Decision Support Systems 66 (2014) 114–122

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

An ontology-based Web mining method for unemployment rate prediction

Ziang Li [a], Wei Xu [a,*], Likuan Zhang [a], Raymond Y.K. Lau [b]

[a] School of Information, Renmin University of China, Beijing, China
[b] Department of Information Systems, City University of Hong Kong, Hong Kong

失业率预测
基于本体的Web数据挖掘方法

ARTICLE INFO

Article history:
Received 15 November 2012
Received in revised form 1 May 2014
Accepted 11 June 2014
Available online 20 June 2014

Keywords:
Unemployment rate prediction
Search engine query data
Domain ontology
Web mining
Neural networks
Support vector regressions

ABSTRACT

Unemployment rate is one of the most critical economic indicators. By analyzing and predicting unemployment rate, government officials can develop appropriate labor market related policies in response to the current economic situation. Accordingly, unemployment rate prediction has attracted a lot of attention from researchers in recent years. The main contribution of this paper is the illustration of a novel ontology-based Web mining framework that leverages search engine queries to improve the accuracy of unemployment rate prediction. The proposed framework is underpinned by a domain ontology which captures unemployment related concepts and their semantic relationships to facilitate the extraction of useful prediction features from relevant search engine queries. In addition, state-of-the-art feature selection methods and data mining models such as neural networks and support vector regressions are exploited to enhance the effectiveness of unemployment rate prediction. Our experimental results show that the proposed framework outperforms other baseline forecasting approaches that have been widely used for unemployment rate prediction. Our empirical findings also confirm that domain ontology and search engine queries can be exploited to improve the effectiveness of unemployment rate prediction. A unique advantage of the proposed framework is that it not only improves prediction performance but also provides human comprehensible explanations for the changes of unemployment rate. The business implication of our research work is that government officials and human resources managers can utilize the proposed framework to effectively analyze unemployment rate, and hence to better develop labor market related policies.

我们的实证研究结果也证实，领域本体和搜索引擎查询可以被利用来提高失业率预测的有效性。#只要发现一点提高就可以作为创新点。

1. Introduction

Unemployment rate can influence the rates of treasury bills and the financial market as a whole. In fact, any unexpected changes of unemployment rate can substantially affect consumers' spending because these changes influence households' perceptions and expectations about the economic conditions [23]. Accordingly, financial analysts can predict the economic trend of a targeted market by analyzing the unemployment rate of the corresponding nation. Moreover, government officials and human resource managers can develop appropriate human resources related policies by analyzing and predicting unemployment rate. Unemployment rate prediction has become increasingly more important in recent years because of the financial turbulence in different continents of the world. Accordingly, unemployment rate prediction has attracted much attention from governments, businesses, and researchers.

A large number of methods have been proposed to predict unemployment rate. Initially, univariate time series models were proposed to predict unemployment rate [3,16,30,32]. For example, autoregressive fractionally integrated moving average (ARFIMA) was developed to predict unemployment rate, and the empirical results showed that ARFIMA had a better predictive performance than the threshold autoregressive (TAR) and the symmetric ARFIMA models [16]. Alternatively, a time deformation model was developed to predict the trend of unemployment, and the experimental results indicated that the proposed model outperformed some classical forecasting models such as the autoregressive integrated moving average (ARIMA) model [32]. Furthermore, economic and social factors including gross national product (GNP), money supply, and interest rates were taken into account to improve the accuracy of unemployment rate prediction [12,14,15,21,25,27]. For example, GNP was applied to construct an unemployment rate prediction model [12], while gross domestic product (GDP) was leveraged to build an alternative prediction model [15]. Moreover, money supply, producer price index, and interest rate were incorporated into a prediction model to forecast unemployment rate [21].

In the era of Web 2.0, user-contributed information on the Web has been regarded as a valuable resource to analyze social or economic hotspots such as the financial market [2,17,28]. A kind of user-contributed data of the Social Web, namely search engine query data, was applied to detect influenza epidemics [10,32] and predict unemployment rate [1]. In addition, a Web-based model that was constructed

* Corresponding author. Tel.: +86 10 82500904.
E-mail address: weixu@ruc.edu.cn (W. Xu).

http://dx.doi.org/10.1016/j.dss.2014.06.007
0167-9236/

based on users' activities on the Internet was applied to identify the relationship between the frequency of Web search and unemployment rate [1]. The empirical experiments showed that the proposed Web-based model had a potential to enhance unemployment rate prediction [1]. Google Index (GI), an Internet job-search indicator, was regarded as one of the leading indicators for unemployment rate prediction [6]; the predictive power of GI was examined in the context of quarterly unemployment rate prediction [7]. Similarly, the huge volume of Web search data captured by Google was applied to predict contemporaneous economic indicators before government officials actually announced the figures [29]. A set of Web search queries was also applied to model unemployment time series, and the empirical results showed that it could significantly improve forecasting accuracy [4,5]. Furthermore, search engine queries were leveraged to predict unemployment rate using neural networks instead of traditional statistical models, and the experimental results demonstrated that the proposed method outperformed traditional statistical prediction models [33,35]. Finally, a hybrid forecasting model that combined search engine query data and time series data was developed to improve the performance of unemployment rate prediction [34].

Among the previous studies that leveraged search engine queries to predict unemployment rate, there are two common methods to retrieve search engine data. The first method is to collect thousands of search engine queries, and then select a subset of relevant queries using some feature selection methods [10,33]. The second one is to directly select relevant queries according to some pre-defined topics [4]. However, the first method suffers from the problem of inefficiency in applying feature selection methods to extract useful features from a large number of queries, while the second method may lead to insufficient number of features extracted based on a small number of pre-defined topics. As for the phase of unemployment rate prediction, statistical methods have been widely used in previous studies [1,4–7]. Comparatively speaking, few data mining tools such as support vector regression were applied to predict unemployment rate in previous research.

One of the main contributions of our research work is the development of a novel ontology-based Web mining framework that exploits search engine query data to enhance unemployment rate prediction. In particular, the proposed framework is underpinned by a domain ontology that captures the prominent concepts and their semantic relationships related to the problem domain of unemployment; the proposed ontology-based method alleviates the weakness of some existing Web-based methods that cannot effectively extract relevant queries from among a large number of possibly noisy search engine queries. The domain ontology also contributes to enhance automated feature selection which aims to reduce the dimensionality of the training query data and improve prediction accuracy. In addition, various data mining methods have been explored in our study, and then the best prediction model is identified via our cross-validation approach. Finally, the most effective prediction model and the best subset of query data are applied to predict unemployment rate.

