Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6714785, 8 pages
https://doi.org/10.1155/2022/6714785

Research Article

Research on the Key Technology of Web Data Extraction and Mining Based on the Probability Distribution

Jinqiao Yang,[1] Binghui Yang,[2] Qi Sun,[3] Shi Yan ,[4] and Yuxin Miao[5]

[1] *China Post Insurance Heilongjiang Branch, Harbin 15000, China*
[2] *School of Software, Yunnan University, Kunming 650504, China*
[3] *The University of Melbourne, Victoria 3010, Australia*
[4] *Modern Educational Technology Center, Mudanjiang Medical University, Mudanjiang, Heilongjiang 157011, China*
[5] *College of Information and Computer, Taiyuan University of Technology, Taiyuan, Shanxi 030000, China*

Correspondence should be addressed to Shi Yan; yanshi@mdjmu.edu.cn

Received 6 April 2022; Accepted 26 April 2022; Published 11 June 2022

Academic Editor: Chia-Huei Wu

In the development of internet technology innovation, the quantity of big data information continues to rise, showing the characteristics of dynamic, heterogeneous, massive, and so on. How to explore valuable and potential knowledge from the network system is the main topic of research and scholars. Based on the understanding of web data extraction and web mining technology, this paper makes a systematic study of web text mining, proposes the most commonly used hidden Markov model in probability distribution, and constructs the corresponding text mining method. The hidden Markov model is a statistical model, which is mainly used to represent a Markov process with unknown parameters. The difficulty in practical application lies in how to define the hidden parameters of the process from observable parameters and then use these parameters to conduct in-depth research. Finally, the practical results show that the hidden Markov model can not only obtain the information of different regions but also deeply analyze the data set, which proves the feasibility of this research technology.

1. Introduction

In essence, the probability distribution refers to the probability law of the values of random variables. The probability of an event represents the probability of an outcome after the experiment. In order to fully understand the experiment, it is necessary to get all the possibilities and probabilities of the experiment, which is also regarded as the probability distribution of random experiment. Assuming that the results of an experiment are represented by the value of the variable X, the probability distribution of a random experiment can be regarded as the probability distribution of a random variable. Normal distribution is a key form of practical exploration, and variables in many biological phenomena obey or approach normal distribution, such as blood glucose content, milk yield, body weight, and hemoglobin content of livestock. Most statistical analysis methods are put forward with normal distribution as the core, so it is necessary to deeply understand and skillfully use relevant concepts in both theoretical research and practical application. Generally speaking, the normal distribution probability counting, wanting to combine standard normal distribution, is analyzed; the reason is that this form is the most simple, and all normal distribution can be reduced to a standard normal distribution is calculated; research of scholars has been combined with standard normal distribution function, compiled into normal distribution table query; this practice helps to simplify calculation steps.

A research mainly from the perspective of the software database, this paper deeply explores the probability distribution as the core of data extraction and mining technology and therefore should not only master the related concepts of probability distribution, to clear the web information extraction technology and the steps of data mining technology, a comprehensive understanding of research scholars at home and abroad of web data extraction and mining technology research, in order to get more technical research experience. To extract mining data from software resource base, it is necessary to select specific technology to mine specific development process and finally assist developers to complete software engineering tasks according to the results.

Data extraction is the basis of web data mining, which refers to a process of extracting measured data from embryonic text. Traditionally, information extraction uses natural language processing technology to process free files according to syntax or semantically limited extraction mode. Because web documents are semistructured text formats that contain a lot of markup and hyperlinks and few complete sentences, traditional natural language processing techniques do not meet the requirements of such information extraction. Web data extraction is regarded as a problem of extracting target information from web pages. Data should be extracted not only from natural language text but also from structured data of web pages. This paper focuses on the latter. Generally speaking, the structured data contained in the web refers to the relevant records obtained from background data, which will be presented in the web page according to a certain template. Extracting this kind of data can provide value-added services to users, such as shopping platforms, metasearch, and information collection, on the basis of integrating data from multiple web pages.

In the research of web data mining, researchers mainly focus on two aspects, one is text clustering, and the other is automatic text classification. Among them, text clustering refers to one of the key technologies for unsupervised document organization, which first projects text into low-dimensional subspace and then uses clustering algorithm for mining and analysis. At present, the most common clustering methods include latent semantic index clustering, mixed model clustering, and spectral clustering. For example, Griffiths et al. proposed a potential Dirichlet distribution method with Gibbs sampling as the core in practical exploration, which does not require direct estimation of relevant parameters, but uses the Markov random process to simulate the distribution process of lexical topics and calculate the distribution of topics indirectly. At the same time, they also proposed the potential Dirichlet distribution method model and the dynamic potential Dirichlet distribution model, which are mainly used to deal with the problem of topic number selection and online evolution of topic clustering.

