Towards a Vocabulary for Incorporating Predictive Models into the Linked Data Web

Evangelos Kalampokis, Areti Karamanou, Efthimios Tambouris, Konstantinos Tarabanis

http://islab.uom.gr
Objective

“To propose an RDF Schema vocabulary, named the Linked Statistical Models (limo) vocabulary, that will enable the incorporation of descriptions of predictive models into the Linked Data Web and establish links to other resources such as datasets, other models, academic articles and studies.”
The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
Data analytics

- **Easy access** to large amounts of data
- Combine data and perform **data analytics**
- Create **statistical or data mining models** for understanding and describing various problem areas and domains

http://www.flickr.com/photos/seandreilingger/2587606107
Can Twitter predict the future?

Internet forecasting: Businesses are mining online messages to unearth consumers’ moods—and even make market predictions

Jun 2nd 2011 | From the print edition
Controversial results

- 11 models aiming at predicting elections results using Social Media (SM) related variables
- Only 3 of them included sentiment related variables
- Only 1 of them employed predictive analytics evaluation methods
- 6 supported SM predictive power while 5 challenged it

Understanding the predictive power of SM

- 52 empirical studies that exploit Social Media for predictions
- The predictive power of a model is directly related to:
  - Selected predictors
  - Statistical or data mining method used
  - Evaluation method employed
  - Datasets selected
  - Approaches used to collect, filter and process data

- Why don’t we **reuse** all this information?
Social Media Data Analysis Process for Predictive Analytics

Data Conditioning Phase

Raw Data Collection and Filtering
✓ Time window determination
✓ Location identification
✓ User profile's characteristics identification
✓ Search term selection

Predictor Variables Computation
✓ Predictor variables selection
✓ Predictor variables measurement
✓ Predictor variables computation

Predictive Model Creation
✓ Evaluation data identification
✓ Method selection
✓ Non-SM predictor variables selection and usage

Predictive Analysis Phase

Predictive Performance Evaluation
✓ Prediction baseline specification
✓ Evaluation method selection

Reuse of Descriptions of Predictive Models

- Discover **variables** that a predictive relationship between them have been suggested by a model
- Discover **predictor variables** that are connected to the same response
- Discover **statistical or data mining methods** used in certain cases
- Discover **datasets** used or could be reused in existing or new models
- Discover **predictive models** that could be reused (e.g. for baseline predictions or with different data)

- This is where **Linked Data** comes in …
Raw Data Collection and Filtering
- Time window determination
- Location identification
- User profile's characteristics identification
- Search term selection

Predictor Variables Computation
- Predictor variables selection
- Predictor variables measurement
- Predictor variables computation

Predictive Model Creation
- Evaluation data identification
- Method selection
- Non-SM predictor variables selection and usage

Predictive Performance Evaluation
- Prediction baseline specification
- Evaluation method selection

http://www.flickr.com/photos/mbiskoping/6075387366
A vocabulary for describing predictive models as Linked Data

- A simple vocabulary that enables the creation of description of predictive models based on linked data principles.
Relevant endeavors - PMML

- The **Predictive Model Markup Language (PMML)** is a standard for XML documents which express trained instances of analytic models.

- **Main goal**: cross-platform interoperability.

- PMML contains **over 700 elements**.
PMML’s Model Element

ModelElement

AssociationModel  BaselineCell  ClusteringModel  GeneralRegressionModel  MiningModel

NaiveBayesModel  NearestNeighborModel  NeuralNetwork

MiningSchema
Linked Statistical Models Vocabulary (LIMO)

- LIMO will enable the creation of predictive models descriptions adhering to the Linked Data principles
- First unofficial draft in:
  - [http://www.purl.org/limo-ontology/limo](http://www.purl.org/limo-ontology/limo)

Linked Statistical Models Vocabulary (LIMO)

A Vocabulary for Incorporating Predictive Models into the Linked Data Web

Unofficial Draft 15 October 2013

This version: [http://www.purl.org/limo-ontology/limo/2013/vocab-limo-20131015](http://www.purl.org/limo-ontology/limo/2013/vocab-limo-20131015)

Latest Published version:
[http://www.purl.org/limo-ontology/limo](http://www.purl.org/limo-ontology/limo)

Previous version:

Authors:
Evangelos Kalarpeakis, CERTH/ITI and University of Macedonia,
Areti Kaminina, CERTH/ITI and University of Macedonia,
Efthimios Tzirakis, CERTH/ITI and University of Macedonia,
Konstantinos Tarabanis, CERTH/ITI and University of Macedonia.

This document is licensed under a Creative Commons Attribution License. This copyright applies to the Limo Specification and accompanying documentation in RDF.

Abstract

Predictive modeling reflects the process of using data and statistical or data mining methods for predicting new observations. The predictive models that are created out of this process could be reused in different applications in the same sense that open data is reused. Towards this end, a few standards have been proposed in order to enable transfer of predictive models across platforms and applications. In this paper we suggest the need for incorporating predictive models into the Linked Data Web. Towards this end, we propose an RDF Schema vocabulary that will enable the creation of predictive models descriptions adhering to the Linked Data principles. The incorporation of these descriptions into the Linked Data Web could create new potentials beyond cross-platform model reuse. In particular, it will enable (a) easy discovery and reuse of appropriate models at a Web Scale and (b) creation of more accurate models exploiting connections of models to other models, datasets and other resources on the Web.
Future Work

- Finalize the model
- Create a dataset with predictive models described using LIMO
- Develop LIMO descriptions exporter
Future Work

- **OpenCube**: Publishing and Enriching Linked Open Statistical Data for the Development of Data Analytics and Enhanced Visualization Services

- *FP7-ICT-2013-SME-DCA No 611667*

- Start date: **1 November 2013**

- Duration: **24 months**
Thank you for your attention!!

http://kalampok.is