ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

An improved interior-exterior informative similarity measure for web document clustering

[1]Reka, M. and [2]Dr.N. Shanthi

[1]Research Scholar, K.S. Rangasamy College of Technology, Tiruchengode, Tamil Nadu,India
Rekamca2010@gmail.com
[2] Professor &Dean, Department of CSE, Nandha Engineering College Erode,Tamil Nadu, India
shanthimoorthi@yahoo.com

Abstract

The World Wide Web grows at each fraction with number of documents. Such growth introduces challenge in clustering the documents. There are number of clustering algorithms has been discussed earlier but suffers to achieve clustering efficiency. To overcome the deficiency, the proposed algorithm introducedan efficient clustering algorithm which consider the relevancy of documents to be measured with internal and external documents. The method first computes the informatic similarity measure with all clusters and selects a higher one. In the second stage, the method compute the internal informative similarity and external informative similarity to compute the Informative weight. Based on computed informative weight the method assigns the class label for the web document. This algorithm uses 5,00,000 web documents for evaluation and 70 percent as training set and 30 percent as testing set. The method produces higher classification accuracy with less time complexity.

Index Terms

Clustering, Web Documents, Informative Measure, Interior-Exterior Similarity.

INTRODUCTION

The web is the large medium which contains enormous number of documents of different topics and categories. The number of documents in the web is growing all the time and lookup process facing many challenges due to the dimension. The web document contains many information regarding many topics and for each topic there will be number of documents present in the web. So identifying the related documents at the requirement is becoming a challenging task. The people use the web for many purpose and they always search for some information about any topic in the web. In order to provide them in efficient manner the documents of the web must be organized in proper manner. Also it is not necessary that the web document should speak about a specific topic but it can speak about many. So grouping the documents of the web is highly required one.

The clustering is the process of grouping the web documents into different category or classes. The popular K-means algorithm computes the distance between the documents of the class to perform clustering. There are number of other clustering algorithms available to group the documents under different classes. The problem with the earlier approach is the dimensionality, the k-means algorithm computes distance between the points and it cannot handle high dimensions. Similarly each algorithm has different issues in grouping the documents. The efficiency of clustering is highly depending on the measure being used.

The informative measure is the major measure being used to perform document clustering. The informative measure represents how depth a particular subject is covered in the document. For example, if you compute the informative measure for the subject “Network”, the informative measure Im can be computed as follows:

Im = Number of terms covered in the document/ Total number of terms belongs to the subject.

This is very much similar to the term frequency measure used in text cluster. More than that the informative measure can be computed using the taxonomy of words related to many subjects. Using the pure informative measure will not help in improving the clustering performance. To improve the clustering performance a new measure is introduced in this paper.

Interior informative measure (IIM) is the value which is computed based on the taxonomy of words extracted from the list of documents of the class. If there exists N number of documents in the class C, then the interior informative measure is computed based on the number of terms from the documents set of class C, and the number of terms from the input document. This represent how depth the document is describing the topic and how it is close to the document set of the class C.

Similarly the exterior informative measure (EIM) is computed based on the taxonomy of words being extracted from the document set of the class C and the terms set being extracted from the document given as input. The exterior informative measure is the value which represent the Document closure to the document set of the other class C. By combing both the measure the document clustering can be computed in efficient manner.

RELATED WORKS

There are number of clustering approaches has been discussed for the problem of web document clustering and this section discuss about some of the methods relate to web document clustering.

Agent-based document clustering [1] uses semantic ontology and fuzzy rule sets. The method extracts the features from the document and feature reduction is performed with the help of synset values obtained from wordnet. The method obtains the related terms for each of the semantic class with the help of wordnet and the mapping is performed by the semantic ontology. The method improves the performance of clustering by reducing the feature set and reduces the time complexity as well.

Self-Organizing Map -based Document Clustering Using WordNet Ontologies [3], propose a semantic text document clustering approach that using WordNet lexical and Self Organizing Maps. The proposed approach uses the WordNet to identify the importance of the concepts in the document. The SOM is used to cluster the document. We use this approach to enhance the effectiveness of document clustering algorithms. The approach takes the advantages of the semantics available in knowledge base and the relationship between the words in the input documents. Some experiments are performed to compare efficiency of the proposed approach with the recently reported approaches.

