



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:

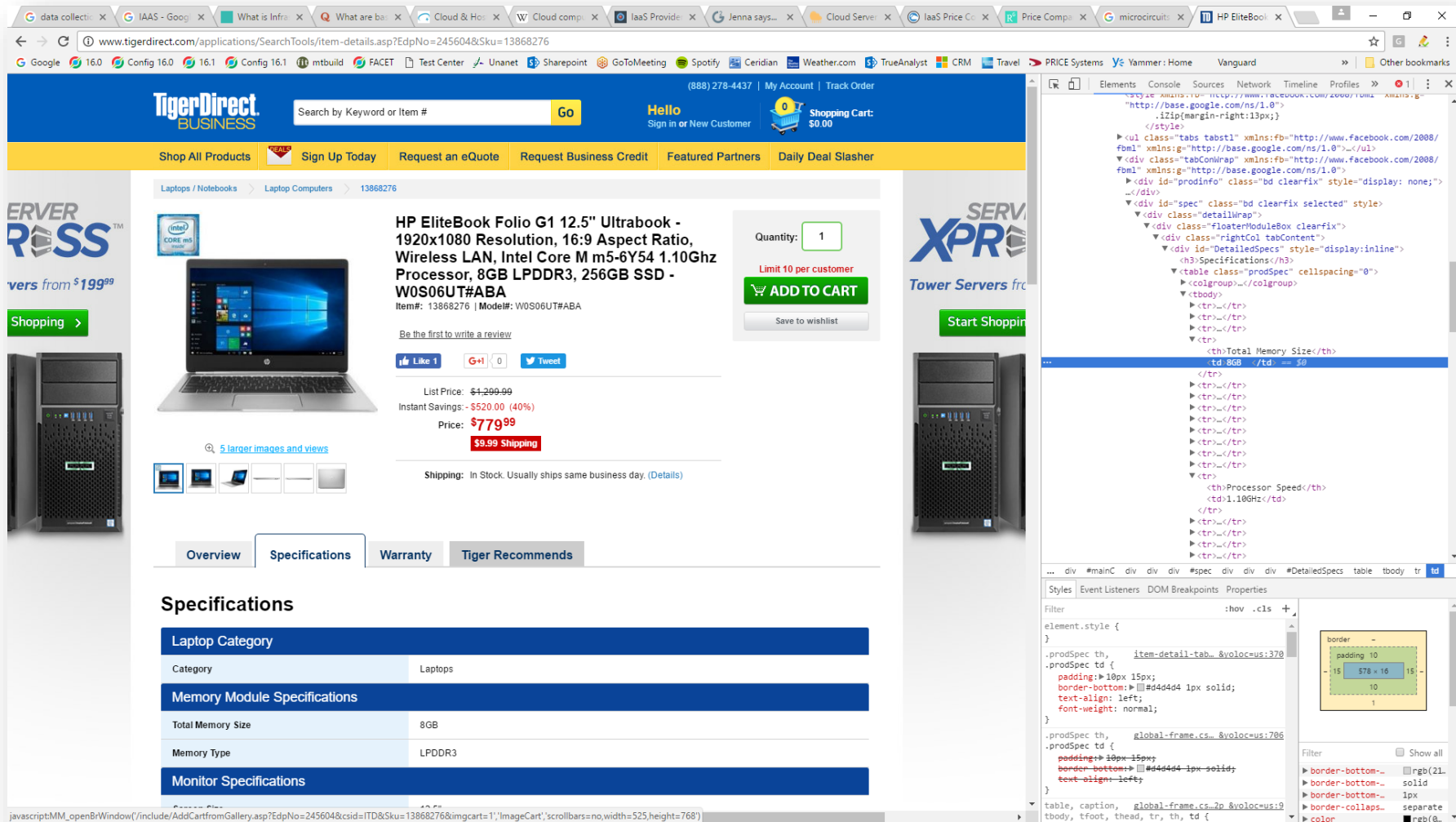
<http://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/>

Introduction to Web Crawling

HTML source code

– While the website layout may change, html source stays consistent

- Ex: looking for “memory” on webpage



The screenshot shows a web browser displaying a product page for an HP EliteBook Folio G1 12.5" Ultrabook. The browser's developer tools are open on the right, showing the HTML source code. The code includes a table with the following specifications:

Total Memory Size	8GB				
Processor Speed	1.10GHz				
<table border="1"> <tr> <td>Total Memory Size</td> <td>8GB</td> </tr> <tr> <td>Processor Speed</td> <td>1.10GHz</td> </tr> </table>		Total Memory Size	8GB	Processor Speed	1.10GHz
Total Memory Size	8GB				
Processor Speed	1.10GHz				

The table also shows a 'Limit 10 per customer' message and an 'ADD TO CART' button. The browser's developer tools show the HTML source code for the table, highlighting the 'Total Memory Size' row.

