A web content mining application for detecting relevant pages using Jaccard similarity

Ahmed Adeeb Jalal, Abdulrahman Ahmed Jasim, Amar A. Mahawish
Computer Engineering Department, College of Engineering, Al-Iraqia University, Baghdad, Iraq

Article Info

Article history:

Received Sep 8, 2021
Revised Jun 22, 2022
Accepted Jul 18, 2022

Keywords:

Data mining
Jaccard similarity
Query suggestions
Text mining
Web content mining

ABSTRACT

The tremendous growth in the availability of enormous text data from a variety of sources creates a slew of concerns and obstacles to discovering meaningful information. This advancement of technology in the digital realm has resulted in the dispersion of texts over millions of web sites. Unstructured texts are densely packed with textual information. The discovery of valuable and intriguing relationships in unstructured texts demands more computer processing. So, text mining has developed into an attractive area of study for obtaining organized and useful data. One of the purposes of this research is to discuss text pre-processing of automobile marketing domains in order to create a structured database. Regular expressions were used to extract data from unstructured vehicle advertisements, resulting in a well-organized database. We manually develop unique rule-based ways of extracting structured data from unstructured web pages. As a result of the information retrieved from these advertisements, a systematic search for certain noteworthy qualities is performed. There are numerous approaches for query recommendation, and it is vital to understand which one should be employed. Additionally, this research attempts to determine the optimal value similarity for query suggestions based on user-supplied parameters by comparing MySQL pattern matching and Jaccard similarity.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Ahmed Adeeb Jalal
Computer Engineering Department, College of Engineering, Al-Iraqia University
Baghdad, Iraq
Email: ahmedadeeb@aliraqia.edu.iq

1. INTRODUCTION

Text mining is a very effective technique for using the enormous quantities of unstructured textual data that are commonly accessible in big data analytics, to find new information and discover important patterns and correlations buried in the data. Unstructured data accounts for more than 80% of the information accessible on web sites. As a result, it is more difficult to evaluate and search. As a consequence, attempting to detect and extract the feature properties from many heterogeneous formats of stored data has significant obstacles [1], [2]. Textual data (email messages, word documents, and PDF files) and non-textual data (MP3 music, photos, and videos) are both considered unstructured data on the Internet. Unstructured data refers to any kind of not-indexed data, either in a relational database or via another indexing technique.

The phrase "text mining" refers to the process of extracting usable information from human language texts via the analysis of enormous amounts of data [3], [4], such as those included in electronic documents, online web pages, and vehicle advertisements. Classification and organization of data by text mining algorithms is based on lexical or linguistic patterns. Calculating the similarity of texts is critical for a variety of activities, including text summarization, machine translation, text clustering, and text classification

[5]. As a result, we try to identify one keyword or a sequence of keywords that may be used to construct queries across retrieval systems with the aim of collecting the data we seek. Information retrieval systems are used in a variety of applications, including social media analytics, plagiarism recognition, and search engines, all of which make use of text mining for predictive analytics and opinion mining [6]. As such, text mining is a broad term that covers a number of fields, including information extraction, text analysis, and knowledge discovery.

Text mining began in the 1980s via the use of manual approaches. Since then, the storage capacity of digital information is increasing every month. Thus, storage capacity per month was 122 Exabyte in 2017, and rose to 156 Exabyte in 2018; Cisco predicts a continuous rise in storage capacity over the next few years [7]. Between 2016 and 2020, the capacity of information storage doubled, and it is expected to increase as well in 2021 [8]. This growth renders manual text mining approaches obsolete and prohibitively costly. Thus, the huge increase in data volume necessitates pre-processing that may have a discernible effect on achieving good findings. As a result, many text mining algorithms consider text processing activities to be a critical component of addressing the documents. Thus, the emphasis was on automating the processing of data during the development of special-purpose programs in a variety of fields [9]. Text mining programs are utilized in a wide variety of fields, including sociology, education, research, medicine, and marketing. The programs are being developed at a breakneck speed, in lockstep with the evolution of technology.

One of the most effective computer science techniques is regular expressions [10] and is heavily used in natural language processing. Regular expressions are a subset of linguistic notation that are determined by a defined search pattern in conjunction with a supplied sequence [10]. While regular expressions have certain limits, they are an incredibly powerful tool for defining a wide range of protocols, formats, and tiny textual languages. This advantageous language is used to find words, terms, or text strings that fit a pattern in textual data. Typically, patterns are composed of a series of characters. Regexps are a kind of algebraic notation used to define and characterize a collection of specific strings. Regexps are now extensively employed throughout a range of scientific disciplines, including information technology, medicine, and sociology. For example, identify network protocol [11], malware detection [12], web mining [13], summarize the events that led to childcare outside the home [14], and determining the sequences of deoxyribonucleic acid (DNA) [15]. Despite this, it is not easy to comprehend the regular expressions to reuse them, owing to the absence of abstract procedures that would otherwise result in the fast expansion of regular expressions. Additionally, it is quite difficult to locate the appropriate regular expression and apply it to a given situation.

We provide a robust set of self-contained regular expressions in this paper to solve the lack of understandability and usefulness. Additionally, the suggested regular expressions' syntax gives increased coverage while minimizing complexity. Thus, this strategy has a considerable influence on the discovery of relevant information and the enhancement of database search ability through user queries. As a result, we suggested a regular expression-based semantic analysis method for addressing advertisement pages and user questions. The suggested system leads users through the structured database domain by way of their queries.

Typically, users confront several obstacles when it comes to learning new languages due to their fast lifestyles. As a result, one of the objectives of this paper is to develop a search engine capable of learning patterns described by regular expressions. Then, based on advanced learning and the user's requirements, the search engine can detect and discriminate between two languages: Turkish and English.