The rest of this paper is organized as follows. Section 2 introduces the basic concepts of ontology and the data mining tools such as neural networks (NNs) and support vector regressions (SVRs) examined in our study. The ontology-based Web mining framework for unemployment rate prediction is then illustrated in Section 3. For the assessment of the efficiency and effectiveness of the proposed framework, empirical experiments are performed and the experimental results are reported in Section 4. Finally, we offer concluding remarks and summarize the future directions of our research work.

2. Theoretical foundations

In this section, the theoretical foundations of some computational methods, which are applied to construct the proposed ontology-based Web mining framework, are briefly described. More specifically, the basic concept and the formal definition of domain ontology are first introduced. Then, some data mining methods such as neural networks and support vector regressions are illustrated. These computational methods underpin the development of an effective and efficient ontology-based Web mining framework for unemployment rate prediction.

2.1. The formal definition of domain ontology

The concept of ontology is first explored in the field of philosophy; ontology is often represented by a hierarchy of concepts and other semantic information. According to existing literature, domain ontology captures concepts and their relationships to a specific domain as well as representing the axioms (e.g., rules) and constraints that define the prominent features of the domain [8,9,19,20,37]. It is a formal and generic way to represent a set of related concepts of a domain so that different people can reuse and apply this domain knowledge. Ontology is popular in describing domain knowledge due to its distinct advantage of promoting reusability. Although various definitions of ontology are proposed by scholars in different fields, there is no confusion upon the usage of ontology from the perspective of data and knowledge engineering.

Ontology has been widely used in the field of information systems; it has been applied to construct causal maps [37], supporting research management [13], and enhancing adaptive learning [18]. In a previous study, a domain ontology that formally captures the concepts of financial news articles, market participants, issuers, and financial instruments was built to examine the relationships between financial news articles and financial instruments [38]. Our proposed framework is grounded in the notion of domain ontology. In particular, the semantic relationships among different concepts of the domain of labor economics are first identified. Then, a causal map that can be used as a basis to explain different events (e.g., increasing or decreasing unemployment rate) pertaining to the labor market is constructed and represented in the form of a semantically rich domain ontology. Finally, the domain ontology is applied to extract relevant search engine queries and select useful features for unemployment rate prediction. The proposed unemployment ontology is built and refined using well-known and effective knowledge engineering tools such as Protégé [24]. The proposed domain ontology that represents various concepts of labor economics is formally defined as follows.

Definition 1. Domain ontology

A domain ontology is a septuple $Ont = \langle X, A, C, R_{XC}, R_{AC}, R_{CC}^{CAS}, R_{CC}^{NCAS} \rangle$, where X,A,C represent finite sets of objects, attributes, and concepts, respectively. The relation $R_{XC} : X \times C \mapsto [0, 1]$ maps the set of objects X to the set of domain concepts C for all $x_i \in X$, $c_i \in C$.The relation $R_{AC} : A \times C \mapsto [0, 1]$ defines the mapping between the set of domain concepts C and the set of attributes A applied to describe these concepts. The relation $R_{CC}^{CAS} : C \times C \mapsto [0, 1]$ maps the finite set of domain concepts C through the causal relations, and the relation $R_{CC}^{NCAS} : C \times C \mapsto [0, 1]$ defines the association (i.e., non-causal) relations among the finite set of concepts C.

Fig. 1 is a snapshot view of a segment of the proposed domain ontology applied to perform unemployment rate prediction. For our domain ontology model, the set of attributes A refers to linguistic terms applied to describe the set of concepts C. For instance, the concept of "Unemployment Situation in the U.S." is represented by the single term (attribute) "Unemployment" at the center of Fig. 2. Moreover, the set of objects X refers to the set of relevant Web queries pertaining to the set of unemployment related concepts C. For instance, the Web query "U.S. Department of Labor on Age Discrimination" is about the concept "Age and Employment in the U.S." that is represented by the single attribute "Age" in Fig. 1. Directed arrow lines indicate causal relationships, that is, from a concept denoting a possible cause to the concept describing

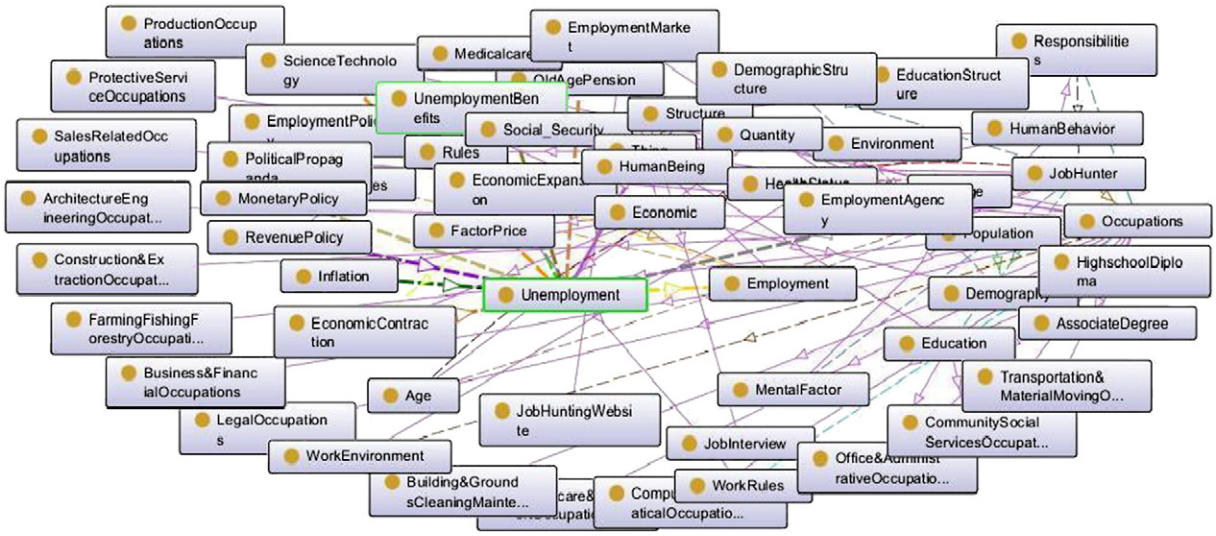

Fig. 1. A segment of domain ontology for unemployment rate prediction.

the likely consequence. On the other hand, undirected lines represent simple association relationships among concepts.

