Turkish Journal of Computer and Mathematics Education Vol.12 No. 9 (2021), 3166-3171

Research Article

Genetic based method for Mining Association Rules from Text

Suad abd alwahab[1], Hadeel Jameel Nassr[2] , Zanbaq Hikmet Thanon[3] Baydaa Jaffer Al-Khafaji[4]

Ministry of Education , Iraq[1]
Ministry of Education , Iraq2
Iraqi Commission for Computers & Informatics Informatics Institute for Postgraduate Studies, Iraq[3]
Computer Science Department College of Education for Pure Science/ Ibn Al-Haitham University of Baghdad, Iraq[4]
Email :bayda.khafaji@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: In general, text mining is the process of discovering very important and unknown information from a semi-structured or unstructured data set, while data mining deals with structured data. Where natural language processing technology was used: coding, stop word deletion, derivation and indexing of the document, convert text into structured form, and in order to extract correlation rules, we used the genetic algorithm (GA) to solve the correlation rule problem by overcoming the limitations of using the priori algorithm (a coding system, storage capacity). Were the estimates are based or based on a recommendation to use the individual variable length within the population. The rules for extracted relationships contain the essential features included in a document compilation. The proposed system aims to solve the problems facing traditional techniques as it works to address two main issues: Computational complexity, where the number of rules increases greatly with the number of elements in the database. Second, the rules related to interests must be collected within the place of creation of the rules that usually suggest a higher algorithm. The use of a genetic algorithm to find a relationship in the unorganized document between keywords is powerful and very effective in solving this problem because it effectively cuts the search space for threshold based consistency measurements.

Keywords: Genetic algorithm, Text mining, Association rules, AIS, STEM, Coronavirus.

Introduction

For data mining is to discover confidential data and information and the purpose of this DM is to focus the existing set of databases, create new algorithms and publish statistical outputs to discover possible knowledge. It is used traditionally in business organizations and financial intelligence for analysts and is being used increasingly in the scientific field to extract information by applying it from huge data sets [1].

Data mining works with unstructured or textual information of massive volumes of text to retrieve information and experience. It is needed for effective analysis and exploration of the information available in textual form. The text should be translated into data and then analyzed by other statistical techniques. Most of the time the data that we get from different sources is too large to be able to read and interpret it manually, so we need (TM technology) to handle this data. [2] The two sources are assets with the English introduction. Text exploration is characterized by being a process of detecting and analyzing huge numbers of unorganized texts through the use of a program that can identify key words, concepts and other basic features in texts and as it was called text analyzes . [3]are considered mining techniques Texts are a subset of great importance in extracting texts that work to extract knowledge from semi-structured and unstructured textual data in addition to applications that work on text analysis and processing[4].

The "American Professor Don Swanson" has done research in exploring the text by combining slices of information from medical documents that appear irrelevant or unrelated as this was in the mid-1980s. Thus it was possible to obtain new hypotheses [5] as A two-dimensional method was introduced to prospecting texts through the analysis of bibliometric and networks, in order to know the academic trends of the main branches[6]. A general and clear idea was presented about the tools, applications and matters that happen to mine the text [7].The different methods of analyzing and classifying various text documents were also discussed [8].

The various techniques, applications and tools of text mining were also reviewed [9].

Text mining is the extraction of hidden, unknown information and the process of extracting this information is automatic and then linking this information with each other to find new facts and hypotheses.There are many traditional experimental methods. Text mining does not match what is known in web search, as it has the advantage of using a number of algorithms to convert unstructured texts into structured texts and using quantitative methods to analyze this information[10]..

The main objective of the data mining process is to extract information from the original text documents and to manage the processes of extraction, classification, summarization and retrieval, and that is under supervision and collection operations without the need for supervision[11].

Assouation rule and text mining

The rules for societies are defined according to the applications and disciplines. The main goal is to discover the qualitative rules that distinguish the links between the groups[12].

One of the best techniques for knowledge discovery is the search for association rules, which represent a branch of great importance in monitoring or managing information and knowledge. It works to create important relationships between information encoded in huge databases[13]. It has been used in decision-making in many smart applications, for example Intention to buy from customers[14], Analysis of commercial sales[15] Market basket analysis[16], Customs risk management [17].

Uses of association rules in data mining

Correlation rules are of great use in data mining for forecasting and analyzing customer behavior, and they also have a great role in store planning, catalog design and product grouping, and market basket analysis. The correlation rules have been used by programmers to create machine learning programs. It is one of the types of artificial intelligence that works to create more efficient programs when they are programmed in an unclear way.[20].[21].

