International Journal for Rapid Research in Engineering Technology & Applied Science
Volume V Issue II February 2019
ISSN (Online): 2455-4723

A Data Mining Technique for Tourist Destination Brand Image Building

Alok Aamle*, Prof. Mohit Jain**

Department of Computer Science & Engineering, BM College Indore, India,

alokraamle@gmail.com* bmctmohitcs@gmail.com**

Abstract: **The destination image branding is the domain of tourism industry where the facts and information is collected and evaluated for finding the credibility of a target tourist destination. Manual collection and processing of collected information accurately is a complicated and time consuming task therefore a data mining model is suggested in this presented work that collect and evaluate the destination image accurately and based on evaluation can make the recommendations about visits of tourist. In order to perform this task data mining techniques are applied on text data source. In first the data is extracted from the Google search engine and it is preprocessed for make it impure. In further the data is labeled based on the positive and negative words available in the collected facts. Finally the clustering and classification of text is performed. For clustering of data FCM (fuzzy c means) clustering algorithm and for classification the Bayesian classifier is used. Based on final classification of text data the decision is made for the destination visits. The implementation of the proposed technique is performed using JAVA technology and their performance is defined using positive and negative probability of prediction. In addition of that the time and space a requirement of the data evaluation is measured which is also acceptable for the proposed application.**

Keyword: Brand building, destination image building, data mining techniques, classification, clustering, FCM

I. INTRODUCTION

Data mining and their techniques provides us the ability to analyze the data automatically using the computational algorithms. Additionally grab the outcomes of analysis for decision making, classification, predictions or other essential task. In this presented work the data mining is performed on the unstructured data i.e. text documents. Therefore the proposed work is intended to demonstrate the technique of text mining. The text mining is the sub-domain of data mining that deal with the text data. Using the text mining approaches in this work the destination branding of the tourist places is obtained using text mining techniques. Basically when someone plans to visit some place as tourist he/she not know all the prospects of the particular place.

Therefore sometimes the visitor is trapped in various kinds of issues such as misleading place, inappropriate visiting conditions, risk of thief and others. Therefore the visitors collect the information from web to know basics of the particular place. But in most of online resources only the common or basic overview about the places are available. That information is not complete in terms of to make decision to visit the place strongly recommended or not. Therefore in this presented work using the different source of data analysis a new model is proposed that investigates about the places to visit. Additionally by analysis of the data it produces the strong recommendations to the users to visit the place or not. In this context different source of data such as news, blog and other source of data is investigated and analyzed using the data mining algorithms. The analysis of the data results the patterns of data and using the recovered patterns the suggestions are made to visit the place or not.

II. PROPOSED WORK

The sentiment analysis and supervised learning techniques help to investigate the user's mood and the reviews about the different products. In this presented work the review analysis of visiting places are performed using the data mining techniques for developing effective branding of any tourist place.

A. System Overview

In literature a number of applications and examples are exist where the data mining approaches are employed for reviewing the text for obtaining the review about product and services. In this presented work the sentiment analysis is applied on NEWS data and blog data which are collected using the Google search results to develop strong recommendation about the place to visit. This process is termed here as destination image branding. Basically when we search about any tourist destination from web the available source of information provides us the overview of the particular destination.

Additionally data is not updated frequently or regularly in available sources such as blog or other informative sites. But the place and their conditions are influenced by the different political event, natural events and other human made disasters (i.e. terrorism) and others. Therefore the image of particular place is also changing over time. In this context for obtaining the clear and strong review about the tourist destination the fresh data is required to process or analyze. In this presented work the Google search technique is implemented for finding fresh information from web data source. Additionally by using this fresh data the analysis is made to recommend the destination image. In addition of that a hybrid data model is proposed for performing the data analysis. The proposed technique is a combination of two data mining algorithms FCM (fuzzy c means) clustering and the Bayesian classifier. Using both the technique the data is processed and the essential features are recovered for making recommendations about the tourist destinations. In this section the need of proposed system is described. In next section the methodology of the proposed technique is explained.

