithm to test the effect of initialization in the precision results of cosine and geodesic in k-means algorithm. Moreover, an alternate distance out of cosine and geodesic is calculated by taking the minimum and harmonic average of the given distances.

III. Geodesic Distance

The distinguishing property of the proposed geodesic distance is that local curvature values are considered in the calculation of the distance between the objects. The intuition behind the approach comes from the Riemannian geometry where a local curvature of uniform sign across the manifold implies strong global properties. Thus, we take into consideration the sign of the curvature in the algorithm and are in the pursuit of such global behaviors.

This intuition in mind, we come up with a calculation scheme for geodesic distance. The proposed scheme is based on the relationship between Euclidean and geodesic distances on the unit circle as shown in Figure 1. In order to introduce the geodesic similarity measure for the linked text documents, the formula is formed using the relationship between Euclidean and geodesic distances on the unit circle as shown in Figure 1. The line length between two points on a unit circle represents the Euclidean distance while the arc length between those points represents the geodesic one.

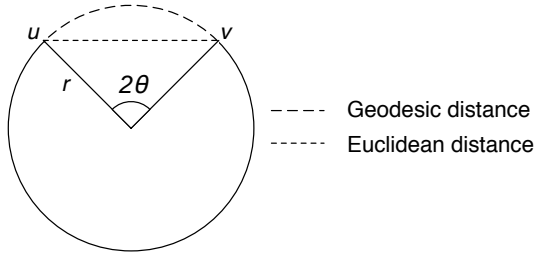

Fig. 1. Euclidean and geodesic distances on a circle.

The line length is computed using the respective triangle and can be stated as follows:

$$d_E(u,v) = 2r\sin\theta \tag{1}$$

The arc length is given by the following formula:

$$d_g(u,v) = 2r\theta \tag{2}$$

The sine in the Euclidean distance formula can be approximated using the Maclaurin series:

$$d_E(u,v) = 2r(\theta - \frac{1}{6}\theta^3 + ...) \tag{3}$$

Substituting for θ obtained from the geodesic distance, we have

$$d_E(u,v) = d_g(u,v) - \frac{d_g^3(u,v)}{24r^2} \tag{4}$$

Finally, radius is represented in terms of the curvature of the circle as follows;

$$r = \frac{1}{\kappa} \tag{5}$$

and the resulting equation is solved for the geodesic distance.

$$d_g(u,v)^3 - 24\frac{1}{\kappa^2}d_g(u,v) + 24\frac{1}{\kappa^2}d_E(u,v) = 0 \tag{6}$$

In this equation, the parameters are dependent on κ and $d_E(u,v)$ values respectively. Thus, geodesic distances can be calculated in terms of κ curvature values and $d_E(u,v)$ Euclidean distances.

As we work with linked text documents in our context, we compute the clustering coefficient values from the link graph to substitute for κ curvature value and replace $d_E(u,v)$ Euclidean distance by the cosine text similarity measure.

The Maclaurin series expansion in equation 3 can be extended by one more term such as the following:

$$d_E(u,v) = 2r(\theta - \frac{1}{6}\theta^3 + \frac{1}{120}\theta^5...) \tag{7}$$

Again by substituting for θ obtained from the geodesic distance, the equation becomes:

$$d_E(u,v) = d_g(u,v) - \frac{d_g^3(u,v)}{24r^2} + \frac{d_g^5(u,v)}{1920r^4} \tag{8}$$

Lastly, radius is replaced by the curvature equivalent and the following quintic equation is obtained:

$$d_g^5(u,v) - \frac{80}{\kappa^2}d_g^3(u,v) + \frac{1920}{\kappa^4}d_g(u,v) - \frac{1920}{\kappa^4}d_E(u,v) = 0 \tag{9}$$

The clustering coefficient is defined as $C_i = 2n/(k_i(k_i-1))$, where n denotes the number of direct links connecting the k_i nearest neighbors of node i [17]. It is the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. The coefficient represents the local connectivity of a document by giving a measure of the degree of interconnectedness in the neighborhood of a node. A node whose neighbors are all

connected to each other has $C = 1$, whereas a node with no links between its neighbors has $C = 0$.

