Big Data Analysis Using Web Control

Ahmed, Tanvir
Talukder, Ahmed Rohan
Rahman, Md. Mahbubur
Tahsin, Anika

Bachelor of Science in Computer Science

Department of Computer Science

Faculty of Science and Information Technology

American International University – Bangladesh

December 2014

Declaration

We proudly declare that this thesis is our unique work and not even submitted in any form of degree or diploma at any university or other institute of branch in education. Each and every evidence consequent from the published and unpublished work of others has been permissive in the transcript and a list of references is given.

Ahmed, Tanvir
Bachelor of Science in Computer Science

Talukder, Ahmed Rohan
Bachelor of Science in Computer Science

Rahman, Md. Mahbubur
Bachelor of Science in Computer Science

Tahsin, Anika
Bachelor of Science in Computer Science

Approval

The thesis titled “Big Data Analysis Using Web Control” has been submitted to the following respected members of the board of examiners of the department of computer science in partial contentment of the requirements for the degree of Bachelor of Science in Computer Science on 20th December, 2014 by Ahmed,Tanvir (ID: 11-18782-1),Talukder, Ahmed Rohan (ID: 11-18754-1),Rahman, Md. Mahbubur (ID: 11-18614-1) and Tahsin, Anika (11-18554-1) and has been acknowledged as pleasing.

Mohammad Saidur Rahman
Assistant Professor & Supervisor
Department of Computer Science
AIUB

Dr. Dip Nandi
Assistant Professor & Head
Department of Computer Science
AIUB

Prof. Dr. Tafazzal Hossain
Dean
Faculty of Science and Information Technology
AIUB

Dr. Carmen Z. Lamagna
Vice Chancellor
AIUB

Acknowledgement

At the very beginning we would be obleeged to our decent supervisor Mohammad Saidur Rahman and Dr. Md. Rafiqul Islam for their vast support, motivation and cooperative effort. It is a matter of regret that, we get an opportunity to work with them as well as so much proud to be successfully fulfill our requirements. Without their nonstop help it is almost impossible for us to complete this thesis within this short time period.

We are also thankful to our honorable teachers of American International University Bangladesh (AIUB).

We wish to express our gratefulness to American International University Bangladesh (AIUB) for providing an outstanding environment for research.

Our deepest thanks to our dear parents.

Last, but by no means the least, we thank almighty Allah for the gifts and capabilities we were given that made it possible to complete this research.

Abstract

Now a day's big data analysis is a hot topic in every research area. Data scrapping using web content is one of the important part of this. To maintain this huge data is a big challenge. Every day we use huge data in the internet and also store them by local or cloud database. But this data is not structured properly though we may face many problems when need this data. There are so many structured and unstructured data in web fields. We can easily get the structured format data into our database but the critical problem is about unstructured data because they are not in sequence. So it is a vital fact to store this unstructured data without occurring any damages. It is a matter of our motivation that how we can make an asynchronous technique to easily classified this raw data's. To maintain these circumstances one of the key aspect is to use web scrapping. Among all other tools and techniques for efficiently web scrapping, we think our technique has given a better solution and enriched this field to generate a new era.

Contents

Chapter 1

Introduction

Big data analysis with web scrapping is very vital and elegant for modern computation. The chapter is basically an elaboration of the web scrapping. And thereby, proposing a method to overcome these limitations. Before coming to the proposal of this thesis, there will a brief discussion about other schemes that have been proposed to solve those issues, but may not be up to the mark.

1.1 Introduction

In modern Statistics and Data Science, Web technologies have become an essential tool for accessing, collecting and organizing data from the web. Well-known companies, including Google, Amazon, Wikipedia, Facebook and increasingly all data providers, are providing APIs that include specifications for variables, data structures and object classes which facilitate the interaction with other software components such as R.

There have been a number of techniques to collect data from several web content. But the main problem is the data classification and normalization. This thesis is about overcoming this situation of sacrifice and developing a scheme that will attack all the previous algorithms and come up with solution to all at once without requiring one to lag behind. Thus the new scheme shall be proposed in a way that is provides fluent data collection, minimal user waiting time, maximize data reuse capability, categories different data type and differentiate similar words.

The field of Web Content Mining applies data mining techniques to the discovery and extraction of information available on the Web. Web Content Mining comprises several research fields such as Information Extraction or Natural Language Processing which research related technique's that are used to extract data from web documents.

Approaches to the problem of extracting information out of HTML documents considers processing either the DOM tree or the resulting rendering information. The first approach involves defining an extractor or wrapper that selects the relevant information out of the DOM tree. The latter is a vision-based approach that attempts to provide a more general solution to the problem by assuming that similar content types have similar visual features.

Our preliminary goal was to collect each and every raw data's from websites. By collecting this huge number of data set we face some strategic problem at the beginning. While data set was stored successfully the upcoming challenge comes to our hand to classify them with proper type. For text storage we have manipulated some logic to differentiate same word and also count them as well. On the other hand only auxiliary verbs, articles and other unmeaning full words are also been prohibited. Images and Meaningful Texts are stored in a particular path of the computer aquatically and also serialized themselves programmatically.

Chapter: 2 Literature Review

Big Data Analysis Using Web Control

2.1 What is Big Data?

Data is a collection of information. Now a day Data is increasing day by day. At a time warehouse of dataset was 500 terabytes which was claimed by Wal-Mart in 2004.In 1995 Yahoo have created hard drives of 170 petabytes which is 8.5 times of all hard drives created in 1995.Mainly, data size is growing exponentially due to machine generated data (data records, web-log files, sensor data), digital pictures and videos, the usage of social sites, cell phone GPS signals and purchase transaction records to name a few. The data usage is increasing at a high rate that it will never stop. Data is too big or it moves too fast that there must be an alternative way to process it. Scientists define it as "Big Data".

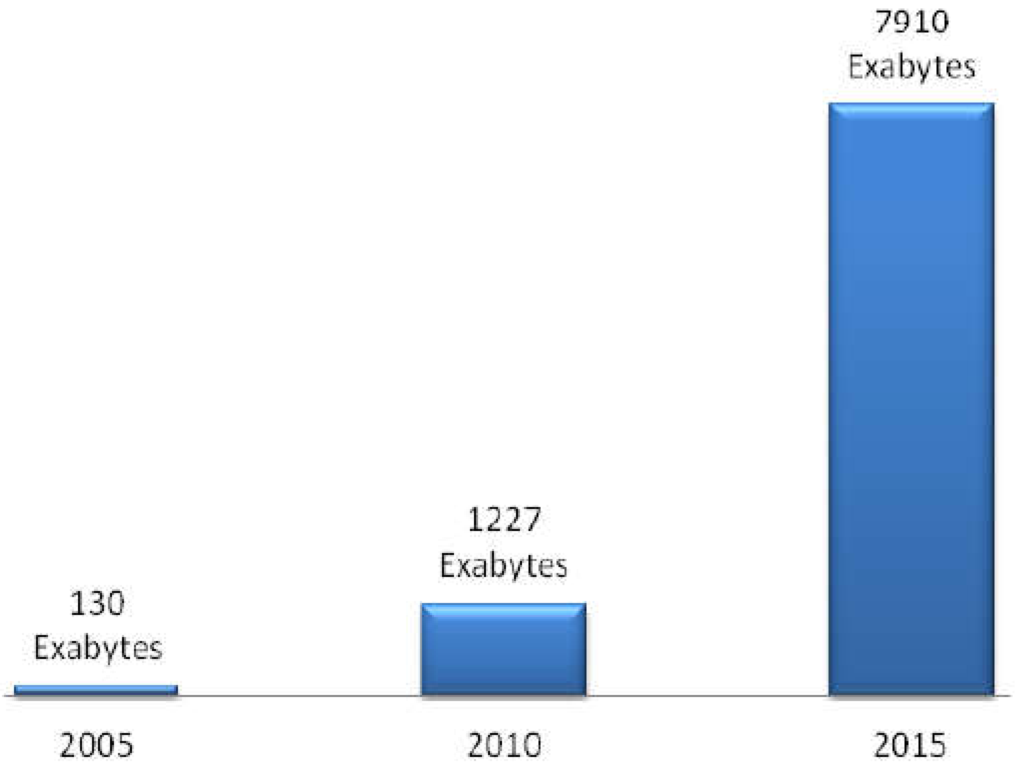


Figure:-Data growth storage in Exabyte's

According to O'Reilly, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it."

According to McKinsey, "Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyses. There is no explicit definition of how big a dataset should be in order to be considered Big Data".

Big Data is not just about the size of data but also includes data variety and data velocity.

