International Journal of Engineering & Technology, 7 (4) (2018) 5699-5704

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.22914
Research paper

Onto based cluster labelling and incremental system for information retrieval

Seifedine Kadry [1] *, Irfan Ahamed Mohammed Saleem [2], Lakshmana Kumar Ramasamy [3], N. Kannammal [3]

[1] Department of Mathematics and Computer Science, Faculty of Science, Beirut Arab University, Lebanon
[2] Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
[3] Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[4] Surya Engineering College, Perundurai, Erode, India
**Corresponding author E-mail: s.kadry@bau.edu.lb*

Abstract

Document clustering is utilised for data retrieval, past task of labels to cluster individuals enhances quick retrieval, and Existing framework doles out labels because of standard terms show in archives. However, semantic marks are first taking into Document semantic relationship, the incremental calculation for the versatile framework. The proposed work allocates onto mark because of scientific categorisation, i.e. ontology-based; word net Synsets and gleams coordinating and incremental dynamicity is accomplished through naming. The assessment is done utilising f-measure and figuring speed, contrasted and benchmark K-Means, K-Means without labels. Thus semantic labelling is designed more efficient than traditional document clustering methodologies and can be implemented for real-time internet document clustering applications.

Keywords: *Semantic Document Clustering; K-Means; Labels; Synsets.*

1. Introduction

Document clustering has turned out to be one of the fundamental strategies for sorting out the broad measure of document s into a little number of significant groups, which assumes a vital part in data recovery. Customary clustering calculations are generally depending on the Bag of Words approach, and a conspicuous detriment of the pack of the word is that it overlooks the semantic relationship among words so that can't precisely speak to the significance of archives. Conventional Document clustering calculation utilises highlights like words, expressions, and arrangement to make a cluster. Conventional archive clustering techniques utilise vector space display. In this model, Document is spoken to as a vector utilising term recurrence based weighting plan. In any case, term recurrence based weighting plan can catch the quantity (Li et.al.2009) of events of the terms in an archive; along these lines this model can't impeccably use semantic relationships between's Document substance.

Semantics concentrates on the connection between signifiers like words, expressions, signs, and images. Semantics checks the several routes in which the implications of words can identify with each other to comprehend the connections between them. Semantic Clustering (Rong et.al.2011) is a strategy to create important keywords by focusing significantly on keywords and keyword expresses that are firmly related and cooperative. Semantic clustering worries with apportioning purposes of the information set into unmistakable gatherings (groups) in a way that two focuses from one cluster are semantically like each other yet two focuses from particular clusters are not at all like each other.

As of late, various semantic based clustering methodologies are being created. Be that as it may, there still exist a few difficulties for expanding the clustering quality.

1) The more significant part of existing Document clustering calculations doesn't consider the semantic connections which create poor clustering comes about.
2) Many space particular ontologies are not accessible, so mapping of the idea with that area is impractical.
3) When Word Sense Disambiguation method is utilised, the nature of the clusters is profoundly reliant on the accuracy of that strategy.

Allot recognised and significant depiction for the created clusters. With a specific end goal to advantageously perceive the substance of every cluster, it is fundamentally to appoint small and elucidating marks to help investigators to translate the outcome. By and by, great arrangements of doling out point labels to clusters for simplicity of examination, acknowledgement, and elucidation are still uncommon.

Semantically grouped document s require post-task of descriptive titles to help clients translate the outcomes. Existing systems frequently appoint labels to cluster construct just concerning the terms that the grouped Documents contain, which may not be adequate for a few applications. It is alluring to additionally recommend nonspecific theme terms for simplicity of examination, particularly in the applications where document s cover an extensive variety of area knowledge.

The document can be clustered in a group mode, or they can be clustered incrementally. In group clustering, every one of the document s should be accessible at the time-clustering begins. At that point, the clustering calculation emphasises various circumstances over the dataset and enhances the nature of clusters it shapes. Be that as it may, in some crucial situations archives arrive persistently with no certain limit as to where the accumulation procedure can be ended, and document s can be clustered. Consequently, an incremental clustering arrangement is required in these cases.

Copyright © 2018 Seifedine Kadry et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Literature review

Naming a grouped arrangement of document s is a particular task in content clustering applications. Programmed naming strategies predominantly depend on separating critical terms from clustered document s, where the term essentialness can e figured uniquely in contrast to clustering calculations.