The clustering coefficient provides an approximation of the scalar curvature in the sense that $C = 0$ implies that the scalar curvature at that vertex is negative, while $C = 1$ means that the scalar curvature is positive, with $C = 1/2$ the borderline case of vanishing scalar curvature. In our case, each document is thus assigned one clustering coefficient value.

The geodesic distance equation given in equation 6 is a special cubic equation in which the coefficient of the squared term is zero. In this cubic form there are two complex roots and one real root and the average of the roots of the equation is zero. For solving the equation Cardano's method for cubics is utilized [18]. In our experiments, the real root is used as the geodesic distance. The quintic geodesic distance equation, on the other hand, can be solved numerically by Newton's method.

Geodesic distance can be seen as a weighted distortion measure in the clustering context in that the weights are taken from the link graph and the cosine distance is used as a base distortion measure. Distortion measures are used to evaluate the results of the clustering along with the ground truth categorization. The weighted distortion measures are defined as

$$D^{\alpha}(x, x') = \sum_{l=1}^{m} \alpha_l D_l(F_l, F'_l) \tag{10}$$

where D is the distortion measure, α the weight, x the cluster set, x' the category set, and F_l and F'_l lth feature vectors in the cluster and category sets respectively. The vector of weights is called feature weighting.

Document-cluster set membership and document-category set membership matrices are shown in Table I. At the last row, the corresponding feature vectors are depicted.

TABLE I
THE FEATURE WEIGHTING FOR CLUSTER AND CATEGORY SETS RESPECTIVELY

	Cluster Set					Category Set			
	x_1	x_2	...	x_m		x'_1	x'_2	...	x'_m
d_1	w_{11}	w_{12}	...	w_{1m}	d_1	w_{11}	w_{12}	...	w_{1m}
d_2	w_{21}	w_{22}	...	w_{2m}	d_2	w_{21}	w_{22}	...	w_{2m}
:	:	:	...	:	:	:	:	...	:
d_n	w_{n1}	w_{n2}	...	w_{nm}	d_n	w_{n1}	w_{n2}	...	w_{nm}
	F_1	F_2	...	F_m		F'_1	F'_2	...	F'_m

Distortion measures should be utilized in a way that within category distances must be smaller. In the k-means algorithm, it means that document to cluster centroid distances must be smaller that is k disjoint clusters $\pi_1^{\dagger}, \pi_2^{\dagger}, ..., \pi_k^{\dagger}$ are generated in the manner to minimize the objective function which is given below:

$$\{\pi_u^{\dagger}\}_{u=1}^{k} = \underset{\{\pi_u\}_{u=1}^{k}}{\operatorname{argmin}} \left(\sum_{u=1}^{k} \sum_{x \in \pi_u} D^{\alpha}(x, c_u)\right) \tag{11}$$

c_u thereby denotes the cluster centroids vector.

IV. EXPERIMENTAL EVALUATION

The experiments are conducted on the Wikipedia XML Corpus [4], which is composed of hyperlinked Wikipedia articles. Two category files are also included in the data set. One contains id_document, id_category pairs and the other lists the category names for the defined category ids. In total, there are 659,388 documents in 72 English portal categories in the collection. In Table II, link related statistics (indegree, outdegree, and clustering coefficient) of the data set are provided:

TABLE II
LINK RELATED STATISTICS OF THE WIKIPEDIA XML CORPUS

	min	max	mean	median	stdev
indegree	0	74950	20.9016	4	289.0161
outdegree	0	5176	20.9016	12	37.3416
ccoef	0	1	0.2493	0.2	0.1875

Link related statistics say that the indegrees follow a power-law distribution and the clustering coefficient values have a tendency to be smaller than 0.5, which means that the curvature is mainly negative in the inherent document space.

We randomly selected 10 categories to test our approach. In Table III you find the selected category names along with the corresponding document counts.

The clustering coefficient values are calculated based on the global link graph rather than the link graph for the selected categories because the clustering coefficient values begin to converge when the node count increases. In fact, you cannot get clustering coefficient values other than NaN using the category link graphs as in-category links are quite rare. As for the text part, each document is considered as a multi-dimensional vector and bag-of-words approach with tf-idf is utilized to form final document vectors. In the experimental scenario, as we deal with high dimensional data, clustering algorithms that have to face the curse of dimensionality would not fit the scheme. Thus, the popular k-means algorithm has been chosen. k-means is a good choice because we use two feature sets, namely a) curvature values and b) term-document vectors and there exists an abstract framework for integrating multiple feature spaces in the k-means algorithm [19].

