https://doi.org/10.33472/AFJBS.6.6.2024.9325-9335

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Advanced Clustering Techniques: Graph Partitioning and Adaptive Algorithms for Enhanced Document Classification

Dr. M .Meena Krithika*

*Assistant Professor, Department of Computer science, NGM College, Pollachi

Article Info

Volume 6, Issue 6, September 2024

Received: 26 July 2024

Accepted: 28 August 2024

Published: 19 September 2024

doi: 10.33472/AFJBS.6.6.2024.9325-9335

ABSTRACT:

Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters. The goal is to create clusters that are coherent internally, but substantially different from each other. In plain words, objects in the same cluster should be as similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in the other clusters. Automatic Text clustering has played an important role in many fields like information retrieval, data mining, etc. The aim of this thesis is to improve the efficiency and accuracy of document clustering. We discuss two clustering algorithms and the fields where these perform better than the known standard clustering algorithms. The first approach is an improvement of the graph partitioning techniques used for Text clustering. In this we preprocess the graph using a heuristic and then apply the standard graph partitioning algorithms. This improves the quality of clusters to a great extent. The second approach is a completely different approach in which the words are clustered first and then the word cluster is used to cluster the documents. The adaptive adjustment of the damping factor to eliminate oscillations (called adaptive damping), adaptive escaping oscillations, and adaptive searching the space of preference parameter to find out the optimal clustering solution suitable to a data set (called adaptive preference scanning). With these adaptive techniques, adaptive AP will outperform SAP algorithm in clustering quality and oscillation elimination, and it will find optimal clustering solutions. This reduces the noise in data and thus improves the quality of the clusters.

Keywords: Automatic Text clustering, document clustering, partitioning techniques, Text clustering

1. Introduction

Clustering, a vital unsupervised learning technique, plays a significant role in organizing objects into groups or clusters where members of the same group are similar, while members of different groups are distinctly dissimilar[1]. This technique has become increasingly important in various fields, especially those dealing with large, unstructured data sets, such as information retrieval, natural language processing, and data mining. In the realm of text data, document clustering is a crucial application, used to automatically organize vast amounts of textual information into meaningful groups. By doing so, it enhances efficiency in tasks such as search engine optimization, topic modeling, and document categorization[2].

However, traditional clustering algorithms like k-means, hierarchical clustering, and others often face limitations when applied to textual data[13]. The high dimensionality, sparsity, and inherent noise in text data can lead to poor cluster quality, inefficiency, and sensitivity to parameter choices. These challenges have prompted ongoing research to develop more robust, accurate, and efficient clustering techniques tailored for text data[14].

The focus of this thesis is to explore and propose two novel approaches to improve the efficiency and accuracy of document clustering. The first approach is an enhancement of graph partitioning techniques, which are widely used in text clustering due to their ability to model relationships between documents as graphs[15]. In this approach, a heuristic-based preprocessing step is introduced to refine the graph before applying standard partitioning algorithms. By optimizing the graph structure in this way, the quality of the resulting clusters is significantly improved, leading to more meaningful and coherent document groupings[16].

The second approach takes an innovative direction by first clustering words, rather than documents, and then using these word clusters as the basis for document clustering[17]. This word-first clustering method aims to reduce noise and capture more fine-grained semantic relationships within the text. By focusing on word clusters, it addresses the issue of high dimensionality in text data and ensures that documents are grouped based on more meaningful content relationships[18].

In addition to these two primary approaches, the thesis also introduces adaptive mechanisms into Affinity Propagation (AP), a clustering algorithm known for its ability to identify exemplars without requiring a predetermined number of clusters[19]. By incorporating adaptive techniques such as adjusting the damping factor to eliminate oscillations and adaptively searching for the optimal preference parameter, the proposed adaptive AP method achieves higher clustering quality while reducing computational issues that affect standard AP, such as oscillation and suboptimal clustering solutions.

The contributions of this work lie in developing these adaptive and heuristic-based techniques to not only improve clustering accuracy but also enhance the ability to handle noisy and complex text data. By leveraging these methods, this thesis aims to advance the state of document clustering and provide more effective solutions for organizing and analyzing large-scale textual data.

