

Big Data Analysis in Hotel Customer Response and Evaluation based on Cloud

Ming-Shen Jian, Yi-Chi Fang, Yu-Kai Wang, Chih Cheng

Department of Computer Science and Information Engineering, National Formosa University,

No.64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan

E-mail: jianms@nfu.edu.tw, ken801227@gmail.com, fengkui1216@gmail.com, macmax2011@gmail.com

Abstract—Traveling has always been a popular way of relaxation. With the ongoing stress cause by work and daily labours, the need of reducing stress via traveling has grown immensely, making trips a major part of human life. The core of vacation is to relax not only the body, but also the soul. However, there are so many information scattered around the internet, making it very hard to pin point the perfect hotel. This paper provides a solution based on cloud computing combined with big data. By gathering and sorting the information on the web, then analysing and ranking the results, they can be shown in a way that is easy to read and understand. The research started with the implementation of a cloud server and installing software, then gather information on the internet for processing, the next step is programing the code for cloud computation, and at last, design the web UI for data presentation. At first, the cloud server is equipped with Hadoop [13] for distributed computing. After installation, WebCrawler [15] is used to gather information and review from the internet, then using neural network and unsupervised learning to teach the server how to break down the reviews. Through analysing these results to understand the emotions behind the sentences, transforming it into data for Hadoop to do distributed computing. Lastly, using K-means [6] algorithm to sort all the data, and updating the database every day. This solution combines a lot of popular technologies, sorting the cluttered reviews of hotels on the internet, making them easier to read and understand. Allowing travelers to easily pick out the perfect hotel for their needs.

Keywords— WebCrawler, Neural-Network, K-means, Lexical Emotional Evaluation, Unsupervised Learning.

I. INTRODUCTION

Nowadays, due to innovations in the industry, “cloud” becomes a popular topic among many companies. Cloud platform can be used to establish the public cloud, private cloud, hybrid cloud, or social cloud. In addition, service models like infrastructure as a service (IaaS), platform as a service(PaaS), and software as a service (SaaS) are all examples of application of clouds.

Most data today are stored on public clouds like Google drive or Dropbox. However, since users are using services provided by a company, the privacy and data protection is a major issue.

In addition, based on the progression of cloud, big data becomes a major topic. In many categories, the data for further processing is too huge for a single computer to calculate.

Therefore, applying cloud computing and clustering multiple computers to enable parallel process has become the solution.

The rest of this paper is organized as follows. Section 2 presents the literature review. Section 3 shows the proposed system structure of solution for review collection and analysis. Implementation is explained in section 4. The conclusion is given in section 5.

II. LITERATURE REVIEW

A. Hadoop

Hadoop [13] is a cloud computation platform used to process and store massive amount of data [10]. it is part of Apache’s Top-Level Project. Hadoop includes Hadoop Distributed File System (HDFS), MapReduce structure, and HBase.

HDFS [12] will organize scattered data into an environment with high fault tolerance, great efficiency, and massive capacity.

MapReduce [10] is divided into 2 parts, Map and Reduce. Map turns the inputs into many parts, distributing them to multiple nodes for computing. Reduce then collects the parts and combining them back together for output.

HBase [8] is a NoSQL database that runs under HDFS, it has high availability, efficiency, and easy expansion of capacity. Figure 1 shows the structure of Hadoop.

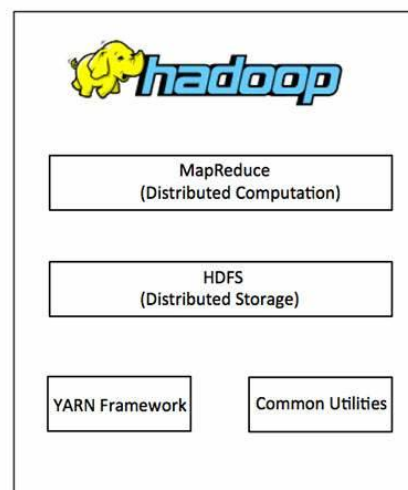


Figure 1. Structure of Hadoop [7].

