Multi-Label Naïve Bayes Classifier for Identification of Top Destination and Issues to Accost by Tourism Sector

Sapna Sinha, Amity Institute of Information and Technology, Noida, India
Vishal Bhatnagar, Ambedkar Institute of Advanced Communication Technologies and Research, Delhi, India
Abhay Bansal, Amity School of Engineering and Technology, Noida, India

ABSTRACT
This article describes how the tourism sector plays very important role in the social and economic growth of any country. Technologies like the internet and mobile technologies have changed the marketing strategy of the sector. It has been observed that business operators and the workforce employed in the tourism industry do not have sufficient knowledge, tools and/or strategy to implement technology correctly. Technology can be used in all the dimensions of tourism for gaining a competitive edge and providing a wide range of services to customers. For better customer satisfaction, the industry should know the major issues confronting tourism. Experiences shared by tourists on social media, i.e. Facebook, Twitter, Instagram and YouTube, can be used and analyzed to gain insight on customer needs. In this article, the authors propose a unified framework and have used tweets shared by tourists for the identification of major issues faced by tourism sectors. Identified issues are categorized into four main categories. The obtained results will help players of tourism sector for improving the services and growth of the sector.

KEYWORDS
Big Data, Flume, HDFS, Mapreduce, Naïve Bayes Multi-Label Classifier, Twitter Analytics

1. INTRODUCTION
Tourism Sector across the world has started embracing technology for simplification of their process and for providing complete customer satisfaction. Big Data impact and applicability is not hidden any more to the world. Big data can be also used in the tourism sector for getting insight of data available in abundance. The tourism sector is also trying to use big data up to some extent, but failed to utilize it fully due to unavailability of any standard framework for big data processing. The data acquired is available in different silo’s, traditional data is stored in traditional database management system, Social media data is available in unstructured format on the web, which needs different treatment for gaining insight and performing analytics over it. Temporal data can also be incorporated in the routine analytics process. Analytics performed on traditional databases are by using SQL or any

DOI: 10.4018/JGIM.2018070104

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
query processing language, but for abundant data available on social media sites and portals, tweets or posts are first needed to be extracted followed by cleaning of data and feature extraction, structure is given to unstructured data for performing analytics.

Social media is a web technologies and mobile based media that allows individuals or organizations to create application based profiles, share and exchange contents. It includes blogs, media sharing, business networks, product/review sharing, virtual worlds, social gaming, social networks etc. If Facebook is considered as a country, it will be 3rd largest population after China and India (Safranek, R., 2012). Twitter itself observed growth of more than 300 million users. These stats give glimpse and reach of social media across the world. Peoples across the world are using social media and it has been observed that 35% to 45% of youth of any country in the world is using two social media platforms to express its view and share it with other users. The messages or post shared generally includes perception, images, emotions and geo-locations of an individual (Carr and Hayes, 2015). The data generated provides opportunities for researches to extract meaning full insight out of the contents shared online.

Twitter is a microblogging site which allows its users to share their views with limitation of 140 characters. It has been observed that retweeted tweets reach on an average 1000 users, regardless of number of followers hence it is the fastest mode of disseminating information (Kwak et.al., 2010). Traditional methods used to gain perception of users were surveys, feedbacks and interviews are found not suitable in present scenario due to its limitations. Alternatively, social media mining can be used to break the barriers imposed by traditional methods. Twitter has emerged as one of the prominent platform for researchers for data analysis in their field (Tumasjan et al., 2010). For example, tweets are used for finding epicenter of earthquakes and trajectory of typhoons. The information about earthquake were send to all registered users, which found to be faster than the broadcast by Japan Meteorological Agency (Sakaki et.al, 2010). A twitter tweet also helps to predict stock market (Bollen et al., 2011). Twitter tweets are used to analyze the outbreak of influenza and other communicable diseases and prediction of presidential election (Tumasjan et al., 2010).

Analyzing social data includes several challenges due to its nature. Tweets falls in the category of big data due to 5Vs like: Volume, Velocity, Value, Veracity and Variety (Zikopoulos and Eaton, 2011). In this paper, authors have used tweets for analysing various factors affecting tourism sector. Tweets are found suitable for gaining feedback of tourists for the places visited by them. Tweets are extracted based on twitter hashtags like: #travelgram, #vacation, #visiting, #instatravel, #instago, #trip, #holiday, #travelling, #tourism, #tourist, #instatraveling, #mytravelgram, #travelingram, #travelgoals, #travel, #traveling, #travelingproblems, #travelingstress. The issues identified is categorized into four main categories and based on these categories Multi-Label Naïve–Bayes classification algorithm is implemented in MapReduce programming and results were evaluated and compared.

The structure of the remaining sections is as follows. Section 2 included the proposed unified framework for tourism sector for big data processing, methodology and the different algorithms used in the proposed approach. Section 3 represented the results and discussion. Finally, the conclusion is included in section 4.

2. METHODOLOGY

In this paper, we have proposed unified framework for big data processing for tourism sector and methodology for exploiting social media data for finding the most visited place and the factors that is of concern by tourism sector. Figure 1 shows diagrammatical representation of proposed unified framework for big data processing for tourism sector.

Unified framework is required to integrate data from different sources to centralized data warehouse and analytics can be performed on it to gain insight from it. Analytics can be broadly categorized into three categories: Descriptive Analytics – for summarization of data and standard report generation, Predictive Analytics – for predicting future based on historical data. Prescriptive analytics
The Cultural Construction of Information Technology
Vanessa Dirksen (2001). *Journal of Global Information Management* (pp. 5-10).
www.igi-global.com/article/cultural-construction-information-technology/3548?camid=4v1a

The Satisfaction of Saudi Customers Toward Mobile Banking in Saudi Arabia and the United Kingdom
www.igi-global.com/article/the-satisfaction-of-saudi-customers-toward-mobile-banking-in-saudi-arabia-and-the-united-kingdom/190492?camid=4v1a