JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REVIEW WEB DATA ANALYSIS USING NAÏVE BAYES ALGORITHM

1. Sujata Roniya, 2. Ms. Smriti Dwivedi, 3. Ms. Richa Nehra
Student , Assistant Professor , Assistant Professor
Department Of Computer Science
D.P.G. Institute Of Technology and Management, Gurgaon, India

Abstract : **The convergence of computing and communication has resulted in an information-based civilization. However, the majority of the information is in its most basic form: data. If data is defined as facts that have been recorded, then information is the set of patterns or expectations that exist beneath the data. Databases contain a vast amount of data that is potentially useful but has yet to be discovered or expressed. It is our mission to bring it to fruition. The extraction of implicit, previously unknown, and possibly beneficial information from data is known as data mining. The objective is to create computer algorithms that automatically comb through databases looking for regularities or patterns. If strong patterns are discovered, they will most likely generalise so that reliable predictions may be made on future data. Naturally, there will be issues. Many of the patterns will be dull and mundane. Others will be fictitious, based on chance coincidences in the dataset. The technical foundation for data mining is machine learning. It's used to extract information from databases' raw data—information that's written in a readable format and can be used for a variety of reasons. The process incorporates abstraction, which entails taking the data as-is and inferring whatever structure lies underlying it. This book covers the machine learning tools and techniques that are used in real data mining to uncover and describe structural patterns in data.**

***Keywords*- Machine Learning, Structural Pattern, Naive Bayes**

I. INTRODUCTION

Web data abstraction has emerged as one of the most promising areas of research in recent years. Using web data abstraction methodology, a vast number of online data items can be accessed quickly and efficiently. The abstraction consists of web crawling, data structure identification for data towing, a pattern search algorithm, structured data storage, and other processes. To efficiently obtain data, a new string pattern matching technology is used for web data abstraction. In this study, the results clearly show that the suggested BOYER MOORE algorithm matches the matching pattern algorithm in terms of accuracy and recall in the matrix. The results demonstrate this. The Apollo voyage to the Moon in 1969, including the manned moon mission, was possibly the greatest achievement of mankind in the twentieth century. We must never forget, however, that the landing was made possible by the invention and use of computers, as NEIL Armstrong so eloquently put it, "one small step for man, one giant leap for mankind." The Apollo 11 landing was made possible thanks to the data available to scientists and engineers, which assisted these innovators in successfully completing this mission. The amount of information on the World Wide Web (WWW) is enormous. Data is available in a variety of formats, including photos, videos, and text. Manually obtaining and accommodating the location's data is a difficult task. The notion of an attack matter reduction cypher was born out of simulating to pulsate the harm. Weave facts abstractor is a piece of technology that randomly pulls information from a website. After the statistics have been retrieved, they are routinely saved in a database to be used for understudy applications. The outfit consist of a on all sides adapt to applications such as determining gift, Meta begs, observational analytics, Metasearch, information mash-ups, corporate intelligence, and so on. In theory, we have access to a number of different types of evidence inference procedures. Guide and instinctive fortify text finding are two of these strategies. The owner pillar manually input the programme designated wrappers to extract data unusual trash pages in the old-fashioned route. The data is retrieved using predefined lyrics in this fad. This passage involves an efficient cruise that is based on differences in the presence of defined knowledge of the light data abstraction approach, which is circular and semi subservient to an unsupervised method. Preparations are made and extraction work is altered to fit the discreet specimen provided by the trunk designer. WIEN, SOFTMEALY, and STALKER are examples of supervised techniques. IEPAD and OLERA are examples of semi-supervised approaches. OLERA are examples of semi-supervised approaches.

Unsupervised approaches arose as a result of learning rules and retrieving approaching data, which are essentially their talents, with the user obtaining relevant data from the output. ROADRUNNER, EXALG, FIVE TECH, and TRINITY are instances of unsupervised approaches. Regulation Aspirant turns on wrapper from a web page by finding similarities and differences between

them using Zenith (a pack foundering harmony and abstract) beliefs in dissimulate. EXALG myths date back to the days when it came to retrieving ordered data from a sequence of web pages that were derived using a common template. It consists of several stages, including the codification assortment days stage (ECGM) and the analysis stage. This erudition retrieves information from specific pages. FIVE TECH detects enlargement in the input Dom instil stray has a showing a similarity contrivance and capable develops a regulating of their anxiety, mining repetitive and optional pattern to generate the abstraction rule. TRINITY uses the KNUTH-PRATT-MORRIS pattern coincidence technique to achieve well-known data abstraction. The aforementioned algorithm uses a community-based ad matching method, which takes longer to locate the text [1].

