

Social Media Blog Analysis using Watson Analytics

Sindhu B. Jigali*, Nirmala C.R.

Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology,
Davanagere, Karnataka, India

Abstract

Analytics is defined as the discovery, interpretation, and communication of meaningful patterns in data. A blog is a discussion or informational website published on the world wide web consisting of discrete, often informal diary-style text entries and may contain unstructured or semi-structured data. The analysis of unstructured data types is current challenge, getting attention in industry. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization and information privacy. Watson analytics is a cloud based smart data discovery tool from IBM analytics. Watson solutions aim to enhance, scale, and accelerate human expertise, targeting a wide range of complex challenges. AlchemyAPI is a collection of APIs that offer text analysis through natural language processing. The keyword extraction API works on URL's, HTML documents and plain text. It automatically detects the language of the content and then performs the appropriate analysis. AlchemyAPI's keyword extraction algorithm employs sophisticated statistical algorithms and natural language processing technology aided with machine learning techniques to analyze the content and identify the relevant keywords. The proposed blog analyzer application uses the Alchemy API's keyword extraction service, accepts the URL's of the blogs and processes the content by employing sophisticated statistical algorithms and natural language processing aided with machine learning techniques, to retrieve the related/relevant and rank the keywords of the input content.

Keywords: Blog, keywords, Watson analytics, Alchemy API, keyword extraction service

***Author for Correspondence** E-mail: sindhubjigali94@gmail.com

INTRODUCTION

Big data applies to the information that cannot be processed or analyzed using traditional processes or tools. Organizations have access to a wealth of information but they don't know how to get value out of it because it is sitting in its most raw form, unstructured or semi-structured format [1]. According to the survey conducted by IBM, it is found that half of the business leaders today realize that they don't have access to the insights that they need to fulfill their jobs [1]. Characteristics of Big data are; volume, velocity, variety, validity, veracity, volatility and viability.

Volume

It is the quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not [2]. The volume of big data is very high and it is much difficult to manage the data with traditional database systems.

Velocity

Velocity refers to the rate at which the data is being generated continuously from various sources. Now a days social networking sites are producing data in high volume with high velocity, which requires a fast and real time processing to manage this data.

Variety

Big data is very heterogeneous in nature. This data can be structured, semi-structured and unstructured. Data being generated from different sources in different formats are stored to process and analyze [1].

Validity

As the volume of the data is very high, validity is the most important requirement to fetch the accurate and relevant data for analysis purpose. This valid data would enhance the decision support system [1].

Veracity

The quality of captured data can vary and affect the accurate analysis.

Volatility

Data storage is the critical issue for Big data. How long the data is valid and for how much period the data has to be kept in databases, are the issues related with volatility of big data. Volatility issues are to be fixed before the data is analyzed.

Viability

It refers to choose the appropriate factors and relevant features of the data set under consideration which contributes maximum in predicting the result of analysis [1].

A blog (truncation of weblog) is a discussion or informational website published on the world wide web, consisting of discrete, often informal diary-style text entries called as "Posts" [3]. Posts are typically displayed in reverse chronological order, so that the most recent post appears first, at the top of the web page. Earlier, blogs were usually the work of a single individual occasionally of a small group, and often covered a single subject or topic.

Analytics is defined as the discovery, interpretation, and communication of meaningful patterns in data [1]. Especially valuable in areas which are rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight. The analysis of unstructured data type

is another challenge getting attention in the industry.

Watson Analytics: A New Era of Decision Making

Watson analytics is a smart data discovery tool from IBM analytics (Figure 1). Watson solutions aim to enhance, scale, and accelerate human expertise targeting a wide range of complex solution [4]. Three ways were identified to bring Watson into market, they are:

- Natural language processing.
- Hypothesis generation.
- Evidence based learning.

It leverages the cognitive capabilities of Watson technology to accelerate daily analytical activities of business people in three ways:

- Provides semantic recognition in input.
- Identifies the starting points for analysis.
- Enables interaction in natural language.

Watson uses IBM's DeepQA software and the Apache UIMA (Unstructured Information Management Architecture) framework. The system is written in various languages, including Java, C++, and Prolog, and runs on the SUSE Linux Enterprise Server 11 operating system using Apache Hadoop framework to provide distributed computing. Watson services use natural language processing and machine learning techniques [4]. Many challenges in NLP involve: natural language understanding, enabling computers to derive meaning from human or natural language input; and others involve natural language generation.