To calculations. This segment clarifies some current strategies for group marking. Morris and Hirst (1991) were the first to recommend the utilisation of lexical chains to investigate the structure of writings; they utilised different kinds of syntactic classes to make lexical chains between words.

Caraballo (1999) developed a thing chain of the importance of hypernyms consequently from content. The thing chain of importance is built utilising base up clustering approach, gathering things in light of conjunction and relation. With a specific end goal to mark each interior group, an arrangement of conceivable hypernyms of each thing in the cluster is removed from the content utilising an etymological example. The thing that has the most significant number of hyponym relations with the thing in the groups is allocated as the cluster name.

Pantel et al. (2004) consequently alloted name to semantic classes, produced from their clustering calculation. For each semantic class, a subset of ideas in the class that is well on the way to speak to the semantic class is chosen as class agents. These delegate ideas are then used to concentrate name hopefuls utilising some lexical examples. The mark applicant with the most astounding common data with the class delegates is relegated as the class name.

Treeratpituk et al., (2006) introduce a straightforward naming calculation that consequently allows small marks to various levelled groups. The calculation consolidates measurable components of the group, the parent cluster, and the corpus into an apparent score. The calculation depends on the theory that by looking at the word dispersion from various parts of the chain of command, it ought to be conceivable to allow proper marks to every cluster in the pecking order. Many existing methodologies produce marks with the assistance of external databases.

Tseng (2010) proposed a WordNet-based measure that first concentrates particular class terms as cluster descriptors, and afterwards, these descriptors are mapped to bland terms because of a hypernym seek calculation to make non-exclusive titles for groups. Nonetheless, this approach is exceptionally tedious, something that prompts to high execution times to get the required group labels.

Chen et al., (2010) presents a successful Fuzzy-based Multi-name Document Clustering (FMDC) approach that coordinates fluffy affiliation manage to mine with a current philosophy WordNet to enhance the nature of archive clustering comes about. In this approach, the key terms will be separated from the archives set, and the underlying representation of all document s is further improved by utilising hypernyms of WordNet to endeavour the semantic relations between terms. At that point, a fluffy affiliation governs digging calculation for writings is utilised to find an arrangement of exceedingly related fluffy regular itemsets, which contain essential terms to be viewed as the labels of the hopeful groups. Wei et al., (2015) utilise disambiguated ideas from lexical chains in the determination of theme parks for the producer groups. The leading ten most noteworthy weighted components are extricated as the group labels. The weighted ideas in the extricated delegate lexical chains are semantically critical terms in clusters.

Charikar et al., (1997) propose a model-called incremental clustering which depends on a careful examination of the prerequisites of the data recovery application, and which ought to likewise be valuable in different applications. The creator characterises the incremental clustering issue as, for an upgrade grouping of n focuses in M, keep up a gathering of k clusters to such an extent that as every info point is displayed, it is possible that it is allocated to one of the present k groups or it begins off another group while two existing groups are converged into one. George et al., (2005) outline incremental and parallel forms of the co-clustering calculation and utilise it to construct a proficient continuous cooperative separating framework. An element synergistic separating methodology was exhibited that can bolster the section of new clients, things and appraisals utilising a half and half of incremental and clump variants of the co-clustering calculation.

Devender et al., (2015) use the online reference book Conservapedia, to recover the equivalent words of the inquiry term so that from the recovered Documents of the dataset the connected semantic terms of the predefined question term are recognised lastly more comparative document s are ranked in light of semantic relationship comparability. FICA (Fast Incremental Clustering Algorithm) calculation is adjusted for clustering the Documents for element archive corpora, in light of semantic comparability. For each cluster, the top connected ideas from every Document are removed and are kept up as an idea pool. Rather than figuring the difference between document groups and the new Document, the semantic similitude between the new archive and the idea pool is registered, which decreases the calculation overhead.

3. Clustering of k-means and cover co-efficient algorithm

K-means is one of the easiest unsupervised learning calculations that take care of the outstanding clustering issue (Ponnusamy et.al.2018). The system takes after a basic and straightforward approach to group given information set through a specific number of clusters (expect k groups) settled apriori. The primary thought is to characterise k focuses, one for every cluster. These focuses ought to be set shrewdly on account of the various area causes diverse

Outcome. Along these lines, the better decision is to place them however much as could reasonably be expected far from each other. The next stride is to take every direct having a place toward a given information set and partner it to the closest focus. At the point when no point is pending, the initial step is finished, and an early gathering age is finished. Now we have to re-ascertain k new centroids as barycenter of the groups coming about because of the past stride. After we have these knew centroids, another coupling must be done between similar information set focuses and the closest new focus. A circle has been created. Therefore of this circle, we may see that the k focuses change their area well ordered until no more changes are done or at the end of the day focuses don't move any more. At long last, this calculation goes for limiting a target work knows as squared blunder work. The following steps can follow this.

