IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 9, 2020, accepted January 22, 2020, date of publication February 10, 2020, date of current version March 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972586

Automatic Dialogue System of Marriage Law Based on the Parallel C4.5 Decision Tree

CHENG WANG[1,2], **DELEI CHEN**[1], **YANXIA HU**[1], **YU CENG**[1], **JIANWEI CHEN**[3], **AND HAILIN LI**[4]

[1]College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[2]Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Soochow 215006, China
[3]Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182, USA
[4]College of Business Administration, Huaqiao University, Quanzhou 362011, China

Corresponding author: Jianwei Chen (jchen@mail.sdsu.edu)

This work was supported in part by the National Natural Science Foundation of China under Grant 51305142, Grant 51305143, Grant 71571056, Grant 18BTJ031, and Grant DMS-0907710, in part by the General Financial Grant from the China Postdoctoral Science Foundation under Grant 2014M552429, in part by the Project of Quanzhou science and technology plan under Grant 2017G045, Grant 2018C110R, and Grant 2018C114R, in part by the Shanghai Science and Technology Committee under Grant 18411950300, in part by the Postgraduate Scientific Research Innovation Ability Training Plan Funding Projects of Huaqiao University under Grant 1400214012, in part by the Fujian Province Science and Technology Plan under Grant 2017H01010065, and in part by the open project of Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, China, under Grant KJS1839.

ABSTRACT In order to realize real-time marriage legal consultation automatically, a marriage legal dialogue system based on the parallel C4.5 decision tree was designed in this paper. Firstly, the legal consultation problem is transformed into a classification task. Secondly, a legal consultation classification prediction model based on the parallel C4.5 decision tree algorithm was trained with MapReduce by collected data. Finally, a model based on the SVM algorithm, which has a strategy that was designed to provide automatic interaction for users, was designed to extract attribute value from the user's input. When a new user comes to consult, an automatic legal dialogue is launched to respond to user intelligently. The proposed system works well in some real applications, such as the divorce problem, and the experimental results show that it is outperforming the SVM and NB algorithm and more applicable than the other two algorithms. Moreover, the system can return consultation results with fewer questions asked to the user than some automatic legal consultation websites, which improves the efficiency of consultation.

INDEX TERMS Automatic dialogue system, parallel C4.5 decision tree, SVM, attribute value extraction, marriage law.

I. INTRODUCTION

In modern society, the marriage crisis is a common phenomenon, and many people would like to protect their rights in legal means. However, most people know little about the rights that they should protect, let alone protecting them from violating lawfully.

People always maintain their rights in two ways. One way is to refer to many related cases and make comparisons with their cases [1]. Another way is to seek help from lawyers. As far as the current legal consultative service is concerned, the main method is that lawyers provide help and services to the users through a one-to-one communication method, which means that lawyers need much time to help the users, which leads to the inefficiency of consultation [1].

The first way requires searching for many similar cases. Although part of the burden can be alleviated by keywords search, there are still many cases returned from the searching engines, such as Google, Baidu, etc., which can't fully meet the requirements of the users. The second way is to seek help from lawyers. It is well known that lawyers usually ask the same questions to different users and do the same logical reasoning. These cumbersome and repeatable works are urgent to be substituted by machines.

Currently, rule-based reasoning is widely used in most legal expert systems [2]. It has an advantage that both theoretical and practical knowledge of legal experts can be easily collected. In [3], a case-based medical consultation system was proposed, it analyzed a person's complaint (disease) in the form of a sentence or question paragraph. Then the system answered the problem in the form of diagnosis according to the system knowledge, the system uses Case-based

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

TABLE 1. Product status of legal automatic consulting products.

Company	Product name	Functional description	Product shortage
Lvpinghui Technology Co., Ltd.	Legal consultation	Products provide users with single or multiple choices	Single function, Limited ability to analyze and mine;
Qingdun Information Technology Co., Ltd.	Xiaofa robot	Q&A consultation	Only support a round of dialogue process
Wusong Network Technology Co., Ltd.	Fa Xiaotao's case analysis	Search the database by keyword	The returned answer is low in accuracy
Lvban Network Technology Co., Ltd.	Legal consulting Service Platform	User consulting lawyer online.	Online service that requires professional lawyers, can't achieve 24-hour online service, its response is slow, its experience is unfriendly

Reasoning (CBR) and Sorenson coefficient calculations to perform the matching process to find out which cases have the highest matching rate with the new cases. In [4], it proposed a similar case retrieval system, the system uses the iFLYTEK's online speech synthesis technology and natural language processing technology to realize the Legal Knowledge Q & A with relevant context ability, and finally retrieve the most similar cases from database as the conclusion of user consultation. In [5], an algorithm of legal text classification based on feature words was proposed. It took legal judgment as a training corpus to establish the relationship between legal provisions and feature words so that relevant legal provisions can be accurately extracted from the judgment. Then it established the corresponding relationship between legal provisions and feature words by calculating the feature words of documents with TF-IDF. In [6], a semi-automatic ontology construction method was proposed for legal question-and-answer, which provides reasoning support for the legal question-and-answer system by exploring the implication between legal provisions and problem statements, and effectively helped the development of the ontology and rules of criminal law.

With the development of natural language technology and deep learning, natural language technology and deep learning are widely used in simple dialogues in e-commerce customer service, chat, intelligent devices and other fields [7]–[14]. In [15], a question-and-answer system based on knowledge graph reasoning used a knowledge graph to provide well-structured relationship information between entities, and used deep learning to deal with noise in problems and learn multi-skip reasoning at the same time. In [16], a remote supervised open field question-and-answer system used a paragraph selector to filter out the noisy paragraphs and a paragraph reader to extract the correct answers from the de-noised paragraphs. In [17], a question-and-answer system was proposed based on reinforcement learning and collator. Through a new open domain question-and-answer communication model with collation components, the retrieved answers were ranked according to the possibility of extracting the basic factual answers of a given question, and the collator was trained by reinforcement learning.

