APPLICATION OF FUZZY LOGIC FOR USER CLASSIFICATION IN PERSONALIZED WEB SEARCH

S. Sendhilkumar[1], K. Selvakumar[2] and G.S. Mahalakshmi[3]

[1,2,3] Dept. of IST, Anna University, Chennai

ABSTRACT

Classifying web users in a personalised search setup is cumbersome due the very nature of dynamism in user browsing history. This fluctuating nature of user behaviour and user interest shall be well interpreted within a fuzzy setting. Prior to analysing user behaviour, nature of user interests has to be collected. This work proposes a fuzzy based user classification model to suit a personalised web search environment. The user browsing data is collected using an established customised browser designed to suit personalisation. The data are fuzzified and fuzzy rules are generated by applying decision trees. Using fuzzy rules, the search pages are labelled to aid grouping of user search interests. Evaluation of the proposed approach proves to be better when compared with Bayesian classifier.

KEYWORDS

Web search, Personalization, Fuzzy classification, Fuzzy rules and World Wide Web

1. INTRODUCTION

Web search represents a significant portion of Web activity. There are three basic approaches to web search: i) Search Engines ii) Web Directories and iii) Hyperlinks [1]. However neither of these traditional search processes is effectively made to handle user's interest and their shift in search interest. Therefore results returned by existing search techniques do not provide the acceptable relevant search information. The underlying assumption is that user query is static and the search process has iterated the query recursively until the required information is retrieved [2,3].

Yet, there are difficulties encountered in web search which fall into five categories: i) Dynamic changes in data ii) Difficulties during retrieval of data iii) Understanding the context of search requests iv) Continuous drift in search and v) Difficulties in User and system communication [4,5]. Information on the Web is dynamic, heterogeneous and rapidly increasing, modified, and moved out. It is also un-structured and not self organized, so surfing needed information is dangling in nature [6,7]. In reality, during the search process users learn to go through the information, read the titles in result sets, retrieve the documents themselves, viewing a list of topics related to their query terms and navigating through hyperlinked Web sites [8].

User's information needs are not satisfied by a single search process, there is a mismatch between the retrieved documents and the required information. Information encountered at one search point may lead to a new unanticipated direction, so the user's required information and consequently their queries continually shift. The user model has to translate user requirement into a query form within some agreed upon system model. Keeping track of user's choice made during their search process allows them to return to temporarily unwanted contents, jump from one category of the content to another and, to retain information and the context across search sessions [9-12].

Personalized Web search mechanism effectively obtains the user's real time required information [13,14]. Significant research progress exists in this area towards improving search results according to user interest; yet this has not been made into a commercial facility in existing search engines. The reason is that the search results are often interpreted unsatisfactory.

DOI: 10.5121/ijci.2014.3303

People usually spend time to browse, read the irrelevant documents and continue to spend their valuable time until they find satisfactory results. The main reason behind this is that it is very hard to understand user's search intention; also it is fuzzy in nature. There is always a continuous change or drift in their search intension. In order to estimate change of interest of the users, following parameters shall be considered during the search process: i) Explicit (denotative) factors like user thrown queries, keywords, URLs, content referred, log-files, link structure, preference of previous users, etc. ii) Implicit (connotative) factors like semantics, term- list, time-spent on the page, scrolling speed, size of the page, click-through, maximum scrolling-speed, minimum scrolling-speed, time-spent per page-size, etc. [3,15,16].

The above mentioned user based parameters contains fuzziness due to the dynamism in their interest as well as shifting nature of their intension from one category to another. Collection of these parameters in direct manner is unethical [5]. In order to predict the user interest using these fuzzy nature parameters, personalized search process requires special and effective new user model which incorporates a fuzzy approach [17-19]. This fuzzy model shall be utilised to overcome fuzziness from the various user parameters and accurately predict the user's interest class as well as the interest shift between different user needs. This work proposes a fuzzy based user classification of search results in a personalised web search setting.

The rest of the paper is organized as follows. Section-2 provides the related works of previous Web personalization approaches with soft computing techniques. The proposed user model is described in section-3. Data pre-processes and validation of the user's data set, the relationship between the various users' search parameters, the randomness and uncertainty estimation of the user's data using frequency test and various statistical measures, mathematical notations of various fuzzy membership functions on different user search parameters, generation of fuzzy rules using decision tree algorithm and user grouping are shown in sub sections. Section-4 depicts a discussion and evaluation of results.

2. RELATED WORK

Fields like artificial intelligence, web mining and pattern recognition are influenced by many factors. Classification is one such important factor. Heuristic searching techniques forms the basis for majority of common classification algorithms. Finding the subset of the representative rules from the restricted set of training data is the main role of these algorithms [20,21].

Fundamentally there are three objectives: generating user navigation profiles for link prediction; enriching the profiles with semantic information to diversify them, and obtaining global and language-dependent user interest profiles combined with web usage and content mining techniques [22]. In this paper [22], the website content and the standard web usage information (log files from the web server) are taken as input. Semantic structure is automatically extracted and combined with information sources to extract knowledge for many applications. A crisp-based approach is used to assign the interest profiles to clusters. User's interests change dynamically during web traversal. Hence it is essential to detect the dynamic web traversal patterns based upon the user's progress. TFPM (Transitional Frequent Pattern Mining) and TUPM (Transitional Utility Pattern Mining) were proposed to detect the web patterns [23].

Analysing students' study environment to suggest personalised activities for each student [24] is another remarkable work in this area. The recommended suggestions were made to reinforce the students' competences in a subject. Characteristics of competences from each subject and the designed activities modelled by fuzzy linguistic labels compute the entire recommendations. A recommendation system based on fuzzy linguistic modelling has this approach as its core. And this approach allows students to receive personalised activities to practice competences which have to reinforce for passing a subject. The unhandled issue is that it is very difficult to determine one's affinity with other students because this system is not connected under a social networking platform.

In order to conceptualise documents into concepts using FFCM (Fuzzy Formal Conceptualization Model), a novel classification approach was developed [25, 26]. And this approach reduced the impact from textual ambiguity successfully. This work focuses only on single-class based classification due to the wide-ranging applications of text. In order to provide more complete coverage of FFCM applications, multi-class problems need to be handled. The performance of fuzzy classification systems was improved by reducing the decision subspace and its weight with the help of punishment algorithm [27].

In order to answer user consultations using natural language by Information Extraction (IE) in a set of knowledge platform, a novel method was proposed in [21, 28]. This methodology based on fuzzy logic engine, takes advantage of its flexibility for managing sets of accumulated knowledge. These sets are constructed in a form of a tree structure containing hierarchy levels. Designing and implementing an intelligent agent which can manage any set of knowledge from a space where information is imprecise, vague and abundant is the main goal of this system. Fuzzy logics were also used to summarize text for extracting the most relevant sentences [29].

For assisting the user to find the required experts an expert recommendation was proposed [30]. In order to construct the expert profile, a fuzzy linguistic method was adopted in this method. When the document is registered the fuzzy text classifier is used to get the relevant degree of document to each knowledge area, which forms the base for the construction of user profile. To derive the overall knowledge of the user, the user profile consisting of time and relevance factor of the rated documents is constructed. Based on the similarity between the user profile and the expert profile [31], the expert that fulfils the knowledge needs are recommended. For accurate building of user interest profiles, the method requires the user to rate similar to explicit feedback and the user must be reluctant to rate. But this stands to be a common drawback for all content-based recommendation methods. In the recommendation process like consulting history and the collaborative filtering, required information will be considered and filtering may be combined to make the recommendation more precise and comprehensive.

Two key technologies, namely pruning the outliers in the training data by SVMs (support vector machines) by eliminating the influence of outliers on the learning process and finding the fuzzy set with sound linguistic interpretation to describe each class based on AFS (Axiomatic Fuzzy Set) theory are fully capitalised by a classification method [32]. Aspects such as the improvement of learning methods in terms of computational complexity, especially in terms of adaptability with large scale data sets have to be improved further.

Although still a prototype, the idea of detecting emotions from the text, in terms of the sentiment polarity and the emotions has been proposed [33]. A fuzzy- based modelling of emotions enables the identification of linguistic patterns that intensify or reduce such emotions. Fuzzy logic is also used for evaluating the normalization completeness [34]. It was carried out in two steps – finding the relation's normal form and scaling the normalization. Even for software effort estimation, fuzzy approaches like Trapezoidal Membership Function (TMF) and Gaussian Membership Function (GMF) can be used as an alternative to traditional approaches like COCOMO model [35]. The use of fuzzy logic provides more flexibility for modelling the emotions; in addition, fuzzy modifiers are an adequate tool to tune these emotions according to the text.

The above mentioned fuzzy system based web mining research are showing their outstanding performance in the uncertain nature of the Web environment. This also proves to be the right decision for the recommendation process to be made from the vague and the uncertain environment in a robust manner. So it concludes that any Web mining application along with fuzzy methodologies will bring accurate estimation of user needs.

3. PROPOSED WORK: FUZZY BASED USER CLASSIFICATION MODEL

The proposed user model classifies/groups the users based on their interest labels using fuzzy approach. The classification model has the following components: Data collection and Pre-processing, Fuzzification, Rule generation, Label assignment and Grouping.

The customized Web browser [5] is used for collecting user search data such as: Username, IP address of the system, time spent, scrolling speed, URLs, queries, click through, page size, Scrolling speed and pre-processed using linguistic approaches. The data collected is then fuzzified, decision tree constructed and later rule generation is performed [3, 36]. The rules are then used to filter the pages that are not relevant to the user's search queries. The selected Web pages and user queries are mapped to categorical labels from ODP taxonomy. Finally the users are grouped based on the categorical labels.

3.1. Creation of Meta data

Data for experimental purpose is collected on the client side because various factors that affect personalization are not correctly reflected in the server side. For example page-view time recorded in the server logs is affected by network delay. Also cache hits are not recorded accurately in server logs [3]. In addition, it is very hard and complex to identify the specific user log information from the server side. Finally scrolling speed cannot be traced from the server side [3, 5].

After collecting data from the client side, the search data is pre-processed for extracting various factors that affect personalization. The following useful information is drawn out from the search data such as: i) User queries ii) Web pages visited iii) Scrolling Speed iv) Click through and v) Page size.

