Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Intuitionistic Partition based Conceptual Granulation Topic-Term Modeling

By D. Malathi & S. Valarmathy

Bannari Amman Institute of Technology, India

Abstract- Document Analysis represented in vector space model is often used in information retrieval, topic analysis, and automatic classification. However, it hardly deals with fuzzy information and decision-making problems. To account this, Intuitionistic partition based cosine similarity measure between topic/terms and correlation between document/topic are proposed for evaluation. Conceptual granulation is emphasized in the decision matrix expressed conventionally as tf-idf. A local clustering of topic-terms and document-topics results in comparing dependent terms with membership degree using cosine similarity measure and correlation. A preprocessing of documents with intuitionistic fuzzy sets results in efficient classification of large corpus. But it depends on the datasets chosen. The proposed method effectively works well with large sized categorized corpus.

Keywords: document analysis, intuitionistic fuzzy, topic modeling.

GJCST-C Classification: K.6.3

Intuitionistic Partition based Conceptual Granulation Topic-Term Modeling

D. Malathi [α] & S. Valarmathy [σ]

Abstract- Document Analysis represented in vector space model is often used in information retrieval, topic analysis, and automatic classification. However, it hardly deals with fuzzy information and decision-making problems. To account this, Intuitionistic partition based cosine similarity measure between topic/terms and correlation between document/topic are proposed for evaluation. Conceptual granulation is emphasized in the decision matrix expressed conventionally as tf-idf. A local clustering of topic-terms and document-topics results in comparing dependent terms with membership degree using cosine similarity measure and correlation. A preprocessing of documents with intuitionistic fuzzy sets results in efficient classification of large corpus. But it depends on the datasets chosen. The proposed method effectively works well with large sized categorized corpus.

Keywords: document analysis, intuitionistic fuzzy, topic modeling.

I. Introduction

Document model in the information retrieval has three main components, namely Text Pre-processor, Topic Extractor and Corpus category-zation. These components are integrated to deploy knowledge extraction in information system. In spite of this, the growing data and its knowledge recognition complications have considerably encouraging the extensions of machine learning algorithms.

a) Document Model

The text document Modeling is observed as latent topics model. Various prominent approaches in machine learning are used to study the model. Document model is a mixture of topics [4]. Topics are inferred by the collection of correlated words. But unsupervised learning perspective is the pulse of bubbling out the topics. By modeling, varieties of mining range can be established with various subjects. The models try to observe the likely documents and tend to focus on topics. But document models are discriminant because of random words due to linguistic factors such as synonym, hyponym, Polysemy, etc.

b) Text Pre-processor

The functionalities essential for machine learning of document are document pre-processing and corpus representation. Stop words removal, word stemming, filtering to exclude certain words, are done within each document. This process is called pre-processing of documents. Obtained vocabulary is put up in the word-document matrix which is generally called as bag-of-words model. The document representations may be in binary (0, for nonoccurrence and 1 for occurrence of each term in a document), term frequency (tij - number of occurrence of ith word in jth document) and term frequency inverse document frequency (probable occurrence of tij' – distribution of ith word in jth document). Obtained data in this stage is huge in dimension, and lot of techniques [15] have been proposed for dimension reduction.

c) Topic Extractor

A topic model is a probabilistic model that can be considered as a mixture of topics, represented by probability distributions of words in a document. The latent variables or topics are the inferring components of this model. The main objective is to learn from documents the distribution of the underlying topics in a given corpus. Topic model is Text corpora representation by a co-occurrence matrix of words and documents. The probabilistic latent semantic analysis (PLSA) model [10] uses probability of words with given topics and probability of topics in a document, to build a topic model. The Latent Dirichlet Allocation (LDA) model [1], is another probabilistic approach which ties the parameters of all documents through hierarchical generative model.

d) Corpus Categorization

Text Categorization is a classical application of Text Mining [19], and is used in email filters, social tagging and automatic labeling of documents in business libraries. Text mining applications in research and business intelligence include, latent semantic analysis techniques in bioinformatics automatic investigation of jurisdictions plagiarism detection in universities and publishing houses, cross-language information retrieval, spam filters learning, help desk inquiries, measuring customer preferences by analyzing qualitative interviews, automatic grading, fraud detection or parsing social network for ideas of new products [9].

II. Literature Support

The theory of fuzzy set is Consider as a degree of membership assigned to each element, where the degree of non-membership is just automatically equal to

Author α σ: Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India. e-mails: malathisubbu@gmail.com, artmathy@gmail.com

its complement. However, human interpretation often does not express the corresponding degree of non-membership as the complement to 1. So, Atanassov [1][2][3] introduced the concept of intuitionistic fuzzy set that is meant to reflect the fact that the degree of non-membership is not always equal to 1 minus degree of membership, but there may be some hesitation degree.

