Volume 2, Issue 6, June 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Multi-viewpoint Based Similarity Measure and Optimality Criteria for Document Clustering

Saurabh Bhonde[1], P M Chawan[2], Prithviraj Chauhan[3]
Department of Computer Technology
Veermata Jijabai Technological Institute,
Mumbai, India

Abstract—All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a novel multi-viewpoint based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Index Terms—Document clustering, text mining, similarity measure.

1. INTRODUCTION

Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. They can be proposed for very distinct research fields, and developed using totally different techniques and approaches. Nevertheless, according to a recent study, more than half a century after it was introduced; the simple algorithm k-means still remains as one of the top 10 data mining algorithms nowadays. It is the most frequently used partition clustering algorithm in practice. Another recent scientific discussion states that k-means is the favorite algorithm that practitioners in the related fields choose to use. Needless to mention, k-means has more than a few basic drawbacks, such as sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the –art algorithms in many domains. In spite of that, its simplicity, understandability and scalability are the reasons for its tremendous popularity. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity. While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems. A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity among data.

Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. Hence, effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to be data at hand. For instance, the original k-means has sum-of –squared –error objective function that uses Euclidean distance. In a very sparse and high dimensional domain like text documents, spherical k-means, which uses cosine similarity instead of Euclidean distance as the measure, is deemed to be more suitable. In a provocative study, Ahlgren et al. (2003) questioned the use of Pearson's Correlation Coefficient as a similarity measure in Author co Citation Analysis (ACA) with the argument that this measure is sensitive for zeros. Analytically, the addition of zeros to two variables should add to their similarity, but the authors show with empirical examples that this addition can depress the correlation coefficient between these variables. Salton's cosine is suggested as a possible alternative because this similarity measure is insensitive to the addition of zeros (Salton & McGill, 1983). In a reaction White (2003) defended the use of the Pearson correlation hitherto in ACA with the pragmatic argument that the differences between using different similarity measures can be neglected in the research practice. He illustrated this with dendrograms and mappings using

Ahlgren et al.'s own data. Bensman (2004) contributed to the discussion with a letter in which he argued for using Pearson's r for additional reasons. Unlike the cosine, Pearson's r is embedded in multivariate statistics and because of the normalization implied this measure allows for negative values. The problem with the zeros can be solved by applying a logarithmic transformation to the data.

In his opinion, this transformation is anyhow advisable in the case of a vicariate normal distribution. Leydesdorff & Zaal (1988) experimented with comparing results of using various similarity criteria—among which the cosine and the correlation coefficient—and different clustering algorithms for the mapping. Indeed, the differences between using the Pearson's r or the cosine were also minimal in our case. However, our study was mainly triggered by concern about the use of single linkage clustering in the ISI's World Atlas of Science (Small & Sweeney, 1985; Small et al., 1985). The choice for this algorithm had been made by the ISI for technical reasons given the computational limitations of that time. Single linkage clustering is well-known for a tendency to link areas of high density together to one super cluster because of accidental in-between points (Everitt, 1974). I argued that the co citation clusters in the Atlas of Science were sometimes confounded by this so-called effect of 'chaining.' For example, when Small et al. (1985) claimed that the larger part of the natural sciences is 'interdisciplinary' with chemistry 'to be considered the model of an interdisciplinary science,' this result could be considered as an effect of using the wrong algorithm (Leydesdorff, 1987). The differences between using Pearson's Correlation Coefficient and Salton's cosine are marginal in practice because the correlation measure can also be considered as a cosine between normalized vectors (Jones & Furnas, 1987). The normalization is sensitive to the zeros, but as noted this can be repaired by the logarithmic transformation. More generally, however, it remains most worrisome that one has such a wealth of both similarity criteria (e.g., Euclidean distances, the Jaccard index, etc.) and clustering algorithms (e.g., single linkage, average linkage, Ward's mode, etc.) available that one is able to generate almost any representation from a set of data (Oberski, 1988).