RapidMiner

- **Open source, predictive analytics platform**
 - Freely available, may be modified and redistributed

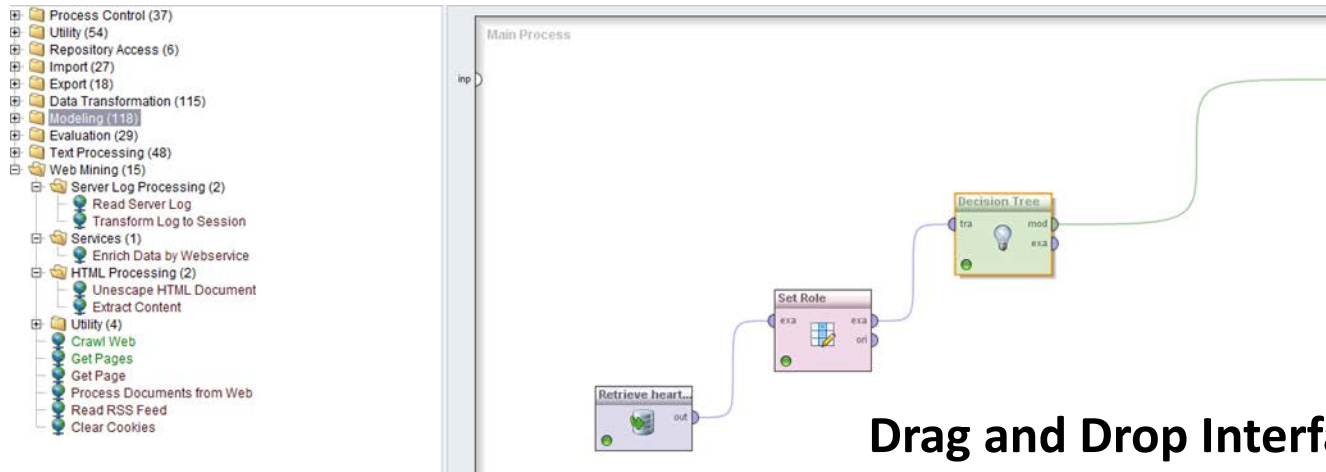
- **Available under the GNU Affero General Public License (AGPL)**
 - Free software license

- **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

Data View Meta Data View Plot View Advanced Charts Annotations

ExampleSet (138 examples, 0 special attributes, 8 regular attributes)

Row No.	Age	Marital_Stat...	Gender	Weight_Cat...	Cholesterol	Stress_Man...	Trait_Anxiety	2nd_Heart_...
1	60	2	0	1	150	1	50	Yes
2	69	2	1	1	170	0	60	Yes
3	52	1	0	0	174	1	35	No
4	66	2	1	1	169	0	60	Yes
5	70	3	0	1	237	0	65	Yes
6	52	1	0	0	174	1	35	
7	58	2	1	0	140	0	45	
8	59	2	1	0	143	0	45	
9	60	2	0	0	139	0	45	
10	51	1	1	0	174	1	40	
11	52	1	0	0	189	1	65	
12	70	2	1	1	147	1	50	
13	52	2	1	2	160	0	40	
14	74	3	1	2	178	0	75	
15	64	2	1	2	236	1	80	
16	69	2	0	1	146	1	50	
17	58	2	0	0	141	0	45	
18	68	1	0	0	172	0	60	

Simple Data Customization

ExampleSet (8 examples, 0 special attributes, 6 regular attributes)

Row No.	Name	URL	Definition Name	Type of Device	Level	Router Unit...
1	AudioCodes -...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	66
2	TP-Link TL-R...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	73
3	Cisco RV042 ...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	168
4	Axio Meory D...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	102
5	AudioCodes -...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	93
6	D-Link Power...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	79
7	Zyxel ADSL2+...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	29
8	D-Link Power...	http://www.tig...	Infrastructure Services New Projects	networkDevice	2	59

(888) 278-4437 | My Account | Track Order

Hello
 Sign in or New Customer

Shopping Cart:
 \$0.00

Shop All Products Sign Up Today Request an eQuote Request Business Credit Featured Partners Daily Deal Slasher

VIEW GUIDED SEARCH

Shop Laptop Computers

Price Range

Manufacturer

Laptops / Notebooks > Laptop Computers

Sort by: Most Popular | Items per page: 10 | 20 | 30
 Results 1 - 10 of 10147 | 1 2 3 ...