2.2. Data mining methods

For the proposed ontology-based Web mining framework, data mining methods are applied to select useful features from search engine queries and predict unemployment rate. Data mining methods have been shown to be useful to discover novel knowledge from both structured and semi-structured data. In order to conduct a comprehensive evaluation of different feature selection methods, both filter-based feature selection method and wrapper-based feature selection method are examined in our study. Moreover, different support vector regression models such as ε-SVR, ν-SVR, and different artificial neural network models such as back propagation neural network and radical basis function neural network are explored to evaluate their effectiveness in predicting unemployment rate.

For the feature selection component of the proposed framework, our wrapper-based feature selection method leverages forward selection, backward selection, and a genetic algorithm to analyze and extract the most discriminative features that affect unemployment rate. For the prediction component of the proposed framework, both neural networks and support vector regressions are applied. Neural network is a computational model that mimics the structure of human neural network [23]. Generally speaking, the interconnected artificial neurons are distributed in different layers, namely input layer, hidden layer(s), and output layer, and the structure of an artificial neural network can be tuned during the training phase. Neural networks have been shown to be effective to support a variety of business applications, especially for detecting the non-linear relationships between input and output data. Among a variety of neural network models, back propagation neural network (BPNN) is a classical model, in which the prediction error is estimated and minimized after iteratively feeding the training data to the network and comparing the predicted and the actual outputs. Meanwhile, the support vector regression that is proposed by

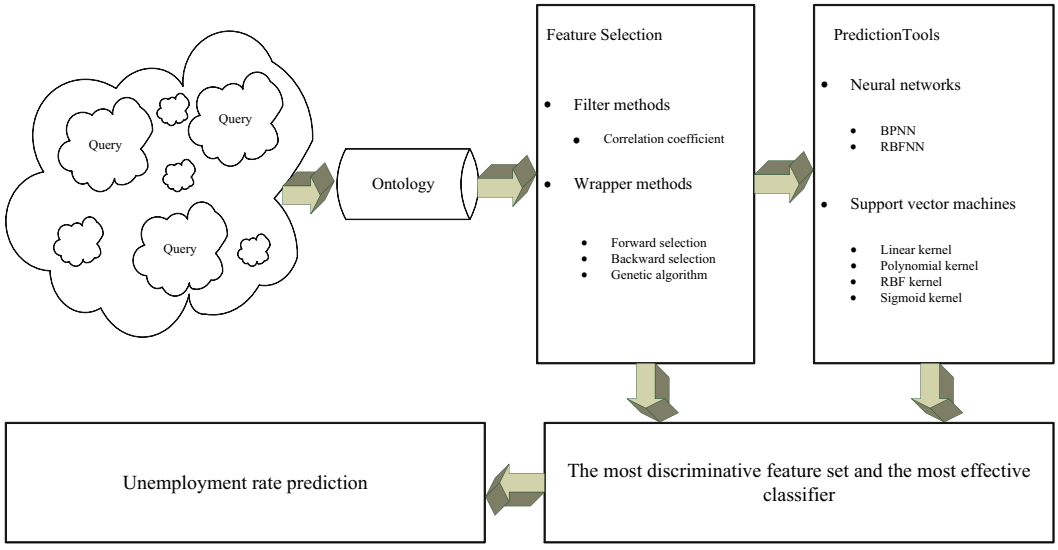

Fig. 2. The proposed ontology-based Web mining framework for unemployment rate prediction.

Vapnik [31] is the extension of the classical support vector machines by incorporating regression functionality. The operating principle of SVR is to first map input data to a high-dimensional feature space, and then using linear regression to find a solution with respect to the high-dimensional feature space.

3. Ontology-based Web mining framework for unemployment rate prediction

In this section, an ontology-based Web mining framework is proposed for unemployment rate prediction. In particular, a semantically rich domain ontology is applied to extract highly relevant queries from the initial set of possibly noisy search engine queries. The proposed feature selection methods are then applied to extract the most discriminative features from the set of relevant search engine queries for unemployment rate prediction. Finally, data mining methods are leveraged to discover the hidden relationships between the features extracted from relevant search engine queries and the movement of unemployment rate. An overview of the proposed framework is illustrated in Fig. 2.

Fig. 2 shows that the relevant search engine queries pertaining to unemployment are first extracted by using a semantically rich domain ontology that captures the prominent concepts of labor economics. The proposed feature selection module that consists of a hybrid filter-based method and wrapper-based method is then invoked to extract the most discriminative features for later prediction. Commonly used data mining methods such as neural networks and support vector regression are invoked to explore the hidden relationships between the features extracted from search engine queries and the movements of unemployment rate. The most discriminative features and the most effective classifier are identified by means of a cross-validation approach. Finally, the most effective classifier and the corresponding set of discriminative features are applied to predict unemployment rate. The computational details of the proposed framework will be illustrated in the following subsections.

3.1. Ontology construction

Several well-known and effective knowledge engineering tools are applied to construct the proposed domain ontology. In this study, Protégé [24], a java-based open source ontology editor developed by the School of Medicine of Stanford University, was adopted as the knowledge engineering tool to build our domain ontology. The main advantage of Protégé is that it has a user-friendly interface, and it supports the widely used Web Ontology Language (OWL). Among the existing ontology languages, the expressive power of OWL is relatively high when compared to other ontology languages such as RDF and XML [22]. OWL is relatively easy to be applied to describe domain concepts even for a novice ontology builder. Accordingly, we adopt OWL to formally represent concepts and their semantic relationships pertaining to the domain of labor economics.

The whole process of domain ontology construction consists of the following steps. The first step is to elicit domain concepts from various knowledge sources. Concept elicitation is extremely difficult because we are not the experts of the labor economics domain. A large number of terminologies pertaining to labor economics are retrieved from reference books covering the topics of economics and labor markets [26]. Fundamental labor economics theories and concepts are also elicited from several domain experts who are the senior scholars or experienced practitioners of the respective fields. Meanwhile, some domain concepts are extracted from relevant search engine queries. In this study, the queries related to unemployment and labor market are gathered from well-known Web search engines such as Google. After eliciting sufficient number of domain concepts, the next step is to identify the semantic relations among these concepts. These mantic relations were elicited from several experts in the field of labor economics. The third step is to encode the domain ontology. Our domain ontology mainly contains three kinds of elements, that is, classes (concepts), attributes (properties), and the relationships among classes. A segment of our domain ontology is presented in Fig. 1. During the ontology refinement stage, the Protégé reasoner, which is an external plug-in of Protégé, is applied to check the integrity and the correctness of the domain ontology. By using the Protégé reasoner, a variety of inference services can be invoked. For instance, we can apply the reasoner to infer the superclass of a given class (concept), or to determine whether one class has certain attributes.

