基于本体的失业率预测 Web 挖掘框架

杨晋青

(大连海事大学 信息科学技术学院，辽宁 大连 116026)

摘 要 失业率是最关键的经济指标之一。通过分析和预测失业率，政府官员可以根据当前的经济形势制定适当的劳动力市场相关政策。因此，失业率预测近年来受到研究人员的极大关注。本文的主要贡献是一个新的基于本体的 Web 开发框架的说明，这个框架可以利用搜索引擎查询来提高失业率预测的准确性。该框架由一领域本体支撑，这一领域本体能够捕获失业相关概念及其语义关系，以便于从相关搜索引擎查询中提取有用的预测特征值。现有的特征选择方法和诸如神经网络和支持向量回归的数据挖掘模型都被用于提高失业率预测的有效性，实验结果表明，所提出的框架的独特优势是它不仅能提高预测性能，而且为失业率的变化提供人类可理解的解释。该框架的研究意义在于政府官员和人力资源管理者可以利用拟议的框架有效的分析失业率，从而更好的指定劳动力市场相关政策。

关键词 失业率；本体；领域本体；数据挖掘；搜索引擎查询

中图法分类号 TP369 文献标识码 A

An Ontology-based Web Mining Method for Unemployment Rate Prediction

Yang Jin-qing

（College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China）

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 empirical findings also confirm that domain ontology and search engine queries can be exploited to improve the effectiveness of unemployment rate prediction. 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.

Key words Unemployment rate; Ontology; Domain ontology ; Data mining; Search engine queries

收稿日期：2016-11-4；修订日期：2016-11-15。
基金项目：国家自然科学基金项目（60972014）。
杨晋青（1995—），女，辽宁大连人，硕士研究生，研究方向为智能信息处理。

1 引言

失业率可以影响国库券和整个金融市场的利率。事实上，失业率的任何意想不到的变化都会严重影响消费者的支出，因为这些变化影响了家庭对经济状况的看法和期望。因此，金融分析师可以通过分析相应国家的失业率来预测目标市场的经济趋势。此外，政府官员和人力资源管理人员可以通过分析和预测失业率来制定适当的人力资源相关政策。近年来，由于世界不同大陆的金融动荡，失业率预测变得越来越重要，已经受到政府，企业和研究人员的极大关注。关于预测失业率的方法也有很多，最初是单变量时间序列模型，然后又有开发时间变形模型来预测事业的趋势。另外还有些模型考虑了经济和社会因素，包括国民生产总值GNP，货币供应量和利率，以提高失业率预测的准确性。GNP 被应用于构建失业率预测模型[1]，国内生产总值被应用于构建替代预测模型[2]。此外，货币供应，生产者价格指数和利率也都被纳入预测模型以预测失业率。

在 Web 2.0 时代，用户贡献的 Web 信息被认为是分析社会或经济热点（如金融市场）的宝贵资源。应用社交网络的一种用户贡献的数据，即搜索引擎查询数据可以检测流感流行病并预测失业率。应用基于用户在互联网上的活动构建的基于 Web 的模型，也可以用来确定 Web 搜索的频率和失业率之间的关系[3]。经实验表明，所提出的模型有可能提高失业率预测效率。

在利用搜索引擎来预测失业率的以往研究中，存在两种检索搜索引擎数据的常用方法。第一种方法是收集数千个搜索引擎查询，然后使用一些特征选择方法选择相关查询的子集[4]。第二种方法是根据一些预定义的主题直接选择相关查询[4]。然而，第一种方法存在应用特征选择方法以从大量查询提取游泳特征的低效率的问题，而第二种方法可能导致基于少数预定义主题提取的特征数量不足的问题。至于失业率预测的阶段，统计方法已在以前的研究中广泛使用。相比之下，以往的研究多使用支持向量回归等少量数据挖掘工具来预测失业率。而这篇论文中开发的主要贡献之一是一个新的基于本体的 Web 挖掘框架，能够利用搜索引擎查询数据来提高失业率预测。所提出的基于本体的方法缓解了一些现有的基于 Web 的方法的弱点，例如不能从大量可能嘈杂的搜索引擎中有效地提取相关查询等。领域本体也有助于增强自动化特征选择，其目的在于降低训练查询数据的维度并提高预测的精度。

2 失业率预测 Web 挖掘框架

在本节中，提出了一种基于本体的用于失业率预测的 Web 挖掘框架。图 1 中示出了所提出框架的略图。该图表明与失业有关的相关搜索引擎查询首先通过使用捕获劳动经济学突出概念的语义丰富的领域本体来提取。然后调用基于混合滤波器的方法和基于包装器的方法组成的特征选择模块来提取最有区别的特征用于以后的预测。最后，应用最有效的分类器和相应的一组特征来预测失业率。

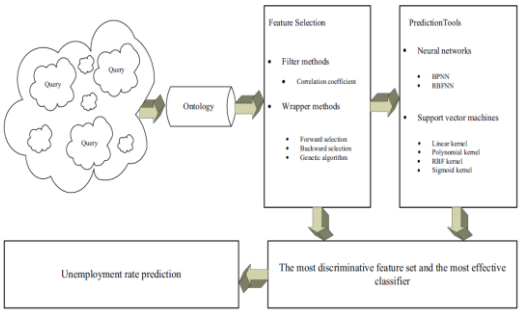

图 1　用于失业率预测的框架

2.1 领域本体构建

本体的概念最早是在哲学领域发现的。本体通常是由概念的层次结构和其他语义信息来表示。根据现有文献，领域本体捕获概念以及一个特定领域中概念和概念之间的关系，且能够表示特定域突出特征的公理和约束[5]。它是一种表示域的一组相关概念的正式且通用的方式，以便能让不同人重用或应用此域的知识。本体在描述领域知识时很受欢迎，因为它具有促进可重用性的独特