This research presents two key innovations: an improved graph partitioning technique with heuristic preprocessing and a word-first clustering approach, both of which aim to address the limitations of traditional algorithms. Additionally, the introduction of adaptive Affinity Propagation offers a dynamic and robust clustering solution that further enhances performance across diverse data sets. These contributions are expected to significantly impact the field of document clustering, offering improved methods for handling complex text data in various real-world applications[20].

2. Existing work

B.J. Frey and D. Dueck [3] proposed Clustering data by identifying a subset of representative examples is important for processing sensory signals and detecting patterns in data. Such "exemplars" can be found by randomly choosing an initial subset of data points and then iteratively refining it, but this works well only if that initial choice is close to a good solution. We devised a method called "affinity propagation," which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. We used affinity propagation to cluster images of faces, detect genes in microarray data, identify representative sentences in this manuscript, and identify cities that are efficiently accessed by airline travel. Affinity propagation found clusters with much lower error than other methods, and it did so in less than one-hundredth the amount of time.

T.Y. Jiang and A. Tuz [4] proposed it is crucial to segment customers intelligently in order to offer more targeted and personalized products and services. Traditionally, customer segmentation is achieved using statistics-based methods that compute a set of statistics from the customer data and group customers into segments by applying clustering algorithms. Recent research proposed a direct grouping-based approach that combines customers into segments by optimally combining transactional data of several customers and building a data mining model of customer behavior for each group. This paper proposes a new micro targeting method that builds predictive models of customer behavior not on the segments of customers but rather on the customer-product groups. This micro-targeting method is more general than the previously considered direct grouping method. We empirically show that it significantly outperforms the direct grouping and statistics-based segmentation methods across multiple experimental conditions and that it generates predominately small-sized segments, thus providing additional support for the micro-targeting approach to personalization.

W.H. Wang, H.W. Zhang, F. Wu, and Y.T. Zhuang [5], proposed E-learning resources increase vastly with the pervasion of the Internet. Thus, the retrieval of e-learning resources becomes more important. However, the typical lexical matching could not satisfy the users' underlying intention. We adapted affinity propagation in MapReduce framework to make semantic retrieval applicable to large-scale data, and with this parallel affinity propagation we proposed an approach to retrieve e-learning materials efficiently, which could retrieve semantically relevant materials utilizing conceptual topics produced in advance.

F. Sebastiani [6], proposed the automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely, document representation, classifier construction, and classifier evaluation.

F. Wang and C.S. Zhang [7], proposed A novel semi-supervised learning approach is proposed based on a linear neighbourhood model, which assumes that each data point can be linearly reconstructed from its neighborhood. Our algorithm, named Linear Neighborhood Propagation (LNP), can propagate the labels from the labeled points to the whole dataset using these linear neigh-borhoods with sufficient smoothness. We also derive an easy way to extend LNP to out-of-sample data. Promising experimental results are presented for synthetic data, digit and text classification tasks.

L. Bottou and Y. Bengio [9], proposed the studies the convergence properties of the well known K-Means clustering algorithm. The K-Means algorithm can be described either as a gradient descent algorithm or by slightly extending the mathematics of the EM algorithm to this hard threshold case. We show that the K-Means algorithm actually minimizes the quantization error using the very fast Newton algorithm.

H. Zha, C. Ding, M. Gu, X. He, and H.D. Simon[10], Proposed the popular K-means clustering partitions a data set by minimizing a sum-of-squares cost function. A coordinate descend method is then used to and local minima. In this paper we show that the minimization can be reformulated as a trace maximization problem associated with the Gram matrix of the data vectors. Furthermore, we show that a relaxed version of the trace maximization problem possesses global optimal solutions which can be obtained by computing a partial eigendecomposition of the Gram matrix, and the cluster assignment for each data vectors can be found by computing a pivoted QR decomposition of the eigenvector matrix. As a by-product we also derive a lower bound for the minimum of the sum-of-squares cost function.

S. Dumais, J. Platt, D. Heckerman, and M. Sahami[11], Proposed the assignment of natural language texts to one or more predefined categories based on their content – is an important component in many information organization and management tasks. We compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed, and classification accuracy. We also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear Support Vector Machines (SVMs) are particularly promising because they are very accurate, quick to train, and quick to evaluate.