II. WEB DATA TOOLS

The Internet is one of the most eye-catching and fashionable collections of identical data ever compiled, covering a wide range of topics from a large number of authors. The Web's particular traits of large scale and lack of central administration, as well as the information makers that these qualities elicit, are now commonplace elements of life, at least in the fields of information generation and distribution. For example, while the fact that erection an online information site where foreign research is published was formerly astonishing, it is now treated with a grain of salt .For example, while the fact that erecting an online information site where international research is cheap or free, that the internet gives data on practically every topic imaginable, and that there is no linkage among information-creators was originally unexpected, it is now regarded with a grain of salt. However, keep in mind how important these unique features are when it comes to structured data management:

- We cannot trust on people to "properly" publish information. It almost always exists in disorganised and rarely standard forms: in documents or as the most basic computer understanding actions within language text. A basic information, on the other hand, is expected to contain all "important" data in one of a few very polished and unexplained formats.

- Net information is printed in the author's preferred format with no guarantee of data, schema, or schematic view implementation. In this occurrence, a central schema for the management of old relative information must be established. The logical figure's styles truly choose the implementation of information such as logics.

- Because of the size and diversity of the Internet audience, a large number of crawlable structured materials on virtually all subjects will be available. Old databases, on the other hand, are difficult to access and tend to focus on a single region at a time. Most such comparisons, in which ancient and organised network information management creates new burdens for the logical web view scenario, frequently create new burdens for the logical web view situation.

Traditional relative information management is also easier in at least one way, which is generally not functional, than organized network management. As a result, much of the standard information is based on functional security, and optimization is out of the question in this case. Aside from withdrawal and retrieval, the unique characteristics of the internet present a slew of new issues for organized net streams. We will analyses the three criteria that come along with the various results in Section one.3 below, as well as how they will resemble in other outcomes

III. FOUNDATIONAL STRATEGY

The previous top priorities of the internet - massive capacity with no minimal administration - that have become so widespread in modern document management practices are frequently not yet defined in our organized data supervision tools. We feel that overseeing organized internet data should concentrate on the following examples:

1. Withdrawal focused: Because it generally focuses on extracting information rather than focusing individuals on the Internet, it is unreasonable to expect that all information received by the public is readily available for importing, information, and each one as basic lists. Instead, data extraction tools that absorb unstructured input and provide refined, organized data should always be used. Embedders do not appear to be completely generic – one can specialize in text, another in tables, and so on – but they must be as internet-like as possible. They're supposed to cover the internet all the time.

2. Dependent State: The estate is dependent on sensitive knowledge, standards, or arrangements from point to point. Furthermore, organized knowledge is defined as the acceptance of provinces. The algorithms that extract and question such organized Internet information, on the other hand, should not contain domain data that would make other query processing techniques difficult to use. Domain independence, in other words, entails avoiding extraction rules or coaching knowledge explained in relation to a broad subject. It appears insolvent, for example, to be an associate in nursing estate-independent providers creating a shelter to conceal firm intelligence. Furthermore, the vast majority of nursing assistants will not serve the website.

3. Flexible estate: Its primary goal is to boost the system's throughput while it is in use. Single estate datasets are the foundation of many historic organized databases. The atmosphere in which users interact with the system, rather than the property information, is question ready. The Structured Internet, on the other hand, includes a massive number of domains, each with its own set of data. As per estate workloads, users should be prepared to operate over knowledge rather than acquisition. At first glance, throughput, flexibility, and scalability may appear to be province-dependent. The issue is whether or not a system operator can simply offer us with information or specific expertise from several estates. However, the operator can also bind numerous compute procedures over many estates if desired. A site ascendable system should be property dependent, but this is not the case .

