International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

New Approach Exploring Unclear Weighted Association Rules Using Weighted Support and Trust Framework by using Data Mining

Bassam Talib Sabri

Submitted: 15/09/2022 **Accepted:** 24/12/2022

***Abstract*—** Automatic classification of text is one of the important applications and search subject ago the Foundation of digital document. Text classification is necessary and the reason excessive number of document text handled daily. This study mainly aims to integrate data mining techniques with classification tasks to build a web text classification (TC) method by 1) influencing the full amount of information contained within web documents for classification and 2) increasing the effectiveness of search processes to identify similar or related information for users. Hypertext classification differs from traditional TC; information beyond web document contents, such as metadata, is a useful source of information. The unification of metadata with text using simple combination methods effectively improves classification performance. Results show that these classifiers are common, accurate, and perform effectively. The accuracy measure is 91%.

***Index Terms*—** Rule Mining, Support Threshold, Term weighting theories, Text categorization, tf–idf

I. INTRODUCTION

With the amount of data growing exponentially, the world wide web (WWW) have become the world's largest and fastest growing information source. According to most predictions, most human being information will be available online in the near future. These huge amounts of information raise the challenge of transforming the web into an increasingly useful information utility. In different context, Trained professionals are used to classify what's new. This process is time consuming and expensive, Limit of viability and resulting in an increase interesting in developing techniques or methods for automate text classification (TC). The growth availability of information source, such as the WWW, Produced the actual need approaches to filtering information. To search engines, directories, and crawlers, such as Yahoo!, users query's is merely a list of keys words and operators. The result a list of pages that is based on the pages' similarity to the query, and keyword-based searches are provided without a full understanding of the content, from which the main part of the information in web documents is obtained. Some www page can't offer just limit assistance to client. Such as, www searching base on key word can back thousands of indicators on the www pages that contains the key word. However, these indicators are nothing to do with the user intended search.

TC, which involves assigning text documents for one or more predefined classification on the basis of their contents, is important components in many information managements tasks. Document classification has become an important task and explore field for searching. Querying massive numbers of documents is a control issue in information retrieval systems that deal with text data. An automatic classification system for documents involves assigning each document label a class from a predefined set classes, which are based on examples or training sets of reclassified documents, to drive a www paper categorized sachet. The classifier is use to categorize new paper. Data mining is typically used for large, highly structured information databases to discover new knowledge. The results of this study show the importance of developing a methodology using data mining techniques to discover effectively integrated knowledge in documents with the benefits of text collection and some of the structured information available in these documents. This search concerns the automatic domain web TC, which integrates information retrieval techniques, word processing with data extraction, machine learning, statistics [1], [2], [3], [4].

Department of Business Information Technology / University of Information Technology and Communications
Baghdad, Iraq
bassam.ali@uoitc.edu.iq

A. *Text Processing*

Through nearly any measurement, the number of information created grows more rapidly from the capability of customers to find and usably this information. organize store an increase number of their internally generate information in accessible-computer form, thereby as well as the opportunities and problems arising from the growth of information. Textual information cannot be effective understood because the relation ship between the sequence of words and content is unclear. Textual information including technic research, memory, manual Ing, electric mailing, documenting, news papers, others forming of texting. In additional, speech, images, videos through textual notes, and other forms of data are also accessed. Contenting-base word process tasks can be dividing into three larging sets.

1. TC involves assigning documents or parts thereof to one or more categories.
2. text understanding involved completing reach to the contenting of word, example extract format information, answer questioning, and summarize or abstract. This researching addressing the second groups [5], [6].

TC is the task of assigned documents presented in natural languages into one or more classes that belong to a predefined set or assigning text word to define categories based on text content. appointment subject labels to documents is one of the many general uses of supervise learn for text [7], [8], [9].

Classification aims to assign a category from a predefined set to an known paper. A sampling case of a categorization problem Ing is the evaluate of topic for a numeric library's. In a explore document, for example, a Regine select the appropriation depend subject that defining the topic of the document. In a digital library, some simple rule base algorithms can be analyzing previous word papers for speech can be using to predicting the suitable topic matters. If a papers including such terminology's as "computing," "algorithms," "accuracy," or "programing," be categorizing as computer Ing systems search newspaper [10].

B. *WWW*

The WWW is a large, hyperlinked, dynamic, global information system; it massive collections of completing un controlling heterogeny nous document is the biggest and widely known repost of hypertext. Hypertext word documents containing text and generally imbed hyperlink to other word document distributed throughout the web. Know, the web including billions of word documents distributing over millions of computers that are connecting by telephone line, optical fiber, and radio modem.

The internet has large numbers of unstructured text based word documents. A prediction regarding future database research states that most data will be usable on the internet within 10 years. The internet covers all widely useable sites and corporation intranet and repositioning [11], [12].

1) *HYPERTEXT MARK-UP LANGUAGE (HTML)*

The main count of web documents is created with HTML, a language that allows the full use of hypermedia, including text, images, graphics, databases, sounds, and other types of multimedia. HTML requires a browser, which is a special software used to access the web. Approximately 1 billion HTML pages are present today, 1 million pages are created per day, and over 600 GB of pages change per month. HTML format files allow these pages to be determined and made available on the internet as fully search documents. Average search engine go over other type, such as PDF files, without indexing them. HTML files have the advantage of accessibility through web browsers and can be obtained and search by any major internet search engine [13], [14].

2) *Current Search Tools*

Useable search tools on the internet belong to two classes, namely, internet searching reengining. Internet direct, as that providing by Hotmail! provided a hierarchal dividing of document; each paper in the directories is associating with a stop of a trees (a leaf's or an internal stop). Proceeding tall the trees, a users accessing a sets of paper that has manually reclassified and located in the branch. Such as, Hotmail! comprises a classifying tree with a deep of 11 or more (depend the path following). Approximately 0–31 branches in each's levels of the branch guide to a lot of a little hundred Ing of 1000s of papers. look in a internet is remarkably conveniently and always lead the users sets of word document sought. However, this search goes to only a small divide of the network. This set covered stems from the slow manual categorization. Search engine, example Alta Vista was a Web search engine and Excite, The user must manually browse the documents to finding the paper(s) required. search engine offering "advancing" searching that enabling Booleans merging of searching term to improved the precisions of the look [15], [16], [17].

