ISSN No: 2454-9614

An Efficient Document Clustering Based on HUBNESS Proportional K-Means Algorithm

R. Saranya, V. Sharmila*, P. Balamurugan

Department Of Computer Science and Engineering, K.S.R. College of Engineering,
Tiruchencode, Namakkal, Tamilnadu, India.

*Corresponding Author: R. Saranya

E-mail: saranyarathakrishnan@gmail.com,

Received: 10/11/2015, Revised: 03/12/2015 and Accepted: 03/03/2016

Abstract

Evaluating similarity between the documents is a main operation in the text processing field. Similarity measurement is used to estimate the relationship between the records or documents.In existing system similarity between two documents can be computed with respect to feature by using Similarity Measure for Text Processing (SMTP). In proposed hybrid SMTP scheme is integrated with hubness based distance analysis scheme, where the KNN search is performed to find out the nearest neighbours based on the similarity. Hubness measure estimation calculates the hubness score. Then Document Clustering performed using Hubness - Proportional K-means (HPKM) algorithm. Semantic analysis method and optimal cluster count estimation mechanism are used to improve the clustering process. Cluster accuracy will be improved by system.

Keywords: Clustering, Document Similarity, K-means algorithm.
Reviewed by* **ICETSET'16 *organizing committee*

1. Introduction

The goal of document clustering is to automatically grouping the similar documents together. It is one of the most key tasks in machine learning and artificial intelligence and has expected much attention in recent eons. There are many number of distance analysis schemes are proposed to deal with the clustering process. The most popular distance analysis schemes are Euclidean distance measure. The K-means algorithm also uses the Euclidean distance scheme, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centres. Generally document dimensionality is considered, that has to be the low-dimensional one. If the document space is high dimensional, then it could be altered to the low dimensional one. So the computational complexity can be reduced. The process of spectral clustering is to make the low-dimensional projection to the documents and the performing the clustering process. It makes the low cost for the computation. Example: Latent semantic indexing (LSI) method.

Because of the high dimensionality of the document space, a certain representation of documents usually resides on a nonlinear manifold embedded in the similarities between the data points **[6].** Unfortunately, there is an algorithm which measures the dissimilarity of the documents rather than similarities between the documents. Thus, it is not able to successfully imprisonment the nonlinear manifold assembly embedded in the similarities between them. If any clustering methodology able to find out the low dimensionality of the document space means that is called as effective clustering method. And that can best preserve the similarities between the data points. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory **[8].** The LPI method spread over a weighted function to each pair wise distance attempting to focus on taking the similarity structure, rather than the dissimilarity structure, of the documents. It does not overwhelm the critical limitation of Euclidean distance. Moreover, the selection of the weighted tasks is often a hard task.

2. Related Work

Short texts vary from outmoded documents in their shortness and sparsity, which makes arithmetical approaches to short texts less in effect. The external knowledge is used by semantics of short text, for example Wikipedia is important one. That is combination of the encyclopedia based wiki. They offer creation, editing of the documents by anyone in the world. It covers wide range of articles; they are named entities, domain specific entities and new entities, in addition to general entities **[1].** Wikipedia also has many structures that are useful for knowledge abstractions, such as condensed link structure among articles, experienced anchor texts and entity disambiguation with URLs.

Wikipedia also has several features that are valuable for knowledge abstractions, such as dense link structure among articles, developed presenter texts and individual disambiguation with URLs. It is prominent that the dump data of Wikipedia can be spontaneously retrieved online. For these details, inquiry on Wikipedia mining has been enhanced.

Most of the effort on Wikipedia-centered short text inquiry concentrated on detailed responsibilities. Ferragina and Scaiella**[6]** proposed a modest and debauched process for entity disambiguation for short texts by Wikipedia.

Meijet al. **[6]** also tackled entity disambiguation by means of several features resulting from Wikipedia for machine learning. Phanet al. [**6]** used concealed subjects attained from Wikipedia for erudition the LDA classifier of short texts. Hu et al. **[6]** subjugated features from Wikipedia for clustering of short texts. Their effort established that Wikipedia was operative as an outside awareness foundation.

Research on expressive semantics of short texts was planned for numerous purposes. Specifically, Explicit Semantic Analysis (ESA) has been broadly charityas of its accessibility and adaptability. ESA was developed for figuring word resemblance as well as text resemblance printed in natural languages. ESA forms a prejudiced up turned guide that records each expression into a gradient of Wikipedia trainings in which it seems, and calculates the resemblance between courses produced from two arguments or texts.