TABLE III
SELECTED CATEGORIES WITH SIZE

Category Name	Size
Bangladesh	393
Colombia	304
Finland	1887
Hong Kong	11056
Morocco	230
Netherlands	1350
New Zealand	2393
Romania	1340
Uganda	232
Venezuela	569
Total	19754

The feature combination is done using a mathematical function to compute geodesic distances by exploiting both

textual information and the link topology. In our approach, we calculate the clustering coefficient values using the whole adjacency matrix of the data set and save these values along with the belonging document ids. As the clustering coefficient values are rough estimates of the curvatures, simple heuristics are applied to them in order to increase their distinctiveness. Algorithm 1 details the heuristic we use to generate the curvature values.

Algorithm 1 generateCurvature algorithm.

```
ccoef: clustering coefficient value
if ccoef ⟩ 0.5 then                    ▷ Curvature is positive?
    ccoef=ccoef-0.5
else
    ccoef=ccoef+1
end if
```

The clustering coefficient values of the documents in the collection are mainly negative with a mean of 0.2. This means that the documents reside in a hyperbolic space. In the heuristic defined in the *generateCurvature* algorithm it is assumed that for negative curvature values which are close to 0.5 (zero curvature), distortion should be greater than for values that are far from 0.5. Thus, 1 is added to the negative curvature values in order to arrange the distortion accordingly. This is consistent with the graph of the distortion of embedding the Internet as a function of the curvature of the embedding space given by [20] in Figure 2. When it comes to positive curvatures, their ordering is preserved and their effect on the centroid curvature is weakened by making a subtraction.

Fig. 2. The distortion of embedding the internet in dimension two as a function of the curvature of the embedding space [20].

After generating the curvature values we need to compute the average of them in order to represent the curvature of the centroids in the k-means clustering algorithm. The centroid curvature is calculated by taking the average of the individual curvature values belonging to the documents that are classified around the same cluster centroid. In this computation, we disregard the NaN values, which are quite rare.

In the experiments, k-means with cosine similarity measure is compared against the k-means with geodesic similarity measure. In order to have a fair comparison we fix the initial cluster assignments. In the cosine case, we run k-means with no initial cluster assignments since the code randomly determines the initial centroids. In the geodesic case, we use the same initial centroids from the cosine case in order to see the effect precisely. In short, in every run the cosine and geodesic share the same initial cluster assignments. However, the initial cluster assignments differ among different runs. We run the experiments 10 times.

We set the number of clusters parameter k as twice the number of categories in order to see the effect more clearly. In the same way, Strehl et al. [12] choose clusters that are twice the number of categories and explain that this setting provides the more natural number of clusters as indicated by preliminary runs and visualization.

The clustering results are evaluated using the metrics rand index [21] and adjusted rand index (AR) that are pair-counting based as well as mutual information [22] [12] and normalized mutual information (NMI), which are information-theoretic measures. In particular, the k-means clustering results are evaluated according to the normalized versions of van Dongen, mutual information and rand index criteria which are stated as the right measures for the algorithm by Wu et al [23]. In the computation of these specified clustering metrics we need the category labels vector and the cluster labels vector as input. As we set the number of clusters for the k-means algorithm to twice the number of categories, while category labels vary between 0 and n, cluster labels have range 0 to 2n. In other words we end up with a contingency table which has n rows (categories) and 2*n columns (clusters). The approach to be taken at this stage to calculate the evaluation metrics for the clustering is complicated. The difficulty lies in determining the criterion to select the n clusters out of 2*n. If you ignore this varying range problem and calculate the metrics accordingly, the clustering quality suffers. If you take the columns (clusters) that have the highest intersection with the categories, it is not fair because in one case the second largest group can be very close in size to the first one whereas in others the gap can be quite big.

The intersection among the selected categories (documents that belong to more than one category) form a small set thus the effect on the clustering can be ignored.