In recent years, there are also some developments in question answering technology related to Chinese law. In [18], a framework was proposed for constructing a network of mixed legal knowledge based on the Chinese encyclopedia and legal judgment. First, it builds a network of legal terms from encyclopedia data. Then, the legal knowledge graph is constructed through Chinese legal judgment to capture the strict logical connection in legal judgment. Finally, a hybrid knowledge network of Chinese law is constructed by combining legal terms network and legal knowledge graph. In [19], it introduced a free Chinese legal technology system (IFly-Legal), which utilizes deep context representation, multiple attention mechanisms and other technologies for legal consultation, multi-channel legal inquiry, and legal literature analysis.

Through the research on the legal auto consulting products on the market in China, it is found that the characteristics of the existing mainstream legal consulting products are shown in Table 1.

Based on the above background, we design and implement a task-oriented automatic dialogue system based on the decision tree for real-time marriage legal consultation. The legal automatic dialogue system can realize multiple rounds of dialogue and provide accurate answers in real-time, which enables users to have a good interaction. Moreover, the method can be extended to other legal consultation easily. The main contributions of this paper are as follows:

1) Based on a parallel C4.5 decision tree, which is built from the case data we collect, an intelligent marriage consultation system is designed and implemented. The system can respond to similar inquiries for users intelligently, with the ability of reasoning.

2) The effect of different training set proportions on the maximum tree depth and the precision of the decision tree model of the legal automatic dialogue system are analyzed in our experiment.

II. THE PROPOSED METHODS

A. DESIGN OF LAW AUTOMATIC DIALOG SYSTEM BASED ON DECISION TREE

There are four modules in the system as shown in Figure 1. The data collection module is to craw data and collect data from web. The data preprocessing module is to fill missing values and discretize continuous value of some attributes; The data learning module is to build a parallel decision

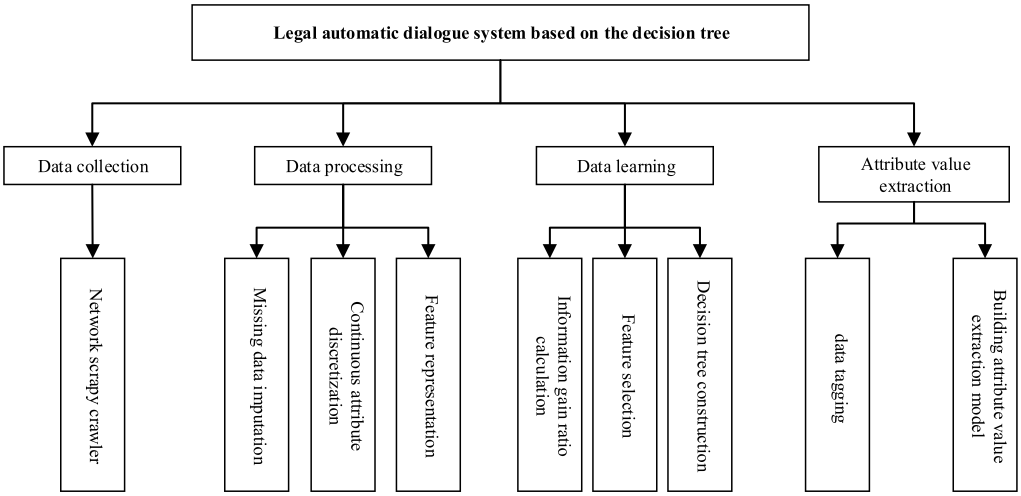

Legal automatic dialogue system based on the decision tree
Data collection
Data processing
Data learning
Attribute value extraction
Network scrapy crawler
Missing data imputation
Continuous attribute discretization
Feature representation
Information gain ratio calculation
Feature selection
Decision tree construction
data tagging
Building attribute value extraction model

FIGURE 1. System function module diagram.

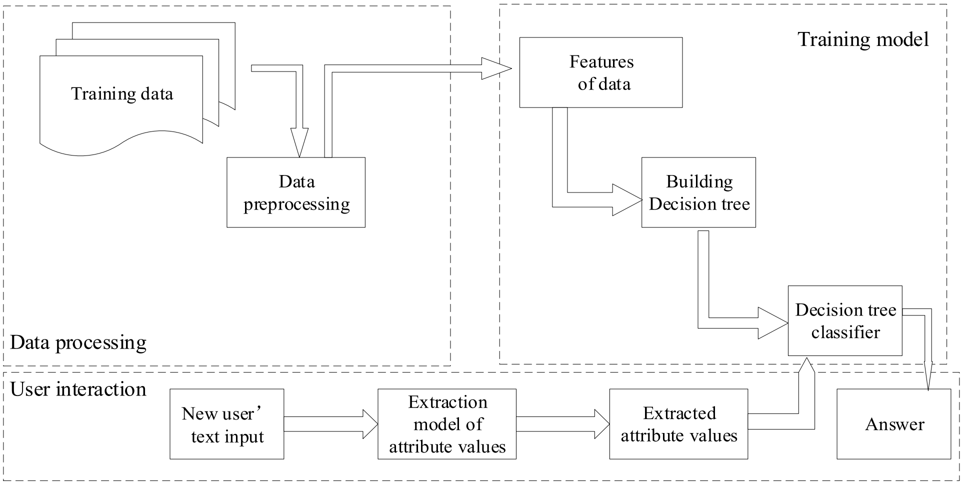


FIGURE 2. System architecture diagram.

tree; attribute value extraction module is to establish models according to different types attributes.