On ontology-driven document clustering using core semantic features [4], discusses that, an ontology can be used to greatly reduce the number of features needed to do document clustering. An Efficient Semantic VSM based Email Categorization Method [5], select related semantic features that will increase the global information, and use them to enrich the semantic feature of an email. The proposed categorization method based on sVSM creates the sementic feature of an email category by both extracting terms of training email and enriching these terms with their concept-chains in WordNet. On Document Representation and Term Weights in Text Classification [6], explore the potential of enriching the document representation with the semantic information systematically discovered at the document sentence level. The salient semantic information is searched using a frequent word sequence method. Different from the classic tfidf weighting scheme, a probability based term weighting scheme which directly reflect the term’s strength in representing a specific category has been proposed.

Topic map based document clustering [6] extracts the useful terms from the document and convert them into a semantic graph. The method maintains a semantic graph, which is generated with the help of semantic ontology. At the testing phase, the method computes a semantic similarity measure computed at each level. Based on the computed semantic similarity measure, a sub space or class of the document is identified.

A clustering approach for scientific literature and news group document [10] uses correlated concepts. The proposed approach discusses the frequency based maximum resemblance. Correlated concept based maximum resemblance document clustering method works using correlation terms. The algorithm has been evaluated for F-measure and purity

All the methods has the problem of poor clustering accuracy and suffers with higher false classification ratio.

INTERIOR EXTERIOR INFORMATIVE SIMILARITY BASED DOCUMENT CLUSTERING

The entire process has been split into four different stages namely preprocessing, IIM estimation, EIM estimation and clustering. This section will discuss all the stages in detail.

Preprocessing

In this stage, the method extracts the term from the input document. The input web document contains different HTML tags. Such presentation tags are removed and only the textual terms are considered. Extracted textual content is split into number of single terms by splitting them according to the space character. The extracted terms are added to a term set and for each term present in the term set, the presence of stop word is verified and removed from the term set. Then for each term from the term set, the method performs stemming operation which extract the pure terms from the document. For tagging the Stanford part of speech tagger is used.

```
Algorithm:
Input: Document Di
Output: Term Set Ts.
Start
    Read Document Di.
    Split text into term set Ts.
    Ts = Split(T, " ");
    For each term Ti
        Ti=$\int_{i=1}^{size(Ts)} Stemming(Ti,\text{ed,ing});$
    End
    For each term Ti
        Ti = $\int POSTagging(Ti)$
    End
Stop.
```

The preprocessing algorithm applies the tagging and stemming process to identify the pure nouns from the document given. The input web document would contain many presentation and textual contents. The preprocessing algorithm extract the terms and identifies the pure terms or root words using stemming and tagging process.

Informative Similarity Measure (ISM)

The informative similarity measure is the value computed based on the taxonomy of words extracted from the document and wordnet. First the method extract the terms present in the input document Di, and extract the term set from each Class of Document set (CDs). For each class of document the method extract the term set and selects the pure nouns by applying the preprocessing technique. Once the term set has been extracted, the method identifies the list of related terms by using the wordnet and odp taxonomy. All the terms obtained from the both taxonomy is used to compute the informative similarity measure for the input document.

```
ISM Algorithm:
Input: Document Di, Class C
Output: Informative Similarity Measure Tsm.
Start
    Read document Di
    Term set Ts = Preprocessing(Di).
    Document Set Ds = $\sum Documents \in C$
    Initialize document term set DTs.
    For each document Di
        DTs = $\sum(Terms \in DTs) \cup Preprocessing(Di)$
    End
    For each term Ti from DTs
DTs = $\sum(Terms \in DTs) \cup Wordnet(Ti) \cup ODP(Ti)$
    End
    Compute ISM.
        ISM = $\frac{\sum Ti(Ts) \in DTS}{size(DTS)} \times \frac{Size(Ts)}{size(DTs)}$
Stop.
```

The informative similarity is the measure which represents the similarity of the topic being discussed in different documents. For any input document the informative similarity can be computed between the documents of the class. The above discussed algorithm compute the informative similarity measure using the wordnet taxonomy and the open directory project taxonomy. This will be used to perform clustering in the final stage.