Intuitionistic fuzzy set is a generalized constructive logic applied in fuzzy set. It is defined on a X of objects, with each object x is described by the degrees of membership and non-membership to a certain property,

$$\{(x,\mu_A(x),\nu_A(x)),x\in X\} \qquad (1)$$

$\mu_A(x)$ represents the degree x belongs to the set A and $\nu_A(x)$ represents the degree x does not belongs to the set A. The model is defined by the restriction

$$0\le\mu_A(x)+\nu_A(x)\le 1 \qquad \forall x\in X \qquad (2)$$

Therefore the degree of non determinacy of the object x with respect to the intuitionistic fuzzy set A is imposed as,

$$\pi_A(x)=\mu_A(x)+\nu_A(x) \quad \forall x\in X \qquad (3)$$

The model is well suited to represent a classification problem with high dimension. The confusion matrix of high dimension can be probably reduced to concept matrix of low dimension. The similarity measures [14] and distance measures [21][20] between two intuitionistic fuzzy sets can be applied in pattern recognition.

In this paper, a Partition based approach [16] inspired by Hierarchical segmentation [8] and topic based segmentation [6] are extended using Intuitionistic fuzzy set approach [23] for local centralization of conceptual words. The intuitionistic fuzzy set theory is applied in conceptual term/topic detection. A cosine similarity and correlation are taken into for defining membership degree and the non-membership degree respectively. The results using this measure found better with respect to the dataset chosen. In literature a intuitionistic fuzzy representation of images for clustering [18] [12] by utilizing a novel similarity metric are defined. But a minimal support is extended for text classification. So, a local centralization of conceptual terms using Intuitionistic logical clustering has been applied in the work.

III. Proposed Model - Intuitionistic Partition based Concept Granulation (IPCG)

Intuitionistic logic is a natural deduction system [13],that have introduction rules μ and elimination rules ν for the logical connectives and quantifiers. The document classification system needs conceptual terms (μ), non deterministic terms or noises (ν) with logics and reasons to quantify concept granules.

Let A be a tf-idf matrix of nXm represents corpus. Each value is associated to

- Set of terms representing the membership of domain $\mu_A(x)$
- Term representing the non membership of domain $\nu_A(x)$

```
Algorithm: IPCG
For each document {
        Lowercase, numbers, special characters from document
        Remove stop list words from document
      Split document into k partitions
      For each segment {
      Find frequency of words
      Prepare matrix with each segment as row and words as
columns
      Include non zero frequency as member
      Cosine similarity distance between each segments
calculated
      Discard the segment with least distance  }
   Single row or vector of a document has been found
   Intuitionistic Correlation to include conceptual terms in topic
   Classify the document and find entropy  }
```

The intuitionistic fuzzy set A is generated by

$$A=\{w_{ij},\mu_i(w_{ij}),\nu_i(w_{ij})\} \quad \text{where} \quad 0<w_{ij}<1 \qquad (4)$$

The similarity between words and on a topic is calculated by the cosine measure. Each document vector is normalized with the weight and length of terms in k partitions. Then the optimal term w_{ij} [16] should be picked from the non sparse term of k partitions. i.e. $\mu_A \in w_{ik}$.

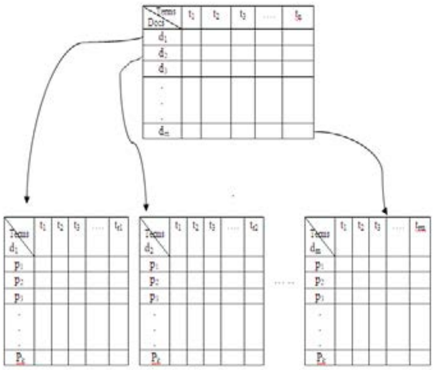


Figure 1 : Partition Model

{ri<=n (i.e. r is random or varies from document to document)(where i=1,2,…m),
k = no. of partitions or segments}

$$\mu_i = \frac{1}{|w_{ik}|}\sum w_{ik} \quad \text{where } w_{ik} > 0 \text{ and } |w_{ik}| >= k/2 \qquad (5)$$

The intuitionistic angular or cosine similarity [22] measure between the m terms in a partitioned set is given as follows:

$$C(A,B) = \frac{\sum_{i=1}^{m}\mu_A(x_i)\mu_B(x_i)}{\sqrt{\sum_{i=1}^{m}\mu_A^2(x_i)}\sqrt{\sum_{i=1}^{m}\mu_B^2(x_i)}} \qquad (6)$$

