International Journal of Electronics Engineering (ISSN: 0973-7383
Volume 11 • Issue 2 pp. 553-558 June 2019-Dec 2019 www.csjournals.com

TF-IDF in Dogri Text: An Analysis

Shubhnandan S. Jamwal
Department of Computer Science and IT, University of Jammu, Jammu
jamwalsnj@gmail.com

Abstract

Term Frequency-Inverse Document Frequency (TF-IDF) has been one of the most highly used information retrieval methods for many years and now TF-IDF is also used in Integrated Digital Assistant (IDA). IDA works to minimize the interaction between user and system. The system will be able to find out information from the outside that is needed by users by searching users' topics through email and social media data. In this paper experiments are carried to measure the performance of the TF-IDF in Dogri documents and observed that TF-IDF is required to be combined with other NLP techniques or replaced by more sophisticated approaches like word embeddings.

Keywords

Term Frequency, Inverse Document Frequency, TF-IDF, Indian languages, NLP, Dogri

Introduction

Term Frequency - Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the relevance of a word to a document in a corpus. TF-IDF is widely used in information retrieval, text mining, and natural language processing tasks to weigh the importance of words in large text datasets. TF-IDF algorithm is also widely used in text feature extraction, in which IDF value demonstrates the importance of a term. When TF-IDF is applied to the processing of web news, the traditional IDF sometimes doesn't work well, especially in a collection divided according to channels. Although there are several variants of TF-IDF optimizing for solving various problems, very few of them considered the properties of the query terms themselves.

L. H. Pramono, A. S. Rohman and D. H. Hindersah [1] conducted exploratory studies on Integrated Digital Assistant when searching and extracting user interest or topics in social media and email data using TF∗IDF weighting modification algorithm named TF∗IDF∗DF. It is also observed that the number of terms that has a value of document frequency more than one increases. On the other hand the computational load is also increasing due to the multiplier factor of df. News taken based on the extracted topic using the TF∗IDF∗DF increased and more diverse. The term from topic extraction result still have noisy text that not appropriate to grammar writing and need to be fixed, so the term that found will be more perfect. Information Retrieval System (IRS) is a software system designed to collect, manage, process, and retrieve information from large datasets or document collections. These systems are fundamental in search engines, digital libraries, and any platform that requires retrieving relevant information based on a user query. An IRS is capable of storage, retrieval, and maintenance of Information. In this context Information can be composed of text (including numeric and date data), images, audio, video and other multi-media objects. In the management of the IRS, TF-IDF is a statistical measure used to evaluate how important a word is to a document. There exist various models for weighting terms of corpus documents and query terms.

Challenges for TF-IDF in Indian Languages

When TF-IDF is applied for Indian languages its presents unique challenges due to linguistic diversity, script complexity, and the characteristics of these languages. Here are some key issues:

- Diversity of Scripts

India has multiple languages with different scripts (e.g., Devanagari for Hindi, Bengali, Tamil script, etc.), making it challenging to develop a unified approach for tokenization and preprocessing. Languages with the same script may also have unique letters and sounds, adding to the complexity.

- Rich Morphology and Inflection

Many Indian languages are morphologically rich, meaning words change forms based on gender, number, tense, and other grammatical features. For example, in Hindi, nouns and verbs can have multiple forms. Similarly the Dogri language is also highly inflectional and very rich in morphology. Moreover the development of the stemmers and lemmatizers for Indian languages are often less developed compared to English.

- Compound Words and Agglutination

Indian languages often use compound words or agglutinated forms, where multiple words are combined into a single one. Tokenizing compound words correctly is crucial for TF-IDF, but segmenting them is challenging due to limited NLP tools in Indian languages.

- Lack of Quality Corpora

To compute meaningful IDF scores, a large corpus of documents is needed. However, high-quality, annotated datasets in Indian languages are often scarce, limiting the accuracy of TF-IDF scores.

- Synonyms and Polysemy

Indian languages have high synonymity and polysemy, where words have multiple meanings or multiple words exist for the same concept. This can affect the weighting of words in TF-IDF. Without advanced disambiguation tools, TF-IDF may not capture the true relevance of terms in documents.

- Encoding and Font Compatibility Issues

Older documents or datasets may use incompatible encoding or fonts, making it hard to process data uniformly. Unicode adoption has improved this, but it’s still an issue in some corpora.