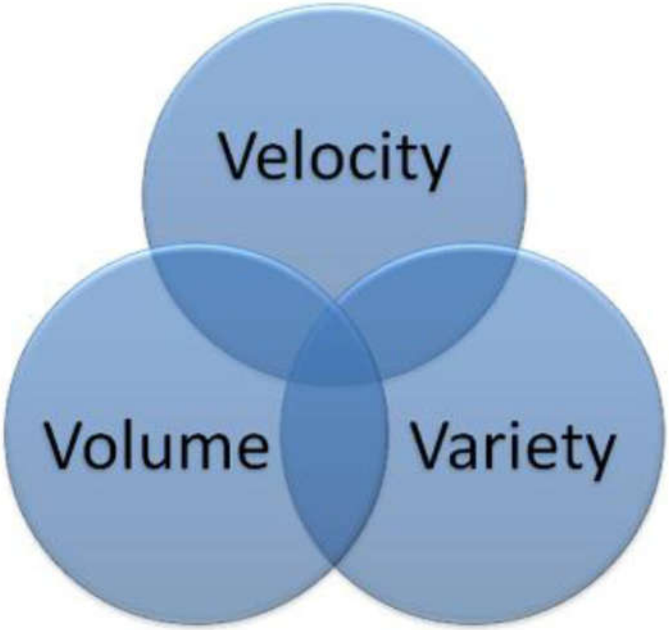


Figure:-Three attributes of Big Data

These three terms define the characteristics of Big Data. Why 3V's is used for big data. 3V's means Velocity, Volume and Variety which judge a big data or characteristics of big data.

Volume: Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes.

Velocity: Social media data streams – while not as massive as machine-generated data – produce a large influx of opinions and relationships valuable to customer relationship management.

Variety: Data can come from a *variety* of sources (typically both internal and external to an organization) and in a variety of types. With the explosion of sensors, smart devices as well as social networking, data in an enterprise has become complex because it includes not only structured traditional relational data, but also *semi-structured* and *unstructured* data.

There are three types of Data which are Structured, Semi-Structured and Unstructured.

So we can define Big Data as it is a progressing terms that huge amount of structured, semi-structured and unstructured data that stored in a warehouse which has 3V characteristics (Volume, Variety and Velocity).

2.2 Big Data Classification:

It's helpful to look at the characteristics of the big data along certain lines — for example, how the data is collected, analyzed, and processed. Once the data is classified, it can be matched with the appropriate big data pattern:

Analysis type — whether the data is analyzed in real time or batched for later analysis. Give careful consideration to choosing the analysis type, since it affects several other decisions about products, tools, hardware, data sources, and expected data frequency. A mix of both types may be required by the use case:

Fraud detection; analysis must be done in real time or near real time.

Trend analysis for strategic business decisions; analysis can be in batch mode.

Processing methodology — the type of technique to be applied for processing data (e.g., predictive, analytical, ad-hoc query, and reporting). Business requirements determine the appropriate processing methodology. A combination of techniques can be used. The choice of processing methodology helps identify the appropriate tools and techniques to be used in your big data solution.

Data frequency and size — how much data is expected and at what frequency does it arrive. Knowing frequency and size helps determine the storage mechanism, storage format, and the necessary preprocessing tools. Data frequency and size depend on data sources:

On demand, as with social media data

Continuous feed, real-time (weather data, transactional data)

Time series (time-based data)

Data type — Type of data to be processed — transactional, historical, master data, and others. Knowing the data type helps segregate the data in storage.

Content format — Format of incoming data — structured (RDMBS, for example), unstructured (audio, video, and images, for example), or semi-structured. Format determines how the incoming data needs to be processed and is key to choosing tools and techniques and defining a solution from a business perspective.

Data source — Sources of data (where the data is generated) — web and social media, machine-generated, human-generated, etc. Identifying all the data sources helps determine the scope from a business perspective. The figure shows the most widely used data sources.

Data consumers — A list of all of the possible consumers of the processed data:

Business processes

Business users

Enterprise applications

Individual people in various business roles

Part of the process flows

Other data repositories or enterprise applications

Hardware — the type of hardware on which the big data solution will be implemented — commodity hardware or state of the art. Understanding the limitations of hardware helps inform the choice of big data solution.

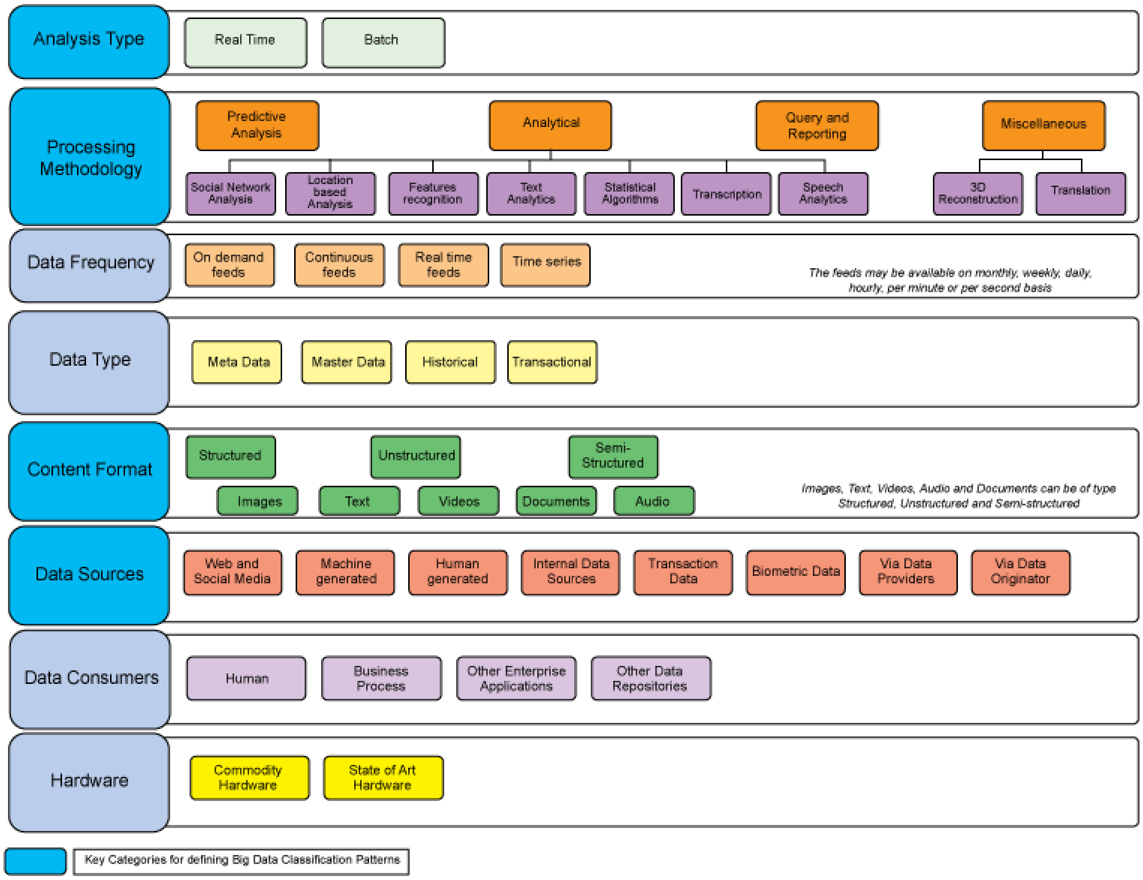


Figure: Various categories for classifying big data

2.3 Why big data is required:

When big data is extracted and analyzed in combination with traditional enterprise data, enterprises can develop a more thorough and insightful understanding of their business, which can lead to enhanced productivity, a stronger competitive position and greater innovation – all of which can have a significant impact on the bottom line. For example, Manufacturing companies install sensors in their products to return a stream of telemetry. In industry, it has to be stored so much information which is used for future forecasting and evaluate the data to analysis production, marketing and so on.

On the other hand, every second satellite has sent millions of terabyte data to the earth which is collected by NASA to store in database and later to evaluate the data. Not only in NASA, our GPS system continuously transmit data to our device which we can use the data. The smart phones and other GPS devices offer advertisers an opportunity to target consumers when they are in close proximity to a store, a coffee shop or a restaurant.

Finally, social media sites like Facebook and LinkedIn simply wouldn't exist without big data. Their business model requires a personalized experience on the web, which can only be delivered by capturing and using all the available data about a user or member. Now the question is how we can manage the huge amount of data. How to use them? How to process the data?

With Big Data databases, enterprises can save money, grow revenue, and achieve many other business objectives, in any vertical.

That's why we need a system that will help to process the data extract the desired output. Now there are so many tools are developing to progression the data to handle our expectations.

2.4 Application of Big Data:

While much of the big data activity in the market up to now has been experimenting and learning about big data technologies, many applications has been focused on also helping organizations understand what problems big data can address.

The Big Data applications are dominated by two classes of technology:

> Systems that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored;
>
> Systems that provide analytical capabilities for retrospective, complex analysis that may touch most or all of the data. These classes of technology are complementary and frequently deployed together.

2.4.1 Operational Big Data:

For operational Big Data workloads, NoSQL Big Data systems such as document databases have emerged to address a broad set of applications, and other architectures, such as key-value stores, column family stores, and graph databases are optimized for more specific applications. NoSQL technologies, which were developed to address the shortcomings of relational databases in the modern computing environment, are faster and scale much more quickly and inexpensively than relational databases.

2.4.2 Analytical Big Data:

Analytical Big Data workloads, on the other hand, tend to be addressed by MPP database systems and Map Reduce. These technologies are also a reaction to the limitations of traditional relational databases and their lack of ability to scale beyond the resources of a single server. Furthermore, Map Reduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL.