Z.H. Zhou and M. Li, [12], Proposed Text categorization is the task of assigning prede-fined categories to natural language text. With the widely used 'bag of words' representation, previous researches usually assign a word with values such that whether this word appears in the document concerned or how frequently this word appears. Although these values are useful for text categorization, they have not fully expressed the abundant information contained in the document. This paper explores the effect of other types of values, which express the distribution of a word in the document. These novel values assigned to a word are called distributional features, which include the compactness of the appearances of the word and the position of the first appearance of the word. The proposed distributional features are exploited by a tfidf style equation and different features are combined using ensemble learning techniques. Experiments show that the distributional features are useful for text categorization. In contrast to using the traditional term frequency values solely, including the distributional features requires only a little additional cost, while the categorization performance can be significantly improved. Further analysis shows that the distributional features are especially useful when documents are long and the writing style is casual.

3. Proposed Methodology

In semisupervised clustering, the main goal is to efficiently cluster a large number of unlabeled objects starting from a relatively small number of initial labeled objects. Given a few initial labeled objects, we would like to use them to construct efficient initial "seeds" for our Affinity Propagation clustering algorithm. To guarantee precision and avoid a blind search for seeds and imbalance errors, we present in the following a specific seeds' construction method, that we named Mean Features Selection. Let N^O, N^F, N^D, and F^C represent, respectively, the object number, the feature number, the most significant feature number, and the feature set of cluster c in the labeled set (they can be searched by viewing each object in cluster c). Suppose F is the feature set and DF is the most significant feature set of seed c (for example, DF of this manuscript could be all the words (except stop words) in the title, i.e., {text, clustering, seed, Affinity, and Propagation}).

The seeds' construction method is prescribed as

1. iff

$$n_{k'} \geq \frac{\sum_{k=1}^{N^F} n_k}{N^o}, \quad f_{k'} \in F; \tag{1}$$

2. iff

$$n_{DK'} \geq \frac{\sum_{k=1}^{N^D} n_{DK}}{N^o}, \quad f_{k'} \in DF. \tag{2}$$

(i) Similarity Measurement

The similarity measurement plays an important role in Affinity Propagation clustering. In order to give specific and effective similarity measurement for our particular domain, i.e., text document, we introduce the following feature sets: the Cofeature Set, the Unilateral Feature Set, and the Significant Cofeature Set. To define these sets, we first detail the computations of the new features. In our approach, each term in text is still deemed as a feature and each document is still deemed as a vector [8]. However, all the features and vectors are not computed simultaneously, but one at a time.

(ii)Seeds Affinity Propagation with exemplar

Clustering by Passing Messages Between Data Points. Science 315, 972 (2007)". It has some advantages: speed, general applicability, and suitable for large number of clusters. AP has two limitations: it is hard to known what value of parameter 'preference' can yield optimal clustering solutions, and oscillations cannot be eliminated automatically if occur.

Adaptive AP improves AP in these items: adaptive adjustment of the damping factor to eliminate oscillations (called adaptive damping), adaptive escaping oscillations, and adaptive searching the space of preference parameter to find out the optimal clustering solution suitable to a data set (called adaptive preference scanning). With these adaptive techniques, adaptive AP will outperform AP algorithm in clustering quality and oscillation elimination. The clusters data, using a set of real-valued pair wise data point similarities as input. Clusters are each represented by a cluster center data point (the "exemplar"). The method is iterative and searches for clusters so as to maximize an objective function, called net similarity.

For N data points, there are potentially N^2-N pairwise similarities; this can be input as an N-by-N matrix 's', where s(i,k) is the similarity of point i to point k (s(i,k) needn't equal s(k,i)).

In fact, only a smaller number of relevant similarities are needed; if only M similarity values are known (M < N^2-N) they can be input as an M-by-3 matrix with each row being an (i,j,s(i,j)) triple. The algorithm automatically determines the number of clusters based on the input preference 'p', a real-valued N-vector. p(i) indicates the preference that data point i be chosen as an exemplar. Often a good choice is to set all preferences to median(s); the number of clusters identified can be adjusted by changing this value accordingly.