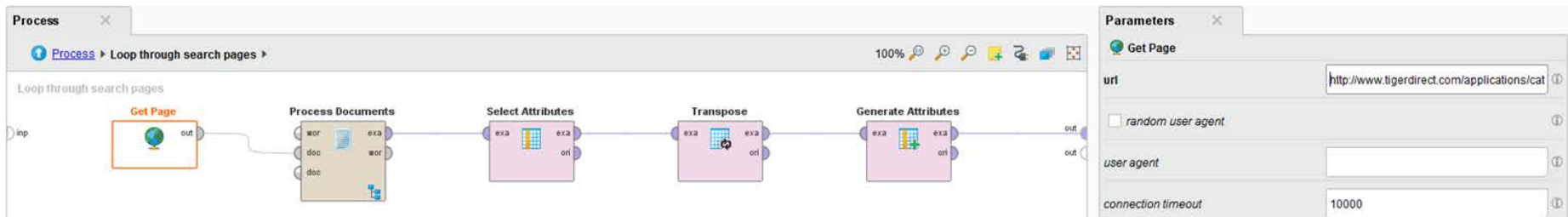
- | | | |
|----------------|--|---|
| <p>Compare</p> | <p>HP 255 G5 15.6" Notebook Laptop - AMD E-Series E2-7110 1.80GHz Quad-core Processor, 4GB DDR3L SDRAM, 500GB HDD, 1366x768 Resolution, AMD Radeon R2 Graphics, HDMI, USB 3.0 - W8W69UT#ABA</p> <p>Item#: 40018298 Model#: W8W69UT#ABA</p> <p>★★★★★ (1 Review)</p> <p>In Stock. Usually ships same business day. (Details)</p> | <p>\$429.99
 \$279.99</p> <p>Save \$150 instantly</p> <p>ADD TO CART</p> <p>Save to wishlist</p> |
| <p>Compare</p> | <p>HP 255 G5 15.6" Notebook - AMD A6-7310, 2GHz, 4GB DDR3L, 500GB HDD, 1366x768, AMD Radeon R4, DVD SuperMulti DL, 802.11ac, Bluetooth, Windows 10 Pro 64-bit - W0S61UT#ABA</p> <p>Item#: 40018294 Model#: W0S61UT#ABA</p> <p>Be the first to review</p> <p>In Stock. Usually ships same business day. (Details)</p> | <p>\$449.99
 \$299.99</p> <p>Save \$150 instantly</p> <p>ADD TO CART</p> <p>Save to wishlist</p> |
| <p>Compare</p> | <p>HP ZBook 14 G2 Mobile Workstation Laptop - 14" IPS Display, 1920x1080, Intel Core i7-5500U Processor, 2.40GHz, 16GB DDR3L, 256GB SSD, 1GB FirePro M4150, USB 3.0, Bluetooth, Windows 7 Pro - X9U28UT#ABA</p> <p>Item#: 40180825 Model#: X9U28UT#ABA</p> <p>Be the first to review</p> <p>In Stock. Usually ships same business day. (Details)</p> | <p>\$1,299.99
 \$879.99</p> <p>Save \$420 instantly</p> <p>ADD TO CART</p> <p>Save to wishlist</p> |
| <p>Compare</p> | <p>Acer Aspire E 15 E5-553-T2XN Notebook - AMD A-series A10-9600P Processor, 2.4 GHz, Win 10 Home, 8GB DDR4 RAM, 1TB HDD, DVDRW, 15.6" Display, AMD Radeon R5 Graphics, Wi-Fi, HDMI, USB - NX.GESAA.004</p> <p>Item#: 40172283 Model#: NX.GESAA.004</p> <p>Be the first to review</p> <p>In Stock. Usually ships same business day. (Details)</p> | <p>\$441.99
 \$441.99</p> <p>ADD TO CART</p> <p>Save to wishlist</p> |