Automatic text classification belongs to the basic condition of data retrieval and mining. In practice, it is necessary to judge the data category according to the text content on the basis of clearly marking the category set in advance. In the development of modern technology innovation, text classification is mainly used in information filtering, information management, natural language processing, and understanding. Generally speaking, automatic text classification consists of text representation, classification method, and classification effect evaluation, among which the vector space model is the most common representation method. In other words, words should be regarded as items and the frequency of items as weights. There are several kinds of typical methods of text classification such as the decision tree classification, Bayesian classification, and support vector machine (SVM) classification; the current research scholars mainly discuss the classification of support vector machine (SVM) method, not only reasonable applied to many fields but also obtained the excellent result in the practical development, such as the fuzzy theory and the maximum entropy model. It should be noted that, in the process of web data classification, the accuracy, equilibrium point, recall rate, accuracy rate, and other indicators should be used for evaluation and analysis, so as to clarify the effectiveness of the application algorithm model.

With the increase of network information, the quantity of web text is more and more difficult to estimate, and because most web texts are unstructured or semistructured, the traditional data extraction and mining techniques are no longer in line with the requirements of web text mining. Therefore, researchers have conducted in-depth research on web text mining. Essentially, the web text mining is with its overlapping discipline, which involves the data mining, mathematical statistics, artificial intelligence, computer language, and other fields, although from a different point of view; the basic meaning of web text mining has a variety of understanding way, but the actual mining process is divided into the following main content, and concretes are shown in Figure 1 [1–4].

First, in text collection and preprocessing, web crawler is used to collect web pages. After inputting the initial URL, text information of this web page is obtained, and related content is searched by combining the hyperlinks of web pages. Finally, data sets are constructed in repeated operations. In order to improve the quality of the data, you also need to preprocess the web text, such as scripts and file purging [5]

Second, in feature representation and extraction, extract metadata representing features from web text, which can be stored in a structured form. Feature extraction is used to reduce the dimension of feature vectors [6].

Third, in data mining, this operation is the extraction, classification, and prediction of specific information in a document file.

Fourth, in mining evaluation, evaluate the knowledge model obtained by mining and provide the content conforming to the standards to the user [7].

Fifth, in information presentation and navigation, visually process data mining results and provide users with information navigation to facilitate them to browse and obtain information [8].

2. Analysis of Web Data Extraction and Mining Technology Based on Probability Distribution

2.1. Probability Model. In order to study text mining process systematically, it is necessary to master the function of probability distribution in data analysis. When constructing the model, it is difficult to ensure the clarity of the final results by predictive analysis based on the input characteristics, and only the probability results closest to the truth with the model as the core can be obtained. Before analyzing the probability, it is necessary to clarify the preconditions for its existence: a random variable is an event that may or may not occur. If you lose that premise, then there is no value in studying probability distributions. Generally speaking, random variables are divided into two categories: discrete random variables and continuous

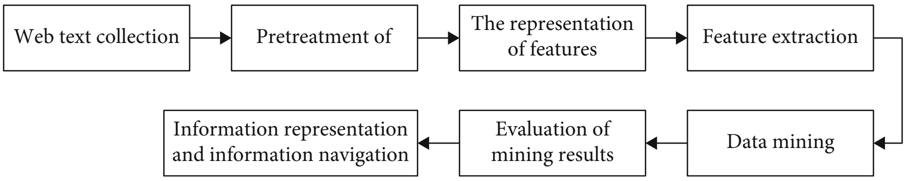
Web text collection
Pretreatment of
The representation of features
Feature extraction
Data mining
Evaluation of mining results
Information representation and information navigation

Figure 1: Flow chart of web text mining.

random variables, which correspond to different probability distribution shapes, as shown in Figure 2 [9].

The defining properties of probability theory are as follows:

Definition 1. The function $P(A)$ can be called A probability if the following conditions are met:

First, for all events A, $P\,(A) \geq 0$. Second, $P\,(\omega) = 1$, where ω represents a necessary event. Third, if multiple events $(A$ $1, A2 \cdots$ An$)$ is not integrated with each other, then the following can be obtained:

$$P(A_1 + A_2 + \cdots) = P(A_1) + P(A_2) +. \cdots \tag{1}$$

Definition 2. Assuming that A and B represent two events that meet the condition of $P(B) > 0$, $P(AB)/P(B)$ can be regarded as the conditional probability of event A under the condition of event occurrence, from which $P(A|B)$ can be obtained.

