Text Mining: Design of Interactive Search Engine Based Regular Expressions of Online Automobile Advertisements

https://doi.org/10.3991/ijep.v10i3.12419

Ahmed Adeeb Jalal
Al-Iraqia University, Baghdad, Iraq
ahmedadeeb@aliraqia.edu.iq

Abstract—Technology world has greatly evolved over the past decades, which led to inflated data volume. This progress of technology in the digital form generated scattered texts across millions of web pages. Unstructured texts contain a vast amount of textual data. Discover of useful and interesting relations from unstructured texts requires more processing by computers. Therefore, text mining and information extraction have become an exciting research field to get structured and valuable information. This paper focuses on text preprocessing of automotive advertisements domains to configure a structured database. The structured database was created by extract the information over unstructured automotive advertisements, which is an area of natural language processing. Information extraction deals with finding factual information in text using learning regular expressions. We manually craft rule-based specific approaches to extract structured information from unstructured web pages. Structured information will be provided by user-friendly search engine designed for topic-specific knowledge. Consequently, this information that extracted from these advertisements uses to perform a structured search over certain interesting attributes. Thus, the tuples are assigned a probability and indexed to support the efficiency of extraction and exploration via user queries.

Keywords—Information Extraction, Information Retrieval, Natural Language Processing, Text Mining, Web Crawler.

1 Introduction

Text mining concepts revolve around extracting information from textual data that is written in unstructured or semi-structured by means of natural language processing. Unstructured data exceeds 80% of the information that available on the web pages. Therefore, it is more difficult to analyze and not easily searchable. Consequently, there are many challenges to discovering and extracting features and characteristics of heterogeneous data types in different formats [1],[2]. Unstructured data can be textual data (e.g. PDF files, Word documents, and email messages) or non-textual data (e.g. images, videos, and MP3 audio files). Unstructured data is all those data that is not indexed in a database or any other type of indexing.

Text mining term refers to extracting useful information from human language texts by analyzing large quantities of information [3],[4] such as in electronic documents, online web pages, and automobile advertisements. Text mining systems classify and organize the data, depend on lexical or linguistic patterns. Therefore, we try to discover a sequence of one or more keywords to define queries within information retrieval systems in an attempt to extract the information [5]. There are many applications of information retrieval systems such as search engines, plagiarism detection, and social media analytics, which use text mining for opinion mining, and predictive [6]. Accordingly, text mining is a multidisciplinary field such as information extraction, text analysis, and information retrieval.

Text mining dates back to the 1980s by using manual techniques [7]. Since then, digital information storage capacity doubles every 40 months [8]. So, the storage capacity was 122 Exabyte (the sixth power of 1000 bytes) per month in 2017, and it increased to 156 Exabyte per month in 2018 [9]. As Cisco forecasts indicate a steady increase in the storage capacity of the coming years [9]. This growing makes these manual techniques for text mining useless and expensive. So, the dramatic expansion in information volume requires pre-processing that may have a noticeable influence to achieve satisfactory results. Accordingly, text processing tasks are considered as one of the key components to address the documents in many text mining algorithms. Thus, the orientation was toward processing the information automatically during the creation of specific-purpose programs in several areas. Text Mining programs apply in many domains including business, sociology, healthcare, government, education, and research. The development of the programs is making great progress every day, sync with the information technology revolution.

Regular expressions are one of the successful standardization in computer science [10] which greatly utilized in natural language processing. Ordinarily, regular expressions define as another type of language notation that depending on a predefined search pattern [10]. Although, regular expressions are limited but it is a powerful language to describe a several types of formats, protocols, and small textual languages. This practical language uses to search in textual data to return all words or text strings that match the pattern. Patterns usually consist of a sequence of characters. Formally, regular expressions, consider an algebraic notation for determined and characterizing a set of specific strings.

Nowadays, regular expressions are widely used in many scientific applications such as computer science, sociology and biology. For example, network protocol analysis [11], signature scanning for virus detection [12], web mining [13], and characterizing events that led to placing of a child in foster care [14], and finding DNA sequences [15]. Notwithstanding, it is difficult to understand and reuse regular expressions, mainly due to the absence of abstract mechanisms that lead to the growth of regular expressions rapidly. Also, it is very difficult to find the correct regular expression and use it to address a particular problem.

In this paper, we offer a powerful range of independent regular expressions to address the lack of understandability and usability. Moreover, the syntax of the proposed regular expressions provides better coverage while avoiding complexity. Thus, this approach has a significant impact to find useful information and improve search-ability in databases via user queries. Consequently, we proposed a semantic analysis system based on regular expressions to address advertisements pages and users queries. The proposed system guides the users by them queries in the domain of structured database.