4. Experimental Results

We have produced for five datasets using two algorithms. The first algorithms are the algorithms proposed by us to be better and the second one is the standard K-Means algorithm. The results for the K-Means algorithm have been generated to compare them with the proposed algorithms.

(i) **Datasets**
The Five datasets are,
1. Wine
2. DocumentSummarization
3. TravelRouting
4. 20 NewsGroups
5. Reuters 21578

(ii) 20 newsgroups
This is a very standard and popular dataset used for evaluation of many text applications, data mining methods, machine learning methods, etc.
Its details are as follows:
- Number of unique documents = 18,828
- Number of categories = 20
- Number of unique words after removing the stopwords = 71,830

(iii)Reuters -21578
This is the most common dataset used for evaluation of document categorization and clustering.
Its details are as follows:
- Number of unique documents = 19715
- Number of categories = 5
- Number of unique words after removing the stopwords = 39,096

(iv) Fitness value (net Similarity)
The cluster location in the multi-dimensional problem space represents one solution for the problem. When a message moves to a new location, a different problem solution is generated. This solution is evaluated by a fitness function that provides a quantitative value of the solution's utility.
The algorithm can be summarized as:
(1) At the initial stage, each message randomly chooses k different document vectors from the document collection as the initial cluster centroid vectors.
(2) For each particle:
(a) Assign each document vector in the document set to the closest centroid vector.
(b) Calculate the fitness value based on equation 5.
(c) Using the velocity and message position to update and to generate the next solutions.
(3) Repeat step (2) until one of the following termination conditions is satisfied.
(a) The maximum number of iterations is exceeded or

(b) The average change in centroid vectors is less than a predefined value.

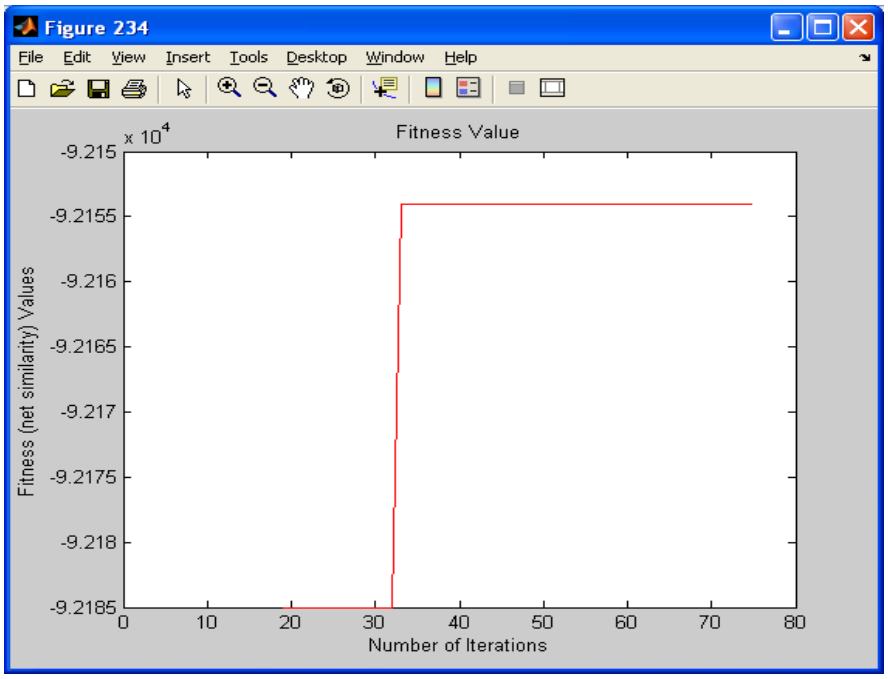

Figure 2: Fitness value(net similarity)

Clustering

Reuters-21578 (Reuters) data set is pre classified manually [36]. This classification information is eliminated before the clustering processes, and is used to evaluate the clustering accuracy of each clustering algorithm at the end of the execution. The original Reuters data consist of 22 files (for a total of 21,578 documents) and contain special tags such as "<TOPICS>" and "<DATE>" among others. The preprocessing phase on the data set cuts the files into single texts and strips the document from the special tags. Then, those documents which belong to at least one topic are selected. At last, after stop words removal, word stemming, and word frequency computation for each document, the data set turns into the form of

$$D = \left\{ \begin{array}{l} \{\langle f_1^1, n_1^1\rangle, \langle f_1^2, n_1^2\rangle, \ldots \langle f_1^{M^1}, n_1^{M^1}\rangle\}, \ldots \\ \{\langle f_N^1, n_N^1\rangle, \langle f_N^2, n_N^2\rangle, \ldots \langle f_N^{M^N}, n_N^{M^N}\rangle\} \end{array} \right\} \quad (6)$$