1) Randomly select "c" cluster focuses.
2) Calculate the separation between every information point, and cluster focuses.
3) Assign the information to indicate the cluster focus whose separation from the group focus is least of all the cluster focuses.
4) Recalculate the new group focus utilising: Where "ci" speaks to the quantity of information focuses in ith group.
5) Recalculate the separation between every information point and the newly acquired group focuses.
6) If no information point was reassigned then stop, generally rehash from step 3.

Coefficient calculation has demonstrated increment in processing speed when contrasted with K-Means.

4. Proposed methodology

Naming a clustered set of Documents is an inescapable task in content clustering applications. Programmed marking strategies mostly depend on extricating noteworthy terms from grouped document s, where the term centrality can be figured uniquely in contrast to clustering calculations to calculations.

Clustering archive accumulations can make it simpler to discover important Documents as clustering unites comparative document sand can make discovering data less demanding and quicker. Customarily, datasets have been static (they don't change), so clustering calculations were created take the preferred standpoint of this. These calculations are known as static calculations, and they cluster the dataset once. Ought to the dataset change (new archives included, Documents erased or altered) then it was important to play out an entire re-clustering.

Albeit incremental calculations are the best technique for dynamic clustering information, they experience the ill effects of issues. Two of these issues are the viability of the general calculation/approach (particularly after some time as the calculation runs the accumulation through various cycles) and the additional request of the new document s into the current collection. Effectiveness decides how right/ specific the outcomes will be. For incremental calculations, this is critical as it influences the outcomes after some time. Since an incremental calculation will be executed an incremental calculation commonly needs not only high introductory adequacy but rather one that can be kept up all through every emphasis, keeping the outcomes precise and significant. The second issue is the additions arrange issue. To some degree, incremental calculations are

Influenced by the request that new document s touch base into be added to the grouped outcomes. In a perfect world, incremental calculations ought to give similar outcomes for a dataset/gathering paying little heed to the request that Documents touch base (altogether free). The objective is to diminish the impact that the inclusion arranges issue has on the outcomes or expel it.

5. Semantic based method

Archive clustering is the task of consequently sorting out content document s into important clusters or gathering, at the end of the day, the Documents in one group have a similar theme, and the Documents in various groups speak to various points. Document clustering has been contemplated seriously as a result of its wide appropriateness in regions, for example, Web mining, Search Engines, Information Retrieval, and Topological Analysis. Unlike in Document grouping, in archive clustering no marked document s are given. Marking groups is a typical issue in content mining and data recovery. For the most part, the strategies discover a rundown of discriminative words that are utilized to encourage the data recovery or the translation of the gatherings. The outcomes could be utilized as the initial step to help in the development of point scientific classification, since the archives are from a particular space and an area pro is included in the task. The point scientific categorization is useful in calculation.

Semantic Labeling: Input:

```
C={C1,C2,C3,… .,Ck} = Document Cluster
Cd = Collection of records with in a bunch
D' n xm: doc x term lattice
```

Yield:

Cluster naming

Arranging archives, for instance, supporting an advanced library or an entry developing. Be that as it may, it exceptionally tedious process. It clarifies the semantic marking and incremental strategies. Marking is done to groups headed by seed document s, separate term for seed doc and cluster individuals is sifted, the term whose idf is more chosen, the separated term is mapped to wordnet synset and gleams, the comparability score is computed, higher score contributes.

```
procedure Labelling_of_the_dictionary
{
 foreach (common Noun of the dictionary)
 {
 (Label, Reliability) = Find_its_label (Noun)
}
 }
procedure Find_its_label (Noun)
 {
 foreach (Sense with Noun Genus/Relator)
{
 if (Genus/Relator labelled)
{
Sense.Label = Genus/Relator.Label
Sense.Reliability = Genus/Relator.Reliability
}
else
{
( Sense.Label, Sense.Reliability) = Find_its_label(Genus)
 }
 #recursion if (Noun.Label != Sense.Label)
{ Noun.Label = [?] } else { Noun.Label = Sense.Label
}
}
# end foreach
Noun.Reliability = ∑ Reliability labelled senses / number of senses
return (Noun.Label, Noun.Reliability)
}
```

```
For every Ci of C ({C1,C2,C3,… .,Ck})
For every report dj of Ci
Discover seed force of dj
Get the most noteworthy seed control archive dj
Take the five terms whose idf is most extreme
For every term ta of dj
Outline wordnet synset
Register likeness get high score term ta
end
end
Give back the term as marking of Cluster Ci
```