Overview of Operational vs. Analytical Systems:

	Operational	**Analytical**
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	Map Reduce, MPP Database

2.5 Tools for Big data analysis:

2.5.1 HIVE Web Based Interface: Hive is a warehouse solution initially developed by Facebook and available now under apache software foundation as an open source framework. The main issue is with Hadoop is to define custom logic for writing map and reduce tasks. Users who even understand what to write in custom logic for analyzing the data cannot write the map and reduce tasks because of lack of programming ability which takes many non-programmer users away from Hadoop and hence they cannot analyze their data without the knowledge the programming language. Hive overcomes this issue by providing the database like structure on top of Hadoop with an interfacing query language called HiveQL12. HiveQL is same like standard database interfacing language SQL13. Users can load their data into Hive by using its data loading facility. Hive communicates with Hadoop and automatically creates the Map Reduce jobs for analysis. Users can also define custom map and custom reduce functions by embedding them into the Hive queries for processing on Hadoop. Figure III shows the architecture of the Hive. In Hive, user issues the queries to the Hive via CLI (command line interface). Hive automatically decide how many mapper and reducers are needed to process the query, hence simplifying the processing wherein Hadoop the users have to define how many mapper and reduces needed to complete the task. Additionally, users can set the number of mapper and reducers needed for the query in the configuration file manually.

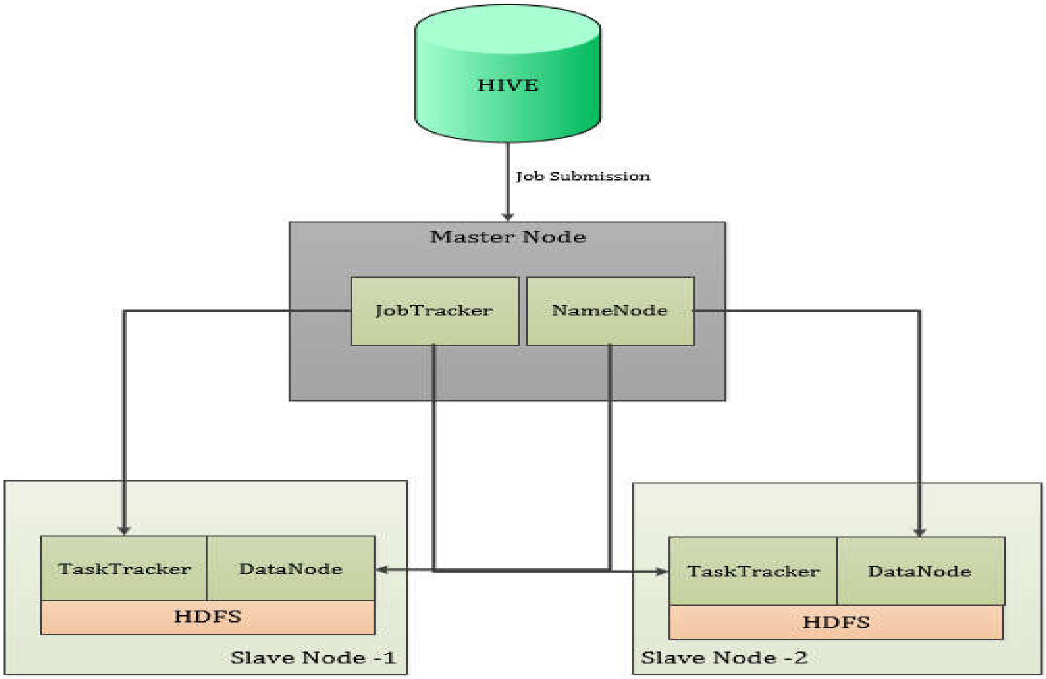


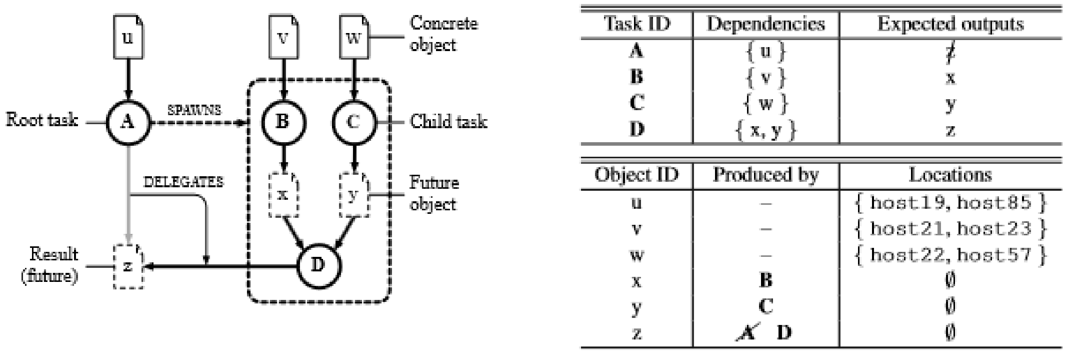
Figure III Hive Architecture [Ref Figure II]

2.5.2 IBM Info Sphere Big Insights:

Info Sphere Big Insights is a Big Data analytics platform from IBM, which support different type of analytics under one roof. Info Sphere is built on top of Hadoop to enhance its capabilities and provides an interactive interface on it for analyzing the Big Data. Info Sphere has built-in analytics capability, including text analytics for getting insights from large textual data, social data analyzer for analyzing social media data, machine data analytics for analyzing machine data such as data from sensors and GPS and Info Sphere also supports the integration with other Big Data technologies. In addition Info Sphere provides a SQL interface, namely as Big SQL and also a spreadsheet like interface called Big Sheets for analyzing andexploring Big Data easily with development tools analytics and security for Big Data operations. Big Sheets and Big SQL modules of Info Sphere are some of the core components of the system. Big SQL provides the facility to user to explore the database schema and provide access to analytics using structured query language. Big Sheets provide Big Data analysis, exploration and manipulation for non-programmers or non-technical users. The system can load data from multiple sources such as databases, web crawlers, text files, Json, csv files etc. and can store it in the distributed file system for processing. Big Sheets provide the user with interactive interface like an excel sheet where users can select or import the data for analysis, such as by applying filters, or by using built-in aggregation functions e.g. sum or average. It also supports user defined functions where users can write their own logic for the computation. Big Sheets also support the visualization of data and analysis results interactively. It supports multiple chart types for rendering the result, such as line, bar and pie charts. Despite of all these benefits and easiness Info Sphere provides, issue with Info Sphere is that, it is not easy to configure the system, even technical computer scientist needs training to install and configure the system, hence Info Sphere taken away the opportunity from non-expert users such as social scientists to use the system without having technical expertise.

2.5.3 CIEL: a universal execution engine for distributed data-flow computing:

Many organizations have an increasing need to process large data sets, and a cluster of commodity machines on which to process them. Distributed execution engines— such as Map Reduce [18] and Dryad [26]—have become popular systems for exploiting such clusters. These says- tems expose a simple programming model, and automatically handle the difficult aspects of distributed com- putting: fault tolerance, scheduling, synchronization and communication. Skywriting is a scripting language that allows the straightforward expression of iterative and recursive task-parallel algorithms using imperative and functional language syntax [31]. Skywriting scripts run on CIEL, an execution engine that provides a universal execution model for distributed data-flow. Like previous sys- tems, CIEL coordinates the distributed execution of a set of data-parallel tasks arranged according to a data-flow DAG, and hence benefits from transparent scaling and fault tolerance. However CIEL extends previous mod- els by dynamically building the DAG as tasks execute. CIEL provides transparent fault tolerance for worker nodes. Moreover, CIEL cantoleratefailuresofthecluster master and the client program. To improve resource until- isation and reduce execution latency, CIEL can memoisetheresultsoftasks. Finally, CIEL supportsthestreaming of data between concurrently-executing tasks.



(a) Dynamic task graph

Task ID	Dependencies	Expected outputs
A	{ u }	~~z~~
B	{ v }	x
C	{ w }	y
D	{ x, y }	z

Object ID	Produced by	Locations
u	–	{ host19, host85 }
v	–	{ host21, host23 }
w	–	{ host22, host57 }
x	B	∅
y	C	∅
z	~~A~~ D	∅

(b) Task and object tables

Figure 2: A CIEL job is represented by a dynamic task graph, which contains tasks and objects (§3.1). In this example, root task **A** spawns tasks **B**, **C** and **D**, and delegates the production of its result to **D**. Internally, CIEL uses task and object tables to represent the graph (§3.3).

2.5.4 Dryad: Distributed Data-parallel Programs:

Dryad is a general-purpose distributed execution engine for coarse-grain data-parallel applications. A Dryad application combines computational “vertices” with communication “channels” to form a dataflow graph. Dryad runs the application by executing the vertices of this graph on a set of available computers, communicating as appropriate through files, TCP pipes, and shared-memory FIFOs. The vertices provided by the application developer are quite simple and are usually written as sequential programs with no thread creation or locking. Concurrency arises from Dryad scheduling vertices to run simultaneously on multiple computers, or on multiple CPU cores within a computer. The application can discover the size and placement of data at run time, and modify the graph as the computation progresses to make efficient use of the available resources. Dryad is designed to scale from powerful multi-core single computers, through small clusters of computers, to data centers with thousands of computers.

