ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

A Proposed Method to Personalize and Extract the Hidden Knowledge Through Web Usage Mining and Pattern Discovery Using Web Usage Mining and Pattern Recovery

Ruchika Kalal, Manish Tiwari

Research Scholar, M. Tech. CS, Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

Assistant Professor, Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

ABSTRACT: Web mining is the application of data mining on web data and web usage mining is an important component of web mining. The goal of web usage mining is to understand the behaviour of web site users through the process of data mining of web data and Web usage mining is to understand the behaviour of web site users through the process of data mining of web Access data. In this paper, we have initiated an approach to acquire knowledge through google. The methods applied in this paper are web usage mining through which we have enhanced web design; introduce personalization service and facilitate more effective browsing the important an application of web mining by extracting the hidden knowledge in the log files of a web server and recognizing various interests of web users. Our novel approach will help in discovering customer behaviour, which is a newly proposed approach of web usage mining. In this paper, we provide an updated focused survey on different pattern discovery techniques of web usage mining.

I. INTRODUCTION

1.1 Project Overview

Google is a very popular and interactive medium for propagating information today. Due to the vast, varied and dynamic nature of web it raises the scalability, multimedia data and temporal issues respectively. The development of the web has been rise to large quantity of data that is freely available for user accessed by different users effectively and efficiently. That is why; the number of researchers in the field of application of Data mining techniques on the web is increasing

Client Level Collection

In this level, data is gathered together by means of java scripts or java applets. This data shows the behavior of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets. The advantage of data collection at client side is that it can capture all clicks including pressing of back or reload button.

Browser Level Collection

Second method of data collection is by modifying the browser. It shows the behavior of single user over multiple sites. The data collection capabilities are enhanced by modifying the source code of existing browser. They provide much more versatile data as they consider the behavior of single user on multiple sites.

Server Level Collection

Web server log stores the behavior of multiple users over single site. These log files can be stored in common log format or extended log format server logs are not able to store cached page views. Another technique used for usage data collection at server level is TCP/IP packet sniffers works by monitoring the network traffic and retrieve usage data directly.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

II. PAST CONTEXTUAL CONSTRUCT MODEL

What is Information Quality?

A. Defining IQ: "Fit-for-Use" (purpose)

Information quality is commonly described in the literature as a multi-dimensional concept (Ballou *et al.*, 1998; Klein, 2001; Pipino, 2002) with varying attributed characteristics depending on an author's philosophical and systems interaction point of view. Most commonly, the term "data quality" (often used synonymously with "information quality") is described as data that is "fit-for-use" (also "fit-for-purpose") (Wang & Strong, 1996), which implies that IQ is *relative*, as information considered appropriate for one use may not possess sufficient attributes for another use (Tayi & Ballou, 1998).

The "fit-for-use" paradigm has been embraced by researchers for a number of reasons. Firstly, it puts into common language the *action* of information quality while still remaining enigmatic and relative like the concept it is used to define. More importantly though, it gives information quality a *context* (Strong *et al.*, 1997a); that is; it suggests that information quality cannot be defined and assessed outside of the reason for which it exists.

Shanks & Corbitt (1999) contend that IQ should be assessed within the context of its generation, while Katerattanakul *et al.* (1999) add that it needs to be assessed according to its intended use. The reason for this contextual approach is both simple and logical, because it recognises that the attributes and dimensions used to assess IQ can vary depending on the context in which the data is to be used (Shankar & Watts, 2003).

B. Investigating IQ: The Information Retrieval environment of the Current Research

The user and information context to be addressed in this Paper is information retrieval in the information environment of the World Wide Web, an information environment devoid of the enforceable standards of quality associated with previous information environments (Hawkins, 1999; Brooks, 2003), where users (information seekers) are largely "on their own" in regards to searching, finding and retrieving target information (Hektor, 2003; Nicholas *et al.*, 2004, 2007). Understanding IQ from the point of view of the user (or searcher) of web-based information, involves understanding the processes of information seeking behaviour within this open system environment.

