ISBN 978-9937-0-4872-9
September 27, 2018

REINVENTING THE INTERNET FOR CUSTOMIZED USER EXPERIENCE

Sujal Paudel

Department of Computer Science and Engineering

School of Engineering, Kathmandu University

Dhulikhel, Nepal

thesujal17@gmail.com

Abstract- The basis of this research was to create a customized environment designed for the internet user, by the Internet Service Provider that would enhance their overall internet surfing experience. The idea was to develop a model that would meet the requirement of the internet for every single user in a distinguish manner. Using cross-platform technologies such as data mining, data analysis, machine learning and content development these environments were created. This study definitively gives enough of the information for Internet Service Providers to create the content and advertisement for customized environments as per needed by their user.

Keywords—customized, environment, Internet Service Provider, advertisement, machine learning, content

I. INTRODUCTION

Internet today connects more than 3 Billion people around the globe [1]. Actually, the internet was invented for the same purpose i.e. for the communication [2]. However, with the rise of 738 million users to 3 Billion people in timespan of 18 years [3], the purpose of internet has moved way ahead from just being a means of communication to a source of entertainment, information, knowledge etc. However, with the presence of such a large population, the contents present in the internet are in the overwhelming quantity. With the variety of contents in it, other means of information/entertainment be it television or radio are far behind in competition with the internet. The orthodox cable television industry has even started shifting towards internet television industry, due to the presence of variety of contents that 3 Billion internet users generate. The content generated off the internet has always been the reason to create a shift of consumers from cable television to the internet consumption, attracting worldwide audience to it. Such shift has created a sense of dilemma for the advertising companies in cable television. The research project tends to design a platform for delivering the content, publishing an advertisement by addressing the actual need of the user through the internet by the Internet Service Provider. The high expense in the television advertisement has leaded the local and small business to confine their presence within paper or radio. Advertisement is about inspiring people, and to inspire, a connection is needed, and with the presence of internet, we have enormous amount of connection present. So, a shift can be implemented in two sections for a particular internet user, first being in the content delivery which is tailor made for the user and second being advertisement schemes that can be generated with machine learning implementation, which in turn helps power those contents. The characteristic that defines the user and those prophecies that would excite the user were discovered as a result. In a nutshell the 'reinvention' of the internet defines development of a targeted approach of providing internet facility to the user by the internet service provider, while 'customized user experience' means providing the contents designed for the user, in a tailor made approach. The system's architecture is comprised primarily of two components, first being the website grouping module fundamentally for categorizing the websites while second being the product recommender module fundamentally for making recommendations to the user for various contents. Both of these modules further consists of submodules to simplify the tasks.

II. WEBSITE GROUPING MODULE

A module is a set of standardized parts or independent units that can be used to construct a more complex structure. There are over 1.5 billion websites in World Wide Web today [4]. A website [5] is collection of related web pages including multimedia content, typically identified with a common domain name, and published on at least one web server. Web Pages, which are the building blocks of websites, are documents, typically composed in plain text with formatting instructions of Hyper Text Markup Language [6]. So, this makes websites collection of text and thus, the internet chunk of text. This module in an overall view, groups the website present in the World Wide Web [7] in 11 different categories, some of them being art, news, entertainment, sports etc. This module further consists of two sub modules.

a) Web Extracting Sub-Module

This sub module extracts text that exists in a webpage following particular Hyper Text Markup Language tags [8], with the use of python [9] library beautifulsoup [10] and then abstracts it. The abstract is generated using the python library gensim [11]. Some basic programming knowledge is required for efficient use of this sub module.

b) Web Categorizing Sub-Module

This sub module comprises of some modern machine learning [12] approach. It needs to be provided with two data set, first training dataset which consist of the abstract of varieties of websites and the category under which those websites fall, while second being the testing dataset that consists of the abstract of website (generated from previous sub–module) whose category is to be found. These two training and testing dataset is then passed into text classification model designed using TensorFlow high level API [13]. The two dataset are read as pandas [14] data frame. Then a Natural Language package of python, Natural Language Toolkit [15] is used to tokenize all of the abstract present in testing data frame. Once the abstract is tokenized, they are passed into stemmer [16]. Stemmers are the logic which converts pronouns and nouns like running, runs into a single word, in this case run. Once the stemmer is completed, a bag of word [17] is created and one hot encoding [18] is applied through the entire abstract and label in the training set, and then deep neural network [19], from TensorFlow high-level API. Softmax is used as output layer in order to get the probability of the labels. When the same model is applied to the test set, all the function that was used for the training set is applied along with the same bag of words that was used for the training set. Finally some of the panda's logic is used to create beautiful data frame.

III. PRODUCT RECOMMENDER MODULE

This module requires of a standard file that is composed of those websites and their category which contribute in maximum traffic of the Internet Service Provider. The file acts as the standard measurement for a website, providing its category, which most of the Internet Service Provider's user view. The file needs to be updated as per the requirement so that it can meet those websites which are in trend.

