Information Retrieval and De-duplication for Tourism Recommender System

Rajesh Thasal[1], Shubhada Yelkar[2], Amit Tare[3], Sharmila Gaikwad[4]

[1, 2, 3] *Student VIII Sem, BE, Computer Engineering, RGIT, Mumbai, India*
[4] *Professor, Computer Engineering, RGIT, Mumbai, India.*

---***---

Abstract - Nowadays many people rely on online services to plan a trip. However, they are usually faced with the problem of being supply with lots of information. In consequence, they have to invest great deal of time to decide what to visit on the basis of their preferences. This paper presents a working of Tourism Recommendation System which helps tourist to find best location suitable for their tourist profile. Recommender system is implemented using hadoop technology. For providing best object with great accuracy system goes through four important phases i.e. scrapping, mapping, de-duplication and recommendation. First the data object is scrapped from different website using jsoup library, after that datasets are mapped using Pig Latin script of hadoop. Mapping maps only required attributes from datasets. In data de-duplication process duplicate copies of same object will be resolved and only unique objects will be saved in datasets. In last phase (Recommendation) the best suitable location is provided with the help of tourist profile which is created by tourist. This tourist profile consists of name, duration of visit, location rating, point of interest, place rating etc.

***Key Word*: Hadoop, Jsoup, Pig-Latin, De-duplication, Recommendation.**

1. INTRODUCTION

Information Retrieval and De-duplication for Tourism Recommender System is the system which recommends the tourist interesting objects on the basis of tourist profile. Tourist profile consists of duration of visit, point of interest, property type and their preferences about different types of objects and events. On the basis of tourist profile, recommendation system is responsible for searching suitable information from database and returns the best objects for the end user. The information present in the database is extracted through scrapper. System uses the Jsoup java library package for extracting necessary information from the web pages. While extracting data sometimes system gets duplicate information which reduces the quality of information. To overcome this problem system uses De-duplication technique. De-duplication is a specialized data compression technique for removing duplicate copies of repeating data. De-duplication improves system recommendation and object information quality of system.

2. OBJECTIVES

The main purpose of Tourism Recommender System is to create a personalization tool that recommends tourist a list of information items that best the individual tastes. A recommender system infers the user preferences by analyzing the available user data, information about hotels and information about the point of interest. The accuracy of the recommendations highly depends on the amount of available information. System gathers information from different web pages and represents only unique information by applying filtering and De-duplication process. De-duplication of data is the important motive of this system.
De-duplication improves system recommendations and object information quality.

3. PROPOSED SYSTEM

Proposed system is designed to overcome the problem of data duplication. Proposed system starts with extracting the data from different websites. While extracting data sometime system get duplicate data which reduces the quality of our system. So to overcome this problem system used de-duplication technique. Before de-duplication mapping is done on datasets and after that filtration process is applied. Proposed system gives recommendation based on user's choice of preferences. Users preferences are compare with the de-duplicate datasets, if correct match is found then recommendation of location with additional information, hotel description and percentage of accuracy are provided to the users.

The proposed system goes through the four phases:

1. Scrapping
2. Mapping
3. De-duplication
4. Recommendation

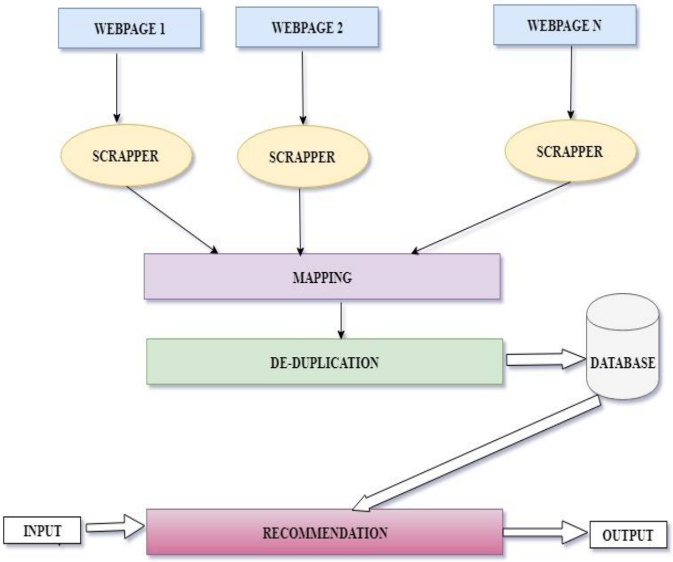


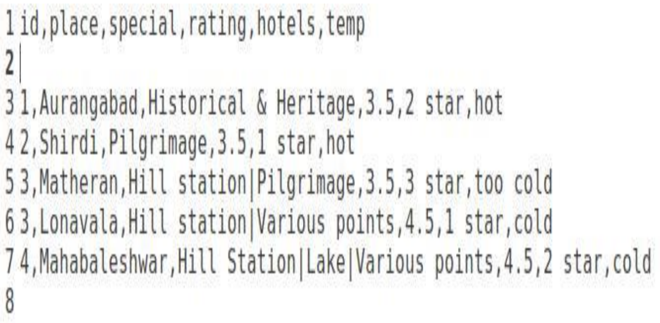
Fig. 3. System Architecture

4. FUNCTIONS OF PHASES:

4.1 Scrapping

Scrapping is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a database. Data is scrapped using jsoup parser. Jsoup is a Java based library to work with HTML based content. Jsoup used to parse HTML document. Jsoup provides API to extract and manipulate data from URL or HTML file.

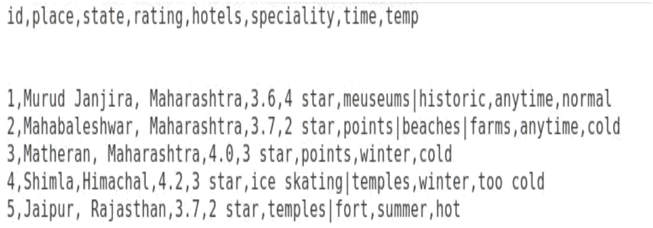
The data is scrapped from the two websites that are MTDC and MAKEMYTRIP. The fig 4.1 and 4.1.2 are the sample datasets of MTDC and MAKEMYTRIP websites.


```
id,place,special,rating,hotels,temp

1,Aurangabad,Historical & Heritage,3.5,2 star,hot
2,Shirdi,Pilgrimage,3.5,1 star,hot
3,Matheran,Hill station|Pilgrimage,3.5,3 star,too cold
3,Lonavala,Hill station|Various points,4.5,1 star,cold
4,Mahabaleshwar,Hill Station|Lake|Various points,4.5,2 star,cold

```

Fig. 4.1.1 sample dataset of MTDC


```
id,place,state,rating,hotels,speciality,time,temp


1,Murud Janjira, Maharashtra,3.6,4 star,meuseums|historic,anytime,normal
2,Mahabaleshwar, Maharashtra,3.7,2 star,points|beaches|farms,anytime,cold
3,Matheran, Maharashtra,4.0,3 star,points,winter,cold
4,Shimla,Himachal,4.2,3 star,ice skating|temples,winter,too cold
5,Jaipur, Rajasthan,3.7,2 star,temples|fort,summer,hot
```

Fig.4.1.2 Sample Dataset of MAKEMYTRIP

4.2 Mapping

The data objects which are scrapped from the different websites need to be mapped according to the attributes. This is done by mapping the attributes of the objects with the help of Pig Latin in Hadoop. In mapping only required fields are taken for further processing such as filtering. Pig uses simple sql-like scripting language called Pig Latin. Pig Latin script is worked as follows:

LOAD: A load statement reads data from the file system.

TRANSFORM: A series of "transformation" statements (FILTER, FOREACH, GENERATE) to process the data.

STORE: A store statement writes output to the file system.