After the refinement step of ontology construction, some prominent concepts pertaining to labor economics are encoded in the proposed domain ontology. For example, the concept “Revenue Policy” is related to “Unemployment”. The reason is that an effective revenue policy may lead to the decrement of unemployment rate. Some factors (concepts) which directly influence unemployment rate are depicted in Fig. 3. By consulting a domain ontology as shown in Fig. 3, relevant search engine queries that are related to unemployment can easily be retrieved.

3.2. Feature selection

By referring to the proposed domain ontology, a set of relevant Web queries is retrieved. However, this set of queries usually contains tens of thousands of terms (i.e., a high-dimensional feature space) which lead to low efficiency in subsequent prediction. Moreover, there are likely many noisy features (terms) which make the data mining tools produce wrong prediction. Therefore, a subset of features should be extracted from the relevant queries to improve prediction accuracy and minimize the computational complexity. Dependent on whether a feature selection method needs to consult the prediction results of a classifier, feature selection (FS) methods can be broadly classified into two categories, namely filter-based method and wrapper-based method [11]. For our study, one filter-based feature selection method and three wrapper-based methods are applied to extract the most discriminative features from search engine queries. For each wrapper-based method, different data mining tools such as NNs and SVRs are employed to estimate the prediction error as measured by Root Mean Square Error (RMSE), Mean Absolute Error (MAE), or Mean Absolute Percentage Error (MAPE).

3.2.1. Filter-based feature selection

The advantage of a filter-based feature selection (FBFS) method is its high efficiency since it does not involve the iterative estimation between feature selection and prediction. The correlation coefficient CC is used in the proposed framework to measure the correlation between each feature and the prediction. Given n pairs of values, including input feature variable x and the prediction value y, CC is defined as follows.

$$CC = \frac{\left(\sum xy - \frac{\sum x \sum y}{n}\right)^2}{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)} \quad (1)$$

The larger the CC value is, the stronger correlation between the input and the prediction is. The CC value of each feature is ranked in descending order, and the first m features are chosen to build the feature subset. From among the features with large CC values in the feature subset, the first k features are then iteratively selected by a classifier to perform prediction.

3.2.2. Wrapper-based feature selection

For wrapper-based feature selection (WBFS), forward feature selection (FFS), backward feature selection (BFS), and genetic algorithm (GA) are often used to extract discriminative features. The computational details of these methods are described in the following subsections.

3.2.2.1. Forward feature selection. The forward feature selection method is based on the simple wrapper-based feature selection approach. It begins with an empty target feature set. For each iteration of the selection process, the features of the initial feature set are progressively selected and added to the target feature set. The features contained in the current target feature set are used by a classifier to make prediction and the prediction error is then calculated. The target set of features that leads to the least prediction error is retained. The iterative feature selection process is terminated if none of the feature from the initial feature set could further improve the prediction performance.

3.2.2.2. Backward feature selection. Similar to forward feature selection, the backward feature selection method is also widely used. The initial target feature set consists of all features extracted from relevant search engine queries. For each iteration of the feature selection process, some features of the target feature set are progressively removed and the remaining features are used by a classifier to make prediction. The reduced target feature set that leads to the least prediction error is retained. The iterative feature selection process is terminated when no more features can be removed from the current target set to further improve prediction accuracy.

3.2.2.3. Genetic algorithm. A genetic algorithm is a mathematical method that mimics the biological reproduction process. A population contains a certain number of chromosomes which represent different feature subsets. In our experiment, each chromosome is represented by a binary vector, in which the number '1' indicates that the corresponding feature is selected, while '0' implies that the corresponding feature is not selected. The first population is generated randomly according to some genetic parameters (e.g., the size of the population). Then, a fitness function (e.g., RMSE) is applied to evaluate the merit of each chromosome. The fitness of a chromosome determines whether it can be selected to produce the new chromosomes of the next generation. From the second generation, the mutation and crossover operators are applied to generated new chromosomes based on the selected chromosomes of the previous generation. Mutation means that a chromosome experiences a sudden change for some of its genes, while crossover refers to the exchange of some genes between some selected chromosomes. The iterative evolution process continues until the maximum number of generations is reached.

4. Empirical analysis

4.1. Data description and evaluation criteria

For our experiments, the number of unemployment initial claims (UIC) released by the US Department of Labor is used as the proxy of the absolute unemployment rate because UIC is updated on a weekly basis and readily available on the Internet. We retrieved the UIC values between Jan.2004 and Mar. 2012 from the official website of the US Department of Labor (http://www.ows.doleta.gov/unemploy/claims.asp). Besides, weekly queries pertaining to the period from Jan.2004 to Mar.2012 were downloaded from Internet search engines. There were 423 relevant search engine queries extracted based on our domain ontology. Moreover, Google AdWords and Google Search Insight were applied to derive the frequency count of each relevant query.

The retrieved search engine queries and UIC values were divided into two groups to build the training set (Jan.2004 to Dec.2010) and the validation set (Jan. 2011 to Mar.2012), respectively. The training set was used for model tuning, and the validation set was applied to assess the performance of the trained model based on un-seen data. For the purpose of model tuning, 5-fold cross-validation was applied to each data mining method. In other words, the training set was divided into five subsets. For each run of the model tuning process, one subset was taken as the test set to assess the performance of a data mining method while the remaining four subsets were used to train the corresponding model (e.g., a NN or SVR model). Then, the average prediction error (e.g., measured by RMSE)of these five runs was taken to evaluate the performance of the particular data mining method.

The genetic parameters of our genetic algorithm were established before the data mining models were tuned. The population size was set as 30; the number of genes of each chromosome was set as 423 (i.e., the number of initial features established based on our domain ontology); and the maximum number of generations was defined as 50. The fitness value of a chromosome was computed by invoking a data mining method.