The problem of how to estimate the number of clusters, factors, groups, dimensions, etc. is a pervasive one in multivariate analysis. If there are no a priori theoretical reasons—as is usually the case in exploratory uses of these techniques—such decisions tend to remain somewhat arbitrary. In factor analysis, methods such as visual inspection of the screen plot or a cut-off at certain eigenvalues are common practice. In cluster analysis and multi-dimensional scaling, decisions based upon visual inspection of the results are common. Small & Sweeney (1985), for example, have proposed 'variable level clustering,' that is, in essence the adaptation of the clustering level to the density of the cluster involved; the search for a formal criterion is thus replaced by a procedural one. This practice was later implemented in the French system for co-word clustering LEXIMAPPE (Callon et al., 1986; Courtial, 1989), but the results of this system could not be validated when using Shannon's (1948) information theory (Leydesdorff, 1992). Information theoretical approach Can an exact solution be provided for the problem of the decomposition? I submit that information theory can be elaborated into statistical decomposition analysis (Theil, 1972) and that this methodology provides us with clear criteria for the dividedness (Leydesdorff, 1991, 1995).

Dividedness and aggregation can both be expressed in terms of bits of information. I shall show that a function for the dividedness can then be maximized. In general, disaggregation of a set in g groups can be described with the following formula:

$$\mathbf{H = H0 + \Sigma g\ Pg\ Hg} \qquad \mathbf{(1)}$$

in which H is the expected information content (probabilistic entropy) of the aggregated distribution, and Pg the probability of each of the groups which as a subset has an uncertainty equal to the respective Hgs. The 'in between group entropy' H0 is a measure of the specificity that prevails at the level of the subsets, and thus it should be possible to use it as a measure for the quality of clustering. The right-hand term of the above equation (Σg Pg Hg) is equal to the entropy of a variable (n) under the condition of a grouping variable (m): H(n|m). The left-hand term of Equation 1, H0, is therefore, equal to H(n) - H(n|m), which is the uncertainty in n that is not attributable to the uncertainty within the groups, or in other words the transmission (mutual information) of the grouping variable m to n (that is to be grouped). The larger this transmission, the more reduction of uncertainty there will be among the groups, and therefore the better the groups will be in terms of the homogeneity of their distributions. However, by definition:

$$\mathbf{H(n|m) = H(n, m) - H(m)} \qquad \mathbf{(2)}$$

Since H0 = H (n) - H (n|m) (see above), this implies:

$$\mathbf{H0 = H(n) + H(m) - H(n, m)} \qquad \mathbf{(3)}$$

In other words, the increase of H0 if one distinguishes an additional group (cluster, factor, etc.) is composed of a part that is dependent only on the grouping variable (H (m)), and a part which is dependent on the interaction between the grouping variable m and the grouped variable n. The interaction between the two variables makes H (n, m) smaller than the sum of H (n) and H (m). Given a number of variables n to be grouped, the question thus becomes: for which value of m does the function {H (m) - H (n, m)}, and consequently H0 as an indicator of the dividedness, reach a maximum? This problem can be solved numerically by recursive reallocation of the
Cases into all possible groupings.

The (normalized) maximization of the H0 thus provides an unambiguous criterion for the quality of the attribution. Q.e.d. Let me formulate the argument also more intuitively: If we divide one group into two subgroups i and j, using Hij = H0 + Pi Hi + Pj Hj, the

aggregated Hij may be larger than both Hi and Hj, or larger than one of them and smaller than the other. (The two groups cannot be both larger than Hij, since the 'in-between group uncertainty' H0 is necessarily larger than or equal to zero.) The case of Hi < Hij < Hj corresponds to the removal of the more than average heterogeneous case (s) into a separate subgroup: therefore, this new subgroup has a higher uncertainty, and the remaining subgroup becomes more homogeneous than The original group. This is always possible, but it is not clustering. Clustering entails by definition the notion of reducing uncertainty in both subgroups. Therefore, we may define 'divisive clustering' as the case where both new subgroups have a lower expected information content than the undivided group. Note that the above justification of the division is based only on the right-hand term of the formula for disaggregation (ΣgPgHg in Equation 1). The value of the left-hand term (H0), however, is sensitive both to the number of groups—since each further division adds to H0 unless the two groups have similar Hgs—and to the quality of the attribution of cases to groups given a certain number of groups.

These two questions:

1. Concerning the number of groups.

2. Concerning the attribution of cases to groups—can be studied Independently

2. EXISTING SYSTEM

Data compression can reduce the storage and energy consumption for resource-constrained applications. In [1], Distributed source coding uses joint entropy to encode two nodes data individually without sharing any data between them; however, it requires prior knowledge of cross correlations of sources. Other works, such as [2, 4], combine data compression with routing by exploiting cross correlations between sensor nodes to reduce the data size.

