Identification of Career Interest using Text Mining Techniques

Dr. Sai
Assistant Professor of Statistics
Vedavyasa College of Arts & Science

Dr. Chandra Mohan
Associate Professor of Statistics
Govt. Arts & Science College

Abstract:

Career choice is one of the most important decisions a student or a professional has to make in his life. When choosing a certain profession, it is important to carefully find out what are the interests and ambitions of the individual. We can do this by means of career interest inventories. There are plenty of tools available for the identification of career interest, either in paper- pencil form or as online tests. But in the context of artificial intelligence and machine learning, it is very interesting to enquire whether one’s vocational interest can be identified by analysing the posts a person made in social media like Facebook or Twitter. This paper is an attempt in this direction to identify the career taste of a person using text mining techniques. Holland’s RIASEC model is the theoretical base of the method that we derived, and the vocational classification is made on the basis of O*net career database. The text used for text mining is from the tweets made in social media platform, Twitter and the study shows perfect matching of career interest.

Key Words: career interest, vocational interest, text mining

I. INTRODUCTION

When information technology entered in the sphere of communication, it has become an inevitable part of all walks of human life. Now people are using social media for all types of communications ranging from official communications to personal chatting. Social media helps people to increase their contacts, enable them to learn new things, to share their ideas and to interact with new people. The increased use of social media has changed the way in which people think. Reports show that Facebook is the most popular social media platform with more than 2.3 billion users, and YouTube, Instagram and WeChat follow it with more than one billion users. Social media gives opportunity to individual to exhibit his or her behaviour topublic. By analysing these behaviours observed on social media, one can categorise these behaviours into individual and collective behaviour. Individual behaviour is exhibited by a single user, whereas collective behaviour is observed when a group of users behave together (Zafarani and Liu, 2014).

As the information available in public domain of social media can be extracted for different purposes, we can think of whether this information may be used to identify the career interest of the person. If a person's interest on jobs can be predicted from their social media profile, it can be benefited for different purposes like career selection, recruitment etc.

One of the methods of extracting the text information from social media is text mining. Text mining involves the use of statistical and machine learning techniques to learn structural elements of text in order to search for useful information in previously unseen text (Wen, 2001). It can be defined as the process of transforming unstructured text data into meaningful information. It utilizes different artificial intelligence technologies to automatically process data and generate valuable insights enabling researcher to make data driven decisions. Here in this paper, for the purpose of getting meaningful classification of career interests, Holland's theories of vocational choice (1997) is used as a criterion. The theory postulates the congruence between personalities of individuals and work environment leads to job tenure and satisfaction. The six types of personalities proposed in this theory are Realistic(R), Investigative (I), Artistic (A), Social(S), Enterprising (E) and Conventional (C).

II. TEXT MINING

Text Analytics, also known as text mining, is the process of examining large collections of written resources to generate new information, and to transform the unstructured text into structured data for use in further analysis. It is the process to extract interesting and significant patterns to explore knowledge from textual data sources (Ramzan Talib, 2016).

Pre-processing Methods

Pre-processing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process. The key steps of pre-processing are discussed below:

Extraction: This method is used to tokenize the file content into individual word.

Text Cleanup: Text Cleanup means removing of any unnecessary or unwanted information such as remove ads from web pages, normalize text converted from binary formats, deal with tables, figures and formulas.

Stop Words Elimination: Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pro-nouns, etc. that does not give the meaning of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, etc. Stop words are removed from documents because those words are not measured as keywords in text mining applications.

Stemming: This method is used to identify the root/stem of a word. For example, the words connect, connected, connecting, connections all can be stemmed to the word "**connect**". The purpose of this method is to remove various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space (Vijayarani, 2016).

III. HOLLAND'S THEORY

John Holland, professor emeritus at Johns Hopkins University, is a psychologist who devoted his professional life to researching issues related to career choice and satisfaction.

