ISSN 1870-4069

Towards the Construction of a Clustering Algorithm with Overlap Directed by Query

Beatriz Beltrán, Darnes Vilariño, David Pinto, Rodolfo Martínez

Benemérita Universidad Autónoma de Puebla,
FCC - Doctorado Ingeniería del Lenguaje y del Conocimiento, Mexico

{bbeltran, darnes, dpinto}@cs.buap.mx, beetho@gmail.com

Abstract. Clustering algorithms are one of the most important element for the data mining and for the automatic learning, also, they keep on being a research topic because algorithms have some limitations. In addition, most of algorithms do not bear overlaps of the groups that they generate, however, in some problems like user profile, clustering algorithms are required for bearing overlaps. Therefore, it pretends to develop an overlapping clustering algorithm using different information retrieval techniques in big data quantities and avoiding computational costs.

Keywords. Clustering algorithms, data recovery, classification models, overlapping groups.

1 Statement of the Problem to be Solved

There is lots of information on internet, due to most of enterprises automate all their processes and local storage problems disappear because the information is in the web, then, with the generation of groups appear the possibility and need to identify groups for data analysis, data managing efficiently, inspect trends, give recommendations, filter queries or having organized information.

Most of the developed clustering algorithms do not allow to overlap between their groups because many problems demand that groups are separated. According to the kind of grouping obtained [1]: It could have the classification as ***exclusive***, this mean, one element belongs to one group and the rest of them not, for example, films could be classify by their content (AA, A, B, B15, C y D), in ***fuzzy***, when an element belongs to all the groups, but with a certain grade of belonging, for example, a range of a million colors and finally, one which are ***overlapped*** where an element could belong to more than one group, for example, the food tastes of people.

The development of a clustering algorithm with overlap is proposed for lots of information volumes, but using an *information retrieval system* which enable to obtain groupings to apply them. For example, in several problems, mainly in the language natural processing area, such as, user grouping, recommendation system like products, films, vacation places, etc., user profiling, stablish themes for documents, organ-

ize documents on the network, trendy news and its analysis, trend of political campaigns, etc. So, the paper is described as follows, first it is described a brief state of the art, in section two, it is given the methodology, in section three it is provided the main contributions and in section four there are some results and validation.

There are many proposal of clustering algorithm with overlap, Star algorithm is showed in [2], and different extensions of the Star algorithm in [9, 10], with data mixed [11], document groupings [12], among others. Exist different groups to mention, but the algorithm Star [2] is taken to illustrate this, the algorithm Star proposes a grouping through a graph not directed weight G= {V, E, *w*}, where V is the set of vertexes which represents documents or in general, elements, E is the set of edges with the weight *w*, corresponds with the resemblance between two documents or elements. For the resemblance, the cosine metrics is used, initially, G is a complete graph, and the grouping leaves a graph of partial maximal coverage, with those edges of resemblance σ, it can get lots of overlapping groups, as not overlapped.

A clustering algorithm with incremental overlap is called Incremental Clustering by Strength Decisions is shown in [3], which uses a heuristic of coverage graphs getting large groups, reducing computational costs, but keeping the incremental structure of the grouping. The corpus used, were pre-processed through eliminating of stop words, lemmatization and the representation used was the Vector Space Model. Cosine was used as a measure of similarity.

The clustering algorithm CLOPE [4] is used for categorical data, being fast, scalable and memory optimization. The similarity measure, proposed, take an intuitive idea as a base to increase the weight proportion of the histogram grouping, using principally this measure due to it is easier and more effective than Jaccard or Dice measures. In the experiments made, the algorithm is very effective, gives interesting groups although a dissimilarity measure intra-groups is not given in specific.

In [5] shows a clustering algorithm of documents, called *Generalized Star (GStar)*, which is a generalization of the Star algorithm, this includes a new concept for Star, allowing a new star form with overlapping groups and defining few groups being easy in implementation and efficient, for the tests, Jaccard was used as a similarity measure, the algorithm is useful for groupings that could be required in the organizing information, surfing, tracking of topics and detecting new topics.

Spatial databases also require clustering algorithms, so an algorithm is necessary to do this task, DBSCAN [6], this has minimum requirements of domain knowledge to determine the input parameters, the discovering of arbitrary form groups. DBSCAN is a clustering algorithm based on density, where the results of the experiments show that it is more effective to discover arbitrary form groups than others algorithm.

In [7] shows the incremental algorithm for overlapping groups, called Incremental Clustering by Strength Decision (ICSD), overlapping and dense groups are getting from a heuristic of coverage graph, reducing the computational time, but keeping the grouping incremental structure. The algorithm builds, as it was mention already, from a set of overlapping and dense groups applying the heuristics, this reach the execution time and do more efficient the managing of the multiple insertions in incremental environments.