Literature Review

Addressing the challenges of TF-IDF may require preprocessing pipelines tailored for specific languages, access to high-quality datasets, and possibly combining TF-IDF with other NLP techniques for more robust text representations. M. Xu, L. He and X. Lin [2] proposed a refined IDF schema named Channel Distribution Information (CDI) IDF, which is based on the information among the IDF values of each channel collections. According to the statistical features, the Top terms and the meaningless terms could be identified. They conducted experiments on a manual labeled test set which indicated that, related to the traditional TF-IDF, the CDI TF-IDF increases the Recall, Precise and F0.5 measure by 2.71%, 3.07% and 3.00%. R. Xu [3] observed that when people type out a query, usually the verbs and the nouns are the primary keywords that directly define the query. The adjectives and adverbs are generally the secondary keywords, which describe the query more accurately. Other terms might not be as important as the terms just mentioned and could be the tertiary keywords. Based on this fact, they proposed an algorithm named POS Weighted TF-IDF algorithm. The proposed algorithm takes every query term's part of speech (POS) into account and assigns each query term frequency a different weight value according to the POS of that term. A. Mishra and S. Vishwakarma [4] carried out work to analyze and evaluate the retrieval effectiveness of vector-space model while using the new data set of FIRE 2011. They conducted experiments with TF-IDF and its variants. For all experiments and evaluation the open search engine, Terrier 3.5 was used and observed that TF-IDF model gives the highest precision values with the new corpus dataset. Q. Liu, J. Wang, D. Zhang, Y. Yang and N. Wang [5] observed that separating words with the same or similar meanings will result in the loss of partial information when text feature were extracted. The representation of words needs to extract the similarity of words, and the similarity among words needs to be obtained by the meaning of words in texts. In order to improve the accuracy of text feature extraction, they used word2vec model to train the word vector in the corpus to obtain its semantic features. After excluding words with low TF-IDF value, the density clustering algorithm is used to cluster the remaining words according to word vector similarity. As a result, similar words are clustered together and can be represented to each other. Experiments show that using the TF-IDF algorithm again, constructing a VSM (vector space model) with these clusters as feature units can effectively improve the accuracy of text feature extraction. Y. Wang, D. Zhang, Y. Yuan, Q. Liu and Y. Yang [6] observed that the solution proposed by many scholars only solves the problems of distribution ratio and does not solve the problem that the domain keywords have unreasonable weights. The problem has led to the use of domain-specific applications where relevant keywords in some areas have not been given appropriate weights. They proposed an improved method based on domain knowledge graph which mainly consider the application of the legal field, and use the legal knowledge graph to make improvements to the TF-IDF algorithm, so as to achieve the reasonable weight assigned to the domain-related keywords in text feature extraction. Experiments show that this method can effectively improving the accuracy of the extraction. S. Bahri, S. Sumpeno and S. M. S. Nugroho [7] proposed hybrid model that combines TF*IDF and LSI to fix some limitations of both. From the experimental results, it is found that the proposed model outperform as compared to TF*IDF model and LSI model that stand alone.

Haoying Wu and Na Yuan [8] proposed an improved feature weighting algorithm FDCD-TF-IDF based on word frequency distribution information and category distribution information. The improved algorithm introduces the concept of word frequency distribution and class distribution to describe the weight of the feature item more accurately. The word frequency distribution is mainly aimed at the correlation between feature items and categories, and the category distribution can better reflect category information of feature items. The improved algorithm can accurately reflect the differences between different text categories. The experimental results show that the improved algorithm can achieve better classification results on both balanced and unbalanced text data sets. Ted Tao Yuan and Zezhong Zhang [9] discussed keyword expansion candidate selection using word embedding similarity, and an enhanced tf-idf formula for expanded words in search ranking. El Barakaz Fatima and El Moutaouakkil Abdelmajid [10] proposed a new approach of unstructured text classification using quantitative variable; this variable has a strong correlation with the text attribute. They classified the text attribute using the bags of words to get the keywords of our corpus, and defined classes to which we wish affecting terms, then they apply the naive Bayes classifier which is not highly efficient in an unstructured text classification case. So they proposed a new classifier combining two algorithms principles: term frequency- inverse document frequency (TF-IDF) and k-nearest neighbor (KNN). The objective is to achieve better classification of an unstructured text with a high level of efficiency. The result of the proposed classifier method was very satisfactory, especially since it enriches the dictionary content each time we use it. Vaishali Ingle and Sachin Deshmukh [11] analysed works on online news data for prediction of stock market states such as high, low etc. The Hidden Markov Model along with features extracted such as TF-IDF is used to find out next day's stock market value for group of companies. The method can be further extended to adjustment of probability values to calculate tuned model for prediction. Jiaul H. Paik [12] proposed a novel TF-IDF term weighting scheme that employs term frequency normalizations to capture two different aspects of term saliency. One component of the term frequency is effective for short queries, while the other performs better on long queries. The final weight is then measured by taking a weighted combination of these components, which is determined on the basis of the length of the corresponding query. Experiments were conducted on a large number of TREC news and web collections demonstrate that the proposed scheme almost always outperforms five state of the art retrieval models with remarkable significance and consistency. The experimental results also show that the proposed model achieves significantly better precision than the existing models.

