Salary Prediction using Big Data

M.Yasmin[1] K.Kavinilavurajan[2]

Abstract—The idea is to help the employers and jobseekers to figure out the market worth of different job positions by building a prediction engine for the salary of any Indian job ad. In this way, we would bring more transparency to this important market .So that Employers would determine more reasonable salary for a position. Also Employees could find more jobs based on their background information by using our recommendation system. " www.payscale.com" is the only website right now providing salary range for various domains. Still it is not accurate We can collect data from the many online job portals. Using web-scrapping method we can collect the data's from the job portals and store it in the hadoop file system. The Hadoop file system enables you to process the large number of datasets at very quick time.

Key words: Salary Prediction, Hadoop

I. How is this idea different from what is already there

Payscale website have salary ranges for certain categories but it is fixed. The problem is, it won't show the current salary range. For example if an experience developer gets the salary of 40k now in particular field, it may change next month depending on the current trend. The process involves collecting the required data from different job portals, and predicts the salary ranges for the particular domain areas. Finally we can develop an app and website depending on the current market situation that displays an accurate salary ranges for individual job sectors, so that the job seekers would be benefited both financially and also they can select the comfortable job location. I guess, it would rock the market for sure.

II. What is big data

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become.

More accurate analyses may lead to more confident decision making. And better decisions can mean greater operational efficiencies, cost reductions and reduced risk.

III. Where is big data coming from?

Before you begin to make sense of your data, it's important to know its origins. The sources of big data are increasing every year, but they generally fall into one of three categories.

A. Streaming Data:

Also called the Internet of Things, this includes data that reaches your IT systems from a web of connected devices. Your organization can analyze this data as it arrives and make decisions on what data to keep, what not to keep and what requires further analysis. Read more about understanding data streams in this white paper.

B. Social Media Data:

The data on social interactions is an increasingly attractive set of information, particularly for marketing, sales and support functions. This data is often in unstructured or semi-structured forms, so besides the sheer size of the data, it poses a unique challenge when consuming and analyzing this information. See how one company is marketing to mobile and social customers.

C. Publicly Available Sources:

Massive amounts of data is available through open data sources like US government's data.gov, the CIA World Fact book or the European Union Open Data Portal. Learn how SAS is helping people visualize 300+ million rows of global UN trade data.

IV. Types of analytics in big data

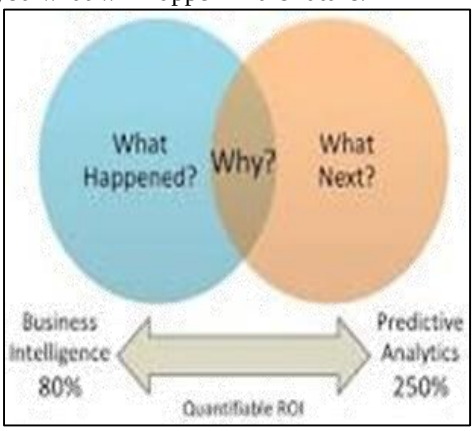
1) Descriptive
2) Predictive
3) Prescriptive

A. Descriptive Analytics:

Descriptive analytics is the most common type of analytics used by just about every organization in every industry. It serves as a foundation for more advanced analytics. When it comes to data analysis, you need to start by fully understanding what has happened and what is happening now.

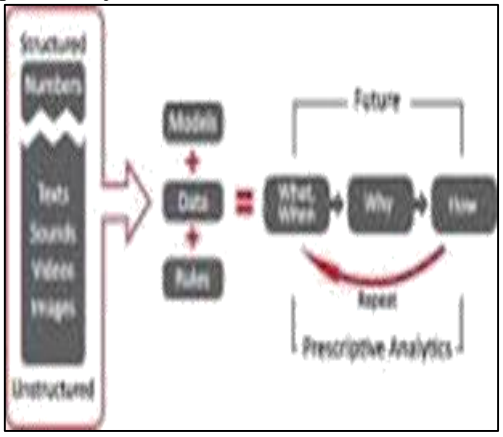
B. Predictive Analytics:

Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. Predictive analytics does not tell you what will happen in the future.



C. Prescriptive Analytics:

Prescriptive analytics is the third and final phase of business analytics (BA) which includes descriptive, predictive and prescriptive analytics.



For salary prediction we need to use predictive analysis only.

V. REQUIREMENTS

1) Scrapping tool
2) Hadoop file system
3) Database
4) R – Studio

VI. HOW TO COLLECT DATA'S FOR SALARY PREDICTION

We can collect data's from the various job portals like nakuri.com, shine.com, timesjob.com etc.