There is a hierarchical clustering algorithm, call *Hierarchical Compact Algorithm* and its dynamic version make up a framework [8], these work with dynamic and static data sets. It can obtain different hierarchical agglomerative algorithms specifying a similarity measure inter-groups, one sub-graph β-similarity and another coverage algorithm. It could use for tasks which require dynamic clustering, such as organizing information, document taxonomy and hierarchical topic detection.

2 Research Methodology

Taking into consideration that a typical information retrieval system [13], according to the lots of information given, is required having precise and fast access of itself through one question (query). Manually, perform the retrieval of the information brings as mainly consequence that the most important information is ignored, due to it does not have the necessary precision on the process because one person wants to do fast the tasks, then, he usually ignores this information. Also, in many cases physically is impossible to do this task because there are so much information. For this reason, an information retrieval system pretends solve, by mean of the automatic solution of, given a query (user need) and having much information, the system will recover the most relevant elements for the query (Fig. 1).

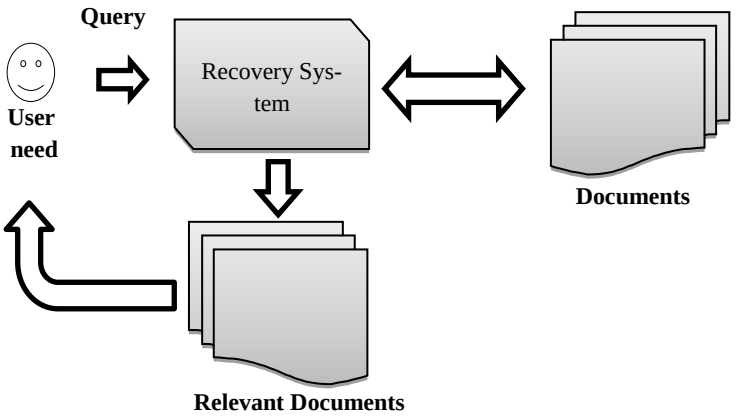

Fig. 1. Typical Information Retrieval System.

In the problem to develop, the query changes to a document to cluster, the documents will be with some presentation and the information retrieval system will propose any similarity measure. Getting the similarity with the documents, the output system will be the similarity with the documents, taking into consideration those one which are over the definite threshold and those one will supply the clustering, this proposed system is showed in the Fig. 2.

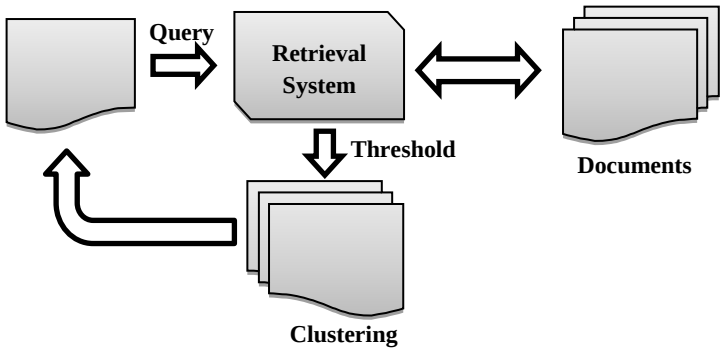

Fig. 2. Proposed System to Generate Clustering.

For all the above reasons, the following methodology is proposed (Fig. 3): first, it is necessary to compile a corpus, for this, it will perform a search of labeled corpus and they should have any overlap in their data. At the same time, a deep analysis of the clustering algorithms proposed is required, from disjointed clustering algorithms like those ones proposed for clustering with overlap, in this analysis, the revision, of the similarity measures used in the different reported cases, will do considering the results obtained.

Design the clustering algorithm with overlap for big information volumes with information retrieval system mechanism. Defining firstly, one or different documents representations, similarity measure to use and define a threshold resemblance.