Organizations such as educational institutions, government offices, and private companies call for a specific use of search engines for their own privacy, known as web site search engines. These software systems have become increasingly important by providing detailed information that internet (WWW) search engines cannot offer; they have better accuracy

results and less irrelevant information, thereby minimizing time and cost in searching only private web sites.

Index-based web explore engines suffer from the following deficiency.

First: topic may contain hundreds or thousands of documents.

Second: Many documents applicable to the topics may not contain key words that define them. For example, the key word “data mining” may be in much pages relate to other forms of mining, and many related web pages that contain “statistics,” “knowledge discovery,” or “machine learning” cannot be identified because they do not contain the keyword “data mining.”

WWW services provide key word based searched without understand the full content. Web pages can offer only limited help to client; in some cases, the search results comprise hundreds of unrelated documents [18], [19], [20], [21].

C. Literature Survey

Automated text classification has been very an curacy function and looking for lecture since the foundation of numeric words paper. This technology is presently necessary because of the extremely large number of text word documents handled daily. Many algorithm and techniques for automatic text categorization have been prepared and proposed in the literature. Lots word text classifiers have been proposed by researchers use machine learn technique and probability models. These classifiers often different in the approaching adopte, namely, decided trees, NaiveBayes, rules induction, neutral’s net works, nearest neighbor, and supports vectors machining (SVM).

The apple of datamining for build categorization modeling is news. recently study propose using associational rule in build dividing for numeric information. The categorization system explorer strong connotation rule in databases and apply classifiers. Such methods are summarized in the following section.

Maron method: Categorize founded on probability replicas have been propose, start to be present in the works by Maron’s in 1962, which was followed by naive Bayes, which performs effectively [22],[23], [24], [25].

Associative General Classifiers, CMAR, and CBA: In addition to these classification methods, recently, a newly technique that build associatory general divided has been propose. state, method is representing by associate law Minnie.

primary knowledge behind Ing this approaching is determine stronger designs are association the session markers. The following stage is to income advantaging of designs that a classifyer is developing recently items are considered appropriate modules. Dual of the reproductions, namely, CMAR [35] and CBA [26], have been presented in the literature. Although both models are effective and have high accuracy, they certain limited. bother models perform only single class classifications and have not been implement in word text categorization. However, in many functions (particularly word text categorization), multiple class classification is requiring. In the present study, we attempt to overwhelmed limited build an associatory categorized typical that lets only manifold lesson classifications of word manuscript papers.

D. Data Mining

A large amount of machine-readable data has been provided in recent years in the form of files and databases; web texts are important usable files. Meanwhile, data mining is a popular area of interest that refers to operation on information along multiple dimensions; it mines or discovers additional information in terms of forms or rules from that vast amount of data. Datamining procedure of discovery designs in information and is relevant to data retrieval (IR) and acquaintance sighting in catalogues (KDDs). This procedure is a collection of different functions, namely, sequential analysis for time dependent data, link analysis that attempts to set relationship among data, summarize that describes subsets of data sets by compute medians and accept deviations, categorization that maps data sets to one or more pre defined classes, and cluster analysis that, like to classification, groups data sets into cluster use similar metrics.

Data mining (DM) usually deal with structure data, but text is usually fair un structured. The heart of the word text mining problem can then be viewing as impose structure on text to make it agree able to the analytic technique of datamining. The problem is conceptual as extract meta data from word text. earlier data mining technique can be apply, the word documents should be preprocessed to create a document index with frequency and weight of the position of the document terms (title, abstract, keywords, or body of the text) [27], [28], [29].

II. MINING WEB TAXONOMY

The taxonomy of web mining is founded on the part of the internet to be mined, and it comprises the three areas shown in Fig. 1 [30], [31], [32].

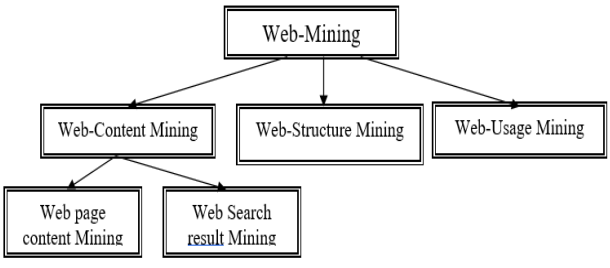

Fig. 1. Catalogue of Mining Network.

A. *Content -Web Withdrawal*

Content Network withdrawal the discovered of useful information from internet documents. Web content comprises numerous kinds of information, example as text, images, auditory, filmed, meta data, and hyper links. Search on mining multiple different types of data is now term as multi media data mining, which can be considered a form of content web mining. Content web information contain of un structured data word text, structured trailer truck- data as HTMLS document, and highly organized data as table and information base-generate HTMLS pages. Content web- mining mainly aims to help in or improve information search or filter information. Queries that are more sophisticated than keyword-based search can be submitted when developing a new modeling of data on the internet. Content web- mining is divide into two type, namely, content web page mining and result web mining [33], [34], [35]. The heterogeneity of the WWW and the absent of structuring have led several research to mine sub sets of known word documents or data from document that are relevant to a assumed subject. Single sub set examination effect of a enquiry directed to a pursuit machine, such as Google or Alta's Vistas. Scheme use exploration engine to retriever pertinent word papers and collection data after the word permits or facts providing by server, such as the uniform resource locator (URLs), label, type content, and modificative type [36], [37], [38].

B. *Structure Web- Mining*

Structure web- mining attempts to detect the model under lying the link structure on the web of internet. The modeling is base on the topological of hyper links with or without description. The modeling can be use to classification web internet sheets is suitable for produce info, as the comparation connection among net internet locations [39], [40], [41].

C. *Usage Net- Mining*

Internet net practice withdrawal attempts to understand the data generate by the internet web behavior. Content web- and structure web- withdrawal apply actual or main facts on internet net. By contrast, tradition web- withdrawal pits minor facts, which are derive after the behave of operators while interact the internet net. The information includes those from server net entree log, substitution servers log, fuels browsers, client outlines, registrations facts, users' meetings or transactional, cookie's, data book mark, and other facts derive from a person interaction with the internet web [42], [43], [44].