In addition to the calculation of probability, it is assumed that ABC represents three events contained in the same sample space; then it can be obtained: first, $P(A + B) = P(A) + P(B) - P(AB)$. Second, $P(BC) - P(AC) + P(ABC)\,;^{P(A+B+C)=P(A)+P(B)+P(C)-P(AB)-}$, and finally, if multiple events $(A1, A2 \cdots$ An$)$ are not compatible, then we can get

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n). \tag{2}$$

In the subtraction calculation of probability, let A and B represent two random events; then the following can be obtained:

$$P(B - A) = P(B) - P(AB). \tag{3}$$

Theorem 3. *Under the condition of $B \subset A$, we can get*

$$P(B - A) = P(B) - P(A). \tag{4}$$

In the calculation of probability multiplication formula, it is assumed that $A_1, A_2, \cdots, A_n$ represents N events and meets the condition of $P(A_1, A_2, \cdots, A_{n-1}) > 0$; then the following can be obtained:

$$\begin{aligned} P(A_1, A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots \\ P(A_n|A_1A_2 \cdots A_{n+1}). \end{aligned} \tag{5}$$

In the calculation and analysis of the full probability formula of probability, it is assumed that $A_1, A_2, \cdots, A_n$ events constitute a complete set of events and meet the condition of $P(A_i) > 0$, $i = 1, 2, \cdots$; then for all events B, the following can be obtained:

$$P(B) = \sum_i P(A_i)P(B|A_i). \tag{6}$$

Lemma 4. *In the probability Bayesian formula analysis, it is assumed that $A_1, A_2, \cdots, A_n$ can constitute a complete set of events and meet the condition of $P(A_i) > 0$, $i = 1, 2, \cdots$; then for all events B and $P(B) > 0$, the following can be obtained:*

$$P(A_i|B) = -\frac{P(A_j)P(B|A_j)}{\sum_i P(A_i)P(B|A_i)}, (j = 1, 2, \cdots). \tag{7}$$

Corollary 5. *As a kind of statistical signal model, the hidden Markov model can use the obtained training samples for adaptive learning. In the late 1960s, Baum et al. proposed the basic theory of the model in empirical research. In the 1970s and 1980s, Baker et al. applied the model to the field of signal processing. It was not until 1990s that the model began to realize data extraction and mining combined with web text features. The specific definition is as follows [10]:*

On the one hand, suppose that $\{X(t), t \in T\}$ represents a random process and E represents the state space. If for any $t_1 < t_2 < \cdots < t_n < t$ arbitrary $x_1, x_2, \cdots, x_n, x \in E$, the conditional distribution function of random variable $X(t)$ in the known condition $X(t_1) = x_1, X(t_2) = x_2, \cdots, X(t_n) = x_n$ is only related to $X(t_n) = x_n$ and has no relationship with $iX(t_{n-1}) = x_{n-1}, \cdots, X(t_2) = x_2$, then the conditional distribution function meets the following requirements:

$$F(x, t|x_n, x_{n-1}, \cdots, x_2, x_1, t_n, t_{n-1}, \cdots, t_2, t_1) = F(x, t|x_n, t_n). \tag{8}$$

This process is known as a Markov process.

Conjecture 6. *On the other hand, the random sequence Xn can be in state $S_1, S_2, \cdots, S_N$ at any time, and the probability of its state $qm + k$ at $m + K$ is only related to its state Qm at m and has no relation to the state before M, so the following can be obtained [11]:*

$$\begin{aligned} P(X_{m+k} = q_{m+k}|X_m = q_m, X_{m-1} = q_{m-1}, \cdots, X_1 = q_1) \\ = P(X_{m+k} = q_{m+k}|X_m = q_m). \end{aligned} \tag{9}$$

In the above formula, Xn can be regarded as a Markov chain if the condition $q_1, q_2, \cdots, q_m, q_{m+k} \in (S_1, S_2, \cdots, S_N)$ is met.