6. Incremental clustering method

Incremental clustering is an approach for making groups on-line. Incremental calculation successively forms document s utilizing a pre-determined request. The present archive is contrasted with every current cluster, and it is converged with the most comparative group if the closeness surpasses a specific limit, else it begins its own particular group. The incremental calculation brings about speedier handling than the agglomerative progressive clustering calculation.

Coefficient-based Incremental Clustering Methodology (C2ICM) calculation. In whatever remains of the paper the images m' and m'', separately, demonstrate the quantity of included and erased document s; comparatively Dm', and Dm'', individually, show the arrangement of included and erased Documents. Dm shows the present document database.

C_2 ICM

Register nc and the cluster seed forces of the Documents in the upgraded document database, Dm = Dm U Dm', – Dm'', and pick the group seeds. (When all is said in done m' > m'') Decide Dr, the arrangement of document s to be clustered. Cluster this document s by doling out them to the group of the seed that spreads them most.

On the off chance that there were document s not secured by any seed, then gathering those together into a ragbag cluster. Apply the above strides for every database overhaul.

The set Dr, comprises of the newcomers, the individuals from the ragbag cluster of the past stride, and the individuals from the adulterated old groups. An old group is characterized to be false if(1) its seed is not a seed any longer (erased seeds would have their

clusters distorted likewise); or (2) one or more of its non-seed document s turns into a seed after a database upgrade.

7. Results and discussion

A clustering calculation is assessed utilizing (i) some interior assessment measure like attachment, division, or the outline coefficient (tending to both, union and partition), (ii) some outer assessment measure like exactness, accuracy, or review regarding some given class-structure of the information. Now and again, where assessment in view of class labels does not appear to be reasonable, (iii) cautious (manual) examination of clusters shows them to be a some way or another significant gathering of clearly some way or another related article.
The proposed framework is assessed by standard measurements of pertinence, for example, F-measure and virtue in the field of data recovery for quality evaluation.

The f-measure is computed as,

$$F = 2 \times \frac{P.R}{P+R}$$

$$P=Precision=\frac{TP}{TP+FP}$$

And

$$R=Recall=\frac{TP}{TP+FN}$$

Where
TP = True Positive
TN= True Negative
FP= False Positive
FN= False Negative

To process immaculateness, every cluster is allocated to the class which most successive in the group and after that exactness of this task is measured by checking the quantity of effectively doled out document s and separating by N.

$$purity(\Omega, C)=\frac{1}{N}\sum_{k}\max_{j}\left|w_k \cap c_j\right|$$

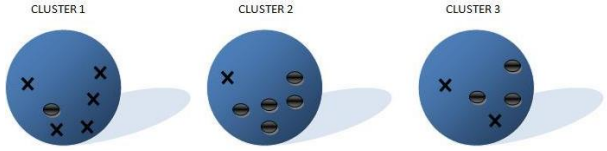


Fig. 1: Clustering Purity.

The immaculateness of this group is figured as take lion's share number of individuals in every cluster. In cluster 1, the greater part is x =5, group 2 o=4 and cluster 3 =3. The immaculateness of this group is
Terrible clustering have immaculateness values near 0, an impeccable clustering has a virtue of 1. Marking is done to groups headed by seed Documents, discriminant term for seed doc and cluster individuals is filtered, the term whose idf is more chosen, the sifted term is mapped to WordNet synset and gleams, the likeness score is figured, higher score contributes . at the point when new archive is included, its term vector is separated, measurement decrease system is connected, contrasted with marks , with firmly comparative group doc is included, term imperative is discovered, mapped to WordNet, naming as past stride, Dataset.