From the architecture point of view, the four modules are integrated into three parts as shown in Figure 2: data preprocessing model is to obtain characteristic representation of data; training model is to use characteristic representation data to build a parallel C4.5 decision tree; user interaction model is to extract attributes from user's input, and replies to users.

B. THE PROCESS OF BUILDING A PARALLEL C4.5 DECISION TREE

The decision tree of the legal automatic dialogue system is built based on the information gain ratio of attribute as below[20].

1) SPLIT ATTRIBUTE SELECTION

Due to information gain ratio is based on information gain, and as we know that the information gain is based on an idea to decrease entropy for a data-set by splitting it on an attribute, and building a decision tree is all to select attribute that returns the highest information gain. Hence, we do in a way as below.

Suppose that the samples set D is divided by attribute A with domain $\{a_1, a_2, \ldots, a_j, \ldots, a_v\}$. Firstly, we should calculate the split information which measures the data distribution of split attribute. The split information of attribute A is:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log \frac{|D_j|}{|D|} \tag{1}$$

Then, the information gain is calculated. The information gain means the difference of information entropy after splitting with attribute A. The entropy measures the uncertainty of attributes, so the larger the information gain is, the better the split is. However, the information gain tends to select the attributes which have more values. The information gain of

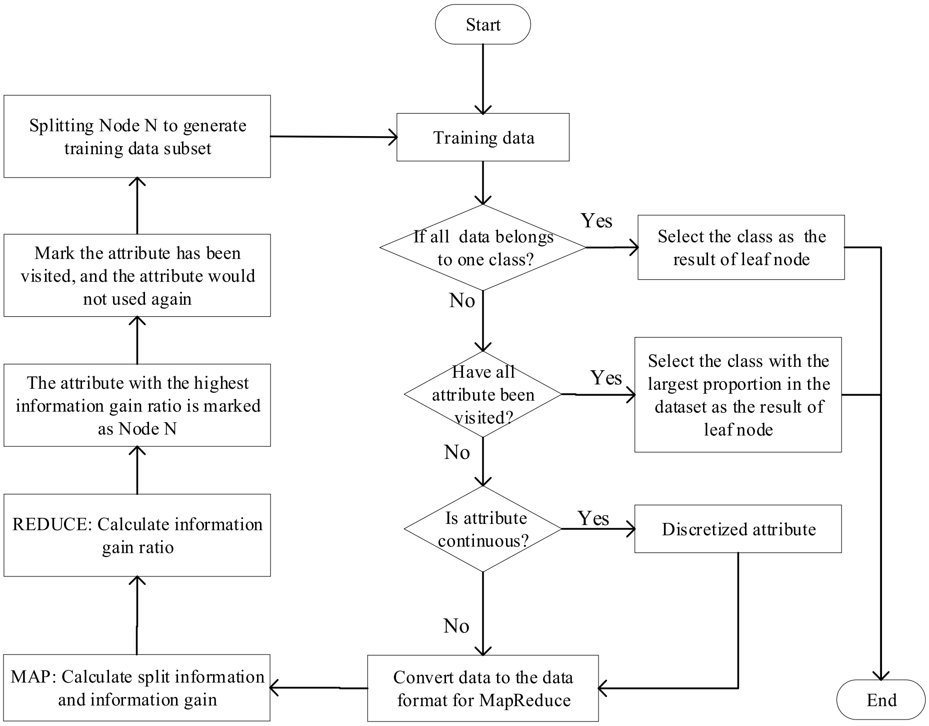

Start
Splitting Node N to generate training data subset
Training data
If all data belongs to one class?
Yes
Select the class as the result of leaf node
No
Mark the attribute has been visited, and the attribute would not used again
Have all attribute been visited?
Yes
Select the class with the largest proportion in the dataset as the result of leaf node
No
The attribute with the highest information gain ratio is marked as Node N
Is attribute continuous?
Yes
Discretized attribute
REDUCE: Calculate information gain ratio
No
MAP: Calculate split information and information gain
Convert data to the data format for MapReduce
End

FIGURE 3. The process of building a parallel decision tree.

set D after splitting with attribute A is:

$$Gain(A) = Ent(D) - \sum_{i=1}^{v} \frac{|D^j|}{|D|} Ent(D^j) \tag{2}$$

where $Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$ is information entropy, $|y|$ is the number of category, p_k represents the proportion of class k.

Hence, the information gain ratio of set D after splitting with attribute A is:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \tag{3}$$

Because information gain ratio considers both the data distribution and the information gain when selecting attributes, which avoid the disadvantage of information gain, it is reasonable to select an attribute A with the maximum $GainRatio$ to be the split attribute.

2) TREE CONSTRUCTION BASED ON MAPREDUCE

The overall process is shown in Figure 3. We can see that it is an iterative process, in order to speed-up the process, the tree is built in parallel by MapReduce[21], [22] in view of the problem that the system would slows down after data expansion. MapReduce is only carried out in the parallel phase of building decision tree model. and the process of MapReduce can be described as following: Firstly, the data should be transformed to the formation for MapReduce. Secondly, procedure MAP is to calculate split information and information gain. Thirdly, procedure REDUCE is to calculate information gain ratio. Finally, the information gain ratio obtained by MapReduce will be used for the selection of split attribute.