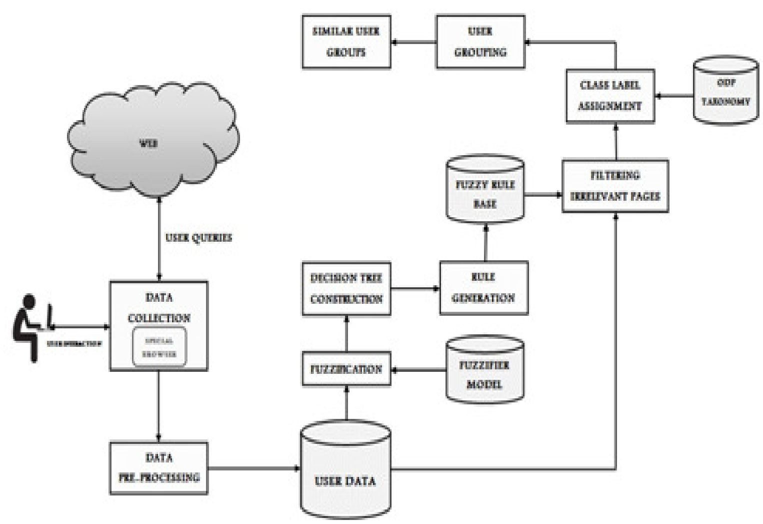


Figure 1. Fuzzy based Classification model

User queries which are the direct indications of user's interest are directly extracted from the input textbox [refer figure 1]. The Web pages visited by the users are parsed. The text content thus extracted is linguistically pre-processed for stop-word removal and stemming. The root words thus identified are used for indicating that Web page in the database.

Each page visited by the user consists of set of scrolling speeds recorded by the browser [5]. The average, maximum and minimum scrolling speed on a page visited by the user is computed from the set of scrolling speeds. The click through on a page is calculated whenever there was the change in address box in the browser. Change in the content of the address box occurs whenever the user clicks a link/URL available in the page that is currently being visited. The Time/Page-Size ratio is derived from the list of pages visited and time spent by the user in the page. A sample raw-data and pre-processed data is shown in Figure 2 & Figure 3 respectively.

Table 1. Experimental dataset information

User Data Set information			
Sl.No	**Factors in user data**	**Total**	**Ratio**
1	Number of Users	10	-
2	Number of search sessions	79	7.9 Search Session/User
3	Number of search queries	486	48.6 Search Queries/User
4	Number of Web pages visited	524	52.4 Web pages visited/User

Data collection involved 10 distinct users and their search data (524 pages) visited during their search process. A summarization of the data collected is shown in Table 1. The table also highlights average search sessions, search queries issued and Web pages visited per user.

```
Url:http://www.annauniv.edu
IpAddress:10.6.156.247
Time Spent:11.062 seconds
Scrolling Speeds encountered:76.54921020656137 12600.0 756.0
10250.0 345.5210237659963 12600.0 5906.25 12600.0
931.0344827586206 549.4186046511628 11812.5 25200.0 23625.0
748.0584474885845 147.54098360655738 11812.5 12600.0 11812.5
11812.5 403.8461538461538 11812.5 10250.0
Average speed:5240.347747397878 pixels/second


href="http://www.annauniv.edu/tanca2011/" target="_new"><span
class="style10"> GATE / Non-GATE Rank Enquiry  <img
src="newitem.gif" border="0"> </span></a><br>
                              <a
href="http://www.annauniv.edu/pdf/programme.pdf" target="_new">
<span class="style10">Workshop on Software Testing and Tools in
Dept.of Information Science and Technology</span></a><br>
                              <a
href="http://www.annauniv.edu/pdf/doublemaster.pdf"
target="_new"><span class="style10">Double Masters Degree
programme of Anna University and Ecole Centrale De Nantes,France
</span></a><br>
                              <a
href="http://www.annauniv.edu/pdf/austri.pdf" target="_new"><span
class="style10">Research & Scholarship Opportunity at The
University of Western Australia </span></a><br>
                              <!--a
href="http://www.annauniv.edu/pdf/bookexh.pdf" target="_new">
<span class="style10">3 Days Book Exhibition</span></a><br-->
                              <a
href="http://www.annauniv.edu/cia/adm.php " target="_new"><span
class="style10">Application form for  Masters' Degree Programmes
- Due date extended upto 19 August 2011 under  Children of Indian
workers in Gulf countries (CIWGC) / Non-Resident Indians (NRI) /
Foreign Nationals (FN) categories for the academic year 2011-2012
</span></a></img><br>
```

Figure 2: Snapshot representation of sample user data before pre-processing

	A	B	C	D	E	F	G	H	I	J	K	L
1	PAGE_NO	USER_NAME	UERY TYPE USE	URL_REFERRED	SAMPLE KEYWORDS	LICK TROUGH	TIME_SPENT(SEC	AVG_SCROLLING	MAX SCROLL	MIN SCROLL	SIZE OF THE PAGE(KB	TIME/PAGE SIZE
2	1	USER_1	Q1	http://www.ask.com/web?qsabout+we	Web, Mining,ask, service,browse,mining,(	1	99.828	508.0923354	2384.615385	30.20951762	143	4.733607527
3	2	USER_1	Q1	http://searchcrm.techtarget.com/defi	web, mining,relationship,web,mining,wor	1	85.735	1061.177159	6000	6.721695837	59	8.777677419
4	3	USER_1	Q1	http://www.ask.com/web?qsabout+we	Web, Mining,ask, service,browse,mining,(	1	36.547	946.5634291	12400	6.721695837	160	22.31612903
5	4	USER_1	Q1	http://www.ask.com/web?qsabout+we	Web, Mining,ask, service,browse,mining,(	1	11.171	946.5634291	12400	6.721695837	129	19.19187097
6	5	USER_1	Q1	http://www.ask.com/web?qsabout+we	Web, Mining,ask, service,browse,mining,(	1	10.906	882.6134444	11625	17.15392419	182	7.695032258
7	6	USER_1	Q1	http://informatics.indiana.edu/fil/Class	Web, Mining,ask, service,browse,mining,(	2	6.657	850.3021823	12400	6.721695837	7	1.041419355
8	7	USER_1	Q1	http://informatics.indiana.edu/fil/Class	web,mining,usage mining, patterns, log, c	2	69.298	810.5354534	12400	6.721695837	4	0.595096774
9	8	USER_1	Q1	http://www.imap.websemanticsjourna	semantic, web mining,www,ontogpqwsed	1	262	729.5507003	12400	6.721695837	1675	249.1967742
10	9	USER_1	Q1	http://www.ask.com/web?qsabout+we	Web, Mining,ask, service,browse,mining,(	1	10.031	738.2063203	12400	6.721695837	128	19.04309677
11	10	USER_1	Q1	http://www.cse.iitb.ac.in/soumen/min	contents,books,catalog,Unsupervised,lea	1	48.266	713.7313274	12400	6.721695837	9	1.338967742
12	11	USER_1	Q1	http://www.cs.uic.edu/~liub/WebMini	WebContentMining,Web mining,structure	2	108.39	646.1727913	12400	6.721695837	1191	177.1900645
13	12	USER_1	Q2	http://www.ask.com/web?qsweb+AND	web,mining,class,search,Effects,Negative,	1	66.016	339.7438681	2384.615385	16.46454811	125	7.592069892
14	13	USER_1	Q2	http://www.ask.com/web?qsweb+AND	web,mining,class,search,Effects,Negative,	2	49.422	324.2375758	2384.615385	16.46454811	114	6.923967742
15	14	USER_1	Q2	http://www.ask.com/web?qsweb+AND	web,mining,class,search,Effects,Negative,	2	29.25	316.870609	2384.615385	16.46454811	114	6.923967742
16	15	USER_1	Q2	http://www.ask.com/web?qsweb+AND	web,mining,class,search,Effects,Negative,	1	24.094	301.2777468	2384.615385	16.46454811	114	6.923967742
17	16	USER_1	Q2	http://weblab.com.cityu.edu.hk/blog/	Web,social,network,Lab,text,Mining,Univ	2	111.984	364.7874279	2384.615385	5.719914837	36	6.2938
18	17	USER_1	Q2	http://webcontentmining.com/	webcontentmining,web,mining,data,know	1	76.953	584.0221515	11250	5.719914837	61	10.66449444
19	18	USER_1	Q2	http://boston.lti.cs.cmu.edu/classes/1	browser,automatically,web,mining,data	1	9.235	576.6908373	6960	5.719914837	47	8.216905556
20	19	USER_1	Q2	http://www.cs.cmu.edu/~callan/Teach	business,intelligence,data,Mining,custom	1	4.875	557.7648687	6960	7.973251029	5	0.627096774
21	20	USER_1	Q3	http://www.ask.com/web?qsweb+mini	web,mining,data,WebMiningBook,knowle	1	37.625	427.607176	1488	24.04809619	121	5.001583333
22	21	USER_1	Q3	http://www.ask.com/web?qsweb+mini	web,mining,data,WebMiningBook,knowle	1	1.047	427.607176	1488	24.04809619	126	5.2395
23	22	USER_1	Q3	http://www.ask.com/web?qsweb+mini	web,mining,data,WebMiningBook,knowle	1	10.906	507.8498725	1488	24.04809619	111	4.61575
24	23	USER_1	Q3	http://www.ask.com/web?qsweb+mini	web,mining,data,WebMiningBook,knowle	1	50.828	506.5196757	2307.692308	24.04809619	114	4.7405
25	24	USER_1	Q3	http://www.ask.com/web?qsweb+mini	web,mining,data,WebMiningBook,knowle	1	205.735	531.6981686	2307.692308	24.04809619	115	4.782083333
26	25	USER_1	Q3	http://infolab.stanford.edu/~ullman/cs	Mining,Web,stanford,introduction,assoc	1	81.625	474.3123509	2307.692308	8.230107893	34	1.699419355
27	26	USER_1	Q3	http://www.ask.com/web?qsweb+mini	web,mining,data,WebMiningBook,knowle	1	10.735	458.4567582	2307.692308	8.230107893	115	13.95951613
28	27	USER_1	Q3	http://www.informatik.uni-trier.de/~le	Database,Systems,sigir2008,medical,Stat	1	10.938	456.9732062	2307.692308	8.230107893	160	18.20806452
29	28	USER_1	Q3	http://www.dais.unive.it/~dm/New_Sli	data, mining, classification, supervised,le	2	180.219	1736.645911	12400	8.230107893	5290	643.1088387
30	29	USER_1	Q3	http://www.dais.unive.it/~dm/	data, mining, extraction, knowledge, patte	2	90.203	1736.645911	12400	8.230107893	6	0.728322581
31	30	USER_1	Q4	http://www.ask.com/web?q=%22web+	mining,web,applications,Social,Networks,	1	96.422	830.8607586	2384.615385	12.50420168	128	10.23658854
32	31	USER_1	Q4	http://www.ask.com/web?q=%22web+	mining,web,applications,Social,Networks,	1	30.141	1073.072032	11625	12.50420168	122	9.75672043
33	32	USER_1	Q4	http://www.ask.com/web?q=%22web+	mining,web,applications,Social,Networks,	1	19.797	1487.048007	11625	12.50420168	124	9.916666667
34	33	USER_1	Q4	http://www.springer.com/computer/ai	book,Web,Mining,Usage,Analysis,Worksh	1	157.641	1207.484593	11625	2.010893106	75	37.14833333
35	34	USER_1	Q4	http://www.ask.com/web?q=%22web+	mining,web,applications,Social,Networks,	1	18.531	1117.2042	6000	2.010893106	103	51.01704444
36	35	USER_1	Q4	http://www.articlesbase.com/internet	pattern, knowledge,numbers,discover, da	1	61.75	1119.008258	11625	2.010893106	5	2.476555556
37	36	USER_1	Q4	http://www.ask.com/advancedsearch?	mining,web,applications,Social,Networks,	1	30.391	1079.210087	11625	2.010893106	118	58.44671111
38	37	USER_1	Q4	http://www.ask.com/web?q=%22web+	mining,web,applications,Social,Networks,	1	92.484	1263.953687	11625	2.010893106	118	58.44671111
39	38	USER_1	Q4	http://onlinelibrary.wiley.com/doi/10.	ebooks,computer,science,data,mining,co	1	204.14	1814.771198	11625	2.010893106	16	7.924977778
40	39	USER_1	Q4	http://www.dcc.uchile.cl/docs/CC73-2	web,mining,content,structure, usage, pat	1	120.406	1676.54894	11625	2.010893106	16	7.924977778