Song et al. **[6]** demonstrated the accessibility of ESA for short text clustering, i.e., calculating semantic aloofness among short texts by means of ESA. Banerjee et al. also hired a comparable style to ESA for the determination of clustering short texts. Sun et al. **[2]** developed ESA to shape short texts with a support vector machine (SVM), which is an administered machine learning practice. Thus, ESA has been established to be operative for determining semantic similarity for short texts. ESA has a problem in its increment organization when it comes to investigating real-world noisy short texts. Wiki Narrate by Strube and Ponzetto useful several simple procedures that have been settled for Word Net to Wikipedia.

Given two Wikipedia courses, they definitely calculate the detachment in the category assembly or the connection point between texts. They established the usefulness of Wikipedia-based approaches on standard datasets for resemblance dimensions and co orientation determination tasks. Milne et al**. [9]** proposed WLM that professionally calculates the resemblance between two courses using the overlay degree of their received and departing links. Graph-based methods **[6]** construct a graph in which nodes are Wikipedia articles and edges are links between articles. Using the graph, they create a vector of objects **[12]** or straight invention correlated entities.

Ito et al. **[1]** projected connection co-occurrence examination to quickly build a suggestion vocabulary. Hassan and Mihalcea exploited cross-language associations of Wikipedia to calculate the resemblance across languages. More recently, hybrid methods have shown to be more perfect. Yazdani and Popescu-Belis utilized both text substances and links in trainings, and Taiebet al. leveraged text contents, categories, Wikipedia category graph, and redirection to achieve competitive or sometimes better results.

3. Clustering on Text Document

Clustering is the process of grouping related documents together, where the similar documents belongs to one cluster, dissimilar objects are belongs to other cluster. In clustering, the characteristics of the similarity of data are doesn’t know previously. By using the statistical concepts, we split the datasets such that the sub datasets have similar data. Since training sets are not used, we describe this as unsupervised learning.

For example, in hospitals grouping patient records with similar symptoms without knowing what the symptom actually indicates. Unsupervised learning is the process of trying to find the hidden structure in the unlabelled data. The process performs without knowing the training data and the response variable.

A document is signified as a vector. In document vector each module designates the value of the equivalent feature in the document. The feature value can be term frequency, relative term frequency. High dimensionality and sparsity can be a unadorned contest for similarity degree. Similarity Measurement for Text Process (SMTP) is used to calculate the resemblance between two documents with esteem to a feature **[2].**Presents and absence of the features in both documents are charity to evaluation the similarity values. The SMTP is stretched to guesstimate resemblance among two set of documents. The SMTP scheme is used with text clustering and classification task. The following concerns are recognized from the current text document clustering scheme.

- Term associations are not measured
- Dimensionality decrement is not accomplished
- Document connection information is not measured
- Partial clustering correctness

4. Ontology Theory

Ontology's belong to the information depiction methodologies that goal to afford a collective accepting of a dominion both for the computers and for the humans. Thus, ontology designates a province of awareness in such a prescribed way that computers can process it. The result is that the computer system knows about this dominion. Ontology is a recognised organization schema, which has a hierarchical command and which is related to some domain. Ontology contains the reasonable essential of a "Knowledge Base". Typically, a knowledge base consists of ontology, some data and also an implication mechanism. Ontology, comprising the logical component of the knowledge base, defines rules that formally describe the field of interest looks like. The data can be any data connected to this field of attention that is extracted from various incomes such as databases, document collections, the Web etc. The extrapolation mechanism would deploy rules in form of axioms, restrictions, logical instants and other numerous methods based on the formal description in the ontology over the definite data to harvest more information.

5. Hub-Based Clustering

Hubness is an aspect of the obscenity of dimensionality relating to nearest neighbours which has only newly come to gentility, unlike the much deliberated distance attention phenomenon. Let $D \subset IR^d$ be a set of data points and let $N_k(x)$ denote the number of k-occurrences of point $x \in D$, i.e., the amount of times x occurs in k-nearest neighbour lists of other opinions from D. As the dimensionality of data rises, the distribution of k-occurrences develops substantially twisted **[3].** As significance, some data points, which will refer to as hubs, are incorporated in many more k-nearest-neighbour lists than other opinions. The number of k-occurrences of point $x \in D$ as its hubness score. It has been shown that hubness, as a sensation, performs in high-dimensional data as characteristic assets of high dimensionality and is neither an artefact of predictable examples nor a individuality of some explicit data sets. Noticeably, the particular unit of hubness might silent vary and is not individually unwavering by dimensionality**.[3]**