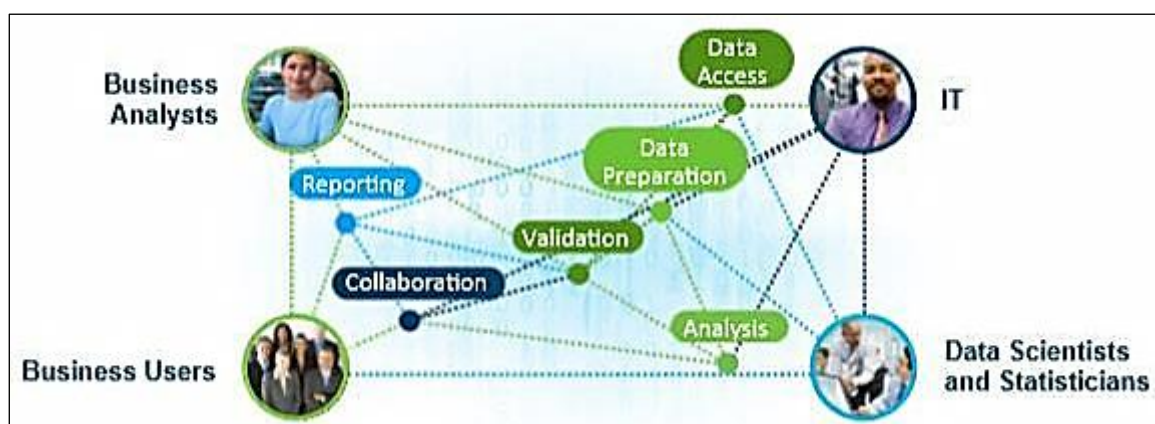


Fig. 1: Analytics Made Easy by Watson Analytics.

NLP comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory, and linear algebra. The technology for statistical NLP involves learning from data.

Machine Learning is a part of AI that continuously observes a series of actions performed over a period of time, and puts the obtained knowledge to use by devising the ways to perform similar processes better, in a new environment [4].

The services provided by Watson analytics are AlchemyAPI, conversation, document conversion, language translator, natural language classifier, personality insights, retrieve and rank, speech to text, tone analyzer, visual recognition and trade off analytics.

LITERATURE SURVEY

Big data analytics makes use of natural language processing techniques, association rule mining, machine learning techniques etc. to process and analyze the data [5]. These techniques help to mine the unstructured data. Many frameworks have been introduced with Hadoop functionality, extended with machine learning techniques to get the accurate analysis of the input data [6, 7]. The earlier proposed frameworks and analytics approaches, costs for the computing resources and the platforms; they lag in drawing the semantics from the data using statistical machine learning algorithms. They don't provide visualization for the results of analytics which makes difficult in analyzing the outcome of analytics [8]. It is very difficult to build the ontology for analytics with the existing approaches because integration of various analytics processes and approaches are difficult with platform dependency and knowledge database can't be easily built [9, 10].

These drawbacks can be overcome by the use of Watson analytics provided by IBM Bluemix cloud. Watson analytics is a collection of services which is built on Apache UIMA framework [11, 12]. It is a cloud based smart data discovery tool that learns from data using natural language processing, statistical algorithms and machine learning techniques etc. [4]. It provides services for cleaning,

transforming, creation of conclusions and decision making on true meaning of the data. It makes analysis easier with integrated services. Watson analytics is able to analyze the data just in four steps:

1. Analysis of question;
2. Hypothesis generation;
3. Evidence scoring; and
4. Final merging and ranking.

The services of Watson analytics can be easily used with the API_keys provided for each service. The services of the Watson analytics can be integrated with other services, which enables user to draw valuable insights from the data.

Problem Statement

The earlier proposed approaches lag in drawing the semantics from the data using statistical machine learning algorithms. They don't provide visualization for the results of analytics which makes difficult in analyzing the outcome of analytics. It is very difficult to build the ontology for analytics with the existing approaches because integration of various analytics processes and approaches are difficult with platform dependency and knowledge database can't be easily built.

Objective

The objective is to design an application that would efficiently extract the relevant keywords of the social media blogs and rank them.

