IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 4, April 2017

Survey on Popularity Prediction of Movies

Prof D. D. Gatade[1] Nishigandha Aware[2], Sonali Mehetre[3], Anjali Khaire[4], Prathmesh Gunavat[5]

Assistant Professor, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India [1]
Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India [2]

Abstract: Now a day's online social media networking used in large amount to access information, share experiances, and express opinions. As a result large amount of data are generated in every day from social media channel such as twitter, facebook, youTube etc. From that, the concept of big data has arrived with large volume, complex and growing data. It becomes important issues in many areas like data mining, computational intelligence, the semantic web and social networking. Big data is very impotant to analyse and predict the future behaviour. Analysis of data is important for identifying the valuable data and to discover the useful knowledge and it help to improve decision making of individual user and companies. The Popularity prediction of movies before thetrical release is main challange and it require high level of artifical intelligence. For that we use the data of social media like YouTube which contains large information about people's preferences.

Keywords: Data Mining, Big Data, Sentiments, Cosine Similarity etc.

I. Introduction

Online social media networking used in large amount to access the information, share experiances and express opinion. As a result huge data is generated. From that concept of big data has arrived with large volume, complex and growing data. Big data is essential to analyse and predict future behaviour. For analysing social media data, we use big data framework such as Apache Hadoop to allow efficient application of data mining methods and machine learning algorithms. Hadoop allow efficient application of data mining method and machine learning algorithm.

Predicting the popularity of movie is main challange and it require high level of artificial intelligence. For that we used data of social media like YouTube which contains large information about people's preferences. We will do data analysis on comments, dislikes, likes, views about movie and sevearl aspects of movie popularity.

Posting reviews online has become an increasingly popular way for people to express opinions and sentiments toward the movies released. Analyzing the large volume of online reviews available would produce useful actionable knowledge that could be of economic values to vendors and other interested parties. In

this paper, we conduct a case study in the movie domain, and tackle the problem of mining reviews for predicting movie popularity. Social networking websites are used by a large sector of people, of almost all age groups across the globe. People,ranging from Celebrities to a common man, use it to connect with other people and express their views on topics such as politics, economy, entertainment, etc. Social media is seen as a means of gathering insights into human behaviors. YouTube provides a wide platform for collecting mass opinion. People from these sites can be used to make various predictions. Movie success can be predicted using the knowledge from these sites. As it is said, more the movie is talked about, more money it will make.

Sentiment Analysis can be done on data extracted from social media, the results of which can be used to determine whether the people are in favor of or against a particular thing. Such knowledge proves usefull in predicting the success of a movie. Audience can also decide a movie that they will watch based on the prediction given using YouTube data. But processing YouTube data is difficult because of its ungrammatical structure.

II. Related Work

Lakkaraju, Praveen, Susan Gauch, and Mirco Speretta The literature related to computing similarity between data objects can be divided into two broad categories. The first uses the semantic information included in the data objects whereas the second makes use of the extraneous information about the data objects such as the concepts to which they belong and the structural context in which they occur. Since we make use of the concepts associated with the documents,but the concepts are automatically associated using the document contents, we are a hybrid of these two approaches. Thus, we review some of the related work in both these categories and some research projects that have made use of the Tree Edit Distance algorithm to compute similarity. Our baseline algorithm uses the vector space model wherein each document is represented as a vector of words and their associated tf*idf weights. The similarity between two documents is computed by calculating the cosine similarity between document vectors.

DOI10.17148/IJARCCE.2017.6453

Abdulhussain, Maysa I., and John Q. Gan.(CEEC) This paper proposes a PCA method based on cosine similarity between pairs of feature vectors. Experiments were conducted with performance evaluated by comparison with PCA based on covariance and correlation matrix. This paper describes the basic principles of PCA, correlation, cosine similarity, and the proposed approach. Experimental results have demonstrated the advantages and usefulness of the proposed method in text classification in high-dimensional feature space, in terms of the number of features required to achieve the best classification accuracy. The proposed approach in this paper employs the cosine similarity and correlation defined to compute the similarity and correlation between each two vectors of features in the training dataset, generating the correlation matrix and similarity matrix of the training data, which are used to replace the covariance matrix in the standard PCA algorithm.

Sohangir, Sahar, and Dingding Wang.Semantic Computing (ICSC), This paper proposed a sqrt-cosine method for similarity measur**e.** Finding an effective and efficient way to calculate text similarity is a critical problem in text mining and information retrieval. One of the most popular simialrity measures is cosine similarity, which is based on Euclidean distance. It has been shown useful in many applications. However, cosine similarity has its limit and not ideal due to the Euclidean distance nature. In this paper, we propose a new similarity measurement technique called improved sqrt-cosine (ISC) similarity which is based on Hellinger distance. We compare the performance of ISC with other most popular and most effective techniques for measuring text similarities in various document understanding tasks. Through comprehensive experiments, we observe that ISC performs favorably when compared to other similarity measures.

