International Journal of Control and Automation
Vol. 13, No.2, (2020), pp. 1004 - 1010

A Model for Document Retrieval Using Earth Mover Distance

Akash Bhattacharyya[1], Siddharth S. Rautaray[2], Manjusha Pandey[3]

[1, 2, 3] *School of Computer Engineering, KIIT, Deemed To Be University, Bhubaneswar, India*

[1]*akash210994@gmail.com*, [2]*siddharthfcs@kiit.ac.in*, [3]*manjushafcs@kiit.ac.in*

Abstract

The recent rise in the amount of unstructured data in digital format has given rise to the use of natural language processing techniques to understand the data. Organizations have started recognizing the potential in the unstructured textual data. Data from the internet as well as from the organizations' internal repository can help them to gain more insight into the market. Information collected from such sources provides valuable decision-making probability for the organizations. Keyword-based search is the most helpful form of the document retrieval process. The paper discusses such processes used in the current scenario as well as propose a new form of technique to be used for the process of document retrieval.

Keywords: *Text Mining, Natural Language Processing, Word Movers Distance, Document Retrieval*

1. Introduction

The process of representation of the distance between two documents has been a far-reaching research application in the information retrieval domain. Processes such as categorization of news, identification of songs and matching of multilingual document matching require the use of information retrieval techniques [1]. The most common way a document is represented is the use of a bag of words notation and term frequency-inverse document frequency (TF-IDF) factor. However, it has been found that these factors are not suitable for document distance due to their near orthogonal [2]. Another significant disadvantage of these processes is the unavailability of distances among individual words. In the context of natural language processing, there needs to be the presence of distance between sentences and words rather than only documents [3]. Measuring the similarity between two words is much easier than comparing two words.

2. Research Objectives

The process of document retrieval has been used in various forms and across various fields of research. However, the research pertains to the process of using primitive methods of the TF-IDF process along with simple classification algorithms. The main aim of this research is to develop a framework for the process of extraction of related documents based on keywords. The objectives for the completion of this research are:

- To perform pre-processing on the text files
- To perform entity extraction on the text
- To return relevant documents based on Word Movers Distance
- To determine better of the two algorithms: proposed method and TF-IDF method.

3. Distance calculation metrics

3.1. TF-IDF

ISSN: 2005-4297 IJCA

The TF-IDF is a weightage calculated based on the importance of a word in a document present in a collection of documents. The words counted are taken as normalized value to prevent biasness in longer documents which helps in giving importance to documents better [4]. Thus, the term frequent is given by

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad ..(1)$$

In the above formula ni, j is the occurrence of a particular term ti in a particular document dj and the denominator is the total summation of the number of words in the document considered. The advanced section of inverse document frequency is the measure of the importance of the term [3]. This is calculated by taking the logarithmic value of the ratio between the total numbers of documents to the total number of documents consisting of the term.

$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|} \quad ..(2)$$

To have a proper non-zero value for the inverse document frequency the denominator can be considered as $1 + |\{d: t_i \in d\}|$. Thus, the final weightage is given as $(tf - idf)_{i,j} = tf_{i,j} \times idf_i$.

A high weighted frequency can be achieved when there are a high term frequency and a low inverse document frequency for the whole collection of documents [5]. This formulation thus allows the removal of most common terms. It has been seen that the value of TF-IDF is more than or equal to zero.

The basic form of any similarity measure is the use of Euclidean distance. Euclidean distance can only be taken into consideration if the data is measured on a similar scale. A distance between two vectors can be given as the square root of the summation formed from the difference of the elements of each of the two vectors. The mathematical form is as follows:

$$d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad ..(3)$$

3.2. Earth Movers distance

Another form of distance measurement between documents can be calculated based on words stored in a lexical database [6]. This form of distance calculation is known as earthmovers distance (EMD). The database stores the semantic distances between the words. The distances are collected and the similarity between the documents is calculated [7]. The distance is calculated as many-to-many matching as a collection of words from one document can have a similar meaning to a single word in another document. The earthmovers distance helps in the process of solving transportation problems. For example, let us consider m as a set of suppliers and n as a set of warehouses. The transportation cost needs to be minimum and would be able to transport all goods from m to n [7]. Taking into account the various constraints as follows:

- The transport can be done only unidirectional (m to n).
- Total cargos leaving cannot exceed the capacity of m
- Total cargos received cannot exceed the capacity of n

- The maximum number of transportation allowed is the minimum between the cargos in m to that of in n.