In [5], a tailed LZW has been proposed to address the memory constraint of a sensor device. Summarization of the original data by regression or linear modeling has been proposed for trajectory data compression [3, 6]. However, the above works do not address application-level semantics in data, such as the correlations of

A group of moving objects, which we exploit to enhance the compressibility.

Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize clustering algorithms published every year. Existing Systems greedily picks the next frequent item set which represent the next and some remaining item sets. In other words, the clustering result depends on the order of picking up the item sets, which in turns depends on the greedy heuristic. This method does not follow a sequential order of selecting clusters. Instead, we assign documents to the best cluster. The cosine similarity in Eq. (3) can be expressed in the Following form without changing its meaning:

$$\mathbf{Sim(d_i, d_j) = Cos(d_i-0, d_j-0) = (d_i-0)^t\,(d_j-0)}$$

where 0 is vector 0 that represents the origin point. According to this formula, the measure takes 0 as one and only reference point. The similarity between two documents di and dj is determined w.r.t. the angle between the two points when looking from the origin. To construct a new concept of similarity, it is possible to use more than just one point of reference. We may have a more accurate assessment of how close or distant a pair of points is, if we look at them from many different viewpoints. From a third point dh, the directions and distances to di and dj are indicated respectively by the difference vectors (di − dh) and (dj − dh). By standing at various reference points dh to view di, dj and working on their difference vectors, we define similarity between the two documents as:

$$\mathbf{Sim(di, dj)} = \frac{\mathbf{1}}{\mathbf{n - nr}} \sum_{\mathbf{dh \in s/sr}} \mathbf{(Sim(di - dh, dj - dh)}$$

As described by the above equation, similarity of two documents di and dj - given that they are in the same cluster - is defined as the average of similarities measured relatively from the views of all other documents outside that cluster. What is interesting is that the similarity here is defined in a close relation to the clustering problem.

A presumption of cluster memberships has been made prior to the measure. The two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster. We call this proposal the Multi-Viewpoint based Similarity, or MVS. From this point onwards, we will denote the proposed similarity measure between two document vectors di and dj by MVS(di, dj |di, dj∈ Sr),
or occasionally MVS(di, dj) for short. The final form of MVS in depends on particular formulation of the individual similarities within the sum. If the relative similarity is defined by dot-product of the difference vectors, we have:

$$\mathbf{MVS(di, dj\,|di,\ dj \in Sr),}$$

$$= \frac{\mathbf{1}}{\mathbf{n-nr}} \sum_{\mathbf{dh \in s/sr}} \mathbf{(di - dh)\ (dj - dh)}$$

$$= \frac{\mathbf{1}}{\mathbf{n-nr}} \sum_{\mathbf{dh}} \mathbf{(cos(di - dh, dj - dh)\,||di - dh||dj - dh}$$

The similarity between two points di and dj inside cluster Sr, viewed from a point dh outside this cluster, is equal to the product of the cosine of the angle between di and dj looking from dh and the Euclidean distances from dh to these two points. This definition is based on the assumption that dh is not in the same cluster with di and dj. The smaller the distances _di−dh_ and _dj

–dh_ are, the higher the chance that dh is in fact in the same cluster with di and dj, and the similarity based on dh should also be small to reflect this potential. Therefore, through these distances, Eq. also provides a measure of inter cluster dissimilarity, given that points di and dj belong to cluster Sr, where as dh belongs to another cluster.

The overall similarity between di and dj is determined by taking average over all the viewpoints not belonging to cluster Sr. It is possible to argue that while most of these viewpoints are useful, there may be some of them giving misleading information just like it may happen with the origin point. However, given a large enough number of viewpoints and their variety, it is reasonable
to assume that the majority of them will be useful.
Hence, the effect of misleading viewpoints is constrained and reduced by the averaging step. It can be seen that this method offers more informative assessment of similarity than the single origin point based similarity measure.