Table 1: Performance Metric for BBC Dataset

Algorithm	F-Measure	Cluster Purity	Computing Time(sec)
K-Means	0.72	0.33	2.217
K-Means+onto	0.79	0.44	1.484
K-Means+ onto+clusterlabel	0.81	0.64	1.215
CC+onto	0.83	0.79	1387.913
CC+onto+clusterlabel	0.85	0.81	1176.876
Algorithm	F-Measure	Cluster Purity	Computing Time(sec)

Table 2: Performance Metric for R8 Dataset

K-Means	0.60	0.37	0.793
K-Means+onto	0.75	0.45	0.287
K-Means+ onto+clusterlabel(proposed)	0.78	0.67	0.238
CC+onto	0.84	0.78	472.67
CC+onto+clusterlabel	0.87	0.84	398.12

Table 3: Performance Metric for News Group 20

Algorithm	F-Measure	Cluster Purity	Computing Time(sec)
K-Means	0.76	0.40	1.249
K-Means+onto	0.83	0.53	0.864
K-Means+ onto+clusterlabel	0.84	0.69	1.092
CC+onto	0.86	0.73	1289.344
CC+onto+clusterlabel	0.88	0.81	1083.258

Table 1 to 3 displays the result of calculating F-measure , Cluster purity , and the processing time take with three different datasets respectively Dataset -1 : BBC , Dataset -2 : R8 , Dataset -3 : News Group 20. To test the consistency of the propose methodology it is compared with five different Algorithms respectively K-Means, K-Means+onto, K-Means+ onto+clusterlabel, CC+onto, CC+onto+clusterlabel

Table 4: Accuracy Value- BBC Dataset

Algorithm	Database size	Accuracy	Searching time (ms)
K-means + Incremental	25	0.84	181
	50	0.83	615
	75	0.82	1164
	100	0.81	1376
CC + Incremental	25	0.95	67025
	50	0.97	106003
	75	0.95	128672
	100	0.96	1178627

Table 5: Accuracy Value- R8

Algorithm	Database size	Accuracy	Searching time (ms)
K-means + Incremental	25	0.80	55
	50	0.84	129
	75	0.82	305
	100	0.83	675
CC + Incremental	25	0.85	21463
	50	0.91	26644

Table 6: Accuracy Value - News

	75	0.91	106665
	100	0.97	1278921
	Data set		

Algorithm	Database size	Accuracy	Searchingtime (ms)
K-means+ Incremental	25	0.84	189
	50	0.82	307
	75	0.81	718
	100	0.81	975
CC + Incremental	25	0.88	25175
	50	0.92	261045
	75	0.94	359011
	100	0.98	1934813
	Group 20		Searching

Table 4 to 6 displays the result of calculating Database size Accuracy time (ms) with three different datasets respectively Dataset - 1: BBC, Dataset -2: R8, Dataset - 3: News Group 20. To test the consistency of the propose methodology it is compared with two different Algorithms respectively K-means + Incremental, CC + Incremental.

Table 7: K-Means + Onto + Cluster Label Performance

Data set	F-Measure	Purity
BBC	0.81	0.64
R8	0.78	0.67
NewsGroups	0.84	0.69

Table 8: CC + Onto + Cluster Label Performance

Data set	F-Measure	Purity
BBC	0.85	0.81
R8	0.87	0.78
NewsGroups	0.88	0.81

Table 7 and 8 displays the result of calculating F-Measure and Purity with three different datasets respectively Dataset -1 : BBC , Dataset -2 : R8 , Dataset -3 : News Group 20.

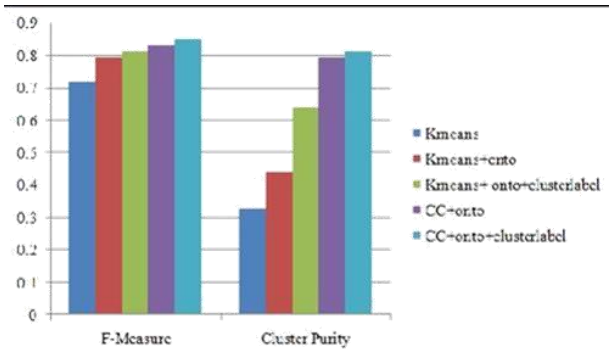


Fig. 2: F-Measure and Purity Comparison-BBC dataset.