Overall, the following equations can be stated:

$$f_{i,j} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n \quad ..(4)$$

$$\sum_{j=1}^{n} f_{i,j} \leq w_{pi}, 1 \leq i \leq m \quad ..(5)$$

$$\sum_{i=1}^{n} f_{i,j} \leq w_{qj}, 1 \leq j \leq n \quad ..(6)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = min\left\{\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj}\right\} \quad ..(7)$$

From the above equation, the symbols are as follows:

- P is the set of origin
- Q is the set of destination
- F(I,j) is the flow from node i to node j
- M and n are the number of origin and destination respectively
- W(I,j) is the number of transport from node i to node j

Taking into consideration the optimal flow F. the linear formulation can be written as:

$$\text{EMD(P, Q)} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}\, d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}} \quad ..(8)$$

3.3. Word mover's distance

Word movers distance makes use of word embedding's to calculate the distance even though there is no common word between them [6]. Based on the study from [8] the word centroid distance can be defined as

$$\sum_{i,j=1}^{n} T_{ij}\, c(i,j) = \sum_{i,j=1}^{n} T_{ij} \left\| x_i - x'_j \right\|_2 \quad (9)$$

$$= \sum_{i,j=1}^{n} \left\| T_{ij}(x_i - x'_j) \right\|_2 \geq \left\| \sum_{i,j=1}^{n} T_{ij}(x_i - x'_j) \right\|_2 \quad (10)$$

$$= \left\| \sum_{i=1}^{n} \left(\sum_{j=1}^{n} T_{ij} \right) x_i - \sum_{j=1}^{n} \left(\sum_{i=1}^{n} T_{ij} \right) x'_j \right\|_2 \quad (11)$$

$$= \left\| \sum_{i=1}^{n} d_i\, x_i - \sum_{j=1}^{n} d'_i\, x'_i \right\|_2 = \left\| Xd - Xd' \right\|_2 \quad (12)$$

This word centroid distance can be represented with the help of a weighted average of the word vector of a document. The word mover's distance is a special case of earthmovers distance [15]. A tight bound word movers distance can be obtained with the help of word centroid distance by removing one of the constraints:

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{ij}\, c(i,j) \;..(13)$$

$$subject\ to: \sum_{j=1}^{n} T_{ij} = d_i \,\forall \in \{1, \ldots., n) \;..(14)$$

Thus the optimal solution required for the movement of a word in document d to its most similar form in d' would be given as matrix as follows:

$$T_{ij}^{*} = \begin{cases} d_i \; if \; j = argmin_j c(i,j) \\ 0 \; otherwise \end{cases} \;..(15)$$

The major assumption made is that similar words have similar vectors. Some of the intriguing properties of word mover's distance are:

- It is easy to understand and hyper-parameter free.
- It is easy to estimate as the distance among the documents can be broken down and plotted concerning a sparse distance among the individual words of the documents

It makes use of word2vec knowledge data which leads to high accuracy in the processes.

3.4. Entity extraction

Entity extraction is a form of information retrieval technique, which is used for the process of identification and classification of major key elements form a large text into some predefined categories [9]. The file formats range from documents, web pages, spreadsheets, and unstructured social media text. By understanding the various entities such as people, organizations, places and numerical expressions (time, date, currency, and phone numbers) provides a simpler way of understanding the information contained in them [6]. Entity extraction applied to semantic techniques can be used to disambiguate meaning as well as understand the context of the file thereby increasing the number of useful operations in context to various business and intelligence.

4. Related Work

There have been several research conducted on the process of enhancing the methodology to be followed for the process of retrieval of documents. In [10] the authors have discussed the use of KNN classification with the help of the TF-IDF method. The model has been executed with the help of map-reduce. They concluded that their methodology was able to produce higher precision than the traditional TF-IDF method. In [11] the authors have discussed the use of the entity extraction process to develop a model for CV parsing. The output generated is in XML format. In [12] the authors have proposed a model that would be able to compile the entity extraction process on textual data. The model would be able to provide a part of speech trainee and classify the words based on the part of speech tagging to extract meaningful content from the text. In [13] the authors discuss their model of using MapReduce to effectively cluster documents with the help of the k-Means algorithm.

In [6] the authors have explained the development of the word movers distance process

to be used in natural language processing. The system changes the words into vectors to be represented in a vector space. Similar words are closer to each other based on the vector values [14]. Distance between words in the vectors provides the similarity between two sentences. The authors have compared their proposed model with various primitive models and have been able to prove that their model is more effective than other methods.