VII. DATA COLLECTION TOOLS

There is lot of tools available for data collection. Since we need to collect the data's from the websites, we can use web scrapping by using Watin in visual C# programmatically or any scrapping tools like web scrapper or Data miner.

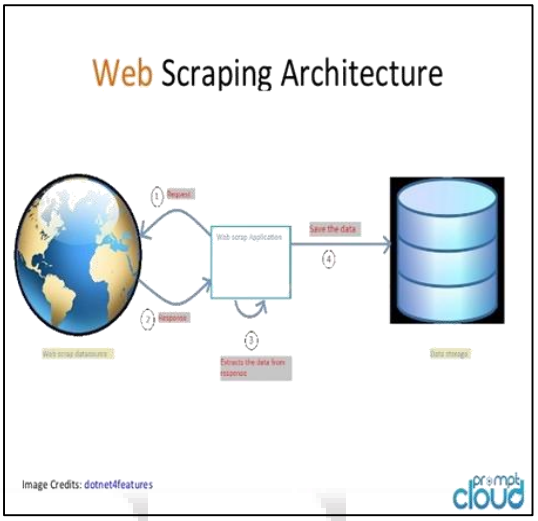
VIII. WHAT IS WEB SCRAPPING?

Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites. Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox.

Web scraping is the process of automatically collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text

Processing semantic understanding, artificial intelligence and human-computer interactions. Current web scraping solutions range from the ad-hoc, requiring human effort, to fully automated systems that are able to convert entire web sites into structured information, with limitations.

IX. WEB SCRAPPING ARCHITECTURE



X. WEB SCRAPPING TOOLS

There are lot of open source tools and plugins available for the web scrapping. Even by using some of the programming languages we can do web scrapping.

Top plugins for web scrapping,

1) Web scrapper (google plugin)
2) Scraper (goggle plugin)
3) Scrapy (firefox plugin)

Top tools for web scrapping

1) Import.io
2) Cloud scrape
3) Data tool

XI. DATA QUERY AND ANALYSIS USING HIVE

warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis. While initially developed by Facebook, Apache Hive is now used and developed by other companies such as Netflix. Amazon maintains a software fork of Apache Hive that is included in Amazon Elastic Map Reduce on Amazon Web Services.

Apache Hive supports analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 filesystem. It provides an SQL-like language called HiveQL with schema on read and transparently converts queries to map/reduce, Apache Tez and Spark jobs. All three execution engines can run in Hadoop YARN. To accelerate queries, it provides indexes, including bitmap indexes.

By default, Hive stores metadata in an embedded Apache Derby database, and other client/server databases like MySQL can optionally be used.

Currently, there are four file formats supported in Hive, which are TEXT FILE, SEQUENCE FILE, ORC and RFILE Apache Parquet can be read via plugin in versions later than 0.10 and natively starting at 0.13.
Other features of Hive include:

- Indexing to provide acceleration, index type including compaction and Bitmap index as of 0.10, more index types are planned.
- Different storage types such as plain text, RCFile, HBase, ORC, and others.
- Metadata storage in an RDBMS, significantly reducing the time to perform semantic checks during query execution.
- Operating on compressed data stored into the Hadoop ecosystem-using algorithms including DEFLATE, BWT, snappy, etc.
- Built-in user defined functions (UDFs) to manipulate dates, strings, and other data-mining tools. Hive supports extending the UDF set to handle use-cases not supported by built-in functions.
- SQL-like queries (HiveQL), which are implicitly converted into Map Reduce or Tez, or Spark jobs.

XII. PREDICTIVE ANALYSIS AND DATA VISUALZATION USING R

A. What Is R?

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

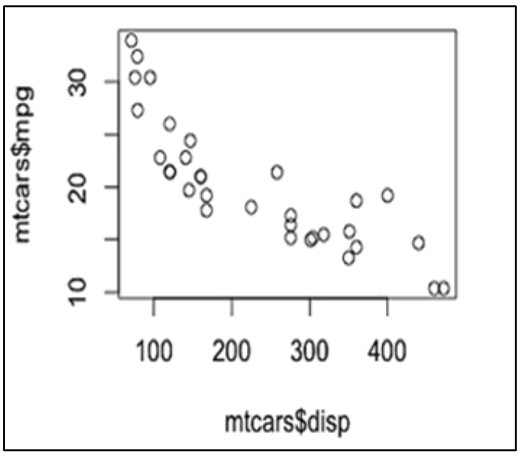
One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.
R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

B. Data Visualization Using R:

One of the most appealing things about R is its ability to create data visualizations with just a couple of lines of code.