The procedure of searching for advertising sites is to locate the most relevant advertising based on the user's description or keywords. For conventional ad searching, the user enters terms and then presses the button to see the results. This paper advances MySQL pattern matching and Jaccard similarity for user-supplied query based on the following parameters: response time, proximity of keywords to the suggestion, proximity ratings and sorting from closest data, data retrieval strength based on the number of words entered, and data retrieval strength among a number of data that can be suggested. Several studies have been conducted to show the various approaches for using regular expressions and determining the degree of similarity between words, phrases, paragraphs, and texts. As such, we provide some reviews of relevant literature on regular expressions and similarity measures in various applications.

Bhatia *et al.* [16] created a method for analyzing unstructured automobile ads on the website (https://www.kbb.com) of Kelley Blue Book. This method put together a variety of techniques of natural language processing (NLP) as well as manual rules, feature engineering, and maximum entropy classifiers. As a result of this information, the relational database is populated with values for interesting attributes. Thus, people can do structured searches on specific interesting characteristics with ease. Michelson and Knoblock [17] suggested analyzing postings to develop reference sets for classifying car ads on the website (https://www.craigslist.org) of Craigslist. As a result, these sets of references can serve as the foundation for building tables in a relational database that contain entities of postings and their characteristics. For example, a reference collection for cars could contain information about the vehicle's model, brand, and trim. This type of information enables structured search on unstructured data. Kabasakal and Soyuer [18] provide research that satisfies a demand in a university–industry partnership web portal project (EGEVASYON) by matching a list of projects with a relevant subset of keywords through the Jaccard similarity measure. Each similar subset of keywords has various use-cases in accordance with the web portal's requirements. It is necessary to process a given collection of keywords that may refer to another project, a researcher's areas of interest, or simply a user-supplied query. The challenge is equivalent to picking more relevant sets based on an input set in each of the three cases outlined. Intuitively, recommendations should come from the projects that most closely match the visitor's profile.

With 492 articles, Rinartha and Suryasa [19] created a methodology for finding the right article based on the user description or keywords. This method for article searching makes use of MySQL pattern matching and Jaccard similarity. Users simply need to type a few characters or words to search, and the system will automatically generate search options. When the user finishes typing in the input field area or when the user picks the supplied recommendation, the system will halt offering search. The study's findings indicate that although Jaccard similarity query suggestion delivers more accurate search results, it takes longer to execute than MySQL pattern matching.

It is important to note that natural language processing is a broad term that includes many fields such as artificial intelligence, machine learning, and linguistic processing. It is used to describe the ability of computers to understand human languages through computational methods [20]. Due to their daily concerns, the majority of users lack the time necessary to learn even a few basic lines in a new language. The consequence was the development of natural language processing, which was designed to make the user's work easier while also satisfying the user's desire to interact with a computer in natural language. Therefore, we will concentrate on gathering relevant information by crawling the website (https://www.arabam.com) [21] of car advertisements in order to assist consumers throughout their search process.

As shown above, we summarized some of the literature reviews on extracting and analyzing information from unstructured data and on resource semantic annotation. The second part discusses techniques that include suggested approaches, web content mining, and text processing. The third part discusses the proposed approach and the methods that were utilized to implement it. Finally, this paper aims to facilitate the search process for users by writing a few words.

2. RESEARCH METHOD

Text mining is a broad term that encompasses a range of approaches, including web data mining, information extraction, and natural language processing, which are used to explore and identify patterns and interrelationships. Developing a range of techniques based on various mathematical and statistical approaches is necessary to extract high-quality information from unstructured data [22]. Several of these strategies are aimed at enhancing language, pattern recognition, and mathematical capabilities.

Unstructured data contains useful information in the form of statements and phrases. Thus, text mining should detect semantic patterns in order to extract and utilize this information that allows for the search ability of the stored database. Structured data requires the development of a database model that represents the information kinds that will be captured, processed, stored, and retrieved. The primary objective is to convert unstructured data into analyzable data (structured data) using natural language processing analytical methodologies. Because structured data is simple to manage and handle, it may be inputted, stored, analyzed, and searched for concurrently. Thus, structured data reduces the high costs and performance constraints associated with storage, memory, and processing.

The text mining process flow is iterative, examining data acquired from web pages using a series of keywords/patterns of regular expressions to get the best results. This phase might result in clusters of several words being stored in the relational tables. Additionally, the search query process evaluates user queries to generate a cluster of various keywords from which a database search may be conducted to uncover information. The fundamental concept of the search process is to leverage the numerous phrases entered by the user to locate the most relevant advertising sites based on MySQL pattern matching and Jaccard similarity. The system is divided into three major phases that correspond to the primary tasks that are involved in typical text mining techniques that are the foundation of the system work. These stages are (web content mining, text processing, and text similarity measures) to provide logical conclusions when determining the similarity between two texts.

2.1. Web content mining

Web content mining is the process of extracting, analyzing, and integrating valuable data, information, and knowledge from the content of web pages. The researchers created a web crawler tool to collect the necessary data from the web. For instance, apartments' information could be gathered from online real estate listing websites to create a new dataset [23]. Mining is always attempting to deduce the website's structure in order to convert it into a database. Normally, a crawler either has a specific domain (URL) for which it is primarily responsible for page harvesting or has no domain at all. A web crawler's primary objective is to efficiently get current data from relevant pages. Thus, crawlers are increasingly considering distributed crawling as a means of circumventing the capability restrictions of global search engines by spreading the crawling operation among users, queries, or even client machines [24]. Prior to text mining, one first identifies the HTML document's coding standard and converts it to an inner code. Then, using various text mining techniques, one may discover important knowledge and patterns.

2.2. Text processing

Text summarization is a long-standing difficulty for applications of text mining. Due to the fact that these applications must summarize lengthy text documents in order to provide a concise summary of the subject [25], it is hoped that the specifics will be omitted while the keywords will be retained. As a result, a text summary is the act of gathering and composing a condensed version of the original text document that includes relevant information for the user.

The deluge of digital information has outstripped our capacity to comprehend it. Thus, semantics or meaning interpretation may be used for computer literacy in order to get access to knowledge and information. Computational intelligence at a higher level, natural language processing, and text summarization all need a judgment of the appropriateness of the returned findings. Text summarizing approaches may be broadly classified into two types: extractive summarizing, which extracts fundamental information from the original text, and abstractive summarizing, which generates new information from the original material [26].