B. Proposed Methodology

The proposed methodology of the system design to find recommendations about the tourist places are described in figure 2.1. In addition of the components of the given model is explained in same section.

User Destination: the data mining and machine learning system requires some initial inputs to process data and recovers the essential information. The proposed system also involve the concept of data mining techniques therefore to find the fresh data about the places a provision is developed for accepting the initial input. Using this input provision user can provide the place name (destination) for visit.

Google Search: the user input place name is produced to the Google search API for finding the information from web. Google search collect the information from internet source and generate a list of results.

Search results: the generated results from Google search is collected in a file for utilizing with the further process.

Data preprocessing: the data preprocessing is an essential step of data mining and knowledge discovery. The preprocessing of data is performed for cleaning the target data and makes it suitable for use. Therefore using the preprocessing here we remove the unwanted data from the generated results through the Google search. In this context the stop words and special characters from the text data is removed. To remove the stop words and special characters from collected data a list of words and characters are prepared. Additionally a function is developed that remove all the data from text which is available in the prepared stop word list and special character list.

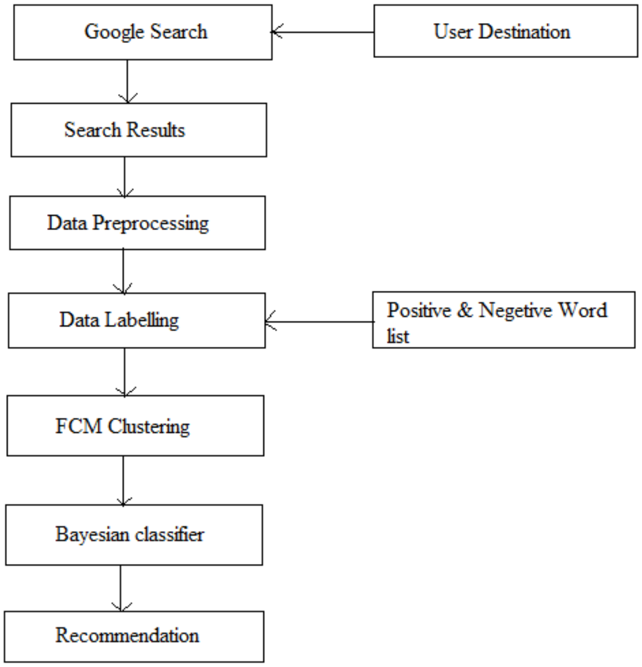

Figure 2.1 proposed system architecture

Positive & negative word list: a list of positive and negative words that are used in sentiment analysis or orientation mining is taken from web. This list of word contains both kinds of words negative as well as positive labels.

Data labeling: using the positive and negative word list initial labels of the collected data is prepared. In order to perform the labeling of data all sentences are evaluated using this word list and if the sentence contains multiple positive words then the sentence is labeled as positive otherwise the sentence is labeled as negative sentence.

FCM clustering: after successfully labeling of data it is used with the FCM (fuzzy c means) clustering algorithm. That clusters all the sentences more precisely. The FCM is an unsupervised learning technique that evaluates data using the fuzzy optimization function to make more clear clusters from the data. The process of data clustering is given as follows:

The fuzzy c means clustering works on the basis of the following objective function:

$$O_n = \sum_{i=1}^{n}\sum_{j=1}^{k} P_{ij}^{n}\left\|d_i - k_j\right\|^2$$

Where n is a real number and must be grater then 1

P_{ij}^{n} is the degree of membership

d_i is ith element of data object or instance of data (i.e sentences)

k_j is cluster centroid

For computing the partitions the iterative optimization process is called. Therefore it is required to compute the membership of data and new cluster centers.

$$P_{ij} = \frac{1}{\sum_{l=1}^{k}\left(\frac{\left\|d_i - k_j\right\|}{\left\|d_i - k_l\right\|}\right)^{\frac{2}{n-1}}}$$

$$k_j = \frac{\sum_{i=1}^{N} P_{ij}^{n} * d_i}{\sum_{i=1}^{N} P_{ij}^{n}}$$

The algorithm is terminated when the following condition reached.

$$\left\{\left|P_{ij}^{(k+1)} - P_{ij}^{k}\right|\right\} < \varepsilon$$

ε is the error or difference between previous membership value and current membership value.