TABLE I. Measurement file for website

Website	***Art***	***Business***
http://www.nepalstock.com/	0	1
https://www.gsmarena.com/	0	0
https://www.acetravels.com/	0	0
https://www3.gomovies.sc/	0	0
http://www.educatenepal.com/	0	0
https://www.facebook.com/	0	0
https://www.youtube.com/	0	0

This module is possible after computing results from first module i.e. Website grouping module. It comprises of three sub-sections.

a) Knowing the user

The feature of the user can be known with their browsing history, considering the fact that the user provides the permission, as it is intensely confidential. The log generated by the user is then compared with the standard file, which consist of the category in which the website present in user's log falls. This then leads the algorithm to know in which category the user actually falls i.e. either he/she is interested in technology, business, sports, arts or any other. Knowing the user gives the system an opportunity to present the user with customized environment that he/she might prefer i.e. a technology lover can be provided with a customized environment that would consist of technology related content be it articles, news, or video contents. Similarly the user with higher interest in arts can be delivered with the artistic contents be it certain painting, notice about auctions or such events that are happening locally. The contents help to promote the global and local content as well.

b) Advertising the user

The contents needs to be powered by some sort of finance, this can be done with the advertisement model designed for the user individually. We use collaborative filtering [20], to make tailor made advertisements for the user. Collaborative filtering helps the system to predict the probability of what a user might like, based on his/her interest on things. Working with the mechanism of collaborative filtering, this module enables to overcome the orthodox manner of advertising in cable television, which has high level of irrelevancy, thus making advertisements efficient and targeted. The advertisement model is built upon the collaborative filtering, which is user based, making this mathematical technique more sophisticated, accurate and relevant.

c) Customized Environment

These are basically the platforms that are designed for the delivery of the content that a user will view in his/her interface. These environments are the result that the system needs to generate, taking leads from above two modules. Hence, this can be considered as the final product, for which all the system is designed. The environments can comprise of various contents be it articles, videos, podcasts, pretty much anything, but designed for the user, and addressing his/her field of interest. As, mentioned above the environment would also consist of advertisements, that would power the productions of the contents financially and those advertisements would be tailor made for the user. The identity of the user here can be considered as the mac

address of a device which is unique and thus perfect for the system to work.

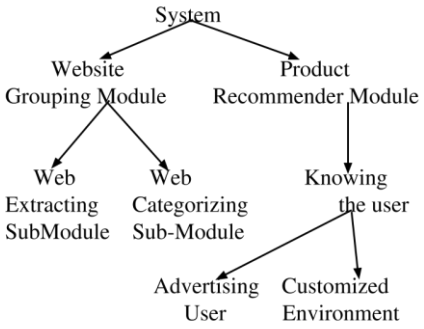

Fig 1: Representation of system

IV. AN EXPERIMENT

A practical implementation was done to accomplish the goal of the whole system. For the website grouping module, under web extracting sub module 15 different websites, were extracted and the contents of the website were labelled under 3 different categories, the categorization was done manually, as this was the training data set. Then for the web categorizing sub module, the system was trained with that similar training dataset, and as the testing dataset 3 different websites were extracted, they were then tested. The grouping module categorized the given test websites into the average of 0.33 for each of the three categories. The precision was not upto the mark nevertheless on further training of the dataset, with higher amount of input data, the precision can indeed be enlarged. For the product recommender module, a user's characteristic i.e. his interest in various contents by the categories of the websites he surfed was taken as the input, which was in the number. Finally the result obtained was also in the number, which signified the predictions for those websites the user didn't surf, considering the fact either he was unaware of those websites or had no interest. As, shown in the Fig 1 the Frequency_x represents the log of the user x, i.e. the websites he/she have visited while the Frequency_y represents the log of the user y. The collaborative filtering needs to be done between these two users

	Website	Frequency_x	Frequency_y
0	https://www.youtube.com/	13.0	12.0
1	https://www.facebook.com/	19.0	0.0
2	https://www.fashionnepal.com/	1.0	0.0
3	http://www.hamrobazaar.com/contactus.php/	1.0	0.0

Fig 2: Frequency of user's visit on various websites.

The Fig 3 represents the prediction about the preference of the user. In the Fig 2 the third row of Frequency_y column has the value 0, which signifies either the user y has no interests to visit the website https://www.fashionnepal.com or is unaware of such website, following up in the Fig 3 the value of third row of second column is -2.53, which is below average and signifies the user y is not interested about this particular website, while in the case of second row of third column of Fig 2, about the website https://www.facebook.com, the user is positive and attains the score of 30.74, as shown in second row of second column of Fig 3, which signifies though the user is unaware of the website nevertheless, he/she might be interested in using the website. The result obtained defined the personality of the user, expressing that the user is more interested to deal with social contents rather than fashion related items, thus environment that has higher relation to social contents can be designed for that particular user. In a nutshell, the system met the expectation kept with it, which in sequence helped to build up a customized environment for the user.

```
[25.499999934780682, 24.50000040425474]
[38.00000018483858, 30.746612192729046]
[2.0000000835199163, -2.5385473869091397]
[2.0000000835199163, -2.5385473869091397]
[2.0000000835199163, -2.5385473869091397]
[12.500000155956995, 5.5000000226885035]
[34.0000000834353, 16.373257624430046]
[2.000000075460705, 1.999999958086236]
[1.9999999999947136, 1.1819878132086992]
```

Fig 3: Probability of user to like particular content.