DUMP: a dump statement displays output to the screen

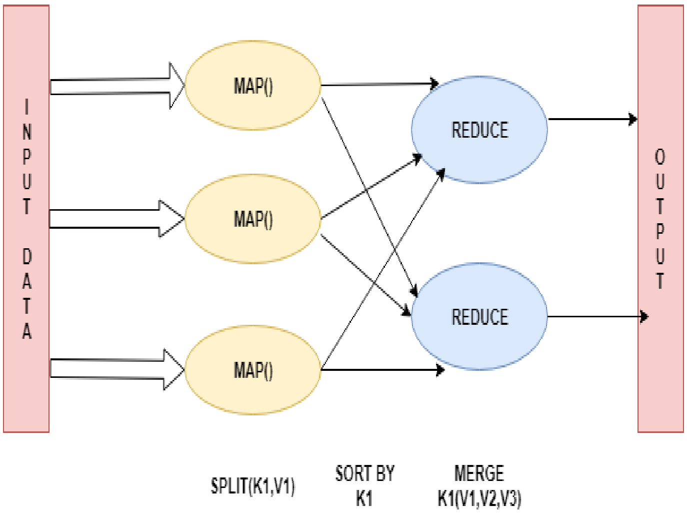


Fig . 4.2.1 Mapping

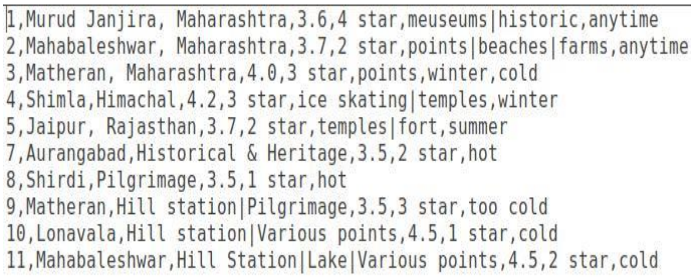

```
1,Murud Janjira, Maharashtra,3.6,4 star,meuseums|historic,anytime
2,Mahabaleshwar, Maharashtra,3.7,2 star,points|beaches|farms,anytime
3,Matheran, Maharashtra,4.0,3 star,points,winter,cold
4,Shimla,Himachal,4.2,3 star,ice skating|temples,winter
5,Jaipur, Rajasthan,3.7,2 star,temples|fort,summer
7,Aurangabad,Historical & Heritage,3.5,2 star,hot
8,Shirdi,Pilgrimage,3.5,1 star,hot
9,Matheran,Hill station|Pilgrimage,3.5,3 star,too cold
10,Lonavala,Hill station|Various points,4.5,1 star,cold
11,Mahabaleshwar,Hill Station|Lake|Various points,4.5,2 star,cold
```

Fig 4.2.2 Mapped and Merged dataset

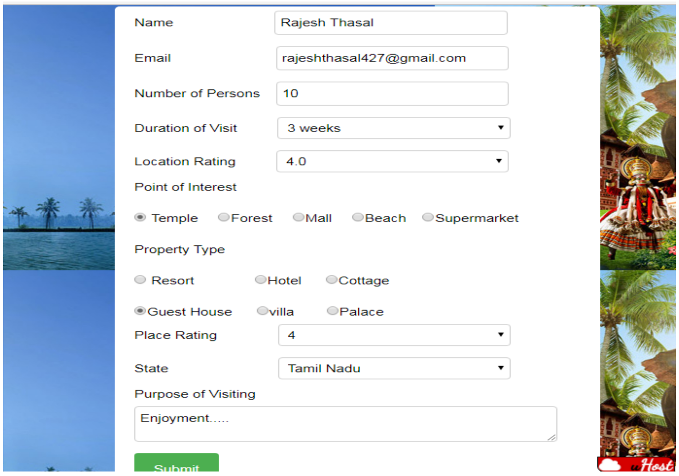
4.3 De-duplication

De-duplication is a specialized data compression technique for removing duplicate copies of repeating data. In this technique system search for duplicate copies of data if duplicate is found during analysis then only single copy of that attribute is stored in dataset and other duplicates copies are removed using Pig Latin script.