Research Scope: - The initial identification of appropriate academic literature for review was plagued with problems relating to research scope. This was reflected in the research proposal document written during the first six months of the Paper. The development of the research questions then became an important contributing factor in deciding how to micro-manage which literature would provide the best theoretical foundation for the research. The literature review then became topic-driven, rather than discipline-driven, which better suited the inter-disciplinary nature of the research. To a degree,

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

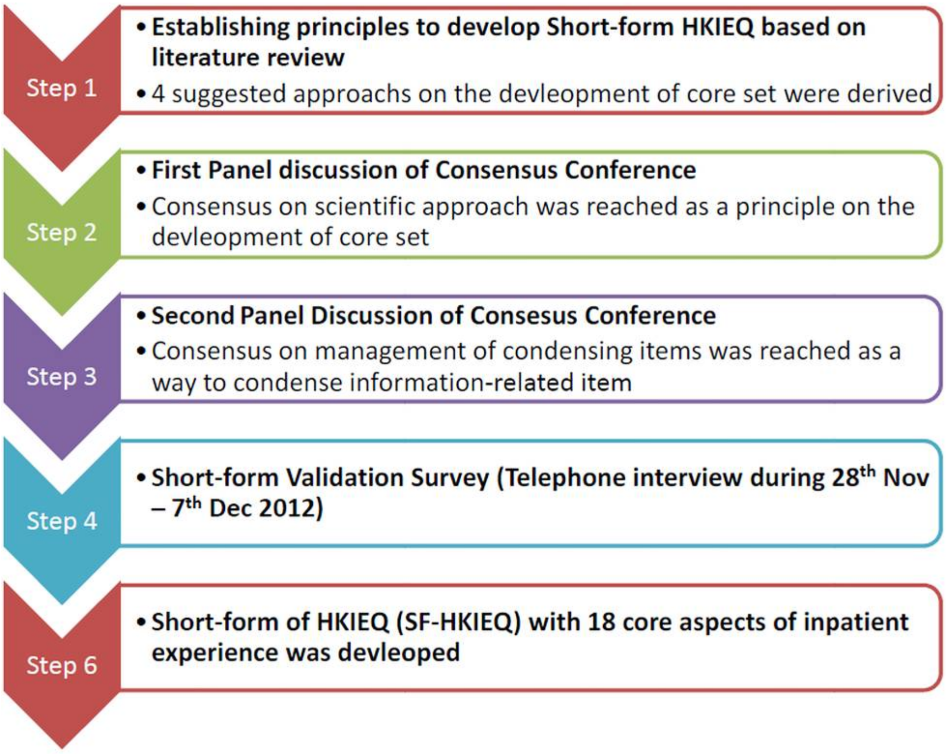

Figure 1– Illustration of layout chosen for the released TAM surveys

Data Analysis

The data analysis phase of the research, illustrated in figure 4.6, involved the synthesis and (2) analysis of user results, within a framework of (3) exploration; (4) confirmation; and finally (5) induction; processes. The following section describes the various strategies undertaken to handle and analyses the collected user results.

Bing: -

2011	2012	2013	2014	2015	2016	2017	2018
0.1	0.2	1.3	0	0.9	1.464	0.45	0.097
0.1	0.2	1.2	0	0.378	6.287	0.476	0.072
0.1	0.2	1.1	0	0.236	6.264	0.58	0.1
0.1	0.2	0.9	0	0.232	5.886	0.57	0.031
0.1	12.3	0.8	0	0.211	4	0.33	0.044
0.1	39.7	0.9	0	0.315	3.863	0.27	0.039
0.1	21.9	0.9	0	0.258	3.061	0.17	0.034

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

UC Browser: -

2011	2012	2013	2014	2015	2016	2017	2018
319.6	619.1	110	286.3	1655	869.38	1796.959	77.596
235.7	440	115	328.3	1701	1255.2	1210	68.784
190	632.6	110.7	416.6	2962	1140.9	1400.347	65.71
153.1	2424.6	103.8	414	2682	777.8	900.015	75.747
160	3721	104	424.4	2400	716.8	630.336	90.082
179.1	2929.3	100.5	480.3	1751	572.8	502.754	90.38
223.4	1255.3	100.33	450	1489	602.89	439.498	93.258