Measures of association rules

There are two basic parameters by which to measure the strength of the rules of correlation:

1- Support: Support means the number of times a specific database appears in the database that is subject to mining.

2- Trust: The number of times a specific rule is proven to be true in practice.

A rule can appear to be strongly correlated in the data set, as it appears more often, while it occurs in a smaller number upon application. This case represents a case of high support while confidence is low. On the contrary, the rule may not appear in particular in the data set, while upon continuous implementation it appears to be a frequent occurrence. In this case the confidence is high and the support is low.

These measures enable analysts to distinguish between causality and correlation, and enable them to properly evaluate a specific rule. .[20],[21].

The third parameter of the value, which is known as the lift value, is the ratio of confidence to support.

If the lift is positive, then there is a positive correlation between the data points. If the lift is negative, then there is a negative correlation. .[22],[23].

But if the lift value is equal to 1, this means that there is no correlation between the data points[18].

Association rule algorithms

The algorithms most commonly used to use correlation rules are: SETM, AIS, Apriori, and others.

AIS algorithm: This algorithm builds groups of elements and count them during data examination. This algorithm determines the large elements present in the transaction and then finds new candidate elements by combining the large group of elements with other groups in the transaction data.

STEM algorithm: This algorithm works to find groups of candidate elements during the data examination process, as in the ACE algorithm, but it counts the groups of elements after the scanning process. And the negative drawbacks of both algorithms are that they count many of the candidate small elements according to the author of the real Time Data mining for Dr. saeid Siad.[18].

APRIORI algorithm: In this algorithm, the groups of candidate elements are found only by adopting the large elements and then linking this group to itself to build groups of elements that are one larger and then deleting the small subgroups and the remaining elements are the candidate elements. By using this method, this algorithm reduces the number of candidates by Detecting groups of elements is over-supported, according to the d.siad[18,19].

The Proposed System and the Experimental Part

Many texts will get the particular inputs to the approach of Text Mining . text records are a source of knowledge that is unstructured and unplausible for standard file techniques to be processed.Therefore, thedistillation of structured data or possibly information from unstructured text mining by defining references for namedentities in addition to defined relationships between these choices is the object of data extraction (IE).

Another important things of the system of data mining may be the KDD, which takes account of the use of record and techniques for machines learning to explore new connections in relation sources.

Both KDD and IE subjects are synonymous with extensive work conducted within the subjects in order to f ormulate novel methods for text mining and also to solve
issues that are confronted with conventional techniques alone.
Our aim of our study is to develop good approaches for the mining of possible
knowledge and interesting groups across large collections associated
with text documents for 200 research addresses in a certain object using
Genetic Algorithm (GA).

The system proposed have two phases:

1- Pre rocessing phase: The purpose of this step is to optimize the output of the
next step. These processes consist of:
A-Tokenization:
that splitting research address into words calling(tokens). Filtration:
When this is not present in the predefined stop words database, a word is selected as a keyword. The list of stopwords contains posts, pronunciations,
An example of tokenization is:
Research address : (Coronavirus Replication and Reverse Genetics)
After tokenization will be : (Coronavirus Replication Reverse Genetics
B-Stemming,
In order to save memory space and processing time, attempts to minimize everything. Not only does it confl ate variations of a word with a single descriptive type, but it also reduces the amount associated with specifi c words required to describe a collection associated with documents.
For instance (Genetics) can be stemmed to (Genetic).
C- Eliminating all keywords that appear in only one document.
1. Coronavirus, patient, pneumonia, China
2. Drag, treatment, option , Coronavirus.
3. patient, Coronavirus, SARA_COV_2, Wuhan, China.
4. scientist, Coronavirus, drug, animal, Wuhan.
5. mechanism, cell, immune, protein.
6. organization, SARS, protein
7. interaction, RNA, protein.
Since the order of the word for each research title is not important.

2- Knowledge distillation phase (Generating association rules through using GA

Constructing Itemset

In" I ={i1, i2, ...in}" we know the number and quote all the objects. , by their indexes . We may presume, in other words, that the universal object set" I ={A1, A2, . ., An}"..

We construct the collection of items in our work from the 200 research address through the set of items.