2.6 What is Scrapping?

In spite of the increasing presence of Web facilities, only a limited amount of the available resources in the Internet provide a semantic access. Recent initiatives such as the emerging Linked Data Web are providing semantic access to available data by porting existing resources to the semantic web using different technologies, such as database-semantic, mapping and scrapping. The principle goal of web scrapping is to make more computational data structure and also reduce times. It presents a comprehensive review of software tools for social networking media, wikis, really simple syndication feeds, blogs, newspapers, chats and news feeds. For completeness, it also includes introductions to social media scrapping, data cleaning and sentiment analysis. The demand of web scrapping is increasing rapidly. By maintaining big data in the web content, maintaining huge social data store scrapping is mandatory. This also important for research into computational social science that investigates questions (Lazer et al. 2009) using quantitative techniques (e.g. computational statistics, machine learning and complexity) and so-called big data for data mining and simulation modeling (Cioffi-Revilla 2010). The field of web Content Mining applies data mining techniques to the discovery and extraction of information available on the Web. Web Content Mining comprises several research fields such as Information Extraction or Natural Language Processing, which research related techniques that are used to extract data from web documents. Approaches to the problem of extracting information out of HTML documents considers processing either the DOM tree or the resulting rendering information.

2.7 How it helps to Collect Data (Analyzing unstructured Data Scrapping)

Natural Language Processing—(NLP)

NLP is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages. Specially, it is the process of a computer extracting meaningful information from natural language input and or producing natural language output.

News Analytics—

The measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories. Some of these attributes are: sentiment, relevance and novelty.

Opinion Mining—

Opinion mining (sentiment mining opinion/sentiment extraction) is the area of research that attempts to make automatic systems to determine human opinion from text written in natural language.

Scraping

Collecting online data from social media and other Web sites in the form of understanding text and also known as site scraping, web harvesting and web data extraction.

Scraper

A scraper is an automatic agent that is able to extract particular fragments out of the web.

Semantic Analysis—

Sentiment analysis refers to the application of natural language processing, computational linguistics and text analytics to identity and extract subjective information in source materials.

Text Analytics—

Text analytics involves information retrieval (IR), lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization and predictive analytics.

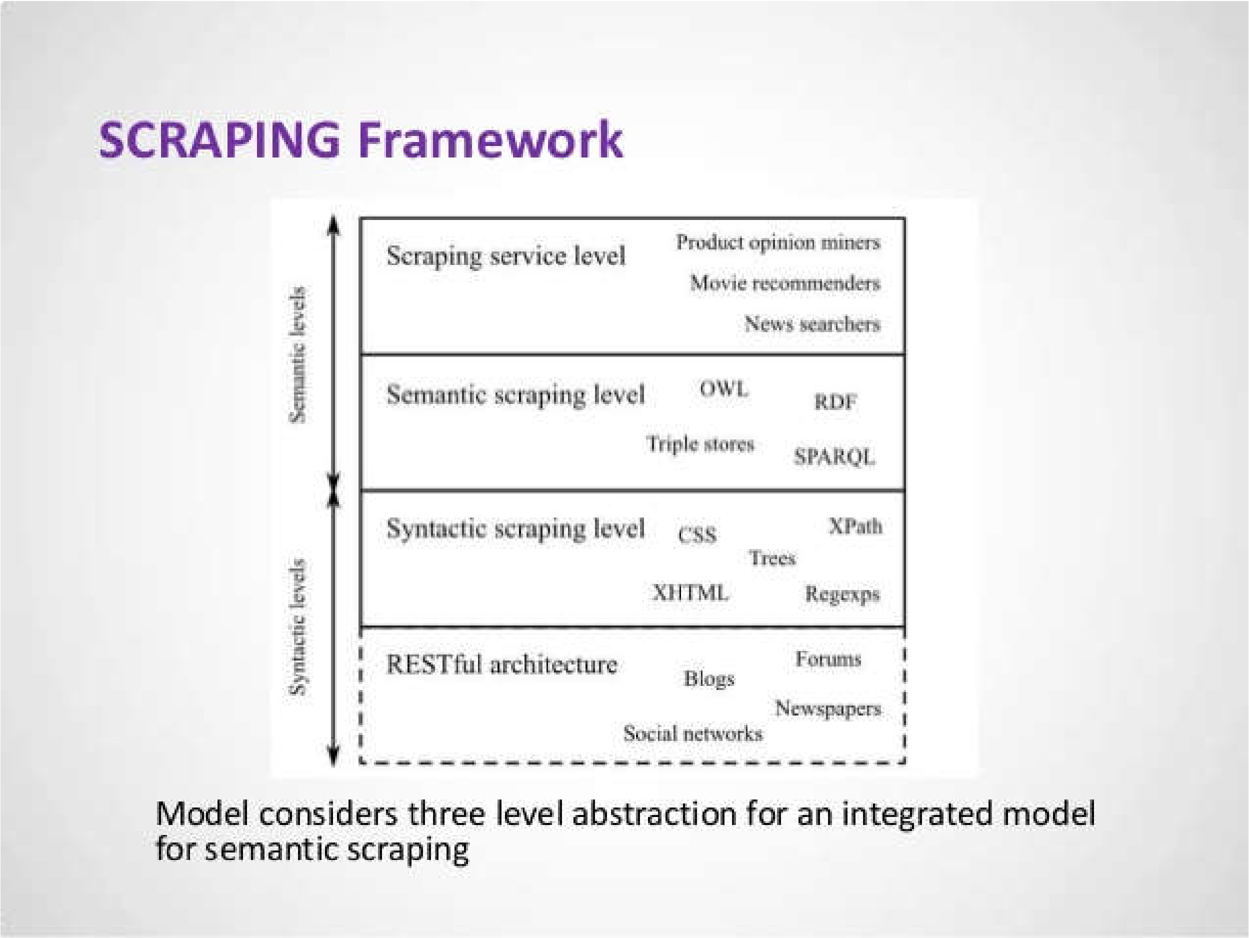
Fragment—
Any element of an HTML document. It serves to represent and traverse a whole sub tree of a document.

Selector—
A condition that indicates which this element is. Different selector terms are defined for each selector type. Selectors can be X-Path expressions, CSS selectors, URI selectors, etc. Selectors are means to identify a web document fragment.

Mapping—
The mapping between a fragment and an RDF resource or blank node. An identifier is defined to map the fragment to a URI. A predicate between the parent's mapped fragment and this is defined to produce an RDF triple. Also, an RDF class can be assigned to the mapped resource of this fragment.

2.8 Data scrapping and their using tools:

Here define a framework for web scraping for the extraction of RDF graphs that represent content in HTML documents. This framework allows defining services based on screen scraping by linking data from graphs with contents defined in HTML documents. They have used this model to build a semantic scrapper that uses RDF-based extractors to select fragments and data from web documents and build RDF graphs out of unstructured information. The model enables to generate graphs in different representations by keeping the original sources in the resulting graph.



SCRAPING Framework

Semantic levels

Scraping service level
Product opinion miners
Movie recommenders
News searchers

Semantic scraping level
OWL
RDF
Triple stores
SPARQL

Syntactic levels

Syntactic scraping level
CSS
XPath
Trees
XHTML
Regexps

RESTful architecture
Forums
Blogs
Newspapers
Social networks

Model considers three level abstraction for an integrated model for semantic scraping

Fig 1: Semantic scrapping framework

First, the framework for scraping of web resources is defined in section 2. The development of a scenario with semantic scrapers that uses the model is described in section 3. Finally, related work is compared with the one presented in this paper, and the main conclusions and future research lines are presented.

2.8.1 SEMANTIC SCRAPING FRAMEWORK

In this paper, a framework for using semantic extracted data from the web is defined, which is shown in figure 1. The model considers three levels of abstraction in order to provide an integrated model for semantic scraping:

- **Scraping service level**

This level comprises services that make use of semantic data extracted from unannotated web resources. Possible services that benefit from using this kind of data can be opinion miners, recommenders, mashups that index and filter pieces of news, etc.

- **Semantic scraping level**

This level defines a model that maps HTML fragments to semantic web resources. By using this model to define the mapping of a set of web resources, the data from the web is made available as knowledge base to scraping services. This level provides semantics to the syntactic scraping capabilities of the level below.

- **Syntactic scraping level**

This level gives support to the interpretation to the semantic scraping model. Wrapping and Extraction techniques such as DOM selectors are defined at this level for their use by the semantic scraping level.

2.8.2 SCRAPING SERVICE LEVEL—

This level comprises all services and applications that make use of the semantic scraping level by providing value to an end user. Services such as opinion miners, recommenders, mashups, data mining applications or any agent-based service benefit from an increased level of knowledge. Other approaches that make use of the semantic scraping facilities can be automatic service composition for automatic generation of new applications out of existing services. Scraping technologies allow getting wider access to data from the web for these kinds of services.

The paradigm behind scraping services has subtle differences from that behind traditional Semantic Web applications or knowledge-based systems. While annotated data in the Semantic Web allows automatic knowledge extraction and retrieval by automatic agents, data in unstructured web documents require prior supervision of some kind to allow information extraction. This implies that when designing a scraping service some of the following steps might be required:

- **Scraping Data Identification**

  Data that wants to be scraped and merged with other knowledge is identified in this task. Target web sites and resources are identified for fragment extraction.

- **Data Modeling**

A model to represent the extracted data is defined in this task. Either existing ontologies might be available or new ones should be defined. The result from this step is an ontology that fits the data that needs to be extracted. A bounded context, i.e. a conceptual context where a domain model has a non-ambiguous meaning, should be identified in order to separate domain models of similar fields. Methodologies for the definition of ontologies can be useful for this task.