In the evaluation part, we calculate the precision numbers in order to measure the overlap between a given clustering and the ground truth classification. In our case the ground truth classification is given as Wikipedia categories. We comparatively analyze the clustering results for the k-means with cosine and k-means with geodesic with the real categories. The precision computations are done based on the methods provided by [19]. Their work establishes the framework for integrating multiple feature spaces in the k-means clustering algorithm. Thus, valid comparisons between single feature spaces and multiple feature spaces in the k-means case can be best accomplished using the framework's defined precision metrics rather than the traditional clustering metrics for the k-means namely NMI and AR. In our experiments, we also calculated NMI and AR values. The results verify that the order of the NMI and AR values in the cosine and geodesic cases is in accordance with the order of the defined precision

metric values in both cases for every run.

To meaningfully define precision, we convert the clusterings into classification using the following simple rule: identify each cluster with the class that has the largest overlap with the cluster, and assign every element in that cluster to the found class. The rule allows multiple clusters to be assigned to a single class, but never assigns a single cluster to multiple classes.

Suppose there are c classes $\{\omega_i\}_{i=1}^{c} = 1$ in the ground truth classification of n objects. Precision is defined using the following equations where a_i denotes the number of data objects that are correctly assigned to the class ω_i, b_i the documents that are incorrectly assigned to the class ω_i, and c_i denotes the documents that are incorrectly rejected from the class ω_i.

$$p_i = \frac{a_i}{a_i + b_i} \quad and \quad r_i = \frac{a_i}{a_i + c_i}, \quad 1 \leq i \leq n \tag{12}$$

The precision is defined per class. In order to capture the performance averages across classes micro-precision (micro-p) values are calculated as follows:

$$micro - p = \frac{1}{n} \sum_{i=1}^{c} a_i \tag{13}$$

The experimental results (micro-precision values) are shown in Table IV. The first column lists the values belonging to k-means with cosine, the second column k-means with geodesic, the third column k-means with a geodesic derivative, the fourth, min of cosine-geodesic pair, and finally the last one harmonic mean of cosine-geodesic pair respectively. The difference between the two geodesic approaches is in the calculation of the average centroid curvature values. The former one sums the curvature values without paying attention to the signs of the curvature. In the latter one the summation operation takes into account the signs that is the positive ones are added to the sum whereas the negative values are subtracted from it.

k-means' performance function aims at minimizing the total within-cluster variance by the way of minimizing the total mean squared distance for each point and the closest centroid. The closest centroid assignment of a point implies that the algorithm implicitly assigns every point to exactly one cluster, imposing a hard membership for points. k-harmonic means [24], on the other hand, uses the distances to all centroids in order to assign weights to the points and before the final convergence phase there's no assignment to any particular clusters. Therefore, k-harmonic means utilizes soft membership and has the capability of moving points to other cluster centers in the case of high locally dense data points and centers [25].

Inherently, k-means has sensitivity to initialization and k-harmonic means is said to be essentially insensitive to initialization due to the above mentioned capability over k-means. Therefore, it's a good starting point to investigate the effect of initialization on the k-means algorithm in pursuing the factors related to the performance of geodesic over cosine in the experiments. Both cosine and geodesic approaches were run on a k-harmonic means implementation, but no distinguishing difference was observed. Then the effect of initialization can be disregarded in comparing the effectiveness of cosine and geodesic similarity measures in k-means clustering applications.

k-means' performance function given in equation 11 can be rewritten as:

$$\{\pi_u^{\dagger}\}_{u=1}^{k} = \sum_{i=1}^{N} \min(|x_i - c_u|^2 \,|u = 1, ..., k) \tag{14}$$

This new representation is the result of a unified view of the k-means and k-harmonic means' performance functions [24]. The part that comes right after the min, represents the distance function. The min assigns the documents to clusters according to minimum distances. In the k-harmonic case, the harmonic averages (HA) of the distances from each data point to the centers are computed as components to the relative performance function.

Taking inspiration from this rewritten form of the performance function, the min and HA can be evaluated as operators that are applied to the succeeding distance functions. In the context of this paper, these operators can be moved inside and be applied directly to the distance function part as well. As we have cosine and geodesic distances in the experimental setting, by calculating the minimum and harmonic averages of the two distances, an alternate distance form can be generated to be useful. In Table IV, the calculations for both of these variations are listed as the fourth and fifth columns respectively.

In order to compare the effects of the different distance measures, we perform the nonparametric Friedman's test. The test is conducted on three different triples: "cosine-geodesic-geodesic derivative", "cosine-geodesic-min", and "cosine-geodesic-harmonic". The rows are the different runs. The resulting p values of the Friedman's test for the triples are as follows: 0.0608, 0.0450, and 0.0273.