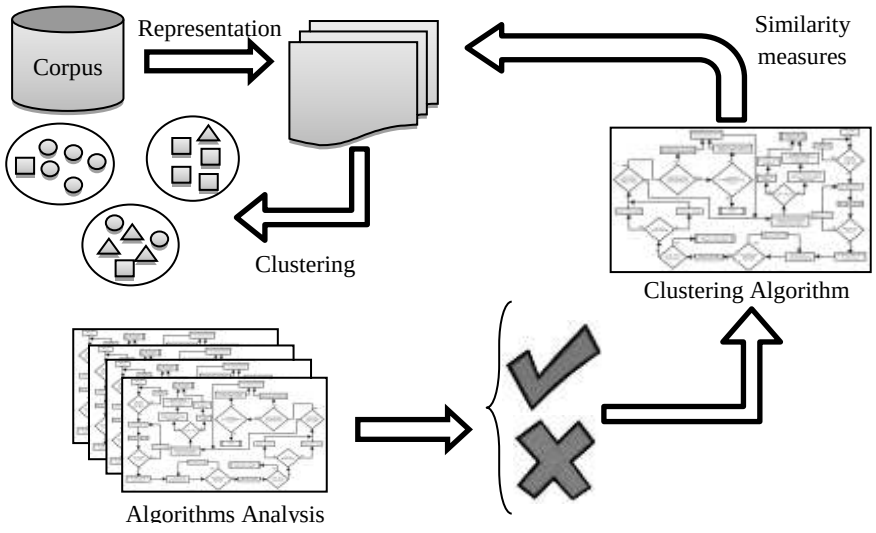

Fig. 3. Methodology proposed.

Implementing the algorithm and do proof to consider with the different representations, similarity measure and thresholds.

Analyze the evaluation ways with the clustering algorithms to check the groups quality and propose one form for evaluating which allow to verify if the algorithm proposed generates competitive groups with the reported algorithms and evaluate the algorithm proposed.

Make a comparison between different algorithm and the proposed to check the quality of the groups generated, considering the execution time facing the big information volumes.

Make an analysis of the results obtained and if it is possible to determine the limitations that could be determine within the algorithm.

In the development of the present research, the main contribution is to use the recovery information system mechanisms as a methodology to model a generalization of the design of one clustering algorithm with overlap.

3 Main Contribution

The contribution is a clustering algorithm with overlapping, using information retrieval system. The first proposal to use an information retrieval system is to use posting list for trigrams and tf-idf. For this algorithm, was applied to corpus of PAN 2017.

The results are significant for cluster algorithms, but this is only a proof and it is necessary more proofs with other corpora, implements some techniques of information retrieval system and it is important to check different threshold.

4 Results and Validity

Some proof was made about the procedure described already, the corpus competence was taken from the PAN[1] 2017 in the author clustering tasks. This consist in 60 problems in 3 languages (English, German and Greek), each problem is made up of 20 texts and have the gold standard. In this paper, the algorithm only was probed with documents in English, but the algorithm could be executed with other languages and only it is presented some results.

The procedure carried out, in the proof made, explain the following algorithm, the input is a query training document and de output is a set of documents with a similarity. The clusters are obtained with those documents with a similarity great or equal to 0.05. this allow the possibility of overlap. The threshold was taken with value great or equal to 0.05, in a way empirical for this analysis. If the threshold is changed, the clusters are modified.

The representation of the training documents and the query was for trigram words, any pre-processing was not made in the documents. The similarity measure was the cosine, the change in this similarity measure will change the groups created.

[1] http://pan.webis.de/

```
SRI Algorithm
```

Input: Query training document
Output: Documents with a threshold greater or equal to 0.05 of similarity.

```
1. PL ← Posting List (training)
2. For each trigram in the query
     Wtq ← (1+log(tfq(trigram)))*log(N/df(trigram))/log
     10
     For each DocId in PL
       Wtd ← (1+log(tf))*log(N/df(trigram))/log 10
       Score[DocId] ← Score[DocId]+(Wtd*Wtq)
     End
3. End
4. For each elem in Score
     Score[elem] ← Score[elem]/TotalDoc
     Write elem, Score[elem]
5. End
```

Different proofs were made, and they show some results. The proof obtained show positive results, the table 1 shows the groups given by the standard gold for the problem 15.

Table 1. Example of group of problem 15.

Groups	Groups of the gold standard
1	document0004 document0011
2	document0001 document0002 document0003 document0006 document0010 document0013 document0016
3	document0005 document0014
4	document0007 document0008 document0009 document0012 document0015

Table 2 shows a query made with the document0001 like a test, the resemblance is presented with a threshold greater than 0.05, the elements of the gold group could be observed that four elements of six are recovered.

Table 2. Similarity between document0001 and the great similarity.

Document	Similarity
document0001	0.695751
document0016	**0.0866128**
document0003	**0.0637452**
document0013	**0.0583805**
document0010	**0.0519872**

Table 3 explains a query made with the document 0007, this show a threshold greater than 0.05, in this case, it could be observed that all documents of the gold standard group, are recovered.

Table 3. Similarity between document0007 and the great similarity.

Document	Similarity
document0007	0.758995
document0015	**0.120644**
document0008	**0.09973**
document0012	**0.0799576**
document0009	**0.0664196**

In the following, the results for the problem 002 are showed, table 4 display the groups supplied by gold standard for this problem.

Table 4. Example of group of problem 002.