优势。文中提出的框架是基于领域本体的概念。特别地，首先确定劳动经济学领域的不同概念之间的语义关系。然后，构建可以用作解释与劳动力试产有关的不同事件的基础因果图，并且以语义丰富的领域本体的形式表示。最后，应用领域本体来提取相关的搜索引擎查询并选择用于失业率预测的有用特征。几个知名的和有效的知识工程工具都可被应用于本体的构建。

领域本体的构建具体由以下步骤组成。第一步是从各种知识源中引出领域的概念。在获得足够数量的领域概念之后，下一步是确定这些概念之间的语义关系。第三步是对领域本体进行编码。我们的领域本体主要包括三种类型的元素，即类，属性，还有类与类之间的关系。失业率预测的领域本体的一部分如图 2 所示。

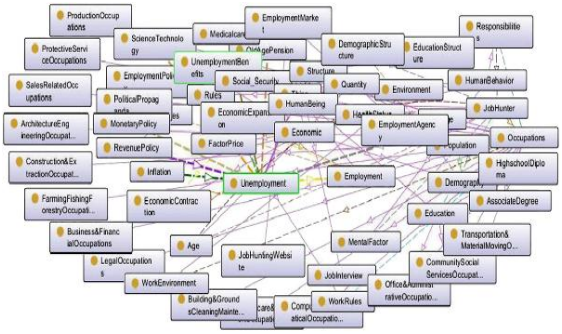

图 2　失业率预测的领域本体的一部分

在经过本体构建的细化步骤之后，一些关于劳动经济学的突出概念被编码在所提出的领域本体中。例如，概念“收入政策”与“失业”相关。原因是有效的收入政策可能导致失业率下降。直接影响失业率的一些因素即概念如图 3 所示。通过咨询如图 3 所示的领域本体可以很容易的检索出与失业相关的搜索引擎查询。

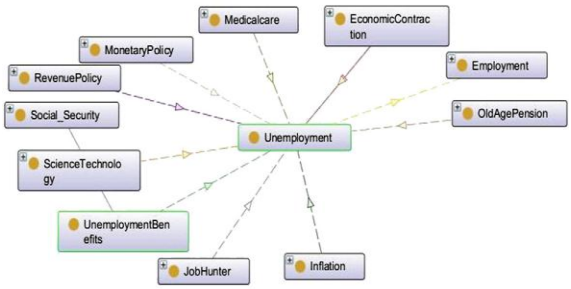

图 3　直接影响失业的因素

2.2 特征选择

通过参考提出的领域本体，检索出一组相关的 Web 查询。然而，这组查询中通常包含导致后续预测低效率的数万个词组（高维特征空间），也可能存在许多噪声特征，使得数据挖掘工具产生错误的预测。因此，应当从相关查询中提取特征的子集以提高预测准确度并最小化计算复杂度。根据特征选择方法是否需要参考分类器的预测结果将特征选择方法分为两类，即基于过滤器的方法和基于包装器的方法。研究中基于过滤器的特征选择方法和基于包装器[6,7]的方法均可使用，再应用诸如 NN 和 SVR 的不同数据挖掘工具来估计预测误差。

3 实验分析

实验中分析基于分类器和基于包装器的特征选择方法在框架中的运行结果，将训练集划分为五个子集，在模型调整过程中每次运行将其中一个子集作为测试集以评估数据挖掘方法的性能，而剩余的四个子集用于训练相应的模型。然后，采取这五次运行的平均预测误差以评估特定数据挖掘方法的性能。实验结果如下图所示。

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

图 4　基于过滤器特征选择方法的结果

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

图 4　基于包装器特征选择方法的结果

4 结论

本文阐述了一个新的基于本体的 Web 挖掘框架，该框架将用于失业率的预测。更具体地，领域本体首先通过使用公认的知识工程工具来构建。在提出的领域本体的引导下，提取高度相关的搜索引擎查询以增强随后的预测过程。另外，在特征提取中，基于过滤器和包装器的特征选择方法可以从相关查询中选择最有区别的特征以引导预测性能。几种先进的数据挖掘方法，如 NN 和 SVR 可用于确定失业率预测的最有效的方法。所提框架对于失业率的预测是十分有效的，并且它要优于经典的基于时间序列的预测模型。

当前的框架中仍有一些不足有待改进。首先，应用一组数据挖掘的方法作为一个整体，而不是单独的方法来进一步提高失业率据测的有效性。其次，探索附加的 Web 信息源和语言特征以改进预测性能。最后，原型系统继续加强，可以做现场有政府官员或金融分析师指挥的测试。另外，拟议的框架可扩大到支持其他领域的决策。

参 考 文 献

[1] 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.

[2] 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.

[3] N. Askitas, K.F. Zimmermann, Google econometrics and unemployment forecasting, Applied Economics Quarterly 55 (2) (2009) 107–120.

[4] 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.

[5] J. Du, L. Zhou, Improving financial data quality using ontologies, Decision Support Systems 54 (1) (2012) 76–86.

[6] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[7] 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.