ALTERNATIVE DATA SOURCES FOR PRICE STATISTICS

Susanna Tåg & Anna-Riikka Pitkänen, Statistics Finland

INTRODUCTION

WHAT?

- Study, develop and test the capabilities to utilize alternative data sources, e.g. scanner data and web scraping, in statistics production

WHY?

- Aim to improve the coverage and quality of statistics, increase efficiency, and lower the response burden of enterprises

HOW?

- Statistics Finland initiated a project "Web Collection on Consumer Prices" aka "WEBHI" granted by EU on the modernization of price data collection
- Consumer Price Index, Producer Price Indices for Services and Building Cost Index were involved in the project

OUTPUTS

SCANNER/SALES DATA

CASE 1: DAILY CONSUMPTION GOODS

Scanner data was received from a major supermarket chain in Finland

CASE 2: PHARMACEUTICAL PRODUCTS

Sales data was obtained from a centralized drug information organization

WEB SCRAPER

CASE 3: HARDWARE ONLINE-SHOPS

- Web stores where a large collection of items are divided into several product categories and with a fairly stable set of price and product data

In the online hardware stores a web scraper begins navigation on the starting page of the webstore and follows the product category links systematically through the whole product selection. For each product the web scraper collects all available product and price data, as well as additional metadata.

CASE 4: FLIGHT TICKETS

- Web stores, where the price is determined according to parameters selected by the buyer

Web scraper allows the end user to save any number of different search configurations in the system. The configurations include not only the flight data (e.g. destination, day, ticket class) but also metadata regarding the time of data capture (e.g. exactly one week before departure). The web scraper is running constantly and it automatically collects price data according to each active configuration at the predetermined time of data capture.

CLASSIFICATION

- Key challenge for the utilization of alternative data in statistics production is the classification of massive amounts of observations

CASE 5: CLASSIFICATION SYSTEM

- Solution for product data classification is integrated with the web scraping system

The classification system is based on machine-learning model which requires teaching data. The system reads and analyses the product data and based on the teaching data is able to deduce the most likely category for each product. A human user can utilize an annotation interface to browse through the suggestions made by the classification system. The user can either confirm them or correct them. Each new linkage confirmed by the user is added to the teaching data and utilized to classify new products.

FINDINGS

Promising Results

- Project has successfully negotiated with two companies about delivering **scanner/sales** data in Statistics Finland (CASE 1 & 2)
- Software for web scraper is designed and developed, and at the end of the project it is operational on approx. 40 different webstores (CASE 3 & 4)
- A novel classification solution based on machine-learning enables classifying large amounts of data (CASE 5)

Lessons Learned

- Find out possibilities to (1) obtain scanner/sales data or (2) get an interface with enterprises' data or (3) retrieve data by web scraping
- Observing "netiquette" ➔ it is good to inform the enterprise about web scraping in advance
- Many challenges in web scraping, e.g.
  - Blocking the web sites
  - Web sites change irregularly ➔ maintenance needed
  - Observing the actual prices that are paid by the buyers
  - Quality changes
  - Information for weights

FUTURE PLANS

- To be able to utilize alternative data sources in statistics production, Statistics Finland will continue to test and develop the solutions ➔ evaluation phase
- Recognition the required know-hows
- Permanent team responsible for alternative data ➔ to be established in 2017
- Expandability of the solutions to other data sources and statistics

Statistics Finland