Furthermore, three performance measures, namely Root Mean Square Error, Mean Absolute Error, and Mean Absolute Percentage Error were applied to measure the performance of different feature selection methods and data mining models. Given n pairs of actual value (A_i) and predicted value (P_i), each of these performance measures is defined as follows.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (A_i - P_i)^2 \Big/ n} \tag{2}$$

$$\mathrm{MAE} = \sum_{i=1}^{n} |A_i - P_i| \Big/ n \tag{3}$$

$$\mathrm{MAPE} = \sum_{i=1}^{n} \frac{|A_i - P_i|}{A_i} \Big/ n \tag{4}$$

4.2. Experiments for model tuning

4.2.1. Filter-based approach

According to the proposed framework described in Section 3, the filter-based feature selection method is applied to the correlation

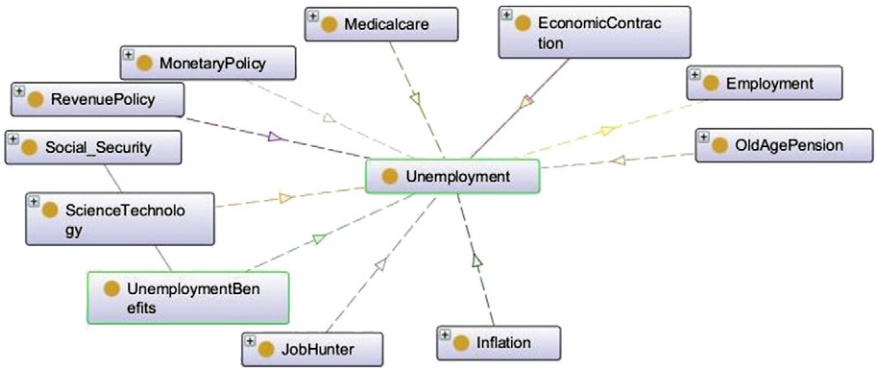

Fig. 3. The factors directly influencing unemployment.

Table 1
Performance of filter-based feature selection method.

Data mining model	Top n	Average RMSE	Average MAE	Average MAPE
CC	64	48,963	36,813	9.18
BPNN	76	68,151	50,638	12.71
RBFNN	34	91,875	63,509	15.05
ε-SVR(Linear)	22	49,878	37,037	9.50
ε-SVR(Polynomial)	8	67,519	49,464	11.82
ε-SVR(RBF)	92	51,390	35,825	8.92
ε-SVR(Sigmoid)	100	85,897	64,399	16.43
ν-SVR(Linear)	80	**47,419**	35,810	9.15
ν-SVR(Polynomial)	22	69,592	49,404	11.63
ν-SVR(RBF)	92	51,156	**34,239**	**8.32**
ν-SVR(Sigmoid)	100	76,565	58,151	15.00

The bold valuesare the minimum in terms of different error measurements.

coefficient model [1], two neural network models, and two support vector regression models in which four different kernel functions are implemented. For each data mining model, the average RMSE, average MAE and average MAPE based on five model tuning runs are calculated. The prediction performance achieved by each data mining model based on the best feature subset identified by filter-based feature selection method is depicted in Table 1.

According to Table 1, the ν-SVR(Linear kernel) model achieves the best performance in terms of RMSE with the first 80 features having the highest CC value, while the smallest MAE and MAPE are obtained when the first 92 features having the highest CC value are applied to the ν-SVR(RBF kernel) model. In contrast, the RBFNN model, the ε-SVR(Sigmoid kernel) model, and the ν-SVR(Sigmoid kernel) model perform poorly as evaluated according to RMSE, MAE, and MAPE, respectively.

4.2.2. Wrapper-based approach

Three wrapper-based feature selection methods are applied to each of the data mining method to predict unemployment rate. The average RMSE, MAE, and MAPE as computed based on five model tuning runs are listed in Table 2. In Table 2, FS stands for feature selection method; GA stands for genetic algorithm; FFS stands for forward feature selection, and BFS stands for backward feature selection.

Table 2
The performance of wrapper-based feature selection methods.

Data mining model	FS	Average RMSE	Average MAE	Average MAPE
BPNN	GA	77,786	57,639	14.62
	FFS	65,426	49,469	12.64
	BFS	69,611	52,297	13.31
RBFNN	GA	117,660	89,627	22.60
	FFS	80,761	58,598	14.66
	BFS	123,194	92,351	22.01
ε-SVR(Linear)	GA	68,618	50,633	8.51
	FFS	46,586	35,461	9.13
	BFS	44,084	34,427	8.91
ε-SVR(Polynomial)	GA	72,673	50,653	11.84
	FFS	49,667	37,069	9.34
	BFS	75,804	53,913	12.63
ε-SVR(RBF)	GA	52,629	37,295	9.34
	FFS	42,844	31,716	8.07
	BFS	48,035	35,329	9.02
ε-SVR(Sigmoid)	GA	56,405	40,950	10.32
	FFS	77,801	55,736	13.48
	BFS	50,365	36,392	9.18
ν-SVR(Linear)	GA	52,989	39,092	9.95
	FFS	45,570	34,516	8.84
	BFS	67,252	52,564	14.15
ν-SVR(Polynomial)	GA	73,535	51,009	11.91
	FFS	50,990	37,119	9.20
	BFS	76,005	54,068	12.69
ν-SVR(RBF)	GA	52,777	36,943	9.13
	FFS	**42,463**	**31,267**	**7.92**
	BFS	46,503	32,940	8.10
ν-SVR(Sigmoid)	GA	57,266	40,738	10.14
	FFS	47,092	34,769	8.88
	BFS	50,004	34,803	8.52

The bold valuesare the minimum in terms of different error measurements.

According to the result of this experiment, we can conclude that the forward selection approach outperforms the other two wrapper-based feature selection approaches. The forward selection approach leads to the lowest RMSE, MAE and MAPE for eight data mining models, except the ε-SVR (Linear kernel) model and the ε-SVR (Sigmoid kernel) model. The average performance of the backward selection approach is close to that of the genetic algorithm based feature selection method. Furthermore, only the forward selection approach coupled with the best data mining model can outperform the filter-based approach, while the other two wrapper-based approaches are not superior to the filter-based feature selection approach.

Generally speaking, the prediction performance of SVR-based data mining models is superior to the NN-based data mining models; the prediction errors of the BPNN and the RBFNN models are almost the highest for each of the feature selection approach. In contrast, the ν-SVR (RBF kernel) model achieves the best prediction performance, especially for the forward selection approach. With this particular configuration, the ν-SVR (RBF kernel) model achieves the lowest average RMSE, MAE and MAPE at 42,463, 31,267, and 7.92, respectively.