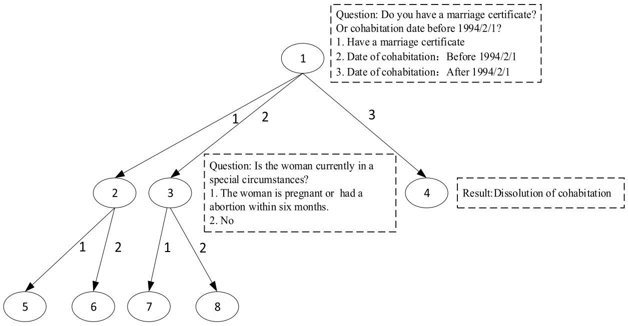

Question: Do you have a marriage certificate?
Or cohabitation date before 1994/2/1?
1. Have a marriage certificate
2. Date of cohabitation：Before 1994/2/1
3. Date of cohabitation：After 1994/2/1
1
1
2
3
2
3
Question: Is the woman currently in a special circumstances?
1. The woman is pregnant or had a abortion within six months.
2. No
4
Result:Dissolution of cohabitation
1
2
1
2
5
6
7
8

FIGURE 4. The example of building decision tree.

As Figure 4 shows, we can see that the decision tree summarizes decision rules from data, and presents these rules with tree structure, where each non-leaf node means a judgment on an attribute, and each leaf node represents a classification result. Hence, it is applicable to classify data for law consultation.

Another example is given as shown in Figure 4, when a customer asks a question about divorce, the system will ask the user whether he/she has a marriage certificate or cohabited before 1994/2/1. The system reaches the leaf node (Node 4) and returns a result which means dissolution of cohabitation if the user doesn't satisfy both conditions,

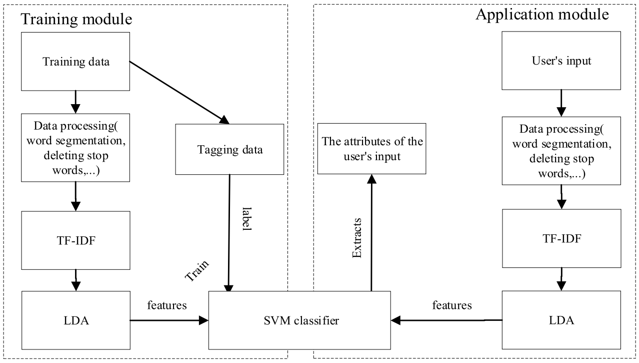

Training module
Training data
Data processing(word segmentation, deleting stop words,...)
TF-IDF
LDA
Tagging data
label
Train
features
SVM classifier
Application module
User's input
Data processing(word segmentation, deleting stop words,...)
TF-IDF
LDA
features
The attributes of the user's input
Extracts

FIGURE 5. The process of attribute value extraction.

otherwise, it will visit the next branch node and ask other question until leaf node is touched. For example, if the user has no marriage certificate and cohabited after 1994/2/1, then the system reaches the node 3, and asks the user whether the woman currently in a period of pregnancy or abortion within six months.

When the user consults the issue of law, the automatic process of legal automatic dialogue will be launched based on decision tree as shown in Figure 5. Firstly, the system starts from the root node of the decision tree according to the input of user. Secondly, after the corresponding attribute value of user's input is extracted, the system would judge whether the attribute value is reasonable for current node. If the value is appropriate, it would reach the next branch node of the decision tree and ask user the question which is related to the attribute value extracted above, otherwise, the system would ask user the same question. Finally, the dialogue is terminated if the current node is a leaf node, and return a result to the user.

C. ATTRIBUTE VALUE EXTRACTION

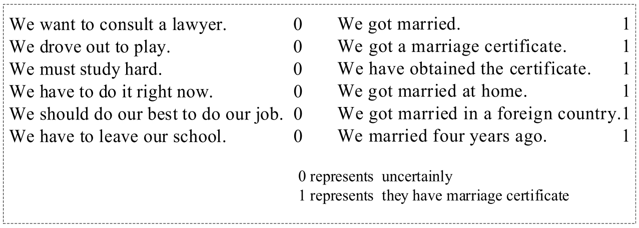
In order to remove noise for the user's input, and obtain the user basic attributes in user's input. Therefore, it is necessary to establish a discriminant model for each basic attribute to extract the key attribute value from the user's input accurately. The process of attribute value extraction is shown in Figure 5, which includes training module and the application module.

1) THE TRAINING MODULE

Firstly, the elements of training data are tagged with the corresponding labels. Secondly, the module is to preprocess the collected data, which includes removing the noise including word segmentation, deleting stop words, and eliminating low frequency words. Thirdly, the term frequency–inverse document frequency (TF-IDF) is calculated to obtain the feature vector of document. Fourthly, topic features are extracted by Latent Dirichlet Allocation (LDA) with the feature vector. Finally, SVM is applied to train the extracted topic features with the tagged labels above.

2) THE APPLICATION MODULE

When the user consults the legal issues, the user's input is preprocessed and extracted by LDA and TF-IDF, and then the trained SVM classifier will return a result which indicates the attributes of the user's input.



We want to consult a lawyer.	0	We got married.	1
We drove out to play.	0	We got a marriage certificate.	1
We must study hard.	0	We have obtained the certificate.	1
We have to do it right now.	0	We got married at home.	1
We should do our best to do our job.	0	We got married in a foreign country.	1
We have to leave our school.	0	We married four years ago.	1

0 represents uncertainly
1 represents they have marriage certificate

FIGURE 6. Example of attribute value extraction.

Taking whether there is a marriage certificate between the couple as an example, Figure 6 shows the result of extracting attributes from a dialog which is launched by a couple who consult divorce issue. 1 represents a positive class indicating that the marriage certificate is mentioned in the text, and 0 means the negative-class indicating that the marriage certificate is not mentioned in the text. When the user inputs text, the classification model will judge the category of input. If the input is the positive-class, the system would judge the existence of marriage certificate in user's the input by the recognition of affirmative and negative sentences. Otherwise, the dialogue system will ask the user again until the user's input is a positive class.