After clustering the data is subdivided into two groups one group contains the positive sentences and other group contains the negative sentences.

Bayesian classifier: the categorized data using the FCM (fuzzy c means) clustering is used with the Bayesian classifier. The classifier basically cross validates the clustered data instances by again classification of all the data. That improves the quality of prediction which is performed after carry out the classification of Bayesian classifier. The Bayesian classifier works in the following manner.

Bays classifier computes the probability of an event occurrence, which can be defined using a formula as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A and B are the events

Basically we are trying to find the probability of A with respect to given event B here event B is termed as evidence.

P(A) is priori of A and the evidence is an attribute value of an unknown instance. $P(A|B)$ is posterior probability of B

Recommendation: after classification with bays classifier the data is divided into the negative and positive instances of data. if the positive score of the instances are higher then the negative instances the recommendation is made to visit the place otherwise the system is not suggesting the place to visit.

C. Proposed Algorithm

The proposed methodology of the destination image branding is explained in previous section. In this section the methodology is transformed into the process steps using table 2.1 as algorithm.

Input: user destination D, Negative and positive word list W_{list}

Output : recommendation R

```
Process:
    1. $S_n = GoogleAPI.Searc□(D)$
    2. $for(i = 1; i \leq n; i++)$
        a. $P_i = RemoveStopWord(S_i)$
        b. $P_i = RemoveC□aracter(P_i)$
    3. $end\ for$
    4. $L_{data} = FindClasses(P_n, W_{list})$
    5. $[centroid, index] = FCM.CreateCluster(L_{data}, 2)$
    6. $C_{data}[Negative, Positive] = Bays.Classify(L_{data}, index)$
    7. $if(Positive > Negetive)$
        a. $R = visit\ Place$
    8. Else
        a. $R = Not\ visit\ Place$
    9. End if
```

```
10. Return R
```

Table 2.1 Proposed algorithmc

III. RESULT ANALYSIS

The given section includes the performance analysis of the implemented algorithms for the proposed web content mining approach of destination branding and image. Therefore some essential performance parameters are obtained and listed with their obtained observations.

A. Positive Probability

Positive probability is the probability of the words which are belonging to outcome of the positive reviews. Positive probability can be calculated by using following formula:

$$\text{Positive Probability} = \frac{\text{Number of Positive Words}}{\text{Total Number of Words}}$$