Tokenization is often the first step in text preprocessing [22]. Tokenization is the process of transforming a series of characters into tokens. For the most part, researchers employ words as tokens and white spaces to divide strings. However, phrases or paragraphs might likewise be used as tokens. By using word tokens and converting them to a document-term matrix, the 'bag-of-words' is created.

One of the unsung achievements of computer science standards is the regular expression (Regex), a language for describing text search strings. This useful language is embedded in almost every computer language, word processor, and text processing application [10]. In its most basic form, a regular expression is an algebraic notation for describing a collection of strings. They are especially useful for textual searches when we have a pattern to look for and a corpus of texts to search through. Thus, the majority of what we will perform with unstructured data will require first separating or tokenizing words using regular expressions to transform them into more searchable data that is stored in a relational database.

2.3. Text similarity measures

Information retrieval is a process of finding and retrieving relevant information linked with a given collection of keywords in order to meet the user's specific information request [27]. As a consequence, this method has often been focused on easing access to information rather than on processing, analyzing, and summarizing data in order to uncover hidden patterns. The retrieval idea is based on the notion of questions and answers, with the objective of finding documents that contain the answers to the questions. Thus, after the text summarizing step, information retrieval may be focused on the user's inquiry.

In text mining, similarity measures are defined as functions that increase with the number of common characteristics and decrease with the number of dissimilar characteristics. Word similarities may be quantified in two ways: lexical and semantic. When words have a similar character sequence, they are lexically similar, but, when they have the same topic, they are semantically similar. Typically, string-based algorithms are used to determine lexical similarity. However, knowledge-based and corpus-based methods are used to determine semantic similarity [28].

Numerous techniques may be employed to generate query recommendations. Two search applications were constructed in this research employing MySQL pattern matching and Jaccard similarity, and the resulting results were compared. MySQL supports both traditional SQL pattern matching and pattern matching using extended regular expressions. MySQL's pattern matching functionality allows you to use it to match any single character or match any number of characters [29]. Jaccard similarity is a technique for comparing members (words/phrases) of two sets in order to determine which members are common and which are unique. Jaccard similarity is often used in member comparison to determine the members' similarity value. Jaccard similarity is utilized in query recommendations to provide good results for the user throughout the search process.

In this paper, the crawler is programmed to retrieve all the pertinent URLs from the site's primary title in order to extract and store the content of the pages in the repository. Text processing is applied to all pages crawled and saved in the repository to summarize and tokenize them. Our methodology is based on predetermined regular expressions to be utilized in Arabam websites ad collection (https://www.arabam.com)

[21]. We create and produce these regular expressions in Turkish to parse the content with the regex. These expressions are meant to get matching information according to Arabam advertisement pages. A structured database is the outcome of identifying the best matches to be indexed in it. Thus, depending on the similarity approaches employed, the database and the user-entered terms are utilized to present the search results.

Figure 1 illustrates the phases of the suggested techniques for computing the similarity degree and matching. The system is capable of handling both short and long Turkish texts crawled from websites. Additionally, this system enables searching and matching in both English and Turkish by translating the collected terms into English. The translation tool increases the user's search options by allowing him to search and query in two languages.

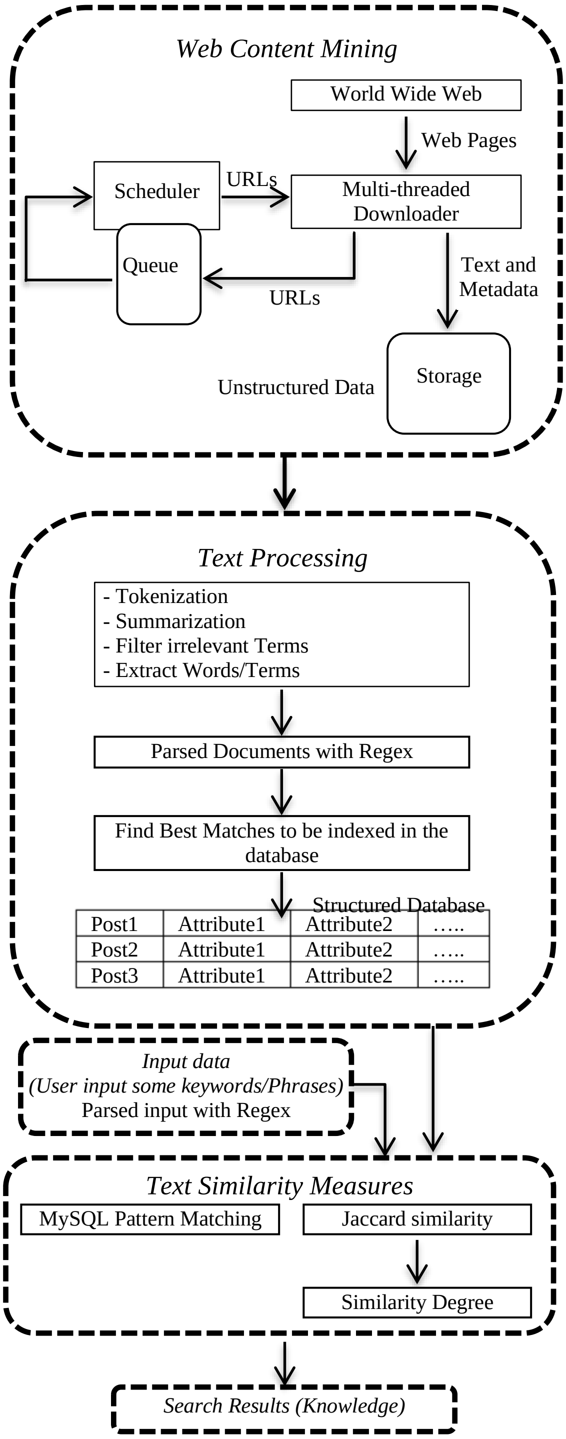


Figure 1. Process flow of system architecture for text mining techniques

As seen in Figure 1, the first step (web content mining) is concerned with collecting and extracting unstructured textual material from visited websites. Regarding the second stage (text processing), it is the step in which unstructured texts from the previous stage are processed in order to extract matching terms using regular expressions. The system's final stage (text similarity measures) involves intersecting the matching of user-entered keywords (after further processing) to the database created during the text processing phase. The recommended methodology is as follows:

a. When a web crawler is used to extract the content of a collection of posts P that correspond to training samples of Arabam posts, the collection of P web pages may be represented as (1):

$$P = \{p_1, p_2, p_3, \dots \dots, p_n\} \quad (1)$$

b. To identify the qualities, use specified regular expressions R, which might be a collection of different words/terms in the posts. The following terms are used to describe regular expressions in R.