4.3. Comparative evaluation

For this experiment, we aim to evaluate the proposed ontology-based Web mining method when compared to traditional time series-based forecasting methods such as ARIMA [1,4,6,36]. In addition, we would like to evaluate the performance of the proposed framework with various configurations such as with or without the use of search engine queries, and with or without the use of domain ontology. The result of our comparative evaluation is reported in Table 3. Since traditional time series-based forecasting methods such as ARIMA do not utilize search engine queries, all the prediction methods depicted in Table 3 are not empowered by search engine queries. In contrast, the comparative performance of various prediction methods empowered by search engine queries is shown in Table 4. In Table 3 and Table 4, SEQ stands for search engine queries. Moreover, to demonstrate the distinct advantage of employing the proposed domain ontology, a comparative evaluation based on the best performing data mining method (i.e., ν-SVR with RBF kernel) is shown in Fig. 4.

Table 3 shows that the proposed data mining methods such as ν-SVR with RBF kernel significantly outperform the traditional time series-based prediction methods such as AIRMA in terms of all the three performance measures. The performance of various data mining methods empowered by search engine queries as shown in Table 4 is better than that of their counterparts which are not enhanced by Web queries. For example, the ν-SVR with RBF kernel and BFS feature selection method achieves a RMSE of 44,461, whereas its counterpart that is not enhanced by Web queries only achieves a RMSE of 53,120 as shown in Table 3. On the other hand, Fig. 4a–c shows that the best data mining method (i.e., ν-SVR with RBF kernel) empowered by the proposed domain ontology consistently outperforms its counterpart in which a domain ontology is not employed with respect to most of the feature selection methods. The reason is that the domain ontology contributes to extract a set of relevant Web queries in which discriminative features

Table 3
The comparative performance of various prediction methods without search engine queries.

Model	Ontology	SEQ	FS	Average RMSE	Average MAE	Average MAPE
ARIMA	No	No	–	92,345	64,016	15.88
BPNN	No	No	–	80,491	59,792	15.23
RBFNN	No	No	–	160,556	87,550	21.72
ε-SVR(RBF)	No	No	–	59,395	45,150.17	12.03
ν-SVR(RBF)	No	No	–	**53,120**	**37,815**	**9.58**

The bold valuesare the minimum in terms of different error measurements.

Table 4
The comparative performance of various prediction methods empowered by search engine queries.

Model	Ontology	SEQ	FS	Average RMSE	Average MAE	Average MAPE
CC	No	Yes	FBFS	53,390	37,825	9.63
BPNN	No	Yes	FBFS	71,028	57,323	13.84
	No	Yes	GA	68,314	56,038	13.78
	No	Yes	FFS	65,276	50,623	12.63
	No	Yes	BFS	76,953	56,837	14.49
RBFNN	No	Yes	FBFS	98,726	65,738	16.86
	No	Yes	GA	90,345	62,016	15.03
	No	Yes	FFS	85,684	61,024	14.63
	No	Yes	BFS	128,063	93,291	23.64
ε-SVR(RBF)	No	Yes	FBFS	49,138	35,867	9.17
	No	Yes	GA	52,629	36,295	9.34
	No	Yes	FFS	46,697	35,073	8.21
	No	Yes	BFS	48,286	35,822	9.23
ν-SVR(RBF)	No	Yes	FBFS	52,019	35,827	8.86
	No	Yes	GA	52,777	36,943	9.13
	No	Yes	FFS	45,517	33,553	8.25
	No	Yes	BFS	**44,461**	**31,835**	**8.15**

The bold valuesare the minimum in terms of different error measurements.

can possibly be selected to support unemployment rate prediction. For the GA-based feature selection method, the ontology-based ν-SVR model only achieves a comparable performance with respect to its counterpart in which a domain ontology is not used. The reason is that the GA-based feature selection method is not able to identify the most discriminative features under our experimental setting. As a result, both ontology-based ν-SVR model and the ν-SVR model without ontology perform poorly, and there is not a significant difference between the prediction performances of these two models. As a whole, this series of experiments confirm that the proposed ontology-based Web mining method that leverages relevant search engine queries can bootstrap the performance of unemployment rate prediction.

4.4. *Evaluation based on the validation data set*

According to our experimental results discussed in the previous subsections, the ν-SVR (RBF kernel) model with forward selection method achieves the best prediction performance. Accordingly, this model is applied to our validation data set to simulate the scenario that the proposed ontology-based Web mining framework is deployed to the real-world setting where prediction is conducted based on unseen data. In particular, the most discriminative features chosen by the forward selection method are depicted in Table 5. Moreover, the model's prediction performance as evaluated based on the validation set is plotted in Fig. 5.

Fig. 5 shows that the UIC values predicted by the proposed model are close to the actual values for the period between Jan.2011 and Mar.2012. For instance, there was a considerable reduction of the number of unemployment initial claims in the first quarter of 2011, and the proposed model can predict such a trend in general. The actual UIC values were relatively small in the remaining three quarters of 2011, and the proposed model can also make a correct prediction. There was a sharp increment of UIC values in Jan. 2012 followed by a sharp decrement in Feb. 2012; such a considerable change of the number of unemployment initial claims within a relatively short time is accurately predicted by the proposed model as well.

Apart from the merit of making an accurate unemployment rate prediction, the proposed model can also facilitate government officials or financial analysts to make timely analysis and prediction of unemployment rate. As a crucial indicator of a nation's macro economy, government officials or financial analysts tend to continuously monitor the nation's unemployment rate in order to make timely adjustments with respect to labor market related policies or financial investment strategies. However, constrained by the time-consuming processes of conducting labor market surveys and manually filtering the data collected from multiple sources, unemployment rate predictions are often made based on possibly outdated data. As a result, the accuracy of the traditional methods for unemployment rate prediction may be affected. In contrast, the proposed framework can leverage up-to-