- **Extractor generalization**

In order to perform massive extractions, enough samples need to be collected to generalize an appropriate extractor. This collection of samples needs to be provided to a human administrator or an automated or semi-automated module. Using this data set, one or more extractors are defined at the semantic scraping level and serve to provide additional knowledge to the scraping service.

2.8.3 SEMANTIC SCRAPING LEVEL

The paper allows to reference HTML fragments in RDF and define web content extractors, being a basis for the programmatic definition of extractors for screen scraping. This requires bridging the gap between both RDF and HTML's data models. HTML is a markup language for documents with a tree-structured data model. On the other hand, RDF's data model is a collection of node triples, defined by a subject, a predicate, and an object. Each node can be a text literal, a resource (identified by a URI) or a blank node.

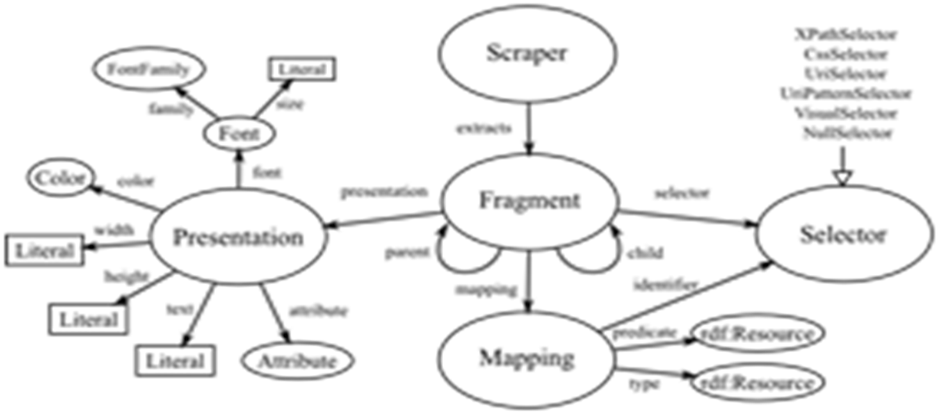


Fig 2: Semantic scraping RDF model

2.8.4 SYNTACTIC SCRAPING LEVEL

This level defines the required technologies to extract data from web resources. It provides a basis for an interpretation of the semantics defined in the semantic scraper level's RDF model. Some of the considered scraping techniques in this level are the following:

Content Style Sheet selectors. Content Style Sheets define the visual properties of HTML elements. These visual properties are mapped to elements through the use of CSS selectors, defined through a specific language. Therefore, CSS is one technology that serves to select and extract data.

URI Patterns:

URI patterns allow to select web resources according to a regular expression that is applied on the resource's URI. While XPath or CSS selectors are able to select an element at document level, URI patterns allow selecting documents, i.e. resources representations, according to the resource's URI.

Visual Selectors:

Visual information can be used to select nodes. HTML nodes are rendered with a set of visual properties given by the used browser. It is common that human users prefer uniform web designs. Web designers thus make elements of a same kind to be rendered with similar visual properties to help identification. A visual selector is a condition that combines several visual properties of an element to identify the element's class.

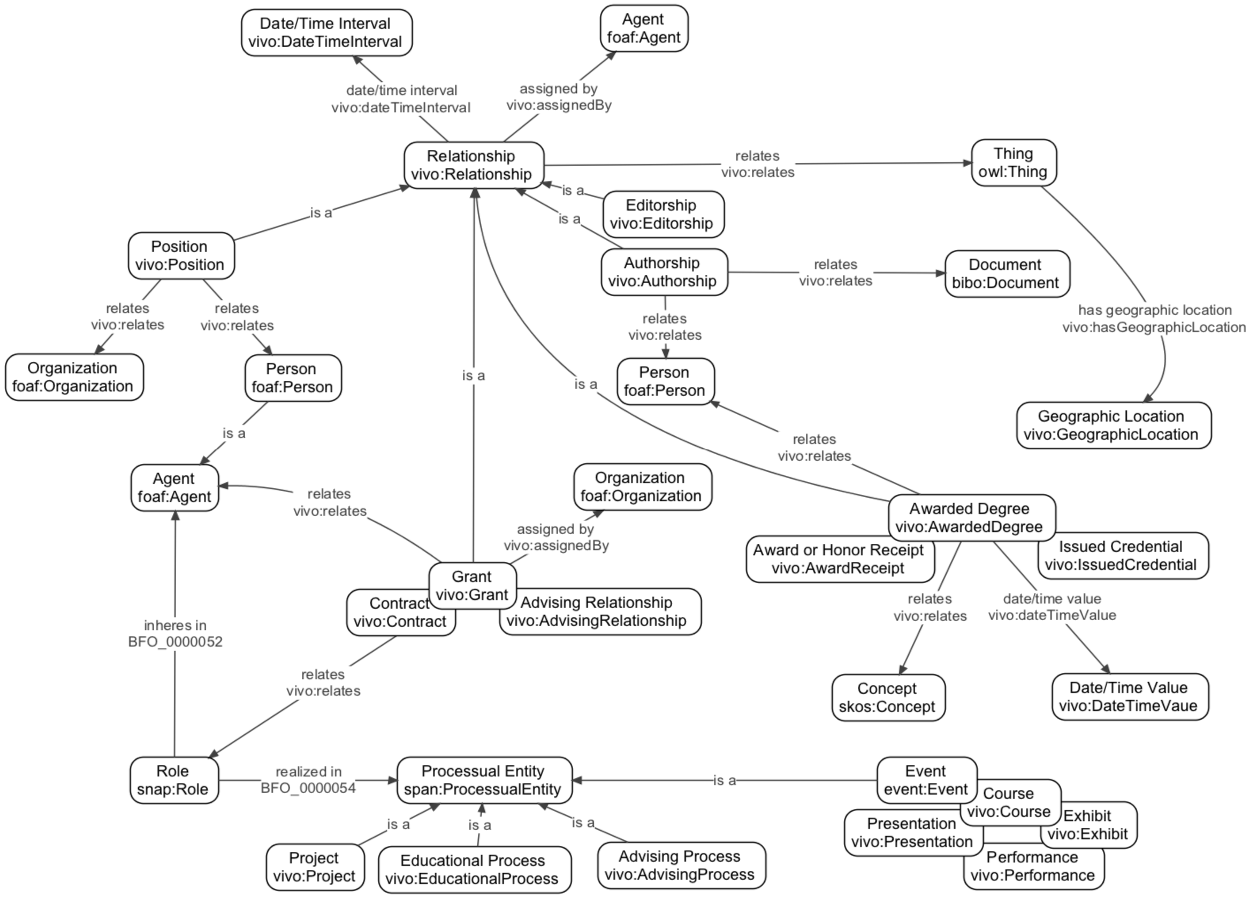


Figure: example of Semantic Scrapper

Chapter 3 an Application to Collect Data for Analysis

Proposed Scheme with Data Scrapping using Web Control:

3.1 **Introduction:**

There are different types of way for data scrapping. Data scrapping is normally used to collect data from forecasting-commerce site, text labeling, Natural Language Processing (NLP), Various Social media sites etc. Data scrapping mainly focused on data collection so that user can easily get various types of data within a short. Dhaka Metropolitan Police(DMP), all kind of national defense, various types of banking organizations, Mobile Company operators, GPS/Satellite handles a huge amount of data daily using data scrapping for various purposes.

DMP basically collects data for security purpose, finding criminal, predicting for future crimes etc.by data scrapping from Social Media, National Database and so on.
National defense uses data scrapping from various sources for issue of national security. For example, national defense can collect personal data from Facebook, Twitter and any kind of social media for national security issue.
Banking organization mainly used data scrapping to predict yearly income of the company, profit, future investment etc.

Now a day's data scrapping from Web resources is increasing gradually. Data scrapping gives us a huge dataset that sometimes we cannot find our appropriate result. Huge data is there on various web resources. To find our particular data or information, data classification is needed. In previous page we have described the classification of various data types. Here in our thesis, we have worked with content types of data like structured, unstructured data, image, text, videos & audio. We have tried to implement such system that will collect data from a web resource. It can be social site, educational site like University, School website. After collecting these huge data, it will classify these data and will give output according to user interest. The system will distinguish these data in text, image and video format. So, user can easily find his choice able data using this system. In modern technology, data usage has been increased at a high rate that data classification is genuinely needed. For example, a user can send e-mail using yahoo account containing with text, image, and video to other. So, everyday yahoo server will be updated with huge set of text, image and video data. So, if we want to exactly find image data from yahoo server, data classification is needed. Exactly, we have tried to create such a system that can easily classify these data and put different data in different folder to access.

3.2 Our Techniques:

Data scrapping processing using web control and their tools have described below. Mainly we used C# with its package and library. Now question is why we choose this platform. Because of there are lots of useful resources and libraries in C#. But the useful way to scrapping data is python, because python is more reliable and faster to scrap data from web control.

For the implementation of these schema we have used several tools and libraries in C# is described below:

Html Agility Pack
Data Web Client
Regular expression
Word counter function
Threading

3.2.1 HTML Agility Pack:

This is an agile HTML parser that builds a read/write DOM and supports plain XPATH or XSLT. It is a .NET code library that allows you to parse "out of the web" HTML files. The parser is very tolerant with "real world" malformed HTML. The object model is very similar to what proposes System.Xml, but for HTML documents (or streams).