$$R = \{r_1, r_2, r_3, \dots \dots, r_k\} \quad (2)$$

c. The structured database organizes the frequency with which regular expressions R occur in P posts. The structured database can then be referred to by $D(P, R)$. Where the P posts set denotes the rows and R regular expressions the columns.

$$D(i,j) = \int_{j=1}^{k} R_j \in \int_{i=1}^{n} P_i \quad (3)$$

d. Create SQL statements using the search query attributes entered by the user that can be denoted by U. Thus, the results are shown by matching the SQL statement to the structured database.

$$SQL_{ST}(U) \in D \quad (4)$$

e. The Jaccard similarity coefficient is used to determine comparable words between two texts by dividing the number of crossing words between them by the total number of words in both texts. Thus, the similarity degree is calculated by dividing the number of crossing terms between U and D by the union of all their words.

$$J(U, D) = \frac{U \cap D}{U \cup D} \quad (5)$$

Algorithm 1 is a representation of our technique that is shown in Figure 1, for creating a structured database utilizing Regex matches. As seen in algorithm 1, users can input any search queries about the cars to receive the top relevant information depending on their query words. Algorithm 1 illustrates how to analyze user queries in order to uncover important properties inside a structured database based on similarity methods.

Algorithm 1 show results based on MySQL pattern matching and Jaccard similarity

```
Input:
   $\vec{P} \leftarrow < p_1, p_2, \dots., p_n >$
   $\vec{R} \leftarrow < r_1, r_2, \dots., r_K >$
   $\vec{U} \leftarrow < u_1, u_2, \dots., u_a >$
Initialize:
   Regex Match Vector $\overrightarrow{RMV}$={}
Output:
   Search Results $S$.
   for $i$=1:$n$
        for $j$ =1:$k$
          if $r_j \in p_i$
                 $D(i,j) \leftarrow RMV$
         end
        end
   end
   for $i$=1:$a$
        $SQL_{ST}$(U) $\leftarrow u_i$
   end
   $S \leftarrow SQL_{ST}(U) \in D$
   $S \leftarrow J(U,D) = \frac{U \cap D}{U \cup D}$
```

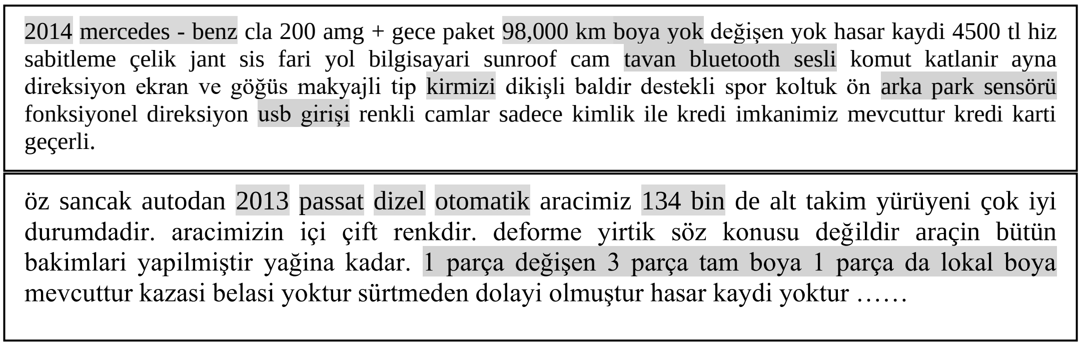
3. RESULTS AND DISCUSSION

The automated text mining methods discussed earlier are in use in a variety of sectors for the purpose of identifying patterns and information in unstructured data. So, numerous computer languages, including Java, Python, and C++, support the creation of regular expressions. Regular expressions establish a search pattern for textual data using a series of letters. There are several regular expression formats, such as in the regular expressions library (http://www.RegExLib.com), which now indexes 4,149 expressions uploaded by 2,818 contributors worldwide and has garnered widespread popularity for its syntactic simplicity. Nonetheless, regular expressions have certain drawbacks. As a result, strategies for validating regular expressions [30] or testing their use in applications [31] are being sought to expose potential flaws in regular expressions.

These flaws are classified into three categories: complexity, mistakes, and version proliferation, so regexes are difficult [32]. These issues make it rather difficult to reuse regular expressions. Several of the repository's regular expressions surpassed 4,000 characters. As a result, it is not easy for users to comprehend what these long sequences of expressions do, and it is practically impossible to verify them. Additionally, there are many regular expressions strewn across online pages, which sometimes experience minor lapses. Typically, these gaps or flaws are so precise that they are difficult for people to notice. Additionally, other variants of regexps are produced and maintained in various repositories to accomplish the same task. As a result, locating and selecting the appropriate version for a certain activity is challenging. These three issues make it very difficult to deduce the purpose of the regular expression syntax. As a consequence, it is quite cumbersome to determine if we can choose and reuse this regular expression among a large set of regexps stored in repositories. The aim of this paper is to demonstrate how to automate knowledge discovery by formulating regular expressions for text analysis and classification. These expressions will be meticulously formulated to be able to extract keywords from targeted web pages seamlessly by supplying a collection of regular expression strings to detect relevant pages.