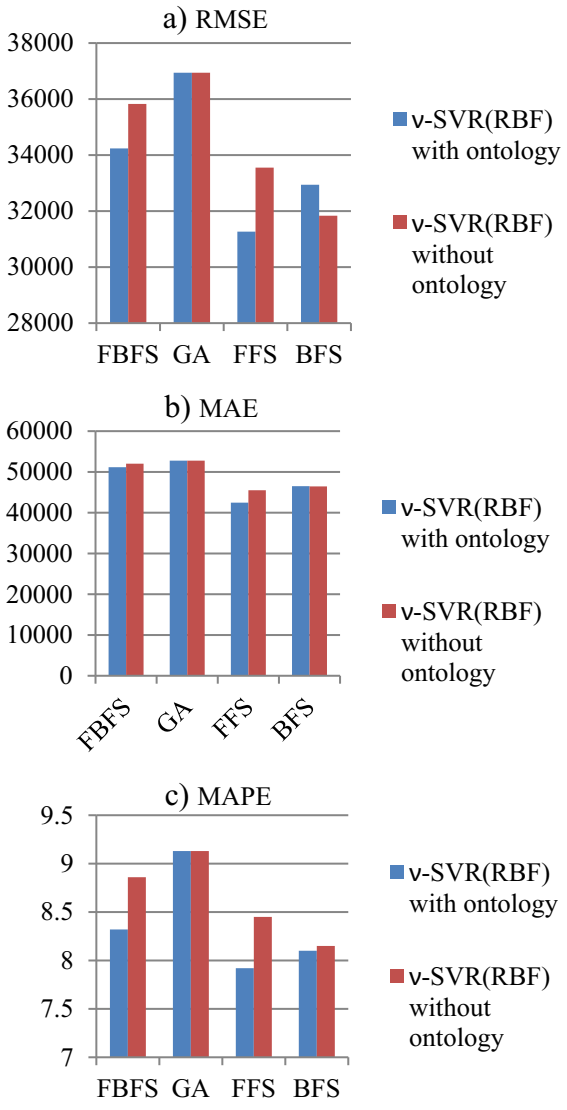

Fig. 4. Comparative prediction performance of the ν-SVR (RBF) model.

Table 5
The most discriminative features chosen by the forward selection method.

No.	Features	No.		No.	Features
1	Michigan unemployment	6	Working capital	11	Family budget
2	Coupons	7	Labor	12	Hospital pharmacist
3	Florida unemployment	8	Engineering manager	13	Economy
4	Personal finance	9	Interest rate	14	Alaska unemployment
5	Layoff	10	Forester	15	Economy current

dated search engine data instantly collected from the Web to make unemployment rate prediction in real-time. Thereby, the problem of making a prediction based on outdated data is alleviated. Moreover, government officials and financial analysts are concern not only about the trend of unemployment rate but also the underlying reasons of triggering such a trend. Traditional purely quantitative models can identify the trend of unemployment but they are weak in providing qualitative information to explain the causes of such a trend. In contrast, the proposed model can provide useful clues for the analysis of the possible increment or decrement of unemployment rate based on the human comprehensible features extracted from search engine queries. Examples of these human comprehensible clues are depicted in Table 5. As a summary, the proposed ontology-based Web mining framework for unemployment rate prediction can alleviate some of the weaknesses of classical prediction models, and it represents a viable alternative to perform real-time unemployment rate prediction in practice.

5. Conclusions

This paper illustrates a novel ontology-based Web mining framework for unemployment rate prediction. More specifically, a domain ontology is first constructed by using well-proven knowledge engineering tools. Guided by the proposed domain ontology, highly relevant search engine queries are extracted to enhance the subsequent prediction process. Several filter- and wrapper-based feature selection methods are proposed to select the most discriminative features from relevant queries to bootstrap prediction performance and yet reduce the computational complexity. Moreover, several state-of-the-art data mining methods such as NNs and SVRs are explored in order to identify the most effective method for unemployment rate prediction. Based on real-world data crawled from Web search engines and the US Department of Labor, the results of a series of empirical experiments reveal that the proposed framework is effective for the prediction of unemployment rate, and it outperforms the classical time series based prediction models such as ARIMA. Our empirical experiments also suggest that the ν-SVR (RBF kernel) prediction model together with wrapper-based forward selection method is the most effective one under our experimental setting. Apart from a higher prediction effectiveness, another merit of the proposed framework lies on its ability to provide human comprehensible qualitative clues to explain the underlying reasons of changing unemployment rates.

Some limitations of our current framework should be addressed in future research. First, an ensemble of an array of data mining methods instead of individual methods will be applied to further improve the effectiveness of unemployment rate prediction. Second, additional Web information sources and linguistic features (e.g., n-grams) will be explored to improve the prediction performance. Third, the prototype system will be enhanced and field tests which involve government officials and financial analysts will be conducted. Finally, the proposed framework will be extended to support decision-making in other fields, especially for society hotspots such as crude oil market, real-estate market, and foreign exchange market.

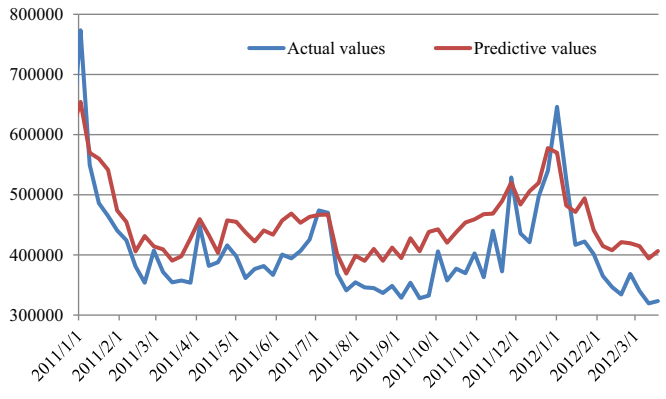

Fig. 5. The prediction performance of the ν-SVR(RBF kernel) model.

Acknowledgment

The authors would like to thank the editors and the anonymous reviewers for their insightful comments and valuable suggestions such that the quality of this manuscript is enhanced. This research work was partly supported by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No. 71001103, 91224008, 91324015), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No. Z131101000413058). Lau's work is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 115910) and the Shenzhen Municipal Science and Technology R&D Funding — Basic Research Program (Project No. JCYJ20130401145617281).