To store each candidate term, use the data structure to measure its frequency over

The total number of addresses for research and our threshold will be 50 percent because .The Word's Weight (Drag) = (number of documents in which the term is written Drug appeared) / 200 (total number of used documents) < 50%, so it is taken

As a keyword in a collection of objects.

Our itemset={ Patient, Wuhan, SARS, China, drug, protein, genetic, treatment, Knowledge, interaction, epidemic, evolution, system, immune, animal, disease, cells, …….}.

We discard a word which is its frequent=1/200 in research title and others which appear in more than half in research also The set of keywords will increase according to the number of researches.

Individual Representation

Given an association" k-rule $X \rightarrow Y$, where X, $Y \subset I$, and $X \cap Y = 0$", we encode it into an individual as:

J	*A*1	...	*Aj*	*Aj*+1	…..	*Ak*

Where j is an indicator that separates the antecedent from the consequent of the rule. That is," X={A1, . . . , Aj } and Y={Aj+1, . . . , Ak}; 0 < j < k. Therefore, a k-rule X $\rightarrow\rightarrow$ Y is represented by k+1 positive integers".

One of rules looks like :

{disease, cell} $\rightarrow$ {immune,drug}

Genetic Algorithm for Association Rules:

Constructing Initial Population:

1. Specify the chromosome length at random.
2. Indicate the number of objects in the context.
3. Choosing the words randomly from the set of items in the created chromosome, since each word occurs only once.
4. If a chromosome is redundant, discard it.

Example :

Length of chromosome chosen : 4
Length of an antecedent : 3
The chosen words from the set of all words are :

{ drug ,system ,immune, SARS } is a chromosome in an initial population and will represent the rule :

immune, drug ——— system, SARS .

An instance of such generated rules in the first population are :

epidemic, drug ——— treatment
human, infection ——— MERS
COV_19 ——— Protein, disease
Cell, China ——— Pathogene, structure, animal

1- Selection operator:
ranked selection method used since functions as a chromosome buffer with their fitness and probability considerations of selection according to chromosome fitness.

2- Crossover operator:
by picking two random elements from the rule container to replicate
Off spring chromosomes with a probability of pc=0.6 from the population on the basis of classic genetic operators and then return a new population.
The single point crossover is generated at random, so that any chromosome segment may be selected. We avoid falling into logical contradictions like inconsistency, then new laws are considered.

An example of single point crossover will be :

drug, immune | system — disease, disease | system
disease, disease | infection — drug, immune | infection

(a) before crossover — (b) after crossover

3- Mutation operator: periodically shifts chromosome c genes to a likelihood of pm=0.001, in addition to considering c fitness as an additional weight .

4- Fitness function: Our goal is to search for the most interesting rules of association. (considering minimum support as 0.4 and confidence 50%) as :

$$\textbf{Fitness(c)} = \frac{\mathbf{supp(A1....Ak) - supp(A1....Aj)supp(Aj+1.....Ak)}}{\mathbf{supp(A1.....Aj)(1-supp(Aj+1.....Ak))}}$$

$$\frac{\mathbf{supp(A1....Ak) - supp(A1....Aj)supp(Aj+1.....Ak)}}{\mathbf{supp(A1.....Aj)(1-supp(Aj+1.....Ak))}}$$

Where" c =(j, A1, . . ., Aj, Aj1, . . . , Ak)" is a given chromosome the relative confidence of the corresponding association rule"{A1, . . ., Aj} → → {Aj +1,…,AK}".The fitness of c is, in fact,

As a result from our experiment we get interesting rules with high fitness which reflects the important association between keywords in 200 research address some of these rules are:

{Disease, immune} ——— {system}
{cell, RNA} ——— {virus, protein}
{infection, treatment} ——— {system}
{disease} ——— {epidemic}

Conclusions:

During these tasks, there are so many research studies aimed at developing novel text mining methods as well as solving problems facing conventional techniques. Setting the law of the text mining association gives rise to two major issues that need to be addressed: algorithmic complexity as the number of rules increases dramatically with the number of items in the database. Second, interest-related rules should be collected inside the place of generating rules that typically suggest a top algorithm. Utilizing Genetic Algorithm to find a relation in unorganized document between keywords is very powerful and effective in solving this issue as it effectively cuts Search space for threshold-based consistency measurements of rules.