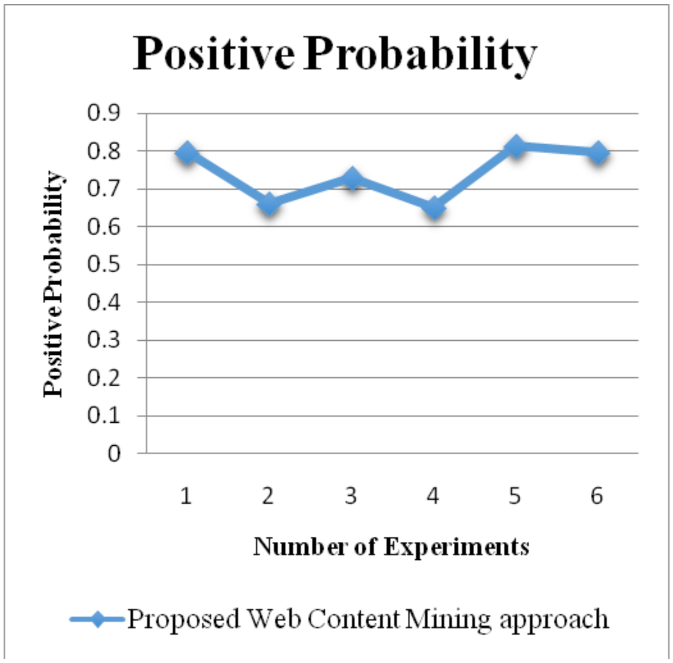

Figure 3.1 Positive Probability

The positive probability of the proposed algorithm of destination branding image is represented using table 3.1 and figure 3.1. The given graph figure 3.1 contains the positive probability of the implemented algorithms. The X axis of the diagram shows the different experiments and Y axis contains the obtained performance. To demonstrate the performance of the proposed technique is representing using blue line. According to the obtained results the performance of the proposed model provides more accurate results. Additionally the positive probability of the proposed model is varying as the number of experiments increase.

Table 3.1 Positive Probability

Number of Experiments	Proposed Web Content Mining Approach
1	0.7954
2	0.6595
3	0.7272
4	0.65
5	0.8125
6	0.7955

B. Negative Probability

The negative probability which is shows the negative reviews of the outcome of the classification process. The negative probability can be estimated by using following formula.

$$\text{Positive Probability} = \frac{\text{Number of Negative Words}}{\text{Total Number of Words}}$$

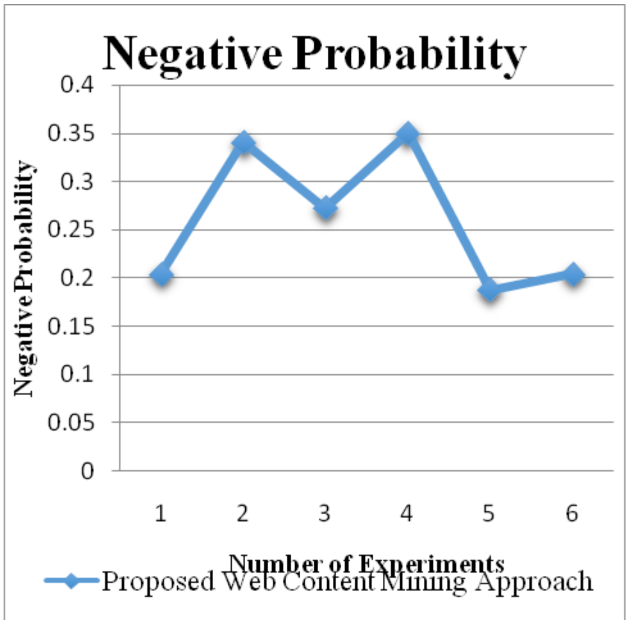

Figure 3.2 Negative Probability

The figure 3.2 and table 3.2 shows the error rate of implemented proposed approach. In order to show the performance of the system the X axis contains the experiments and the Y axis shows the performance in terms of negative probability. The performance of the proposed web content mining technique is given using the blue line. The performance of the proposed word classification is effective and efficient during different execution and reducing with the amount of data increases. Thus the presented classifier is more efficient and accurate than the other text classifier of classification

Table 3.2 Negative Probability

Number of Experiments	Proposed Web Content Mining Approach
1	0.2045
2	0.3404
3	0.2727
4	0.35
5	0.1875
6	0.2045

C. Memory Usage

Memory consumption of the system also termed as the space complexity in terms of algorithm performance. That can be calculated using the following formula:

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

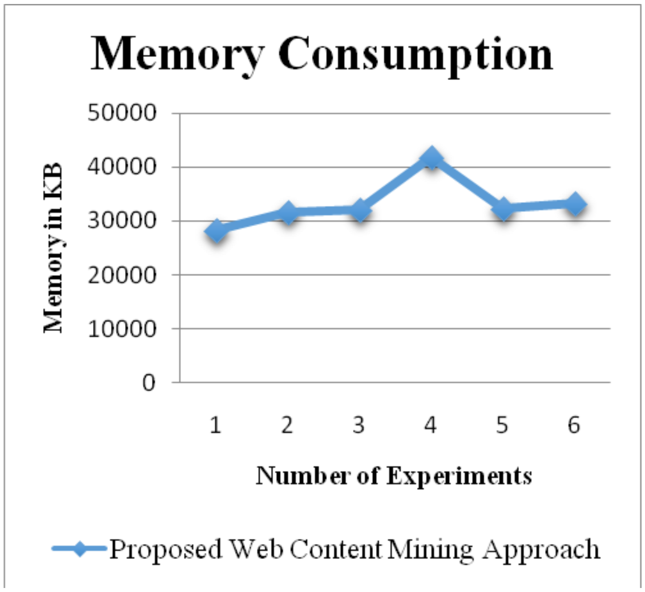

Figure 3.3 Memory Consumption

The amount of memory consumption depends on the amount of data reside in the main memory, therefore that affect the computational cost of an algorithm execution. The performance of the implemented web content mining approach of destination branding is given using figure 3.3 and data is numerically show by table 3.3. For clarification of the result, X axis of figure contains the different amount of code execution and the Y axis shows the respective memory consumption during execution in terms of kilobytes (KB). According to the obtained results the performance of algorithm demonstrates similar behavior with increasing size of data, but the amount of memory consumption is decreases with the amount of data. This consumed memory represents the required space by which algorithm of word classification of dictionary is executed and produces efficient output.

Table 3.3 Memory Consumption

Number of Experiments	Proposed Web Content Mining Approach
1	28187
2	31611
3	31955
4	41764

5	32124
6	33151

D. Time Consumption

The amount of time required to classify the negative and positive words of dataset is known as the time consumption of the system. That can be computed using the following formula:

$$\text{Time Consumed} = \text{End Time} - \text{Start Time}$$

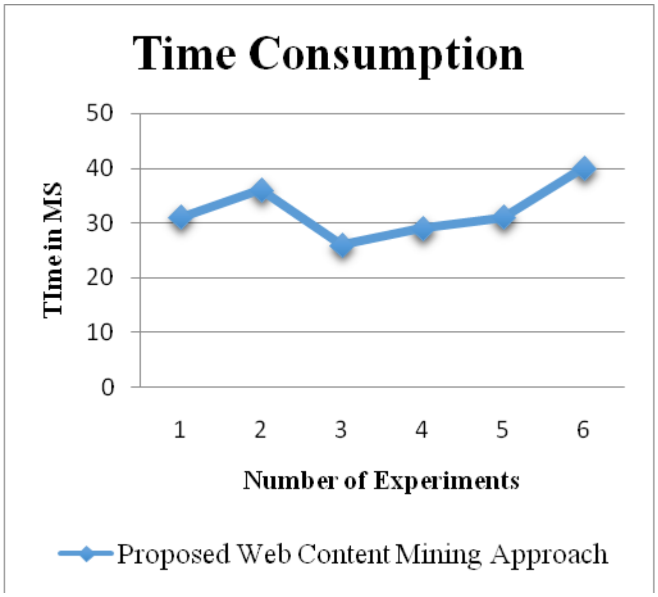

Figure 3.4 Time Consumption

The time consumption of the proposed algorithm is given using figure 3.4 and table 3.4. In this diagram the X axis contains the program execution of the system and the Y axis contains time consumed which is measures in milliseconds. According to the evaluated performance of the proposed technique is process the word to classify their nature. For processing algorithm consume time which is illustrated in table 3.4 in numerically. But the amount of time is increases in similar manner as the amount of data for analysis is increases.

Table 3.4 Time Consumption

Number of Experiments	Proposed Web Content Mining Approach
1	31
2	36
3	26
4	29
5	31
6	40

IV. CONCLUSION & FUTURE WORK

The proposed work is intended to design and develop a data mining based technique that helps to suggest the credibility of the tourist destination. Therefore a model is developed using the data mining algorithm. This chapter provides the summary of conducted efforts for developing the required data model.