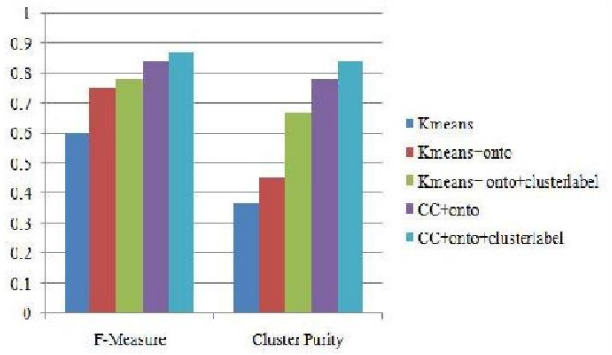


Fig. 3: F-Measure and Purity Comparison-R8 dataset.

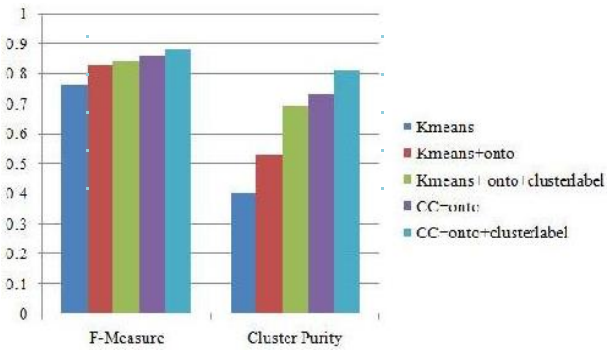


Fig. 4: F-Measure and Purity Comparison-News Groups dataset.

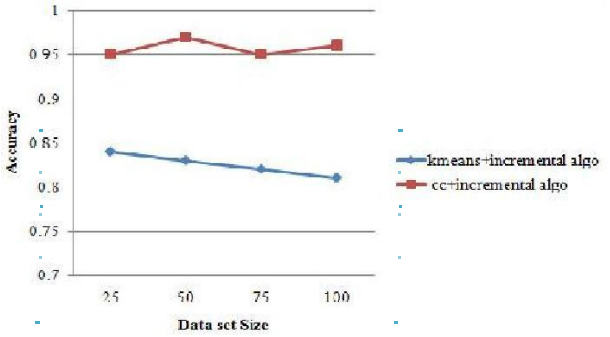


Fig. 5: Accuracy Comparison - BBC Dataset.

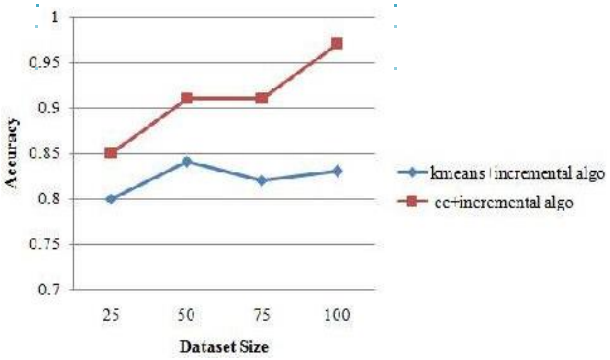


Fig. 6: Accuracy Comparison - R8 Dataset.

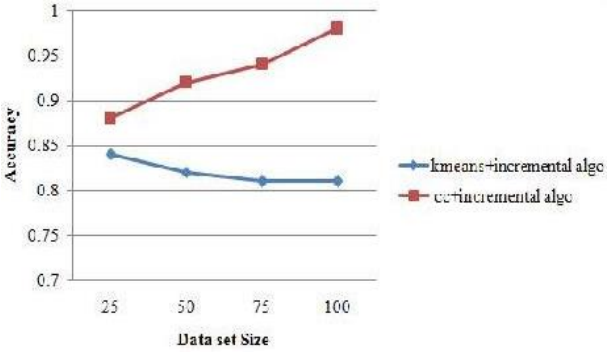


Fig. 7: Accuracy Comparison - News Group 20 Dataset.

