Automated Data Collection Using Open Source Web Crawling Technology

Anna Foote
PRICE Systems, LLC
Anna.Foote@pricesystems.com
Agenda

- Introduction and Motivation
- Data Collection Challenges
- Introduction to Web Crawling
- Introduction to RapidMiner
- RapidMiner - Crawling the Web
- Implementation
- Future Directions
- Questions/Discussion
Introduction and Motivation

- **Data Collection is a necessary evil for cost estimators**
  - To support the creation of Cost Estimating Relationships (CERs)
  - To support estimating by analogy
  - To support selection of input values for cost estimating models

- **Data collection is hard**
  - Data is often hard to find
  - Data is often hard to mine as the process is tedious and time consuming
  - Data is often very noisy making it hard to understand and extract from

- **Manual data collection is unreliable**
  - Inconsistency between data collectors
  - Prone to human and technical errors
  - Unable to automatically analyze data
Introduction and Motivation

Motivation for automated data collection:
- TruePlanning® Information Technology Services Cost Model

Many of the models we developed for IT Services required commodity pricing information

Commodity pricing is hard to estimate over time because
- Prices are constantly changing
- Many companies have negotiated agreements with specific vendors
- There are many things that drive commodity pricing outside of the scope of a typical cost estimating relationship
Data Collection Challenges

- Finding the right data
  - Accurate pricing data
  - Significant technical and specification information
  - Normalization across multiple vendors

- Keeping the data up to date
  - Commodity prices change frequently – based on market factors, supply and demand, etc.
  - Good pricing data from last quarter is unlikely to be relevant in this quarter

- Need a solution that is:
  - Repeatable
  - Consistent
  - Can be accomplished quickly with the push of a button
  - Can be updated regularly without extensive time investment
Introduction to Web Crawling

- **What is a web crawling?**
  - Automated process that browses the World Wide Web in a methodical manner
  - Used to provide up-to-date data
  - Used to gather/store specific information from a website

- **Web Crawling Tools**
  - Tableau
  - RapidMiner
  - Mozenda
  - Knime
  - Weka
  - Orange, etc.

*Top 50 open source web crawlers for data mining:*

Introduction to Web Crawling

- **HTML source code**
  - While the website layout may change, html source stays consistent
  - *Ex: looking for “memory” on webpage*
- Open source, predictive analytics platform
  - Freely available, may be modified and redistributed

- Available under the GNU Affero General Public License (AGPL)
  - Free software license

- User friendly, graphical user interface that allows for:
  - data collection and data mining
  - data loading and transformation
  - predictive analytics and statistical modeling
  - data preprocessing and visualization
Summary of RapidMiner Capability

- **Data Access**
  - Read from/write:
    - Files
    - Database
    - Applications
    - Cloud Storage

- **Transformation**
  - Generate and set attributes
  - Filter and sort examples
  - Normalization

- **Modeling**
  - Correlation
  - Clustering
  - Predictive analytics

- **Validation**
  - Significance testing
  - Regression
  - Visualization

- **Text Processing**
  - Extract data
  - Filter and transform data
  - Create documents

- **Web Mining**
  - Decode URLs
  - Get webpages
  - Crawl web
Introduction to RapidMiner Capabilities

Drag and Drop Interface to Build Processes

Easy import of data from Excel

Simple Data Customization
RapidMiner for Web Crawling

Version 1:

- **Get Page**
  - Identify a webpage URL
  - RapidMiner sends a GET request via HTTP
  - Returns the webpage as a document
    - *RapidMiner can crawl and scrape this document*

- **Process Documents**
  - Generates word vectors from a text object
RapidMiner for Web Crawling

Version 1:

- **Extract Information**
  - Extracts information from (Get Page) document
  - Create list of product URLs
RapidMiner for Web Crawling

Version 1:

- **Get Pages**
  - Sends RapidMiner out to each product link that it grabbed in Process Documents

- **Process Documents from Data**
  - Generates word vectors from string attributes

- **Extract Information**
  - Extracts information from (Get Page) document
  - XML Path Language (XPath) queries extract specifications from the document
Web Crawling: Version 2

Business Desktops, Workstations and All-In-One PCs

OptiPlex
For business
The world’s most secure, manageable and reliable business desktops that fit any workspace.

Precision Fixed Workstations
Starting at $549.00
For professional creators

Inspiron
Starting at $279.00
For home and home office

XPS
Starting at $749.99
For the ultimate experience

OptiPlex

3000 Series
Starting at $479.00
For everyday businesses needs and budgets. Featuring basic management and security features.

5000 Series
Starting at $529.00
Commercial desktops available in tower and small form factor for advanced performance, security and manageability.