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

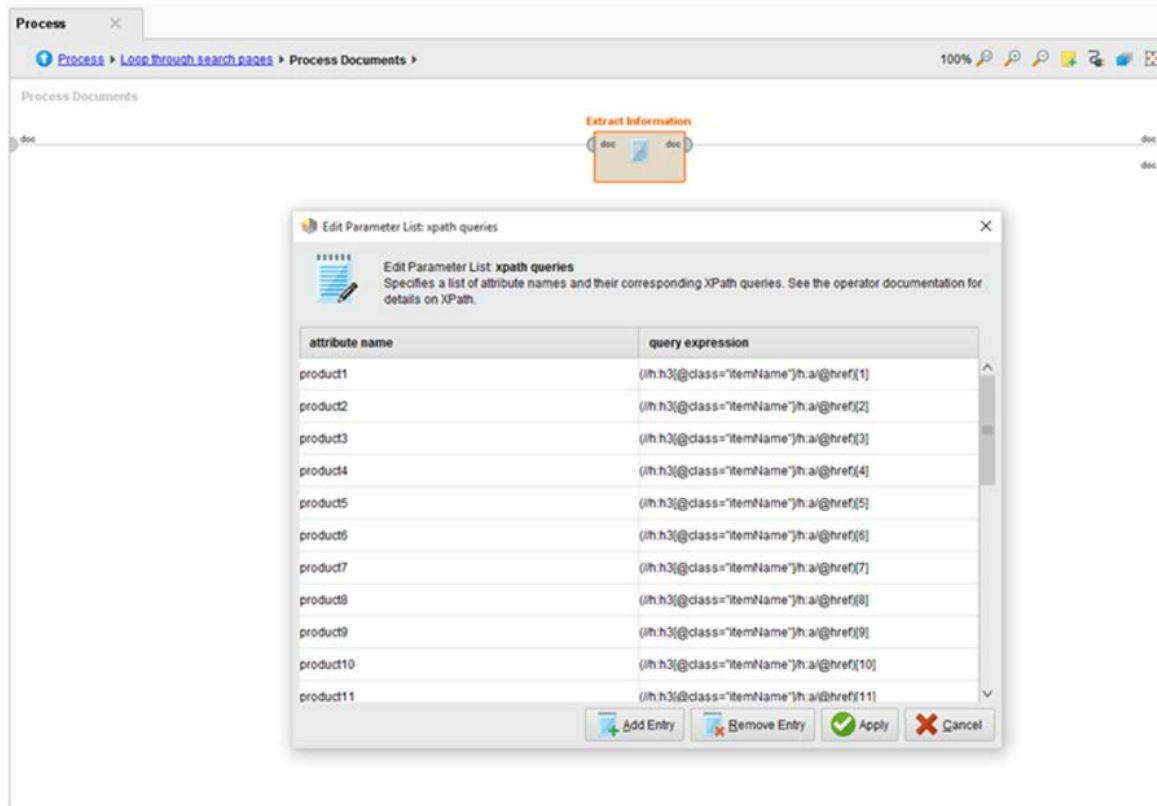
- Generates word vectors from a text object

RapidMiner for Web Crawling

Version 1:

■ Extract Information

- Extracts information from (Get Page) document
- Create list of product URLs



The screenshot displays the RapidMiner interface. The main window shows a process flow with an 'Extract Information' operator. A dialog box titled 'Edit Parameter List: xpath queries' is open, showing a table of attribute names and their corresponding XPath queries.

attribute name	query expression
product1	<code>(/h:h3[@class="itemName"]/h:a)[@href[1]</code>
product2	<code>(/h:h3[@class="itemName"]/h:a)[@href[2]</code>
product3	<code>(/h:h3[@class="itemName"]/h:a)[@href[3]</code>
product4	<code>(/h:h3[@class="itemName"]/h:a)[@href[4]</code>
product5	<code>(/h:h3[@class="itemName"]/h:a)[@href[5]</code>
product6	<code>(/h:h3[@class="itemName"]/h:a)[@href[6]</code>
product7	<code>(/h:h3[@class="itemName"]/h:a)[@href[7]</code>
product8	<code>(/h:h3[@class="itemName"]/h:a)[@href[8]</code>
product9	<code>(/h:h3[@class="itemName"]/h:a)[@href[9]</code>
product10	<code>(/h:h3[@class="itemName"]/h:a)[@href[10]</code>
product11	<code>(/h:h3[@class="itemName"]/h:a)[@href[11]</code>

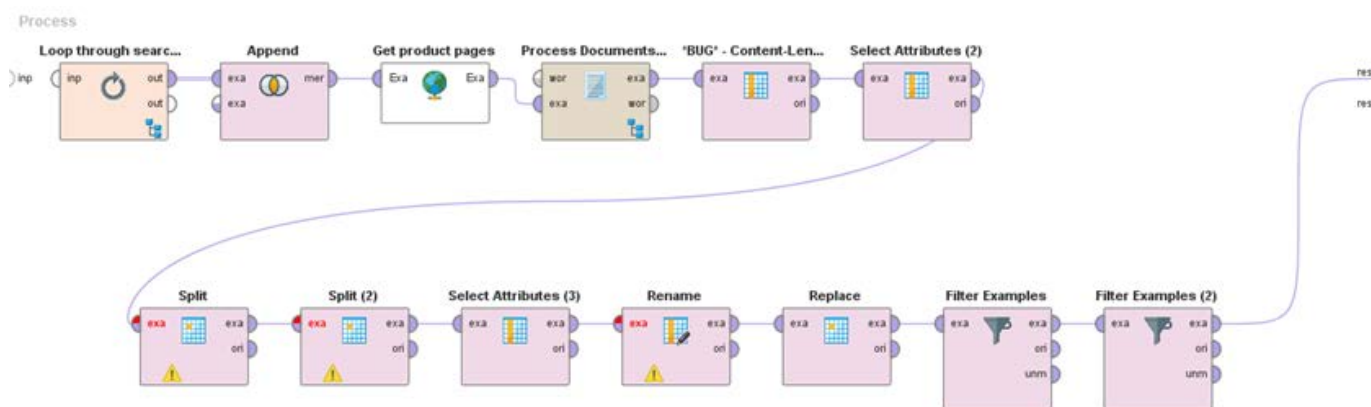
At the bottom of the dialog box, there are four buttons: 'Add Entry', 'Remove Entry', 'Apply', and 'Cancel'.