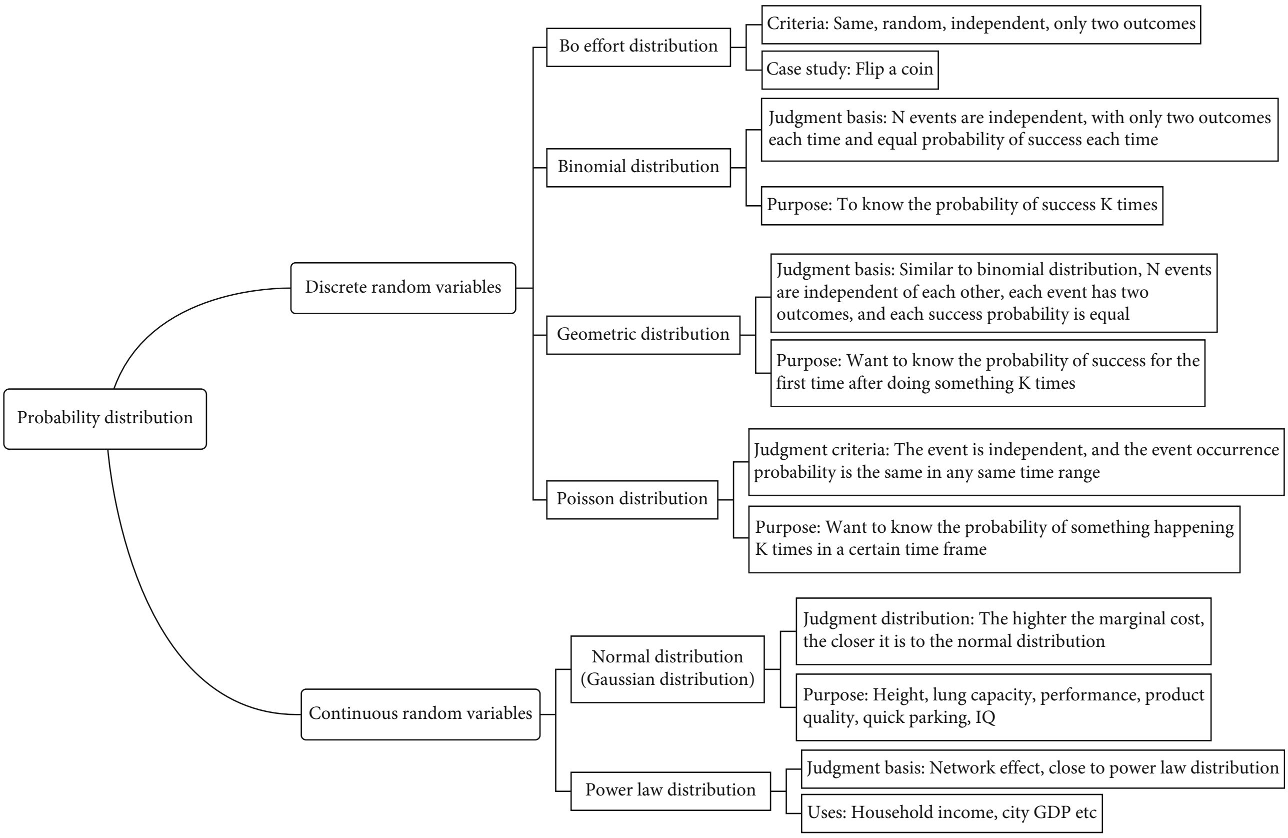
Probability distribution
Discrete random variables
Bo effort distribution
Criteria: Same, random, independent, only two outcomes
Case study: Flip a coin
Binomial distribution
Judgment basis: N events are independent, with only two outcomes each time and equal probability of success each time
Purpose: To know the probability of success K times
Geometric distribution
Judgment basis: Similar to binomial distribution, N events are independent of each other, each event has two outcomes, and each success probability is equal
Purpose: Want to know the probability of success for the first time after doing something K times
Poisson distribution
Judgment criteria: The event is independent, and the event occurrence probability is the same in any same time range
Purpose: Want to know the probability of something happening K times in a certain time frame
Continuous random variables
Normal distribution (Gaussian distribution)
Judgment distribution: The highter the marginal cost, the closer it is to the normal distribution
Purpose: Height, lung capacity, performance, product quality, quick parking, IQ
Power law distribution
Judgment basis: Network effect, close to power law distribution
Uses: Household income, city GDP etc

Figure 2: Content analysis of random variables.

According to the structural diagram analysis of the Markov model as shown in Figure 3, during data extraction and mining, the random process is regarded as a series of continuous transfer of states, and the state at time T is QT, which can be any result of N state sets [12].

Combined with the structural diagram analysis of the hidden Markov model, it is found in Figure 4 [13].

2.2. Problem Algorithm

Theorem 7. *First, evaluate the problem. This problem requires the analysis of a given model λ, a sequence of observed values $O = (O_1, O_2, \cdots, O_T)$, and the specific study of probability $P(O|\lambda)$. There are two common algorithms:*

On the one hand, the variables of the preceding algorithm are shown as follows:

$$\alpha_t(i) = P(O_1, O_2, \cdots, O_r, q_t = s_i|\lambda) 2 \le t \le T. \tag{10}$$

The initialization result is

$$\alpha_l(i) = \pi_i b_i(O_l) l \le i \le N. \tag{11}$$

The recursive result is

$$\alpha_{t+l}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+l})\ 1 \le t \le T-1, 1 \le j \le N, \tag{12}$$