ARCHITECTURE

The proposed application workflow is designed using the MVC [Model, View, Controller] architectural pattern. It consists of a view, controller and Bluemix cloud providing Alchemy_API service. It is one of the services of Watson analytics (Figure 2).

The view component accepts the URL as input and forwards it to the controller. The controller is responsible for forwarding the request to the server. The API_key for keyword extraction service is added to the controller. Then the request is sent by controller to server/Bluemix cloud running keyword extraction service of Watson analytics.

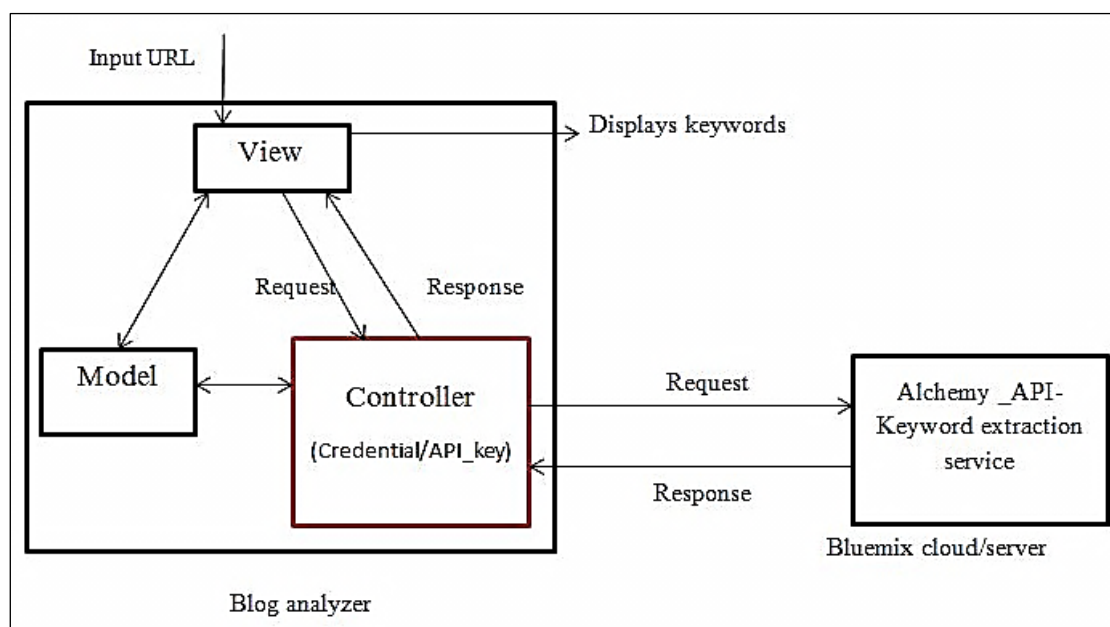


Fig. 2: Blog Analyzer Application Using Watson Analytics.

The request associated with the API_key is sent to the service in the form of a function call with the input URL as parameter of the function, where in the definition of the function is embedded within the service. Keyword extraction service processes input and extracts the keywords relevant to the input URL and rank them. It sends the response to the controller in the form of JSON object containing relevant/related keywords to the input. The object contains the keywords in ranked order. This object is forwarded to the view.

View processes the object and displays the extracted keywords that are related to the input URL to the user. The keywords are displayed in the ranked order.

IMPLEMENTATION AND PROTOTYPE

The following are the procedures to be followed in designing the blog analyzer application. It requires an API_key of the keyword extraction service that is running on Bluemix cloud [4]. Once the key is integrated with the application, the application can make use of the service. The service processes the input content and provides the response to the request made by the application.

Getting the Alchemy_Language API Key

```
var express = require('express');
var router = express.Router();
```

```
var watson = require('watson-developer-cloud');
```

```
var alchemy_language =
  watson.alchemy_language({
```

```
  api_key:
    '81e32a4cc260b79a714c23f4f3e7b1a833332e2f'});
```

The API key for the keyword extraction service of Alchemy_language is added to the blog analyzer application so that the application is able to use the service running on the server.

Processing the Input URL

```
router.get('/', function(req, res, next) {

  res.render('index', {"url": '', "data": ''});

  router.post('/analyze', function(req, res, next) {

    var parameters = {};

    parameters = {url: req.body.url,

      knowledgeGraph: 1};
```

The URL is parameter for the function alchemy_language.concepts. This function is defined by the alchemy_language service. The URL is sent to the service in the function call of alchemy language service.