D. THEORETICAL ANALYSIS AND COMPARISON OF DECISION TREE CLASSIFIER MODEL

We use the C4.5 decision tree in the proposed system. In this subsection, the decision tree model is compared with Naive Bayesian (NB) and SVM in theory. The premise of the NB classifier model is that the attributes are independent of each other [23], so NB is not applicable in this system; The accuracy of SVM model often depends on the selection of support vector, and when there is a lot of data noise, it will seriously affect the performance of the SVM model [24]; Decision tree has strong comprehensibility and interpretability, and the characteristics of branch facilitate the generation of dialogue process of consultation [25]. The advantages and disadvantages of decision tree classifiers and other classifiers [26] is shown in Table 2.

III. EXPERIMENT

A. EXPERIMENTAL DATASET

In our experiment, we test our algorithm in judging divorce issues, and the related data set comes from website lvpin (https://ai.lvpin100.com), which is a legal consulting website, and the data on this website has been sorted out by many professionals. The data cardinality is 2304, and the data formation is shown in Table 3(all data have been translated from Chinese to English for understanding). In addition, the data formation after preprocessing, which is the training data of the classifier, and an example of classification are shown in Table 4 It is observed that there are four categories and eight basic attributes, and each attribute may have multiple

TABLE 2. Comparison of decision tree classifiers and other classifiers (M represents the number of features. S represents the number of categories. Epoch represents the number of iterations. C represents the number of hidden nodes. D represents the depth of the decision tree).

Classifiers	interpretability	Scale of data	Time complexity	Space complexity
NB	Good	Small-Medium	$O(N*M)$	$O(M*S)$
SVM	Poor	Small	$O(N^2)$	$O(N^3)$
Decision tree	Good	Small-Medium	$O(N*M*D)$	$O(N+M*S)$

TABLE 3. Original format of the sample data.

Attributes	Attribute value	Sample
Gender of the user	Man/woman	Man
Marriage certificate	Have a marriage certificate / no marriage certificate	Have a marriage certificate
Date of cohabitation	Before 1994/2/1 After 1994/2/1	After 1994/2/1
Date of marriage	date	2009/9/1
Special circumstances of the woman	1. The woman is pregnant 2. The woman will give birth to the child in one year 3. The woman had a miscarriage within six months 4. Neither	1
Military marriage	1. The man is an active non-civilian soldier 2. The woman is an active non-civilian soldier 3. Both sides are active non-civilian soldiers 4. Neither side is an active non-civilian soldier	4
Divorce attitude	1. You want to divorce, another party does not want to divorce or the attitude is unknown. 2. You don't want to divorce, another party want to divorce 3. Both sides want to divorce	3
Reasons for divorce	1. Derailed 2. Domestic violence 3. Living with others for more than six months 4. bigamy 5. Abandoning lover 6. Abuse 7. None of the above	2
Answer(output)	Both parties want to divorce, so both parties can go to the Civil Affairs Bureau to agree on a divorce or go to the place of court to litigation divorce.	

values. Moreover, the all attributes have been discretized into Numbers, and clustered Manually. Each attribute corresponds to a question that will be thrown to user if the decision tree traverses corresponding node, as shown in Table 5.

In the experiment of attribute value extraction, the data is collected by manual collection, and annotated by professor. The data cardinality is 753, and the data formation is the same as figure 6.

B. BTHE METHOD OF EVALUATION

The method of K-fold cross-validation is used to verify the experimental results, and obtain a reliable and stable validation model which prevents the model from over fitting [27].

C. METRICS FOR EVALUATION

The results of the classification are evaluated by the precision (P), recall rate (Recall, R) and F1-score. The formulas are:

$$P = \frac{TP}{TP+FP} \tag{4}$$

$$R = \frac{TP}{TP+FN} \tag{5}$$

$$F1 = \frac{2P*R}{P+R} \tag{6}$$

Meanwhile, the average number of questions is also used to evaluate our system. The fewer questions required, the faster the system can understand the real intention of the user. The metric can be reflected by the depth of the decision tree, and the formula is:

$$h_{avg} = \frac{\sum_{i=1}^{N} deep_i - 1}{N} \tag{7}$$

where i is the i^{th} leaf node, N is the total number of leaf nodes, $deep_i$ is the depth of the i^{th} leaf node.

D. MACHINE CONFIGURATION

The system is written in Python, and experiments are conducted on windows 10 with Intel(R) Xeon(R) CPU @ 2.30GHz and 12G RAM.

E. RESULTS OF ATTRIBUTE VALUE EXTRACTION

In this subsection, we conduct experiment on the data in eight-fold cross validation method to compare LSTM model and SVM model in attribute value extraction. The LSTM method takes word vector as word feature after data preprocessing, the number of LSTM layer is 2, and the number of units is 128. There are 602 data as training samples, 151 as test set in this experiment. The performances of the SVM method and its competitor are shown in Table 6 (Avg means average, std represents standard deviation). It is observed that all average scores of the SVM model are above 97%, which outperforms LSTM model, indicating that the SVM model can be better applied to the extraction and discrimination of attribute values for further processing by decision tree. Moreover, the standard deviation of SVM model is lower than

TABLE 4. Data format after data preprocessing.