III. DESIGN OF AUTOMATIC WEB TEXT CLASSIFICATION (TC)

This chapter discusses a practical solution proposed for accomplishing automatic web TC through data mining association rules. The work is partitioned into the following steps: document collection, document preprocessing, indexing, constructing document vectors, classifier construction, and class prediction. HTML documents occupy the main part of web pages. HTML important features approach is presented to handle the issue. The proposed classifier is designed in three steps, namely, training, testing, and application. Each step consists of many stages or modules, such as lexical text analyzer, string tokenizer, stop word eliminator, and term stemming module, which play important roles in the project. A stemming algorithm is proposed for high-precision results. The final stages include building a categorize and the predict module for new word documents.

A. *Automatic Web TC*

A newly classification method is proposed for web text. This technique taking advantage of HTML features and data mining tasks. The proposed method incorporates the associate rule mining task with the categorization problem. The following sections of this chapter clarify the tasks of the proposed method. Fig. 2 depicts a simple overview of the processes and tasks.

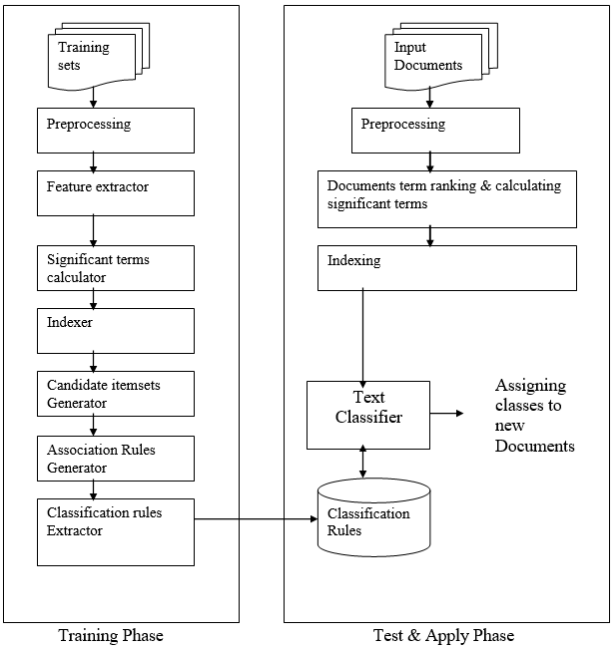

Fig. 2. Overview of Processes and Tasks.

The five processes of the proposed method are as follows.

1. DATA COLLECTION PHASE
   - Downloading of web document
   - Document cleansing
2. PREPROCESSING PHASE
   - Lexical text analysis and string tokenizer

– Stop word elimination
– Simple stemming method (SSM)
– Term tokenizing

3. TERM INDEXING PHASE

– Feature extraction
– Document term ranking
– HTML document ranking (HDR) method
– Term pruning

4. CLASSIFIER CONSTRUCTION PHASE

– Association rule generation method
– Association rule pruning (ARP) technique

5. CLASS PREDICTION PHASE

– Optimal association-based classifier rules (OABCR)

For further explanation, Fig. 3 depicts the architecture of the proposed method of the five mining phases.

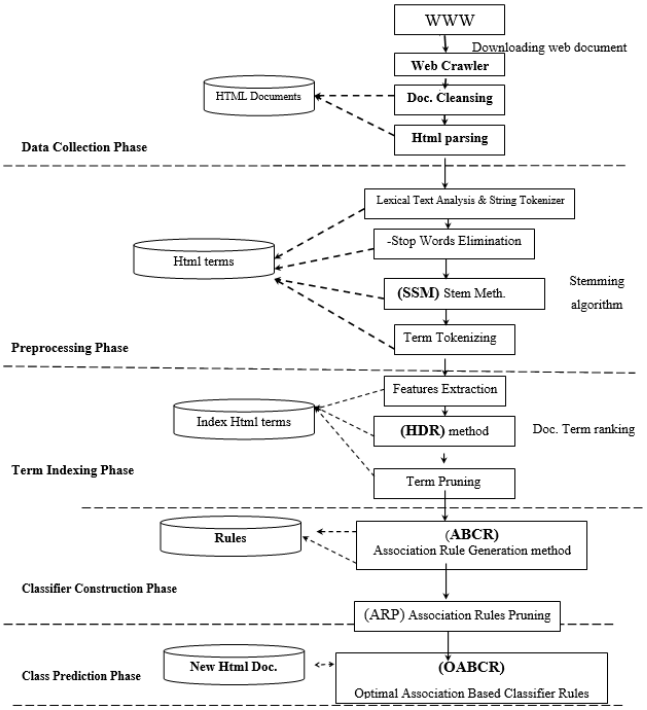

Fig. 3. Web Text Analysis and Classification Architecture.

B. *HTML Document Feature Extraction*

The basic structure of web HTML pages is as follows.

<HTMLs,>

<heads>

<titles> Titles that is displaying highest of the browser net

</titles>

<metals names="explanation" contents ="explanation of web location speak by examination machines">

<metals names="keyword" satisfied ="key key words of web location read by search engines">

</head>

<body>

This is a new web page please visit …..again.