3.2.2 Data Web Client:

Web Client downloads files. Found in the System.Net namespace, it downloads web pages and files using the C# language targeting the .NET Framework. This class makes it possible to easily download web pages for testing.

3.2.3 Regular Expression:

A regular expression is a pattern that could be matched against an input text. The .Net framework provides a regular expression engine that allows such matching. A pattern consists of one or more character literals, operators, or constructs.

3.2.4 Threading:

Threads: Threads are often called lightweight processes. However they are not process.es A Thread is a small set of executable instructions, which can be used to isolate a task from a process. Multiple threads are efficient way to obtain parallelism of hardware and give interactive user interaction to your applications.

3.3 Proposed Scheme

3.3.1 Algorithm

We have been implementing this algorithm for data scrapping from web pages. Our proposal algorithm is being given below:

A words list W, A header H, title T and Text Tt are in web browser. This algorithm finds every sentences, words and count same words of H and T in Tt by web browser. Also delete same words from W.

1. Find index of H from web browser by its tag name.
2. Repeat while find all head from web page (Enter loop).
3. Find index of T from web browser by its tag name.
4. Repeat while find all title from web page (Enter loop).

5. Take Tt from T.
6. Find W from words (Enter loop).
7. Check the Tt is not empty.
8. When found W format those words and move forward to next.
9. Replace string and put this into Tt.
10. End of loop (Step 6)
11. Make the word count from Tt.
12. Add this word into sentence put this into T.
13. End of loop (Step 4)
14. Repeat step 4 to 11
15. Add this word into sentence put this into H
16. End of loop (Step 2)

This algorithm is working for collecting words from title, header, and also different dib and tables. For collecting images and videos following algorithm is being approached:

An image tag Img. This algorithm finds every image from Img tags.

1. Find index of img from web browser by its tag name.
2. Repeat while find all image from web page (Enter loop).
3. Make the image count from img tag.
4. Add this count into counter put this into count.
5. End of loop (Step 4)
6. Repeat step 4
7. End of loop (Step 2)

Chapter 4

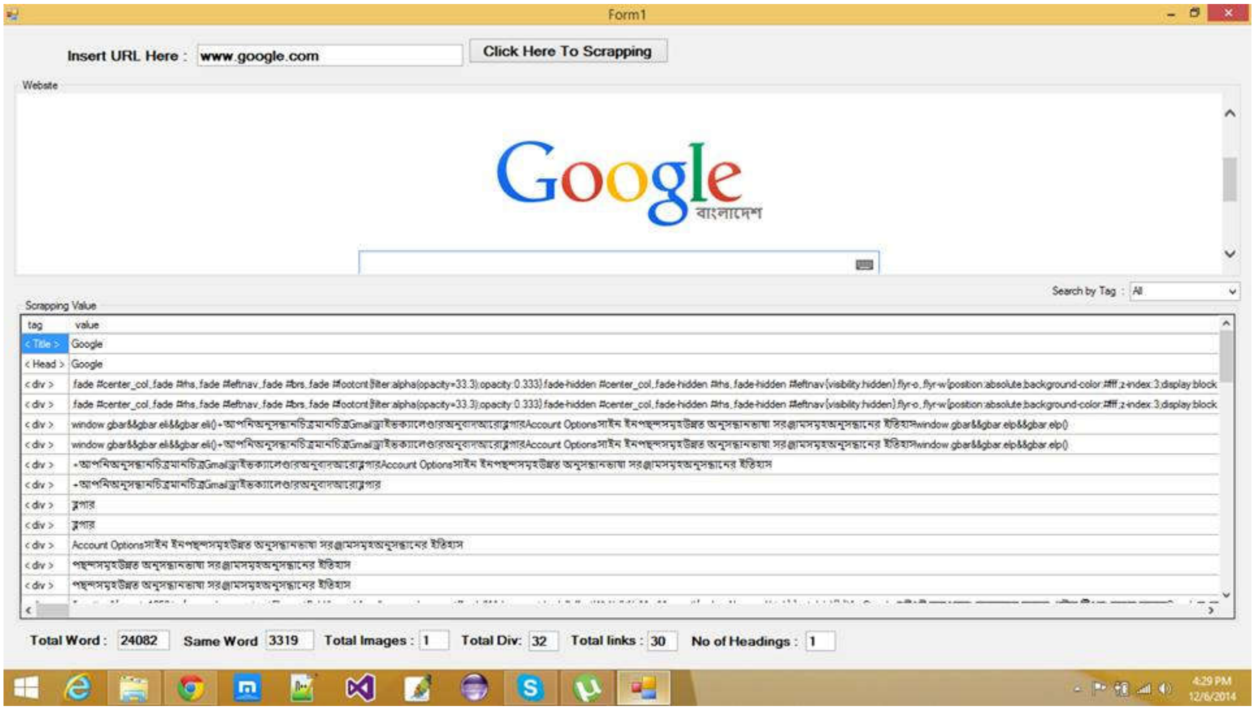
Result

4.1 Introduction

In our schema we have try to collect data from web pages and try to differentiate them with their types and properties. Similar type of words is being distinguished and also count themselves. For the word classification only evocative words are stored from the web. Images are saved with their tag names and also videos as well.

4.2 Graphical Illustrations

Here are some illustrations to back the accomplishments of the proposed scheme based on running different web pages through our algorithms. First we test our algorithm in www.google.com and our resulted schema looks like bellow:



Insert URL Here : www.google.com

Search by Tag : All

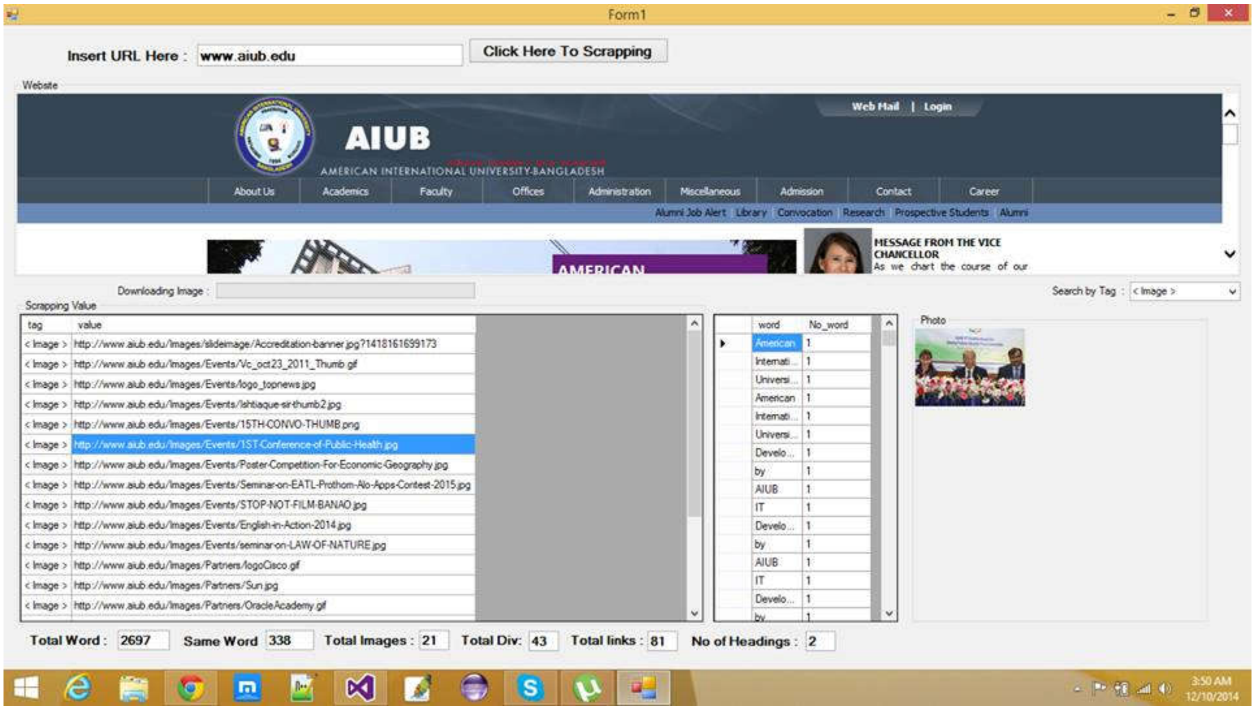
Scrapping Value

tag	value
< Title >	Google
< Head >	Google
< div >	.fade #center_col,.fade #rhs,.fade #leftnav,.fade #brs,.fade #footcnt{filter:alpha(opacity=33.3);opacity:0.333}.fade-hidden #center_col,.fade-hidden #rhs,.fade-hidden #leftnav{visibility:hidden}.flyr-o,.flyr-w{position:absolute;background-color:#fff;z-index:3;display:block
< div >	.fade #center_col,.fade #rhs,.fade #leftnav,.fade #brs,.fade #footcnt{filter:alpha(opacity=33.3);opacity:0.333}.fade-hidden #center_col,.fade-hidden #rhs,.fade-hidden #leftnav{visibility:hidden}.flyr-o,.flyr-w{position:absolute;background-color:#fff;z-index:3;display:block
< div >	window.gbar&&gbar.eli&&gbar.eli()+আপনিঅনুসন্ধানচিত্রমানচিত্রGmailড্রাইভক্যালেন্ডারঅনুবাদআরোব্লগারAccount Optionsসাইন ইনপছন্দসমূহউন্নত অনুসন্ধানভাষা সরঞ্জামসমূহঅনুসন্ধানের ইতিহাসwindow.gbar&&gbar.elp&&gbar.elp()
< div >	window.gbar&&gbar.eli&&gbar.eli()+আপনিঅনুসন্ধানচিত্রমানচিত্রGmailড্রাইভক্যালেন্ডারঅনুবাদআরোব্লগারAccount Optionsসাইন ইনপছন্দসমূহউন্নত অনুসন্ধানভাষা সরঞ্জামসমূহঅনুসন্ধানের ইতিহাসwindow.gbar&&gbar.elp&&gbar.elp()
< div >	+আপনিঅনুসন্ধানচিত্রমানচিত্রGmailড্রাইভক্যালেন্ডারঅনুবাদআরোব্লগারAccount Optionsসাইন ইনপছন্দসমূহউন্নত অনুসন্ধানভাষা সরঞ্জামসমূহঅনুসন্ধানের ইতিহাস
< div >	+আপনিঅনুসন্ধানচিত্রমানচিত্রGmailড্রাইভক্যালেন্ডারঅনুবাদআরোব্লগার
< div >	ব্লগার
< div >	ব্লগার
< div >	Account Optionsসাইন ইনপছন্দসমূহউন্নত অনুসন্ধানভাষা সরঞ্জামসমূহঅনুসন্ধানের ইতিহাস
< div >	পছন্দসমূহউন্নত অনুসন্ধানভাষা সরঞ্জামসমূহঅনুসন্ধানের ইতিহাস
< div >	পছন্দসমূহউন্নত অনুসন্ধানভাষা সরঞ্জামসমূহঅনুসন্ধানের ইতিহাস

Total Word : 24082 Same Word 3319 Total Images : 1 Total Div: 32 Total links : 30 No of Headings : 1

4:29 PM 12/6/2014

Next we check our university official webpage www.aiub.edu and found this result:



Insert URL Here : www.aiub.edu

Downloading Image :

Search by Tag : < Image >

Scrapping Value

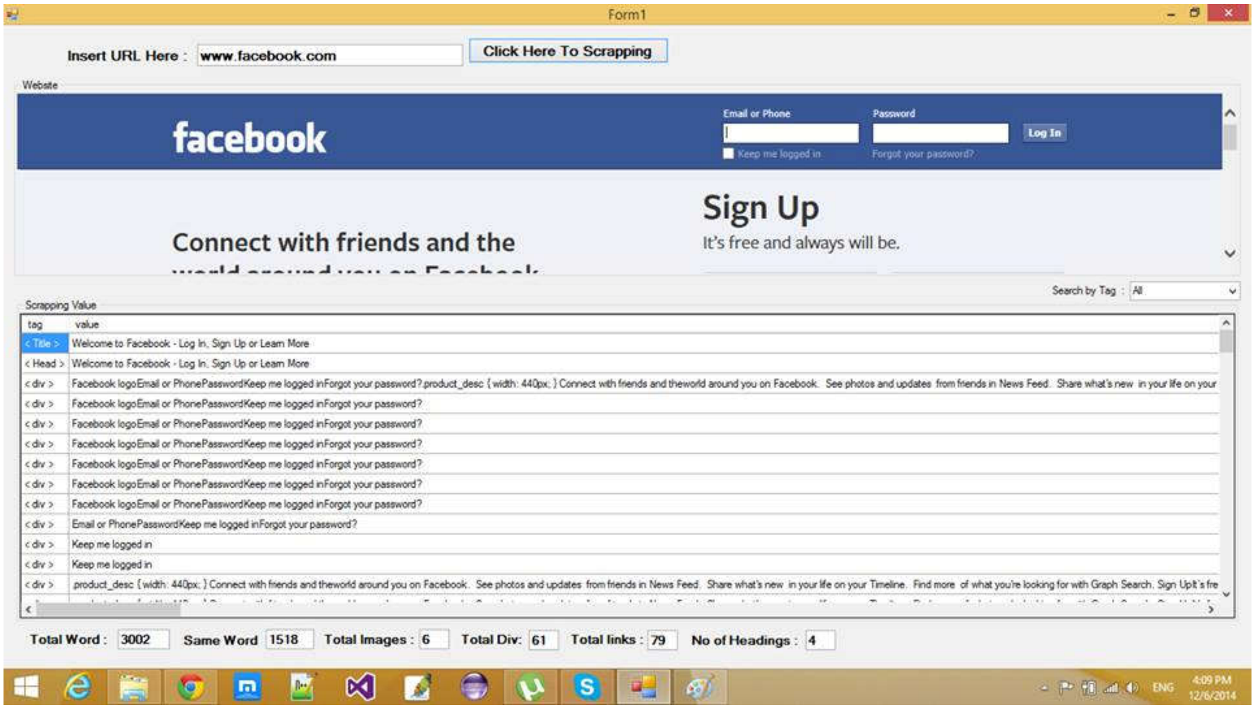
tag	value
< Image >	http://www.aiub.edu/Images/slideimage/Accreditation-banner.jpg?1418161699173
< Image >	http://www.aiub.edu/Images/Events/Vc_oct23_2011_Thumb.gif
< Image >	http://www.aiub.edu/Images/Events/logo_topnews.jpg
< Image >	http://www.aiub.edu/Images/Events/Ishtiaque-sir-thumb2.jpg
< Image >	http://www.aiub.edu/Images/Events/15TH-CONVO-THUMB.png
< Image >	http://www.aiub.edu/Images/Events/1ST-Conference-of-Public-Health.jpg
< Image >	http://www.aiub.edu/Images/Events/Poster-Competition-For-Economic-Geography.jpg
< Image >	http://www.aiub.edu/Images/Events/Seminar-on-EATL-Prothom-Alo-Apps-Contest-2015.jpg
< Image >	http://www.aiub.edu/Images/Events/STOP-NOT-FILM-BANAO.jpg
< Image >	http://www.aiub.edu/Images/Events/English-in-Action-2014.jpg
< Image >	http://www.aiub.edu/Images/Events/seminar-on-LAW-OF-NATURE.jpg
< Image >	http://www.aiub.edu/Images/Partners/logoCisco.gif
< Image >	http://www.aiub.edu/Images/Partners/Sun.jpg
< Image >	http://www.aiub.edu/Images/Partners/OracleAcademy.gif

word	No_word
American	1
Internati...	1
Universi...	1
American	1
Internati...	1
Universi...	1
Develo...	1
by	1
AIUB	1
IT	1
Develo...	1
by	1
AIUB	1
IT	1
Develo...	1
by	1

Total Word : 2697 Same Word 338 Total Images : 21 Total Div: 43 Total links : 81 No of Headings : 2

3:50 AM 12/10/2014

Here we see total word number, total similar word number, total image number, total div number, total link number and also total header number. We can easily get specific type of data by selecting data type get in the right upper corner of the image. Also we test various social media sites like www.facebook.com or www.twitter.com as well as test result is being given bellow:

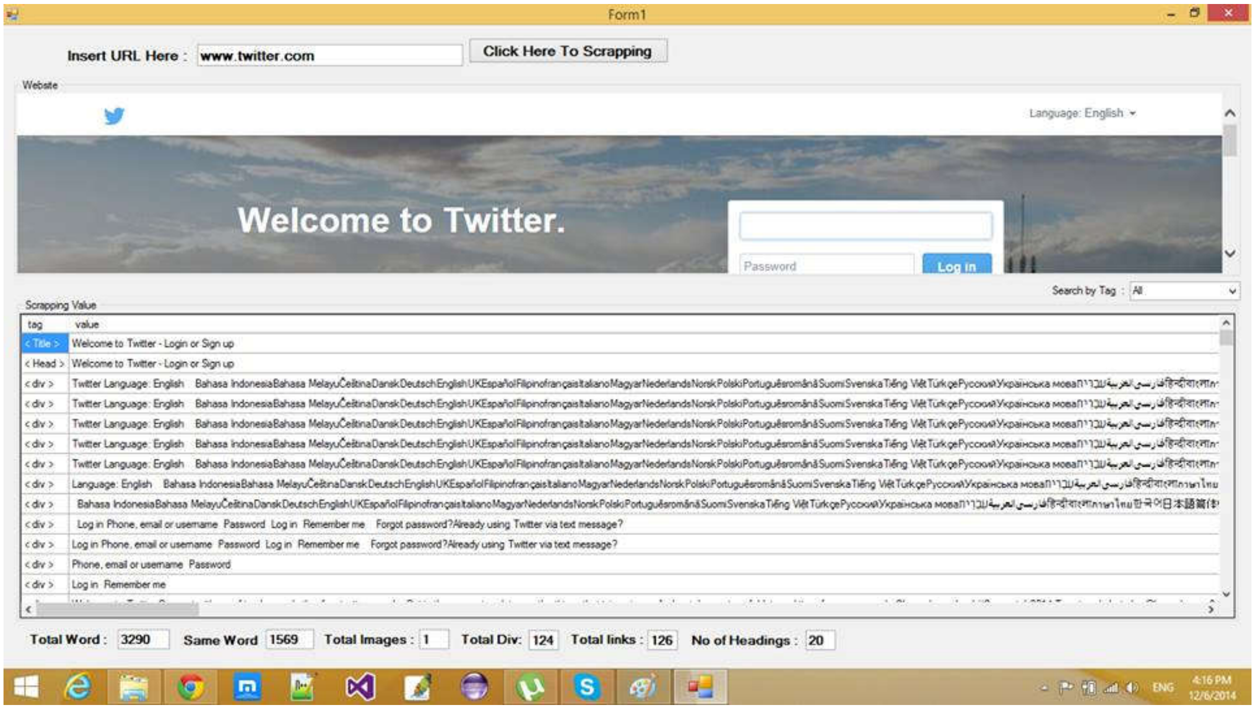


Insert URL Here : www.facebook.com

Search by Tag : All

tag	value
< Title >	Welcome to Facebook - Log In, Sign Up or Learn More
< Head >	Welcome to Facebook - Log In, Sign Up or Learn More
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?.product_desc { width: 440px; } Connect with friends and theworld around you on Facebook. See photos and updates from friends in News Feed. Share what's new in your life on your
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?
< div >	Facebook logoEmail or PhonePasswordKeep me logged inForgot your password?
< div >	Email or PhonePasswordKeep me logged inForgot your password?
< div >	Keep me logged in
< div >	Keep me logged in
< div >	.product_desc { width: 440px; } Connect with friends and theworld around you on Facebook. See photos and updates from friends in News Feed. Share what's new in your life on your Timeline. Find more of what you're looking for with Graph Search. Sign Up!It's fre