The Naïve Bayes Algorithm: -

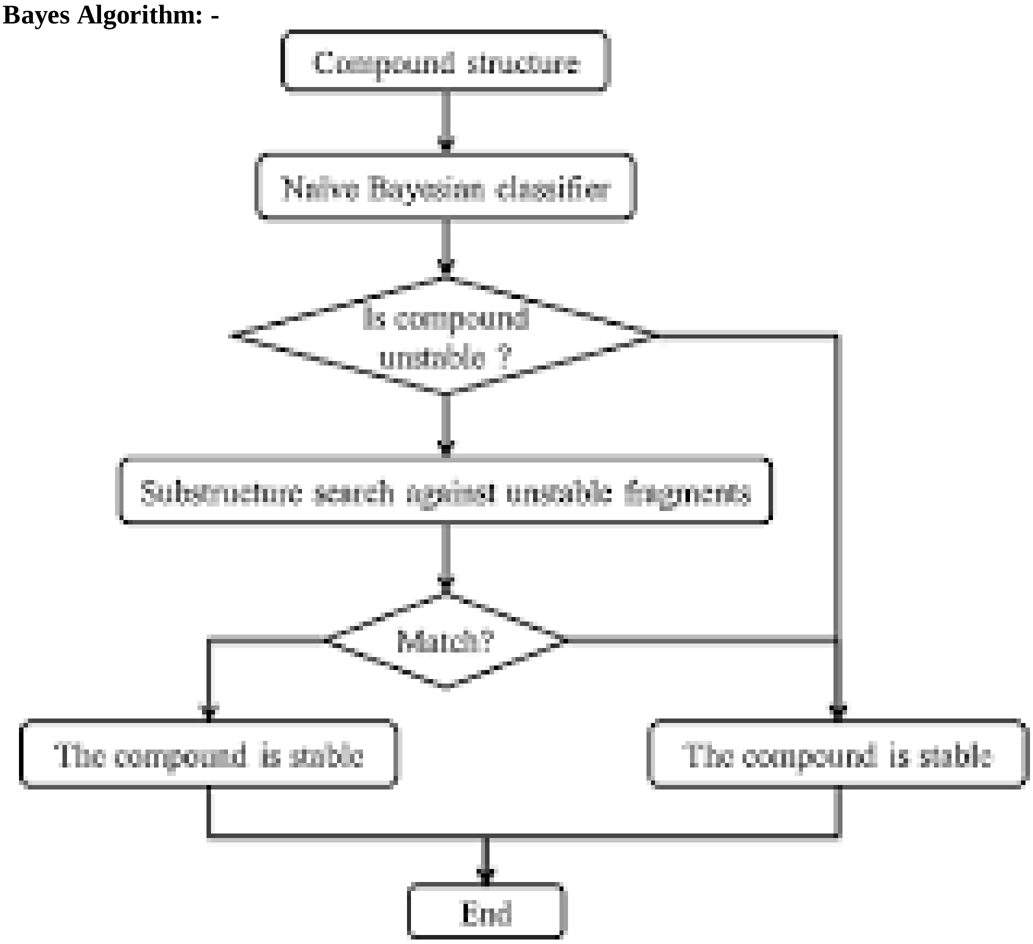

Figure 2 – Brief Flow chart for Naive Bayes

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

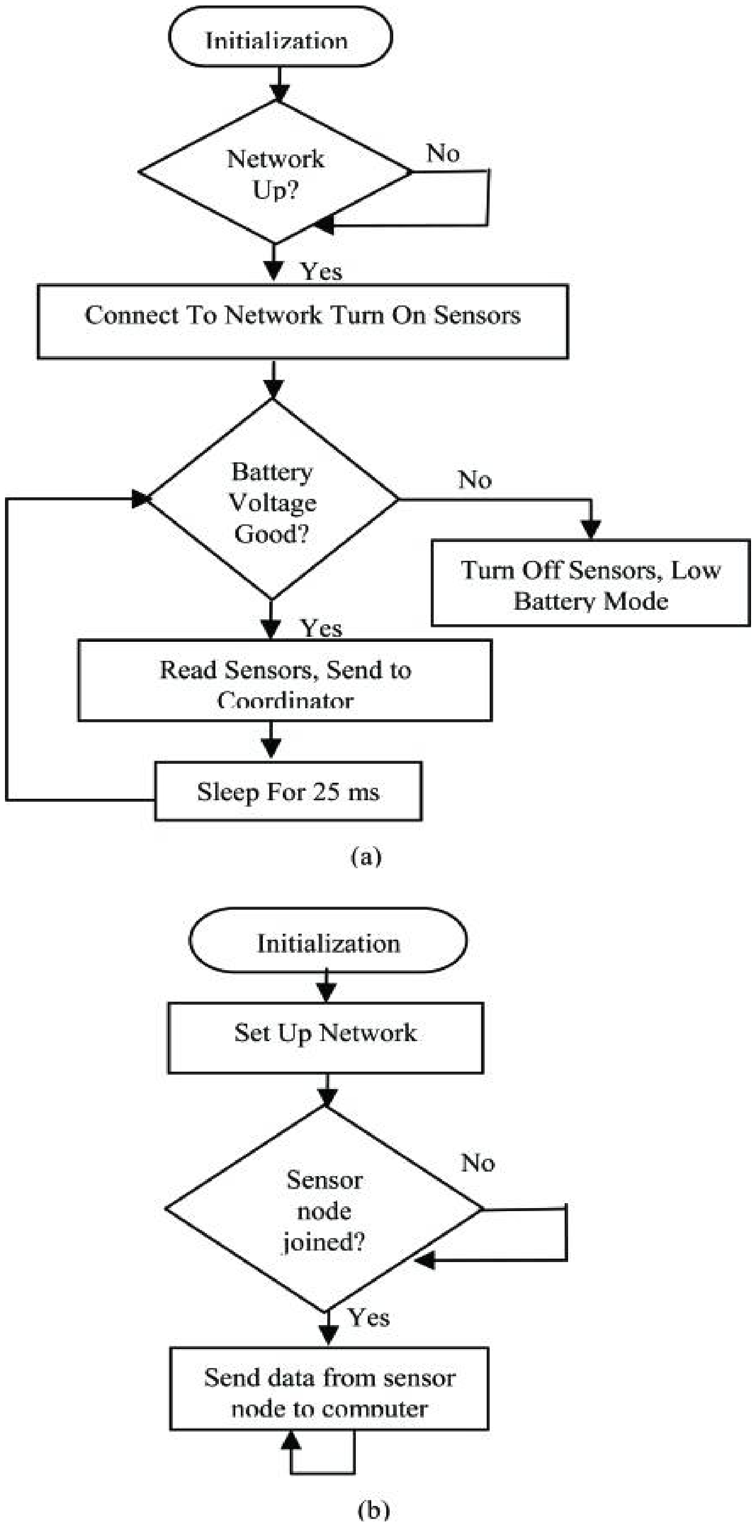

Figure 3 – Comprehensive Flow chart for Naïve Bayes

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

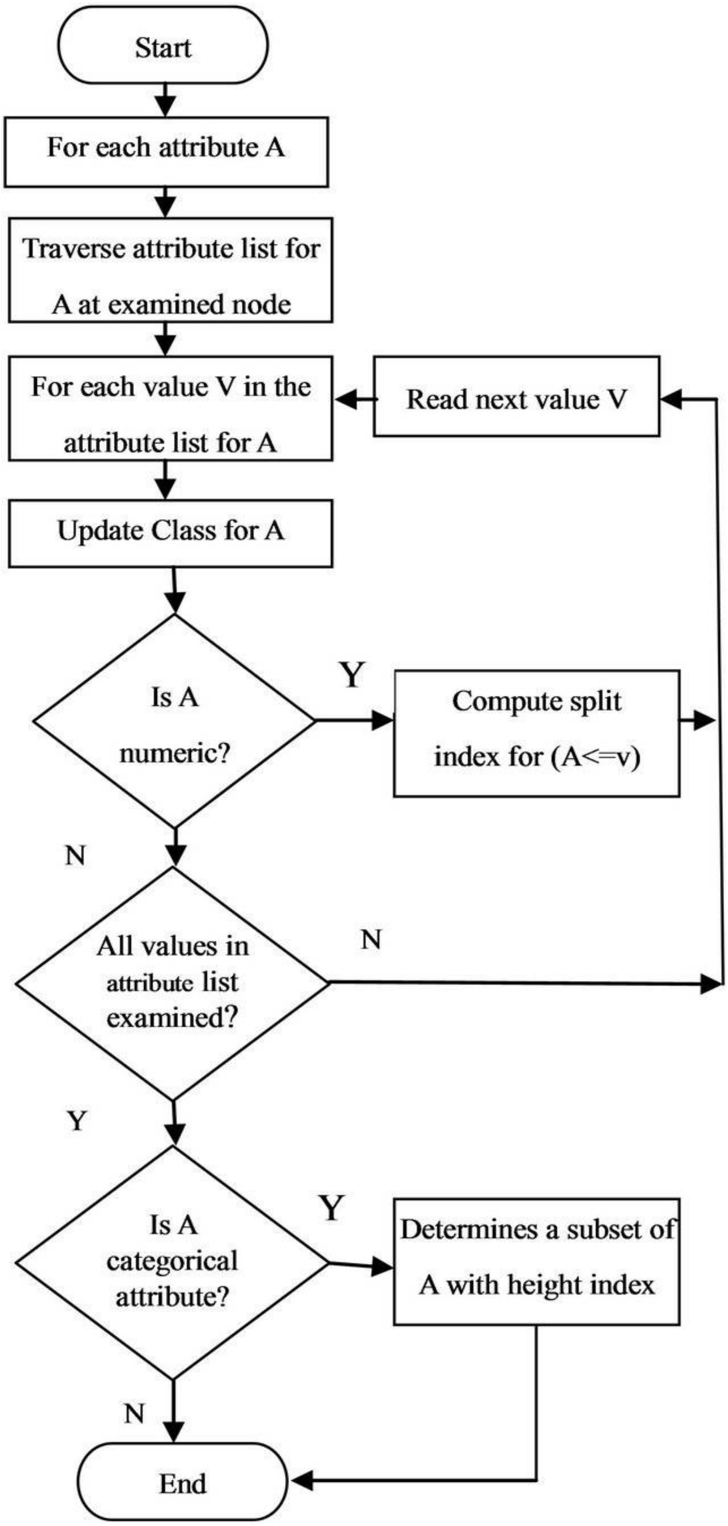

Figure 4 – Flow chart for Naive Bayes for feature extraction through Node Internet

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

IV. SIMULATION RESULT AND ANALYSIS

The Group-Case: Constructing a context

The goal of contextualising results is essentially to establish meaning to those results. Chapter 3 (*Research Methodology*) discussed the interpretivist view of how investigating a phenomenon within a context provides a back-drop by which meaning of participant results can be better understood. From a *big