Processing by Alchemy Language

Alchemy language's keyword extraction service uses sophisticated natural language processing techniques aided with machine learning to process the user's input and drive insights from it. Input to the service is the URL of the blog that is to be analyzed. It automatically identifies the supported language then identifies and ranks the keywords in the input content. The list of keywords is sent to controller in JSON object.

Response from Alchemy_Language Service

```
alchemy_language.concepts(parameters,
function (error, response) {
if (error)
console.log('error:', error);
else
console.log(JSON.stringify(response, null, 2));
res.render('index',{ "url": req.body.url, "data":
response});
```

```
module.exports = router;
```

The service returns the response in a function containing JSON object. It contains the keywords related to the blog in an order.

Display of Result

```
<% data.concepts.forEach(function(data) {%>
<li><%= data.text %></li><br/>
```

The JSON object is processed by the above loop to display the keywords related to the analyzed blog.

Software requirements:

1. Node.js
2. CLI-cloud foundry tool for Windows operating system.

RESULTS

Results are given in Figures 3 to 6.

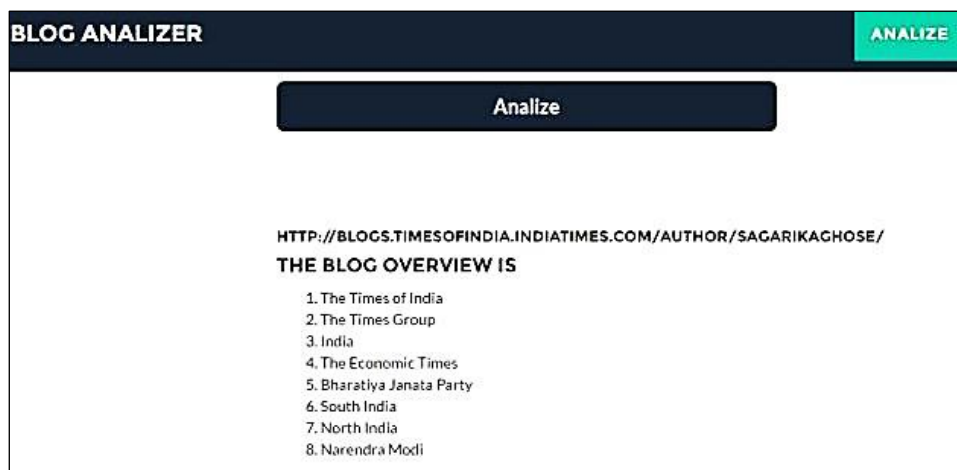


Fig. 3: Overview of Blog



Fig. 4: Page analyzes the blog and specifies relevant keyword about the blog.

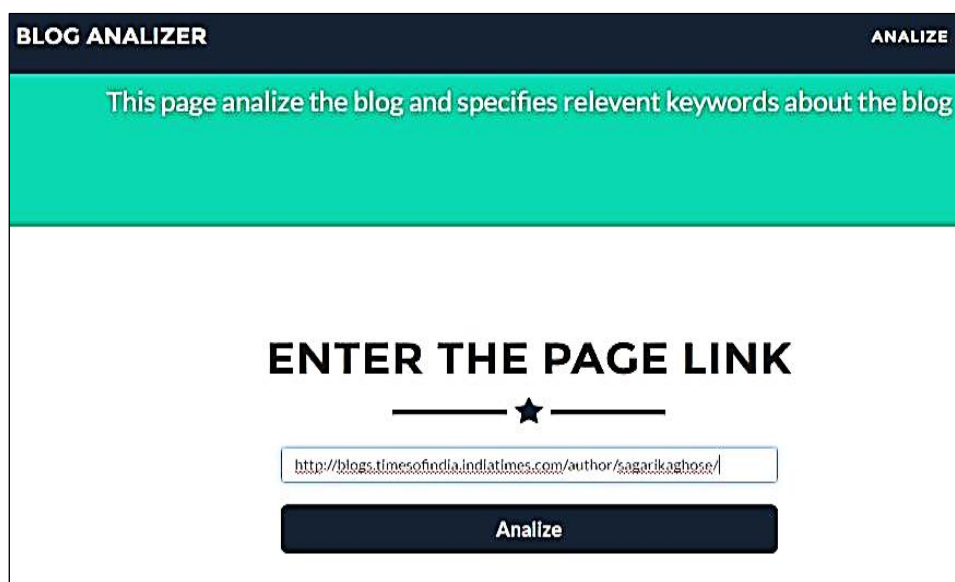


Fig. 5: Required the Name of Page.