The intuitionistic correlation [7] of rows all fuzzy numbers are included from the samples of tf-idf (Partition Model). The crisp set is modified intuitionistically with the sample mean and variance of membership function as:

$$CR_I(A,B) = \frac{\left(\sum_{i=1}^{n}(\mu_A(x_i)-\bar{\mu}_A)(\mu_B(x_i)-\bar{\mu}_B)\right)}{\sqrt{\sum_{i=1}^{n}(\mu_A(x_i)-\bar{\mu}_A)}\sqrt{\sum_{i=1}^{n}(\mu_B(x_i)-\bar{\mu}_B)}} \qquad (7)$$

The effectiveness of the intuitionistic classification of corpus is approximately studied and analyzed using the following entropy [22] specifically used for Intuitionist Fuzzy Set 'A'.

$$E = \frac{1}{n}\sum_{i=1}^{n}\frac{\min(\mu_A(x_i),v_A(x_i)) + \pi_A(x_i)}{\max(\mu_A(x_i),v_A(x_i)) + \pi_A(x_i)} \qquad (8)$$

IV. Datasets

a) *Newspaper Article collection*

The newspaper articles under different topics are collected. The categories are marked. The training and testing documents are randomly chosen. The growing social media made essential to include newspaper article collection to include in this work. News are generally categorized by topic area ("politics," "business," etc.) written in clear, correct, "objective," and somewhat schematized language [5]. This would pave way to extend the research towards social networking and marketing. The collection includes about 780 documents with 25 categories. All new social relevant topics ("mobile","opinion", etc.) are included for categorizing.

b) *Reuters-21578 Data Set*

The Reuters-21578 Data Set collection provides a classification task with challenging properties. There are multiple categories, the categories are overlapping and non exhaustive, and there are relationships among the categories. There are interesting possibilities for the use of domain knowledge. There are many possible feature sets that can be extracted from the text, and most plausible feature/example matrices are large and sparse [11].

c) *Movie Review Dataset*

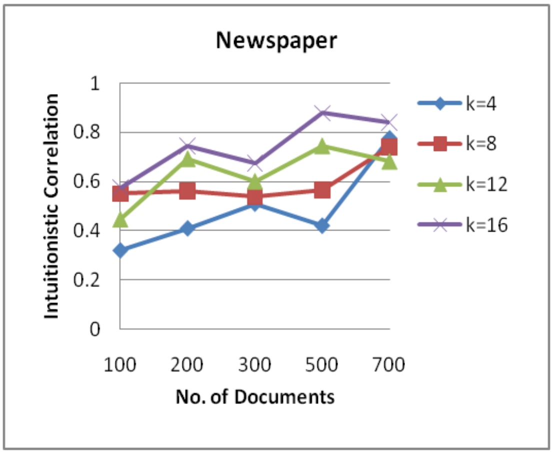
The Movie Review Dataset, Polarity dataset v0.9 with 900 positive and 900 negative reviews is used. Using movie reviews as data, the problem of classifying documents using standard machine learning techniques definitively outperform human-produced baselines processed reviews [17]. The training cases are chosen randomly from each class about 100 documents. Which means about 500 cases are considered for training.

V. Results and Analysis

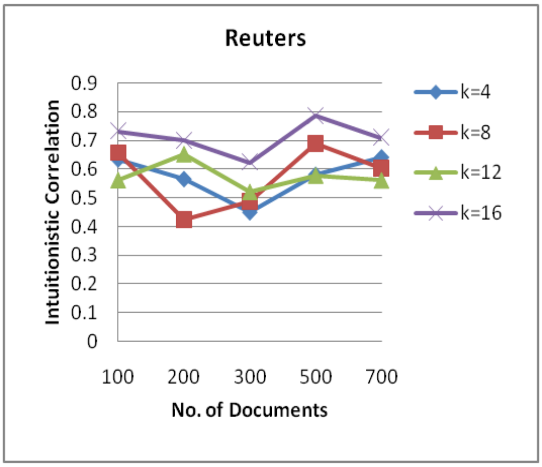
The machine learning classification methods, such as Bayesian, Naïve Bayes, J48, Support Vector Machines, LMT are strong enough to support classifications.

In the case of concept granulation in document classification, the feature selection is fine tuned to achieve categories strictly connected to the human perception. Before imposing the features into the classifier, some form of selection must be chosen. The proposed method, selects the features according to the intuitionist logic. The features tf-idf matrix has been transformed into intuition based feature model. The proposed approach is modeled as a probability distribution over the set of Topic/Words represented by the vocabulary. These distributions are sampled from multi-nominal distributions.