The Friedman's test evaluates the hypothesis that the column effects are all the same against the alternative that they are not all the same. The first result says that the three distance measures are not the same within the 90 % confidence interval. The other two prove that these variations introduce statistically important effects to the already computed values within the 95 % confidence interval. In other words, the methods affect the clustering effectiveness.

When we have a look at the micro-p values given in Table IV, we see that the worst performance for the geodesic cases is in the last run. In order to find out the reason behind that, we expanded Maclaurin series approximation in the equation 3 by one more term ending up with the equation 7. We used Newton's solver to numerically estimate the root of the quintic equation given in equation 9. The clustering results we get show that the quintic geodesic equation improves the results in favor of some specific categories whereas it works against the remaining ones resulting in almost the same micro-p value we have. When we analyze the contingency tables for cosine and

geodesic in every run, we also realize that the geodesic runs' improvements are the results of great performances on those specific categories. Thus, the geodesic approach's effectiveness must have some relation with some category-specific attribute. However, we have not clarified it yet.

According to the mean and standard deviation of the different distance measurements provided in Table IV, the geodesic better expresses the inherent clustering structure of the data (due to higher mean) and at the same time it is more robust as it has less variance.

V. Conclusion

In this work, we propose a novel distance measure for clustering hypertext documents, which is based on both textual information and the link topology of the hypertext document collection. It is useful to highlight the basic components of our approach:

- The basic assumption is that clustering coefficient values that indicate the local connectivity structure of the documents can be used as curvatures. This assumption is based on the fact that clustering coefficients are rough estimates of curvatures [3]. They are computed on the global link graph of the data set.
- The notion of geodesic distances in curved spaces is used to define a mathematical function to do feature combination. For this purpose the geodesic distance calculation scheme on the 0-sphere is utilized.
- Text features are combined with the generated curvature values in order to improve the clustering results in the k-means case. This means integrating multiple feature spaces in the k-means algorithm. One needs a framework that covers the comparative analysis of multiple feature spaces over single feature spaces in the k-means algorithm to test the effectiveness of the candidate functions for feature combination. The abstract framework provided for the feature weighting in the k-means algorithm [19] defines the context and the evaluation methodology for the work.

The experiments are conducted on the Wikipedia XML Corpus English subset [4]. The evaluation metrics are based on the ground truth classification provided as the Wikipedia categorical information. The results show that the curvature values calculated based on the link graph of the data set can be used to fine-tune the similarity values so that the objective function for the clustering can be minimized. Furthermore, the k-means algorithm has proven to be suitable for the proposed geodesic method because of the centroid concept. Using centroid curvature value rather than the individual document clustering coefficient values to fine-tune the cosine is more reasonable as centroid curvature value is a better indicator of locality. Thus, the geodesic approach can be transferred to contexts where there is a multiple feature space, in which one feature can represent curvature, and there is a cumulative calculation potential for this feature.

A. Future Work

The experimental results show that in some runs the geodesic approaches perform better than cosine whereas in some others they are slightly worse. The next task to be done is to discover the factors related to the success or failure of the geodesic method.

In pursuing the factors related to the performance of geodesic over cosine in the experiments, the effect of initialization on the k-means algorithm has been investigated. As previously denoted, cosine and geodesic approaches' outcomes do not have any relationship with the choice of initialization in the k-means algorithm. K-harmonic means implementations were run to analyze the initialization sensitivity and no important improvements were observed in both similarity measures and one's performance in comparison to the other's.

The results motivate alternative computation schemes for geodesic distances. We use geodesic distance computation formula for 0-sphere (circle) in this work. Alternatively, great-circle distances (1-sphere) can be utilized as geodesics. On the other hand; rather than assuming that the space is spherical, taking into consideration the fact that the clustering coefficient average for the data collection coincides with a negative curvature value, the underlying space can be assumed as hyperbolic. Hyperbolic distance calculation schemes in accordance with the given parameters can be devised.

We believe that the heuristics that are applied to the clustering coefficient values in order to generate curvatures can be systematically studied and improved. Furthermore, other linked data sets can be used to further evaluate the effectiveness of the geodesic distance measure.