Ends as follows:

$$P(O|\lambda) = \sum_{i=l}^{N} \alpha_T(i). \tag{13}$$

On the other hand, the variables of the latter algorithm are shown as follows:

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \cdots O_T | q_t = s_i, \lambda)\ l \le t \le T-1, \text{other}\ \beta_T(i) = I. \tag{14}$$

The initialization result is

$$\beta_T(i) = I\ 1 \le i \le N. \tag{15}$$

The recursive result is

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+l}(j)\ t = T-1, T-2, \cdots CI, \ne i \le N. \tag{16}$$

End as follows:

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_l) \beta_l(i). \tag{17}$$

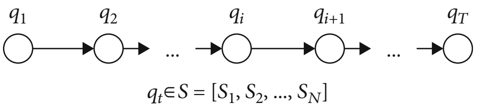

Figure 3: Structure diagram of the Markov model.

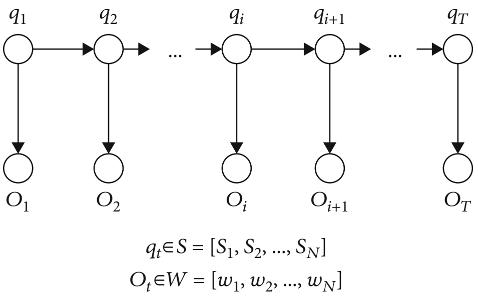
q_1
q_2
q_i
q_{i+1}
q_T
O_1
O_2
O_i
O_{i+1}
O_T
$q_t \in S = [S_1, S_2, ..., S_N]$
$O_t \in W = [w_1, w_2, ..., w_N]$

Figure 4: Structure diagram of hidden Markov model.

Second, in learning problems, the Baum-Welch algorithm should be used to deal with such problems. When $\xi_t(i, j)$ represents the given training sequence O and model λ, the Markov chain is in state θi at time T and state θj at time $t + L$. The corresponding probability calculation formula is shown as follows:

$$\xi_t(i, j) = P(O, q_t = \theta_i, q_{t+1} = \theta_j | \lambda) = [\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+l}(j)] / P(O|\lambda). \tag{18}$$

The posterior probability function is defined as follows:

$$\gamma_t(i) = p(q_t = s_i | o, \lambda). \tag{19}$$

According to the definition, the relationship between them can be found as follows:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_l(i, j). \tag{20}$$

Example 8. By defining the above probability function and reevaluating the model parameters, the following results can be obtained:

$$\begin{aligned} \bar{\pi}_i &= \gamma_l(i), \\ \bar{a}_{ij} &= \frac{\sum_{t-1}^{T-1} \xi_t(i, j)}{\sum_{t-1}^{T-1} \gamma_t(i)}, \\ \bar{b}_j(k) &= \frac{\overset{\sum^{T-1} \gamma_t(j)}{t = l\, SZO_t = W_k}}{\sum_{t=1}^{T-1} \gamma_t(j)}. \end{aligned} \tag{21}$$

Third, in the decoding problem, using the Viterbi algorithm to deal with the decoding problem, it is necessary to determine the best state sequence $Q^* = (q_1^*, q_2^*, \cdots, q_l^*)$ on the basis of clearly observing the numerical sequence $O = (O_1, O_2, \cdots, O_T)$ and model λ, where it is initialized as

$$\begin{aligned} \delta_I(i) = \pi_i b_i(O_I), 1 \le i \le N\, \delta_I(i) = \pi_i b_i(O_I), 1 \le i \le N, \\ \phi_I(i) = 0, 1 \le i \le N \phi_I(i) = 0, 1 \le i \le N. \end{aligned} \tag{22}$$

The recursive result is

$$\begin{aligned} \delta_t(j) &= \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), 2 \le t \le T, 1 \le j \le N, \\ \phi_t(j) &= \arg\max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], 2 \le t \le T, 1 \le j \le N. \end{aligned} \tag{23}$$

The final result is

$$\begin{aligned} P^* &= \max_{1 \le i \le N} [\delta_T(i)], \\ q_T^* &= \arg\max_{1 \le i \le N} [\delta_T(i)]. \end{aligned} \tag{24}$$

The result of state sequence is

$$q_t^* = \underset{1 \le i \le N}{\phi_{t+1}} (q_{t+1}^*), t = T - 1, T - 2, \cdots, l. \tag{25}$$

2.3. Framework Analysis. This paper studies the use of a website as a tool for web record extraction, and the overall process is divided into two stages: On the one hand, it is necessary to study the HTML tree formed by all the input web pages, including images, formats, and free text, and then use an entropy to evaluate the similarity of the HTML subtree to estimate the positioning data and divide the data into data segment sequence, which is the numerical content of the next stage. On the other hand, record templates that are expected to be presented are correctly populated, sorted sequentially. In this process, in order to complete the correct classification, two conditional models should be selected, one is MaxEnt and the other is MEMM, which can be used to train and observe different types of context information.