Holland's theory rests on four basic assumptions that describe how occupational interests are developed. The first assumption states that individuals can be categorized into Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C) types. The second assumption asserts that environments (e.g., places of employment) are also categorized into these same six types. The third assumption is that individuals tend to choose environments that fit with their personality. The fourth assumption highlights the importance of one's personality being congruent with his or her environment. It states that behaviour is determined by the fit between an individual's personality and the environment by which he or she is surrounded (Holland, 1997). The six Holland's occupational personality types are described below:

A) Realistic

- Likes to work with animals, tools, or machines; generally avoids social activities like teaching, healing, and informing others;
- Has good skills in working with tools, mechanical or electrical drawings, machines, or plants and animals;

- Values practical things you can see, touch, and use like plants and animals, tools, equipment, or machines; and
- Sees self as practical, mechanical, and realistic.

B) Investigative

- Likes to study and solve math or science problems; generally avoids leading, selling, or persuading people;
- Is good at understanding and solving science and math problems;
- Values science; and
- Sees self as precise, scientific, and intellectual.

C) Artistic

- Likes to do creative activities like art, drama, crafts, dance, music, or creative writing; generally avoids highly ordered or repetitive activities;
- Has good artistic abilities -- in creative writing, drama, crafts, music, or art;
- Values the creative arts -- like drama, music, art, or the works of creative writers; and □ Sees self as expressive, original, and independent.

D) Social

- Likes to do things to help people -- like, teaching, nursing, or giving first aid, providing information; generally avoids using machines, tools, or animals to achieve a goal;
- Is good at teaching, counselling, nursing, or giving information;
- Values helping people and solving social problems; and □ Sees self as helpful, friendly, and trustworthy.

E) Enterprising

- Likes to lead and persuade people, and to sell things and ideas; generally avoids activities that require careful observation and scientific, analytical thinking;
- Is good at leading people and selling things or ideas; □ Values success in politics, leadership, or business; and □ Sees self as energetic, ambitious, and sociable.

F) Conventional

- Likes to work with numbers, records, or machines in a set, orderly way; generally avoids ambiguous, unstructured activities
- Is good at working with written records and numbers in a systematic, orderly way;
- Values success in business; and

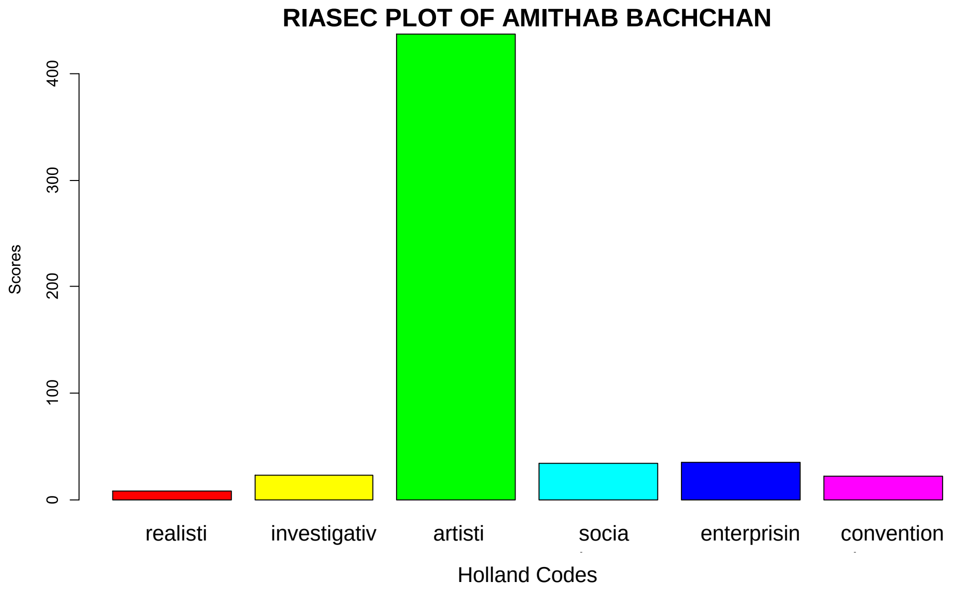
- Sees self as orderly, and good at following a set plan.

The importance of the RIASEC hexagon is to emphasize that one can fall on the border of different interests. People will have stronger preference in certain categories over others. Everyone is their own unique combination of the RIASEC interest types. The personality types closest to each other are more alike than those farther away.