Certain languages, such as Turkish (research domain), include a large number of suffixes, making word tokenization more complex. As a consequence, we need to create innovative regular expressions for segmenting or tokenizing words in order to describe patterns in discovered texts. Regular expressions may be used to specify specific words from a body of text to extract. On the other hand, regular expressions are essential for converting textual data to a more convenient standard format. Additionally, we will analyze user-generated queries in order to give more comprehensive semantic explanations that may be utilized to do repository searches. As a result, the search procedure will be simple and accessible through the application we created. This application enables querying by defining desired characteristics, names, or strings in order to get the best results with the most efficiency, accuracy, and speed.

The dataset includes automotive advertisements obtained from the Arabam websites. As a result, we extracted postings' content from their original HTML format by crawling all the sub-sites. The domain and characteristics we picked posed several hurdles in terms of deciding which candidate values would be of interest to consumers. As is generally known, it is difficult for individuals new to Turkey to acquire the language, as well as other foreign languages, in a timely way. As a result, we are attempting to provide a method for facilitating the search process inside the advertisements accessible on the English-language website. Additionally, we ensured that the search query was available in both English and Turkish, with the option of writing and drafting questions. As seen in Figure 2, we should define some features of the automobile, such as if it has colored components, Bluetooth, or sensors. Even though these traits are not often expressed in detail, the suggested methodology assists in analyzing descriptions of automobiles in order to identify certain key aspects.



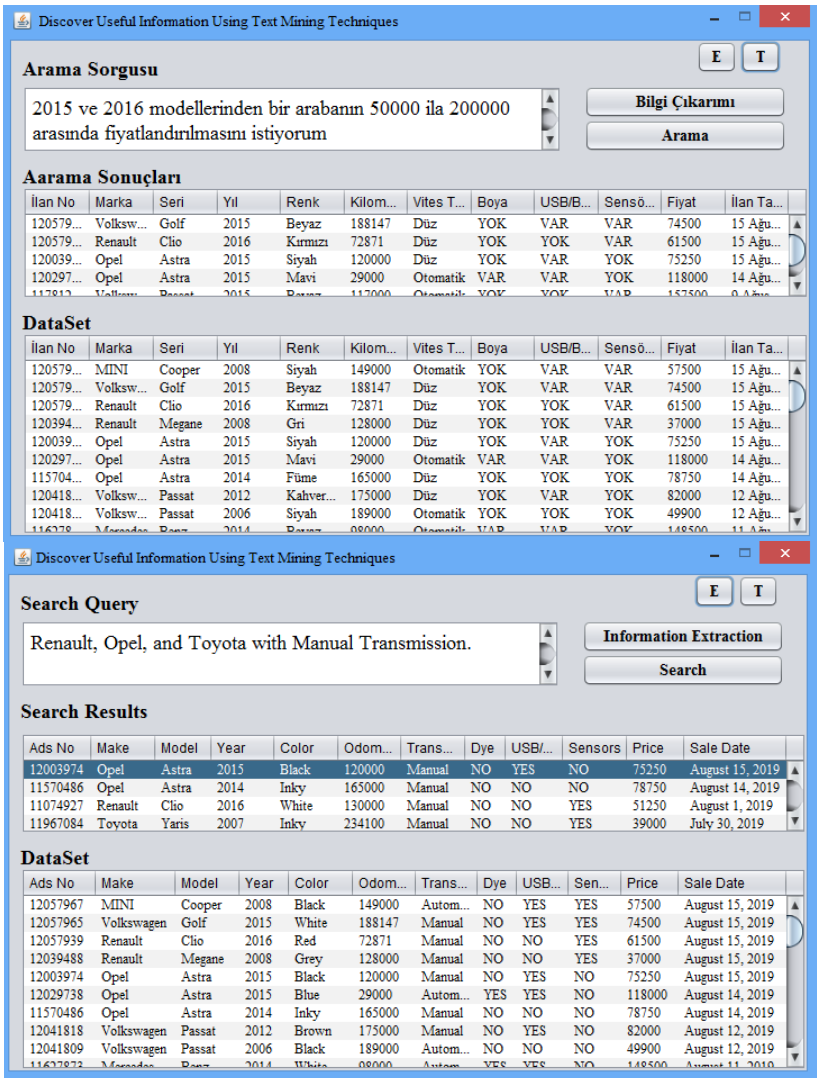
2014 mercedes - benz cla 200 amg + gece paket 98,000 km boya yok değişen yok hasar kaydi 4500 tl hiz sabitleme çelik jant sis fari yol bilgisayari sunroof cam tavan bluetooth sesli komut katlanir ayna direksiyon ekran ve göğüs makyajli tip kirmizi dikişli baldir destekli spor koltuk ön arka park sensörü fonksiyonel direksiyon usb girişi renkli camlar sadece kimlik ile kredi imkanimiz mevcuttur kredi karti geçerli.

öz sancak autodan 2013 passat dizel otomatik aracimiz 134 bin de alt takim yürüyeni çok iyi durumdadir. aracimizin içi çift renkdir. deforme yirtik söz konusu değildir araçin bütün bakimlari yapilmiştir yağina kadar. 1 parça değişen 3 parça tam boya 1 parça da lokal boya mevcuttur kazasi belasi yoktur sürtmeden dolayi olmuştur hasar kaydi yoktur ……

Figure 2. Example of attributes detection

We created a search engine that is dynamic and is based on regular phrases found in online vehicle ads. The goal of this project is to extract structured data from internet adverts in order to do search and query operations. Regular expression methods were used to circumvent the language barrier and extract significant characteristics from Arabam posts. Instead of using generalization patterns, this program searches for candidates for extraction using regular expressions.

The data collected and stored in the relational tables were in Turkish. As a consequence, we trained the suggested application using Turkish and English regular expressions. We analyze two different types of user search queries in Turkish and English, as shown in Figure 3. The first query in Turkish uses regular expression matching to determine the vehicle's production year and price. Algorithm 1 formulates the SQL query using these observed properties to get search results. After matching the text analysis for the second inquiry in English, we discovered the distinguishing qualities of the automobile, which are the model and the gearbox. Similarly, the SQL query is formed.