This paper studies the use of the hidden Markov model to obtain specific information, combined with the training data set to build the corresponding model, and based on the model to complete the mining of specific information in the test data, the specific framework is shown in Figure 5.

First, in data extraction, this operation is the basic condition of data mining, which refers to the process of extracting specific data information from web text. Traditionally, information extraction takes advantage of natural language processing technology and regards syntax or semantics as constraints, so as to construct data extraction mode and process web text information freely. From a practical point of view, however, this technique has fallen short because of the semistructured nature of web text, which contains not only large numbers of tags and hyperlinks but also very few full sentences.

Nowadays, the common extraction operation is divided into three forms: the first is the manual method, requiring programmers to observe the web page and source code based on the selection of mode to write programs to extract target data, and the selected mode is usually used to describe the location of tags. Secondly, it refers to wrapper induction, which belongs to a supervised learning method. It needs to first make use of manual annotation to identify the data records to be extracted and then regard the annotated web pages as learning samples, from which a group is extracted as learning rules, and then used to extract data information similar to web pages. Finally, it refers to automatic extraction, which belongs to unsupervised learning method. One or more web pages should be regarded as input content, and then mode or grammar should be automatically selected from them to complete information extraction. According to the structure diagram analysis shown in Figure 6, it is found that the automatic extraction method is very suitable for processing large quantities of webpage information extraction, so it is the main topic of current scientific research.

Second, in data mining, web mining tasks are divided into three types: the first is content mining, the second is structure mining, and the last is using mining. From the perspective of information retrieval, the task of web text mining is to help users filter page information and improve the efficiency and quality of information retrieval. From the perspective of database, the task of Web text mining is to build and integrate data model and to support complex query, not simple keyword search. At present, the most common research on web text mining can be divided into two aspects. On the one hand, it refers to text clustering, and on the other hand, it refers to automatic text classification. The former, as the key content of unsupervised method to organize documents, should first project the text into low-dimensional subspace and then use clustering algorithm for processing. The latter, as the key content of information retrieval and text mining, needs to give a set of category tags and then judge the specific classification according to the text content.

After obtaining record data, all records should be briefly processed to remove nonstandard content. Since my research is based on the text of marking training data for mining, word segmentation and word state should be marked for the recorded content. The specific record processing results are shown as follows:

```
<title>Implementing </title> <title> Distributed</titlc><title> Server </titlc>
<title>Groups </title> <title> for </title><title> the </title><title>World</title>
<title>Wide</title><title>Web</title><author>Michael</author> <author>
Garland</author><author>Sebastian </author><author>
Grassia</author><author> Robert</author><author>Monroe
</author><author>Siddhartha</author> <author>Puri</author> <date>
25</date> <date> January</date> <date> 1995 </date> <affiliation>School
</affiliation><affiliation> of </affiliation><affiliation> Computer
</affiliation><affiliation> Science</affiliation> <affiliation> </affiliation> Carnegie
<affiliation>Mellon <affiliation> University </affiliation>
```

2.4. Parameter Design. First, $N = 4$ involves author, date, title, and affiliation and is represented by $S1$, $S2$, $S3$, and $S4$. Second, M represents the number of words in the text sequence; Thirdly, $PI\,I$ represents the probability of being in the Si state at the beginning of the experiment. Fourth, AIJ represents the specific probability of transition from state Si to state Sj. Fifth, $bj\,(k)$ represents the actual probability of the word wk appearing at time t in the state.

After the model structure is defined, the three parameters $\pi\,I$, AIj, and $BJ\,(k)$ are used for analysis. As the training data set studied in this paper has been marked, maximum likelihood can be directly used for evaluation in the analysis, as shown below:

$$\begin{aligned} p_i &= P(q_i = s_i) = \frac{\text{The } \textit{number} \text{ of } \textit{times} \text{ the } \textit{first word} \text{ in } \textit{a text sequence belongs} \text{ to } \textit{state } s_i}{\sum_{i=1}^{4} \text{The } \textit{number} \text{ of the } \textit{first word} \text{ in } \textit{a text sequence belongs} \text{ to } \textit{state } s_i}, \\ a_{ij} &= P\left(q_{t+1} = s_j | q_t = s_i\right) = \frac{\textit{Number} \text{ of } \textit{transitions from state } s_i \text{ to } \textit{state } s_j}{\sum_{j=1}^{N} \textit{Number} \text{ of } \textit{transitions from state } s_i \text{ to } \textit{state } s_j}, \\ b_j(k) &= P\left(O_t = W_k | q_t = s_j\right) = \frac{\text{The } \textit{number} \text{ of } \textit{times state } s_j \textit{ releases} \text{ the } \textit{word } w_k}{\text{The } \textit{total number} \text{ of } \textit{occurrences} \text{ of } \textit{state } s_j \text{ in } \textit{a text sequence}}. \end{aligned} \tag{26}$$