Hubs also exist in clustered data, nursing to be located in the closeness of cluster centers. In tally, the grade of hubness does not be contingent on the implanting dimensionality, but rather on the essential data dimensionality, which is observed as the insignificant number of variables wanted to account for all pair wise reserves in the data. Hubs are points x having Nk(x) more than two ordinary divergences higher than the estimated value k. In most trials that follow, will only concern ourselves with one main hub in each cluster, i.e., the point with the uppermost hubness score.

6. Deterministic Approach

A simple way to occupation hubs for clustering is to routine them as one would generally use centroids. In addition, this agrees us to make a straight judgment with the K-means method.

The algorithm, referred to as K-hubs, is given in

```
initializeClusterCenters();
Cluster[ ] clusters = formClusters();
repeat
    forall Cluster c $\in$ clusters do
        Document h = findClusterHub(c);
        setClusterCenter(c, h);
    end for
    clusters = formClusters();
untilnoNewDocument
return clusters
```

Algorithm 1.K-hubs

After initial assessment on artificial data, it developed strong that even though the procedure achieves to discovery worthy and smooth best arrangements frequently, it is fairly complex to initialization. To growth the possibility of discovery the universal best, we resorted to the stochastic approach. Even though K-hubs exhibited low stability, it converges to cluster configurations very rapidly, in no more than four repetitions on all the data sets recycled for testing, most of which contained around 10,000 data instances.

7. Probabilistic Approach

Even though points with maximum hubness scores are without doubt the key candidates for cluster centers, there is no need to disregard the information about hubness scores of other points in the data. The temperature factor was introduced to the algorithm, so that it may start as being totally probabilistic and ultimately end by performing deterministic K-hubs repetitions.

The aim why hubness-proportional clustering is possible in the background of high dimensionality lies in the skewness of the circulation of k-occurrences. Specifically, there exist many data points having low hubness scores, making them bad applicants for cluster centers. Such points will have a small probability of being selected. To further highlight this,use the square of the actual hubness score in its place of making the likelihoods directly proportional to N_k (x).

8. A Hybrid Approach

The algorithms do not require knowledge of data/object demonstration, so all that is essential is a distance/similarity degree well-defined for each pair of data objects. If the demonstration is also available such that it is possible to expressively compute centroids, there also exists a third alternative: use point hubness scores to guide the examiner, but choose a centroid-based cluster conformation in the end. We will refer to this algorithm as hubness-proportional K-means (HPKM). It is nearly like to HPC, the only modification being in the deterministic phase of the repetition, as the conformation cools downcast through the hardening method: instead of relapsing to K-hubs, the deterministic phase executes K-means updates.

```
initializeClusterCenters();
Cluster[] clusters = formClusters();
repeat
  float $\theta$ =getProbFromSchedule(t);
  forall Cluster c $\in$ clusters do
    ifrandomFloat(0, 1) < $\theta$ then
      Document h = findClusterCentroid(c);
      setClusterCenter(c, h);
    else
      for all Document x $\in$ c do
        setChoosingProbability(x, $N_k^2$ (x));
      end for
      normalizeProbabilities();
      Document h = chooseHubProbabilistically(c);
      setClusterCenter(c, h);
    end if
  end for
  clusters = formClusters();
untilnoNewDocument
```

20. **return** clusters

Algorithm 2. HPKM

There are, indeed, cases when HPKM might be desirable to the pure hubness-based approach of K-hubs and HPC. Even although our initial tests propose that the main hubs lie near to limited cluster means in high dimensional data, there is no potential that this would grip for all clusters in each probable data set. It is sensible to assume there to be disseminations which main to such native data structure where the main hub is not amongst the greatest dominant points. Also, a perfect cluster formation on an agreed real-world data set is occasionally unbearable to attain by points as centers, mean while centers may need to be situated in the unfilled space among the points.