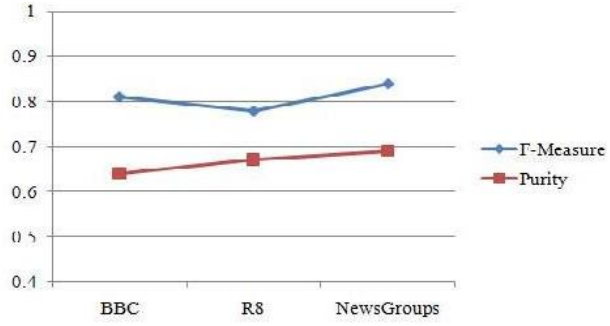


Fig. 8: F-Measure and Purity for K-Means + Onto + Cluster Label

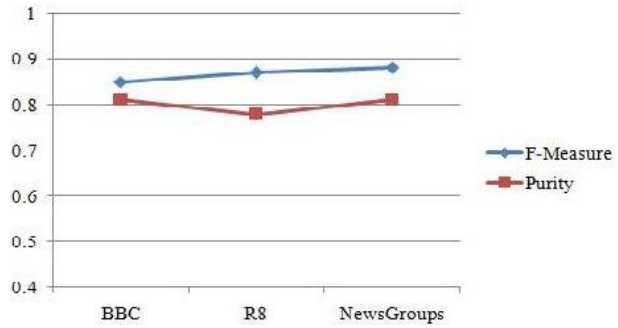


Fig. 9: F-Measure and Purity for CC + Onto + Cluster Label.

8. Conclusion

Increase of web administration increments and speed in looking of result for question is especially fundamental in web usage, the dynamicity in web use and also shrinking and extending store needs exceptional consideration in proposing incremental group algorithm. Existing a few endeavors has understood and accomplished speed recovery and proficient adaptability on utilizing a few techniques. The speed is accomplished by diminishing pursuit space and versatility by incremental calculation. Be that as it may, just few has been founded on semantic knowledge. Joining semantics at all phase of clustering and seeking enhances the aggregate framework. The proposed work by discovering cluster marks on extricating semantic knowledge from head document s and individual group part. Updating in group labels when new Document is included or erased proves significant improvement in clustering. The clustering metric f-measure and registering rate is assessed against benchmark K-Means, K-Means with semantic based dimensionality diminishment. Labels when new document is added or deleted. The clustering metric f-measure and computing speed is evaluated against baseline K-Means, K-Means with semantic based dimensionality reduction. Hence it can be concluded that the proposed methodology is more efficient than the previous traditional clustering. In future the research can be widened on different degrees of semantic onto labeling.

References

[1] Ponnusamy, R., Degife, W. A., & Alemu, T. (2018). Recommender Frameworks Outline System Design and Strategies: A Review. In Knowledge Computing and its Applications (pp. 261-285). Springer, Singapore. https://doi.org/10.1007/978-981-10-8258-0_12.

[2] Li, Y., Hsu, D. F., & Chung, S. M. (2009, November). Combining multiple feature selection methods for text categorization by using rank-score characteristics. In Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on (pp. 508-517). IEEE.

[3] Rong, C. (2011, November). Using Mahout for clustering Wikipedia's latest articles: A comparison between k-means and fuzzy c-means in the cloud. In Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on (pp. 565-569). IEEE.

[4] Ackerman, M & Dasgupta, S 2014, 'Incremental clustering: The case for extra clusters', In NIPS, pp. 307-315.

[5] Altingovde, S, Subakan, NÖ & Ulusoy, Ö 2013 'Cluster searching strategies for collaborative recommendation systems', Information Processing and Management, vol. 49, no. 3, pp. 688-697. https://doi.org/10.1016/j.ipm.2012.07.008.

[6] Bilge & Polat, H 2013 'A scalable privacy-preserving recommendation scheme via bisecting k-means clustering', Information Processing and Management, vol. 49, no. 4, pp. 912-927.

[7] Caraballo, S 1999, 'Automatic Acquisition of a hypernym-labeled noun hierarchy from text', In Proceedings of the Association for Computational Linguistics Conference.

[8] Charikar, M, Chekuri, C, Feder, T & Motwani, R 1997, 'Incremental clustering and dynamic information retrieval', 29th Symposium on Theory of Computing, pp. 626-635.

[9] Chen, X, Liu, X, Huang, Z & Sun, H 2010, 'Regionknn: A scalable hybrid collaborative filtering algorithm for personalized Web service recommendation', Proc. Eighth Int'l Conf. Web Services ICWS '10, pp. 9-16. https://doi.org/10.1109/ICWS.2010.27.

[10] Cutting, DR, Karger, DR, Pedersen, JO & Tukey, JW 1992, 'Scatter/gather: a cluster-based approach to browsing large document collections', In SIGIR '92, New York, NY, USA, ACM, pp. 318-329. https://doi.org/10.1145/133160.133214.