Figure 3: Snapshot representation of sample user data after pre-processing

3.2. Statistical analysis of pre-processed search data

The pre-processed user search data are analyzed using SPSS (Statistical Package for the Social Sciences) and StatPlus mathematical tools. The collected user data contains the randomness as well as uncertainty due to drift in their search process. It is tested and measured through frequency test and various statistical measures. A hypothesis test use sample data to test a hypothesis about the population from which the sample was taken. It makes inference about one or more population when sample data are available. Hypothesis test on user data based on time-spent on a Web page is considered based on statistical measures (mean, median and mode, etc) such as H_0 and H_1 called Null hypothesis and Alternative hypothesis respectively. Here H_0, H_1 are assumed and represented below

H_0: Data collected highlights randomness as well as uncertainty.
H_1: H_0 is not true

The results show that there is an uncertainty that exists in the users' browsing data. In order to resolve these issues fuzzy approach is incorporated with this model to effectively perform interest label based classification on the users. The randomness and uncertainty that exist in the Time-spent on a single page visited by the user in various sessions are shown in Table 2.

Table 2: Randomness and uncertainty existing in Time-spent on single-user's single page visit across various sessions

SESSION	PAGE_NO	TIME_SPENT(Sec)
I	P_1	99.828
	P_1	85.735
	P_1	11.171
	P_1	10.906
II	P_1	66.016
	P_1	49.422
	P_1	29.25
	P_1	24.094
	P_1	111.984
III	P_1	18.531
	P_1	61.75
	P_1	30.391
	P_1	92.484

3.3. Fuzzification on User search data

In order to illustrate uncertainty and vagueness mathematically, fuzzy concept is specifically designed. Everything is a matter of degree in fuzzy logic and knowledge is interpreted as a collection of elastic or equally fuzzy constraint on a collection of variables [32-38]. Concept of Linguistic variable plays an important role in fuzzy logic and their values are words or sentences in natural language [39, 40]. Any relation between two linguistic variables can be expressed in terms of fuzzy if-then rules [41, 42].

Fuzzy sets have been applied in many fields [18] in which uncertainty plays a key role. In particular, Web search area is an excellent example of vagueness and uncertainty [19, 43, 44]. Here the user interest is uncertain and fuzzy in nature; it is not simply classified using binary class labels (like Yes or No). In order to resolve this fuzziness in user interest, the pages visited by users are classified under labels like Not-Interested, Weakly-Interested, Mediumly-Interested and Strongly-Interested.

3.3.1. Fuzzification

The process of transforming crisp values into grades of membership for linguistic terms of fuzzy sets is known as fuzzification [25-26, 28, 45-47]. In this work, linguistic variables are evaluated using both triangular and trapezoidal membership functions and accompanied by degree of membership ranging from 0 to 1. The user data and its distribution is compared with various fuzzification membership functions. The collected user data like time-spent and scrolling-speed resembles triangular and trapezoidal membership functions respectively. Therefore this model utilizes both Triangular and Trapezoidal fuzzifiers for the fuzzification of user data as shown in Eq.1 & 3. Fuzzification of given user data can be performed by opting input parameters into X-axis (Horizontal line) and representing Y-axis (Vertical line) with the upper limit of the membership function for estimating the degree of membership.

3.3.1.1. Fuzzification of time-spent using Triangular Fuzzy membership function

Triangular function is defined by a lower limit a, an upper limit b, a value c, and x is current time-spent value, where $a < b < c$. It is a fuzzy number represented by three points as follows:

$$\mu_{Time\_Spent}(x,a,b,c) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ \frac{c-x}{c-b} & b \le x \le c \\ 0 & c \le z \end{cases} \qquad (1)$$

Here the entire user time-spent values are divided into three bins such as B_1, B_2 and B_3 and their mean values are calculated and assigned to a, b, c respectively The Triangular fuzzifier is applied on Time-Spent value of the user and the boundary values are defined based on bin-mean approach. The number of bin values is chosen based on the number of boundary values of the specific fuzzifier function.

For example, for 524 Web pages visited by an individual, the time-spent value is sorted and divided into three bins of sizes 174, 174 and 176 each. Here the first bin B_1 values and its mean value is represented. The below bin B_1 contains 174 time-spent values in sorted form, and its mean value is calculated as 16.43631. The boundary values of this fuzzifier are shown in the Figure 4.

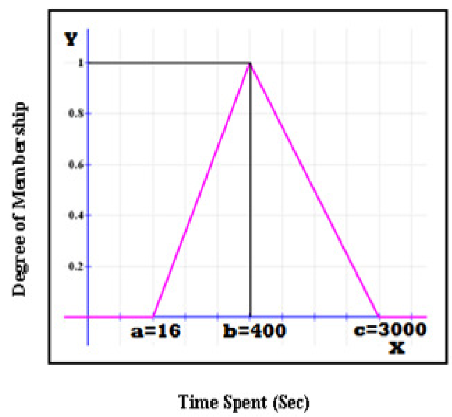


Figure 4: Boundary values of Triangular fuzzifier

B_1= [1.016, 1.031, 1.047, 1.094, 1.125, 1.406, 2.031, 2.485, 3.485, 4.844, 4.875, 5.203, 5.563, 5.719, 5.781, 6.297, 6.657, 7.125, 7.234, 7.375, 7.391, 7.516, 7.531, 7.594, 8.125, 8.453, 8.625, 8.641, 8.781, 8.859, 8.984, 9.235, 9.39, 9.953, 9.969, 10.031, 10.032, 10.047, 10.062, 10.11, 10.375, 10.39, 10.391, 10.641, 10.656, 10.735, 10.781, 10.844, 10.906, 10.906, 10.907, 10.938, 10.985, 11.031, 11.063, 11.094, 11.094, 11.125, 11.125, 11.125, 11.171, 11.203, 11.297, 11.437, 11.438, 11.625, 12.062, 12.062, 12.063, 12.109, 12.188, 12.312, 12.328, 12.375, 12.891, 13, 13.203, 13.328, 13.375, 13.5, 13.718, 13.719, 13.734, 13.75, 13.906, 13.953, 14.125, 14.188, 14.484, 14.734, 14.734, 14.844, 14.953, 15.469, 15.578, 16.234, 16.312, 16.359, 16.406, 16.515, 16.64, 17.516, 17.516, 17.813, 18.016, 18.531, 19.5, 19.594, 19.703, 19.797, 19.875, 20.265, 20.578, 20.687, 20.71, 21.047, 21.188, 21.203, 21.218, 21.297, 21.922, 21.984, 22.046, 22.079, 22.188, 22.281, 22.625, 22.625, 22.641, 23.313, 23.344, 23.562, 23.875, 23.907, 24.094, 24.141, 24.39, 24.531, 24.563, 25.094, 25.438, 25.484, 25.531, 25.75, 26.031, 26.141, 26.437, 26.625, 26.671, 26.797, 27.329, 27.969, 28.093, 28.344, 28.359, 28.859, 28.875, 29.156, 29.25, 29.984, 30.015, 30.141, 30.25, 30.296, 30.797, 30.844, 30.984, 31.125, 31.265, 31.281, 31.5, 31.687, 31.75, 32.438]

$$B_1\,Mean = \frac{\sum_{i=1}^{N} timpe - spent_i}{N} \qquad (2)$$

B_1 *mean* = 16.43631

Table 3: Time-spent parameter equivalent to Triangular fuzzy function based membership values

Page No.	Time-Spent(Sec)	Triangular Fuzzifier
P_1	99.828	0.028092
P_2	85.735	0.02337
P_3	36.547	0.006886
P_4	11.171	0
P_5	10.906	0
P_6	6.657	0
P_7	69.093	0.017793
P_8	262	0.08244
P_9	10.031	0
P_{10}	48.266	0.010813
P_{11}	106.39	0.030292
P_{12}	66.016	0.016761
P_{13}	49.422	0.0112
P_{14}	29.25	4.44E-03
P_{15}	24.094	0.002712
P_{16}	111.984	0.032166
P_{17}	76.953	0.020427
P_{18}	9.235	0
P_{19}	4.875	0
P_{20}	37.625	0.007246984

Similarly other two bin-mean values B_2 and B_3 are estimated and assigned to the corresponding boundary values $\boldsymbol{b}$ and $\boldsymbol{c}$ respectively. The remaining parameters such as page-size and time/page-size are approached in the same method and the membership values of these personalization parameters are derived. Time-spent parameter equivalent to Triangular fuzzy function based membership values for twenty Web pages visited (Pages P_1-P_{20}) by a single user during his/her Web search is shown in Table 3.

