January 1, 2014

Using Web Scraping Technology to Automatize Cost Data Collection for Calculating Construction Cost Indexes

Joseph Shrestha, Iowa State University
H. David Jeong, Iowa State University

Available at: https://works.bepress.com/joseph_shrestha/7/
Using Web Scraping Technology to Automatize Cost Data Collection for Calculating Construction Cost Indexes

JOSEPH SHRESTHA
CONSTRUCTION ENGINEERING
04/04/2014
Construction Cost Index - Definition

- Price index = \( \frac{\text{Price at time } t_2}{\text{Price at time } t_1} \)

- Construction cost index
  - Accounts for industry specific cost movement

- Available Indexes
  - ENR BCI, ENR CCI, UCLA, NHCCI, state CCIs

- Mathematical indexes
  - Laspeyres, Paasche, Fisher
Application

- Cost Estimation
- Financial Planning for owners
- Trend of Cost Growth
- Comparing the cost growth experiences between different states
Web Scraping - Definition

- **Definition**
  - Similar to web crawler
  - Can collect data from websites automatically

- **Available tools**
  - Scrapy (Python based)
  - Selenium (Java, C#, Ruby, Python, JavaScript based)
Current Methodology

- Data collection
  - ENR Indexes:
    - Input cost index
    - contacting the suppliers
  - DOTs:
    - Output cost index
    - their own bid data

- Analysis:
  - Limited items used
Proposed Methodology

- **Proposed Method**
  - Web Scraping Algorithm
  - Dynamic Method with Data Mining and Predictive Analytics
  - Construction Cost Index

- **Current Method**
  - Manual communication
  - Static Method with Predefined Formula
  - Construction Cost Index
Methodology (continued)

Work done so far

- Algorithm for “Scrapy” for Home Depot Website
- Database design to collect and store data

Remaining tasks

- Algorithm for more website
- List of data to be collected
- Algorithm to calculate the cost index using data mining
- Algorithm for forecasting the cost index using predictive analytics
Advantages

- Reduced cost
- Reduced time
- Automatize
- As frequent as required
- Add more items than just the cost (e.g. tax rate)
- One system to calculate CCI for different locations
Limitations

- Site specific coding
- Change in website structure
- Online price vs on-site price