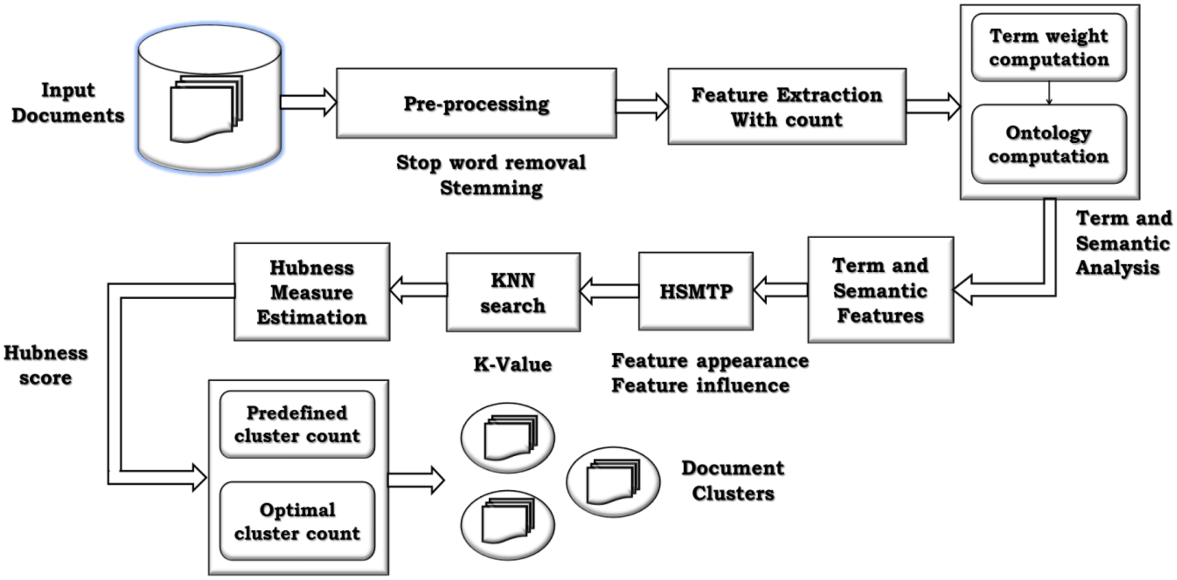

Figure 1.HPKM clustering Process

9. Integration of SMTP with HUBNESS Relationship for Document Clustering

The Similarity Measure for Text Processing (SMTP) scheme is integrated with hubness based distance analysis scheme. Document clustering is performed using Hubness-proportional K-means (HPKM). Semantic analysis methods are integrated to improve the document clustering process. Optimal cluster count estimation mechanism is used to improve the clustering process.

The system is designed to cluster the text documents with statistical and conceptual relationships. Weight values are integrated in the presence and absence based similarity analysis mechanism. Hubness relationship based partitioning mechanism is adapted in the system. The system is divided into six major modules. They are Document Preprocess, Term and Semantic Analysis, Document Similarity Estimation, KNN Search, Hubness Measure Estimation and HPKM Clustering Process.

Document parsing, stop word elimination and stemming process are carried out under document pre-process module. Weight estimation is carried out under the term and semantic analysis. Document similarity estimation is performed with the appearance and influence factors. Nearest neighbour data values are identified using the KNN search module. Hubness score values are estimated with the KNN search results.

Hubness-proportional K-means (HPKM) Clustering process is applied to partition the documents. Figure 1 depicts the process of HPKM algorithm.

- *Document Preprocess*

Document pre-process is performed to parse the text documents into words. Document cleaning is applied to remove stop words. Stemming process is applied to detect the base term. Terms are updated with their frequency values.

- *Term and Semantic Analysis*

The term analysis is performed to estimate the term weight values. Statistical method is used for the term weight estimation process. Term frequency (TF) and Inverse Document Frequency (IDF) are used for the term weight estimation process.Tf measures how frequently a term occurs in a document. IDF measures how important a term is. Term frequency is often divided by the document length. Dimensionality reduction process is carried out to remove infrequent feature values. The semantic analysis is performed to identify the concept relationships. Ontology is constructed for the selected domains. Terms and associated concept relationships are identified using the Ontology. Semantic weights are assigned with reference to the concept relationship type. Ontology considers meaning, logical part and its type.

- *Document Similarity Estimation*

Hybrid Similarity Measure for Text Process (HSMTP) mechanism is used for the distance analysis process. Feature appearance and feature influence details are integrated in the distance estimation process. Feature influence weight value is used in the distance estimation process. Distance analysis is carried out with the term and semantic features.