Arama Sorgusu

2015 ve 2016 modellerinden bir arabanın 50000 ila 200000 arasında fiyatlandırılmasını istiyorum

Aarama Sonuçları

İlan No	Marka	Seri	Yıl	Renk	Kilom...	Vites T...	Boya	USB/B...	Sensö...	Fiyat	İlan Ta...
120579...	Volksw...	Golf	2015	Beyaz	188147	Düz	YOK	VAR	VAR	74500	15 Ağu...
120579...	Renault	Clio	2016	Kırmızı	72871	Düz	YOK	YOK	VAR	61500	15 Ağu...
120039...	Opel	Astra	2015	Siyah	120000	Düz	YOK	VAR	YOK	75250	15 Ağu...
120297...	Opel	Astra	2015	Mavi	29000	Otomatik	VAR	VAR	YOK	118000	14 Ağu...

DataSet

İlan No	Marka	Seri	Yıl	Renk	Kilom...	Vites T...	Boya	USB/B...	Sensö...	Fiyat	İlan Ta...
120579...	MINI	Cooper	2008	Siyah	149000	Otomatik	YOK	VAR	VAR	57500	15 Ağu...
120579...	Volksw...	Golf	2015	Beyaz	188147	Düz	YOK	VAR	VAR	74500	15 Ağu...
120579...	Renault	Clio	2016	Kırmızı	72871	Düz	YOK	YOK	VAR	61500	15 Ağu...
120394...	Renault	Megane	2008	Gri	128000	Düz	YOK	YOK	VAR	37000	15 Ağu...
120039...	Opel	Astra	2015	Siyah	120000	Düz	YOK	VAR	YOK	75250	15 Ağu...
120297...	Opel	Astra	2015	Mavi	29000	Otomatik	VAR	VAR	YOK	118000	14 Ağu...
115704...	Opel	Astra	2014	Füme	165000	Düz	YOK	YOK	YOK	78750	14 Ağu...
120418...	Volksw...	Passat	2012	Kahver...	175000	Düz	YOK	VAR	YOK	82000	12 Ağu...
120418...	Volksw...	Passat	2006	Siyah	189000	Otomatik	YOK	YOK	YOK	49900	12 Ağu...

Search Query

Renault, Opel, and Toyota with Manual Transmission.

Search Results

Ads No	Make	Model	Year	Color	Odom...	Trans...	Dye	USB/...	Sensors	Price	Sale Date
12003974	Opel	Astra	2015	Black	120000	Manual	NO	YES	NO	75250	August 15, 2019
11570486	Opel	Astra	2014	Inky	165000	Manual	NO	NO	NO	78750	August 14, 2019
11074927	Renault	Clio	2016	White	130000	Manual	NO	NO	YES	51250	August 1, 2019
11967084	Toyota	Yaris	2007	Inky	234100	Manual	NO	NO	YES	39000	July 30, 2019

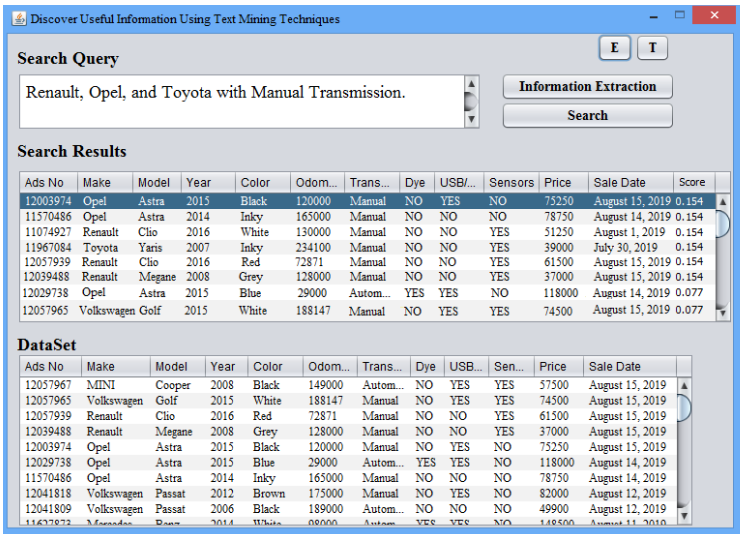
DataSet

Ads No	Make	Model	Year	Color	Odom...	Trans...	Dye	USB...	Sen...	Price	Sale Date
12057967	MINI	Cooper	2008	Black	149000	Autom...	NO	YES	YES	57500	August 15, 2019
12057965	Volkswagen	Golf	2015	White	188147	Manual	NO	YES	YES	74500	August 15, 2019
12057939	Renault	Clio	2016	Red	72871	Manual	NO	NO	YES	61500	August 15, 2019
12039488	Renault	Megane	2008	Grey	128000	Manual	NO	NO	YES	37000	August 15, 2019
12003974	Opel	Astra	2015	Black	120000	Manual	NO	YES	NO	75250	August 15, 2019
12029738	Opel	Astra	2015	Blue	29000	Autom...	YES	YES	NO	118000	August 14, 2019
11570486	Opel	Astra	2014	Inky	165000	Manual	NO	NO	NO	78750	August 14, 2019
12041818	Volkswagen	Passat	2012	Brown	175000	Manual	NO	YES	NO	82000	August 12, 2019
12041809	Volkswagen	Passat	2006	Black	189000	Autom...	NO	NO	NO	49900	August 12, 2019

Figure 3. Text analysis for structured search process using MySQL pattern matching [33]

When MySQL pattern matching and Jaccard similarity are used, users need to input a few search terms and the program will provide results. Figure 4 shows the application's derived scores after evaluating a collection of user-selected texts. The scores indicate the degree of similarity as determined by the Jaccard coefficient. When texts from the same class or a neighboring class are compared, the Jaccard coefficient shows a rise in similarity degrees. However, when texts from other classes are compared, the Jaccard coefficient shows a reduction in similarity degrees.

