Review of "A Comparison of Two Novel Algorithms for Clustering Web Documents"

George Stoica

Abstract

"A Comparison of Two Novel Algorithms for Clustering Web Documents"[1] investigates different approaches to clustering web documents using the k-means clustering algorithm. It studies the global k-means method and the graphs instead of the vector model. The results are compared with previous baselines, the k-means with random initialization and the vector model.

1 Overview

1.1 Selected problem

The problem that the authors of the article try to solve is the clustering of web documents. Due to the huge collection of web documents available it is virtually impossible to organize them in a supervised manner, therefore here comes the need of an unsupervised method of ordering such documents such that they are easily browsed by users.

1.2 Place in literature

In this article, the authors approach the problem of clustering web documents using the k-means algorithm. The k-means algorithm consists of assigning elements to k clusters using distance or similarity as a criterion of selection, then calculating each cluster's centroid based on its elements, and repeating these two steps until convergence.

In the classical k-means, the first k centroids are randomly initialized, which may lead the k-means clustering to a local minima. The global k-means[2], which appeared in 2003, the same year as [1], comes with a method of initializing the centroids such that it improves the clustering done by the k-means algorithm. The initialization is done by randomly selecting the first centroid, and subsequently selecting centroids that are far from all the previously selected centroids. This leads to a better initial distribution of elements in each cluster which may avoid local minima.

In previous works, the elements which were clustered using k-means were mostly represented using a vector of features. However, a new method of representing data using graphs[3] was introduces in 2003, by the same authors of this article. In [3], what they did was only to introduce the method and examine one dataset with a single performance measure.

Therefore, in the current presented work[1], the authors combine the global k-means from [2] with the graph-based representation from [3] in order to do a comparison of this and the classical approach. Moreover, this marks the first time this two methods were used together for clustering web documents while previous works they mention[4][5] use k-means with random initialization.

2 Approach

The experiments are performed on two web data sets called F-series (slightly altered) and J-series. F-series consists of 93 web documents from four classes, and J-series consists of 185 web documents from ten classes.

In order to measure performance, the authors use two different criteria. The first criterion is the Rand index[6], in which all possible pairs are examined and divided into two classes, agreements and disagreements. An agreement is when the two objects of the pair are in the same class, or when the two objects are in different clusters both the ground truth and the examined clustering. Disagreements are when the two objects are in the same cluster in the ground truth and are not in the same cluster in the examined cluster, or vice-versa. The second criterion is mutual information[7].

Experiments have shown that global k-means consistently outperforms the random initialization one, while the graph based approach performs better in the majority of cases even when using ten-nodes graphs. However, execution time is always shorter for the vector-based representation when using random initialization, but when using global k-means, the ten-nodes graphs have a lower execution time than vectors, due to the fact that once the global clusters were computed, they can be re-used for incremental clustering.

3 Critique

I consider this research to be worthwhile as it combines two new approaches in the literature at that time and it provides practical usage of the two theoretical methods. Concrete results often prove theoretical results to be useful and may guide other people to use or improve similar methods.

As I previously mentioned, the authors claim to be the first to combine the two methods and give an actual comparison, and the current article comes as a continuation of their prior work [3] in which they introduce a new method of graph-based representation.

Nevertheless, I think that the authors should have provided results of another approach apart from k-means, for a wider perspective of the problem, as one cannot understand whether k-means is the best method of clustering web documents or whether there are other possible algorithms which can give good results.

4 Conclusion

In conclusion, "A Comparison of Two Novel Algorithms for Clustering Web Documents" illustrates results of combining two variants of the k-means clustering algorithm, respectively the global k-means and the graph-based representation. Combining these two provides considerable improvement in accuracy and comes with an improvement in time efficiency for datasets which need to be reclustered frequently.

References

[1] Adam Schenker, Mark Last, Horst Bunke, and Abraham Kandel. A comparison of two novel algorithms for clustering web documents. In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, pages 71–74, 2003.

[2] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[3] Adam Schenker, Mark Last, Horst Bunke, and Abraham Kandel. Clustering of web documents using a graph model. In *Web Document Analysis: Challenges and Opportunities*, pages 3–18. World Scientific, 2003.

[4] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Workshop on artificial intelligence for web search (AAAI 2000)*, volume 58, page 64, 2000.

[5] Gerard Salton. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, 169, 1989.

[6] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[7] Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of Information Theory*, 1(1):279–335, 1991.