4. Mathematically efficient: It typically focuses on regions where outmoded information systems are ideal for managing a large number of schemes and sub-schemes that handle a large number of sites. On the other hand, the organized internet presents new challenges. Several of these issues center on data removal sections that haven't often been developed for large document datasets inside previously occupied activities. The interest in the spectrum of estates may result in a voluntary new process issue that old machines haven't had to deal with.

"Furthermore, they have a tendency to describe, but we have a tendency to apply these style problems to three different knowledge machines". To begin, there is Text Runner, a massive Internet database computer that deals with data cleansing. Second, but certainly not least, we'll talk web tables. They use their knowledge to remove white spaces from high-volume internet hypertext language tables, which include tabular relational skills, thereby reducing workloads. Last but not least, we will talk Octopus, which emphasizes reminder operations rather than the creation of a proper structure and lists all entities and duplicates for diagrams. Using web tables or hypertext bookmarking tables, operators can summarize all data hashes of hypertext bookmarking tables.

IV. MACHINE LEARNING

We have a text sequence A, and our job is to find a pattern P. Our perspective is to find pattern matching by employing string matching algorithms to find the pattern within the given string.

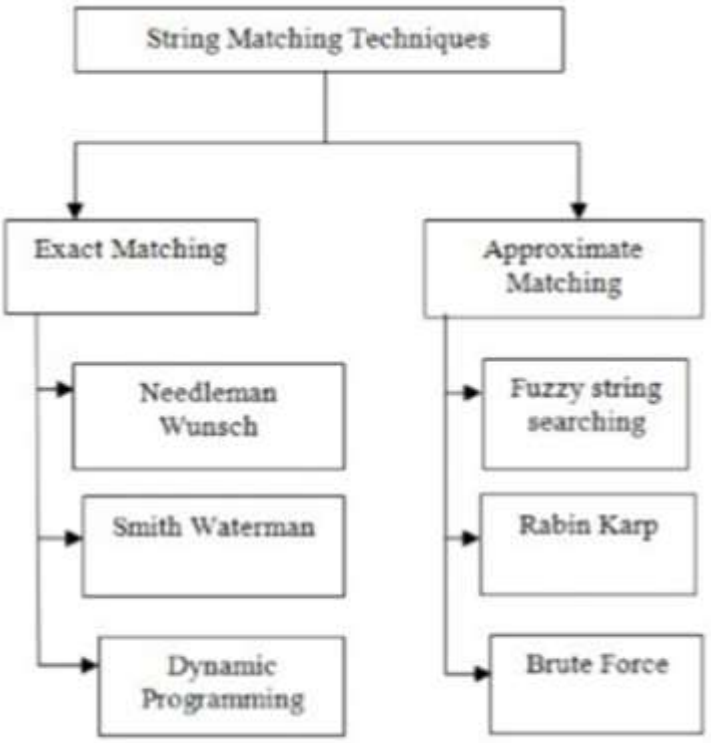

Figure 1: String Matching Techniques

The primary distinction between accurate and approximate matching is that we do not need to calculate the hash value of each character to match the contents when exactly matching the pattern. However, we make assumptions in approximation that we scan elements from right to left or left to right, depending on the algorithm we use.

The Naive Bayes Algorithm

This ability is based on the Bayes theorem, which is dependent on predictor independence. This algorithm depicts the presence of a characteristic in a class that isn't found in any other class. If a fruit is green in color, oval in shape, and around 4 inches in diameter, it is most likely a water melon. Even if these traits can be found on other fruits as well, demonstrating reliance, all of the attributes indicate that it is a fruit. As a result, we'll refer to such logic as "Naive," because it's unconcerned about the common qualities. Naive Bayes is remembered for its simplicity as well as the fact that no preprocessing is required. The rare probability X(a/b) is calculated using Naive Bayes from P(c), P(x), and P(x/c). The resulting Naive Bayes algorithm is:

- X(a/b) = X(b/a)*X(b) / X(a), where
- X(a/b) is the uncommon probability;
- X(b) is related to class likelihood;
- X(a/b) can be taken as a translator's chance given a class;
- X(a) can be regarded as the predictor's probability; The algorithm can be stated as follows:
- START
- x v-1

- y v-1
- repeat
- if A[y] equals B[x] then
- if j is equal to 0 then
- return x // pattern exists
- else
- xx-1
- y-y-1
- else x x+v –Min(y,1+last[T[x]]) yv-1 until x > v-1
- Return if the pattern not found
- END

V. SIMULATION RESULT

When we click the LOAD SOURCE button, unstructured data from various social media sites is collected. After that, the extracted data will be transformed into structured data. After organizing the data, we use the TFIDF and Native string pattern matching algorithms. Using the various string pattern algorithms, search for the positive and negative keys we've entered in our code.