3.3.1.2. Fuzzification of time-spent using Trapezoidal Fuzzy membership function

A trapezoidal function is defined by a lower limit a, an upper limit d, a lower support limit b, an upper support limit c and current scrolling-speed x, where $a < b < c < d$. It is specified by four parameters a, b, c, d follows:

$$\mu_{Scrolling-Speed}(x,a,b,c,d)=\begin{cases}0 & x<a\\ \frac{(x-a)}{(b-a)} & a\le x\le b\\ 1 & b\le x\le c\\ \frac{(d-x)}{(d-c)} & c\le x\le d\\ 0 & d<x\end{cases} \qquad (3)$$

The Trapezoidal fuzzifier is applied on Scrolling-Speed of the user parameter and its boundary values are initialized to a =100, b =1000, c =2000 and d =3000 using bin-mean approach. In this case boundary values are four therefore the entire user dataset is divided into four bins and its mean values are estimated accordingly. Initialization of boundary values of this fuzzifier is shown in Figure 5. Table 4 highlights the fuzzified scrolling speed collected on a sample of 20 Web pages (Pages P_1-P_{20}) visited by an individual user.

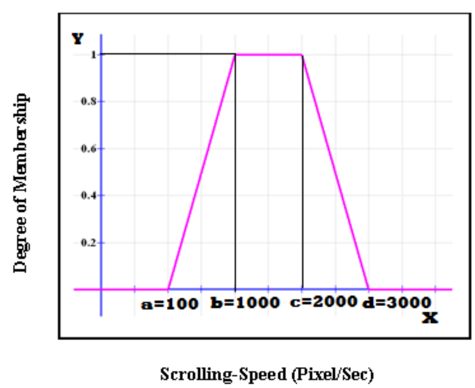


Figure 5. Boundary values of Trapezoidal fuzzifier

Table 4. Scrolling-speed parameter equivalent to Trapezoidal fuzzy function based membership values

Page No.	Scrolling-Speed (Pixel/Sec)	Trapezoidal Fuzzifier
P_1	508.0923354	0.78521456
P_2	1061.177159	0.494117285
P_3	946.5634091	0.554440311
P_4	946.5634091	0.554440311
P_5	882.6134444	0.588098187
P_6	853.3021823	0.603525167
P_7	810.5354534	0.626033972
P_8	729.5507003	0.668657526
P_9	738.2065203	0.664101831
P_{10}	713.7313274	0.676983512
P_{11}	645.1727913	0.713066952
P_{12}	339.7438681	0.873819017
P_{13}	324.2375758	0.881980223
P_{14}	316.870609	0.885857574
P_{15}	301.2777468	0.894064344
P_{16}	364.7874279	0.860638196
P_{17}	584.0221515	0.745251499
P_{18}	576.6906373	0.749110191
P_{19}	557.7648687	0.759071122
P_{20}	427.607176	0.827575171

3.4. Fuzzy Rule Generation

In this paper, the linguistic variable values which are used for representing various input personalized parameters are low, low-medium and medium and high. Not-interested [NI], Weakly-interested [WI], Mediumly-interested [MI] and Strongly-interested [SI] are the output parameters respectively. A decision tree technique is used to generate fuzzy if-then rules for classifying the user interest. Before generating fuzzy rules the entire personalized user parameters are fuzzified and assigned linguistic labels based on their fuzzy membership values discussed from the previous section 3.3.

The fuzzified values for 20 Web pages visited by an individual are given in Table 6. The fuzzified values and its equivalent linguistic labels are represented in the Tables 7. Here LMD, MED, HGH labels represents Low-medium, Medium and High respectively.

Table 5: Normalized information about the various user parameters

SL.NO	ATTRIBUTE	NORMALIZED VALUE	LINGUISTIC LABELS	FUZZIFIER APPROACHED
1	Time Spent	<16 16-400 400-3000 >3000	Low [1] Low Medium [2] Medium [3] High [4]	Triangular
2	Scrolling Speed	<100 100-1000 1000-2000 >2000	Low [1] Low Medium [2] Medium [3] High [4]	Trapezoidal
3	Click Through	<1 1-4 >4	Low [1] Medium [2] High [3]	Trapezoidal
4	Page Size	<10 10-105 105-203 >203	Low [1] Low Medium [2] Medium [3] High [4]	Triangular
5	Max Scrolling	<1936 1936-8582 8582-15790 >15790	Low [1] Low Medium [2] Medium [3] High [4]	Trapezoidal
6	Min Scrolling	<4.5 4.5-30.20 >30.20	Low [1] Medium [2] High [3]	Trapezoidal
7	Time Spent/Page Size	<0.098 0.098-12.27 12.27-19.05 >19.05	Low [1] Low Medium [2] Medium [3] High [4]	Triangular
8	Type of Interest (Class Label)	-----	Not-Interested [1] Weakly-Interested [2] Mediumly-Interested [3] Strongly-Interested [4]	----

3.3.1. Fuzzy Decision tree based Fuzzy Rule Generation

Proposed work analyzes some possible variants of making classification rules from a fuzzy decision tree based on cumulative information. Decision trees, which make use of fuzzy sets and fuzzy logic for solving the introduced uncertainties, are called Fuzzy decision trees (FDTs) [48-50]. Fuzzy decision trees mixes part of symbolic and sub-symbolic approaches. Fuzzy sets and symbolic logic permit modeling language-related uncertainties: whereas providing a symbolic framework for data quality. This work projected a brand new interpretation of Fuzzy C4.5, which relies on accumulative data estimate.

Table 6: Fuzzified user data

Page no.	Time spent	Avg scroll	Click through	Page size	Max scroll	Min scroll	Time/page size ratio
P_1	2	1	1	1	1	1	2
P_2	2	2	2	1	3	2	3
P_3	2	3	2	3	3	1	2
P_4	2	4	2	2	3	1	4
P_5	1	1	1	1	1	1	4
P_6	1	1	1	1	1	1	4
P_7	4	3	1	4	4	1	2
P_8	3	1	1	1	1	1	3
P_9	1	1	1	1	1	1	4
P_{10}	2	1	1	1	1	1	2
P_{11}	3	2	2	4	3	2	2
P_{12}	4	3	1	4	4	1	2
P_{13}	2	1	1	1	1	1	2
P_{14}	2	1	1	1	1	1	2
P_{15}	4	1	1	1	1	1	4
P_{16}	3	2	2	3	3	2	1
P_{17}	2	3	1	1	3	2	3
P_{18}	2	3	1	1	3	1	1
P_{19}	1	1	1	1	1	1	4
P_{20}	2	1	1	1	1	1	2

Table 7. Linguistic labels based user data with feedback

Page no.	Time spent	Avg scroll	Click through	Page size	Max scroll	Min scroll	Time/page size ratio	Feedback class label
P_1	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_2	LMD	LMD	MED	LOW	MED	MED	MED	WI
P_3	LMD	MED	MED	MED	MED	LOW	LMD	WI
P_4	LMD	HGH	MED	LMD	MED	LOW	HGH	NI
P_5	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_6	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_7	HGH	MED	LOW	HGH	HGH	LOW	LMD	WI
P_8	MED	LOW	LOW	LOW	LOW	LOW	MED	MI
P_9	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_{10}	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_{11}	MED	LMD	MED	HGH	MED	MED	LMD	MI
P_{12}	HGH	MED	LOW	HGH	HGH	LOW	LMD	MI
P_{13}	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_{14}	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_{15}	HGH	LOW	LOW	LOW	LOW	LOW	HGH	WI
P_{16}	MED	LMD	MED	MED	MED	MED	LOW	MI
P_{17}	LMD	MED	LOW	LOW	MED	MED	MED	WI
P_{18}	LMD	MED	LOW	LOW	MED	LOW	LOW	NI
P_{19}	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_{20}	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI

C4.5 is a propagation of ID3 that improves computing potency, deals with continuous values, handles attributes with missing values, avoids over fitting, and performs different functions [51-54]. Fuzzified 524 Web pages of user data is provided as an input to C4.5 algorithm. Both testing and training data sets are divided using bootstrap approach in order to generate the accurate decision tree. The decision tree thus constructed is given in figure 6.

User given feedback label based sample training dataset is shown in Table 7.This approach correctly classified 497 instances out of 524 and incorrectly classified instances are 27 and also its mean absolute error, root mean squared error and relative absolute error, etc are shown in Table 8.

Table 8: Various measures conducted on C4.5 decision tree generation process

Sl.No	Various Measures	Percentage
1	Correctly Classified Instances	94.8473%
2	Incorrectly Classified Instances	5.1526%
3	Mean absolute error	0.0462%
4	Root mean absolute error	0.152%
5	Relative absolute error	14.5805%
6	Root Relative absolute error	38.413%
7	Coverage of cases (0.95 level)	100%
8	Total Number of Instances	524

3.4.2. Fuzzy classification rule induction

In this empirical research work, we applied the above procedure on the various user attributes and generate 25 rules. The set of sample rules are shown in the Table 9. The same rules are applied for 10 different users in order for checking its completeness and consistency. Each user's interest may vary according to their different and depends on several factors. This user model considers the user interest as the decision variable.

The attributes that are playing major role in the user interest classification and the notations that are used in this work is normalized and presented in Table 9 also C4.5 algorithm generated tree equivalent rules are represented in Figure 7.