Ordinarily, users face many challenges to learning other languages according to accelerated lifestyle. Therefore, one of our goals in this paper is, design a search engine that is able to learn patterns defined by regular expressions. Then, it can recognize and distinguish between Turkish and English languages to discover the user's requirements based on advance learning.

The second section briefly reviews the literature reviews related to extracting and analyzing information through unstructured automotive advertisements and semantic annotation of resources. The third section focuses on the methodology and explains the proposed methods in detail such as, web crawler, information extraction, and information retrieval. The fourth section highlights on the proposed system and the algorithms used to implement it. Finally, this research outlines the challenges of regular expressions and provides a user-friendly search engine to guide the users.

2 Literature Reviews

In this section, we review some examples of applying regular expressions in various domains. Three widely accepted standard methods of machine learning algorithms in language processing to extract a specify string from a document [10], that describe below:

1. Hand-written Regular Expressions (Regex).
2. Classifiers:
   - Naive Bayes is a generative classifier that builds a model for classification tasks. It based on generates some input data with strong independence assumptions between the features.
   - Maximum Entropy Language Models consider a discriminative classifier based on logistic regression classifiers that recognize the features of most useful inputs to distinguish between the different possible classes. Consequently, it uses a lot of features to help predict the upcoming words.
3. Sequence models:
   - Hidden Markov Models (HMM) are a probabilistic sequence model. It calculates the probability distribution for a series of units such as words, sentences, or letters, to choose the best sequence for assigning a label or class.
   - Maximum Entropy Markov Model (MEMM) is a discrimination model that combines features of HMM and maximum entropy model. MEMM offers more freedom in choosing features on successive words by using the class of the prior word as a feature in the classification of the next word.
   - Conditional Random Fields (CRFs) are a type of discriminative undirected graphical model that compute log-linear functions over a clique at each time step. The CRFs are commonly used in conjunction with IE for varied tasks as extracting information from research papers to extracting navigation instructions.

Bhatia et al. [16] built a system to perform information extraction through unstructured automotive advertisements from the Kelley Blue Book website (https://www.kbb.com). This system combines many natural language processing techniques including manual rules, maximum entropy classifiers, and feature engineering. Accordingly, this information is used to fill the relational database with values of interesting attributes. Consequently, they can easily perform structured search over certain attributes that are interesting.

Michelson and Knoblock [17] proposed analysing posts to build a reference sets to classify the advertisements for cars from the Craigslist website (https://www.craigslist.org). Thus, these reference sets can be the kernel to create relational tables of entities and their attributes. For instance, the reference set about cars would include attributes such as a car make, a car model, and a car trim.

Rubens and Agarwal [18] applied a combination of classification algorithms of natural language processing and machine learning to extracting attributes for online automotive which classified on the Craigslist site. The study was motivated by the difficulty one author encountered when trying to find a used car on the Internet. This type of information can be used to facilitate structured search on unstructured data.

Abderrahman et al. [19] proposed a semantic annotation system for information extraction from dispersed online educational resources such as electronic documents. This information is stored in a warehouse that consists of two parts: database and descriptors. The database uses to store learning objects to be easily found by LOM metadata and semantic descriptors. The proposed system based on intelligent agents combining ontology engineering, semantic web and multi-agent systems to improve information retrieval.

As mentioned in the examples above, natural language processing considers a tract of artificial intelligence fields and linguistic processing, so that computers can understand human languages using a range of computational techniques [20]. Most users do not have enough time to learn even some simple sentences for new language because their everyday worries. So, natural language processing emerged to ease the user's work and to satisfy the wish to communicate with the computer in natural language. Consequently, this paper aims to extracting valuable information from the website (https://www.arabam.com) for online automobile advertisements to facilitate the search process for users.

3 Methodology

Text mining involves a variety of techniques such as web mining, natural language processing, information extraction, document classification, and information retrieval [6]. Text mining techniques are used in text analysis to study and discover patterns and their interrelationships. Accordingly, Extraction of high-quality information and relevant data from unstructured data requires developing various techniques based on diverse mathematical and probabilistic algorithms. Some of these techniques aim to improve linguistic, pattern recognition, and mathematical techniques [21].