Attributes	Attribute value	Sample
Gender of the user	1. Man 2. woman	1
Marriage certificate and Date of cohabitation	1. Have a marriage certificate 2. Date of cohabitation：Before 1994/2/1 3. Date of cohabitation：After 1994/2/1	1
Special circumstances of the woman	1. The woman is pregnant or gave birth to the child within one year or had a miscarriage within six months 2. No	1
Military marriage	1. The man is an active non-civilian soldier 2. The woman is an active non-civilian soldier 3. Both sides are active non-civilian soldiers 4. Neither side is an active non-civilian soldier	1
Divorce attitude	1. The man wants to divorce 2. The woman wants to divorce 3. Both sides want to divorce	1
Prosecuted in court	1.Yes 2.No	1
Reasons for divorce	1. The man is derailed, domestic violence, living with others for more than six months, bigamy, abandonment, abuse, or the whereabouts are not clear for two years. 2. The woman is derailed, domestic violence, living with others for more than six months, bigamy, abandonment, abuse, or the whereabouts are not clear for two years. 3. Both have these 4. Neither side has these	1
Separated situation	1. Separated for two years or more 2. Not separated or separated for less than two years	1
Classification(output)	1. Agreement divorce 2. Litigation divorce 3. Can't divorce 4. Dissolution of cohabitation	3

TABLE 5. The questions corresponding to the attributes.

Attributes	Corresponding question
Gender of the user	Are you a man or a woman?
Marriage certificate and Date of cohabitation	Do you have a marriage certificate? Or cohabitation date before 1994/2/1?
Special circumstances of the woman	Is the woman currently in a special circumstance?
Military marriage	Who is a non-civilian soldier in active service?
Divorce attitude	What is your attitude towards divorce?
Prosecuted in court	Have you ever been to the court to sue for divorce?
Reasons for divorce	Are you currently separated and separated for two years or more?
Separated situation	Whether the two sides have derailed, domestic violence, living with others for more than six months, bigamy, abandonment, abuse, and the whereabouts of two years?

TABLE 6. Result of attribute value extraction.

Algorithm	Cross-validation	Precision	Recall	F1-score
SVM	Avg	0.978	0.977	0.977
	Std	0.017	0.019	0.017
LSTM	Avg	0.97	0.968	0.968
	Std	0.15	0.011	0.13

LSTM model, which intends that the SVM model is more stable than LSTM model.

F. COMPARISONS OF C4.5 DECISION TREES WITH OTHER ALGORITHMS

Due to the good performance of C4.5, we mainly use it as the decision tree in the proposed method. In this subsection, some experiments are conducted to compare C4.5 decision tree with other classification algorithms, such as SVM and NB.

The first experiment is to analyze the influence of the different proportions of the training set. The result of precision is shown in Figure 7, the result of training time cost is presented in Figure 8, and the time cost of 1000 times predictions on the same test set is shown in figure 9.

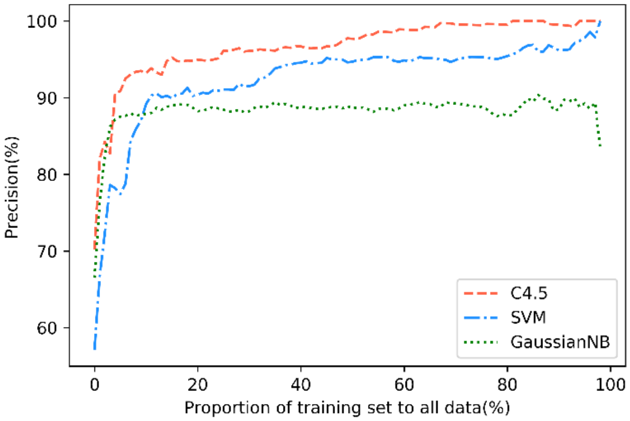


FIGURE 7. Classification precision at different scales.

In the view of Figure 7, the larger the proportion of training set is, the higher the precision we obtain. Moreover, the precision of the C4.5 decision tree model is higher than SVM and NB after 5% with 90% proportion of training set, which indicates that the decision tree needs less data to achieve better results than the other models.

As can be seen from Figure 8, the training time is increased by the proportion of training set. The shorter is the model,

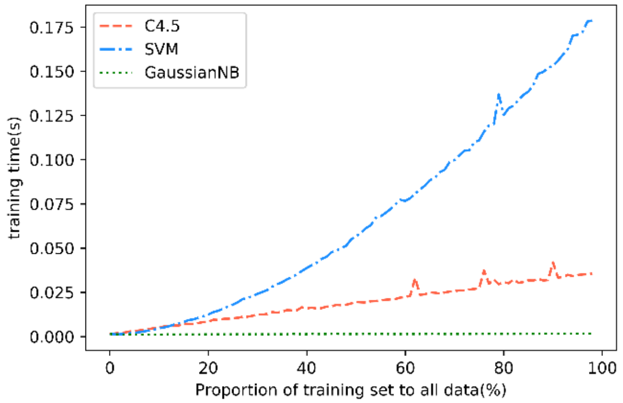


FIGURE 8. Training time cost at different scales.

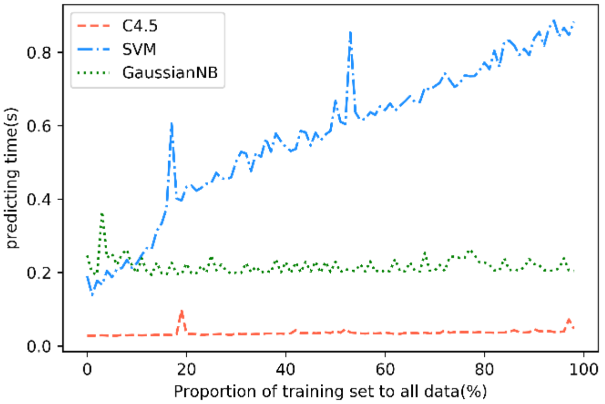


FIGURE 9. Prediction time cost at different scales.