References:

1. J. Manimaran, T. Velmurugan ," A Survey of Association Rule Mining in Text applications ",IEEE International Conference on Computational Intelligence and Computing Research At: India,pp.698-699,2013. DOI: 10.1109/ICCIC.2013.6724258.
2. Hashimi. H, et al. "Selection criteria for text mining approaches". Computers in Human Behavior (2015).http://dx.doi.org/ 10.1016/j.chb.2014.10.062
3. (N.Padhy,P. Mishra , and R.Panigrahi, "TheSurvey of Data Mining Applications And Feature ", Engineering and Information Technology (IJCSEIT), Scope" International Journal of Computer Science, Vol.2, No.3, June 2012.
4. Nie, B. and Sun, S.," Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research", Appl.Sci. 7(4),401,2017. https://doi.org/10.3390/app7040401.
5. Tarequl Islam Manir ," Application of Text Mining on the Editorial of a Newspaper of Bangladesh" International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 11,pp.23-24, May 2019 .
6. Kaushik, A. and Naithani, S. " A Comprehensive Study of Text Mining Approach", International Journal of Computer Science and Network Security. 16(2), 69-76,2016.
7. Roul R.K., Varshneya S., Kalra A., Sahay S.K ." A Novel Modified Apriori Approach for Web Document Clustering". In: Jain L., Behera H., Mandal J., Mohapatra D. (eds) Computational Intelligence in Data Mining - Volume 3. Smart Innovation, Systems and Technologies, vol 33. Springer, New Delhi. Doi: 10.1007/978-81-322-2202-6_14,2015.
8. Sumathy, K. and Chidambaram, M. "Text mining: Concepts, applications, tools and issues-an overview". International Journal of Computer Applications. 80(4), 29-32. Doi: 10.5120/13851-1685,2013.
9. Jhanji, D. and Garg, P." Text Mining. International Journal of Scientific Research and Education". 2(8), 1642-1648,2014.
10. Shah Neha K, "Introduction of Text mines and an Analysis of Text mining Techniques", PARIPEX, ISSN: 2250-1991, Volume 2, Issue-2, 2013.
11. Dang, D. S. and Ahmad, P. H.," A review of text mining techniques associated with various application areas". International Journal of Science and Research (IJSR). 4(2), 2461–2466,2015.
12. Kunal Bhatia," Visualizing Association Rules for Text Mining"towards data scienc,pp.1-2,jun23,2019.[12]Kunal Bhatia," Visualizing Association Rules for Text Mining"towards data scienc,pp.1-2,jun23,2019.
13. g products for an assembly oriented product family identify," Association rules mining in R for product performance management in industry 4.0 " 11th CIRP Conference on Industrial Product-Service Systems,Volume 83, Pages 699-704,2019.
14. Kim, Jong Woo, Song-Yi Han, and Dong Sung Kim. "Association rules application to identify customer purchase intention in a real-time marketing communication too"l, in Fourth International Conference on Ubiquitous and Future Networks (ICUFN). 2012.
15. Bing, H. and. Ye-Bai. Li," Research and Application of Association Rules Methods in Data Mining for Commercial Sales Analysis. in International Conference on Networking and Digital Society", 2009.

16. A., R. Langer, and S. Conrad. TARtool: A Temporal Dataset Generator for Market Basket Analysis. in 4th International Conference on Advanced Data Minin and Applications (ADMA). 2008.
17. Wang, Yaqin,and Song Yuming. Classification Model Based on Association Rules in Customs Risk Management Application. in International Conference on Intelligent System Design and Engineering Application. 2010.
18. Ben Lutkevich; Mark Labbe,"association rules",Margaret Rouse, WhatIs.com, September 2020.
19. Adekanmbi 'Yosola," Association Rule Mining - Apriori Algorithm",the journal blog,Dec 17, 2018 .
20. B.J AlKhafaji, M Salih, S Shnain, Z Nabat ,improved technique for hiding data in a colored and a monochrm images,2020, Periodicals of Engineering and Natural Sciences 8 (2), 1000-1010
21. BJ AlKhafaji, MA Salih, S Shnain, Z Nabat,segmenting video frame images using genetic algorithms,2020 Periodicals of Engineering and Natural Sciences 8 (2), 1106-1114
22. Al-Khafaji, B.J. Proposed Speech Analyses Method Using the Multiwavelet Transform , Ibn Al-Haitham Journal For Pure And Applied Science.2014, vol.27 No. 1 .
23. Al-Khafaji, B.J. Detect The Infected Medical Image Using Logic Gates. Ibn Al-Haitham Journal For Pure And Applied Science.2014, 27, 2, 260-267.