Shen, Yung-Chi, et al, This paper proposed a cosine similarity algorithm and ORCLUS algorithm for clustering. After document collection, the obtained texts were preprocessed, such as unified synonyms, removing stopwords, and stemming. After preprocessing text, 3,031 and 6,217 words from patent and media report corpuses, respectively, were obtained. To compare the two different data sources, the ORCLUS algorithm is applied to cluster patents and Internet report documents. Cosine similarity is then used to determine the semantic similarity.

Bilenko, Mikhail, and Raymond J. Mooney, In this paper, we present a framework for improving duplicate detection using trainable measures of textual similarity. We propose to employ learnable text distance functions for each database field, and show that such measures are capable of adapting to the specific notion of similarity that is appropriate for the field's domain. We present two learnable text similarity measures suitable for this task: an extended variant of learnable string edit distance, and a novel vector-space based measure that employs a Support Vector Machine (SVM) for training. Experimental results on a range of datasets show that our framework can improve duplicate detection accuracy over traditional techniques.

We present in the following table some prominent works in short along with our observations and inferences.

No.	Title Paper	Author Name	Claims by Author	Observations and Inferences
1	Document Similarity based on Concept Tree Distance	Lakkaraju, Praveen, Susan Gauch, and Mirco Speretta,2008.	Tree Edit Distance algorithm is proposed to compute similarity.	The similarity between two documents is computed by calculating the cosine similarity between document vectors. The results of our user study showed that the concept tree similarity measure with a propagation factor of 0.25 outperformed the cosine similarity measure by 15- 43%.
2	An Experimental Investigation on PCA Based on Cosine Similarity and Correlation for Text Feature Dimensionality Reduction	Abdulhussain, Maysa I., and John Q. Gan.(CEEC), 2015 7th. IEEE, 2015.	This paper proposes a PCA method based on cosine similarity between pairs of feature vectors.	This paper proposes PCA based on similarity/correlation criteria instead of covariance to gain low-dimensional features with high performance in text classification.
3	Document Understanding Using Improved Sqrt-Cosine Similarity	Sohangir, Sahar, and Dingding Wang.Semantic Computing (ICSC), 2017 IEEE	This paper proposes a sqrt-cosine similarity method.	We proposed a new similarity measure sqrt-cosine based on Hellinger distance. In this paper, we improve the sqrt-cosine similarity to fix its flaws as a similarity measure.

DOI10.17148/IJARCCE.2017.6453

4	A Cross-Database Comparison to Discover Potential Product Opportunities Using Text Mining and Cosine Similarity	Shen, Yung-Chi, et al,2017.	This paper proposed a cosine similarity algorithm and ORCLUS algorithm for clustering.	After document collection, the obtained texts were preprocessed, such as unified synonyms, removing stopwords, and stemming.To compare the two different data sources, the ORCLUS algorithm is applied to cluster patents and Internet report documents. Cosine similarity is then used to determine the semantic similarity.
5	Adaptive Duplicate Detection Using Learnable String Similarity Measures	Bilenko, Mikhail, and Raymond J. Mooney, ACM, 2003.	We propose to employ learnable text distance functions for each database field.	An adaptive approach that learns to identify duplicate records for a specific domain has clear advantages over static methods.

III. CONCLUSION

This paper represents an extension of the research on the influence of online communities on the success of movies. The wide spread use of online reviews as a way of conveying views and comments has provided a unique opportunity to understand the general public’s sentiments and derive business intelligence. In this paper, we have explored the predictive power of reviews using the movie domain as a case study, and studied the problem of predicting sales performance using sentiment information mined from reviews. Also, we focus mainly on big data processing and predicting the popularity of movies with the help of big data available on site like youTube.

ACKNOWLEDGMENT

We would like to express our sincere gratitude towards our guide Prof. D. D. Gatade for her invaluable guidance and supervision that helped us in our research. She has always encouraged us to explore new concepts and pursue newer research problems. I credit our project contribution to her. Collectively, we would also like to thank our project committee members Prof M. P. Wankhade and Prof. T. P. Vaidya for their time, suggestions, and for graciously agreeing to be on our committee, and always making themselves available. We cannot thank them enough.

REFERENCES

[1] Lakkaraju, Praveen, Susan Gauch, and Mirco Speretta. "Document similarity based on concept tree distance." Proceedings of the nineteenth ACM conference on Hypertext and hypermedia. ACM, 2008.

[2] Abdulhussain, Maysa I., and John Q. Gan. "An experimental Investigation on PCA based on cosine similarity and correlation for text feature dimensionality reduction." Computer Science and Electronic Engineering Conference CEEC), 2015 7th. IEEE, 2015.

[3] Sohangir, Sahar, and Dingding Wang. "Document Understanding Using Improved Sqrt-Cosine Similarity." Semantic Computing (ICSC), 2017 IEEE 11th International Conference on. IEEE, 2017.

[4] Shen, Yung-Chi, et al. "A Cross-Database Comparison to Discover Potential Product Opportunities Using Text Mining and Cosine Similarity." (2017).

[5] Bilenko, Mikhail, and Raymond J. Mooney. "Adaptive duplicate detection using learnable string similarity measures” Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2003.

DOI10.17148/IJARCCE.2017.6453