AN EFFICIENT CONCEPT-BASED MINING MODEL FOR ENHANCING TEXT CLUSTERING USING WORD CO-OCCURENCE

1. Abstract :

Text clustering is a fundamental task in data mining, aiming to identify inherent groupings within a collection of documents. Traditional methods often rely on frequency-based approaches to represent text data, neglecting the underlying semantic structure. This project introduces an innovative Concept-Based Mining Model for enhancing text clustering using word co-occurrences. The proposed model goes beyond conventional frequency-based representations by capturing the semantic structure of each term within sentences and documents through word co-occurrence analysis. Three distinct measures are employed to analyze concepts at the sentence, document, and corpus levels, thereby improving the accuracy and effectiveness of the clustering algorithm.

2. Existing System :

The existing text clustering methods predominantly utilize traditional data mining techniques such as decision trees, conceptual clustering, data summarization, statistical analysis, neural networks, inductive logic programming, and rule-based systems. These approaches often rely on frequency-based representations of terms within documents, without explicitly considering word co-occurrence. While effective to some extent, these methods may not fully capture the nuanced semantic relationships between terms the limitations of the existing systems underscore the need for a more advanced approach that considers the semantic structure, particularly word co-occurrence, for improved clustering performance.

3. Proposed System:

The proposed Concept-Based Mining Model addresses the shortcomings of traditional text clustering methods by introducing a novel approach to feature representation that explicitly incorporates word co-occurrence. Instead of relying solely on term frequency within documents, the model captures the semantic structure of each term within sentences and documents through comprehensive word co-occurrence analysis. This nuanced representation enables a more accurate understanding of the underlying concepts in the text. Three key

measures, including word co-occurrence, are computed to analyze concepts at different levels sentence, document, and corpus. By leveraging word co-occurrences, the proposed system enhances the clustering process, leading to higher intra-cluster similarity and lower inter-cluster similarity. The emphasis on semantic information ensures that the clustering algorithm produces more meaningful and contextually relevant groupings of text documents.

4. Software Requirements :

- **Operating System:** Windows 10
- **Coding Language**: Java/Python
- **IDE** : Eclipse/ Net-beans/Vs Code IDE

5. Hardware Requirements :

- System : Single / Dual Core Processor and above.
- Hard Disk : 256GB
- Ram : 8GB.

Project Activity	**Recommended Duration**	**Suggested Duration By Faculty**
STAGE-I		
Topic Identification	2weeks	2 weeks
Literature Survey	8 weeks	8 weeks
Abstract Finalization	2 weeks	2 weeks
Requirement Analysis and feasibility	2 weeks	2 weeks
STAGE-II		
Modeling/Design	3 weeks	3 weeks
Implementation/Experimentation	7 weeks	7 weeks
Testing and results	2 weeks	2 week
Documentation	2 weeks	7 weeks

Mohd. Khaleel Uddin Siddiqui
20841A1211

P. Madhu
21845A1202

A.Shraddhanand
20841A1201

K.Sampath raju
20841A1223

Mrs. Durga Pavani
Name of the Guide

LITERTAURE REVIEW - II

Paper I: Extending The Scope Of Co-Occurrence Embedding

Authors: - Jinhong Mi, Jiren zhu

Published in year 2021

Abstract

The paper proposes a model that generalizes the scope of co-occurrence based embedding techniques to capture semantic and syntactic information of natural language. The model treats a natural language sentence as a continuous flow of possibly overlapping signals, and counts the co-occurrence of these signals to factor the co-occurrence matrix into vectors using the GloVe method. The model extends embedding to n-grams and synsets and POS-tags, and trains "gram vectors" and "combined-feature vectors" on the latest Wikipedia dump corpus and a smaller dataset, respectively. The results show that learning embedding for word and n-grams together preserves and improves the structures of word vectors, and the trained gram vectors integrate well into the sentiment analysis task.

Background

Word embedding techniques have shown to capture semantic and syntactic information of natural language and improve performance of various downstream tasks. However, they suffer from some intrinsic disadvantages, such as not properly understanding language and not capturing multi-word phrases. The paper aims to generalize the scope of co-occurrence based embedding techniques to capture semantic and syntactic information of natural language.

Methodology

The paper uses the GloVe method to factorize the co-occurrence matrix into vectors. It extends embedding to n-grams and synsets and POS-tags, and trains "gram vectors" and "combined-feature vectors" on the latest Wikipedia dump corpus and a smaller dataset, respectively. The model treats a natural language sentence as a continuous flow of possibly overlapping signals, and counts the co-occurrence of these signals to factor the co-occurrence matrix into vectors.

Gap Identification

The paper identifies the gap in current word embedding techniques that they do not properly understand language and do not capture multi-word phrases. It aims to generalize the scope of co-occurrence based embedding techniques to capture semantic and syntactic information of natural language.