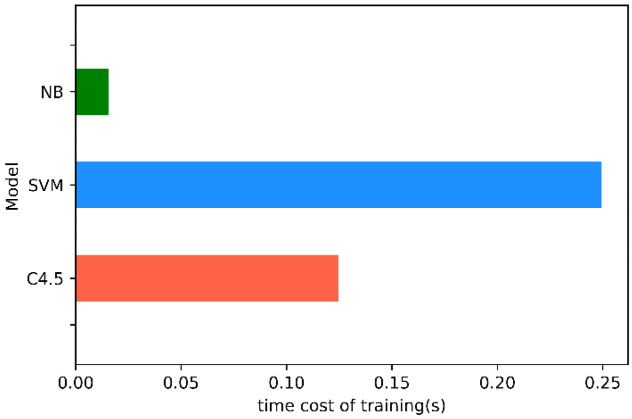


FIGURE 10. Training time cost of different model.

the better the model is. The training time of SVM model has the largest growth, followed by C4.5, and NB is the most stable, which shows that in the case of big data, SVM training cost is far more than C4.5 and NB.

Figure 9 shows that the prediction time of NB and C4.5 does not increase with the proportion of training set, but SVM increases with it. For the real-time requirement of the system, because of the time cost of SVM in the case of big

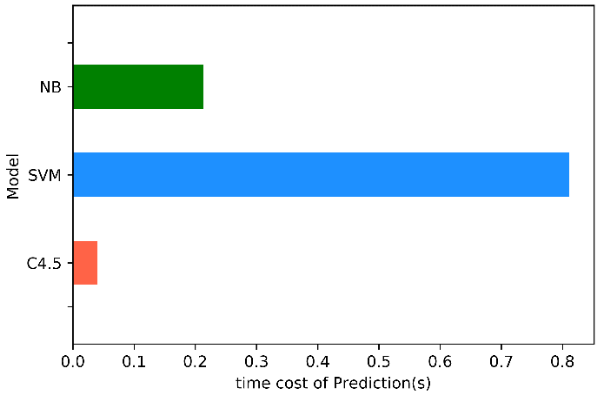


FIGURE 11. Prediction time cost of different model.

TABLE 7. The scores of the decision tree, SVM, NB.

Algorithm	Cross-validation	precision	recall	f1-score
C4.5	Avg	0.969	0.988	0.976
	Std	0.028	0.011	0.022
SVM	Mean	0.943	0.914	0.923
	Std	0.016	0.020	0.12
NB	Mean	0.875	0.851	0.862
	Std	0.033	0.044	0.26

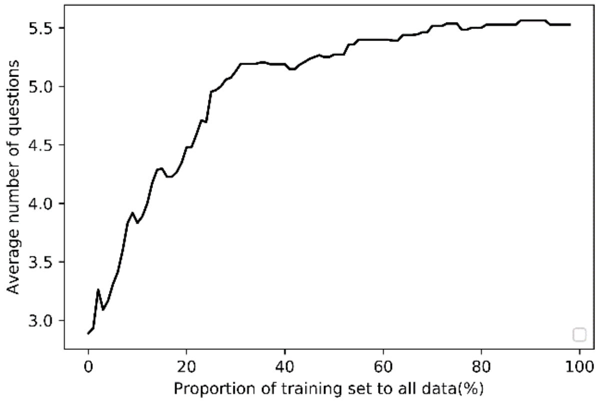


FIGURE 12. Average number of questions at different scales.

data, C4.5 and NB are more suitable for this system, and the real-time performance of C4.5 is the best.

The second experiment is conducted to verify the performance of the decision tree with 90% proportion of training set, and we also compare C4.5 with SVM and NB model. The scores of the C4.5 decision tree, SVM, NB are shown in Table 6, the training time costs are shown in Figure 10, and the predicting time costs are shown in Figure 11.

According to Table 7, the scores of the C4.5 decision tree model are all better and more stable than SVM and NB. Moreover, from Figure 10 and 11 we can see that the training time of the C4.5 decision tree is less than SVM model, but longer than NB model, and predicting time cost is better than other two algorithms which indicates better real-time performance. Comprehensively, the C4.5 decision tree is more applicable than the other two algorithms.

The third experiment is to analysis the effects of different proportion training data on the depth of decision tree, and the experimental results are plotted in Figure 12.

According to Figure 12, it is observed that the higher the proportion of training set, the deeper the depth of decision tree. In addition, it can be found that the average depth of the decision tree is about 5.5, which indicates that the automatic dialogue system can return the consultation result after 5-6 questions averagely. It is fewer than that of website lvpin (the average number of questions is about 8) and the SVM, NB model which should need all attributes to make a predict. Therefore, the decision tree model reduces some useless questions, which improves the efficiency of consultation.

IV. CONCLUSION

In order to realize real-time legal consultation automatically, we design an automatic legal marriage consultation system based on the parallel C4.5 decision tree for divorce issues. It responds to users intelligently with the ability of reasoning, which yields higher accuracy than SVM and NB model. Compared with some automatic legal consultation websites, the proposed method needs fewer questions asked to the user during a dialog, which improves the efficiency of consultation.

However, due to the low efficiency of attribute extraction experts' manual tagging, and the metric data in this case may not correspond to the user opinion. Therefore, it is suggested that the process be crowdsourced and labeled by users themselves to reduce the cost of manual labeling and improve the accuracy of attribute value extraction.

Our future work is to develop a new version of the proposed method by using fast clustering [28]–[30] and CNN [31] based time series data mining to deal with complex consultation. Also, we would optimize the depth of the decision tree, and prevent the tree building from overfitting.

REFERENCES

[1] L. Chen, J. M. Jose, H. Yu, and F. Yuan, "A hybrid approach for question retrieval in community question answerin," *Comput. J.*, vol. 60, no. 7, pp. 1019–1031, 2016.

[2] B. Morgan, "Lawyers, legal advice and relationality in sustainable economy initiatives," *Oñati Socio-Legal Ser.*, vol. 7, no. 7, p. 22, 2017.