B. WebCrawler

WebCrawler [15] is a unified search engine. It combines many different search engines and presents a single search interface. It can tap into multiple databases for results. The WebCrawler can be based on Python and allows more variant results.

WebCrawler, first formed from Google's search engine, is low-threshold and old technology. It uses HTTP protocol's search command and regular expressions [7]. Then, a site's useful data can be picked out. The flow chart is shown as Figure 2.

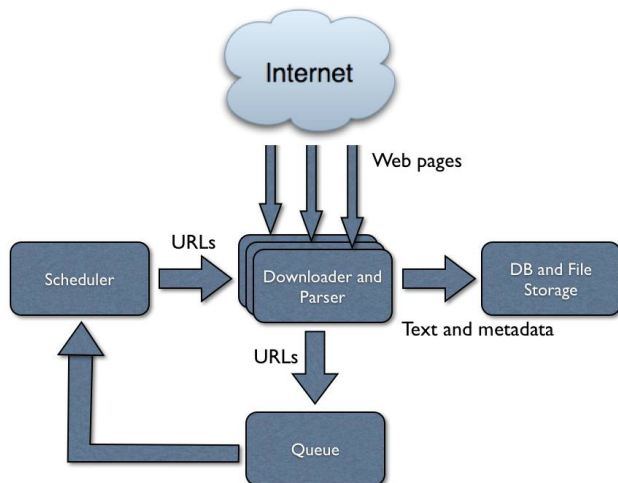


Figure 2. WebCrawler flow chart [14].

C. Python

Python is an interpreted and object oriented programming language. It can be executed under most operating system [5].

All functions, modules, numbers, strings are objects in Python. It means that these function, etc., can be inherited, overloaded, and polymorphed. It also enhances the usability of the original code. Compared to function-based languages such as Lisp, Python has the limited support towards function design. With only two standard libraries, functools and itertools, having function-like tools.

Even though Python can be considered as a script language, a lot of big scale projects like Google are still using Python. Python's followers often call it a high-level dynamic programming language, since most script languages such as shell script or VBScript are only used for simple tasks. Python's structure is shown in Figure 3.

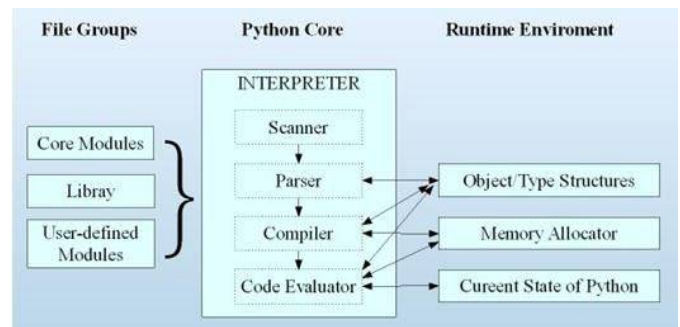


Figure 3. Structure of Python [3].

D. Neural Network

A neural network is a calculation model based on how a biological-neural network would process and calculate data. It has the ability to do parallel processes, a high tolerance to faults, and Associative Memory. Recurrent and Feed forward are the two main structure of neural network.

A recurrent structure's process is as follows:

- 1) During the learning stage, it will detect the part of the question that needs to be memorized.
- 2) Decide the weight of it through a Hebbian Learning machine.
- 3) At the recall stage, it will find if the presented question has something in common with the data in its memory.

A feed forward structure [14] works like this:

- 1) During the learning stage, users will provide a question and an answer.
- 2) By adjusting weight using steepest-descent method, the system will attempt to output the right answer.
- 3) At the recall stage, an untrained question will be proposed. By sorting and predicting the algorithm needed, it will output the most possible answer.

Its structure is as shown in Figure 4.