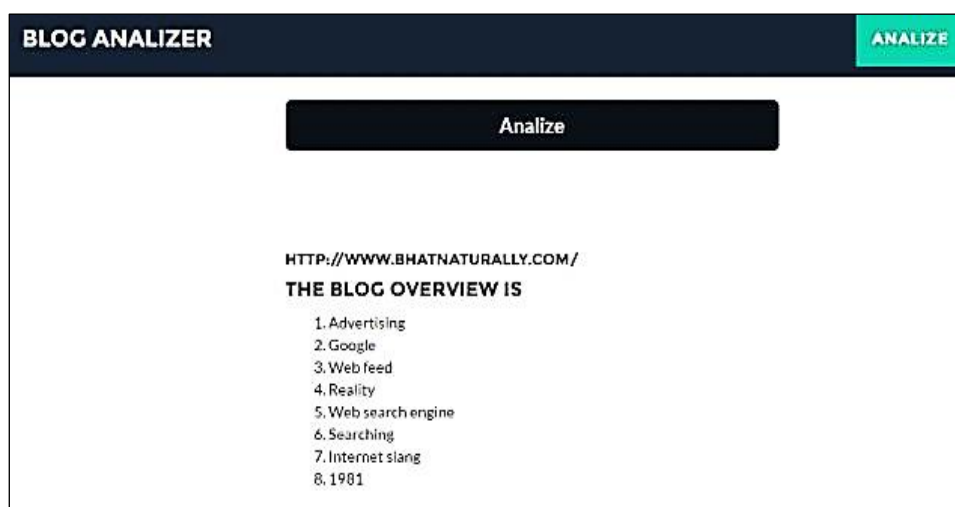


Fig. 6: Result of Entered Page Link.

CONCLUSION

Big data analytics is widely growing area for the analysis and forecasting of trends and also to derive decisions based on the data. Blog is a one such example that contains wide variety of data; and analyzing the content of the blogs and deriving insights from the content is not an easy job. The proposed blog analyzer application is able to identify the keywords related to the blog and rank them. Blog analyzer uses keyword extraction service of Watson analytics provided by IBM Bluemix Cloud. One of the advantages is that it is a cloud based data discovery tool built on Apache UIMA framework that makes analysis easier by providing API interface to the user, to integrate the services to the application. The future scope of the proposed work is that the

blog analyzer can be extended with functionality to identify the entities like people, places etc. and would be able to extract the concept of the blog, sentiment like opinion and feelings in the blog by utilizing the appropriate services of Watson analytics.

REFERENCES

1. Understanding Big Data. IBM.
2. https://en.wikipedia.org/wiki/Big_data
3. <https://en.wikipedia.org/wiki/Blog>
4. www.ibm.com/cloud-computing/bluemix/
5. Yunus Yetis, et al. *Application of Big Data Analytics via Cloud Computing*. IEEE; 2016.
6. Alisa Sotsenko, et al. *Using a Rich Context Model for Real-Time Big Data Analytics in Twitter*. IEEE; 2016.

7. Awais Ahmad. Defining Human Behaviors using Big Data Analytics in Social Internet of Things. *2016 IEEE 30th International Conference on Advanced Information Networking and Application*.
8. Omesh Kumar, et al. *Visualization: A Novel Approach for Big Data Analytics*. 2016.
9. Nural Mustafa V, et al. *Using Semantics in Predictive BigData Analytics*. 2015.
10. Ravi Vatrpu, et al. *Social Set Analysis: A Set Theoretical Approach to Big Data Analytics*. 2010.
11. Leung Carson K, et al. *A Data Science Model for Big Data Analytics of Frequent Patterns*. 2016.
12. Gloria Bordogna. *Clustering Geo-Tagged Tweets for Advanced Big Data Analytics*. 2016.

Cite this Article

Jigali Sindhu B, Nirmala CR. Social Media Blog Analysis using Watson Analytics. *Journal of Artificial Intelligence Research & Advances*. 2016; 3(3): 15–21p.