For text clustering problem, Cofeature Set can be viewed as a two-tuples set. Each term in the set consists of one word that exists both in di and dj, and the word's frequency in d_j. The terms in the Unilateral Feature Set, on the other hand, consist of the words that only exist in di and their frequencies in di. Moreover, there are some words that exist in the title, abstract, or in the first sentence of each paragraph in dj and they can also be found in d_i. These words and their frequencies at important positions of d_j can be viewed as the two-tuples of the Significant Cofeature Set (we used the words (except stop words) in the title). For the construction of seeds, in order to quickly find out the representative features, the proposed Mean Features Selection strategy is applied.

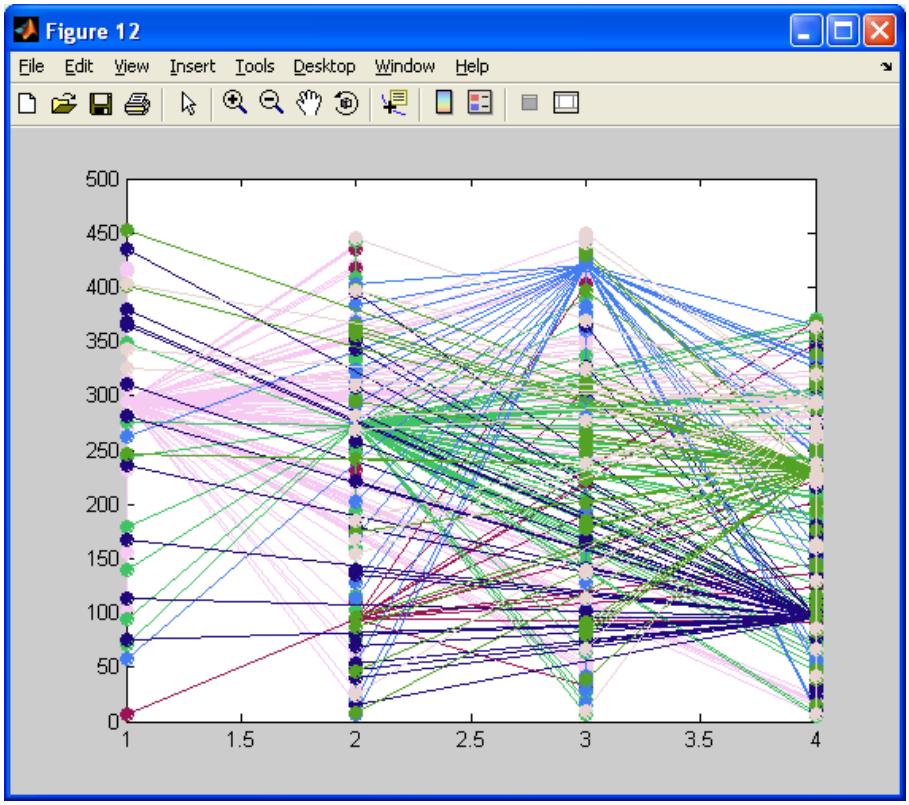

Figure 3: Clustering (Message passing between data points)