This technique is used to improve storage utilization and speed of execution.

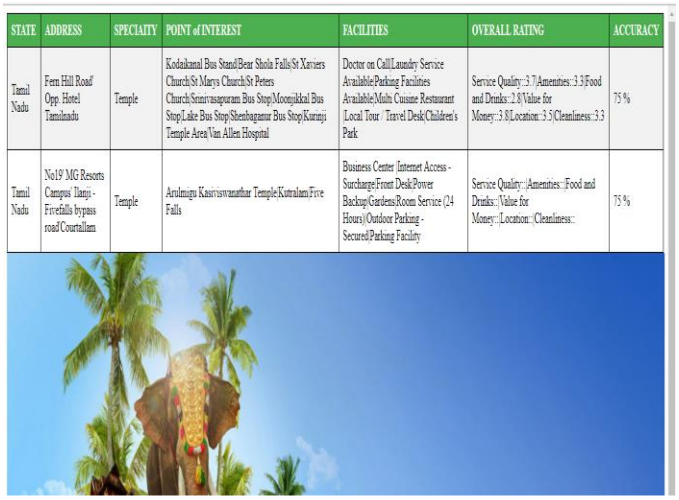
4.4 Recommendation

In recommendation first tourist will fill up the tourist profile form which includes point of interest, property type, duration of visit, location rating and place rating. After that this choice of preferences are compared with de-duplicated dataset. If perfect or nearby match is found then that location will be given as recommendation with percentage of accuracy.



Name Rajesh Thasal
Email rajeshthasal427@gmail.com
Number of Persons 10
Duration of Visit 3 weeks
Location Rating 4.0
Point of Interest
● Temple ○ Forest ○ Mall ○ Beach ○ Supermarket
Property Type
○ Resort ○ Hotel ○ Cottage
● Guest House ○ villa ○ Palace
Place Rating 4
State Tamil Nadu
Purpose of Visiting
Enjoyment.....

Fig. 4.4.1 Tourist Profile



STATE	ADDRESS	SPECIALTY	POINT of INTEREST	FACILITIES	OVERALL RATING	ACCURACY
Tamil Nadu	Fern Hill Road' Opp. Hotel Tamilnadu	Temple	Kodaikanal Bus Stand|Bear Shola Falls|St Xaviers Church|St Marys Church|St Peters Church|Srinivasapuram Bus Stop|Moonjikkal Bus Stop|Lake Bus Stop|Shenbaganur Bus Stop|Kurinji Temple Area|Van Allen Hospital	Doctor on Call|Laundry Service Available|Parking Facilities Available|Multi Cuisine Restaurant |Local Tour / Travel Desk|Children's Park	Service Quality::3.7|Amenities::3.3|Food and Drinks::2.8|Value for Money::3.8|Location::3.5|Cleanliness::3.3	75 %
Tamil Nadu	No19' MG Resorts Campus' Ilanji - Fivefalls bypass road'Courtallam	Temple	Arulmigu Kasiviswanathar Temple|Kutralam|Five Falls	Business Center |Internet Access - Surcharge|Front Desk|Power Backup|Gardens|Room Service (24 Hours)|Outdoor Parking - Secured|Parking Facility	Service Quality::|Amenities::|Food and Drinks::|Value for Money::|Location::|Cleanliness::	75 %

Fig. 4.4.2 Recommendation

5. TECHNOLOGY USED FOR IMPLEMENTATION

5.1 HADOOP

Hadoop is an open-source framework that allows users to store and process large amount of Data in a distributed environment across clusters of computer using programming models. It is designed to scale up from single servers to several of machines with high degree of fault tolerance. Data in a Hadoop cluster is broken down into smaller chunks and distributed throughout the cluster like the Map and Reduce functions that are executed on smaller chunks of larger data sets, and this provides the scalability required for Big Data processing.