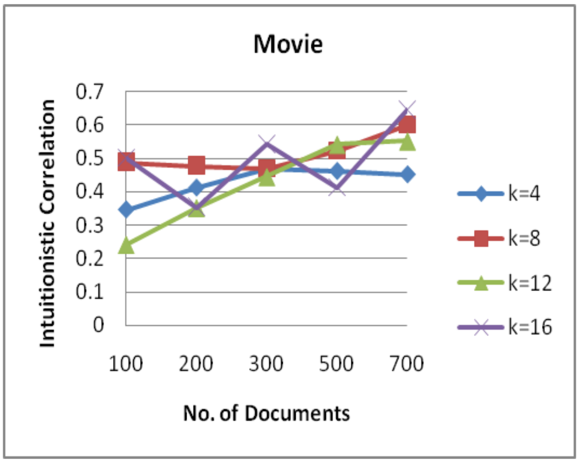
The proposed Concept Granulation Using Intuitionistic Partition Based Classification Model is implemented administered in the Java based system and analyzed for its significance. The intuitionistic correlation is applied to the specified datasets. In which the chosen dataset and the partitions play the very important role in finding the result of the model. The tf-idf-IP is favorable for Reuter dataset than for Newspaper and Movies. This is represented in the Figures 2(a) (b) (c). Reuters in which documents are well organized behaves highly significant to the model. In Newspaper collection, the documents are synthetically collected and organized. But due to the nature of news along with the temporal parameters, it is moderately supported by the model. The least support is favored by the movie dataset. This is due to the heterogeneity of the documents/terms/topics.



(a)



(b)



(c)

Figure 2 : Intuitionistic correlation Vs The number of training documents

The perplexity is depicted in Figure 3 and Table1. So the analysis can be interpreted or inferred in the following ways:

Intuitionistic approach is in favor of the classified documents or corpus chosen

Partition plays the important role in the proposed model. Out of four types of partition, k=8 plays a smoothened strong support for the proposed model

k=16, the highest partition yield only a very moderate result and more confusions.

k=4, the least partition model yield the smooth but less significant support for all the datasets.

k=8, yield the partially smooth but supportive significant for the movie dataset. (Than other partitions)

The results are focused to average training datasets and micro f-measure (Table 2) to show up the IPCG performs better with dimension reduction for categorization of corpus. Every datasets chosen for analysis behaves to the pull and push of various stages of the proposed model.

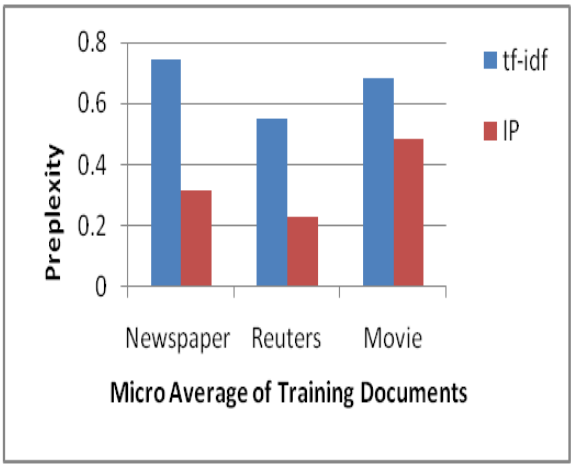


Figure 3 : Confusions in classification

Table 1 : Confusions in classification

Training with 300 Doc	Dimension Reduction	Perplexity	Correlation
Newspaper	26%	0.231	0.582
Reuters	22%	0.311	0.520
Movie	16%	0.483	0.480

Table 2 : Micro Evaluation of F-measure with average training sets

Dataset Classifiers	tf-idf			IPCG		
	Reuters	News Paper	Movie	Reuters	News Paper	Movie
SVM	0.482	0.422	0.321	0.844	0.841	0.799
NB	0.401	0.369	0.297	0.872	0.834	0.810
J48	0.400	0.399	0.381	0.798	0.797	0.784
Bayes'	0.541	0.411	0.399	0.831	0.854	0.829
LMT	0.442	0.541	0.587	0.878	0.798	0.722

VI. Conclusions

In this paper, we have proposed a intuitionistic partition based concept granulation topic-term model for a nominal tf-idf vector space model which is often used in information retrieval, topic analysis, and automatic classification. The cosine distance and correlation treatment to the tf-idf reduces the dimension and improves the efficiency of bag of words/terms in topics. However, it is priory treated using the intuitionistic partition for fitting the model into decision-making problems. To account this, Intuitionistic partition based cosine similarity measure between topic/terms and correlation between document/topic are included. The proposed fuzzy model is tailored with normal combinational approach to fetch intuitionistic fuzzy crisp set. Yet, it is observed the model is well behaving and promising for the categorized documents and not so bad support for the low inference corpus collections like movie review. So, this make us clear that the social media documents should be specially treated before introducing this model. It is felt that aggregation of social media topic-terms is needed. This is taken for future work or extension of the proposed work.