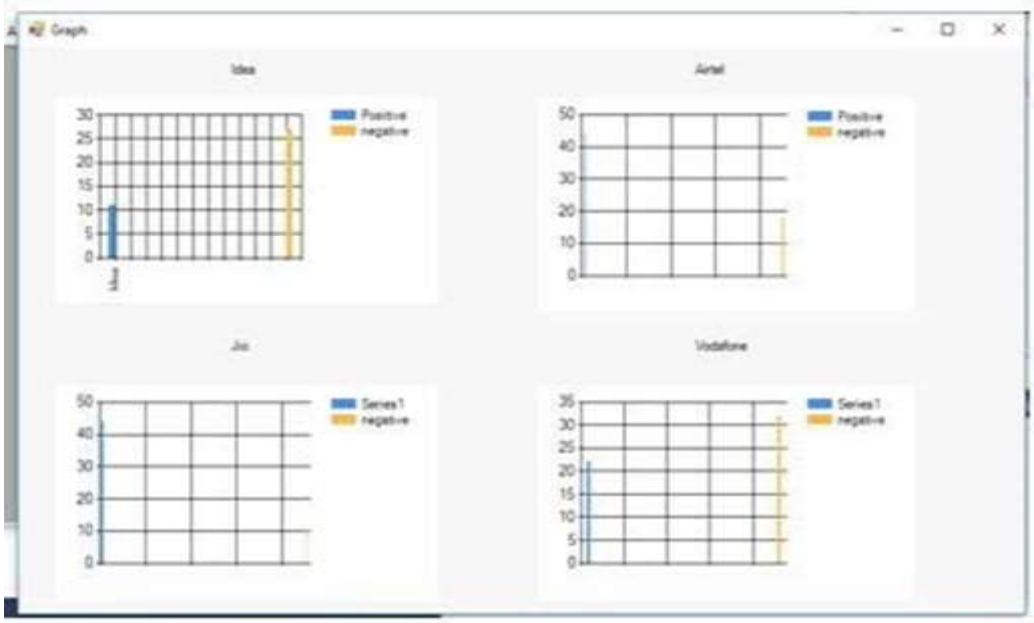

Fig 2: Comparison of positive and negative keys using naïve bayes algorithm

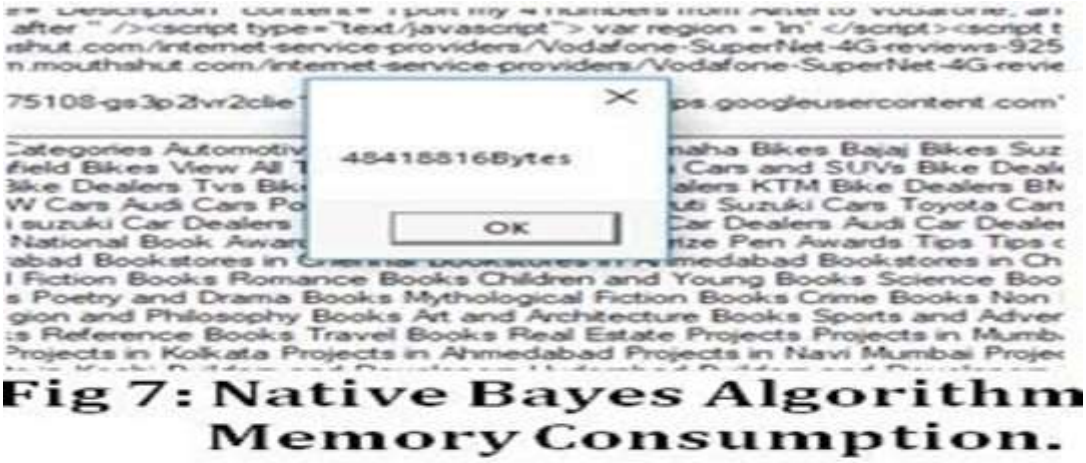

Fig 7: Native Bayes Algorithm Memory Consumption.