3. PROPOSED WORK

The main work is to develop a novel hierarchal algorithm for document clustering which provides maximum efficiency and performance. It is particularly focused in studying and making use of cluster overlapping phenomenon to design cluster merging criteria. Proposing a new way to compute the overlap rate in order to improve time efficiency and "the veracity" is mainly concentrated. Based on the Hierarchical Clustering Method, the usage of Expectation-Maximization (EM) algorithm in the Gaussian Mixture Model to count the parameters and make the two sub-clusters combined when their overlap is the largest is narrated.

In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. The generalization of optimization theory and techniques to other formulations comprises a large area of applied mathematics. More generally, optimization includes finding "best available" values of some objective function given a defined domain, including a variety of different types of objective functions and different types of domains. Such a formulation is called an optimization problem or a mathematical programming problem (a term not directly related to computer programming , but still in use for example in linear programming Many real world and theoretical problems may be modeled in this general framework. Problems formulated technique as energy minimization, speaking of the value of the function f as representing the energy of the system being modeled. Typically, A is some subset of the Eucliden Rn, often specified by a set of constraints , equalities or inequalities that the members of A have to satisfy. The domain A of f is called the search space or the choice set, while the elements of A are called candidate (maximization), or, in certain fields, energy function, or energy function. A feasible solution that minimizes (or maximizes, if that is the goal) the objective function is called an optimal solution. By convention, the standard form of an optimization problem is stated in terms of minimization. Generally, unless both the objective function
and the feasible region are convex in a minimization problem, there may be several local minima, where a local minimum x* is defined as a point for which there exists some $\delta > 0$ so that for all x such that
$\|X - X^*\| <= \partial$;
the expression
$f(x^*) <= f(x)$
holds; that is to say, on some region around x* all of the function values are greater than or equal to the value at that point. Local maxima are defined similarly. A large number of algorithms proposed for solving non-convex problems – including the majority of commercially available solvers – are not capable of making a distinction between local optimal solutions and rigorous
optimal solutions, and will treat the former as actual solutions to the original problem. The branch of applied mathematics is concerned with the development of deterministic algorithms that are capable of guaranteeing convergence in finite time to the actual optimal solution of a non-convexproblem is called global optimization.

4. CONCLUSIONS

The future methods could make use of the same principle, but define alternative forms for the relative similarity or do not use average but have other methods to combine the relative similarities according to the different viewpoint. In future, it would also be possible to apply the proposed criterion functions for hierarchical clustering algorithms. It would be interesting to explore how they work types of sparse and high-dimensional data.

References

[1] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg,"Top 10 algorithms in data mining", Knowl. Inf. Syst., Vol. 14, No. 1, pp. 1–37, 2007.

[2] I. Guyon, U. von Luxburg, R. C. Williamson, "Clustering: Science or Art?", NIPS'09 Workshop on Clustering Theory,2009.

[3] I. Dhillon, D. Modha,"Concept decompositions for large sparse text data using clustering", Mach. Learn., Vol. 42, No. 1-2, pp. 143–175, 2001.

[4] S. Zhong,"Efficient online spherical K-means clustering", in IEEE IJCNN, 2005, pp. 3180–3185.

[5] A. Banerjee, S. Merugu, I. Dhillon, J. Ghosh,"Clustering with Bregman divergences", J. Mach. Learn. Res., Vol. 6,
pp. 1705–1749, Oct 2005.

[6] E. Pekalska, A. Harol, R. P. W. Duin, B. Spillmann, H. Bunke,"Non-Euclidean or non-metric measures can be informative", in Structural, Syntactic, and Statistical Pattern
Recognition, ser. LNCS, Vol. 4109, 2006, pp. 871–880.

[7] M. Pelillo,"What is a cluster? Perspectives from game theory", in Proc. of the NIPS Workshop on Clustering Theory,2009.

[8] D. Lee, J. Lee,"Dynamic dissimilarity measure for support based clustering", IEEE Trans. on Knowl. and Data Eng., Vol. 22, No. 6, pp. 900–905, 2010.

[9] A. Banerjee, I. Dhillon, J. Ghosh, S. Sra,"Clustering on the unit hypersphere using von Mises-Fisher distributions", J. Mach. Learn. Res., Vol. 6, pp. 1345–1382, Sep 2005.

[10] W. Xu, X. Liu, Y. Gong,"Document clustering based on nonnegative matrix factorization", in SIGIR, 2003, pp. 267–273.