Due to the absence of a counting mechanism, proximity scores for MySQL pattern matching cannot be computed. Rather than that, MySQL pattern matching is dependent on word matches, and without a match, search possibilities disappear. However, the program shows a range of values between 0 and 1 for Jaccard similarity. If the keywords are not comparable to the data in the database, the score for the Jaccard index is 0, and if the keywords are 100 percent similar to the data in the database, the score is 1. This enables the user to choose the result that is the most closely related to his query field. The conclusion of this paper is that Jaccard similarity query suggestion provides better accurate search results than MySQL pattern matching but requires more processing time.



Search Query

Renault, Opel, and Toyota with Manual Transmission.

Search Results

Ads No	Make	Model	Year	Color	Odom...	Trans...	Dye	USB/...	Sensors	Price	Sale Date	Score
12003974	Opel	Astra	2015	Black	120000	Manual	NO	YES	NO	75250	August 15, 2019	0.154
11570486	Opel	Astra	2014	Inky	165000	Manual	NO	NO	NO	78750	August 14, 2019	0.154
11074927	Renault	Clio	2016	White	130000	Manual	NO	NO	YES	51250	August 1, 2019	0.154
11967084	Toyota	Yaris	2007	Inky	234100	Manual	NO	NO	YES	39000	July 30, 2019	0.154
12057939	Renault	Clio	2016	Red	72871	Manual	NO	NO	YES	61500	August 15, 2019	0.154
12039488	Renault	Megane	2008	Grey	128000	Manual	NO	NO	YES	37000	August 15, 2019	0.154
12029738	Opel	Astra	2015	Blue	29000	Autom...	YES	YES	NO	118000	August 14, 2019	0.077
12057965	Volkswagen	Golf	2015	White	188147	Manual	NO	YES	YES	74500	August 15, 2019	0.077

DataSet

Ads No	Make	Model	Year	Color	Odom...	Trans...	Dye	USB...	Sen...	Price	Sale Date
12057967	MINI	Cooper	2008	Black	149000	Autom...	NO	YES	YES	57500	August 15, 2019
12057965	Volkswagen	Golf	2015	White	188147	Manual	NO	YES	YES	74500	August 15, 2019
12057939	Renault	Clio	2016	Red	72871	Manual	NO	NO	YES	61500	August 15, 2019
12039488	Renault	Megane	2008	Grey	128000	Manual	NO	NO	YES	37000	August 15, 2019
12003974	Opel	Astra	2015	Black	120000	Manual	NO	YES	NO	75250	August 15, 2019
12029738	Opel	Astra	2015	Blue	29000	Autom...	YES	YES	NO	118000	August 14, 2019
11570486	Opel	Astra	2014	Inky	165000	Manual	NO	NO	NO	78750	August 14, 2019
12041818	Volkswagen	Passat	2012	Brown	175000	Manual	NO	YES	NO	82000	August 12, 2019
12041809	Volkswagen	Passat	2006	Black	189000	Autom...	NO	NO	NO	49900	August 12, 2019
11627873	Mercedes	Benz	2014	White	98000	Autom...	YES	YES	NO	148500	August 11, 2019

Figure 4. Text analysis for structured search process based on scores of Jaccard similarity

4. CONCLUSION

Through semantic analysis and translation of user inquiries, we are attempting to validate a new language-based method. This paper describes a successful application for correctly collecting data from internet classified ads. Our application's present architecture makes use of a fairly basic method for extracting data from web pages. As a result, this method depends on natural language processing's regular expressions to extract useful information from postings. As a result, considerable progress has been achieved in attribute-based research. Nonetheless, we intend to include more attributes to enhance the system's utility for users and to facilitate searching over a broader range of attributes. This could help to increase involvement in its characteristics. Additionally, we will expand the system's coverage by including more car advertising websites in several languages.

REFERENCES

[1] M. R. Murty, J. V. Murthy, P. P. Reddy, and S. C. Satapathy, “A survey of cross-domain text categorization techniques,” in *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*, Mar. 2012, pp. 499–504, doi: 10.1109/RAIT.2012.6194629.

[2] X. Zhou *et al.*, “A survey on text classification and its applications,” *Web Intelligence*, vol. 18, no. 3, pp. 205–216, Sep. 2020, doi: 10.3233/WEB-200442.

[3] M. U. Maheswari and J. G. R. Sathiaseelan, “Text Mining: Survey on Techniques and Applications,” *International Journal of Science and Research*, vol. 6, no. 6, pp. 1660–1664, 2017.

[4] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, “Text mining: techniques, applications and issues,” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, 2016, doi: 10.14569/IJACSA.2016.071153.

[5] S. G. Cho and S. B. Kim, “A data-driven text similarity measure based on classification algorithms,” *International Journal of Industrial Engineering : Theory Applications and Practice*, vol. 24, no. 3, 2017, doi: 10.23055/ijietap.2017.24.3.2451.

[6] L. Faty *et al.*, “News comments modeling for opinion mining: the case of senegalese online press,” in *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, Jun. 2020, pp. 1–5, doi: 10.1109/ICACCE49060.2020.9155069.

[7] T. Barnett, S. Jain, U. Andra, and T. Khurana, “Visual networking index,” Cisco, 2018.

[8] K. Mlitz, “Data center storage capacity worldwide: consumer and business segments 2016-2021,” Statista, 2021.

[9] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, “Text mining in big data analytics,” *Big Data and Cognitive Computing*, vol. 4, no. 1, pp. 1–34, Jan. 2020, doi: 10.3390/bdcc4010001.