Themes and Trends

Within the domain of language and text analysis, the paper delves into the intricate themes of natural language processing, word embedding, and co-occurrence-based embedding techniques. A pivotal contribution lies in its alignment with a prevailing trend—namely, the concerted effort to broaden the applicability of embedding techniques. This trend specifically aims to enrich these techniques with the capacity to capture both semantic nuances and syntactic intricacies inherent in natural language.

Synthesis Analysis

The paper engages in a thorough synthesis of experimental outcomes, revealing that the simultaneous learning of embeddings for both individual words and n-grams serves to not only preserve but also improve the underlying structures of word vectors. Notably, the trained n-gram vectors seamlessly integrate into the complex landscape of sentiment analysis tasks, showcasing a harmonious fusion of learned representations.

Discussion

Moving into a comprehensive discussion, the paper elucidates key experiment results, accentuating the model's prowess in generalizing the scope of co-occurrence-based embedding techniques. This generalization empowers the model to effectively capture the rich tapestry of semantic and syntactic information within natural language.

Conclusion

In its conclusive remarks, the paper affirms that the proposed model not only extends the scope of co-occurrence-based embedding techniques but does so with a nuanced proficiency in capturing the intricate dance between semantic and syntactic elements in natural language. The seamless integration of trained n-gram vectors into sentiment analysis tasks serves as a testament to the model's potential.

Paper II: Corpus-Based Topic Diffusion For Short Text clustering

Authors: - Chu Tao Zheng, Cheng Liu, Hau San Wong

Published in year 2018

Abstract

In this paper, we propose a novel corpus-based enrichment approach for short text clustering. Since sparseness brings about the problem of insufficient word co-occurrence and lack of context information, previous researches use external sources such as Wikipedia or WordNet to enrich the representation of short text documents, which requires extra resources and might lead to possible inconsistency. On the other hand, corpus-based approaches use no external information in mining short text data. By introducing a set of conjugate definitions to characterize the structures of topics and words, and by proposing a virtual generative procedure for short texts, we perform expansion on short text data. Specifically, new words which may not appear in a short text document were added with a virtual term frequency, and this virtual frequency is obtained from the posterior probabilities of new words given all the words in that document. The complete procedure can be regarded as mapping data points (documents) from the original feature space to a hidden semantic space (topic space). After performing semantic smoothing, data points are then mapped back to the original space. We conduct experiments on two short text datasets, and the results show that the proposed method can effectively address the sparseness problem.

Background

The paper addresses the challenge of sparseness in short text data, which arises from insufficient word co-occurrence and lack of context information in short texts. Traditional text clustering algorithms are less effective in the short text setting. Various methods have been proposed to deal with this issue, including the use of external information sources to enrich the representations of short texts. However, these methods raise concerns about the correctness of enrichment and the structural coherence between the external source data and the original data. The paper introduces a corpus-based expansion technique that does not rely on external resources, aiming to overcome the sparseness problem by mapping data points.

Methodology

The methodology proposed in the paper involves a corpus-based expansion technique to address the sparseness problem in short text data. This technique does not rely on external information sources for enrichment. Instead, it introduces a set of conjugate definitions to characterize the structures of topics and words, along with a virtual generative procedure for short texts. The expansion on short text data is performed by adding new words with a virtual term frequency, which is obtained from the posterior probabilities of new words given all the words in the document. The method achieves comparable performance with methods based on enrichment with external information source.

Gap Identification

The absence of a rigorous comparison with existing state-of-the-art methods limits insights into the proposed technique's relative efficacy. Additionally, the paper emphasizes its comparable performance with methods using external information sources, but it lacks a detailed exploration of potential challenges or limitations associated with its corpus-based approach. Future research could enhance the paper's impact by conducting a more extensive comparative analysis and providing a nuanced discussion on the practical implications, challenges, and potential areas for improvement related to the proposed expansion technique. This would strengthen the paper's contribution and provide a more comprehensive understanding of its applicability in diverse contexts.

Theme and Trend

The paper addresses the theme of addressing the sparseness problem in short text data, which has become increasingly important due to the prevalence of short texts in social media applications, microblogs, search result snippets, online commercial commons, and online advertisements. The trend in this area of research involves the development of methods to improve the effectiveness of text clustering algorithms for short text data. The paper contributes to this trend by introducing a corpus-based expansion technique that does not rely on external information sources for enrichment, aiming to effectively address the sparseness problem in short text data.

Synthesis Analysis

The paper presents a corpus-based expansion technique to address the sparseness problem in short text data. The method does not rely on external information sources for enrichment. Instead, it introduces a set of conjugate definitions to characterize the structures of topics and words, along with a virtual generative procedure for short texts. The expansion on short text data is performed by adding new words with a virtual term frequency, which is obtained from the posterior probabilities of new words given all the words in the document. .