IV. THE PRESENT STUDY

As a preliminary work, for the construction of a method to identify the career interest of person using text mining in R programming, we conducted an essay writing competition among 12th standard school students. The topic of the essay was ‘My Dream Job’. The students wrote about their dream job, hobbies and interests. Using the text mining methods in R programming, we classified each student in the basis of Holland’s codes. The 20 text documents are treated as one corpus. That is here our corpus is a collection of 20 text documents. Then the pre-processing steps like removing punctuations, removing digits, converting to lowercase, removing stop words, removing single words, removing extra white spaces and stemming the document are followed. We can construct a word cloud of the corpus and also a comparison word cloud to Plot a cloud comparing the frequencies of words across documents. Then the package ***stringr*** is used to split up a string into pieces. Here we used a lexicon-based classification of the text documents. That is 6 lexicons of the Holland codes REALISTIC, INVESTIGATIVE, ARTISTIC, SOCIAL, ENTERPRISING and CONVENTIONAL are created using the Holland’s theory and found the score matching of each document to these lexicons. These lexicons are made using the theory of RIASEC. The scores are arranged in descending order to find each student’s Holland code. Also, bar plot which shows the RIASEC scores for each student is drawn. Using the list of occupations grouped by the primary interest area (RIASEC code), taken from the Department of Labor’s O*NET database we can assign job to their interest.

Finally, the work was done using social media Twitter. Twitter has made the task of analysing tweets posted by users easier by developing an Application Programming Interface (API) which people can use to extract tweets and underlying metadata. This API helps us extract twitter data in a very structured format which can then be cleaned and processed further for analysis. To create a Twitter app, we first need to have a Twitter account. Once we have created a Twitter account, apply for a developer account using the new developer portal at developer.twitter.com. Write the basic details such as application name, description along with a website name. We may enter any test website name as well. Once we have entered these details, we will get the consumer Key (API Key), Consumer Secret (API Secret), access Token and access Token Secret. These keys and tokens will be used to extract tweets from Twitter into R.These tweets are treated as a corpus and RIASEC score is found using with the help of R programming software. As an example, the RIASEC plot of a person using the tweets he/she posted in twitter is given below:



V. R-CODE FOR DATA EXTRACTION

The R packages used in this analysis are tm, SnowballC, wordcloud, stringr, plyr. Lexicons of each of the personality types are created using Holland's theory. The match() function in R returns a vector of the position of first occurrence of each of the document in each of the lexicon.Using this vector RIASEC scores of each document can be found.Then RIASEC plot is drawn using barplot() function.

For Twitter analysis, use the package twitteR and RCurl in addition to the packages installed early. We will get the consumer Key (API Key), Consumer Secret (API Secret), access Token and access Token Secret after creating Twitter API. These keys and tokens will be used to extract data from Twitter in R.

VI. CONCLUSION

The study shows that using the text mining method, one can identify the career interest of a person using data uploaded in social media. We could classify each one according to their interests so that they can be matched with appropriate careers. Each of the RIASEC plot clearly shows the interest area of the person. According to the descriptions of Holland's personality types the types closest to each other on the hexagon have the most characteristics in common. Those types that are furthest apart, i.e., opposites on the hexagon, have the least in common. But in practical situations, these theories are partially contradicted.

VII. REFERENCES

[1] Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

[2] Kasper Welbersa, Wouter Van Atteveldtb, and Kenneth Benoit(2017), Text Analysis in R, *Communication Methods and Measures Vol 11, NO. 4, 245–265*

[3] Ramzan Talib, Muhammad Kashif Hanify, Shaeela Ayeshaz, and Fakeeha Fatimax(2016), Text Mining: Techniques, Applications and Issues, *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11,414-418*

[4] Vijayarani S. Preprocessing Techniques for Text Mining - An Overview, *International Journal of Computer Science & Communication Networks, Vol 5(1),7-16*

[5] Wen, Yingying (2001), Text Mining using HMM and PPM, *PhD Thesis*, Department of Computer Science, University of Waikato, New Zealand.

[6] Zafarani Reza and Liu Huan (2014). Behavior Analysis in Social Media, *IEEE Intelligent Systems, Vol, 29, No,4*