Unstructured data contain sentences and phrases which represent valuable information. So, text mining must recognize semantic patterns to extract and use this information, which makes the stored database searchable. Structured data depends on creating a database model of the data types that will be recorded, stored, processed, and accessed. The essential goal is to turn unstructured data into structured data (data for analysis) via analytical methods of natural language processing. Because structured data is very easy to deal with and process, so that it is entered, stored, analyzed, and queried at the same time. Thus, structured data reduces the high cost and performance limitations of storage, memory, and processing. As a consequence, it leads to improved machine learning.

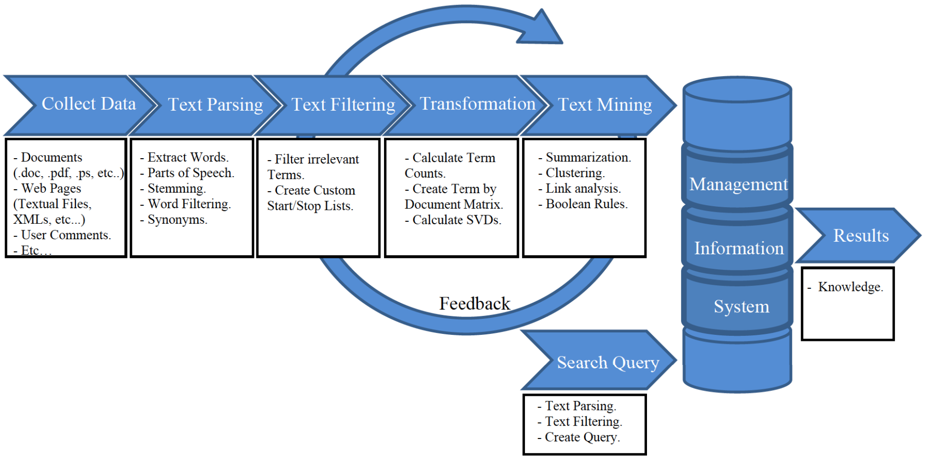


Fig. 1. Text Mining Process Flow [22] and Search Query Process

Two processes are shown in Figure 1: text mining process flow and search query process. Text mining process flow involves an iterative process of analyzing the data collected from web pages by using regular expressions while including and excluding terms for better results. The outcome of this step can be clusters of multiple terms that are stored in the relational database. Also, search query process analyzes user queries to obtain a cluster of multiple terms to perform a database search to discover knowledge. A typical text mining technique involves the following major tasks:

3.1 Web crawler

A web crawler or web spider is software that can guide the navigation of websites automatically [23]. The architecture of crawler consists of a number of components, besides the URL frontier. Ordinarily, a crawler has a certain topic to primarily focus on harvesting pages or has no certain domain. A general purpose of web crawler is providing up-to-date data from relevant pages efficiently. Web crawlers use systematical manner to copy the page's content. The downloaded pages process by a search engine to make it a searchable. Thus, crawlers increasingly consider as a way to address the capability limitations of universal search engines, by distributing the crawling process across users, queries, or even client computers [24].

3.2 Information extraction

An information extraction (IE) is a task of scan and analysis a set of machine-readable documents to extract specific portions of textual data which store in a structured database [22]. Most of the useful information is extracted without a proper understanding of the text [25] such as automobile information. Ordinarily, information extraction serves as a starting point for other text mining algorithms, including question-answering, visualization, and data mining. In the case of a domain specific search engine as in this paper, the automatic identification of important information can increase the accuracy and efficiency of a directed search**.**

3.3 Text summarization

One of the old challenges confronting applications of text mining is text summarization. Since these applications need to summarize large text documents for a brief overview of the topic [26], hoping to remove the details while retaining the keywords. Therefore, text summarization is a process of collecting and producing a brief representation of the original text documents that provide useful information to the user.

The overload of digital information goes beyond our ability to understand it. Therefore, the interpretation of semantics or meaning can be applied to computer literacy to access knowledge and information. Higher levels of computational intelligence, natural language processing, and text summarization require an assessment of the appropriateness of the results to be returned. In general, text summarization techniques can be divided into two types: extractive summarization consisting of basic information extracted from the original text, and abstractive summarization providing new information that can be derived from the original document [27].

3.4 Information retrieval

Information retrieval (IR) is the activity of extracting the relevant patterns associated with a particular set of keywords to satisfy the particular information sought by the user [28]. Therefore, information retrieval often focused on facilitating access to information instead of processing, analysis and summarize of information to discover the hidden patterns. The idea of information retrieval is based on questions and answers to finding the documents which contain answers to the questions. Thus, text summarization stage can be followed by information retrieval that focuses on the user's query. Although, information retrieval is a relatively old field of research, it has gained increasing interest with the emergence of the World Wide Web, which is needed for sophisticated search engines.