Discussion

The paper discusses the challenges posed by insufficient word co-occurrence and lack of context information in short texts, making traditional text clustering algorithms less effective. The proposed method involves a corpus-based expansion technique that does not rely on external resources, aiming to overcome the sparseness problem by mapping data points from the original feature space to a hidden semantic space and then back to the original space after performing semantic smoothing. The paper conducts experiments on two short text datasets to demonstrate the effectiveness of the proposed method in addressing the sparseness problem, achieving comparable performance with methods based on enrichment with external information sources.

Conclusion

The paper concludes that the proposed corpus-based expansion technique effectively addresses the sparseness problem in short text data, achieving comparable performance with methods based on enrichment with external information sources. The method does not rely on external resources, making it a promising approach for improving the effectiveness of text clustering algorithms for short text data.

Paper III: Text Clustering using Semantics

Authors: - Bhoopesh Choudhary

Published in year 2012

Abstract

The paper proposes a corpus-based approach to address the sparseness problem in short text data. The method involves introducing a set of conjugate definitions to characterize the structures of topics and words and proposing a virtual generative procedure for short texts. New words are added to short text data with a virtual term frequency obtained from the posterior probabilities of new words given all the words in the document. The complete procedure maps data points (documents) from the original feature space to a hidden semantic space (topic space) and then back to the original space after performing semantic smoothing. Experiments on two short text datasets demonstrate that the proposed method can effectively address the sparseness problem, achieving comparable performance with methods based on enrichment with external information sources.

Background

The paper discusses the challenges posed by the sparseness problem in short text data, which arises from insufficient word co-occurrence and lack of context information in short texts. Traditional text clustering algorithms are less effective in the short text setting due to these challenges. The paper proposes a corpus-based approach to address the sparseness problem by introducing a set of conjugate definitions to characterize the structures of topics and words and proposing a virtual generative procedure for short texts. The approach involves adding new words to short text data with a virtual term frequency obtained from the posterior probabilities of new words given all the words in the document. The complete procedure maps data points (documents) from the original feature space to a hidden semantic space (topic space) and then back to the original space after performing semantic smoothing. The paper conducts experiments on two short text datasets to demonstrate the effectiveness of the proposed method in addressing the sparseness problem, achieving comparable performance with methods based on enrichment with external information sources.

Methodology

The methodology proposed in the paper involves a corpus-based approach to address the sparseness problem in short text data. The approach introduces a set of conjugate definitions to characterize the structures of topics and words and proposes a virtual generative procedure for short texts. New words are added to short text data with a virtual term frequency obtained from the posterior probabilities of new words given all the words in the document. The complete procedure maps data points (documents) from the original feature space to a hidden semantic space (topic space) and then back to the original space after performing semantic smoothing. The paper conducts experiments on two short text datasets to demonstrate the effectiveness of the proposed method in addressing the sparseness problem, achieving comparable performance with methods based on enrichment with external information sources.

Gap Identification

While the paper presents a corpus-based approach to address the sparseness problem in short text data, some notable gaps are evident. Firstly, the paper lacks a robust comparison with existing state-of-the-art methods, limiting insights into the relative efficacy of the proposed approach. Additionally, the methodology is primarily validated on two short text datasets, raising concerns about its generalizability to diverse contexts. Future research could enhance the paper's impact by conducting more extensive evaluations against a broader range of benchmarks and exploring the method's adaptability to different types of short text data beyond the datasets considered. This would provide a more comprehensive understanding of the proposed approach's strengths and potential limitations in real-world applications.

Themes and Trends

The paper addresses the theme of addressing the sparseness problem in short text data, which has become increasingly important due to the prevalence of short texts in social media applications, microblogs, search result snippets, online commercial commons, and online advertisements. The trend in this area of research involves the development of methods to improve the effectiveness of text clustering algorithms for short text data. The paper contributes to this trend by proposing a corpus-based approach to address the sparseness problem, which introduces a set of conjugate definitions to characterize the structures of topics and words and proposes a virtual generative procedure for short texts.

Synthesis Analysis

The paper proposes a corpus-based approach to address the sparseness problem in short text data. The approach involves introducing a set of conjugate definitions to characterize the structures of topics and words and proposing a virtual generative procedure for short texts. New words are added to short text data with a virtual term frequency obtained from the posterior probabilities of new words given all the words in the document. The complete procedure maps data points (documents) from the original feature space to a hidden semantic space (topic space) and then back to the original space after performing semantic smoothing.

Discussion

The paper discusses the challenges posed by the sparseness problem in short text data, which arises from insufficient word co-occurrence and lack of context information in short texts. The proposed corpus-based approach introduces a set of conjugate definitions to characterize the structures of topics and words and proposes a virtual generative procedure for short texts. The approach involves adding new words to short text data with a virtual term frequency obtained from the posterior probabilities of new words given all the words in the document. The complete procedure maps data points (documents) from the original feature space to a hidden semantic space (topic space) and then back to the original space after performing semantic smoothing. The paper conducts experiments on two short text datasets to demonstrate the effectiveness of the proposed method in addressing the sparseness problem, achieving comparable performance with methods based on enrichment with external information sources.