Table 9. Fuzzy based rules for classification of users

Rule no	Time spent	Avg scroll	Click through	Page size	Time/page size ratio	Interest
1	Low	Low	Low	Low	High	NOT INTEREST
2	Low Medium	Low	Low	Low	Low Medium	WEAKLY INTEREST
3	Medium	Low	Low	Low	Medium	MEDIUM INTEREST
4	High	Low	Low	Low	High	STRONGLY INTEREST
5	Low Medium	Medium	Medium	Medium	Low Medium	WEAKLY INTEREST
6	Medium	Low Medium	Medium	High	Low Medium	MEDIUM INTEREST
7	High	High	Medium	High	Medium	WEAKLY INTEREST
8	Medium	High	Medium	Low	Medium	NOT INTEREST

9	Medium	Medium	High	Medium	High	WEAKLY INTEREST
10	Low Medium	High	Medium	Low Medium	High	NOT INTEREST
11	High	Low Medium	High	High	Medium	STRONGLY INTEREST
12	Low Medium	Low Medium	Medium	Low	Medium	WEAKLY INTEREST
13	Medium	Low Medium	Medium	Medium	Low	WEAKLY INTEREST
14	High	Low Medium	High	Medium	Medium	MEDIUM INTEREST
15	Medium	Low	High	High	Low Medium	WEAKLY INTEREST
16	Low Medium	Medium	Low	Low	Low	NOT INTEREST
17	High	Low Medium	Low	Low	Medium	MEDIUM INTEREST
18	Low Medium	Medium	Low	Low	Medium	NOT INTEREST
19	High	Medium	Medium	Medium	Medium	MEDIUM INTEREST
20	High	Medium	Low	High	Low Medium	WEAKLY INTEREST

3.5. User Query mapping to ODP

The proposed user model then categorizes the user referred Web with the help of ODP taxonomy. Table 10 contains different query types of the single user in three sessions mapped to keyword based ODP taxonomical categories. Here the category estimation is performed based on a frequency count of each keyword from the user query which is compared with ODP taxonomy. It represents the number of times that particular keyword is matched with the entire list of the ODP structure. In each session there are five different forms of queries like keyword query, Boolean query, phrase query, proximity query and natural language query and the same are represented by Q_1, Q_2, Q_3, Q_4, and Q_5 respectively.

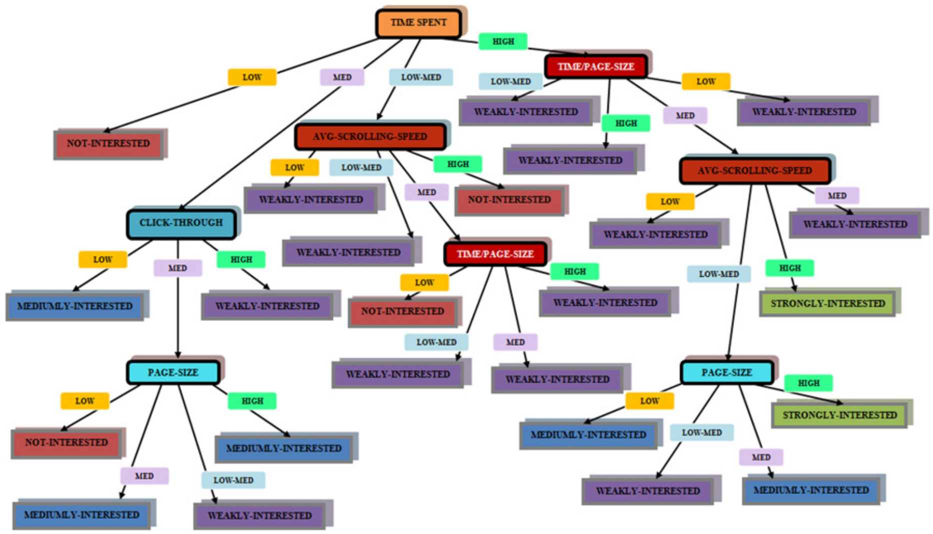


Figure 6. Decision tree for user interest classification

```
TIMESPENT = Low: NI (1.0)
TIMESPENT = LowMedium
|  TIME/PAGE SIZE RATIO = High: NI (1.0)
|  TIME/PAGE SIZE RATIO = LowMedium: WI (2.0)
|  TIME/PAGE SIZE RATIO = Medium
|  |  AVGSCROLL = Low: NI (0.0)
|  |  AVGSCROLL = Medium: NI (0.0)
|  |  AVGSCROLL = LowMedium: WI (1.0)
|  |  AVGSCROLL = High: NI (0.0)
|  |  AVGSCROLL = Medium: NI (1.0)
|  TIME/PAGE SIZE RATIO = Low: NI (1.0)
TIMESPENT = Medium
|  PAGESIZE = Low
|  |  AVGSCROLL = Low: MI (1.0)
|  |  AVGSCROLL = Medium: NI (0.0)
|  |  AVGSCROLL = LowMedium: NI (0.0)
|  |  AVGSCROLL = High: NI (1.0)
|  |  AVGSCROLL = Medium: NI (0.0)
|  PAGESIZE = Medium: WI (2.0)
|  PAGESIZE = High
|  |  AVGSCROLL = Low: WI (1.0)
|  |  AVGSCROLL = Medium: WI (0.0)
|  |  AVGSCROLL = LowMedium: MI (1.0)
|  |  AVGSCROLL = High: WI (0.0)
|  |  AVGSCROLL = Medium: WI (0.0)
|  PAGESIZE = LowMedium: WI (0.0)
TIMESPENT = High
|  PAGESIZE = Low
|  |  AVGSCROLL = Low: SI (1.0)
|  |  AVGSCROLL = Medium: MI (0.0)
|  |  AVGSCROLL = LowMedium: MI (1.0)
|  |  AVGSCROLL = High: MI (0.0)
|  |  AVGSCROLL = Medium: MI (0.0)
|  PAGESIZE = Medium: MI (2.0)
|  PAGESIZE = High
|  |  MAXSCROLL = Low: WI (0.0)
|  |  MAXSCROLL = Medium: WI (0.0)
|  |  MAXSCROLL = High: WI (2.0)
|  |  MAXSCROLL = LowMedium: SI (1.0)
|  PAGESIZE = LowMedium: MI (0.0)
```

Figure 7. C4.5 algorithm generated tree equivalent rules

Table 10. Different query types of the single user in three sessions mapped to ODP categories

Session no	Query type	Original query	Odp taxanomy category
1	Q1	about data mining	data_mining/databases/software/computers
1	Q2	about Web AND data mining	data_mining/databases/software/computers
1	Q3	“Web and data mining”	data_mining/databases/software/computers
1	Q4	Introduction about	data_mining/databases/software/computers

		”Web and data mining”	
1	Q5	definition of Web mining	data_mining/databases/software/computers
2	Q1	about Sachin	tendulkar,_sachin/players/india/full_members/icc/cricket/sports
2	Q2	about Sachin AND his records	tendulkar,_sachin/players/india/full_members/icc/cricket/sports
2	Q3	“Sachin Tendulkar”	tendulkar,_sachin/players/india/full_members/icc/cricket/sports
2	Q4	records of “Sachin Tendulkar”	tendulkar,_sachin/players/india/full_members/icc/cricket/sports
2	Q5	number of centuries hit by Sachin	tendulkar,_sachin/players/india/full_members/icc/cricket/sports
3	Q1	University of madras	university_of_madras/tamil_nadu/india/asia/colleges_and_universities/education/reference
3	Q2	University of madras AND distance education	university_of_madras/tamil_nadu/india/asia/colleges_and_universities/education/reference
3	Q3	“university of madras and its distance education scheme”	university_of_madras/tamil_nadu/india/asia/colleges_and_universities/education/reference
3	Q4	About “university of madras and its distance mode education”	university_of_madras/tamil_nadu/india/asia/colleges_and_universities/education/reference
3	Q5	List of courses in university of madras	university_of_madras/tamil_nadu/india/asia/colleges_and_universities/education/reference

3.6. User Grouping

After the estimation of categories, the proposed fuzzy model approaches fuzzy based classification on the specific query based pages which is referred by the user. Also this model generates the user grouping based on class labels assigned to each page (say, Not-Interest, Weak-Interest, Medium-Interest, and Strong-Interest). Using this model we have examined 10 users and their' accessed 524 pages and performed both interest based classification as well as clustering based on class labels. Also we performed other types of grouping using user queries and related pages as well as categorization based grouping.

In order to simplify the grouping process single user visited 58 pages are processed and grouped according to its interest class labels as well as other groupings. The entire pages have belonged to any one of the class label. Each page is represented as separate edge and there exists a link between the corresponding classes it belongs to. Similarly all other groups are formed and represented as a graph and shown in Figure 8. In order to make query based user accessed page grouping, number of queries thrown by the user is counted and each query related pages are estimated using similarity measure. The following sets represent the number of queries and its related pages accessed by the single user along with the category in which each query belongs to (refer Figure 9).

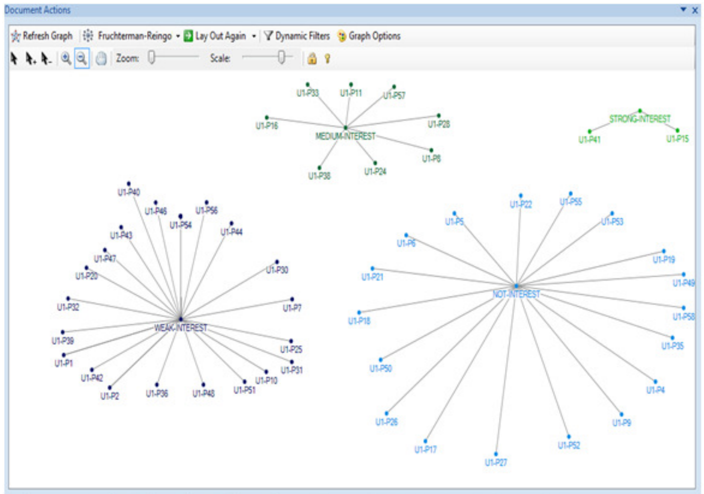


Figure 8. User grouping based on interest class labels

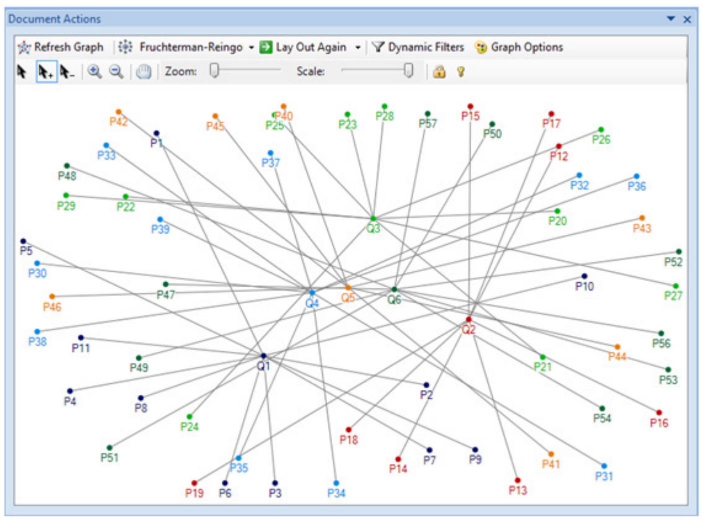


Figure 9. User query based Web page grouping

$Q_1 = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}, P_{11}\}$
$Q_2 = \{P_{12}, P_{13}, P_{14}, P_{15}, P_{16}, P_{17}, P_{18}, P_{19}\}$
$Q_3 = \{P_{20}, P_{21}, P_{22}, P_{23}, P_{24}, P_{25}, P_{26}, P_{27}, P_{28}, P_{29}\}$
$Q_4 = \{P_{30}, P_{31}, P_{32}, P_{33}, P_{34}, P_{35}, P_{36}, P_{37}, P_{38}, P_{39}\}$
$Q_5 = \{P_{40}, P_{41}, P_{42}, P_{43}, P_{44}, P_{45}, P_{46}\}$
$Q_6 = \{P_{47}, P_{48}, P_{49}, P_{50}, P_{51}, P_{52}, P_{53}, P_{54}, P_{56}, P_{57}\}$