Total Word : 3002 Same Word 1518 Total Images : 6 Total Div: 61 Total links : 79 No of Headings : 4



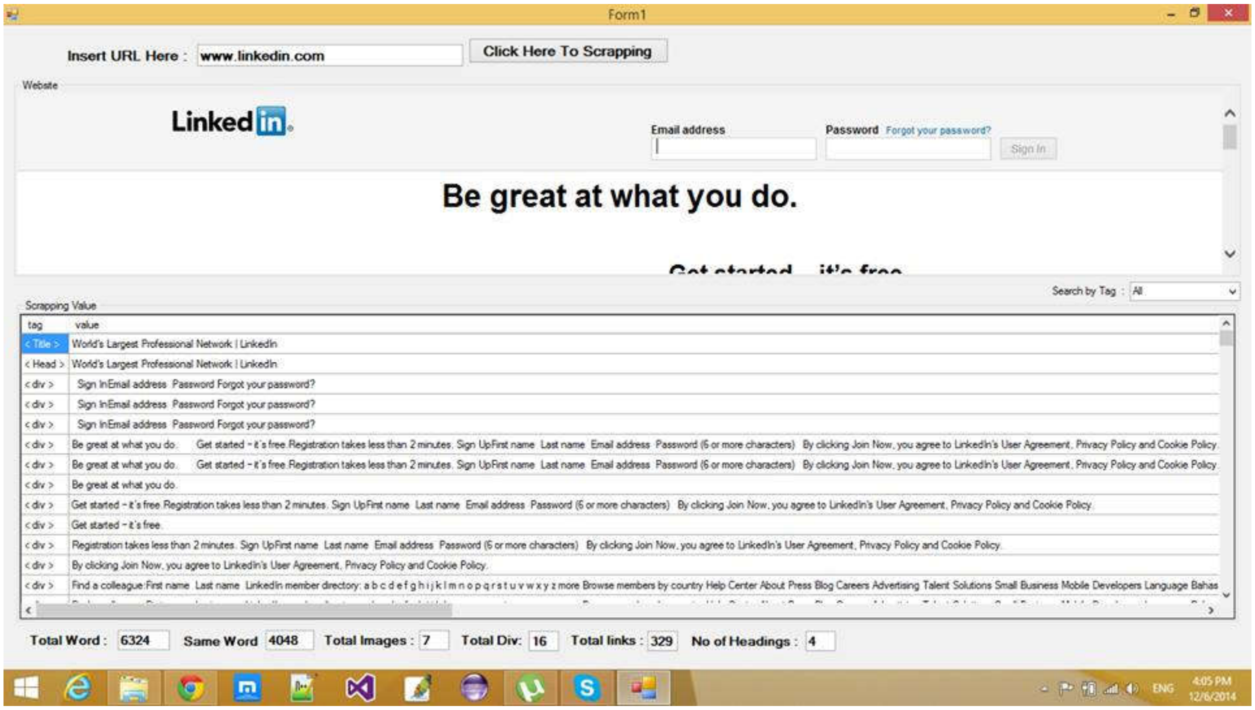
Insert URL Here : www.twitter.com

Search by Tag : All

tag	value
< Title >	Welcome to Twitter - Login or Sign up
< Head >	Welcome to Twitter - Login or Sign up
< div >	Twitter Language: English Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলা
< div >	Twitter Language: English Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলা
< div >	Twitter Language: English Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলা
< div >	Twitter Language: English Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলা
< div >	Twitter Language: English Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলা
< div >	Language: English Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলাภาษาไทย
< div >	Bahasa IndonesiaBahasa MelayuČeštinaDanskDeutschEnglish UKEspañolFilipinofrançaisItalianoMagyarNederlandsNorskPolskiPortuguêsromânăSuomiSvenskaTiếng ViệtTürkçeРусскийУкраїнська моваעבריתالعربيةفارسیहिन्दीবাংলাภาษาไทย한국어日本語簡体
< div >	Log in Phone, email or username Password Log in Remember me Forgot password?Already using Twitter via text message?
< div >	Log in Phone, email or username Password Log in Remember me Forgot password?Already using Twitter via text message?
< div >	Phone, email or username Password
< div >	Log in Remember me

Total Word : 3290 Same Word 1569 Total Images : 1 Total Div: 124 Total links : 126 No of Headings : 20

For each and every search we can get our required data easily and also stored them whenever we need. Also checked www.linkedin.com.



Insert URL Here : www.linkedin.com

Search by Tag : All

tag	value
< Title >	World's Largest Professional Network | LinkedIn
< Head >	World's Largest Professional Network | LinkedIn
< div >	Sign InEmail address Password Forgot your password?
< div >	Sign InEmail address Password Forgot your password?
< div >	Sign InEmail address Password Forgot your password?
< div >	Be great at what you do. Get started – it's free Registration takes less than 2 minutes. Sign UpFirst name Last name Email address Password (6 or more characters) By clicking Join Now, you agree to LinkedIn's User Agreement, Privacy Policy and Cookie Policy.
< div >	Be great at what you do. Get started – it's free Registration takes less than 2 minutes. Sign UpFirst name Last name Email address Password (6 or more characters) By clicking Join Now, you agree to LinkedIn's User Agreement, Privacy Policy and Cookie Policy.
< div >	Be great at what you do.
< div >	Get started – it's free Registration takes less than 2 minutes. Sign UpFirst name Last name Email address Password (6 or more characters) By clicking Join Now, you agree to LinkedIn's User Agreement, Privacy Policy and Cookie Policy.
< div >	Get started – it's free.
< div >	Registration takes less than 2 minutes. Sign UpFirst name Last name Email address Password (6 or more characters) By clicking Join Now, you agree to LinkedIn's User Agreement, Privacy Policy and Cookie Policy.
< div >	By clicking Join Now, you agree to LinkedIn's User Agreement, Privacy Policy and Cookie Policy.
< div >	Find a colleague:First name Last name LinkedIn member directory: a b c d e f g h i j k l m n o p q r s t u v w x y z more Browse members by country Help Center About Press Blog Careers Advertising Talent Solutions Small Business Mobile Developers Language Bahas

Total Word : 6324 Same Word 4048 Total Images : 7 Total Div: 16 Total links : 329 No of Headings : 4

Here we see search result as well as the word and other content numbers. We have also checked other popular web sites in the internet and each and every time our algorithm worked properly.

Chapter 5

Conclusion and Future Works

Site Scraper has meet our goals to make web scraping easy and automatic retraining possible. It has proved a convenient tool for extracting data from web. Our approach based on learning patterns using XPath, allowed us to produce a system that can satisfy user needs with high precision and recall with minimal training.

At first we tried to develop our system using Hadoop and Map Reduce Technology. But for some incomplete resources we moved on another platform (C#) which is very helpful to establish such data crapping system. To complete this process, we faced some data scrapping problems. To cope with this problem, we choose static websites for desirable output, enhance we classified the text and image data accurately from websites.

What we actually tried to establish, we have succeeded to develop such desirable system with some little incompleteness. We actually classified the web data into text and image only. However in future, we will work on classified video as well as differentiate them according to their category. Moreover, we will try to re-establish our system so that it can scrap data from dynamic website. After all, we will mainly focus to develop Data Scrapping API that will be open source. So that later if anyone try to work on scrapping, they can use our API.

Bibliography

[1] Bogdan Batrinca , Philip C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms”, vol.2 pp. 2-5. Article is published with open access at Springerlink.com.

[2] Alberto Cavallo, MIT Sloan, “Scraped Data and Sticky Prices”, Published May 2013.

[3] Jose´ Ignacio Ferna´ndez-Villamor, Jacobo Blasco-Garc´ıa, Carlos A´. Iglesias, Mercedes Garijo, “A SEMANTIC SCRAPING MODEL FOR WEB RESOURCES”, vol. 2, pp. 2-5.

[4] Richard Baron Penman, Timothy Baldwin, David Martinez, “Web Scraping Made Simple with SiteScraper” vol. 4 pp. 4-6.