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational*

Linguistics, and Speech Recognition (3rd Edition draft). Stanford University, 2020.
[11] J. Shi, X. Yu, and Z. Liu, "Nowhere to hide: a novel private protocol identification algorithm," *Security and Communication Networks*, vol. 2021, pp. 1–10, Mar. 2021, doi: 10.1155/2021/6672911.
[12] O. Aslan and R. Samet, "A comprehensive review on malware detection approaches," *IEEE Access*, vol. 8, pp. 6249–6271, 2020, doi: 10.1109/ACCESS.2019.2963724.
[13] N. Wang, Q. Jiao, and Z. Zhao, "A general web page extraction method aiming at online social networks," in *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Jun. 2020, pp. 135–140, doi: 10.1109/ITOEC49072.2020.9141580.
[14] R. Goerge, *Bringing big data in public policy research: Text mining to acquire richer data on program participants, their behaviour, and services,*. Chicago, IL: Chapin Hall at the University of Chicago, 2018.
[15] J. Hon, T. Martínek, J. Zendulka, and M. Lexa, "pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R," *Bioinformatics*, vol. 33, no. 21, pp. 3373–3379, Nov. 2017, doi: 10.1093/bioinformatics/btx413.
[16] N. Bhatia, R. Kumar, and S. Senapaty, *Extraction of structured information from online automobile advertisements*. Stanford University, 2018.
[17] M. Michelson and C. A. Knoblock, "Constructing reference sets from unstructured, ungrammatical text," *Journal of Artificial Intelligence Research*, vol. 38, pp. 189–221, May 2010, doi: 10.1613/jair.2937.
[18] İ. Kabasakal and H. Soyuer, "A Jaccard similarity-based model to match stakeholders for collaboration in an industry-driven portal," in *The 7th International Management Information Systems Conference*, Mar. 2021, pp. 15–24, doi: 10.3390/proceedings2021074015.
[19] K. Rinartha and W. Suryasa, "Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, Aug. 2017, pp. 1–4, doi: 10.1109/CITSM.2017.8089237.
[20] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, Jul. 2017, doi: 10.1016/j.inffus.2016.10.004.
[21] Arabam, "Arabam advertisement website," *arabam.com*, 2022. https://www.arabam.com (accessed Jun. 20, 2022).
[22] A. Payak, S. Rai, K. Shrivastava, and R. Gulwani, "Automatic text summarization and keyword extraction using natural language processing," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Jul. 2020, pp. 98–103, doi: 10.1109/ICESC48915.2020.9155852.
[23] R. Annamoradnejad, I. Annamoradnejad, T. Safarrad, and J. Habibi, "Using web Mining in the analysis of housing prices: A case study of tehran," in *2019 5th International Conference on Web Research (ICWR)*, Apr. 2019, pp. 55–60, doi: 10.1109/ICWR.2019.8765250.
[24] L. Liu, T. Peng, and W. Zuo, "Topical web crawling for domain-specific resource discovery enhanced by selectively using link-context," *The International Arab Journal of Information Technology*, vol. 12, no. 2, pp. 196–204, 2015.
[25] L. Abualigah, M. Q. Bashabsheh, H. Alabool, and M. Shehab, "Text summarization: A brief review," in *Recent Advances in NLP: The Case of Arabic Language*, vol. 874, M. Abd Elaziz, M. A. A. Al-qaness, A. A. Ewees, and A. Dahou, Eds. Cham: Springer International Publishing, 2020, pp. 1–15.
[26] M. Allahyari *et al.*, "Text summarization techniques: A brief survey," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 397–405, Jul. 2017.
[27] R. A. Fink, D. R. Zaret, R. B. Stonehirsch, R. M. Seng, and S. M. Tyson, "Streaming, plaintext private information retrieval using regular expressions on arbitrary length search strings," in *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, Aug. 2017, pp. 107–118, doi: 10.1109/PAC.2017.35.
[28] M. K. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 1, pp. 19–28, Mar. 2016, doi: 10.5121/mlaij.2016.3103.
[29] L. Friedrichsen, L. Ruffolo, E. Monk, J. L. Starks, and P. J. Pratt, *Concepts of database management (10th Edition)*, 10th Editi. Cengage Learning, 2020.
[30] L.-X. Zheng, S. Ma, Z.-X. Chen, and X.-Y. Luo, "Ensuring the correctness of regular expressions: A review," *International Journal of Automation and Computing*, vol. 18, no. 4, pp. 521–535, Aug. 2021, doi: 10.1007/s11633-021-1301-4.
[31] P. Arcaini, A. Gargantini, and E. Riccobene, "Fault-based test generation for regular expressions by mutation," *Software Testing, Verification and Reliability*, vol. 29, no. 1–2, pp. 1–22, Jan. 2019, doi: 10.1002/stvr.1664.
[32] L. G. Michael, J. Donohue, J. C. Davis, D. Lee, and F. Servant, "Regexes are hard: Decision-making, difficulties, and risks in programming regular expressions," in *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Nov. 2019, pp. 415–426, doi: 10.1109/ASE.2019.00047.
[33] A. A. Jalal, "Text mining: Design of interactive search engine based regular expressions of online automobile advertisements," *International Journal of Engineering Pedagogy (iJEP)*, vol. 10, no. 3, pp. 35–48, May 2020, doi: 10.3991/ijep.v10i3.12419.

BIOGRAPHIES OF AUTHORS

Ahmed Adeeb Jalal received the Engineer degree in Software Engineering from Al-Rafidain University College, Iraq in 2002. He received the Master degree in Computer Engineering from Yildiz Technical University, Turkey in 2016. Currently, he is a lecturer of Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq. His research interests include data mining, hybrid recommendation systems design, and web applications. He can be contacted at email: ahmedadeeb@aliraqia.edu.iq.

Abdulrahman Ahmed Jasim received the Engineer degree in Computer Engineering from Dijlah University, Iraq in 2012. He received the Master degree in Electrical and Computer Engineering from Altinbas University, Turkey in 2018. Currently, he is a lecturer of Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq. Now, he is a PhD student at Altinbas University, Turkey. His research interests include data mining, machine learning, and deep learning. He can be contacted at email: abdulrahman.alsalmany@aliraqia.edu.iq.

Amar A. Mahawish received the Engineer degree in Computer Engineering from University of Technology, Iraq in 2005. He received the Master degree in Computer Engineering from Pune University, India in 2014. Currently, he is a lecturer of Computer Engineering Department, College of Engineering, Al-Iraqia University, Iraq. His research interests include data mining, machine learning, and deep learning. He can be contacted at email: amar.mahawish@aliraqia.edu.iq.