Interior Informative Measure

The interior informative measure is the value of closure which is computed between the documents of any specific class. The given document may come closure to any class but when you think about the closeness between the document of the class it may be scatter. So in order to measure the closeness between the documents of the class the IIM is computed. First the terms set are extracted and using the taxonomy the terms of other documents of the class, the IIM measure is computed.

```
IIM Algorithm:
Input: Document Di, Class C.
Output: IIM.
Start
    Read Document Di.
    Term set Ts = Preprocessing(Di).
    For each document Dk of C
        DTS = $\sum(Terms \in DTs) \cup Preprocessing(Dk)$
    End
    For each term Ti from DTs
            End
    Compute interior informative similarity Isim.
    IIM = $\frac{\sum Terms(Ts) \in DTs(Dk)}{size(Ts(Di))} \times \frac{\sum Terms(Ts) \in DTs(Dk)}{number\ of\ documents\ of\ Ci}$
Stop.
```

The above discussed algorithm computes the interior informative measure to be used in clustering the web document..

Exterior Informative Measure

The exterior informative similarity measure representthe topical measure for the input document Di which belongs to the class Ci, towards other class of documents. This is computed by measuring the closeness of the terms and their presence the documents of other classes. The method first identifies the terms and identifies the nouns using the word net taxonomy. Using the terms identified and the nouns identified, the method compute the exterior informative similarity measure. The exterior informative measure is the value computed based on the presence of terms in the other class documents.

```
Input: Input Document D, Target Class TC.
Output: EIM.
Start
    Read Document D.
    Term set ITs = Preprocessing(D).
    For each document Dk of TC
        Target term set TTS = $\sum(Terms \in TTs) \cup Preprocessing(Dk)$
    End
    For each term Ti from TTs
        TTs = $\sum(Terms \in TTs) \cup Wordnet(Ti) \cup ODP(Ti)$
    End
    Compute exterior topical similarity Esim.
        EIM = $\frac{\sum Terms(ITs) \in TTs(Dk)}{size(ITs(D))} \times \frac{\sum Terms(ITs) \in TTs(Dk)}{number\ of\ documents\ of\ TC}$
Stop.
```

The above discussed algorithm computes the exterior informative measure to be used in clustering the web document. The exterior informative similarity measure shows the similarity of the document to the other class than the class being considered.

In this method, the informative similarity measure for each class is computed using the taxonomy. Based on computed measure, the method selects a single class. Then for each class, the method computes the interior and exterior informative similarity measure. Finally ainformative weight is computed based on which the document is assigned to a class.

```
Input: Document D, Classes C
Output: Highest  CumulativeInformative Similarity Value.
Begin
    Identify list of terms and pure nouns.
    For each class
        Compute informative similarity.
        Isim = $\frac{\sum Terms(C)\in Ts}{size(C)}$
    End
    Choose the class with higher informative similarity.
    For the class selected
        Compute Interior informative similarity Isim.
For each document Di from class Ci
Compute interior informative similarity Isim.
    IIM = $\frac{\sum Terms(Ts)\in Ts(Di)}{size(Ts(Di))} \times \frac{\sum Terms(Ts)\in Ts(Dk)}{number\ of\ documents\ of\ Ci}$
        End
        Compute exterior informative similarity
EIM = $\frac{\sum Terms(Ts)\in Ts(Di(Ck))}{size(Ts(Di(Ck)))}$
        End
        Compute cumulative informative similarity
CIs = ITM×ETM
    End
    Choose the class with higher CIs value.
End
```

The above discussed clustering algorithm computes the IIM and EIM measures to identify the class of the document.

RESULTS AND DISCUSSION

The proposed method have been evaluated using different data set to test the performance and effectiveness of the web document clustering. Table-I shows the details of the data set used. Three different data sets have been shown to each one has the number of documents collected at different sessions. The evaluation was done on different data sets with varying size of web pages and the session length.

The outcome of the experimentation demonstrates that Interior-Exterior Informative Similarity Based Web Document Clustering method has achieved 99.37% of accuracy in clustering which is a drastic improvement compared to concept based document indexing(Fatiha et al 2010).