5. Proposed framework

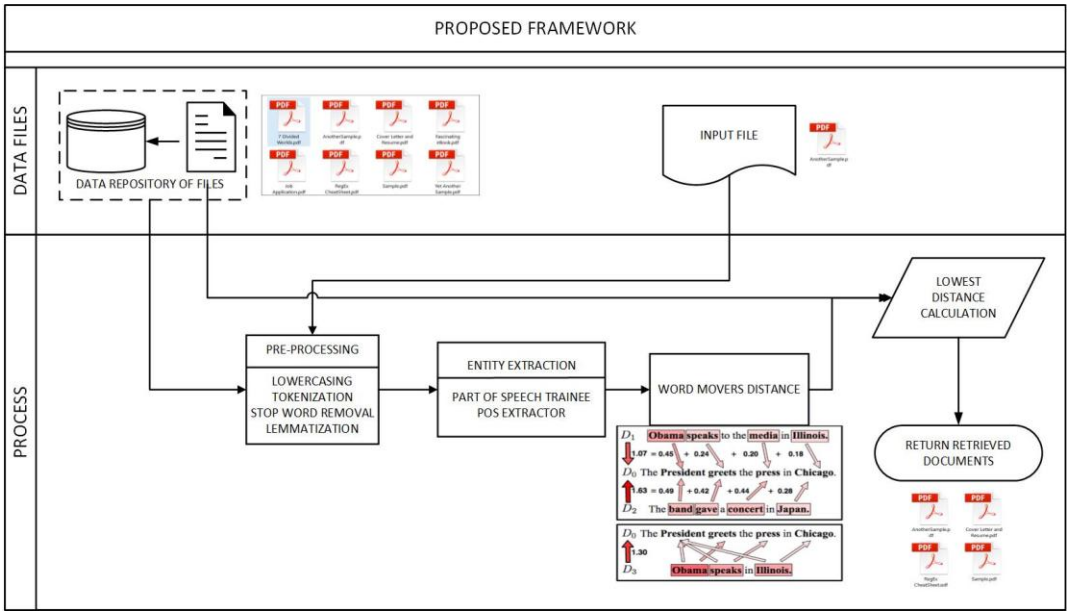

Figure 1: Proposed framework for the model

The data repository consists of various text files in portable document format (pdf). A file location is given as input, which is studied, and top 5 most used words are selected as a list of keywords. The list of keywords would be used to find out relevant documents from the repository.

Pre-processing of the text is done with the help of multiple processes. Firstly, the whole text is converted into lowercase, which helps in the easier interpretation of the words. Mixed case words are harder to differentiate. Two words (Canada and canada) would have the same meaning but would be treated differently because of their case sensitivity. The next process is to tokenize the sentences into individual words. This would help in mapping each word of the text separately. After tokenization has been done, the stop-words in the text would be removed. These words would hamper the mapping of the important words on the vector space. A simple task of lemmatization or stemming is performed as the final step of pre-processing. In this process, the words are changed into their parent or root word so that words with similar root can be mapped together.

The process of using the entity extraction technique would be to remove disambiguation in the words of the documents. This would help in returning documents containing words with similar meaning back to the user rather than dissimilar word meaning. Disambiguation refers to the problem where the same word can have multiple meanings depending on the position it is used. The part of the speech trainee would be used to determine how the words are used in the context of the text.

Word movers distance would be used then to determine the distance of vector representation of the words concerning the sentence in a given document. This would help in understanding the similarity of the words of the document and the words taken as input. The distance calculated would then be used to determine the similarity between the input

text and the document in the repository.

The algorithm does not require any model building methodology or tuning of different parameters. The algorithm is versatile and can be used easily. The lowest distance calculated is the closest to the keywords, i.e. it is most similar to the document given as input. Thus, the documents can then be collected and returned to the user.

6. Result and analysis

Based on the above proposed framework a working program was made where the list of most relatable documents would be returned. The framework takes a pdf file location as input and the repository consists of 188 pdf documents. The program takes 1490 seconds to complete the process of reading the documents and calculating the word movers distance between the input file and the files in the repository. The distances are stored in a CSV data set for easier access. The lowest distance shows the closest related files to the query file. The smallest distance files of 10 distances are selected and displayed.

The existing framework was also designed on the same repository of data available. The files were read and then stored in a CSV data set. This data set was then used to calculate the TF-IDF factor of the files with respect to a predefined file. The data was then used in KNN classification algorithm to find the nearest neighbor to the query file. This algorithm took around 1140 seconds. The neighbors of the file given as input is displayed.

Both the algorithms had been tested with the help of same input file and took files from the same repository. However the runtime of WMD process took longer due to the fact that calculation of the distance vector based on the key words was time consuming. Moreover there was a requirement of preloading a large tool which helps in the calculation of the vector distance among various words. The GoogleNews-vectors-negative300 is a word2vec tool designed and shared by Google. The file is around 3.4 GB in size and needs to be loaded before the program can be executed. Looking at the time it can be said that the TF-IDF method is much faster than the WMD method (1140s<1490s), but the space required for the two program is different. The WMD method creates the data set of size 4 KB, but the TF-IDF method creates a data set of size 5.91 MB. Though this is small in size compared to the modern large scale availability of space, the WMD method can be completed without the use of any CSV files, but with the help of simple data frames.