The above sets represent six different queries thrown by asingle user (USER_1) during the entire search session. It also provides number of pages visited for each query and its corresponding page numbers from P_1 to P_{58} are grouped.

$Q_1 = \{C_1, C_2\}$; $Q_2 = \{C_1, C_3\}$; $Q_3 = \{C_1, C_2\}$;
$Q_4 = \{C_1, C_4\}$; $Q_5 = \{C_1, C_2, C_5\}$; $Q_6 = \{C_6\}$

Similarly six different queries from Q_1-Q_6 are grouped based on the category label C_1-C_6. i.e. Q_1, Q_3 and Q_5 belongs to category C_2 ("web_mining/data_mining/databases/software/computers") which is shown in Figure 10. This query category is performed using ODP taxonomy as well as cosine similarity measure approach. Also each query related information and its relevant category label names are shown. Finally this model provides various relationships between the queries, categories and query related Web pages as well as their categories based on their similarities.

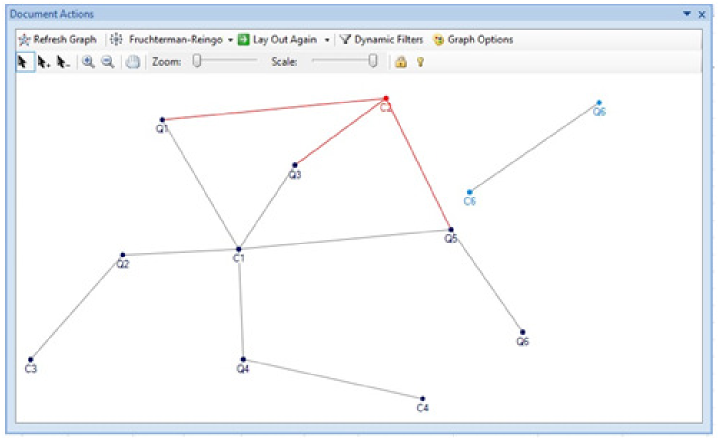


Figure 10. User query based Web page categorisation

C_1 = “data_mining/databases/software/computers”
C_2 = “web_mining/data_mining/databases/software/computers”
C_3 = web_content_mining/web_mining/data_mining/databases/software/computers”
C_4 = “books/object-oriented/software/computers”
C_5 = “social_networking/web_mining/data_mining/databases/software/computers”
C_6 = “tendulkar,_sachin/players/india/full_members/icc/cricket/sports”

Q_1- about data mining;Q_2- definition of Web mining
Q_3-about Web AND data mining; Q_4-Introduction about ”Web and data mining
Q_5- “Web and data mining”; Q_6- records of “Sachin Tendulkar”

4. DISCUSSION

Personalized Web search depends greatly on the user data. Hence data collection becomes an integral part of any personalized Web search research. In order to achieve the proper collection of user search data, a specialized Web browser is utilised [5] and incorporated with the user model. This browser [5] has the ability to collect both the explicit as well as implicit data from the Web user. In particular this browser tracks user Scrolling Speed and other additional parameters such as: Page Size, Click-Through.

4.1. Analysis of correlation model for user parameters

The user search parameters are Time-spent, Scrolling-Speed, Click-through, Page-size, Max-scroll, Min-scroll and Time per Page-size ratio. Among these, the most significant parameters are Time-spent, Average-scrolling speed and Number of queries and also their relationship as well as significance are examined using Pearson correlation with 524 users’ records. Time-spent and Average-scrolling speed are inversely proportional, Average-scrolling speed and number of queries are also inversely proportional and the number of query and Time-spent is directly proportional. In similar manner the other combinations of search parameter also estimated. The Pearson correlation is best suitable for measuring the inversely proportional relationship whereas Spearman Rank-Order Correlation and Kendall's tau-b are suitable for estimating the directly proportional relationship between the other user’s parameters.

4.2. Z-Sample Test based result analysis

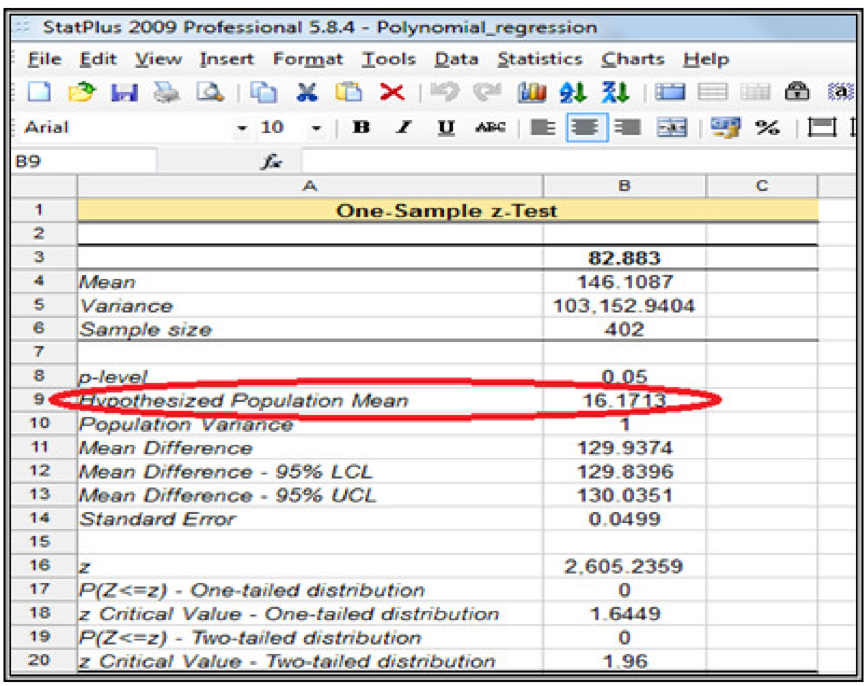
Hypothesis test on user data based on time-spent on a Web page is considered based on statistical measures (mean, median and mode, etc) such as H_0 and H_1 called Null hypothesis and Alternative hypothesis respectively.

H_0: A page is interest to a user if time-spent>= 15 seconds and
H_1: A page is of not-interest to a user if time-spent < 15 seconds.

Based on the above test results, the finalized user's minimum time-spent value will be 15 seconds. It also predicts that if it is greater than 15 seconds, the user is interested on the particular page; otherwise the user is not interested. This resultant value is tested by set of 10 user's data as shown in Figure 11. Similarly average scrolling speeds as well as other parameters like maximum-scrolling speed, minimum scrolling-speed, etc. are initialized using various statistical measures.

4.3. Weighed N-Gram Based User Content Categorization

Here the documents are represented by N-gram frequency profiles. It is set of N-grams arranged based on their frequency of that particular document. The grams are ordered according to their frequency values, in the form of chronological order. Table 11 provides single session query based user content categorization using N-gram approach with the help of ODP Taxonomy.



	A	B	C
1	One-Sample z-Test		
2			
3		82.883	
4	*Mean*	146.1087	
5	*Variance*	103,152.9404	
6	*Sample size*	402	
7			
8	*p-level*	0.05	
9	*Hypothesized Population Mean*	16.1713	
10	*Population Variance*	1	
11	*Mean Difference*	129.9374	
12	*Mean Difference - 95% LCL*	129.8396	
13	*Mean Difference - 95% UCL*	130.0351	
14	*Standard Error*	0.0499	
15			
16	*z*	2,605.2359	
17	*P(Z<=z) - One-tailed distribution*	0	
18	*z Critical Value - One-tailed distribution*	1.6449	
19	*P(Z<=z) - Two-tailed distribution*	0	
20	*z Critical Value - Two-tailed distribution*	1.96	

Figure 11.Snapshot representation of one-sample z-test result of user data

Table 11. Weighed N-gram based categorization of session-1 user content

Sl. no	Gram type	Sample	Frequency count	Category	Frequency count	Similarity measure
1	Trigram	about Web mining	6	data_mining/databases/software/computers	225	0.5814
		discover patterns	3	data_mining/databases/software/computers	225	0.5814
		data mining techniques	3	data_mining/reuse/software_engineering/software/computers	16	0.5643
		data mining methodologies	2	data_mining/reuse/software_engineering/software/computers	16	0.5643
2	bigram	Web mining	24	data_mining/databases/software/computers	225	0.5814
		about Web	6	online_publications/data_mining/databases/software/computers	18	0.5221
		data mining	5	data_mining/databases/software/computers	225	0.5814
3	unigram	Web	71	Web/servers/internet/software/computers	27	0.4872
		Data	16	2008/events/data_mining/databases/software/computers	20	0.4652
		Mining	63	drilling/tools_and_equipment/mining_and_drilling/business	62	0.4361

Here the N - value is restricted up to 3 (i.e., 1, 2 and 3). The N-value is obtained by several experiments. Each gram from the specific N-value based ordered list is compared with ODP taxonomy and matched categories are chosen and listed based on their frequencies and ranked. Finally top ranked category is assigned as resultant category based on the contents referred during their search process.