7000 Series
Starting at $589.00
Premium business desktops designed for optimum performance. The most secure, manageable and reliable solutions.

View all OptiPlex Desktops.
RapidMiner for Web Crawling

Version 2:

- **Crawl Web**
  - Start on the specified URL
  - Crawling rules tell RapidMiner which links to follow
  - Store retrieved pages in an Example Set

![Diagram showing the Crawl Web process in RapidMiner](image)
RapidMiner for Web Crawling

Version 2:

- **Process Documents from Data**
  - Generates word vectors from string attributes

- **Extract Information**
  - Extracts information from a document (stored URLs)
Implementation to Support Commodity Pricing Data Collection

Transforming the data

- **Map**
  - Normalize data by mapping units to conversion factors
- **Replace**
  - Replace specific columns with just numerical values
- **Rename**
  - Rename attribute names to fit your data sets
- **Filter**
  - Filter out missing or unwanted data
- **Reorder**
  - Transform rows to match your ideal data set order

<table>
<thead>
<tr>
<th>Row No.</th>
<th>URL</th>
<th>Name</th>
<th>Definition Name</th>
<th>Hard Drive</th>
<th>Hard Drive Size</th>
<th>Level</th>
<th>Memory</th>
<th>Memory_GB</th>
<th>Operating System</th>
<th>Processor</th>
<th>Type of Device</th>
<th>Unit Purchase Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>Latitude 13 3k</td>
<td>End User Services</td>
<td>M.2 128GB</td>
<td>128</td>
<td>2</td>
<td>4GB (1x4G)</td>
<td>4</td>
<td>Windows 10 Pro 64t</td>
<td>Pentium DC</td>
<td>laptop</td>
<td>699</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>Latitude 12 7k</td>
<td>End User Services</td>
<td>M.2 128GB</td>
<td>128</td>
<td>2</td>
<td>4GB (1x4GB)</td>
<td>4</td>
<td>Windows 10 Pro, 64-6th Generation</td>
<td>Intel® Core</td>
<td>laptop</td>
<td>1049</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>New Inspiron</td>
<td>End User Services</td>
<td>M.2 128GB</td>
<td>128</td>
<td>2</td>
<td>4GB LPDDR</td>
<td>4</td>
<td>Windows 7 Professional</td>
<td>Intel® Core</td>
<td>laptop</td>
<td>1290</td>
</tr>
<tr>
<td>4</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>New Inspiron</td>
<td>End User Services</td>
<td>500GB 5400</td>
<td>500</td>
<td>2</td>
<td>4GB Single</td>
<td>4</td>
<td>Windows 10 Home</td>
<td>Intel® Pentium</td>
<td>laptop</td>
<td>499</td>
</tr>
<tr>
<td>5</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>New Inspiron</td>
<td>End User Services</td>
<td>500GB 5400</td>
<td>500</td>
<td>2</td>
<td>4GB Single</td>
<td>4</td>
<td>Windows 10 Home</td>
<td>Intel® Pentium</td>
<td>laptop</td>
<td>449</td>
</tr>
<tr>
<td>6</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>New Inspiron</td>
<td>End User Services</td>
<td>256GB Solid</td>
<td>256</td>
<td>2</td>
<td>8GB Dual CI</td>
<td>8</td>
<td>Windows 10 Home</td>
<td>6th Generation</td>
<td>laptop</td>
<td>749</td>
</tr>
<tr>
<td>7</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>New Inspiron</td>
<td>End User Services</td>
<td>256GB Solid</td>
<td>256</td>
<td>2</td>
<td>8GB Dual CI</td>
<td>8</td>
<td>Windows 10 Home</td>
<td>6th Generation</td>
<td>laptop</td>
<td>749</td>
</tr>
<tr>
<td>8</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>Precision 15</td>
<td>End User Services</td>
<td>500GB 2.5</td>
<td>500</td>
<td>2</td>
<td>8GB (2x4GB)</td>
<td>8</td>
<td>Windows 7 Professional</td>
<td>Intel® Core</td>
<td>laptop</td>
<td>999</td>
</tr>
<tr>
<td>9</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>Precision 15</td>
<td>End User Services</td>
<td>500GB 2.5</td>
<td>500</td>
<td>2</td>
<td>8GB (2x4GB)</td>
<td>8</td>
<td>Windows 7 Professional</td>
<td>Intel® Core</td>
<td>laptop</td>
<td>1399</td>
</tr>
<tr>
<td>10</td>
<td><a href="http://www.d">http://www.d</a></td>
<td>XPS 15 Laptop</td>
<td>End User Services</td>
<td>500GB 7200</td>
<td>200</td>
<td>2</td>
<td>8GB (1x8G)</td>
<td>8</td>
<td>Windows 10 Home</td>
<td>6th Generation</td>
<td>laptop</td>
<td>999</td>
</tr>
</tbody>
</table>
Processes have been created to crawl Dell, HP and TigerDirect for pricing and performance data for:

- Laptops
- Workstations
- Tablets
- Printers
- Storage Devices
- Servers
- Other Supporting Hardware

These processes create Excel files that are directly importable into the IT Hardware TrueFindings® database.