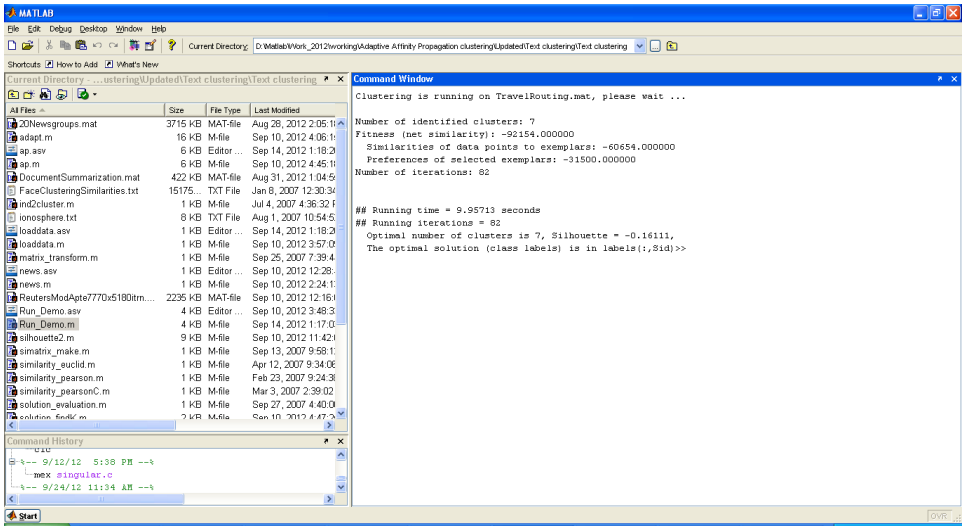

Figure 4: Result for optimal clustering with similarity measure

TABLE 2 Mean Values over All Experiments

	Mean F-Measure
SAP with examplar	**0.625**
SAP	0.599
AP(Tri-set)	0.486
AP(CC)	0.403
K-means	0.416

The table presents the Mean F-Measure scores for various clustering algorithms, highlighting their performance in terms of clustering accuracy. Among the algorithms, **SAP with Exemplar**

achieves the highest score of **0.625**, indicating that this approach, which incorporates exemplar-based clustering, is the most effective in forming coherent and precise clusters. The standard **SAP** method follows closely with a score of **0.599**, demonstrating strong performance but slightly less accuracy compared to the exemplar-enhanced version. The lower scores for **Affinity Propagation (AP) with Tri-set** (**0.486**) and **AP with CC** (**0.403**) suggest that these variants struggle more with forming optimal clusters, potentially due to their handling of cluster boundaries or data characteristics. Meanwhile, the **K-means** algorithm, which is a popular method for partitioning data into clusters, yields an F-Measure of **0.416**, performing better than AP(CC) but still significantly lower than both SAP variants. These results indicate that while K-means and certain AP methods are effective in some scenarios, the **SAP with Exemplar** approach stands out as the best-performing algorithm for clustering, offering the highest accuracy in this evaluation.

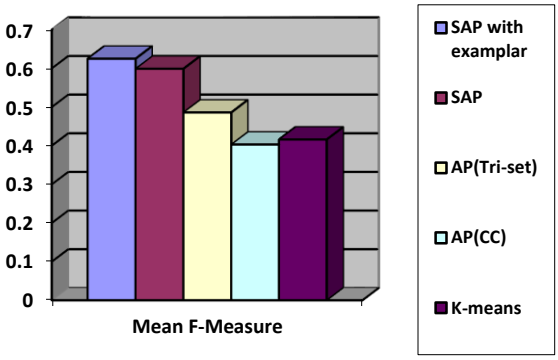

Figure 5: Performance of Frequency measure with existing algorithm

The figure 5 compares the Mean F-Measure values of five different clustering algorithms: **SAP with Exemplar**, **SAP**, **AP (Tri-set)**, **AP (CC)**, and **K-means**. The height of each bar represents the performance of the respective algorithm in terms of clustering accuracy. The highest bar belongs to **SAP with Exemplar**, with a Mean F-Measure of **0.625**, demonstrating its superior performance in generating high-quality clusters. Following closely is the standard **SAP** algorithm, which achieves a slightly lower F-Measure of **0.599**, indicating that while still effective, it does not perform as well as the exemplar-enhanced version. The **AP (Tri-set)** method has a noticeable drop in performance, with a score of **0.486**, suggesting it struggles more with achieving optimal clustering. **AP (CC)** performs even lower, with an F-Measure of **0.403**, indicating significant challenges in producing coherent clusters. **K-means**, a commonly used clustering algorithm, shows a modest performance with a score of **0.416**, placing it between the AP variants in terms of accuracy. Overall, the chart illustrates the clear advantage of **SAP with Exemplar** over the other methods in clustering quality.