If the training data set does not contain the words in the test data combination, then the zero probability problem will arise, which requires probability smoothing. There are many common smoothing methods, such as absolute discount method, linear reduction method, and absolute reduction method. On the basis of fully considering the web text sequence, the probability of the words before and after being in the same state is very large, so a relatively simple way is selected for processing during the experiment. If the probability of a word is zero, then use the probability of subsequent words to replace it. This way, the effect is better and the practical operation is easier.

3. Experimental Analysis

3.1. Evaluation Criteria. According to the performance evaluation index of constant information extraction proposed in the conference in the early 1990s, this paper selects two items for

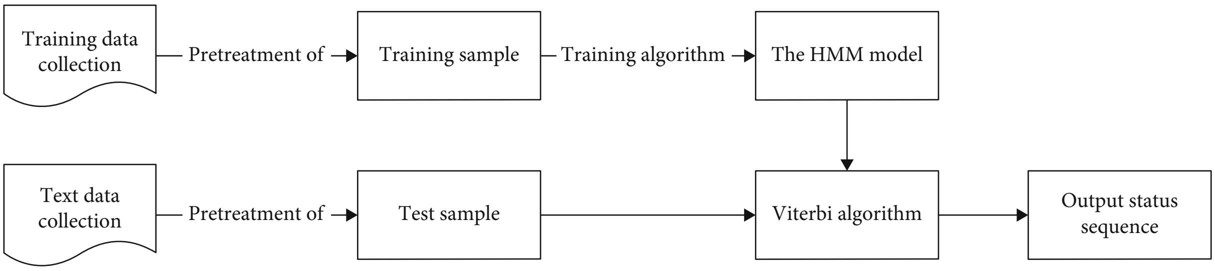

Figure 5: Frame diagram of text data extraction and mining.

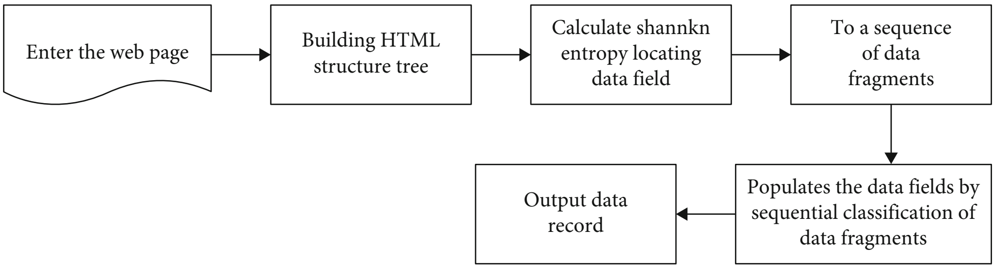
Enter the web page
Building HTML structure tree
Calculate shannkn entropy locating data field
To a sequence of data fragments
Populates the data fields by sequential classification of data fragments
Output data record

Figure 6: Automatically extracted structure diagram.

Table 1: Experimental results of the first group.

The domain name	The number of words			
	Title	Author	Date	Affiliation
Correct-marked words	749	557	85	621
Words in data set	806	621	93	683
Words in extracted result set	835	575	101	691
REC	0.929280	0.896940	0.913978	0.909224
PRE	0.897001	0.968696	0.841582	0.898698

analysis, one is recall rate (REC), and the other is accuracy rate (PRE). The specific calculation formula is as follows:

$$\begin{aligned} \mathrm{REC} &= \frac{\text{Extract the correct number of words marked in a field in the result}}{\text{Total number of words in a field in the smaple}} = \frac{\text{Correct-Marked Words}}{\text{Words in Data Set}}, \\ \mathrm{REC} &= \frac{\text{Extract the correct number of words marked in a field in the result}}{\text{Extract the total number of words in a field in the sturcture}} = \frac{\text{Correct-Marked Words}}{\text{Words in Extracted Result Set}}. \end{aligned} \tag{27}$$

3.2. Result Analysis. In this paper, the VC6.0 programming language and Windows environment were used for experimental operations. The selected experimental data contained the heads of 944 papers. Each time, part of the content was selected as the training data set, and the rest were the test data set. In the experimental analysis, the following results can be obtained: Title represents the title of a paper, author represents the author of a paper, date represents the publication time, and affiliation represents the affiliation.