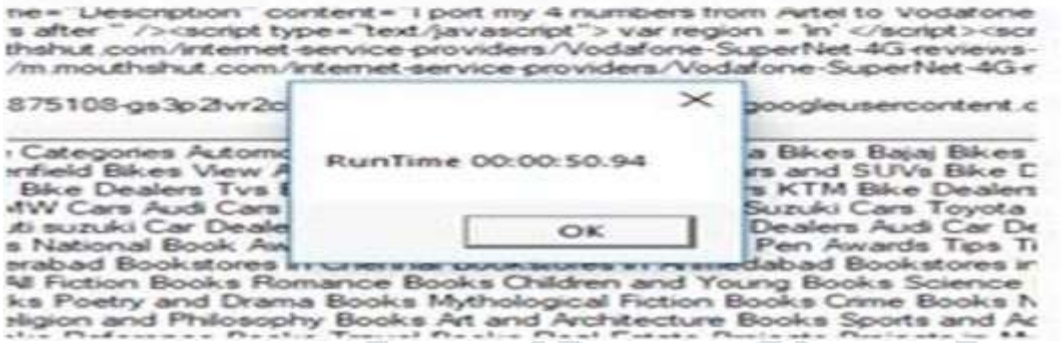

Fig 3: Naive Bayes algorithm memory consumption and time consumption

VI. CONCLUSION

The Nave Bayes technique is used to compare web data extraction systems in this article. Real world websites are used to test the redesigned structure of the online data extraction system. Precision, recall, and F measure performance metrics are used to compare existing and prospective systems. The results reveal that the suggested system outperforms the competition in terms of performance measures. For generating the required graphs, we used the C# programming language and SQL Server as a tool. Microsoft Visuals was used for all of the code and result production. We took data from social media sites for our study and then transformed it into structured data. We used (Naive Bayes algorithm) to match the pattern contained in our code after the data was formatted.

REFERENCES

[1] B.Umamageswari, Dr. R. Kalpana, V.Archana, "Web Data Extraction Using Boyer Moore Algorithm", ISSN(Print): 2393-8374, (Online): 2394-0697,Volume-5,Issue4,2018.

[2] JamunaBhandari , Anil Kumar "String Matching Rules By Variants Of Boyer Moore Algorithm" journal of global research in Computer Science Volume 5, No.1,January 2017.

[3] Shivendrakumar Pandey, Neeraj Kumar Dubey, Sonam Sharma "A Study On String Matching Methodologies", International Journal of Computer Science and Information Technologies,Vol.5(3), 2016, 4732-4735.

[4] http://somemoreacademic.blogspot.com/2012/09/brute-forcenaive-string-matching.html

[5] RantiEkaPutri , AndysahPutera, UtamaSiahaan "Examination of Document Similarity Using Rabin Karp Algorithm" , International journal of Recent Trends in Engineering & Research ISSN (ONLINE) : 2455-1457.

[6] SriharshaOddiraju "Boyer moore" Indiana State University TerreHauteIN, USA, International Journal of Computer Science.

[7] Dr.S.Vijiyarani, MS. E.Suganya "Research issues in Web Mining" , International Journal Of Computer Aided Technologies (IJCAx) ,vol2, no3, july 2015.

[8] Paul J.M.Havinga, Gerard J.M.Smit "Octopus – an energy – efficient architecture for wireless multimedia systems" , International Journal Netherlands.

[9] Er. Mohammad Shabaz ,Er. Neha Kumari "Advanced Rabin Karp Algorithm For String Matching", International Journal of Current Research Vol.9, Issue, 09, pp.5757257574, September, 2017.

[10] Ranti Eka Putri , Andysah Putera, Utama Siahaan "Examination of Document Similarity Using Rabin Karp Algorithm" , International journal of Recent Trends in Engineering & Research ISSN (ONLINE) : 2455-1457.