V. CONCLUSION

This paper, analyzes the new form of content delivery network that can be designed by the Internet Service Providers to their users. The obtained result from the system explained, brings up an ocean of opportunities for the content delivery network and the advertisement industry. The cost of the advertisement can significantly cut down as it being targeted and aired in a limited manner giving a chance for the small and local business to uplift and present themselves to those who are interested on them. The concept of prosumer [21] can arise where the producer of some content can act as the consumer of some other content based upon his/her fields of interest.

ACKNOWLEDGMENT

A sincere thanks to Mr. Bal Krishna Bal and Mr. Santosh Khanal, Department of Computer Science and Engineering, Kathmandu University.

REFERENCES

[1] K. Gordon, "Topic: Internet usage worldwide," Statista. [Online].Available: https://www.statista.com/topics/1145/internet-usage-worldwide. [Accessed: 03-Aug-2018]

[2] "Boutell.Com," WWW FAQs: What is my URL? [Online].Available: https://www.boutell.com/newfaq/history/inventednetwhy.html.[Accessed: 03-Aug-2018].

[3] "Internet," Our World in Data,[Online]. Available https://ourworldindata.org/internet.[Accessed:03-Aug-2018].

[4] "Total number of Websites," Google Search Statistics - Internet Live Stats. [Online]. Available: http://www.internetlivestats.com/total-number-of-websites [Accessed: 04-Aug-2018].

[5] "website," The Free Dictionary. [Online]. Available: https://www.thefreedictionary.com/website.[Accessed: 04-Aug-2018].

[6] "HTML Introduction," W3Schools Online Web Tutorials. [Online].Available:https://www.w3schools.com/HTML/html_intro.ASP . [Accessed: 04-Aug-2018].

[7] "What is World Wide Web (WWW)?-Definition from Whatis.com,"WhatIs.com.[Online].Available:https://whatis.techtarget.com/definition/World-Wide-Web.[Accessed: 04-Aug-2018].

[8] "HTML Tag," W3Schools Online Web Tutorials. [Online].Available:https://www.w3schools.com/tags/tag_html.asp . [Accessed: 04-Aug-2018].

[9] "Welcome to Python.org," Python.org. [Online]. Available:https://www.python.org[Accessed: 4-Aug-2018].

[10] "Beautiful Soup Documentation¶," Beautiful Soup: We called him Tortoise because he taught us.[Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc. [Accessed: 04-Aug-2018].

[11] Radim ehů ek and Petr Sojka, "Gensim—Statistical Semantics in Python" in EuroScipy, 2011, 25–28 August 2011, Paris, France [Online],Available: https://www.fi.muni.cz/usr/sojka/posters/rehurek- sojka-scipy 2011.pdf.[Accessed: 04 Aug. 2018].

[12] E. Alpaydin, "Introduction" in Introduction to Machine Learning, 1st ed. London, England: The MIT Press, ch. 1. pp. 1-3

[13] "TensorFlow Guide | TensorFlow," TensorFlow. [Online].Available:https://www.tensorflow.org/guide/#high_level_apis..[Accessed: 04- Aug-2018].

[14] F. Nelli, "The pandas Library—An Introduction," Python Data Analytics, pp. 63–101, 2015.

[15] S. Bird, "Nltk," Proceedings of the COLING/ACL on Interactive presentation sessions -, 2006.

[16]"Stemmers,"Preface.[Online].Available:http://www.nltk.org/howto/stem.html. [Accessed: 04-Aug-2018].

[17] "A Gentle Introduction to the Bag-of-Words Model," Machine Learning Mastery, 21-Nov-2017. [Online]. Available:https://machinelearningmastery.com/gentle-introduction-bag-words- model .[Accessed: 04-Aug-2018].

[18] "sklearn.preprocessing.OneHotEncoder¶," 1.4. Support Vector Machines - scikit-learn 0.19.1 documentation.[Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEnc oder.html. [Accessed: 04-Aug-2018].

[19] "A Beginner's Guide to Neural Networks and Deep Learning,"Skymind.[Online].Available:https://skymind.ai/wiki/neural -network . [Accessed: 04-Aug-2018].

[20] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76–80, 2003.

[21]"Forbes,"Forbes.[Online].Available:https://www.forbes.com/#2ad4d31f2254 . [Accessed: 04-Aug-2018].