5.2 APACHE PIG

Apache pig is a project which is lies on top of Hadoop, and it provides scripting language to use Hadoop's Map-Reduce functionality, Pig Latin script is used to perform read, filter, transform, join, dump and write data with less programming skills. Pig allows data workers to write complex data transformations without the knowledge of java programming language.

Pig is made up of two main components

a) Pig Latin
b) Runtime environment

A. PIG LATIN

Pig uses simple sql-like scripting language called Pig Latin. Pig Latin is relatively simple language that executes a set of commands. Pig latin statements works with relations (bag with collection of tuple), a Pig relation is similar to a table in a relational database, where the tuples in the bag correspond to the records in a table. Pig Latin statements can be in multiple lines and it must end with a semi-colon.

Pig Latin helps non-java developers as it takes less time to code, for example in a test 5 lines of Pig Latin ≈ 100 lines of java. This takes 3 hours to write in java but 10 minutes in Pig Latin.

Table 1. Pig Commands

FUNCTION	PIG COMMANDS

Command to load data	LOAD()
	PigStorage()
	BinStorage()
	TextLoader()

Command to work with data	Filter Foreach Join Group

	Cogroup Union Split
Command to debug the data	Describe Explain Illustrate
Command to retrieve results	Dump Store

B. RUN TIME ENVIRONMENT

The pig execution environment has two modes of execution they are,

1. **Local mode:** Pig scripts runs on a single machine were Hadoop Map-Reduce and HDFS are not required.

2. **Hadoop or Map-Reduce mode:** pig scripts runs on a Hadoop. Pig statements can be executed in three different ways, in which all the three are executed in both local and Hadoop modes.

1. **Grunt shell:** Allows users to type pig commands manually using pig's interactive shell, grunt.

2. **Script file:** Place pig commands in a script file and run the script file.

3. **Embedded:** Embed the pig statements in a host language and run the statement.

5.3 JSOUP

Jsoup is *a java html parser*. It is a java library that is used to parse HTML document. Jsoup provides API to scrap and change data from URL or HTML file. It uses DOM, CSS and Jquery-like methods for scrapping and manipulating file.

J

soup: parsing string

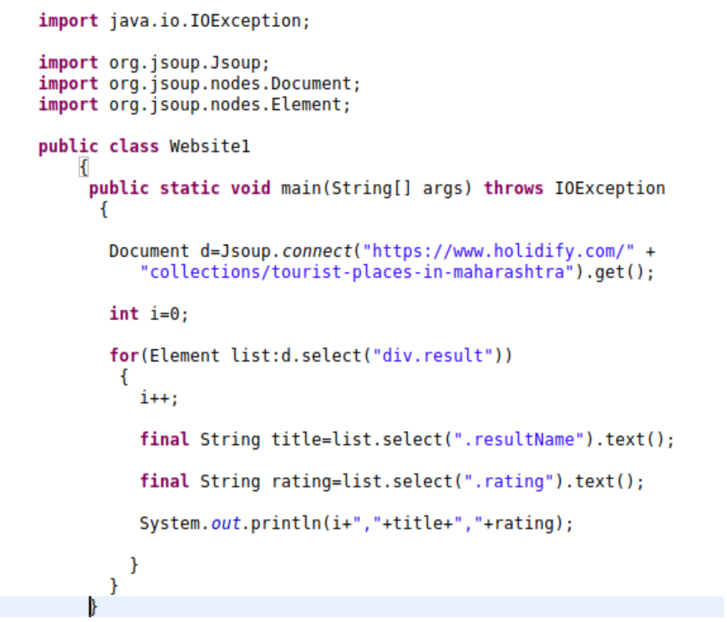
Syntax:

```
Document d = Jsoup.parse(html);
```

Where

- **Document -** document object represents the HTML DOM.
- **Jsoup -** main class to parse the given HTML String.
- **Html -** HTML String.