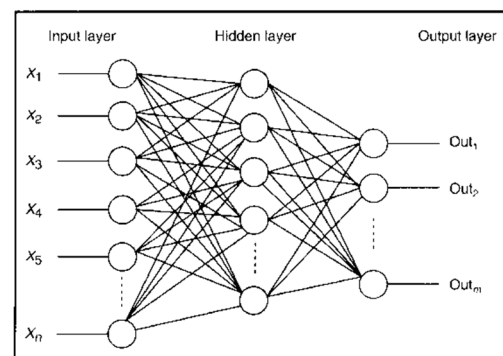


Figure 4. Structure of a Neural Network [5].

E. K-means and Lexical Emotional Evaluation

K-means is a data mining solution based on vector quantization and Cluster analysis.

Its equation is as follows:

$$J = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

K-means will first randomly pick k number of nodes as the cluster centers ($u_1 \sim u_k$). Then, the algorithm adds up the shortest distance between each center to node. By using this result, the algorithm continuously re-group each node to a group. The flow chart is shown as Figure 5.

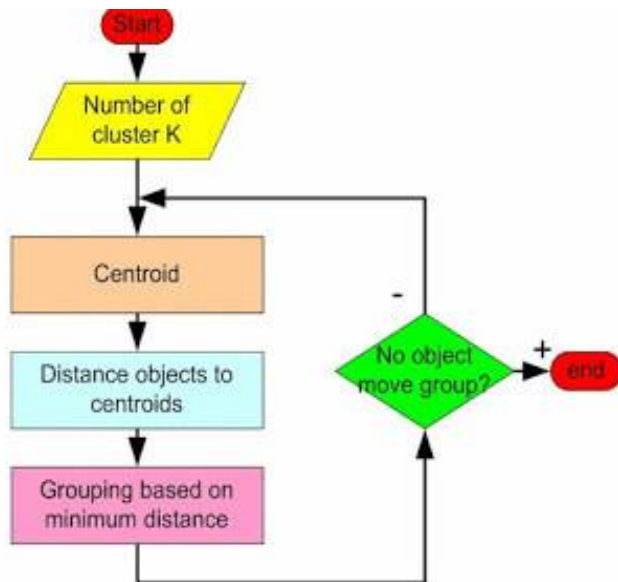


Figure 5. Flow chart of K-means [3].

Because of the continued growth at the data processing field, some researchers tried to add lexical emotional evaluation to improve on the system's accuracy. Based on the emotion presented behind the words, the collected data can evaluate if the sentence has a positive or a negative tone to it.

III. PROPOSED HOTEL EVALUATION ANALYSIS SYSTEM

The process starts with implementing Hadoop environment and Hadoop Multi Node Cluster on the cloud server for it to update its database actively. After implementing, a WebCrawler written with Python will gather reviews on the internet and saving it to the server for further analysis.

The data will first run through jieba text segmentation [2] to allow the system to understand the text and find key adjectives in a review. Then these adjectives will run through Hadoop's WordCount program for counting the times each word is used. Next, using a Feed-Forward structure neural network [1], the system will be provided with some test inputs and their expected solution based on the data, the network will learn to give each adjective a score between +25 and -25 with a positive number being an adjective with a positive meaning and negative number with a negative adjective. After every test, the system will do small adjustments to the weight of each word and the adjustment itself. Things like how much to adjust and

the sensitivity of adjustments are all being changed with each test. After multiple processes, the final answer will start to be much more similar to the expected output. With the system understanding if the word has either a positive or negative tone and how strong the emotion is based on the score. Forming a basic table of the weight of each adjective.

Next, using K-means, the system will group different adjective into clusters. With every word in the cluster having the same weight. By setting the number of clusters(k). To

minimize the result of $J = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2$, meaning sum of the distance between every node(x_j) and the centers (μ_i) must be minimized. The goal is to find the cluster with the best center and node, and as the result, optimize data classification. For this system, this means to optimize the weight of each adjective. Finally calculating a score based on the average score of the hotel and the number of reviews. Lastly, the results will be shown in a website for users to view. The flow chart and structure of the system is shown in Figure 6 and Figure 7.