References Références Referencias

1. D. M. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3(2003), 993-1022.
2. T. Hofmann. Probabilistic Latent Semantic Analysis. *In Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99),* San Francisco, CA, Morgan Kaufmann, (1999), 289–329.
3. F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, 34(2002), 1-47. doi:10.1145/505282.505283.
4. I. Feinerer, K. Hornik and D. Meyer, *Journal of Statistical Software*, 25, (2008), 1-54.
5. K.T. Atanassov, Intuitionistic Fuzzy Sets, Theory, and Applications, Series in Fuzziness and Soft Computing, *Phisica-Verlag*, 1999.
6. K.T. Atanassov, Intuitionistic Fuzzy Set*, Fuzzy Sets System*, (1986) 87–97.
7. K.T. Atanassov, S. Stoeva, Intuitionistic Fuzzy Set, *Polish Symposium on Interval and Fuzzy Mathematics, Poznan*, (1993), 23–26.
8. D. Li and C. Cheng, New Similarity Measures Of Intuitionistic Fuzzy Sets And Application To Pattern Recognition, *Pattern Recognition Letter*, 23(2002) 221–225.
9. E. Szmidt and J. Kacprzyk, Entropy for Intuitionistic Fuzzy Set, *Fuzzy Sets System*, 118 (2001), 467–477.
10. E. Szmidt and J. Kacprzy k, Distance Between Intuitionistic Fuzzy Set, *Fuzzy Sets System*, 114(2000), 505–518.
11. D. Malathi, S. Valarmathy, Domain Classifier using Conceptual Granulation and Equal Partition Approach, *Indian Journal of Engineering, Science and Technology*, 7(2013), 39-43.
12. J. T. Chien and C. H. Chueh, Topic-Based Hierarchical Segmentation, *IEEE Transactions on Audio, Speech and Language Processing*, 20(2012), 55-66.
13. T. Brants, F. Chen and I. Tsochantaridis, Topic-based document segmentation with Probabilistic Latent Semantic Analysis, *in the proceeding of International Conference on Information and Knowledge Management*, (2002), 211–218.

14. Z. Xu, J. Chen and J. Wu. Clustering Algorithm for Intuitionistic Fuzzy Sets. Information Sciences, 178(2008), 3775-3790.
15. N. Pelekis, D. K. Iakovidis, E. K. Evangelos and I. Kopanakis. Fuzzy Clustering of Intuitionistic Fuzzy Data, *International Journal of Business Intelligence and Data Mining*, 3(1), pp. 45-65.
16. D. K. Iakovidis, N. Pelekis, E. K. Evangelos and I. Kopanakis, Intuitionistic Fuzzy Clustering with Applications in Computer Vision. *Advanced Concepts for Intelligent Vision Systems, Lecture Notes in Computer Science,* 5259(2008), 764-774.
17. V. H. Le, C. H. Nguyen and F. Liu, Semantics and Aggregation of Linguistic Information, Based on Hedge Algebras, *The 3rd International Conference on Knowledge, Information, and Creativity Support Systems*, (2013).
18. J. Ye, Multicriteria Decision-making Method Based on a Cosine Similarity Measure between Trapezoidal Fuzzy Numbers, *International Journal of Engineering, Science and Technology,* 3(2011), 272-278.
19. D. A. Chiang, and N. P. Lin, Correlation of fuzzy sets, *Fuzzy Sets and Systems*, 102(1999), 221-226.
20. B. Berendt, Text Mining for News and Blogs Analysis. *In C. Sammut, & G. I. Webb, Encyclopedia of Machine learning*, London: Springer. (2010), 968-972.
21. http://www.daviddlewis.com/resources/testcollections/reuters21578.
22. B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proceedings of EMNLP*, (2002), 79-86.
23. D. Malathi and S. Valarmathy, A Comprehensive Survey on Dimension Reduction Techniques for Concept Extraction from a Large Corpus, *International Journal of Computing Information Systems,* 3(2011), 1-6.