Our methodology depends on predefined some regular expressions that will be frequently used in the collection of Arabam postings. These regular expressions were written and drafted in accordance with the Turkish language. Moreover, these expressions, as noted above, are intended to extract valuable information from these growing pages of postings. The problem of classification is defined as the following:

1. Use the web crawler to download a collection of documents D which denote to the training samples of Arabam postings, then D documents can be represented as:

$$D = \{d_1, d_2, d_3, \dots \dots, d_n\} \quad (1)$$

2. Use the predefined regular expressions L which can be the set of distinct words/terms in the documents to discovery the attributes. The terms of regular expressions L can be denoted by:

$$L = \{l_1, l_2, l_3, \dots \dots, l_k\} \quad (2)$$

3. The frequencies of the regular expressions L that appear in the D documents are organized in the structured database. The structured database can be referenced by $W(D, L)$. Where, the D document set represent the rows and the regular expressions set L represent the columns.

$$W(i,j) = \int_{j=1}^{k} L_j \in \int_{i=1}^{n} D_i \quad (3)$$

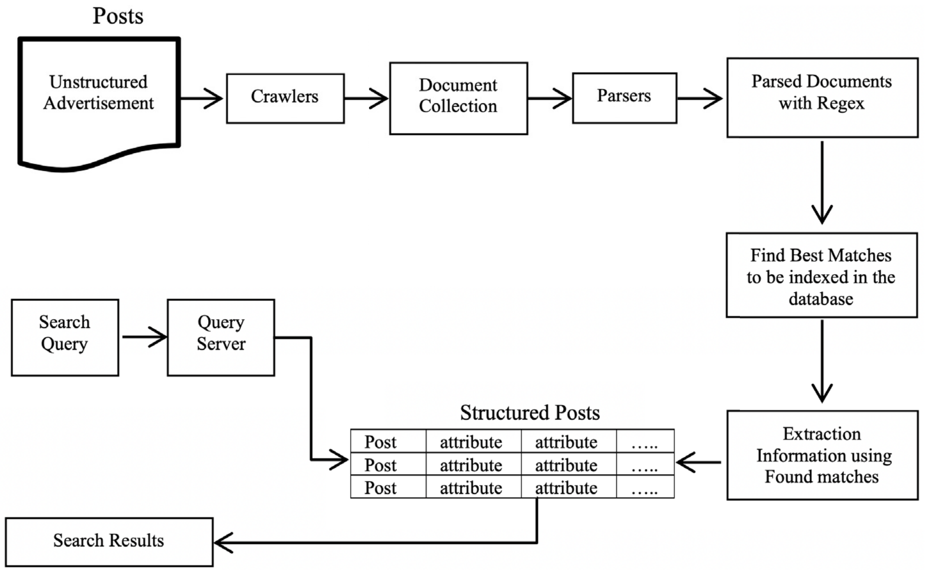

Posts
Unstructured Advertisement
Crawlers
Document Collection
Parsers
Parsed Documents with Regex
Find Best Matches to be indexed in the database
Extraction Information using Found matches
Search Query
Query Server
Structured Posts
Post attribute attribute …..
Post attribute attribute …..
Post attribute attribute …..
Search Results

Fig. 2. Overview of the IE and IR Process [21],[29] with Regex

Figure 2 represents the general schematic of our approach that is explained in Algorithm 1 to build the structured database using Regex match. Now, users can submit search queries to retrieve most relevant information based on query terms, as shown in Algorithm 2. Algorithm 2 shows analysis of user queries to discover the valuable attributes in the structured database.

Algorithm 1 Building the Structured Database using Regex Match

```
Input:
$\vec{\boldsymbol{D}} \leftarrow < d_1, d_2, \ldots., d_n >$ // $\vec{D}$ is the vector of the postings pages.
$\vec{L} \leftarrow < l_1, l_2, \ldots., l_K >$    // $\vec{L}$ is the vector of the regular expressions.
Output:
Regex Match Vector $\boldsymbol{RMV}$.
for $i$=1:$n$
    for $j$ =1:$k$
        if $l_j \in d_i$
            $W(i,j) \leftarrow RMV$ // Organize the matching attributes in the Database.
        end
    end
end
```