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

Refine Your Search

Processor | Memory | Hard Drive | Configurable | Operating System

Form Factor | Brand | Price



OptiPlex
Starting at \$469.00
For business



Precision Fixed Workstations
Starting at \$649.00
For professional creators



Inspiron
Starting at \$279.00
For home and home office



XPS
Starting at \$749.99
For the ultimate experience



Deals
For current Desktop Deals

OptiPlex

For business

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



Learn more about OptiPlex Desktops.

[View all OptiPlex](#)



3000 Series
Starting at \$479.00

[Micro](#) | [Desktops](#)

For everyday businesses needs and budgets. Featuring basic management and security features.



5000 Series
Starting at \$529.00

[Desktops](#) | [Micro](#)

Commercial desktops available in tower and small form factor for advanced performance, security and manageability.



7000 Series
Starting at \$589.00

[Micro](#) | [Desktops](#)

Premium business desktops designed for optimum performance. The most secure, manageable and reliable solutions.

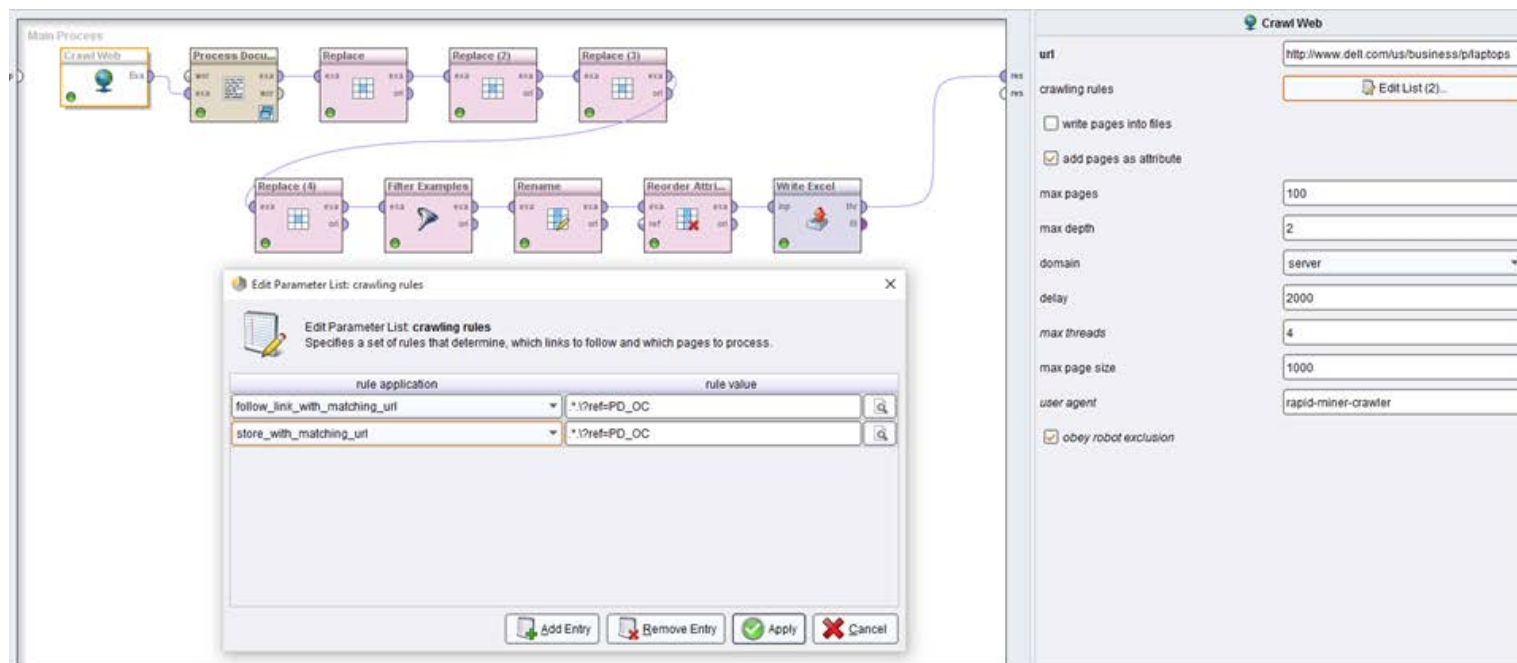


RapidMiner for Web Crawling

Version 2:

■ Crawl Web

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



The screenshot displays the RapidMiner interface. The main process flow includes: Crawl Web, Process Document, Replace, Replace (2), Replace (3), Replace (4), Filter Examples, Rename, Reorder Attributes, and Write Excel. A dialog box titled 'Edit Parameter List: crawling rules' is open, showing a table of rules:

rule application	rule value
follow_link_with_matching_url	*:!?ref=PD_OC
store_with_matching_url	*:!?ref=PD_OC

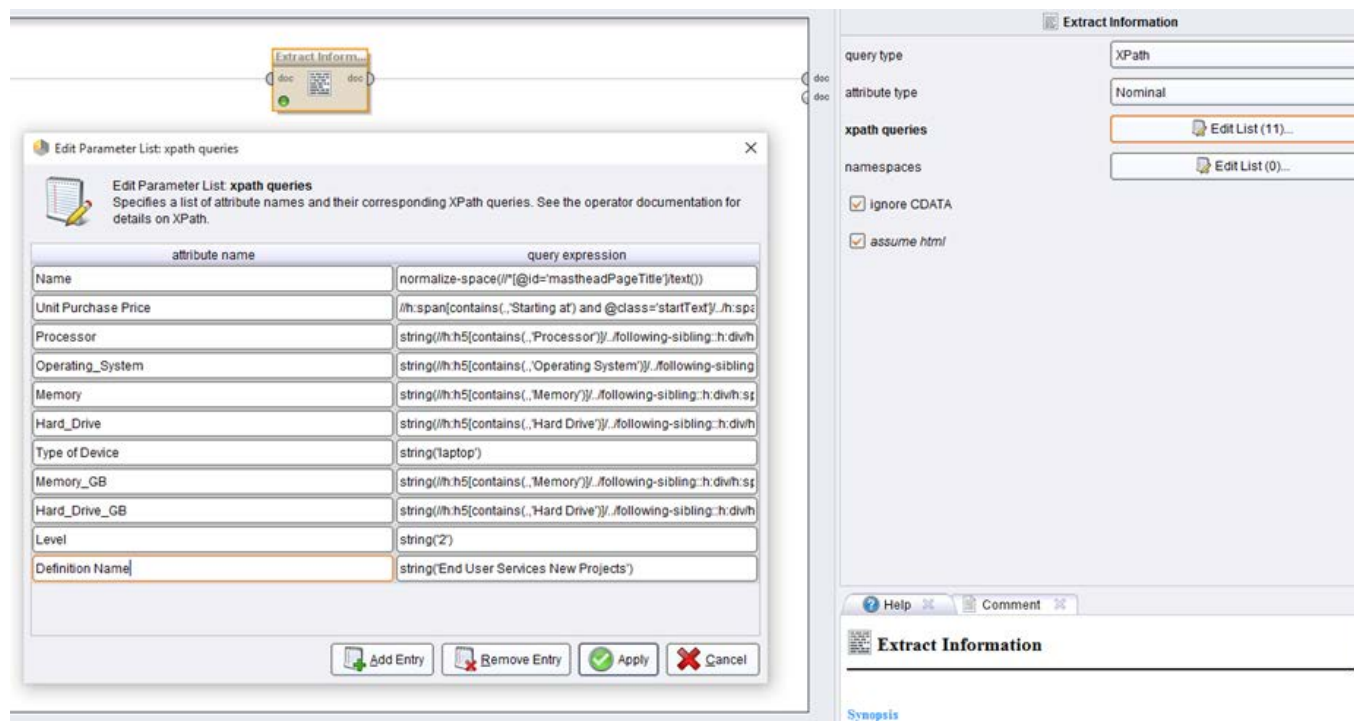
The 'Crawl Web' parameter list on the right includes:

- url: http://www.dell.com/us/business/laptops
- crawling rules: Edit List (2)...
- write pages into files
- add pages as attribute
- max pages: 100
- max depth: 2
- domain: server
- delay: 2000
- max threads: 4
- max page size: 1000
- user agent: rapid-miner-crawler
- obey robot exclusion

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)



The screenshot displays the 'Extract Information' operator in RapidMiner. The main configuration panel on the right shows the following settings:

- query type: XPath
- attribute type: Nominal
- xpath queries: Edit List (11)...
- namespaces: Edit List (0)...
- ignore CDATA
- assume html