Parsing data from Holidiffy.com



```
import java.io.IOException;

import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;

public class Website1
    {
     public static void main(String[] args) throws IOException
      {

        Document d=Jsoup.connect("https://www.holidify.com/" +
           "collections/tourist-places-in-maharashtra").get();

        int i=0;

        for(Element list:d.select("div.result"))
         {
           i++;

           final String title=list.select(".resultName").text();

           final String rating=list.select(".rating").text();

           System.out.println(i+","+title+","+rating);

          }
        }
    }
```

6. FLOWCHART

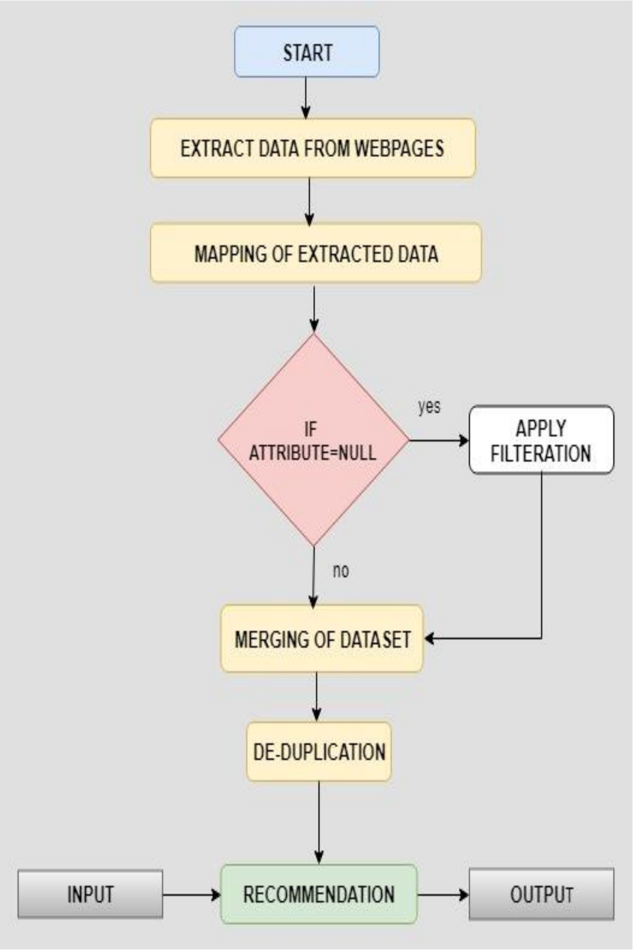


Fig. 6 Flowchart

7. CONCLUSION

This paper gives an overview of implementation of Tourism Recommender System which gives the personalize suggestion to the tourist for selecting best location which is suitable for their profile. This system is implemented using hadoop technology so that processing on huge amount of dataset becomes easy. For implementing mapping and de-duplication Pig Latin script is used. Pig Latin is an extension of map-reduce. Map-reduce functionality of hadoop is implemented using pig. With the help of pig programmers does not have to write lines of code, simply they type command in grunt shell to reduce programming complexity. With the help of pig, recommender system is simple and easy to understand. It also improves the execution speed of recommender system.

8. REFERENCES

[1] Bhosale, Harshawardhan S., and Devendra P. Gadekar. "A Review Paper on Big Data and Hadoop." International Journal of Scientific and Research Publications 4.10 (2014)

[2] Chavan, Ms Vibhavari, and Rajesh N. Phursule. "Survey paper on big data." Int. J. Comput. Sci. Inf. Technol 5.6 (2014): 7932-7939.

[3] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. Swoosh: a generic approach to entity resolution. The VLDB Journal, March 2008. VLDB Journal (Online First) link:
http://dx.doi.org/10.1007/s00778-008-0098-x.

[4] M. Bilenko. Learning to combine trained distance metrics for duplicate detection in databases. Submitted to CIKM-2002, (February):1–19, 2002.

[5] Adomavicius, G & Tuzhilin, A (2005), Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749.

[6] J. L. Herlocker, J. A. Konstan, L. G. Terveen, & J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems, vol. 22, no. 1, pp. 5-53, 2004.