picture* point-of-view, the context of the current research is not just to understand users' perceptions of information quality, it is to understand the manifestation of these IQ perceptions in users' Web-based information retrieval behaviour. In this regard, it could be argued that, at a macro-level, the first case context is "information retrieval on the World Wide Web" and that the research into IQ perceptions is then conducted within this context. At a broad level, this serves to focus the research, enabling it to be compared to previous research and theory (Tsikriktsis, 2002; Chima, 2005) which has examined user perceptions of IQ in similar or different contexts.

In the same way that the broad research context can provide meaning in relation to other research, establishing cases and units of analysis within the research helps to provide meaning to results internally. Cases and units of analysis can be established through:

1.) Imposed existing theoretical frameworks – e.g. the different elements of human information retrieval such as information need; TAM theory; attribution, IQ and ISB theories;

2.) Known characteristic variables between types of users – e.g. gender; user experience; cognitive style; and academic position/role;

3.) The creation of sub-groups of clustered similar results to the same questions – e.g. did users of predominantly "phrase search" techniques (Survey #3, Q.10) have a higher or lower expectation of how often their searches were "successful" (Survey #3, Q.14) than users of predominantly "keywords" techniques? The same unit of analysis could be used to compare answers to other questions such as whether users attribute a "successful search" to their search engine choice or their own search strategies (Survey #3, Q.15)