Conclusion

The paper concludes that the proposed corpus-based approach effectively addresses the sparseness problem in short text data, achieving comparable performance with methods based on enrichment with external information sources. The approach introduces a set of conjugate definitions to characterize the structures of topics and words and proposes a virtual generative procedure for short texts. The approach involves adding new words to short text data with a virtual term frequency obtained from the posterior probabilities of new words given all the words in the document. The complete procedure maps data points (documents) from the original feature space to a hidden semantic space (topic space) and then back to the original space after performing semantic smoothing. The paper contributes to the trend of developing methods to improve the effectiveness of text clustering algorithms for short text data.

Paper IV: Clustering of scientific articles using natural language processing

Authors:- Barbara Probierza, Jan Kozaka , Anita Hrabia

Published in year 2022

Abstract

The paper proposes using natural language processing (NLP) and K-means clustering to automatically cluster scientific articles based on their content. Different NLP measures like TF-IDF are analyzed as well as clustering the abstract vs. introduction. Experiments on over 1,500 articles show TF-IDF gives better clustering than TF or binary measures, and increasing the number of clusters improves connections between articles on similar topics.

Background

The number of online scientific articles has increased exponentially, making it difficult for researchers to find the most relevant papers. Providing keywords and abstracts is not enough. Clustering articles by topic can help with search and recommendations. Other works have tried clustering based on words, citations, etc. but there are still limitations.

Methodology

The proposed approach uses NLP including tokenization, stemming and lemmatization to process the text of abstracts or introductions. TF, TF-IDF and binary measures are extracted as features. Then K-means clustering with a predefined number of clusters is applied. The quality of the clustering is evaluated by comparing connections between articles based on their keyword overlap.

Gap Identification

The presented research on clustering scientific articles using natural language processing (NLP) and K-means clustering has certain gaps that warrant further exploration. Firstly, the study primarily focuses on abstracts and introductions, overlooking the potential insights derived from analyzing the full texts of scientific articles. This limitation could be addressed in future research to enhance the depth of the analysis. Additionally, while the paper mentions evaluating the quality of clustering, it lacks explicit details on the specific metrics employed for this assessment. Clearly defining and elaborating on the evaluation metrics would strengthen the validity of the results.

Themes and Trends

Clustering scientific documents is an active research area in bibliometrics. Different approaches utilize information from words, citations, and combinations thereof. Limitations exist with traditional frequency-based features. Keyword-based and graph-based methods are alternatives being explored.

Synthesis/Analysis

Experiments were conducted on over 1,500 articles, analyzing abstracts and introductions with TF, TF-IDF and binary measures and varying the cluster numbers. TF-IDF gave significantly better and more balanced clusters than the other measures. Increasing cluster numbers improved inter-article connections. Clustering full introductions outperformed just abstracts.

Discussion

The results confirm that TF-IDF and clustering article introductions can effectively group papers by research topic similarity. This could help recommend relevant papers or assign submissions to editors. Future work includes analyzing full texts and adding criteria to balance cluster sizes.

Conclusion

The paper presented a NLP and clustering approach to approximate documents by topic. TF-IDF measure and introductions gave the best clustering. Results show articles in clusters are twice as connected after clustering. This method can improve paper recommendations and assignments

References

[1] Massih R Amini and Nicolas Usunier. (2007). A contextual query expansion approach by term clustering for robust text summarization. In Proceedings of DUC, pages 48–55.

[2] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. (2017). A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of ICLR.

[3] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. (2007). Clustering short texts using Wikipedia. In Proceedings of SIGIR, pages 787–788. ACM.

[4] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. (2003). Model-based clustering and visualization of navigation patterns on a website. Data Mining and Knowledge Discovery, 7:399–424.

[5] Andrew M. Dai, Christopher Olah, and Quoc V. Le. (2015). Document embedding with paragraph vectors. CoRR, abs/1507.07998.

[6] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. (2016). Representation learning for very short texts using weighted word embedding aggregation. Pattern Recogn. Lett., pages 150–156.

[7] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. (2006). Reducing the dimensionality of data with neural networks. Science, pages 504–507.

[8] Andreas Hotho, Steffen Staab, and Gerd Stumme. (2003). Ontologies improve text document clustering. In Proceedings of ICDM, pages 541–544. IEEE.

Mohd. Khaleel Uddin Siddiqui 20841A1211

P. Madhu 21845A1202

A.Shraddhanand 20841A1201

K.Sampath raju 20841A1223

Mrs. Durga Pavani
Name of the Guide