[3] S. Basuki, A. Rizky, and G. W. Wicaksono, "Case based reasioning (CBR) for medical question answering system," *Kinetik: Game Technol., Inf. Syst., Comput. Netw., Comput., Electron., Control*, vol. 3, no. 2, pp. 113–118, Apr. 2018.

[4] Y. Liu and X. Yang, "A similar legal case retrieval system by multiple speech question and answer," in *Proc. 18th Int. Conf. Electron. Bus. (ICEB)*, Guilin, China, 2018, pp. 72–81.

[5] Z. Li, "A classification retrieval approach for English legal texts," in *Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS)*, Jan. 2019, pp. 220–223.

[6] B. Fawei, J. Z. Pan, M. Kollingbaum, and A. Z. Wyner, "A semi-automated ontology construction for legal question answering," *New Gener. Comput.*, vol. 37, no. 4, pp. 453–478, Dec. 2019.

[7] M.-Y. Kim and R. Goebel, "Two-step cascaded textual entailment for legal bar exam question answering," in *Proc. 16th Ed. Int. Conf. Articial Intell. Law (ICAIL)*, 2017, pp. 283–290.

[8] G. Greenleaf, A. Mowbray, and P. Chung, "Building sustainable free legal advisory systems: Experiences from the history of AI & law," *Comput. Law Secur. Rev.*, vol. 34, no. 2, pp. 314–326, 2018.

[9] D. S. Carvalho, D.-V. Tran, V.-K. Tran, and L.-N. Minh, "Improving legal information retrieval by distributional composition with term order probabilities," 2017, *arXiv:1706.01038*. [Online]. Available: http://arxiv.org/abs/1706.01038

[10] N. Bansal, A. Sharma, and R. Singh, "A review on the application of deep learning in legal domain," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.* Cham, Switzerland: Springer, 2019, pp. 374–381.

[11] A. Mishra and S. K. Jain, "A survey on question answering systems with classification," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 28, no. 3, pp. 345–361, Jul. 2016.

[12] K. Höffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A.-C. N. Ngomo, "Survey on challenges of question answering in the semantic Web," *Semantic Web*, vol. 8, no. 6, pp. 895–920, Aug. 2017.

[13] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Understand.*, vol. 163, pp. 21–40, Oct. 2017.

[14] B. Fawei, J. Z. Pan, M. Kollingbaum, and A. Z. Wyner, "A methodology for a criminal law and procedure ontology for legal question answering," in *Proc. Joint Int. Semantic Technol. Conf.* Cham, Switzerland: Springer, 2018, pp. 198–214.

[15] Y. Zhang, H. Dai, Z. Kozareva, A. J. Smola, and L. Song, "Variational reasoning for question answering with knowledge graph," in *Proc. 22nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[16] Y. Lin, H. Ji, Z. Liu, and M. Sun, "Denoising distantly supervised open-domain question answering," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 1736–1745.

[17] S. Wang, "R 3: Reinforced ranker-reader for open-domain question answering," in *Proc. 22nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[18] S. Bi, Y. Huang, X. Cheng, M. Wang, and G. Qi, "Building Chinese legal hybrid knowledge network," in *Proc. Int. Conf. Knowl. Sci., Eng. Manage.* Cham, Switzerland: Springer, 2019, pp. 628–639.

[19] Z. Wang, B. Wang, X. Duan, D. Wu, S. Wang, G. Hu, and T. Liu, "IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), Syst. Demonstrations*, 2019, pp. 97–102.

[20] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4. 5," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 2, p. 13–19, 2014.

[21] W. Dai and W. Ji, "A mapreduce implementation of C4. 5 decision tree algorithm," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 49–60, 2014.

[22] Y. Mu, X. Liu, Z. Yang, and X. Liu, "A parallel C4.5 decision tree algorithm based on MapReduce," *Concurrency Computat., Pract. Exper.*, vol. 29, no. 8, Apr. 2017, Art. no. e4015.

[23] B. Oviedo, C. Zambrano-Vega, J. León-Acurio, and A. Martinez, "Simple Bayesian classifier applied to learning," in *Proc. Int. Conf. Technol. Trends*. Cham, Switzerland: Springer, 2018, pp. 399–409.

[24] R. Wang, W. Li, R. Li, and L. Zhang, "Automatic blur type classification via ensemble SVM," *Signal Process., Image Commun.*, vol. 71, pp. 24–35, Feb. 2019.

[25] Z. Lin, H. Chen, P. Li, X. Guo, and H. Yang, "A review: The effects of imperfect data on incremental decision tree," *Int. J. Inf. Commun. Technol.*, vol. 12, nos. 1–2, p. 162, 2018.

[26] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[27] T.-T. Wong and N.-Y. Yang, "Dependency analysis of accuracy estimates in k-fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2417–2427, Nov. 2017.

[28] Y. Chen, L. Zhou, S. Pei, Z. Yu, Y. Chen, X. Liu, J. Du, and N. Xiong, "KNN-BLOCK DBSCAN: Fast clustering for large-scale data," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published.

[29] Y. Chen, X. Hu, W. Fan, L. Shen, Z. Zhang, X. Liu, J. Du, H. Li, Y. Chen, and H. Li, "Fast density peak clustering for large scale data based on kNN," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104824.

[30] Y. Chen, S. Tang, N. Bouguila, C. Wang, J. Du, and H. Li, "A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data," *Pattern Recognit.*, vol. 83, pp. 375–387, Nov. 2018.

[31] S. Pei, T. Shen, X. Wang, C. Gu, Z. Ning, X. Ye, and N. Xiong, "3DACN: 3D Augmented convolutional network for time series data," *Inf. Sci.*, vol. 513, pp. 17–29, Mar. 2020.