TF-IDF in Dogri Language

The textual information present in the digital world cannot be processed and analyzed manually. The identification of the keywords and phrases are very useful for quickly and efficiently evaluating massive volumes of information. In NLP tasks the keyword extraction is widely used process in information extraction from a document and TF-IDF is used many a times for the categorization of the document on the basis of the keywords. Using TF-IDF is still a new method for Indian regional languages. The experiments are conducted on Dogri text documents for extractive summarization using the TF-IDF approach. The experiments are conducted on the news data items of Dogri with respect to political keywords and items. The political keywords included in the category are names of the local leaders, names of the political parties etc.

	Precision	Recall	F-measure
TF	68.34	68.45	68.40
IDF	73.04	73.10	73.90

Observation and Conclusion

TF-IDF is a popular algorithm used in text processing and natural language processing for evaluating the importance of a word in a document relative to a collection of documents. However, TF-IDF treats words independently, ignoring semantic relationships. TF-IDF considers the frequency of terms without regard for the order or position of words in the document. As a result, it cannot capture the syntactic structure of the text, which can be critical for understanding meaning. TF - IDF method is implemented with respect in Dogri language for the evaluation of the keywords and it is observed that the position information of keyword, word class and the balance of term frequency algorithm, are required to be considered for the improvement of TF-IDF method. In shorter texts, TF-IDF may be less effective because there is limited context, making the calculation of meaningful frequencies challenging. This limits its application in domains like social media or micro-blogging. TF-IDF is required to be combined with other NLP techniques or replaced by more sophisticated approaches like word embeddings (e.g., Word2Vec, GloVe) or transformer models that can better capture semantic meaning and context.

References

[1] L. H. Pramono, A. S. Rohman and D. H. Hindersah, "Modified weighting method in TF*IDF algorithm for extracting user topic based on email and social media in Integrated Digital Assistant," 2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T), Bandung, Indonesia, 2013, pp. 1-6, doi: 10.1109/rICT-ICeVT.2013.6741547.

[2] M. Xu, L. He and X. Lin, "A Refined TF-IDF Algorithm Based on Channel Distribution Information for Web News Feature Extraction," 2010 Second International Workshop on Education Technology and Computer Science, Wuhan, China, 2010, pp. 15-19, doi: 10.1109/ETCS.2010.130.

[3] R. Xu, "POS weighted TF-IDF algorithm and its application for an MOOC search engine," 2014 International Conference on Audio, Language and Image Processing, Shanghai, China, 2014, pp. 868-873, doi: 10.1109/ICALIP.2014.7009919.

[4] A. Mishra and S. Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 2015, pp. 772-776, doi: 10.1109/CICN.2015.157.

[5] Q. Liu, J. Wang, D. Zhang, Y. Yang and N. Wang, "Text Features Extraction based on TF-IDF Associating Semantic," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 2018, pp. 2338-2343, doi: 10.1109/CompComm.2018.8780663.

[6] Y. Wang, D. Zhang, Y. Yuan, Q. Liu and Y. Yang, "Improvement of TF-IDF Algorithm Based on Knowledge Graph," 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA), Kunming, China, 2018, pp. 19-24, doi: 10.1109/SERA.2018.8477196.

[7] S. Bahri, S. Sumpeno and S. M. S. Nugroho, "An Information Retrieval Approach to Finding Similar Questions in Question-Answering of Indonesian Government e-Procurement Services using TF*IDF and LSI Model," 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), Bali, Indonesia, 2018, pp. 626-631, doi: 10.1109/ICITEED.2018.8534856.

[8] Haoying Wu and Na Yuan. 2018. An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information. In Proceedings of the 3rd International Conference on Intelligent Information Processing (ICIIP '18). Association for Computing Machinery, New York, NY, USA, 211–215. https://doi.org/10.1145/3232116.3232152

[9] Ted Tao Yuan and Zezhong Zhang 2018. Merchandise Recommendation for Retail Events with Word Embedding Weighted Tf-idf and Dynamic Query Expansion. In The 41st International ACM SIGIR Conference on Research &amp; Development in Information

Retrieval (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 1347–1348. https://doi.org/10.1145/3209978.3210202

[10] El Barakaz Fatima and El Moutaouakkil Abdelmajid. 2017. A new approach to text classification based on naïve Bayes and modified TF-IDF algorithms. In Proceedings of the Mediterranean Symposium on Smart City Application (SCAMS '17). Association for Computing Machinery, New York, NY, USA, Article 24, 1–5. https://doi.org/10.1145/3175628.3175643

[11] Vaishali Ingle and Sachin Deshmukh. 2016. Hidden Markov Model Implementation for Prediction of Stock Prices with TF-IDF features. In Proceedings of the International Conference on Advances in Information Communication Technology &amp; Computing (AICTC '16). Association for Computing Machinery, New York, NY, USA, Article 9, 1–6. https://doi.org/10.1145/2979779.2979788

[12] Jiaul H. Paik. 2013. A novel TF-IDF weighting scheme for effective ranking. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13). Association for Computing Machinery, New York, NY, USA, 343–352. https://doi.org/10.1145/2484028.2484070