Algorithm 2 Text Analysis for Structured Search Process

```
Input:
Search Query Attributes $\vec{\boldsymbol{S}} \leftarrow < s_1, s_2, \ldots., s_a >$ // Analyze User Query to Discover Valuable Attributes.
Structured Database $\boldsymbol{W}$.
Output:
SQL Statement $\boldsymbol{SQL_{ST}}$
Search Results $\boldsymbol{R}$.
for $i$=1:$a$
    $\boldsymbol{SQL_{ST}} \leftarrow s_i$. // Build SQL Statement Based on Search Query Attributes.
end
$\leftarrow \boldsymbol{SQL_{ST}} \in \boldsymbol{W}$
```

4 Results and Discussion

As mentioned earlier, computerized text mining approaches are currently being applied in a variety of industries to discover knowledge and patterns in unstructured data. Regular expressions define a search pattern by a sequence of characters to matching with textual data. Regular expressions use in many search engines. Also, several programming languages offer the capabilities to write a regular expression such as Java, Python, and C++. Natural language processing applications typically use regular expressions that developed manually by human experts.

There are many regular expressions formats and many extensions of basic expressions as in [30]. Also, the internet regular expression library (http://www.RegExLib.com) indexed 17328 expressions from 2730 contributors around the world, which gained widely popular for simplicity of syntax. Nevertheless, regular expressions are not without weaknesses. Consequently, techniques are sought to validate regular expressions or test their use in applications to exposing possible faults contained in the regular expression [31]. Those weaknesses can be outlined into three principal shortcomings. These problems, make reuse of regular expressions somewhat difficult.

1. **Complexity:** Some of regular expressions in the repository of (http://www.RegExLib.com) exceeded 4000 characters, and the more complex expressions contained more than ten overlapping levels. So, it is very difficult for users to understand what these long sequences of expressions do due to its sheer size and almost impossible to verify it.
2. **Errors:** There are many of regular expressions scattered on the internet sites, which suffer little lapses here and there and not always. Ordinarily, these lapses or faults are so accurate and cannot be easily to detect it by users.
3. **Version Proliferation**: Because many different versions of regular expressions are developed and stored in different repositories, which perform the same purpose. So, finding and choosing the right version for a particular task is difficult.

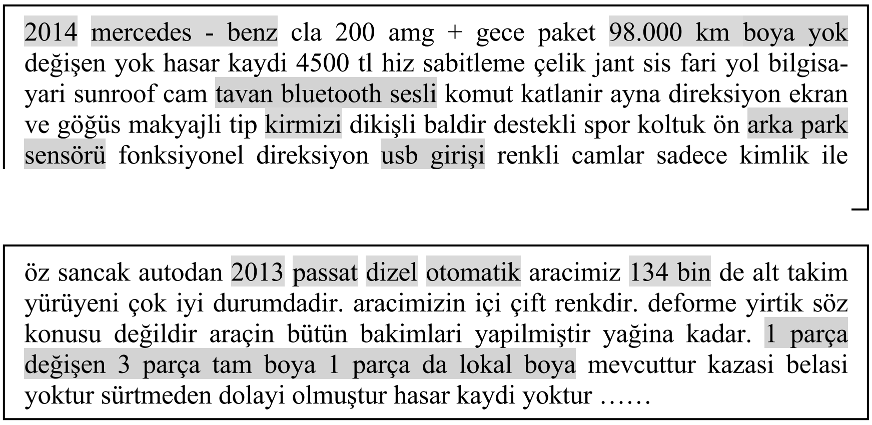
Those three problems mentioned, make it extremely difficult to guess the motivation behind regular expressions syntax. Accordingly, without this knowledge, it is impossible to judge the validity of regular expression syntax. Thus, it is difficult to determine whether this regular expression should be reused in order to choose it from a variety of regular expressions in repositories.

Our goal in this paper is to automate discovery of knowledge by text classification through the creation and utilization of regular expressions. These regular expressions will be powerful enough to extract information from web pages by specifying a set of strings required for a particular purpose.

Some languages such as Turkish (research domain) have many suffixes, so word tokenization becomes more difficult. Consequentially, we need to design novel regular expressions to separating out or tokenizing words to describe patterns of the texts that were discovered. Regular expressions can be used to specify words that we might want to extract from a textual data. Accordingly, regular expressions play an important role to converting textual data into a more convenient standard model. Moreover, we will analyze user-generated queries to providing explanations with more descriptive semantics to exploit it for searching repositories. Consequently, the search process, based on user queries, will be easy and available through the application we design. This application provides the ability to query by specifying desired properties, names, or strings which achieve the best results with maximal efficiency, accuracy and speed.