Data set	No. of Pages	Total Sessions	Avg. Session length(in minutes)
UCI	989818	118,718	6.4
Meme Tracker	5023	172,984	5.5

TABLE I : DETAILS OF THE DATA SET USED FOR THE EVALUATION

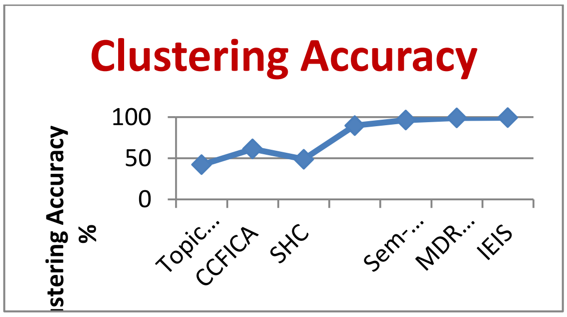

FIGURE 1: RESULT OF CLUSTERING ACCURACY

The clustering accuracy has been predicted and compared among various methods and the number of documents has been assigned with correct class labels. The comparative result on clustering accuracy has been presented in the Figure 1. The result proves that the proposed method has improved the clustering accuracy.

Conclusion and Future Enhancement:

In this proposed work, an efficient interior and exterior informative similarity based web document clustering is presented. The method preprocess the documents to identify the pure terms using the stemming and tagging process. In the second stage, the informative similarity measure is computed towards each category of documents. Based on computed informative similarity measure a single class is selected. Then the clustering is evaluated by computing the interior and exterior informative similarity measure. The method has produced higher clustering accuracy than others and achieves the efficiency upto 99.37 %. Further the performance of document clustering can be improved by computing informative measure for subclasses of each category using multi level informative measure estimation technique.

REFERENCES

[1] Khaled M Fouad and Moataz O Hassan. Article: Agent for Documents Clustering using Semantic-based Model and Fuzzy.*International Journal of Computer Applications* 62(3):10-16, January 2013.

[2] Fellbaum, C. (2010). WordNet. Theory and Applications of Ontology: Computer Applications, 231, PP: 231-243, Springer Science+Business Media B. V.

[3] Gharib, T. , Fouad, M. , Mashat, A. & Bidawi, I. (2012). Self-Organizing Map -based Document Clustering Using WordNet Ontologies, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2.

[4] Liu, Y. (2009). On Document Representation and Term Weights in Text Classification. In: Handbook of Research on Text and Web Mining Technologies. PP: 1-22. DOI: 10. 4018/978-1-59904-990-8. ch001. IGI Global.

[5] B. Fatiha, B. Mohand, T. Lynda, D. Mariam. (2010). Using WordNet for Concept-Based Document Indexing in Information Retrieval, SEMAPRO: The Fourth International Conference on Advances in Semantic Processing, Pages: 151 to 157, IARIA.

[6] Dragoni, M. , Pereira, C. &Tettamanzi, A. (2010). An Ontological Representation of Documents and Queries for Information Retrieval Systems, IEA/AIE 2010, Part II, LNAI 6097, pp. 555–564, Springer-Verlag Berlin Heidelberg.

Mrs.M.Rekareceived the B.Sc. degree in Computer Science from Periyar University in 2002, the M.C.A. degree from Periyar University in 2005, the M.Phil. degree in Computer Science from Periyar University in 2006 and pursuing Ph.D. degree in Anna University from the year 2010.She has been working as an Assistant Professor at K.S.Rangasamy College of Technology, Tiruchengode, since 2008. Her research interests fall in the areas of Data Mining, Networking, Data Communications and Computer Graphics. She is a life member of ISTE (Indian Society for Technical Education).

Dr.N.Shanthi received the B.E. degree in Computer Science and Engineering in 1994 and the M.E. degree in Computer Science and Engineering in 2001 from Bharathiyar University, Coimbatore, Tamil Nadu, India. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in the department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India from 1994 to 2013, and currently working as Professor & Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 27 papers in the reputed international journals and 20 papers in the national and international conferences. She has published 2 books. She is supervising 13 research scholars under Anna University, Chennai. Her current research interest includes Document Analysis, Optical Character Recognition, and Pattern Recognition and Network Security. She is a life member of ISTE and annual member of ACM.