This database can be updated in several hours to support quarterly updates which can be distributed to the community.
IT Hardware for TrueFindings® and TruePlanning®

Presented at the 2017 ICEAA Professional Development & Training Workshop

www.iceaaonline.com/portland2017
These data points can also be accessed via the File New Template Search in TruePlanning 16.0 for immediate drag and drop into a project file.
Future Directions

- **IT Hardware Commodity database**
  - Quarterly updates with the existing processes that have been developed
  - Add additional data sets from new websites based on user suggestions
  - Including more specifications depending on user needs

- **IT Software pricing requirements**
  - We are currently investigating the feasibility of creating similar processes to support this
  - This may be problematic because many software applications require calls to the vendor for quotes – we’re hopeful we may be able to find sources
  - Customers may use specific vendor sites not available to us to crawl

- **Extending processes to Hardware and Microcircuits**
  - Include commodity prices for electronic components
    - Ex: Microcircuits board cost
  - Include Hardware COTS cost estimation
    - Ex: Purchased Hardware Unit Cost
Future Directions

Infrastructure as a service (IaaS): refers to online services that abstract the user from the details of infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc.

- IaaS pricing comparisons
  - Collect IaaS features and pricing
  - Normalize pricing in order to compare
  - Collect data from multiple websites
  - Keep pricing knowledge up to date
Future Directions

### IaaS Providers List: Comparison And Guide

#### IaaS Provider: Windows Azure

Despite the name, Windows Azure is not a Windows-only IaaS. The compute and storage services offered are typical of what you'll find in other IaaS providers, and administrators used to Microsoft platforms will find working with Windows Azure much easier. The IaaS offers ready access to virtual networks, service buses, message queues, and non-relational storage platforms as well.

#### IaaS Provider: Amazon AWS

Amazon Web Services offers a full range of services, including on-demand infrastructure services such as Amazon Elastic Compute Cloud (EC2) with GPU instances, as well as Elastic Block Storage (EBS) with high-performance SSDs on the storage side. Amazon offers infrastructure services such as load balancing, archival storage, in-memory database services, both relational and NoSQL databases.

<table>
<thead>
<tr>
<th>Key Features</th>
<th>Limitations</th>
<th>Pricing</th>
<th>Bonus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy-to-use administration tool, especially for Windows admins. Windows Azure can also be used as a PaaS.</td>
<td>Minimal, easy-to-use portal interface may not be so appealing to command line gurus.</td>
<td>m $0.02 to $1.60 per hour. Storage prices range from $0.07/GB/month to $0.12/GB/month, depending on level of redundancy.</td>
<td>Free 30-day trial with a limit of up to $200 is available for new users.</td>
</tr>
</tbody>
</table>

- **Key Features**: Rich set of services and integrated monitoring tools; competitive pricing model. AWS can also be used as a PaaS.
- **Limitations**: AWS is a complex mixture of services. As your workflows become more complex and you use more services it can be difficult to project expenses. However, Amazon offers a monthly calculator to help estimate your costs.
- **Pricing**: Instances range from $0.113/hour to $6.82/hour, with volume discounts available for reserved instances. Storage prices range from $0.095/GB/month to $0.125/GB/month. Additional charges for application services and data egress may apply.
- **Bonus**: New users can get 750 hours, 30GB storage and 15GB bandwidth for free with AWS’s Free Usage Tier.
Future Directions

- Custom solutions for clients
  - based on their specific purchasing vendors and products
  - crawl noisy excel files for a faster, automated data collection/ transformation

- Using "/robots.txt" test tool to verify crawling is available
  - Ex: GSA Advantage does not allow web crawling
RapidMiner can be downloaded from: 
https://my.rapidminer.com/nexus/account/index.html#downloads
Tableau can be downloaded from: 
https://www.tableau.com/trial/data-mining#form
Mozenda can be downloaded from: 
https://accounts.mozenda.com/signup
Knime can be downloaded from: 
https://www.knime.org/downloads/overview
Weka can be downloaded from: 
http://www.cs.waikato.ac.nz/ml/weka/downloading.html
Orange can be downloaded from: 
http://orange.biolab.si/download/
IaaS Providers List: Comparison And Guide
http://www.tomsitpro.com/articles/iaas-providers,1-1560.html