Our dataset consist of automobile advertisements that collected from the website (https://www.arabam.com). We downloaded the posts preserving HTML format. The domain and attributes that we chose presented many challenges in determining which candidate values that will be interesting by users. As it is known, it is difficult for those arriving in Turkey to learn the language in a timely manner, as well as other international languages. Consequently, this paper aims to provide a system to facilitate the search process within the ads that available on the website in English. Moreover, we made sure to provide the search query in English and Turkish with the possibility of writing and drafting question.

In the following examples that shown Figure 3, our system should specify some specifications attributes such as whether the car has dyed parts, does it have Bluetooth, or does it have sensors. Even though, these attributes are not often mentioned in details but the proposed system helps to analyze masculine description of the car to find some important features.



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

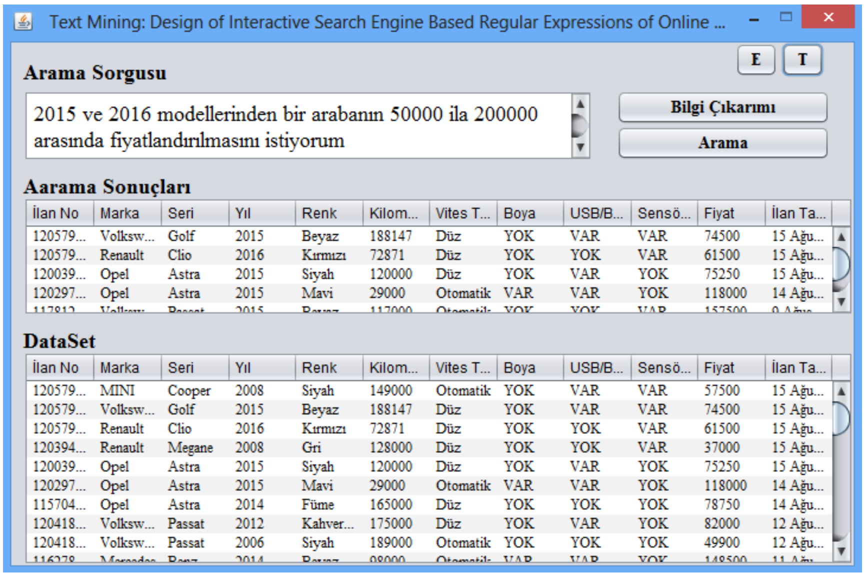
öz sancak autodan 2013 passat dizel otomatik aracimiz 134 bin de alt takim yürüyeni çok iyi durumdadir. aracimizin içi çift renkdir. deforme yirtik söz konusu değildir araçin bütün bakimlari yapilmiştir yağina kadar. 1 parça değişen 3 parça tam boya 1 parça da lokal boya mevcuttur kazasi belasi yoktur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

Fig. 3. Example of Attributes Detection

We designed an interactive search engine based on regular expressions of online automobile advertisements. The principle of this work is to extract structured information from online advertisements to conduct the search and query process. We have applied NLP techniques to avoid the language problem and extract meaningful features from Arabam postings. This system uses regular expressions to find candidates for extraction instead of generalization patterns.

The information that extracted and stored in database was in Turkish language. Accordingly, we train the proposed system with regular expressions in Turkish and English languages.

As shown in Figure 4, we test two types of user search queries in Turkish and English languages. The first query in Turkish, based on regular expression matches, we discover the important attributes which are car's manufacturing year and price. Consequently, the SQL statement is formulated with these detected attributes to get search results, according to Algorithm 2. The second query in English, after matching the text analysis we find the distinctive attributes which are car's model and gearshift. The SQL statement is formulated, in the same manner.


Text Mining: Design of Interactive Search Engine Based Regular Expressions of Online ...
E
T
Bilgi Çıkarımı
Arama

Arama Sorgusu

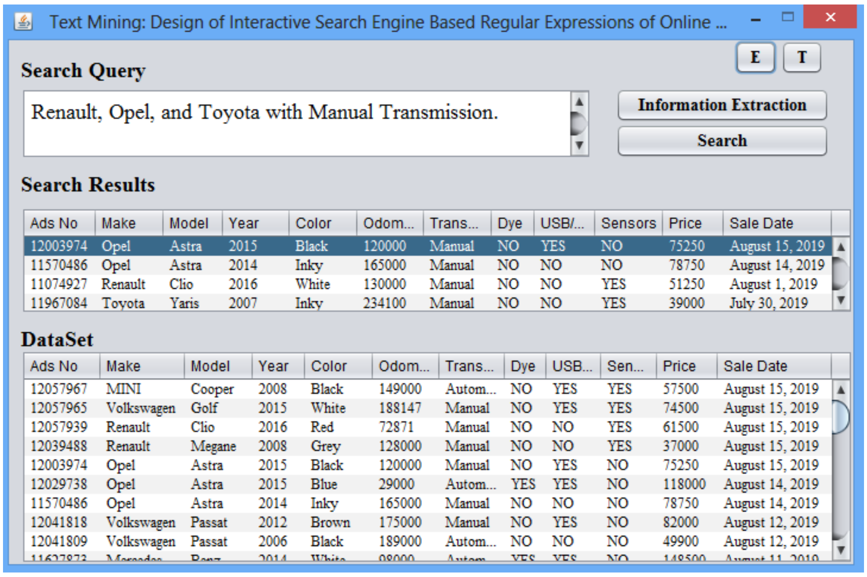
2015 ve 2016 modellerinden bir arabanın 50000 ila 200000 arasında fiyatlandırılmasını istiyorum