References

[1] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern Geometry - Methods and Applications: Part I: The Geometry of Surfaces, Transformation Groups, and Fields (Graduate Texts in Mathematics)*. Springer, 1991.

[2] G. Salton, *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley., 1989.

[3] M. Lou, "Traffic pattern in negatively curved network," Ph.D. dissertation, University of Southern California, 2009.

[4] L. Denoyer and P. Gallinari, "The wikipedia xml corpus," *SIGIR Forum*, vol. 40, no. 1, pp. 64–69, 2006.

[5] J. Kamps and M. Koolen, "Is wikipedia link structure different?" pp. 232–241, 2009.

[6] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[7] M. Hui-Fang, H. Qing, and S. Zhong-Zhi, "Geodesic distance based aproach for sentence similarity computation," in *Machine Learning and Cybernetics, 2008 International Conference on*, vol. 5, 2008, pp. 2551–2557.

[8] J. Goth and A. Skrop, "Varying retrieval categoricity using hyperbolic geometry," *Inf. Retr.*, vol. 8, no. 2, pp. 265–283, 2005.

[9] B. Xiao and E. Hancock, "Geometric characterisation of graphs," in *Image analysis and processing: ICIAP 2005, 13th international conference, Cagliari, Italy, September 6-8, 2005*. Springer, 2005, pp. 471–478.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Technical Report, 1999, previous number = SIDL-WP-1999-0120.

[11] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[12] A. Strehl, E. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *In Workshop on Artificial Intelligence for Web Search (AAAI 2000*, 2000, pp. 58–64.
[13] N. Oikonomakou and M. Vazirgiannis, "A review of web document clustering approaches," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2005, pp. 921–943.
[14] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Gonalves, "Combining link-based and content-based methods for web document classification," pp. 394–401, 2003.
[15] K. Yang, "Combining text- and link-based methods for web ir." in *Proceedings of the 10th Text Rerieval Conference (TREC-10).* Washington, DC.: US Government Printing Office., 2001.
[16] S. Tekir, F. Mansmann, and D. Keim, "Geodesic distances for web document clustering," in *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, april 2011, pp. 15 –21.
[17] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998, 10.1038/30918.
[18] J. White and D. Kalman, "Cardano: An adventure in algebra in 8 parts."
[19] D. Modha and W. Spangler, "Feature weighting in k-means clustering," *Machine Learning*, vol. 52, pp. 217–237(21), September 2003.
[20] E. Begelfor and M. Werman, "The world is not always flat or learning curved manifolds." School of Engineering and Computer Science, Hebrew University of Jerusalem., Tech. Rep., 2005.
[21] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
[22] T. Cover and J. Thomas, *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2006.
[23] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," pp. 877–886, 2009.
[24] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means - a data clustering algorithm," 1999.
[25] D. Turnbull, "K-means and k-harmonic means: A comparison of two unsupervised clustering algorithms." University of California San Diego Department of Computer Science and Engineering, Tech. Rep., 2002.

TABLE IV

THE MICRO-P VALUES WITH THE CLUSTERINGS WITH K-MEANS COSINE, GEODESIC, GEODESIC DERIVATIVE, FIRST TWOS' MIN AND HARMONIC MEANS RESPECTIVELY. THE CORRESPONDING MEAN AND STANDARD DEVIATION VALUES ARE ADDED AS THE LAST TWO ROWS.

run ♯	**cosine**	**geodesic**	**geodesic-derivative**	**min**	**harmonic**
1	0,727448	0,737775	0,743191	0,732712	0,737268
2	0,72355	0,733117	0,738585	0,709122	0,729473
3	0,718285	0,724208	0,724866	0,721120	0,721930
4	0,738585	0,74552	0,740812	0,746178	0,746229
5	0,662752	0,676774	0,684823	0,670902	0,673534
6	0,700618	0,706135	0,696973	0,707958	0,713779
7	0,702288	0,705528	0,703756	0,701023	0,698289
8	0,678394	0,683912	0,685633	0,679660	0,684722
9	0,728612	0,728663	0,720462	0,730687	0,728106
10	0,724309	0,708565	0,690088	0,715298	0,706439
mean	0,7105	0,7150	0,7129	0,7115	0,7140
stdv	0,0243	0,0228	0,0234	0,0234	0,0233