<p/>

<body>

</html>

The </html> tag informs the browser where the html starts. <title> notifies the browser of the address; it is at the top page and used for indexing the page. The </meta name> information data is useful for certain search engine. The “description” tag is used to describe a site. The “keywords” tag is useful for page ranking in several engines. The <body> tag contains the body of the web site or web document. <a href contains the URL, the web site address, anchor link, or hyperlink address. Fig. 4 depicts a fragment of an HTML source code.

```
!>DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN<"
 --!>saved from
url=(0075)http://www.sswug.org/searchresults.asp?keywordstofind=database%20management<--
>HTML><HEAD><TITLE>Help for database management - SQL Server, Oracle, XML,
DB2</TITLE<
>META http-equiv=Content-Type content="text/html; charset=windows-1256<"
>META
content="sql server database, sql server, scripts, data warehousing, olap services, sequel server, sql
scripts, how-to, beginners, SQLSERVER, swynk.com, swynk, sqlservercentral, sequel server "
name=KEYWORDS<
>META
content="SQL Server help and articles, newsletter, list servers, discussion boards, scripts, Oracle
database integration, SOAP links and articles, XSLT articles, XML help, daily news and more ".
name=description<
>META http-equiv=Pragma content=no-cache<
>META content=noarchive name=robots><LINK
href="Help for database management - SQL Server, Oracle, XML, DB2_files/discuss.css "
type=text/css rel=stylesheet><LINK title="SSWUG Articles "
href="http://feeds.feedburner.com/SswugorgArticles" type=application/rss+xml
rel=alternate><LINK title="SSWUG Developer Articles "
href="http://feeds.feedburner.com/sswug/NzDN" type=application/rss+xml
rel=alternate><LINK title="SSWUG DB2 Articles "
href="http://feeds.feedburner.com/sswug/keNv" type=application/rss+xml
rel=alternate><LINK title="SSWUG MySQL/Open Source Articles "
href="http://feeds.feedburner.com/sswug/MfLp" type=application/rss+xml
rel=alternate><LINK title="SSWUG Oracle Articles "
href="http://feeds.feedburner.com/sswug/dVyS" type=application/rss+xml
rel=alternate><LINK title="SSWUG SQL Server Articles "
href="http://feeds.feedburner.com/SswugorgSQLArticles" type=application/rss+xml
rel=alternate><LINK title="SSWUG XML Articles "
href="http://feeds.feedburner.com/sswug/AEMr" type=application/rss+xml
rel=alternate><LINK title="SSWUG Editorials "
href="http://feeds.feedburner.com/SswugorgEditorials" type=application/rss+xml
rel=alternate><LINK title="SSWUG Newsletter via RSS "
href="http://feeds.feedburner.com/SswugorgNewsletterFeed "
type=application/rss+xml rel=alternate<
>SCRIPT language=JavaScript<
--!>
function Wpopup(URL,w,h} (
day = new Date;()
id = day.getTime;()
eval("page" + id + " = window.open(URL, '" + id + "',
'toolbar=no,scrollbars=no,location=no,statusbar=no,menubar=no,resizable=no,width=' + w + ',height='
+ h;(";(" +
{
//if (parent.frames.length > 0} (
    //parent.location.href = self.document.location;
{//

function Pcertify ()
}
function Rcertify ()
```

Fig. 4. Example of HTML Page Source Code.

C. *HTML Feature Analysis*

A strategy must be set to consider the requirements of the lassification task and thus achieve useful text features from existing HTML and natural language properties.

A quantitative analysis is performed to describe the nature of the content to be classified. The results are obtained by analyzing a collection of 100 HTML documents. The best quality and amount of features are extracted according to the tags of the HTML documents. The selection of rich feature location in the document is substantially important for the classification process. A series of examples of web documents is presented. Stand 1 HTML documents with a sure total amount of arguments for apiece kind of HTML label. The amount of disagreements is counted gratified attributes. The tag is also counted without considering the other tags

because of text weakness.

The results are shown in Table 1. In the following diagram, 82% of the test documents contain 1–20 words in the <TITLE>, and 6% of these documents have no title. The <Meta=description> tag exists in 30% of the documents. A total of 70% of the documents have no such tag and therefore have zero words. A total of 15% contains (21–50) words. The <Meta=Keywords> tag exists in 15% of the documents; 85% have no tags and thus have zero words. A total of 10% contains 21–50 words. No dependency exists on the <Meta=Keywords> and <Meta =description> tags, given that the HTML documents have a small word percentage. The main amount of text can be extracted and laid on the <BODY> tag. A total of 17% of the words are in the <BODY> tag, which contains 21–50 words; approximately 66% of the words are in the <BODY> tag, which contains more than 51 words. The body tag can be the main text source in HTML documents. These results reflect the analysis of the experiment samples, which are evidently small and may not be a complete reflection of other web document formats or types. However, this example can be generalized because it includes text-based web documents for classification purposes.

IV. METHODOLOGY

To explain the design of the proposed web TC, it should be partitioned into five phases as follows.

- Data (HTML document) collection
- Data preprocessing
- Term indexing: constructing document vectors
- Classifier construction
- Class prediction

A. Collection Data

The information data collection and preprocessing phase involving the following steps.

1. Downloading the web document: This step involves downloading web document files according to a selected computer science domain to develop a document corpus. Web document files are downloaded from known search engine web sites, such as Yahoo! Directory and Google. The documents are stored at database files designed to be a repository for the data corpus.
2. The registers from the crawl are cleansed to remove non-HTML files because the proposed classifier works on HTML pages only.
3. The downloaded document text is stored in a pretreated database in the text extraction process. Fig. 5 explains the HTML text extraction process from downloaded files.

The final step of the data collection process is splitting the downloaded HTML files into training and test sets; the HTML files should be randomly selected for each set of data.

Algorithm 1: HTML Text Storing Procedure
Procedure Name: Htm-text-store
Input: directory contains html files
Empty database file with text attribute
Output: database file / with html free text
Process:
Begin:

```
1   Array (Faa)          dir (*.html)
2    For i=1 to fcnt
3       Fn=faa(i,1)
4    Select database file
5    Store fn to text attribute
6   End for
7 End
```

TABLE 1.
Sector information

Tag name	No. of Words			
	0	1–20	21–50	51–
<TITLE>	6%	82%	10%	2%
<META=Description>	70%	10%	15%	5%
<META=Keywords>	85%	2%	10%	3%
<BODY>	10%	7%	17%	66%

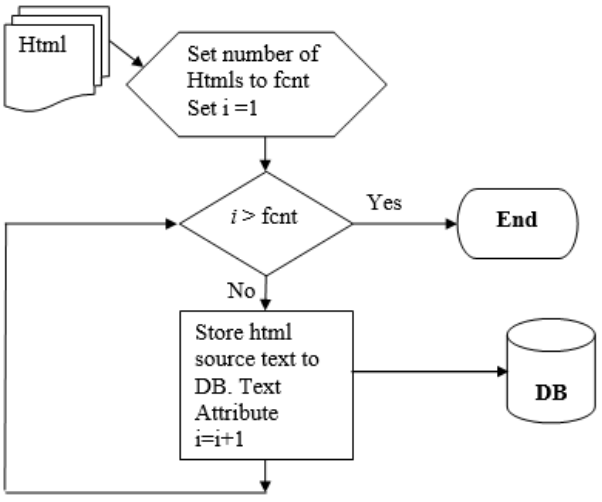

Fig. 5. Text Storing Process.

B. Document Preprocessing

To build a document index for mining, several procedures should be applied in this preprocessing phase. These processes are explanation in detail in the following parts. The web HTML document preprocessing phase contains the following processes.