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...


Text Mining: Design of Interactive Search Engine Based Regular Expressions of Online ...
E
T
Information Extraction
Search

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

Fig. 4. Text Analysis for Structured Search Process

Our education never stops, whether in the classroom, the lecture hall, the normal everyday interactions of life. However, learning is an interactive environment for sharing knowledge and cultures, usually with specific goals in mind. Many aspects of educational dimensions are covered in researches through suitable tools. Such as evoking empathy through games [32], the concept of knowledge and game-based learning from the Gamification concept [33], and online discussions prompts [34]. In the same vein, we are trying to test a novel language-based approach through semantic analysis and translating of user queries. One of the goals of this approach is to create an interactive environment in several languages, taking advantage of the fondness and follow-up of users to online automobile advertisements to teach them some vocabulary in different languages. Consequently, we will combine between users' passion for cars and learning.

5 Conclusion

This paper provides a successful system for accurately extracting information from online classifieds. The current design of our system uses a very simple technique to extract information from web pages. Accordingly, this approach relies on regular expression in natural language processing to extract valuable information from posts. Consequently, we have made significant progress in attribute-based research. Nevertheless, we aim to include a rest of the features to make the system more useful for users and facilitating search over a wider range of attributes. This can help improve the recall for the attributes. Moreover, we make the system more comprehensive through add other automobile advertisements sites in different languages.

6 References

[1] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” Journal of Business, Vol. 70, pp. 263-286, 2017. https://doi.org/10.1016/j.jbusres.2016.08.001

[2] I. Taleb, M. A. Serhani, and R. Dssouli, “Big Data Quality Assessment Model for Unstructured Data,” in Proc. of the 2018 International Conference on Innovations in Information Technology (IIT), 2018. https://doi.org/10.1109/innovations.2018.8605945

[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. K. Hanify, S. Ayeshaz, and F. Fatimax, “Text Mining: Techniques, Applications and Issues,” International Journal of Advanced Computer Science and Applications, Vol. 7, No. 11, pp.414-418, 2016. https://doi.org/10.14569/IJACSA.2016.071153

[5] M. W. Berry and J. Kogan, Text Mining: Applications and Theory, Wiley, Chichester, UK, 2010.

[6] W. He, “Examining Students Online Interaction in a Live Video Streaming Environment Using Data Mining and Text Mining,” Journal of Computers in Human Behavior, Vol. 29, No. 1, pp. 90–102, 2013. https://doi.org/10.1016/j.chb.2012.07.020

[7] J. C. Scholtes, Text-Mining: The Next Step in Search Technology, DESI-III Workshop Barcelona, 2009.

[8] M. Hilbert and P. López, "The World's Technological Capacity to Store, Communicate, and Compute Information," Journal of Science, Vol. 332, No. 6025, pp. 60-65, 2011. https://doi.org/10.1126/science.1200970

[9] T. Barnett, S. Jain, U. Andra, and T. Khurana, "Visual Networking Index," Cisco, 2018.

[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, 2018.

[11] J. Wang, K. Cui, K. Zhou, and Y. Yu, "Based on Regular Expression Matching of Evaluation of the Task Performance in WSN: A Queue Theory Approach," The Scientific World Journal, 2014. https://doi.org/10.1155/2014/654974

[12] N. L. Or, X. Wang, and D. Pao, "MEMORY-Based Hardware Architectures to Detect ClamAV Virus Signatures with Restricted Regular Expression Features," IEEE Transactions on Computers, Vol. 65, No. 4, pp. 1225-1238, 2016. https://doi.org/10.1109/tc.2015.2439274

[13] S. Zhang, L. He, S. Vucetic, and E. C. Dragut, "Regular Expression Guided Entity Mention Mining from Noisy Web Data," in Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1991–2000, 2018. https://doi.org/10.18653/v1/d18-1224

[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, 2017. https://doi.org/10.1093/bioinformatics/btx413

[16] N. Bhatia, R. Kumar, and S. Senapaty, "Extraction of Structured Information from Online Automobile Advertisements," Stanford University, 2008.