For example, it takes just one line of code -- and a short one at that -- to plot two variables in a scatterplot. Let's use as an example the mtcars data set installed with R by default. To plot the engine displacement column disp on the x axis and mpg on y:

plot(mtcars$disp, mtcars$mpg)



C. Simple Scatter Plot Graph Using R:

D. Data Visualization Packages in R:

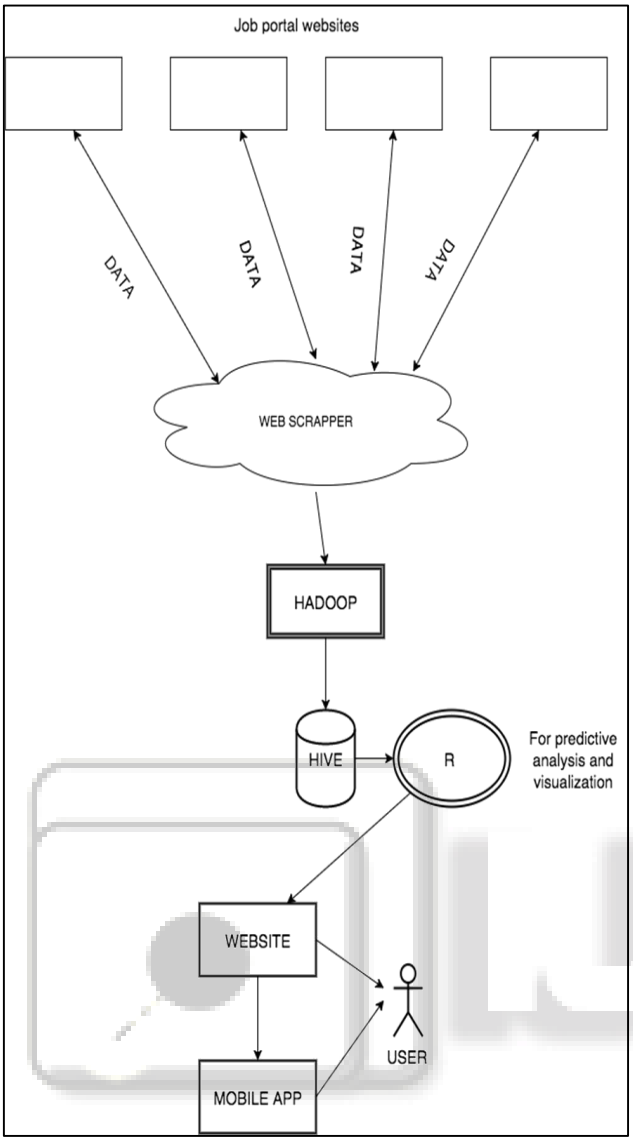
There is lot of open source packages available in r for data visualization. Some of the commonly used packages are

- ggplot2 – one of the best static visualization packages in R
- ggvis – interactive plots from the makers of ggplot2
- rCharts – R interface to multiple JavaScript charting libraries
- plotly – convert ggplot2 figures to interactive plots easily
- googleVis – use Google Chart Tools from R

E. Advantages of Using R:

- R is cross-platform. R runs on many operating systems and different hardware. It is popularly used on GNU/Linux, Macintosh, and Microsoft
- Windows, running on both 32 and 64 bit processors.
- R plays well with many other tools, importing data, for example, from CSV les, SAS, and SPSS, or directly from Microsoft Excel, Microsoft Access, Oracle, MySQL, and SQLite. It can also produce graphics output in PDF, JPG, PNG, and SVG formats, and table output for LATEX and HTML.
- R has active user groups where questions can be asked and are often quickly responded to, often by the very people who developed the environment|this support is second to none. Have you ever tried getting support from the core developers of a commercial vendor?
- New books for R (the Springer Use R! series) are emerging, and there is now a very good library of books for using R.

XIII. OVERALL ARCHITECTURE OF THE PROJECT



REFERENCES

[1] http://blog.fractalanalytics.com/wp-content/uploads/2013/04/Predictive_Analytics_Methdology_Using_R_v1.0.pdf
[2] 2.http://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html
[3] http://www.sas.com/en_us/insights/analytics/big-data-analytics.html
[4] https://en.wikipedia.org/wiki/Web_scraping
[5] http://watin.org/