- Lexical text analysis and string tokenizer
- Stop words elimination
- Term indexing

1) LEXICAL TEXT ANALYZER AND STRING TOKENIZER

Lexical text analysis is the conversion of a stream of characters (text of document) into a stream of documents words to be adopted as index terms. The lexical analysis phase mainly aims to identify the words in the text or text document.

The first step is parsing the HTML page to plain text to remove such HTML tags as "<," "/>,"<HTML>,<HEAD>,<TITLE>,</TITLE>,<H1>, and<P></P>. A tag is a string applied to spot begin or finish of physical rudiments in manuscript. These tags abundantly exist in HTML source codes (Chapter 1). These tag labels are useful despite being considered noisy data [45].

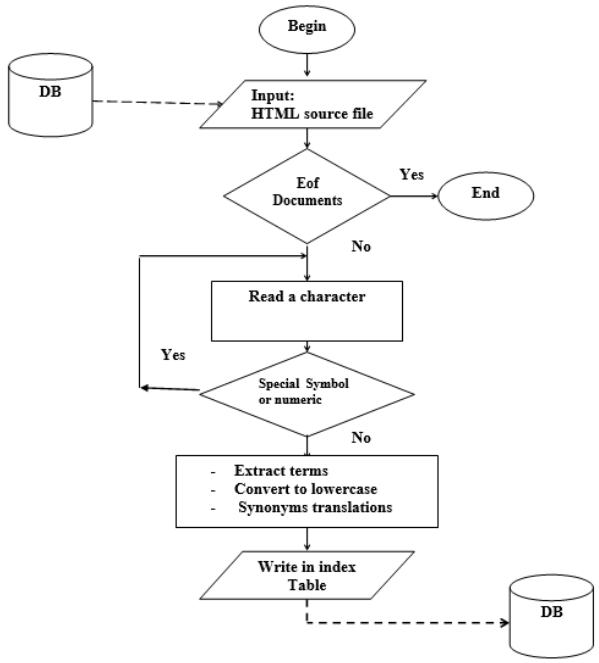

Fig. 6. Lexical Text Analyzer and String Tokenizer Flowchart.

After the HTML text is parsed, it is stored in the database text file. Each tuple is considered a transaction data for building the term index. Then, lexical analysis is performed by ignoring unnecessary special characters, hyphens, commas, and words that begin with the numbers 0–9. The remaining words begin with a–z and A–Z; the other words are ignored. After text cleaning, strings of terms represent a meaningful term of the document, thereby identifying the words in the document. Algorithm 2. explains the process.

Algorithm 2: HTML Text Parser Procedure
Procedure name: html-text-parser
Input: html text
cnt = 1
Output: Free text /stored in text attribute
Process:

```
1 cnt=1
2 sch=""
3 do while cnt < length (text attribute);
4 If buffer(cnt) is not alphabetic or buffer(cnt)
    =space;
5                or buffer(cnt) =Ascii 13
6                sch = sch + buffer(cnt)
7 else
8                write buffer(cnt)
9 End if
10               cnt=cnt+1
11             Loop
12 End while
```

String Tokenizer – After owning a string of words, the terms extracted are to be transformed to lowercase and stored in a document index table. Mutual meaning unify (MMU), which is a combination of abbreviations and synonyms of words, is a preprocessing technique proposed in this research to decrease the document vector space and increase the accuracy of document representations.

2) STOP WORD ELIMINATION

A stop word is a position or item that has small semantically happy; it looks up to arguments have a tall regularity across a collected. Stop words are generally remove from the internal text of a word document because of they appear in many word documents and are not help for retrieve. However, stop word can dependent on context. Such as, the word "computer" will likely be a stop word in a collection of processer discipline newspaper trainings, not in a market list.

After initial indexing, the document index still contains useless terms. The index should be filtered by removing stop words to decrease the number of terms in it. A tilt of halt arguments is prepared in investigate. A total of 1200 words are suggested as stop words, including the ordinary stop words similar to "the," "which," and "is." The extracted or suggested stop words include "repeat," "high," "width," "second," "first," "h1," and "h2." A complete list is shown in Appendix A. In addition to the data cleaning phase, words that are not useful in building the associative classifier should be prepared. An experiment is conducted on a collection of 110 HTML documents and reveals that 46% of the words are stop words. This practical example shows the need for stop word elimination.[46] [47] .

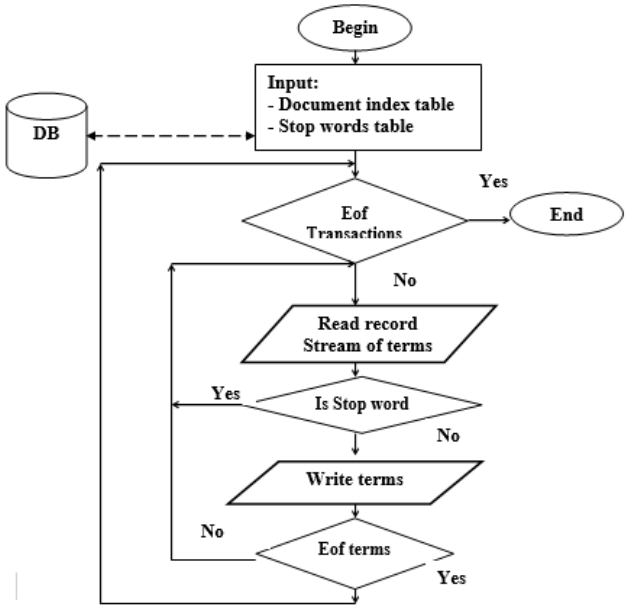

Fig 7. Stop Word Elimination Flowchart.

3) TERM STEMMING

Stemming is technics of reduce words to their grammatic roots. Removing suffixes automatically is operating that is particularly useful in the field of information retrieve and data mining applications. An example of words with suffixes is as follows.