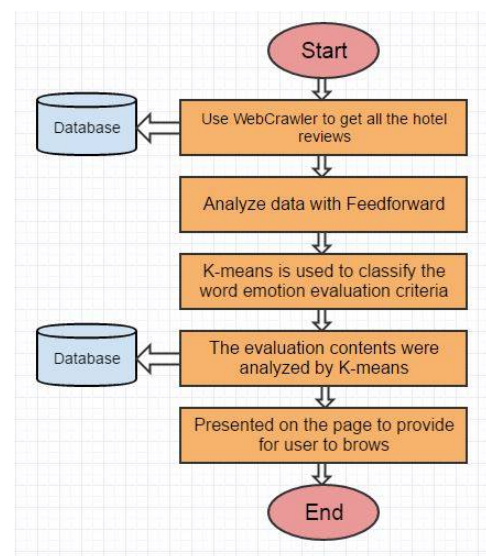


Figure 6. Flow chart of the proposed system.

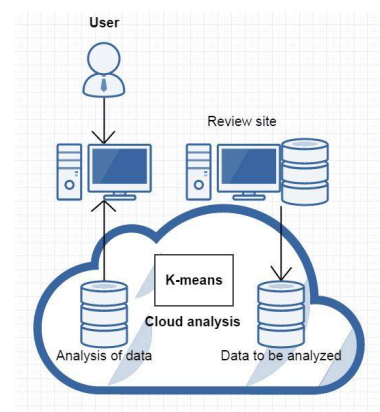


Figure 7. Structure of the proposed system.

IV. IMPLEMENTATION

Stating with WebCrawler, the system will gather reviews on the internet and transforming them into data for Hadoop to process as shown in Figure 8.



Figure 8. Transforming reviews into data using WebCrawler.

Using Hadoop MapReduce in conjunction with jieba text segmentation, the paragraph will be cut into individual vocabularies and counted as shown in Figure 9.

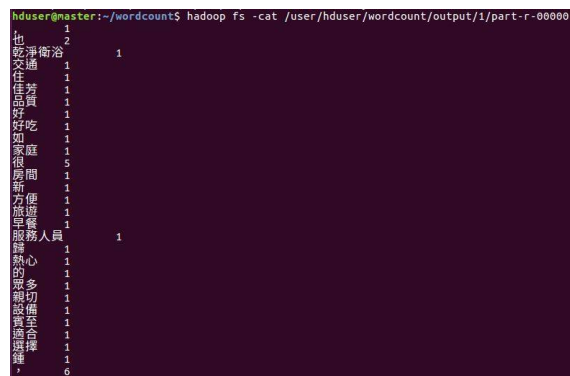


Figure 9. Result of WordCount after jieba text segmentation.

These vocabularies will next run through a neural network for weighting. Words like “Awesome”, “Nice”, and “Great” will be given a positive score while vocabularies like “Awful”, “Bad”, and “Expensive” will be assign with a negative score, K-means will group similar words together for analysis as shown in Figure 10. Figure 11 shows some example scores generated by reviews based on these vocabularies.

25	超級好	推薦
20	極好	值得
15	非常好	讚
10	很好	不錯
5	好	舒服
-5	差	舊
-10	很差	不便宜
-15	非常差	氣死
-20	極差	地雷
-25	超級差	糟透了

Figure 10. Weight given to vocabularies then grouped with K-means.



Figure 11. Example of scores based on reviews.

Lastly, these hotels will be ranked with others in the same area. Figure 12 is the result of the proposed system.



Figure 12. Result of the proposed system using Kaohsiung's hotels.