[11] Devender, A, Srinivas, B & Ashok A 2015, 'Efficient Incremental Clustering of Documents based on Correlation', International Journal of Engineering and Computer Science ISSN: 2319-7242, vol. 4, no. 8, pp. 13704-13709.

[12] Ester, M, Kriegel, HP, Sander, J, Wimmer, M & Xu, X 1998, 'Incremental clustering for mining in a Data Warehousing environment', Proc. of the 24th Int. Conf. on Very Large Databases VLDB'98, New York, USA, pp. 323-333.

[13] Fisher, D 1987, 'Knowledge acquisition via incremental conceptual clustering', Machine Learning, vol. 2, pp. 139-172. https://doi.org/10.1007/BF00114265.

[14] George, T & Meregu, S 2005, 'A scalable collaborative filtering Framework based on co-clustering', in: IEEE International Conference on Data Mining ICDM, 2005, pp. 625-628. https://doi.org/10.1109/ICDM.2005.14.

[15] Geraci, F, Maggini, M, Pellegrini, M & Sebastiani, F 2007, 'Cluster generation and cluster labelling for web snippets: a fast and accurate hierarchical solution', Internet Mathematics.

[16] Gennary P Langley & Fisher, D 1989, 'Model of Incremental Concept Formation', Artificial Intelligence Journal, vol. 40, pp. 11-61.

[17] Glover, E, Pennock, DM, Lawrence, S & Krovetz, R 2002, 'Inferring hierarchical descriptions. In CIKM '02, New York, NY, USA, ACM, pp. 507-514. https://doi.org/10.1145/584792.584876.

[18] Li, Y., Hsu, D. F., & Chung, S. M. (2009, November). Combining multiple feature selection methods for text categorization by using rank-score characteristics. In Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on (pp. 508-517). IEEE.

[19] Lloyd, SP 1982, 'Least squares quantization in PCM', IEEE Transactions on Information Theory, vol. 282, pp. 129-137. https://doi.org/10.1109/TIT.1982.1056489.

[20] Manning, CD, Raghavan, P & Schutze, H 2008, 'Introduction to Information Retrieval', Cambridge University Press. https://doi.org/10.1017/CBO9780511809071.

[21] Morris, J & Hirst, G 1991, 'Lexical cohesion computed by the saural relations as an indicator of the structure of text', Computational Linguistics, vol. 171, pp. 21-48.

[22] Pantel, P & Ravichandran, D 2004, 'Automatically labelling semantic classes', In Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference.

[23] Radev, DR, Jing, H, Stys, M & Tam, D 2004, 'Centroid-based summarization of multiple documents', Information Processing Management, vol. 406, pp. 919-938. https://doi.org/10.1016/j.ipm.2003.10.006.

[24] Rong, C. (2011, November). Using Mahout for clustering Wikipedia's latest articles: A comparison between k-means and fuzzy c-means in the cloud. In Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on (pp. 565-569). IEEE.

[25] Toda, H & Kataoka, R 2005, 'A clustering method for news articles retrieval system'. In WWW '05, New York, NY, USA, ACM, pp. 988-989. https://doi.org/10.1145/1062745.1062832.

[26] Treeratpituk, P & Callan, J 2006, 'Automatically labelling hierarchical clusters', In DG. O '06, New York, NY, USA, ACM, pp. 167-176. https://doi.org/10.1145/1146598.1146650.

[27] Tsai, CF & Hung, C 2012 'Cluster ensembles in collaborative filtering recommendation', Applied Soft Computing Journal, vol. 12, no. 4, pp. 1417-1425. https://doi.org/10.1016/j.asoc.2011.11.016.

[28] Tseng, YH 2010, 'Generic title labelling for clustered documents', Expert Systems with Applications, vol. 373, pp. 2247-2254. https://doi.org/10.1016/j.eswa.2009.07.048.

[29] Wei, T, Lu, Y, Chang, H, Zhou, Q & Bao, X 2015, 'A semantic approach for text clustering using WordNet and lexical chains', Expert Systems with Applications, vol. 424, pp. 2264-2275. https://doi.org/10.1016/j.eswa.2014.10.023.

[30] Wu, J, Chen, L, Feng, Y, Zheng, Z, Zhou, MC & Wu, Z 2013, 'Predicting quality of service for selection by neighborhood-based collaborative filtering', IEEE Trans, Systems, Man, and Cybernetics: Systems, vol. 43, pp. 428-439. https://doi.org/10.1109/TSMCA.2012.2210409.