4.4. Fuzzification Model Selection

This user model analyzes the distribution of various user parameters from the user data and compares it with various fuzzification membership functions such as triangle and trapezoidal fuzzy membership function based graphs. The best suitable fuzzification membership is chosen for user parameters based on the resemblance with the data distribution. This may be the reason for the resemblance of the entire user parameters into either Triangular or Trapezoidal fuzzification functions. And the model is not approached from the server side collected data set. Further, this model will enhance with inclusion of other fuzzifier models for doing fuzzification on user parameters.

4.5. User Interest classification using fuzzy approach

This model considered a set of four interest class labels I, and the expert members defined a set of parameters and constraints P relevant to a particular topical user interest class. $I = \{i_1, i_2, i_3, i_4\}$ where i_1, i_2, i_3, i_4 are representing the four topical interest class label under consideration. $P = \{p_1, p_2, p_3 \ldots\ldots\ p_m\}$ where p_1, p_2, $p_3 \ldots\ldots\ p_m$ represents the parameters and constraints of a particular topical interest class of user.

This research work makes a mild attempt to implement the concept of fuzzy rule based user model that integrates fuzzy techniques in reducing the complexity of identification of users' interest based on their log information. This user model developed and incorporated fuzzy rules for estimating user's interest class.

Table 12. Fuzzy rule base for user interest classification of USER_1

Page no	Rule no	Time spent	Avg scroll	Click through	Page size	Max scroll	Min scroll	Time/ Page size ratio	Interest
P_1	2	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_2	12	LMD	LMD	MED	LOW	MED	MED	MED	WI
P_3	5	LMD	MED	MED	MED	MED	LOW	LMD	WI
P_4	10	LMD	HGH	MED	LMD	MED	LOW	HGH	NI
P_5	1	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_6	1	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_7	20	HGH	MED	LOW	HGH	HGH	LOW	LMD	WI
P_8	3	MED	LOW	LOW	LOW	LOW	LOW	MED	MI
P_9	1	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_{10}	2	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_{11}	6	MED	LMD	MED	HGH	MED	MED	LMD	MI
P_{12}	20	HGH	MED	LOW	HGH	HGH	LOW	LMD	WI
P_{13}	2	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_{14}	2	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI
P_{15}	4	HGH	LOW	LOW	LOW	LOW	LOW	HGH	SI
P_{16}	13	MED	LMD	MED	MED	MED	MED	LOW	MI
P_{17}	18	LMD	MED	LOW	LOW	MED	MED	MED	NI
P_{18}	16	LMD	MED	LOW	LOW	MED	LOW	LOW	NI
P_{19}	1	LOW	LOW	LOW	LOW	LOW	LOW	HGH	NI
P_{20}	2	LMD	LOW	LOW	LOW	LOW	LOW	LMD	WI

This model tested 524 records from 10 different users and assigned class labels using fuzzy approach. It also performed filtering of the Not-interest and Weak-Interest pages and recommended only Medium as well Strong-Interest pages. Finally it performed categorization of the users based on their class labels. Here the class labels are represented by numerical values such as: 1, 2, 3, and 4. the corresponding linguistic labels of each user search parameters are represented in the Table 12.

Table 13. Classification of user interest with Category labels

PAGE_NO	USER_NAME	CATEGORY (DOMAIN)	INTEREST CLASS LABEL
$\mathbf{P_1}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_2}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_3}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_4}$	USER_1	Web_Mining (Computers)	**NOT-INTERESTED**
$\mathbf{P_5}$	USER_1	Web_Mining (Computers)	**NOT-INTERESTED**
$\mathbf{P_6}$	USER_1	Web_Mining (Computers)	**NOT-INTERESTED**
$\mathbf{P_7}$	USER_1	data_mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_8}$	USER_1	data_mining (Computers)	**MEDIUMLY-INTERESTED**
$\mathbf{P_9}$	USER_1	Web_Mining	**NOT-**

		(Computers)	**INTERESTED**
$\mathbf{P_{10}}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_{11}}$	USER_1	Web_Mining (Computers)	**MEDIUMLY-INTERESTED**
$\mathbf{P_{12}}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_{13}}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_{14}}$	USER_1	Web_Mining (Computers)	**WEAKLY-INTERESTED**
$\mathbf{P_{15}}$	USER_1	Web_Mining (Computers)	**STRONGLY-INTERESTED**
$\mathbf{P_{16}}$	USER_1	Web_Mining (Computers)	**MEDIUMLY-INTERESTED**
$\mathbf{P_{17}}$	USER_1	Web_Content_Mining (Computers)	**NOT-INTERESTED**
$\mathbf{P_{18}}$	USER_1	data_mining (Computers)	**NOT-INTERESTED**
$\mathbf{P_{19}}$	USER_1	data_mining (Computers)	**NOT-INTERESTED**
$\mathbf{P_{20}}$	USER_1	data_mining (Computers)	**WEAKLY-INTERESTED**

Here NI, WI, MI, and SI represent Not-Interested, Weakly-Interested, Mediumly-Interested, and Strongly-Interested respectively. Similarly LOW, LMD, MED, and HGH are indicated as Low, Low-Medium, Medium, and High respectively. Table 13 provide the single user accessed 20 pages, its category labels and their corresponding interest class labels. Category label consists of both main (root) category represented within the bracket as well as its sub (ancestor) category of the page which belongs to is shown. i.e. USER_1 accessed Page No: 1 belongs to Main category: Computer and sub-category: Web_Mining.

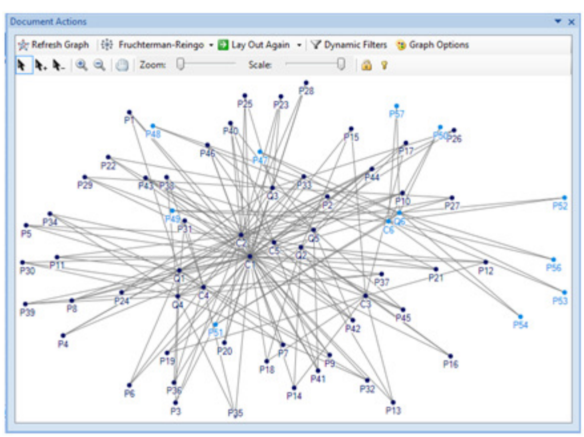


Figure 12. Category based USER_1 accessed Web pages and its query grouping

4.6. Grouping the User Web pages and Queries based on Categories

This model also performs the grouping of user accessed pages and its related queries based on their categories. This approach provides the inference of the users' major interest area in order to provide their anticipated based on their interest. Here the USER_1 accessed 58 Web pages are grouped and the result is shown in Figure 12. This user accessed Web pages mostly belonged to C_1. So it infers that the USER_1 is interested in "data_mining/databases/software/computers" category based contents shown in Figure 13.

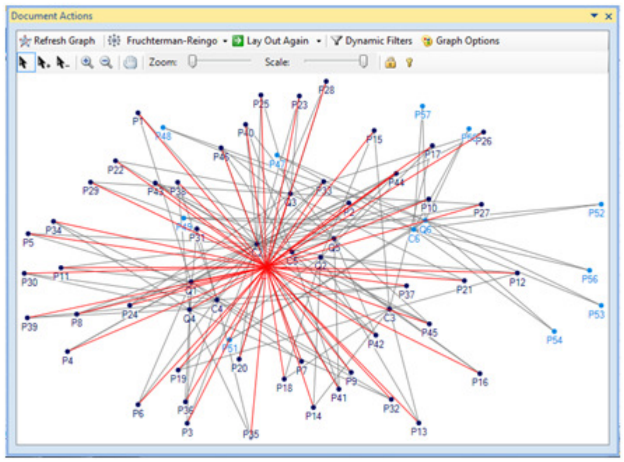


Figure 13. Category C_1 of USER_1 accessed Web pages and its query grouping

Table 14. Various Performance measures of Bayesian and Fuzzy Classifier

Sl. No	Method Approached	Total no. of records	No. of Correct classification	No. of In-Correct classification	Successful classification rate	Failure classification rate
1	Bayesian Classifier	524	398	126	75.9541 %	24.0450%
2	Fuzzy Classifier	524	470	54	89.6946 %	10.3053 %

In order to test the performance of this model, the Bayesian classifier is approached with the same data set (524 records from 10 different users). The results show that the Fuzzy rule based user model provides remarkable results than Bayesian classifier which is shown in Table 14. Compared with Bayesian classifier the proposed model provides less number of failure rate classifications and higher accuracy rate of successful classification.

5. CONCLUSION AND FUTURE WORK

The proposed User interest based classification in a personalized Web search using the fuzzy model delivered the acceptable rate of classification results. Heuristic based approach is incorporated in this model, so it enhances the accuracy of the classification of the user interest. The fuzzification functions are playing a major role for handling uncertainty data in such vague environment. Here the fuzzification is performed based on specific membership function and the selection of a specific membership function is based on the nature of the search data. Therefore caution has to be exercised during the membership function selection process. In future, the same model shall be drawn-out using Artificial Neural Networks (ANN). Genetic algorithm can be used for estimating errors and automatic fuzzifier selection process. In addition, the same user model will be extended to perform the prediction of user's interest fluctuation as well as interest decays.

REFERENCES

[1] Herrera, Mauro Rojas, Edleno Silva de Moura, Marco Cristo, Thomaz Philippe Silva, and Altigran Soares da Silva. "Exploring features for the automatic identification of user goals in web search." Information processing & management 46, no. 2 (2010): 131-142.

[2] Yu, Jie, Jie Gong, and Fangfang Liu. "Generation of Semantic Interactive Environment for Personalized Search." In Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on, pp. 443-448. IEEE, 2011.

[3] Kamalanathan, Selvakumar, and Sendhilkumar Selvaraju. "Collaborative Approaches for Personalized Web Search Using Fuzzy Neural Networks." InGlobal Trends in Information Systems and Software Applications, pp. 367-376. Springer Berlin Heidelberg, 2012.

[4] Delbru, Renaud, Stephane Campinas, and Giovanni Tummarello. "Searching web data: An entity retrieval and high-performance indexing model." Web Semantics: Science, Services and Agents on the World Wide Web 10 (2012): 33-58.