V. CONCLUSIONS

This paper proposed a system to evaluate a hotel based on reviews from multiple websites. By analysing and processing the text, the system can make the reviews into a graph that is easy to read and understand. By intelligent learning, the system allows the user to understand how well the hotel effortlessly. With some tweaking, the system can be customized to the user's needs.

REFERENCES

[1] Yan-Qing Weng, “网页抓取策略研究”, Harbin Engineering University, Harbin, China, 2010.

[2] Wen-Jing Chen, “Evolutionary Training Methods for the Generalized Feedforward Neural Networks”, National Chung Hsing University, Taichung, Taipei, 2011.

[3] Huu Tinh Giang Nguyen, “Design and Implement a MapReduce Framework for Converting Standalone Software Packages to Hadoop-based Distributed Environments”, National Cheng Kung University, Tainan, Taiwan, 2013.

[4] Zhi-Kai Liao, “Fault-Tolerant Management Framework for Hadoop Distributed File System”, Tamkang University, Xinbei, Taiwan, 2013.

[5] Darshan Bhat ,Simple Speech Recognition System Using LPC, Available Online: <http://amateurselectronics.blogspot.tw/2013/07/simple-speech-recognition-system-using.html>

[6] Yu-Chen Wang, “The Structure of the Big Data Based on the Hadoop Platform”, National Chi Nan University, Nantou, Taiwan, 2015.

[7] Jia-Chun Lin, “Study of Job Execution Performance, Reliability, Energy Consumption, and Fault Tolerance in the MapReduce Framework”, National Chiao Tung University, Hsinchu, Taiwan, 2015.

[8] Man-Ning Wu, “Data Analysis Results Storge Design in HBase, Chung Yuan Christian University”, Taoyuan, Taiwan, 2015.

- [9] Ze-An Ye, "Applying Multi-Agent Systems in Constructing Social Network Crawler for Intelligence Gathering: A Case Study of Construction of a Facebook Crawler", National Defense University, Taoyuan, Taiwan, 2015.
- [10] Bin Lin, "Experimental comparison of K-Means text clustering by varied distance calculation methods", Fuzhou Software Technology Vocation College, Fuzhou, Fujian, 2016.
- [11] Jie-An Luo, "Which programming language should students learn first? A comparison of Java and Python", National Taiwan Normal University, Taipei, Taiwan, 2016.
- [12] Flyaway, "Python 剖析 0",
Available online: <http://zhouyichu.com/Python/analysis-python-0/>
- [13] Zlantan Aki Mur, Numerical Example of K-Means Clustering,
Available Online:
<http://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm>
- [14] Bo-Yu Lin, "Text Mining Platform with Python, Fu Jen Catholic University", Xinbei, Taiwan, 2016.
- [15] Xiang-Wei Liu, "Emotion Recognition based on Principal Component Analysis and Neural Network", National Kaohsiung Marine University, Kaohsiung, Taiwan, 2016.



Ming-Shen Jian is an assistant professor at Dept. of Computer Science and Information Engineering in National Formosa University. Ming-Shen Jian's current research interests are in the area related to IOT application, Big Data, optimal solution, Intelligent System, and cloud computing. He received B.S. degree at Dept. of Electrical and Control Engineering in National Chiao Tung University and Ph.D. at Dept. of Computer Science and Engineering of the National Sun Yet-Sen University in Taiwan, 2007, investigating resource management in 3G mobile communication systems.



Yi-Chi Fang is a master degree student at Dept. of Computer Science and Information Engineering in National Formosa University. His current research interests are in the area related to IOT application, Big Data, optimal solution, Intelligent System, and cloud computing. He received B.S. degree at Dept. of Computer Science and Information Engineering in National Formosa University, 2014.



Yu-Kai Wang is an undergraduate student at Dept. of Computer Science and Information Engineering in National Formosa University. His current research interests are in the area related to Big Data, Intelligent System, and cloud computing.



Chih Cheng is an undergraduate student at Dept. of Computer Science and Information Engineering in National Formosa University. His current research interests are in the area related to Intelligent System and cloud computing.