In the first experimental analysis, 600 paper heads were selected for analysis, including 500 training data sets containing 11,472 words in total and 100 test data sets containing 2,203 words in total. The actual analysis results are shown in Table 1.

In the second group, 830 paper heads were selected for analysis, including 600 training data sets, 13,909 words in total, and 230 test data sets, including 4,934 words in total. The actual research results are shown in Table 2.

Table 2: Experimental results of the second group.

The domain name	The number of words			
	Title	Author	Date	Affiliation
Correct-marked words	1705	1182	164	1356
Words in data set	1816	1425	175	1518
Words in extracted result set	1907	1242	232	1609
REC	0.938877	0.829473	0.937143	0.893281
PRE	0.894074	0.951690	0.706897	0.842759

Compared with the above results, it is found that the recall rate and accuracy of web text data are very high. This proves that it is effective to mine marked data information by using the hidden Markov model and maximum likelihood extraction. Among them, title has the highest accuracy rate, with an average value of 0.934. Author had the highest recall rate, with an average value of 0.961. Under the condition that the amount of test and training data keeps increasing, the recall rate of DATE will continue to rise. The reason for this change is that there is not much data information in this area, and the corresponding recall rate will increase as the data information of test set keeps increasing.

4. Conclusion

When analyzing data extraction and mining technologies, researchers in various countries should constantly expand the search scope and fully integrate information data from different fields, so as to get rid of the limitation of traditional technology concepts and find better optimization algorithms. In view of this, this paper systematically understands the extraction and mining of web data, makes clear the importance of key technologies in practical operation, and combines hidden Markov model to carry out in-depth mining, so as to grasp more high-quality research results. Because the hidden Markov model has the advantages of high efficiency, mature algorithm, and in-depth research, the application of maximum likelihood method in web text mining is the main topic of current research and discussion. Combined with the research results of this paper, it is proved that hidden Markov model plays an active role in web text mining. At the same time, the training of professional and technical personnel should be strengthened, and the key technologies of web data extraction and mining should be explored from the perspective of sustainable development.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] C. Shi, W. Luo, and H. Lin, “Application of text mining technology in case analysis of Internet gambling,” *Computer Engineering and Applications*, vol. 47, no. 28, pp. 113–248, 2011.

[2] H. W. Ma, “Application research of web mining based on XML,” *Computer Knowledge & Technology*, vol. 7, no. 32, pp. 7853–7855, 2011.

[3] X. Yao, S. Gao, J. Xue, and M. Lu, “Deep web data extraction and mining in global mode,” *Computer Applications and Software*, vol. 35, no. 2, pp. 91–95, 2018.

[4] X. Sumba, F. Sumba, A. Tello, F. Baculima, M. Espinoza, and V. Saquicela, “Detecting similar areas of knowledge using semantic and data mining technologies,” *Electronic Notes in Theoretical Computer Science (ENTCS)*, vol. 329, pp. 149–167, 2016.

[5] L. Deng, “Application of data mining in Jiangxi tungsten industry technology prediction,” *Cooperative Economics and Science and Technology*, vol. 21, pp. 51-52, 2012.

[6] Y. Ma, X. Meng, and D. Jiang, “Mobile application integration: framework, techniques and challenges,” *Chinese Journal of Computers*, vol. 36, no. 7, pp. 1375–1387, 2013.

[7] Q. Chen, W. Ding, and Q. Shi, “Research on key technologies of deep web data extraction for network public opinion based on cloud computing,” *Computer knowledge and technology*, vol. 12, no. 15, pp. 23–25, 2016.

[8] Y. Gu and H. Feng, “Research on intelligent urban disaster prevention and relief emergency processing support system,” *Computer Engineering and Design*, vol. 6, pp. 1503–1505, 2005.

[9] L. Zhou and L. Lin, “Review of focused reptile technology research,” *Journal of Computer Applications*, vol. 9, pp. 1965–1969, 2005.

[10] C.-Y. Chang, Y.-C. Chuang, E.-J. Chang, and A.-Y. A. Wu, “MulTa-HDC: A Multi-Task Learning Framework for Hyperdimensional Computing,” *IEEE Transactions on Computers*, vol. 70, no. 8, pp. 1269–1284, 2021.

[11] C. Zhang, C. Xiao, W. Zhu, X. Chen, and Z. Li, “Financial intelligence visualization analysis and literature review,” *Financial Monthly*, vol. 3, pp. 24–32, 2019.

[12] S. Huang, “Design of web data mining system model based on XML,” *Journal of tonghua normal university*, vol. 31, no. 12, pp. 35–37, 2010.

[13] A. Tang, K. Gao, and X. Zhang, “Research on Beijing water affairs data fusion technology for big data,” *Water resources informatization*, vol. 6, pp. 9–22, 2019.