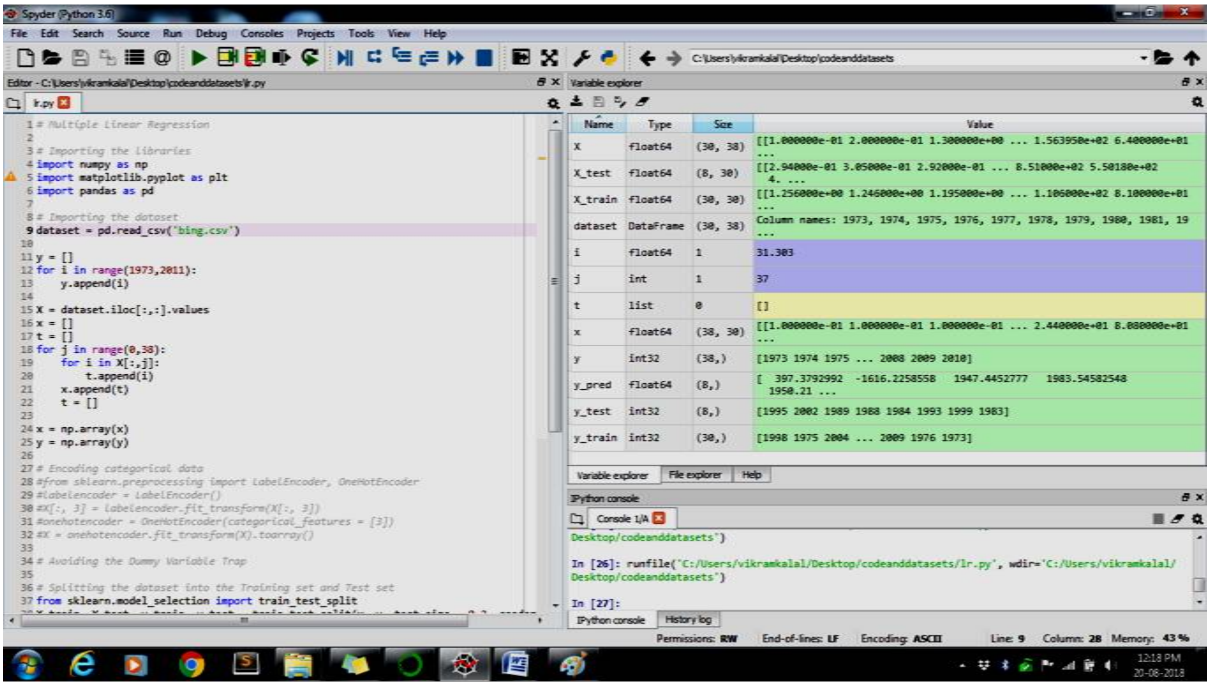

Figure 5 – Result of Bing Datasets

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

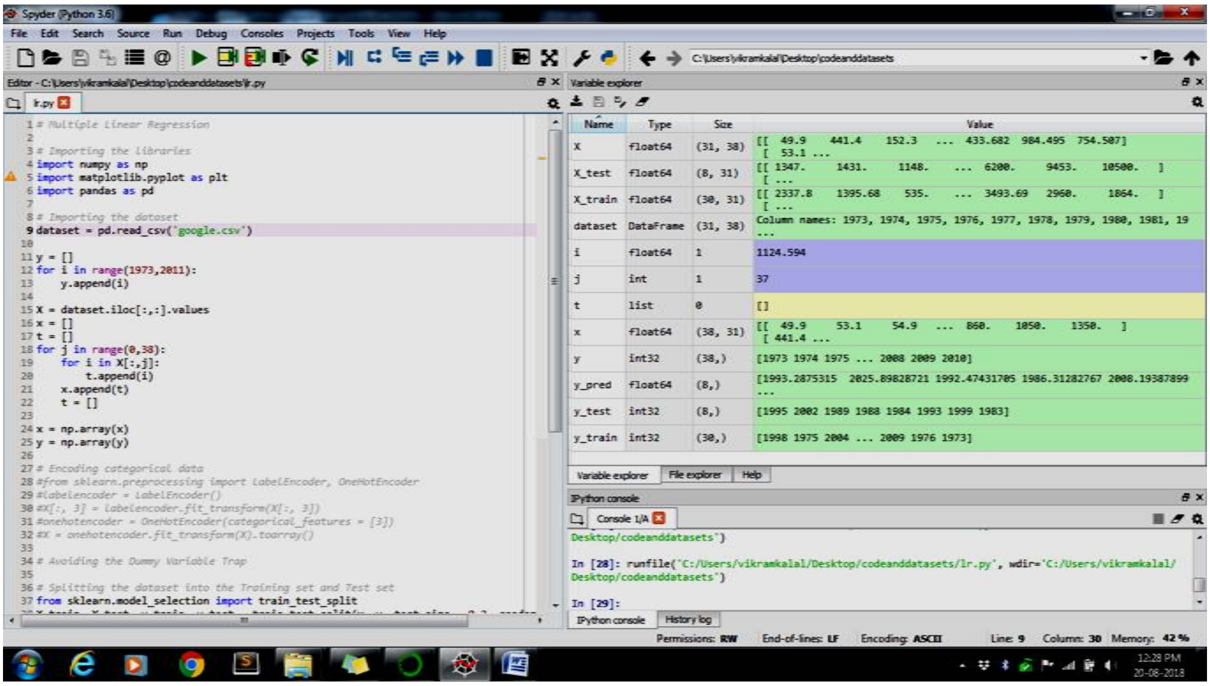

Figure 5 – Result of UC Browser Datasets

V. CONCLUSION AND FUTURE WORKS

Conclusion

This section presents some initial observations regarding the general characteristics of the user-group in this research study. It also presents the various constructed "group-cases" associated with the research, created from clustering sub-groups of users who possess similar characteristics, which will be used, as part of the research analysis framework (figure 4.6), to compare and cross analyze user results.

The user-group did not necessarily have to feel "comfortable" retrieving work/research related information from the Web, but needed to do so relatively regularly and be personally familiar with the process of using the Web as an information retrieval tool for the high quality content associated with their work, research, or both. Users who engage the Web as a means of professional networking, or even entertainment were not excluded from the target user-group. The surveys and questionnaires they completed however, did not relate to these interactions. The goal of the research was to survey a relatively intellectually sophisticated group of users. An assumption was made that academics and postgraduate level students, Honours, Masters and Paper level university students, would possess; (1) a relatively high degree of information quality perception; and (2) the ability to make relevant quality related judgments of the information they encounter on the Web.

REFERENCES

1 Abels, E. G., Liebscher, P., & Denman, D. W. (1996). Factors That Influence the Use of Electronic Networks by Science and Engineering Faculty at Small Institutions. Part I. Queries. *Journal of the American Society for Information Science*, *47*(2), 146-158

2 Ford, N., Miller, D., & Moss, N. (2001). The role of individual differences in Internet searching: an empirical study. *Journal of the American Society for Information Science and Technology*, *52*(12).