References

[1] N. Askitas, K.F. Zimmermann, Google econometrics and unemployment forecasting, Applied Economics Quarterly 55 (2) (2009) 107–120.
[2] N. Blasco, P. Corredor, C. Del Rio, R. Santamaria, Bad news and Dow Jones make the Spanish stocks go round, European Journal of Operational Research 163 (1) (2005) 253–275.
[3] C.I. Chen, Application of the novel nonlinear grey Bernoulli model for forecasting unemployment rate, Chaos, Solitons & Fractals 37 (1) (2008) 278–287.
[4] H. Choi, H. Varian, Predicting initial claims for unemployment benefits, Google Technical Report, 2009.
[5] H. Choi, H. Varian, Predicting the present with Google trends, Google Technical Report, 2009.
[6] F. D'Amuri, Predicting unemployment in short samples with internet job search query data, MPRA Paper No. 18403, 2009, pp. 1–17.
[7] F. D'Amuri, J. Marcucci, Google it! Forecasting the US unemployment rate with a Google job search index, MPRA Paper No. 18248, 2009, pp. 1–52.
[8] J. Du, L. Zhou, Improving financial data quality using ontologies, Decision Support Systems 54 (1) (2012) 76–86.
[9] F. Garcia-Sanchez, R. Martinez-Bejar, L. Contreras, J.T. Fernandez-Breis, D. Castellanos-Nieves, An ontology-based intelligent system for recruitment, Expert Systems with Applications 31 (2006) 248–263.
[10] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (19) (2009) 1012–1014.
[11] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[12] J.L. Harvill, B.K. Ray, A note on multi-step forecasting with functional coefficient autoregressive models, International Journal of Forecasting 21 (4) (2005) 717–727.
[13] J. Ma, W. Xu, E. Turban, S. Wang, O. Liu, An ontology based text mining method to cluster proposals for research project selection, IEEE Transactions on Systems, Man, and Cybernetics 42 (3) (2012) 784–790.
[14] V.I. Keilis-Borok, A.A. Soloviev, C.B. Allegre, A.N. Sobolevskii, M.D. Intriligator, Patterns of macroeconomic indicators preceding the unemployment rise in Western Europe and the USA, Pattern Recognition 38 (3) (2005) 423–435.
[15] H.M. Krolzig, M. Marcellino, G.E. Mizon, A Markov-switching vector equilibrium correction model of the UK labour market, Empirical Economics 27 (2002) 233–254.
[16] A. Lahiani, O. Scaillet, Testing for threshold effect in ARFIMA models: application to US unemployment rate data, International Journal of Forecasting 25 (2) (2009) 418–428.
[17] K.C. Lan, K.S. Ho, R.W.P. Luk, D.S. Yeung, FNDS: a dialogue-based system for accessing digested financial news, Journal of Systems and Software 78 (2) (2005) 180–193.
[18] R.Y.K. Lau, D. Song, Y. Li, C.H. Cheung, J.X. Hao, Towards a fuzzy domain ontology extraction method for adaptive e-learning, IEEE Transactions on Knowledge and Data Engineering 21 (6) (2009) 800–813.
[19] C.S. Lee, Y.F. Kao, Y.H. Kuo, M.H. Wang, Automated ontology construction for unstructured text documents, Data & Knowledge Engineering 60 (2007) 547–566.
[20] S.H. Liao, J.L. Chen, T.Y. Hsu, Ontology-based data mining approach implemented for sport marketing, Expert Systems with Applications 36 (2009) 11045–11056.
[21] C. Milas, P. Rothman, Out-of-sample forecasting of unemployment rates with pooled STVECM forecasts, International Journal of Forecasting 24 (1) (2008) 101–121.
[22] OWL web ontology language guide, Available from http://www.w3.org/TR/owl-guide/.
[23] R.F. Pelaez, Using neural nets to forecast the unemployment rate, Business Economics 41 (1) (2006) 37–44.
[24] Protégé, What is protégé? Available from http://protege.stanford.edu/.
[25] T. Proietti, Forecasting the US unemployment rate, Computational Statistics and Data Analysis 42 (3) (2003) 451–476.
[26] D. Sapsford, Z. Tzannators, The Economics of the Labour Market, Palgrve Macmillan, Hampshire, 1993.
[27] N. Schanne, R. Wapler, A. Weyh, Regional unemployment forecasts with spatial interdependencies, International Journal of Forecasting 26 (4) (2010) 908–926.
[28] R.P. Schumaker, H. Chen, A quantitative stock prediction system based financial news, Information Processing and Management 45 (5) (2009) 571–583.
[29] T. Suhoy, Query indices and a 2008 downturn: Israeli data, Bank of Israel Discussion Paper, 2009.
[30] L.J. Tashman, Out-of-sample tests of forecast accuracy: an analysis review, International Journal of Forecasting 16 (4) (2000) 437–450.
[31] V. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, New York, 1995.
[32] C.P.C. Vijverberg, A time deformation model and its time-varying autocorrelation: an application to US unemployment data, International Journal of Forecasting 25 (1) (2009) 128–145.
[33] W. Xu, Z. Han, J. Ma, A neural network based approach to detect influenza epidemics using search engine query data, Proceeding of the Ninth International Conference on Machine Learning and Cybernetics, 2009, pp. 1408–1412.
[34] W. Xu, T. Zheng, Z. Li, A neural network based forecasting method for the unemployment rate prediction using the search engine query data, Proceeding of the Eighth IEEE International Conference on e-Business Engineering, 2011, pp. 9–15.
[35] W. Xu, Z. Li, Q. Chen, Forecasting the unemployment rate by neural networks using search engine query data, Proceeding of the 45th Hawaii International Conference on System Sciences, 2012, pp. 3591–3599.
[36] W. Xu, Z. Li, C. Cheng, T. Zheng, Data mining for unemployment rate prediction using search engine query data, Service Oriented Computing and Applications 7 (1) (2013) 33–42.
[37] S. Wang, Z. Zhang, K. Ye, H. Wang, X. Chen, An ontology for causal relationships between news and financial instruments, Expert Systems with Applications 35 (2008) 569–580.
[38] X.M. Zhang, Q. Liu, H.Q. Wang, Ontologies for intellectual property rights protection, Expert Systems with Applications 39 (2012) 1388–1400.

Mr. Li is a master student in Colombia University. He got his bachelor degree at School of Information, Renmin University of China. His interests include Web mining and decision support systems.

Dr. Xu is an associate professor at School of Information, Renmin University of China. He is a research fellow at the Department of Information Systems, City University of Hong Kong. He got his bachelor and master degree in Mathematics at Xi'an Jiaotong University and doctor degree in Management Science at Chinese Academy of Sciences. His interests include Web mining, business intelligence and decision support systems. He has published over 30 papers in international journals and conferences, such as Decision Support Systems, European Journal of Operational Research, IEEE Trans. Systems, Man and Cybernetics, and Fuzzy Sets and Systems.

Mr. Zhang is a master student in School of Information, Renmin University of China. His research interests include Web mining and decision support systems.

Dr. Lau is an associate professor at the Department of Information Systems, City University of Hong Kong. He is a senior member of the IEEE and the ACM respectively. His research interests include information retrieval, text mining and e-commerce. He has published research papers in international journals and conferences, such as MIS Quarterly, INFORMS Journal on Computing, Decision Support Systems, and ACM Trans. Information Systems.