- *KNN Search*

Nearest neighbor transactions are identified in the KNN search process. K Nearest Neighbor (KNN) search algorithm is used for the nearest neighbor identification process. KNN search process is applied for all the documents. Similarity measures are used in the KNN search process.

- *Hubness Measure Estimation*

Hubness measure estimation is applied for all the documents. KNN search results are used in the hubness measure estimation process. Hubness measure is assigned for each transaction data. Hubness measure is used for the centroid selection process.

- *HPKM Clustering Process*

Hubness-proportional K-means (HPKM) algorithm is used for the data partitioning process. Clustering process is performed in two ways. Predefined cluster count based data partitioning uses the cluster count collected from the user. Optimal cluster count estimation mechanism is used to identify the feasible cluster count automatically.

10. Conclusion

Similarity measurement is used to estimate the relationship between the records or documents. Similarity Measurement for Text Process (SMTP) scheme is used to estimate the distance values. Statistical weight and concept weight models are used to improve the similarity measurement process. Hubness-proportional K-means (HPKM) algorithm performs the clustering process with SMTP. Concept relationship based similarity analysis is applied in the document relevancy analysis process. Clustering accuracy is improved in the system. The system improves the correctness in quantitative noisy environment. High scalability is supported with hubness relationship mechanism. Automated cluster count estimation mechanism is adapted to perform the document partitioning process.

References

[1] Masumi Shirakawa, Kotaro Nakayama, Takahiro Hara, ShojiroNishio, “Wikipedia-Based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes”, IEEE Trans On Emerging Topics In Compt, Volume 3, No. 2, June 2015.

[2] Yung - Shen Lin, Jung - Yi Jiang and Shie- Jue Lee, “A Similarity Measure for Text Classification and Clustering”, IEEE Trans On Knowledge And Data Engineering,Vol.26, No. 7, pp. 1575-1590, July 2014.

[3] NenadTomasev, Milos Radovanovi, DunjaMladeni and MirjanaIvanovi, “The Role of Hubness in Clustering High-Dimensional Data”, IEEE Trans On Knowledge And Data Engineering, Vol. 26, No. 3, pp. 739-751, March 2014.

[4] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Ou Liu, “An Ontology-Based Text-Mining Method to Cluster Prop for Research Proj Selection,” IEEE Trans on System, Man, Cybernetics—Part A: Systems& Humans, vol. 42, no. 3, pp. 784-790,May 2012.

[5] Taiping Zhang, Yuan Yan Tang, Bin Fang and Yong Xiang, “Document Clustering in Correlation Similarity Measure Space”, IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 6, June 2012.

[6] E. Meij, W. Weerkamp and M. de Rijke, ``Adding semantics to microblog posts," in Proc. ACM International Conferenceon Web Search Data Mining (WSDM), pp. 563_572, Feb. 2012.

[7] Y. Song, H. Wang, Z. Wang, H. Li, W. Chen, ``Short text conceptualization using a probabilistic knowledgebase," in Proc. International Joint Conf on Artificial Intelligence (IJCAI), pp. 2330_2336, July 2011.

[8] Jesus Oliva, Jose Ignacio Serrano, Maria Dolores del Castillo, Angel Iglesias, “SyMSS: A syntax – based measure for short-text semantic similarity,” ELSEVIER Data and Knowledge Engi, pp. 390-405,Jan 2011.

[9] Shady Shehata, FakhriKarray, Mohamed S. Kamel, “An Efficient Concept-Based Mining Model For Enhancing Text Cluster,” IEEE Trans On Knowledge And Data Engineering, vol. 22, no. 10,pp.1360-1371, October 2010.

[10] CongnanLuo, Yanjun Li, Soon M.Chung, “Text Document Clustering Based On Neighbors,” ELSEVIER Data and Knowledge Engineering, pp. 1271-1288, July 2009.

[11] K. M. Hammouda and M. S. Kamel, “Hierarchically distributed peer-to-peer document clustering and cluster summarization,” IEEE Trans on Knowledge and Data Engineering, vol. 21, no. 5, pp.681–698, May 2009.

[12] H. Chim and X. Deng, “Efficient phrase-based doc similarity for cluster,” IEEE Transactions on Knowledge and Data Engineering., vol. 20, no. 9, pp. 1217–1229, Sept 2008.

[13] D. Cai, X. He, and J. Han, “Document clustering using locality preserving indexing,” IEEE Transactions On Knowledge And Data Engineering, vol. 17, no. 12, pp. 1624–1637, December 2005.