[17] M. Michelson and C. A. Knoblock, "Constructing Reference Sets from Unstructured, Ungrammatical Text," Journal of Artificial Intelligence Research, Vol. 38, pp. 189-221, 2010. https://doi.org/10.1613/jair.2937

[18] M. Rubens and P. Agarwal, "Information Extraction from Online Automotive Classifieds," Stanford University, 2002.

[19] C. Abderrahman, O. Aziz, and K. Mohamed, "Semantic Annotation of Resources of Distance Learning based Intelligent Agents," International Journal of Engineering Pedagogy, Vol. 4, No. 1, pp. 69-72, 2014. https://doi.org/10.3991/ijep.v4i1.2845

[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, 2016. https://doi.org/10.1016/j.inffus.2016.10.004

[21] L. Kumar and P. Kalra Bhatia, "Text Mining Concepts Process and Applications," Journal of Global Research in Computer Science, Vol. 4, No. 3, 2013.

[22] G. Chakraborty, S. Garla, and M. Pagolu, "Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS," SAS Institute, 2013.

[23] C. Olston and M. Najork, "Web Crawling," Journal of Foundations and Trends in Information Retrieval, Vol. 4, No. 3, pp. 175-246, 2010. https://doi.org/10.1561/1500000017

[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] V. Gupta and G. Lehal, "A Survey of Text Mining Techniques and Applications," Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, 2009.

[26] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining,” Journal for Computational Linguistics and Language Technology, Vol. 20, No. 1, pp. 19-62, 2005.

[27] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, “Text Summarization Techniques: A Brief Survey,” International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, pp. 397-405, 2017. https://doi.org/10.14569/ijacsa.2017.081052

[28] 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 Proc. of the 2017 IEEE Symposium Conference on Privacy-Aware Computing, 2017. https://doi.org/10.1109/pac.2017.35

[29] M. Shahbaz, P. McMinn, and M. Stevenson, “Automated Discovery of Valid Test Strings from the Web Using Dynamic Regular Expressions Collation and Natural Language Processing,” in Proc. of the 12th International Conference on Quality Software, pp. 79–88, 2012. https://doi.org/10.1109/qsic.2012.15

[30] J. E. Hopcroft, R. Motwani, and J. D. Ullman, Introduction to Automata Theory, Languages, and Computation (3rd Edition), Addison-Wesley Longman Publishing Co., 2006.

[31] P. Arcaini, A. Gargantini, and E. Riccobene, “Fault-based test generation for regular expressions by mutation,” Journal of Software: Testing Verification and Reliability, Wiley Online Library, Vol. 29, No. 1-2, pp. 1-22, 2018. https://doi.org/10.1002/stvr.1664

[32] C. Papoutsi and A. S. Drigas, “Games for Empathy for Social Impact,” International Journal of Engineering Pedagogy, Vol. 6, No. 4, pp. 36-40, 2016. https://doi.org/10.3991/ijep.v6i4.6064

[33] K. Puritat, “Enhanced Knowledge and Engagement of Students through the Gamification Concept of Game Elements,” International Journal of Engineering Pedagogy, Vol. 9, No. 5, pp. 41-54, 2019. https://doi.org/10.3991/ijep.v9i5.11028

[34] L. B. Bosman, N. Duval-Couetil, B. Mayer, and P. McNamara, “Using Online Discussions to Develop the Entrepreneurial Mindset in Environmental Engineering Undergraduates: A Case Study,” International Journal of Engineering Pedagogy, Vol. 9, No. 3, pp. 4-19, 2019. https://doi.org/10.3991/ijep.v9i3.9491

7 Author

Ahmed Adeeb Jalal is a Computer Engineering lecturer. He works at Computer Engineering Department, College of Engineering, Al-Iraqia University in Iraq. He received B.S. degree in Software Engineering from Al-Rafidain University College, Iraq. Additional to, he received Master's degree in Computer Engineering from Yildiz Technical University, Turkey. His research interests include data mining, hybrid recommendation systems design, and web applications. Email: ahmedadeeb@aliraqia.edu.iq

Article submitted 2019-11-18. Resubmitted 2019-12-29. Final acceptance 2020-03-11. Final version published as submitted by the authors.