[5] Sendhilkumar, S., and T. V. Geetha. "Personalized ontology for web search personalization." In Proceedings of the 1st Bangalore annual Compute conference, p. 18. ACM, 2008.

[6] Dinh, Duy, and Lynda Tamine. "Towards a context sensitive approach to searching information based on domain specific knowledge sources." Web Semantics: Science, Services and Agents on the World Wide Web 12 (2012): 41-52.

[7] Kumar, AP Siva, P. Premchand, and A. Govardhan. "Query-based summarizer based on similarity of sentences and word frequency." International Journal of Data Mining & Knowledge Management Process 1, no. 3 (2011).

[8] Lin, Gu-Li, Hong Peng, Qian-Li Ma, Jia Wei, and Jiang-Wei Qin. "A Keyword Based Prototype for Web Search Result Diversification." Journal of Information Science & Engineering 28, no. 3 (2012).

[9] Chen, Tzung-Shi, Yen-Ssu Chou, and Tzung-Cheng Chen. "Mining user movement behavior patterns in a mobile service environment." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 42, no. 1 (2012): 87-101.

[10] Kim, Jin Young, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. "Characterizing web content, user interests, and search behavior by reading level and topic." In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 213-222. ACM, 2012.

[11] Rahul Verma, Kshitij Pathak, “Personalized Concept and Fuzzy Based User Profile from Search Engine Queries”. International Journal of Emerging Technology and Advanced Engineering , Volume 2, Issue 12, December 2012.

[12] Velásquez, Juan D. "Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments." Expert Systems with Applications 40, no. 13 (2013): 5228-5239.

[13] Glorio, Octavio, Jose-Norberto Mazón, Irene Garrigós, and Juan Trujillo. "A personalization process for spatial data warehouse development." Decision Support Systems 52, no. 4 (2012): 884-898.

[14] Guerbas, Abdelghani, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley, and Reda Alhajj. "Effective web log mining and online navigational pattern prediction." Knowledge-Based Systems 49 (2013): 50-62.

[15] Marin, Lucas, David Isern, Antonio Moreno, and Aida Valls. "On-line dynamic adaptation of fuzzy preferences." Information Sciences 220 (2013): 5-21.

[16] Castellano, Giovanna, Anna Maria Fanelli, and Maria Alessandra Torsello. "NEWER: A system for NEuro-fuzzy WEb Recommendation." Applied Soft Computing 11, no. 1 (2011): 793-806.

[17] Ahmad, Mohd Wazih, and M. A. Ansari. "A Survey: Soft Computing in Intelligent Information Retrieval Systems." In Computational Science and Its Applications (ICCSA), 2012 12th International Conference on, pp. 26-34. IEEE, 2012.

[18] Zadeh, Lotfi A. "Fuzzy sets." Information and control 8, no. 3 (1965): 338-353.

[19] Zadeh, Lotfi A. "Fuzzy logic= computing with words." Fuzzy Systems, IEEE Transactions on 4, no. 2 (1996): 103-111.

[20] Verbiest, Nele, Chris Cornelis, and Francisco Herrera. "FRPS: A fuzzy rough prototype selection method." Pattern Recognition 46, no. 10 (2013): 2770-2782.

[21] Brucker, Florian, Fernando Benites, and Elena Sapozhnikova. "Multi-label classification and extracting predicted class hierarchies." Pattern Recognition44, no. 3 (2011): 724-738.
[22] Arbelaitz, Olatz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, and Iñigo Perona. "Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it." Expert Systems with Applications 40, no. 18 (2013): 7478-7491.
[23] Thilagu, M., and R. Nadarajan. "INVESTIGATING SIGNIFICANT CHANGES IN USERS'INTEREST ON WEB TRAVERSAL PATTERNS." International Journal on on Cybernetics & Informatics (IJCI) Vol.2, No.4, August 2013.
[24] Serrano-Guerrero, Jesus, Francisco P. Romero, and Jose A. Olivas. "Hiperion: A fuzzy approach for recommending educational activities based on the acquisition of competences." Information Sciences 248 (2013): 114-129.
[25] Li, Sheng-Tun, and Fu-Ching Tsai. "A fuzzy conceptualization model for text mining with application in opinion polarity classification." Knowledge-Based Systems 39 (2013): 23-33.
[26] Pal, Sankar K., Saroj K. Meher, and Soumitra Dutta. "Class-dependent rough-fuzzy granular space, dispersion index and classification." Pattern Recognition45, no. 7 (2012): 2690-2707.
[27] 26.a. Mansoori, E. G., M. J. Zolghadri, and S. D. Katebi. "Using distribution of data to enhance performance of fuzzy classification systems." Iranian Journal of Fuzzy Systems 4, no. 1 (2007): 21-36.
[28] Ropero, Jorge, Ariel Gómez, Alejandro Carrasco, and Carlos León. "A Fuzzy Logic intelligent agent for Information Extraction: Introducing a new Fuzzy Logic-based term weighting scheme." Expert Systems with Applications 39, no. 4 (2012): 4567-4581.
[29] Kyoomarsi, F., H. Khosravi, E. Eslami, and M. Davoudi. "Extraction-based text summarization using fuzzy analysis." Iranian Journal of Fuzzy Systems 7, no. 3 (2010): 15-32.
[30] Li, Ming, Lu Liu, and Chuan-Bo Li. "An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems." Expert Systems with Applications 38, no. 7 (2011): 8586-8596.
[31] Siddiqui, Fasahat Ullah, Nor Ashidi Mat Isa, and Abid Yahya. "Outlier rejection fuzzy c-means (ORFCM) algorithm for image segmentation." Turkish Journal of Electrical Engineering & Computer Sciences 21, no. 6 (2013).
[32] George J.Klir, Bo Yuan. Fuzzy Sets and Fuzzy Logic Theory and Applications. Prentice Hall P T R Upper Saddle River, New Jersey: Prentice Hall, 1995.
[33] Timothy J. Ross. Fuzzy Logic with Engineering Applications. 3rd Edition, University of New Mexico, USA: John Wiley & Sons, Ltd, 2010.
[34] Qureshi, M. Rizwan Jameel, Mahaboob Sharief Shaik, and Nayyar Iqbal. "Using Fuzzy Logic to Evaluate Normalization Completeness for an Improved Database Design." International Journal of Information Technology and Computer Science (IJITCS) 4, no. 2 (2012): 48.
[35] Batra, Geetika, and Mahima Trivedi. "AFuzzy APPROACH FOR SOFTWARE EFFORT ESTIMATION." International Journal Cybernetics & Informatics (IJCI) Vol.2, No.1, February 2013.
[36] George J.Klir, Bo Yuan. Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems. World Scientific Publishing Co Pte Ltd, Singapore: WSP, 1996.
[37] Ganesh M. Introduction to Fuzzy Sets and Fuzzy Logic. PHI Learning Pvt. Ltd: PHI, 2006.
[38] Bede and Barnabas. Mathematics of Fuzzy Sets and Fuzzy Logic. Studies in Fuzziness and Soft Computing: Springer, 2013.
[39] Ren, Yan, Xiaodong Liu, and Jiannong Cao. "A parsimony fuzzy rule-based classifier using axiomatic fuzzy set theory and support vector machines."Information Sciences 181, no. 23 (2011): 5180-5193.
[40] Loia, Vincenzo, and Sabrina Senatore. "A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content." Knowledge-Based Systems(2013).
[41] Alcalá-Fdez, Jesús, Rafael Alcalá, María José Gacto, and Francisco Herrera. "Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms." Fuzzy Sets and Systems 160, no. 7 (2009): 905-921.
[42] Weber, Siegfried. "A general concept of fuzzy connectives, negations and implications based on< i> t</i>-norms and< i> t</i>-conorms." Fuzzy sets and systems 11, no. 1 (1983): 103-113.
[43] Srikant, Ramakrishnan, Sugato Basu, Ni Wang, and Daryl Pregibon. "User browsing models: relevance versus examination." In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 223-232. ACM, 2010.
[44] Zadeh, Lotfi A. "A fuzzy-set-theoretic interpretation of linguistic hedges." (1972): 4-34.

[45] Samuel, O. W., M. O. Omisore, and B. A. Ojokoh. "A web based decision support system driven by fuzzy logic for the diagnosis of typhoid fever." Expert Systems with Applications 40, no. 10 (2013): 4164-4171.
[46] Djam, X. Y., G. M. Wajiga, Y. H. Kimbi, and N. V. Blamah. "A Fuzzy Expert System for the Management of Malaria." International Journal of Pure & Applied Sciences & Technology 5, no. 2 (2011).
[47] Chang, Robin LP, and Theodosios Pavlidis. "Fuzzy decision tree algorithms."Systems, Man and Cybernetics, IEEE Transactions on 7, no. 1 (1977): 28-35.
[48] Dong, Ming, and Ravi Kothari. "Look-ahead based fuzzy decision tree induction." Fuzzy Systems, IEEE Transactions on 9, no. 3 (2001): 461-468.
[49] Mitra, Sushmita, Kishori M. Konwar, and Sankar K. Pal. "Fuzzy decision tree, linguistic rules and fuzzy knowledge-based network: generation and evaluation."Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 32, no. 4 (2002): 328-339.
[50] LEVASHENKO, Vitaly, and Penka MARTINCOVÁ. "Fuzzy Decision Tree for Parallel Processing Support." Journal of Information, Control and Management Systems 3, no. 1 (2005).
[51] MATIAŠKO, Karol, Ján BOHÁČIK, Vitaly LEVASHENKO, and Štefan KOVALÍK. "Learning fuzzy rules from fuzzy decision trees." Journal of Information, Control and Management Systems 4, no. 2 (2006).
[52] Pulkkinen, Pietari, and Hannu Koivisto. "Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms." International Journal of Approximate Reasoning 48, no. 2 (2008): 526-543.
[53] Abolfazl Kazemia, Elahe Mehrzadeganb. A New Algorithm for Optimization of Fuzzy Decision Tree in Data Mining. Journal of Optimization in Industrial Engineering 2011; 7: 29-35.
[54] Daveedu Raju Adidela, Jaya Suma. G, Lavanya Devi. G. Construction of Fuzzy Decision Tree using Expectation Maximization Algorithm. Journal of Computer Science and Management Research 2012; 1: 416-424.