Connect (the root)
Linked
Linking
Joining
Contacts,

where the suffixation "eds," "ings," "ions," and "ionss" are remove from last four words. The remaining word is the origin or the root of the word, "connect," which is exactly the first word. Therefore, the stemming or word stripping process is necessary for text analysis and document representation by keywords or terms. Stemming methods improve the performance of IR and KDD. Suffixation stripe reducing the of term in the IR scheme and henceforth reducing complication information in the organization. Porter's stemmer, a renowned English word stemming method, is used to remove suffixes from terms and aims to design improved stemming procedures; in this work, it is called SSM. The experimental results show that this method has an accuracy of 85%. In this domain, the method is improved by the following supporting techniques. The first one is extracting the abnormal word vector or non-stemmed terms, it contains the terms which are not valid English words after stemming, it remains as it is in the document. The second one is an important technique that uses an already prepared English dictionary. The stemmed word is matched with the dictionary. If the word exists in the dictionary, then it is deemed an actual English word and the stemming algorithm output is considered true; otherwise, the word cannot be stemmed. This procedure increases the process duration but improves accuracy.

4) SIMPLE STEMMING METHOD (SSM)

To prepare a high ranked document term frequency for efficient representation, a simple applicable stemming method is proposed (SSM). After conducting the experiments, a list of words that cannot be stemmed with their suffixes (either abnormal English word results or a normal English word with different meanings) is obtained. Table 3 presents some terms as examples.

TABLE 3.
SAMPLE TERMS THAT CANNOT BE STEMMED

Terms before stemming	After stemming
Center	Cent
Care	Career
Access	Acces
Banner	Bann

An experiment is conducted on the collected HTML text to collect words that cannot be stemmed and are similar to the words listed in Table 3. The result is a list of 253 words, which is stored in a database table and includes any term compared with a list of non-stemmed words. A word in this list will not be inputted into the stemming procedures. The following rules of SSM show a general process as follows:

Definition:

Lease Ds be a determine of IDs D = {d11, d22, d33,…dns}.

Lease Ts be a determine of document rapports T = {t11, t22, t33,. tns}.

Ts € D.

Rule 1:

The removal of term suffix should be in the following form:

If (condition) S → SW,

where condition refers to the term with suffix S and SW refers to the stemmed word or term after removing a suffix, if the condition is true. For example:

If (engineer) ing → engineer.

A term may end with different suffixes. Certain suffixes are obtained by collection from experience and previous knowledge in English language syntax and morphology rules. Suffixes may be mainly accepted and repeated in text documents. A general determination of word suffix values with slice is listed in the definition as follows:

Definition:

*ED - the stem ends with ED, -> T = left(length(T)) - 2

*S - the stem ends with S, -> T = left(length(T))

- 1
*ER - the stem ends with ER, -> T = left(length(T)) - 2
*ERS - the stem ends with ERS,-> T = left(length(T)) - 3
*ING - the stem ends with ING, -> T = left(length(T)) - 3
*EER - the stem ends with EER,-> T = left(length(T)) - 3
*IER - the stem ends with IER, -> T = left(length(T)) – 3
*LY - the stem ends with LY, -> T = left(length(T)) - 2
*ION - the stem ends with ION-> T = left(length(T)) – 3
*ISE - the stem ends with ISE,-> T = left(length(T)) - 3
*IZE - the stem ends with IZE -> T = left(length(T)) - 3

To explain SSM, rules and examples are shown below.

Rule 2:
If RIGHT (T, 2) = "ed"
If RIGHT (T,4) = "ated"
WSF ← RIGHT (T,4)
Then SW = (T-WSF) + "ate"
End if
End if
Examples:
Let T= "abbreviated"
RIGHT (T,4) = "ated"
WSF = "ated"
SW = ("abbreviated" -"ated") +"ate"
stemmed term (SW)= abbreviate …is true normal English word

activated	active	→ activate
approximated	approxim	→ approximate

Rule 3: The letter before "ed" is a vowel letter "i"
If RIGHT (T,2) = "ed"
If RIGHT (T,3) = "ied"
WSF ← RIGHT (T,3)
Then SW = (T-WSF) +"y"
End if
End if
Examples:
Let T= "classified"
RIGHT (T,3) = "ied"
WSF = "ied"
SW = ("classified" -"ied") +"y"
stemmed term (SW)= classify is true normal English word

Classified	classf	→ classify
Copied	cop	→ copy
Identified	identif	→ identify

Rule 4: the letter before "ed" is a vowel letter "u"
If RIGHT (T,2) = "ed"
If RIGHT (T,3) = "ued"
WSF ← RIGHT (T,2)
Then SW = (T-WSF) +"e"
End if
End if
Examples:
Let T= "argued"
RIGHT (T,2) = "ed"
RIGHT (T,3) = "ued"
WSF = "ed"
SW = ("argued" -"ed") +"e"
stemmed term (SW)= argue is true normal English word

argued	argu	→ argue
catalogued	catalogu	→ catalogue

Rule 5:
If RIGHT (T,2) = "ed"
If RIGHT (T, 4) ="ated" & RIGHT (T, 3) #"ied" & RIGHT (T, 3) # "ued"
WSF ← RIGHT (T, 2)
Then SW = (T-WSF) +"e"
End if
End if
Examples:
Let T= "forced"
Right (T, 2) = "ed"
Right (T,4)#"ated" & Right (T,3)#"ied" & Right (T,3)# "ued"
WSF = "ed"
SW = ("forced" -"ed") +"e"
stemmed term (SW)= force ..is true normal English word

forced	forc	→ force
generalized	generaliz	→ generalize
housed	hous	→ house

Rule 6: The letters before "s" is "ie"
If RIGHT (T,1) = "s"
If RIGHT (T,3) = "ies"
WSF ← RIGHT (T,3)
Then SW = (T-WSF) +"y"
End if
End if
Examples:
Let T= "classifies"
RIGHT (T,3) = "ies"
WSF = "ies"
SW = ("classifies " -"ies") +"y"
stemmed term (SW)= classify ..is true normal English word

classifies	classif	→	classify

abilities abilit → ability

Rule 7: The letter before "s" is "u"
If RIGHT (T,1) = "s"
If RIGHT (T,2) = "us"
Then SW = T
End if
End if
Examples:
Let T= "ambiguous"
RIGHT (T,2) = "us"
WSF = "us"
SW = T
stemmed term (SW)= "ambiguous"…is true normal English word
ambiguous ambiguou → ambiguous
corpus corpu → corpus
heterogeneous heterogeneou → heterogeneous