The 'Edit Parameter List: xpath queries' dialog is open, showing a table of attribute names and their corresponding XPath queries:

attribute name	query expression
Name	normalize-space(//*[d='mastheadPageTitle']/text())
Unit Purchase Price	//h.span[contains(, 'Starting at') and @class='starText']/h.sp
Processor	string(//h.h5[contains(, 'Processor')]/..following-sibling:h.dv/h
Operating_System	string(//h.h5[contains(, 'Operating System')]/..following-sibling
Memory	string(//h.h5[contains(, 'Memory')]/..following-sibling:h.dv/h.s
Hard_Drive	string(//h.h5[contains(, 'Hard Drive')]/..following-sibling:h.dv/h
Type of Device	string('laptop')
Memory_GB	string(//h.h5[contains(, 'Memory')]/..following-sibling:h.dv/h.s
Hard_Drive_GB	string(//h.h5[contains(, 'Hard Drive')]/..following-sibling:h.dv/h
Level	string('2')
Definition Name	string('End User Services New Projects')

Buttons at the bottom of the dialog include 'Add Entry', 'Remove Entry', 'Apply', and 'Cancel'.

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

Row No.	URL	Name	Definition Name	Hard_Drive	Hard_Drive...	Level	Memory	Memory_GB	Operating_System	Processor	Type of Device	Unit Purchase Price
1	http://www.d	Latitude 13 3	End User Services	M.2 128GB €	128	2	4GB (1x4G)	4	Windows 10 Pro 64	Pentium DC	laptop	699
2	http://www.d	Latitude 12 7	End User Services	M.2 128GB €	128	2	4GB (1x4GB)	4	Windows 10 Pro, 64	6th Generat	laptop	1049
3	http://www.d	Latitude 13 7	End User Services	M.2 128GB €	128	2	4GB LPDDR	4	Windows 7 Professi	Intel® Core ⁱ	laptop	1299
4	http://www.d	New Inspiron	End User Services	500GB 5400	500	2	4GB Single €	4	Windows 10 Home,	Intel® Pentii	laptop	499
5	http://www.d	New Inspiron	End User Services	500GB 5400	500	2	4GB Single €	4	Windows 10 Home,	Intel® Pentii	laptop	449
6	http://www.d	New Inspiron	End User Services	256GB Solid	256	2	8GB Dual Cl	8	Windows 10 Home €	6th Generat	laptop	749
7	http://www.d	New Inspiron	End User Services	256GB Solid	256	2	8GB Dual Cl	8	Windows 10 Home €	6th Generat	laptop	749
8	http://www.d	Precision 15 :	End User Services	500GB 2.5 ir	500	2	8GB (2x4GB)	8	Windows 7 Professi	Intel® Core ⁱ	laptop	999
9	http://www.d	Precision 15 :	End User Services	500GB 2.5 ir	500	2	8GB (2x4GB)	8	Windows 7 Professi	Intel® Core ⁱ	laptop	1399
10	http://www.d	XPS 15 Lpto	End User Services	500GB 7200	200	2	8GB (1x8G)	8	Windows 10 Home,	6th Generat	laptop	999

Implementation to Support Commodity Pricing Data Collection



- 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®

	Value
1 Start Date	
2 Device Information	
3 Type of Device	Laptop
4 Service Level	3.00
5 Number of Deployments	Custom - Yearly
6 Quantity Per Next Higher Level	1.00
7 Purchase or Lease	Purchase
8 Service Options	In-House
9 Project Details	
10 Organizational Productivity	1.000
11 Purchase Inputs	
12 Unit Purchase Price	710.03
13 Inventory	
14 Unit Lifetime	
15 Supporting Details	
16 Software Price per Unit	
17 Annual Training per End User	
18 Annual Support per End User	
19 Training per Unit Delivered by Help Desk Analyst	

Name	Value	Method	Type
Laptops - Unit Purchase Price	710.031578947...	Distribution	Mean
Printers - Unit Purchase Price	1552.25946327...	Distribution	Mean
Routers - Unit Purchase Price	6406.29338842...	Distribution	Mean
Servers - Unit Purchase Price	1573.89285714...	Distribution	Mean
Storage Devices - Unit Purchase Price	1369.10394265...	Distribution	Mean
Switches - Unit Purchase Price	389.513089005...	Distribution	Mean
Tablets - Unit Purchase Price	932.086206896...	Distribution	Mean
Workstations - Unit Purchase Price	783.866310160...	Distribution	Mean

Search TrueFindings Database

Your Search

- Type of Device
- laptop

Keyword

Characteristics

- Definition Name
- Hard_Drive
- Memory
- Operating_System
- Processor
- URL

Performance

- Number of Devices (95)
- Unit Purchase Price (95)
- External Integration Complexity (95)
- Total Weight (95)
- Hard_Drive_GB (95)
- Memory_GB (95)