3 Ford, N., Miller, D., & Moss, N. (2002). Web search strategies and retrieval effectiveness: an empirical study. *Journal of Documentation*, *58*(1), 30-48.

4 Ford, N., Miller, D., & Moss, N. (2005). Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes, and approaches. *Journal of the American Society for Information Science and Technology*, *56*(7), 741-756.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

5 Ford, N. (2004). Modeling cognitive processes in information seeking: From Popper to Pask. *Journal of the American Society for Information Science and Technology, 55*(9), p.769-782.

6 Forslund, H. (2007). Measuring information quality in the order fulfilment process. *International Journal of Quality & Reliability Management, 24*(5), 515-524.

7 Foster, A. (2004). A nonlinear model of information-seeking behavior. *Journal of the American Society for Information Science and Technology, 55*(3), p.228-237.

8 Fugmann, R. (1973), "On the role of subjectivity in establishing, using, operating and evaluating information retrieval systems" Information Storage and Retrieval, Vol.9 No.7, p353-72.

9 Fourie, I. (2006). Learning from web information seeking studies: some suggestions for LIS practitioners. *The Electronic Library, 24*(1), 20-37.

10 Fox, E. A. (1987). Development of the CODER system: a testbed for artificial intelligence methods in information retrieval. *Information Processing & management, 23*(4), 341-366.

11 Freudenthal, D. (2001). Age differences in the performance of information retrieval tasks. *Behaviour & Information Technology, 20*(1), 9-22.

12 Frias-Martinez, E., Chen, S. Y., & Liu, X. (2007). Automatic cognitive style identification of digital library users for personalization. *Journal of the American Society for Information Science and Technology, 58*(2), 237-251.

13 Fusilier, M., & Durlabhji, S. (2005). An exploration of student internet use in India: the technology acceptance model and the theory of planned behaviour. *Campus-Wide Information Systems, 22*(4), 233-246.