Rule 8: The letters before "s" is "es"
If RIGHT (T,1) = "s"
If RIGHT (T,2) = "es"
WSF ← RIGHT (T,2)
Then SW = (T-WSF)
End if
End if
Examples:
Let T= "classes"
RIGHT (T,2) = "es"
WSF = "es"
SW = ("classes " -"es")
stemmed term (SW) = class …is true normal English word
classes → class
accesses → access
taxes → tax

Rule 9: The letters before "er" is "i"
If RIGHT (T,2) = "er"
If RIGHT (T,3) = "ier"
WSF ← RIGHT (T,3)
Then SW = (T-WSF) +"y"
End if
End if

Examples:
Let T= "classifier"
RIGHT (T,3) = "ier"
WSF = "ier"
SW = ("classifies " -"ier") +"y"
stemmed term (SW)= classify..is true normal English word
classifier classif → classify
Identifier Identif → Identify
supplier suppl → supply

Rule 10: The letters before "ing" is "tt"
If RIGHT (T,3) = "ing"
If RIGHT (T,5) = "tting"
WSF ← RIGHT (T,4)
Then SW = (T-WSF)
End if
End if
Examples:
Let T= " inputting"
RIGHT (T,4) = "ting"
WSF = "ting"
SW = ("inputting" -"ting")
stemmed term (SW)= "input"..is true normal English word
inputting → input
setting → set
splitting → split
Rule 11: The letters before "ing" is "mm"
If RIGHT (T,3) = "ing"
If RIGHT (T,5) = "mming"
WSF ← RIGHT (T,4)
Then SW = (T-WSF)
End if
End if

Examples:
Let T= " stemming "
RIGHT (T,4) = "ming"
WSF = "ming"
SW = ("stemming -"ming")
stemmed term (SW)= " stem"..is true normal English word
drumm → drum
stemming → stem
trimm → trim

Rule 12: The letters before "ing" is "ate"
If RIGHT (T,3) = "ing"
If RIGHT (T,5) = "ating"
WSF ← RIGHT (T,3)
Then SW = (T-WSF)+"e"
End if
End if
Examples:
Let T= " celebrating "
RIGHT (T,3) = "ing"
WSF = "ing"
SW = ("celebrating -"ing") +"e"
stemmed term (SW)= "celebrate"……is true normal English word
celebrating celebrat → celebrate
operating operat → operate
stimulating stimulat → stimulate

Rule 13: The letters before "ion" is "at"

```
If RIGHT (T,3) = "ion"
  If RIGHT (T,5) = "ation"
    WSF ← RIGHT (T,3)
    Then SW = (T-WSF)+"e"
  End if
End if
```

Examples:

Let T= " administration"
RIGHT (T,3) = "ion"
WSF = "ion"
SW = ("administration -"ion") +"e"
stemmed term (SW)= "administrate"……is true normal English word

administration	administrat	→ administrate
participation	participat	→ participate
association	associat	→ associate

Rule 14: The letter before "ly" is "b"

```
If RIGHT (T,2) = "ly"
  If RIGHT (T,2) = "ly"
    WSF ← RIGHT (T,2)
    Then SW = (T-WSF)+"le"
  End if
End if
```

Examples:

Let T= "assemble "
RIGHT (T,3) = "ly"
WSF = "ly"
SW = ("assembly -"ly") +"le"
stemmed term (SW)= " assemble"……is true normal English word

assembly	assemb	→ assemble
predictably	predictab	→ predictable
responsibly	responsib	→ responsible

Rule 15: The letter before "ly" is "i"

```
If RIGHT (T,2) = "ly"
  If RIGHT (T,3) = "ily"
    WSF ← RIGHT (T,3)
    Then SW = (T-WSF)+"y"
  End if
End if
```

Examples:

Let T= "ordinarily "
RIGHT (T,3) = "ily"
WSF = "ily"
SW = ("ordinarily -"ily") +"y"
stemmed term (SW)= "ordinary"……is true normal English word

ordinarily	ordinar	→ ordinarily
satisfactorily	satisfactor	→ satisfactory

Many expert are performed to prove the effect of the propose stemming methods (SSM), which is reducing the total number of terms by two-thirds.

V. CONCLUSIONS

The automatic evaluation and selection of text documents provide facilities for web users, especially organizations, institutes, and universities, in selecting customized information related to their interest given the large numbers of irrelevant documents retrieved from different search engines. The proposed classifier inherits all properties of data mining systems, such as accuracy, scalability, robustness, and noise tolerance. Hypertext classification differs from traditional TC; information beyond web document contents, such as metadata, is a useful source of information, and the combination of metadata and text effectively improves classification performance. A new approach is proposed to develop accurate classifiers. observational results show that classifiers developed in this manner are common, accurate, and effective. The accuracy measure is 91%.

VI. REFERENCES

[1] Aizawa, A., An information-theoretic perspective of tf-idf measures, Information Processing and Management, Vol. 39, pp. 45–65, (2003).

[2] Bernstein, Phil, et al. The Asilomar report on database research, ACM Sigmod Record, 27(4), December 1998.

[3] B. W. Hsu, and Y. Ma. Integrating Classification And Association Rule Mining, In ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'98), pages 80–86, New York City, NY, August 1998.

[4] Chakrabarti Soumen, Mining The Web Discovering knowledge from hypertext data, Indian Institute of Technology, Bombay, Morgan Kaufmann Publishers, San Francisco, USA 2003.

[5] Cornell University "SMART IR System" (ftp://ftp.cs.cornell.edu/pub/smart),1990

[6] Hull. D. A. Improving text retrieval for the routing problem using latent semantic indexing, In 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR-94), 1994.

[7] Lewis.D. The independence assumption in information retrieval, In 10th European Conference on Machine Learning (ECML-98), pages 4–15, 1998.

[8] Dunham Margaret H., Data Mining, Introductory and Advanced Topics, Department of Computer Science and Engineering Southern Methodist University, Prentice Hall, 2002.

[9] Francesco Balena, Programming Microsoft Visual Basic.Net, wintellut, 2002.

[10] Gerald Gazdar, Natural Language Processing In Prolog, Addison-Wesley Publishing Company1989.

[11] AL-Kafagi Hussien K., Design And Implementation Of Embedded Association Rules Miner, PhD. Thesis, Department Of Computer Science, University Of Technology, Baghdad, Iraq (2002).