A. Conclusion

Destination branding is a subject of tourism industry. In this subject the different source of information is collected for finding the facts about the particular tourist destination. Additionally a brand image is developed on the basis of collected facts about the place goodness. In this context various manual efforts are observed in literature. But the data analysis in manual mode is complicated and time consuming task. Therefore the proposed work include a data mining technique that collect and evaluate the target tourist place for computing the recommendations about the tourist destination brand image. The proposed technique is a data mining model that first employs the Google search engine API for collecting fresh information from web. In next using the positive and negative word list the labeling of data is performed. In final phases two different data mining algorithms are applied for mining and exploring the information and facts. The data mining methods includes the FCM and Bayesian classifier. First using the FCM (fuzzy c means) clustering the data is categorized in two clusters and finally it is classified for finding accurate information about the target place.

The implementation of the proposed tourist destination image branding technique is performed using the JAVA technology. After implementation the performance of the model is also computed. Based on the experimental results the following outcomes are observed as demonstrated in table 4.1.

S. No.	Parameters	Remark
1	Positive probability	The positive probability is fluctuating between 0.65-0.81 which is highly acceptable for recommending the visitors place
2	Negative probability	Negative probability is varies between 0.34-0.09 therefore it is acceptable for recommendations
3	Memory usages	The main memory requirements are depends on the data to be process and it is observed between 28K-41K KB
4	Time consumption	Time consumption for computing the recommendation is varies between 26-40 MS which is acceptable for the data processing

Table 4.1 performance observations

According to the obtained performance of the system the proposed model is an accurate data model for data analysis and can recommend the tourist destination credibility successfully.

B. Future Work

The main aim of the proposed work is to develop the data mining model for recommending the tourist destination credibility according to the fresh data mining. The implementation and their performance evaluation are performed successfully. In near future the proposed model can be extended for the following domains.

1. Current system only extract the data that is appeared on the search results the relevant inner links are not downloaded. Therefore in near future the inner link data extraction is also involved in this technique
2. The proposed system implements the classifier in normal mode in near future the ensemble learning concept is used for improving the classification performance of the system

REFERENCES

[1] Költringer, Clemens, and Astrid Dickinger, "Analyzing destination branding and image from online sources: A web content mining approach", Journal of Business Research 68, no. 9 (2015): 1836-1843.

[2] Berry Michael W., (2004), “Automatic Discovery of Similar Words”, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, pp. 24-43.

[3] Tan, Ah-Hwee, "Text mining: The state of the art and the challenges." In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Volume 8, pp. 65-70. 1999.

[4] Kumar, Lokesh, and Parul Kalra Bhatia. "Text Mining: Concepts, Process and Applications." Journal of Global Research in Computer Science 4.3 (2013): pp. 36-39.

[5] Stavrianou, Anna, Periklis Andritsos, and Nicolas Nicoloyannis, "Overview and semantic issues of text mining", ACM SIGMOD Record 36.3 (2007): 23-34.

[6] S. Vijayarani, J. I lamathi and Nithya, “Preprocessing Techniques for Text Mining: An Overview”, International Journal of Computer Science & Communication Networks, Volume 5(1), pp. 7-16

[7] Vishal Gupta, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Volume 1, Number 1, August 2009

[8] B. Singh, H.K. Singh, “Web data Mining Research”, in the Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010.

[9] Shen zihao, Wang Hui, “Research on E-commerce Application Based on Web Mining”, 2010, IEEE.

[10] Li Haigang and Yin wanling, “Study of Application of Web Mining Techniques in E-Business

[11] Sebastiani, F., “Machine learning in automated text categorization.” ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47, 2002

[12] Sonali Vijay Gaikwad and Archana Chaugule, “Text Mining Methods and Techniques”, International Journal of Computer Applications (IJCA), Volume 85 – No 17, January 2014.