5. Conclusion

The first algorithm is well-suited for document sets where the required classes are related to each other, and a strong basis for each cluster is necessary. This makes the algorithm highly effective in applications like search engines within specific domains or fields.
The second proposed algorithm focuses on feature-based clustering, which is better suited for sets containing documents from very different fields, where the co-occurrence of words plays a crucial role in determining the clusters. Applications such as recommending news articles on news portals can benefit greatly from this approach.

In conclusion, although two algorithms have been proposed for clustering, the problem remains open. Considering the growing complexity of data, further research is needed to improve clustering methods. Both algorithms work with real-valued pairwise data point similarities as input, where clusters are represented by a cluster center or an "exemplar." The method is iterative and seeks to find clusters that maximize an objective function known as net similarity.

6. References

1. L. Larsen, E. Ruspini. J. McDew, D. Walter. and W. Adey. "A test of sleep staging system in the unrestrained chimpanzee," *Brain Res.,* vol. 40, pp. 319-343, 1972.
2. Brian S. Everitt, Sabine Landau, and Morven Leese. *Cluster Analysis*. Oxford University Press, fourth edition, 2001.
3. B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, no. 5814, pp. 972-976, Feb. 2007.
4. T.Y. Jiang and A. Tuzhilin, "Dynamic Micro Targeting: Fitness- Based Approach to Predicting Individual Preferences," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 173-182, Oct. 2007.
5. W.H. Wang, H.W. Zhang, F. Wu, and Y.T. Zhuang, "Large Scale of E-Learning Resources Clustering with Parallel Affinity Propagation," Proc. Int'l Conf. Hybrid Learning 2008 (ICHL '08), pp.1-10, Aug. 2008.
6. F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, pp. 1-47, 2002.
7. F. Wang and C.S. Zhang, "Label Propagation through Linear Neighbourhoods," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 1, pp. 55-67, Jan. 2008.
8. C.J. van Rijsbergen, Information Retrieval, second ed., pp. 22-28. Butterworth, 1979.
9. L. Bottou and Y. Bengio, "Convergence Properties of the KMeans," Advances in Neural Information Processing Systems, vol. 7, pp. 585-592, MIT Press, 1995
10. H. Zha, C. Ding, M. Gu, X. He, and H.D. Simon, "Spectral Relaxation for K-Means Clustering," Advances in Neural Information Processing Systems, vol. 14, pp. 1057-1064, MIT Press, 2001.
11. S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categorization," Proc. Seventh Int'l Conf. Information and Knowledge Management, pp. 148-155, 1998
12. Z.H. Zhou and M. Li, "Distributional Features for Text Categorization," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 3, pp. 428-442, Mar. 2009.
13. S. Gao, W. Wu, C.H. Lee, and T.S. Chua, "A Maximal Figure-of- Merit (MFoM)-Learning Approach to Robust Classifier Design for Text Categorization," ACM Trans. Information Systems, vol. 24, no. 2, pp. 190-218, Apr. 2006.
14. J. Wu, F. Ding, and Q.L. Xiang, "An Affinity Propagation Based Method for Vector Quantization," Eprint arXiv 0710.2037, http:// arxiv.org/abs/0710.2037v2, Oct. 2007.
15. M.J. Brusco and H.F. Kohn, "Comment on 'Clustering by Passing Messages between Data Points,'" Science, vol. 319, no. 5864, p. 726c, Feb. 2008.
16. X.D. Wu et al., "Top 10 Algorithms in Data Mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, Jan. 2008.
17. S. Huang, Z. Chen, Y. Yu, and W.Y. Ma, "Multitype Features Coselection for Web Document Clustering," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 448-458, Apr. 2006.
18. L.P. Jing, M.K. Ng, and J.Z. Huang, "An Entropy Weighting KMeans Algorithm for Subspace Clustering of High-Dimensional Sparse Data," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 8, pp. 1026-1041, Aug. 2007.

19. Z.H. Zhou and M. Li, "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 11, pp. 1529-1541, Nov. 2005.
20. Z.H. Zhou, D.C. Zhan, and Q. Yang, "Semi-Supervised Learning with Very Few Labeled Training Examples," Proc. 22nd AAAI Conf. Artificial Intelligence, pp. 675-680, 2007.