[12] Hammer J., H., Garcia-Molina J, Cho, R. Aranha, Extracting Semistructured Information From The Web,

Department of Computer Science, Stanford University, Stanford CA 94305-9040 , 1998.

[13] Jiawei Han, Michiline komber, Data Mining Concepts, Simon Fraser University, MK Publishers, 2001.

[14] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Mahwah, New Jersey London Simon Fraser University Morgan Kaufmann Publishers 2000.

[15] John M. Pierre, Mining knowledge from text collections using automatically generated, interwoven. Inc. 101 2nd street, san Francisco, CA 94105 jpierre@interwoven.com, 2003

[16] John Wang, Data Mining: Opportunities And Challenges, Montclair State University, Idea Group Inc. USA, , 2003

[17] Lewis, D. and Ringuette, M. Comparison of two learning algorithms for text categorization, In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[18] liu, B., Hsu, and Wang, Classification Using Association Rules" Weaknesses And Enhancements, In Grossman, R, L., Data Mining For Scientific And Engineering Applications. Kluwer Academic Publishers, Machine Learning, Mcgraw-Hill, (1997).

[19] Mehmed Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2003.

[20] Murata M,Ma, Utiyama M., Information Retrieval Using Location And Category Information, Journal Of Association For Natural Language Processing, (2000).

[21] Neel Sundaresan, Using Metadata to Enhance a Web Information Gathering System, IBM Almaden Research Center 650 Harry Rd. San Jose, neel@almaden.ibm.com,1999.

[22] Nong Ye, The Handbook Of Data Mining, Arizona State University,Lawrence Erlbaum Associates, Publishers, 2003.

[23] Paulraj Ponniah, Data Warehousing Fundamentals, Wiley,2001.

[24] Pierre John M., Practical Issues for Automated Categorization of Web Sites, 39 Townsend Street, Suite 100 San Francisco, CA 94107,2000.

[25] Porter, M. An Algorithm for Suffix Stemming, Program 14 (3), pages 130-137, July 1980.

[26] Ramiz Almisry, Fundamentals of database systems, third edition, 2000.

[27] Ricardo Baeza, Modern Information Retrival, Adidison Wesly, 1999.

[28] Riloff, E. and Hollaar, L., Text Databases and Information Retrieval, In Handbook of Computer Science. A.B. Tucker (ed), CRC Press, 1996.

[29] Rick Stout. The World Wide Web Complete Reference, Osborne Mcgraw-Hill, 1996.

[30] Sebastiani Fabrizio, Text Classification for Web Filtering, Institute of information technology, 56124 Pisa, Italy, 2004, http://www.isti.cnr.it/People/F.Sebastiani.

[31] Sebastiani Fabrizio, A tutorial on automated text categorization, Institute of information technology, 56124 Pisa, Italy, 2000, http://www.isti.cnr.it/People/F.Sebastiani.

[32] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Science Department, Stanford University, Stanford,2001 CA 94305, USA.

[33] Stephen Robertson, Understanding Inverse Document Frequency: On Theoretical Arguments for IDF, Microsoft Research 7JJ Thomson Avenue Cambridge CB3 OFB, (1995), London UK

[34] Cohen W. and Hirsch. H. Text classification using whirl, In 4th International Conference on Knowledge Discovery and Data Mining (SigKDD'98), pages 169–173, New York City,USA, 1998.

[35] Han W. Li, J., and Pei. J. CMAR: Accurate and efficient classification based on multiple class-association rules. In IEEE International Conference on Data Mining (ICDM'01), San Jose, California, 2001.

[36] W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization, ACM Transactions on Information Systems, 17(2):141 –173, 1999.

[37] Yanbo Wang, Supervised Text Classification Using Association Rule Mining, Department of Computer Science, University of Liverpool Liverpool L69 7ZF, United Kingdom {jwang}@csc.liv.ac.uk , (2004)

[38] Zhengxin Chen, Data Mining and Uncertain Reasoning, John Wiley & Sons, Inc. 2001.

[39] M. Ma, P. Wang, C.-H. Chu, "Data management for Internet of Things: Challenges approaches and opportunities", Proc. IEEE Int. Conf. IEEE Cyber Phys. Soc. Comput. Green Comput. Commun. (GreenCom) IEEE Internet Things (iThings/CPSCom), pp. 1144-1151, 2013.

[40] M. Aazam, I. Khan, A. A. Alsaffar, E.-N. Huh, "Cloud of things: Integrating Internet of Things and cloud computing and the issues involved", Proc. 11th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), pp. 414-419, Jan. 2014.

[41] S. Mayer, A. Tschofen, A. K. Dey, F. Mattern, "User interfaces for smart things—A generative approach with semantic interaction descriptions", ACM Trans. Comput. Human Interact., vol. 21, no. 2, 2014.

[42] J. C. Vidal, M. Lama, E. Otero-García, A. Bugarín, "Graph-based semantic annotation for enriching educational content with linked data", Knowl. Based Syst., vol. 55, pp. 29-42, Jan. 2014.

[43] S. Blanas et al., "A comparison of join algorithms for log processing in MapReduce", Proc. ACM SIGMOD Int. Conf. Manag. Data, pp. 975-986, 2010.

[44] Bassam S Ali, Osman Nuri Ucan, Oguz Bayat, " A novel approach for ensuring location privacy using sentiment analysis and analysis for health-care and its effects on humans health," Journal of Medical Imaging and Health Informatics, vol. 10, no. 1, pp. 178-184, 2020.

[45] S Bassam, N Osman, S Haitham, " New Methods for Analyzing Spatio-Temporal Simulation Results of Moran's Index" J Computer Eng. Inf Technol 7: 1. Doi: 10.4172/2324, vol. 9307, pp. 2, 2018.

[46] Bassam Talib Sabri, Noaman Ahmed Yaseen AL-Falahi, Isam Adil Salman, " Option for optimal extraction to indicate recognition of gestures using the self-improvement of the micro genetic algorithm"

Journal of international journal of nonlinear analysis and application, vol. 12, no. 2, pp. 2295-2302, 2021.

[47] Bassam S Ali, Osman N Ucan, " Lossy Hyperspectral Image Compression Based on Intraband Prediction and Inter-band Fractal," in 2018 Proceedings of the Fourth International Conference on Engineering & MIS 2018, turkey, Istanbul, 2018/6/19.