7. Conclusion

In this research, we have proposed and executed a framework model, which can be used to retrieve documents based on word mover's distance. The proposed model would be able to achieve much higher accuracy as well as faster execution time and low memory consumption. The use of word mover's distance is helpful in simple word corpora. The research already conducted has been done with the help of the TF-IDF methodology. The latest research about the development of the word mover's distance has been low. Thus, the research would be able to help in the speed-up of the information retrieval process. The next phase of the research would be to design the framework and to be able to work on a larger amount of data and to make the working of the framework faster.

References

[1] M. S. Kheerthana, "Personalized Document Retrieval Using Text Mining," vol. 3, pp. 307–310, 2017.

[2] R. Naidu, S. K. Bharti, K. S. Babu, and R. K. Mohapatra, "Text summarization with automatic keyword extraction in telugu e-newspapers," Smart Innov. Syst. Technol., vol. 77, pp. 555–564, 2018.

[3] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016, pp. 61–66, 2016.

[4] S. Babu and P. G. Scholar, "KNN TFIDF Based Named Entity Recognition 1," vol. 1, no. 12, pp. 35–39, 2020.

[5] M. El Mohadab, B. Bouikhalene, and S. Safi, "Automatic CV processing for scientific research using data mining algorithm," J. King Saud Univ. - Comput. Inf. Sci., 2018.

[6] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From Word Embeddings To Document Distances," in International Conference on Machine Learning, 2015, pp. 957–966.

[7] Y. Rubner, C. Tomasi, Y. Rubner, and C. Tomasi, "The Earth Mover's Distance," Percept. Metrics Image Database Navig., pp. 13–28, 2001.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, "Metric for distributions with applications to image databases," Proc. IEEE Int. Conf. Comput. Vis., pp. 59–66, 1998.

[9] A. Ittoo, L. M. Nguyen, and A. Van Den Bosch, "Text analytics in industry: Challenges, desiderata and trends," Comput. Ind., vol. 78, pp. 96–107, 2016.

[10] Y. Zhao, Y. Qian, and C. Li, "Improved KNN text classification algorithm with MapReduce implementation," 2017 4th Int. Conf. Syst. Informatics, ICSAI 2017, vol. 2018-Janua, no. Icsai, pp. 1417–1422, 2017.

[11] P. Das, M. Pandey, and S. S. Rautaray, "A CV Parser Model using Entity Extraction Process and Big Data Tools," Int. J. Inf. Technol. Comput. Sci., vol. 10, no. 9, pp. 21–31, 2018.

[12] A. Dash, M. Pandey, and S. Rautaray, "Enhanced Entity Extraction Using Big Data Mechanics," in International Conference on Advanced Computing Networking and Informatics, 2019, pp. 57–67.

[13] T. H. Sardar and Z. Ansari, "An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm," Futur. Comput. Informatics J., vol. 3, no. 2, pp. 200–209, 2018.

[14] X. Chen, L. Bai, D. Wang, and J. Shi, "Improve Word Mover's Distance with Part-of-Speech Tagging," Proc. - Int. Conf. Pattern Recognit., vol. 2018-Augus, pp. 3722–3728, 2018.

[15] S. Marinai, B. Miotti, and G. Soda, "Using earth mover's distance in the bag-of-visual-words model for mathematical symbol retrieval," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 1309–1313, 2011.

Authors

Akash Bhattacharyya, is currently pursuing as M.Tech postgraduate student in School of Computer Engineering under KIIT Deemed-to-be University. His areas of research interest include Big Data Analytics, Natural Language Processing, and Big Data Application Domains. He is currently working in document retrieval technique based on document-specific term frequencies. He can be reached at akash210994@gmail.com.

Siddharth Swarup Rautaray, PhD (Computer Science), Member IEEE is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. He has more than a decade of teaching and research experience. Dr Rautaray has published numbers of Research Papers in peer-reviewed International Journals and Conferences. His areas of interest are Image Processing, Data analytics, Human Computer Interaction. He can be reached at siddharthfcs@kiit.ac.in.

Manjusha Pandey, PhD (Computer Science), Member IEEE is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. She has more than a decade of teaching and research experience. Dr Pandey has published numbers of Research Papers in peer-reviewed International Journals and Conferences. Her areas of interest are WSN, Data analytics. She can be reached at manjushafcs@kiit.ac.in.

ISSN: 2005-4297 IJCA