Data (95 rows)

	Name	Definition Name	Type of Device	Number of De...	Unit Purcha...	External Integr...	Total Weight	Ha
1	HP Chromebo...	End User Servi...	laptop	1	179	1	0	16
2	HP Chromebo...	End User Servi...	laptop	1	189	1	0	16
3	HP Chromebo...	End User Servi...	laptop	1	199	1	0	16
4	ASUS 11.6" Eee...	End User Servi...	laptop	1	241	1	0	32
5	Lenovo ThinkP...	End User Servi...	laptop	1	259	1	0	12
6	HP Chromebo...	End User Servi...	laptop	1	279	1	0	16
7	Acer Aspire EI ...	End User Servi...	laptop	1	302	1	0	50
8	Acer Aspire ES...	End User Servi...	laptop	1	302	1	0	50

Distribution Finder

Dependent Variable: Unit Purchase Price

'Unit Purchase Price' Di

Statistics	Data
Min	179
Max	1699
25%	476.5
Mean	710.031578...
Whiskers Percentiles:	15/85

Implementation to Support Commodity Pricing Data Collection

- 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.

The screenshot displays the TruePlanning 16.0 interface. On the left is a blue sidebar with a 'New' menu. The main area shows a 'New' screen with a search bar containing 'IT Hardware'. Below the search bar is a preview of the 'IT Hardware Q1 2017' template, which includes icons for various IT equipment. A large blue arrow points from this preview to a file explorer window showing a directory structure for 'IT Hardware Q1 2017' with subfolders like Laptops, Routers, Switches, Printers, Servers, Storage Devices, and Tablets. The 'Tablets' folder is expanded, listing various tablet models such as HP Elite x2 1012 G1 Tablet, Microsoft Surface Pro 4 Tablet Bundle, and HP SBY Elite X2 1012 G1 12" Tablet. To the right of the file explorer is a detailed data table for the selected 'HP Elite x2 1012 G1 Tablet'.

HP Elite x2 1012 G1 Tablet - Core 3 6Y30 / 900 MHz, Win 10 Pro 64-bit, 4GB RAM, 128GB SSD, 12" IPS tou	
Cost:	\$0
Project Cost:	\$0
Phase Set:	A <Inherited>
Worksheet Set:	A <Inherited>
	Value
1 Start Date	
2 Device Information	
3 Type of Device	Tablet
4 Number of Devices	1.00
5 Service Options	In-House
6 Project Details	
7 Organizational Productivity	1.000
8 Purchase Inputs	
9 Unit Purchase Price	1,049.00
10 Supporting Details	
11 Training per Device	0.00
12 Training per Unit Delivered by Help Desk Analyst	0.00
13 Setup and Installation Time per Device	1.00
14 Work by Systems Administrator	50.00%
15 IT Manager Time per Unit	10.00%
16 Integration Information	
17 Number of Operational Hours	0.00
18 Recovery Time Objective	4.00
19 External Integration Complexity	1.00
20 Total Weight	0.000

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 offerings, including on-demand instances and services such as Amazon Elastic Compute Cloud (EC2), GPU instances, as well as Amazon Elastic Block Store (EBS) performance SSDs on the storage side. AWS offers infrastructure services such as Amazon S3 for passing, archival storage, in-memory services, both relational and NoSQL.

Key Features	Easy-to-use administration tool, especially for Windows admins. Windows Azure can also be used as a PaaS.
Limitations	Minimal, easy-to-use portal interface may not be so appealing to command line gurus.
Pricing	From \$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.
Bonus	Free 30-day trial with a limit of up to \$200 is available for new users.

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



```
User-agent: *
Disallow: /advantage/
Disallow: /advgsa/ebuy_buyer/
Disallow: /advgsa/advantage/profile/
Disallow: /advgsa/advantage/punchout/
Disallow: /advgsa/advantage/sas/
Disallow: /advgsa/advantage/search/
Disallow: /advgsa/advantage/catalog/
Disallow: /advgsa/advantage/checkout/
Disallow: /advgsa/advantage/contractor/
Disallow: /advgsa/advantage/elib/
Disallow: /advgsa/advantage/fedstrip/
Disallow: /advgsa/advantage/information/
Disallow: /advgsa/advantage/logout/
Disallow: /advgsa/advantage/orderhistory/
Disallow: /advgsa/advantage/ordering/
Disallow: /advgsa/advantage/parkcart/
Disallow: /advgsa/advantage/cart/
Disallow: /advgsa/advantage/include/
Disallow: /advgsa/advantage/layouts/
Disallow: /advgsa/advantage/tiles/
```

- Using “/robots.txt” test tool to verify crawling is available
 - Ex: GSA Advantage does not allow web crawling



Resources

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>