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

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.

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