Groups	Groups of the gold standard
1	document0005 document0017
2	document0001 document0006 document0009 document0012 document0014 document0015
3	document0003
4	document0004 document0013 document0018
5	document0002 document0010 document0011 document0016 document0019
6	document0008
7	document0007 document0020
8	document0005 document0017

In the table 5 a query was made with the document0001 as a test, the resemblance shows a threshold greater than 0.05, elements of the standard gold could be observed that three of six documents are recovered.

A query was made with document0002 as table 6 shows, the resemblance has a threshold greater than 0.05. elements of the standard gold could be observed that three of five documents are recovered.

Table 5. Similarity between document0001 and the great similarity.

Document	Similarity
document0001	0.73864
document0015	**0.0827088**
document0019	0.0752238
document0014	**0.0694811**
document0012	**0.0589523**
document0016	0.0582658
document0018	0.0574498

Table 6. Similarity between document0002 and the great similarity.

Document	Similarity
document0002	0.824262
document0006	0.0932512
document0010	**0.0862495**
document0019	**0.0827472**
document0016	**0.0709717**
document0008	0.0661357
document0017	0.0653661
document0009	0.0590648
document0020	0.0540831
document0004	0.0519196
document0005	0.051108

These tests confirm that the problem statement made about this proposal could have positive results. This allow in this way, for more corpora in different areas.

In the table 7 is presented the precision and recall for the experiments were showed.

Table 7. Precision and recall for the experiments.

Problem	Query	Precision	Recall	F-measure
15	document0001	1	0.667	0.8
	document0007	1	1	1
2	document0001	0.5	0.6	0.545
	document0002	0.3	0.75	0.429

5 Conclusion and Future Work

In this paper, it was performed some tests with at least one tagged corpus, which allows to obtain results so far satisfactory. It was obtained an average F-measure of 0.69, and this could have some good results, in general.

An information retrieval system was used, using a technique such as *tf-idf*, and making use of the proposed methodology. An algorithm for clustering with overlap-

ping was proposed, with input a query and the output a cluster, using a similarity measure, like cosine.

As future work requires testing with other corpora and testing other thresholds. In addition to making use of other tools of information retrieval systems.

Acknowledgements. We would like to thank VIEP – BUAP and LKE PhD program for supporting this research work.

References

1. Jain, A. and Dubes M.: Algorithms for Clustering Data. Prentice Hall (1988)
2. Aslam, J., Pelekhov, K. and Rus, D.: Static and dynamic information organization with star clusters. In: Proceedings of the seventh international conference on Information and knowledge management ACM, 208–217 (1998)
3. Pérez, J.A., Martínez, T.J., Carrasco, O.J. and Medina, P.J.: A New Incremental Algorithm for Overlapped Clustering. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications CIARP 2009. Lecture Notes in Computer Science, 5856, 497–504 (2009)
4. Yan, Y., Guan X. and You J: CLOPE: A Fast and Effective Clustering Algorithm from Transactional Data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, SIGKDD'02 ACM, 682–687 (2002)
5. Pérez, J. A. and Medina P. E.: A Clustering Algorithm Based on Generalized Stars. LNAI, Springer-Verlag, 4571, 248–262 (2007)
6. Ester, M., Kriegel, H. P., Sander, J. and Xu, XS.: A Density - Based Algorithm for Discovering Cluster. In: Proceedings KDD-96 AAAI, 226–231 (1996)
7. Pérez, A. S., Martínez, J. F., Carrasco, O. J. and Medina, P. J.: A New Incremental Algorithm ios for Overlapped Clustering. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 14th Iberoamerican Conference on Pattern Recognition, 5856, 497–504 (2009)
8. Gil, G.R., Badía, C. J. and Pons, P. A.: Dynamic Hierarchical Compact Clustering Algorithm. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 10th Iberoamerican Conference on Pattern Recognition, 3773, 302–310 (2005)
9. Gil-García, R.J., Badía-Contelles, J.M., Pons-Porrata, A.: Extended star clustering algorithm. In: Proceedings of the 8th Iberoamerican Congress on Pattern Recognition (CIARP2003), LNCS 2905, 480–487 (2003)
10. Pérez, S.A. and Medina, P.J.: A clustering algorithm based on generalized stars. In: Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007), LNAI 4571, 248–262 (2007)
11. Pons, P.A., Ruiz, S.J., Berlanga, L.R. and Santiesteban, A.Y.: Un algoritmo incremental para la obtención de cubrimientos con datos mezclados. Reconocimiento de Patrones Avances y Perspectivas Research on Computing Science, 405–416 (2002)
12. Zamir, O. and Etziony, O.: Web document clustering: A feasibility demonstration. In: Proceedings of the 21st Annual International ACM SIGIR Conference, 46–54 (1998)
13. Manning, C.D., Raghavan, P. and Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press (2009)