Indian Journal of Science and Technology

Supplementary Article

Web Portal Visits Patterns Predicted by Intuitionistic Fuzzy Approach

A. Kumaravel[1*] and R. Udayakumar[2]

[1]Professor and Dean, School of Computing, Bharath University, Selaiyur, Chennai-600073, India; drkumaravel@gmail.com

[2]Associate Professor, School of Computing Sciences, Bharath University, Selaiyur, Chennai-600073, India; udayakumar@bharathuniv.ac.in

Abstract

Web mining is applied to reflect the importance of webpages and to predict the web domain visits of various users. An IF-inference system is developed for this purpose. This paper presents basic notions of web mining with fuzzy inference systems based on the Takagi-Sugeno fuzzy model. At first, it scans the sequences of visits in database once, finds all the weighted frequent items, and produces a sparse matrix for that sequence item set from which basic statistics like moving averages and exponentially smoothing averages are calculated. Then we construct a IF-inference system using intuitionistic approach for predicting the web domain visits. We obtain the results of such system.

Keywords: Web Mining, IF- inference System, Web Domain Visits, Fuzzy Logic, Prediction, Sparse Matrix, Moving Averages, Exponential Smoothing.

1. Introduction

Web servers are with development of e-business and e-government, as a source of information for web mining. Web mining [1, 2] includes far more than common statistic overviews displaying for example instantaneous number of visitors or most often visited web pages. Web mining try to analyze the following sample queries like, How do various visitors behave?; What are typical sequences of pages walk-throughs?; How long do visitors stay at pages?; How and from where do they leave pages?

These are data mainly about user behavior on the Internet [13]. They can be created as a track of user on web-page or application. That means that all steps and attributes of the user visits. These data are kept in log files. Based on information stated we can identify and filter auto-generated visits of full-text seekers, which are quite common and distort user behavior statistics.

In general, classification and prediction [3] can be realized by FISs. Based on general FIS structure, we can design two basic types - Mamdani type and Takagi-Sugeno type [4]. Both the FISs types differ by means of obtaining the output. Output formulation results in different if-then rules construction. These rules can be designed by user or the user can obtain them through extraction from historical data. Fuzzification of input variables and application of operators in if-then rules are the same in both FISs types.

In this paper we present problem statement with the aim to describe the various users of the msnbc.com web log file and possibilities of its pre-processing of various moving averages of the users in various web pages of site. Further, the paper includes the comparison between prediction results obtained with the FIS characterized in membership functions μ and the FIS characterized in non-membership functions ν, by the IF-inference system. [9]

* *Corresponding author:*
A. Kumaravel (drkumaravel@gmail.com)

The paper is organized with section 2 on problem description, section 3 discussing the design of IF-Inference system, section 4 describing the experiments and results, finally the conclusion.

2. Problem Statement

The datasets for these experiments are from Internet Information Server (IIS) logs for msnbc.com [11].

2.1 Dataset

2.1.1 Dataset Information

The page visits of msn.com for a day by various users. The web pages are homepage, news, on-air, local, opinion, tech, misc., weather, health, living, business, summary, BBS, travel, sports, msn-news, and msn-sports.

Number of users: 989818
Average number of visit per user: 5.7
Number of URLs per category: 10 to 5000

2.1.2 Attribute Information

The categories are associated in-order with an integer starting with "1". For example, 1 indicates homepage, 2 is news, and 3 is tech. Each row below "% Sequences:" describes the hits--in order--of a single user. For example, the user hits news page twice, and the next user hits homepage once.

2.1.3 Pre-processing of data

The data represents the sequence of visits by users which are ordered by sparse matrix [12]. A matrix mostly occupied with zeros is called sparse matrix. Preprocessing of data is realized by means of simple mathematic-statistic methods [5]. Data smoothing algorithm is used to remove noise data in the process of preprocessing data. The generally stated method is represented by a function, in each user, takes certain real value [6] dependent on empirically determined values. The individual methods are stated in Table 1. The preprocessed visit rate of the msn web presentation, msnbc.com, is stated in Figure 1 to Figure 5.

The basic statistics of the users is stated in Table 2. The general formulation of the model prediction of msnbc.com web visit rate can be stated in this manner $y=f(x_1,x_2,\ldots.x_m)$, m=5 where y is daily web visits, x_1 is SMA, x_2 is CMA, x_3 is WMA, x_4 is SES and x_5 is DES of various visitors.

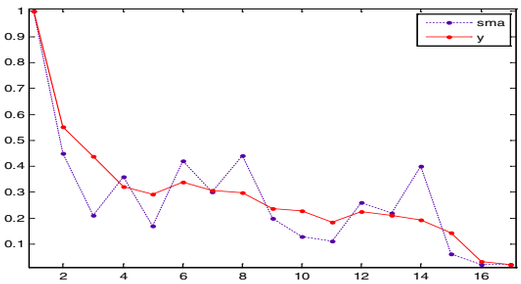

Figure 1. The Pre-Processing of msn page visits by SMA.

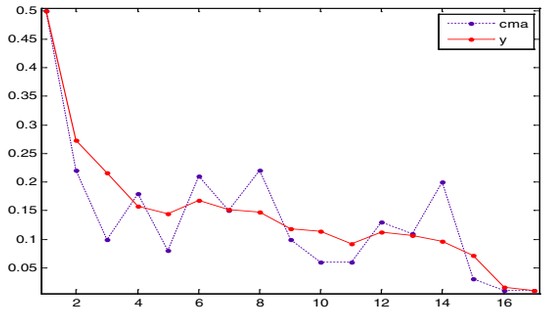

Figure 2. The Pre-Processing of msn page visits by CMA.

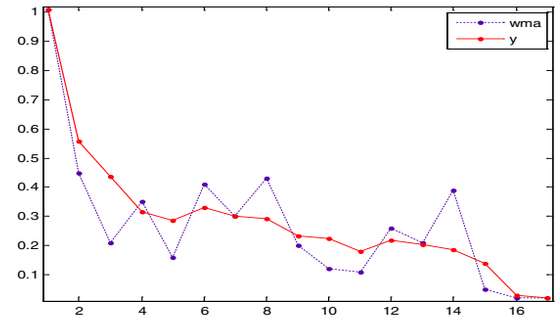

Figure 3. The Pre-Processing of msn page visits by WMA.

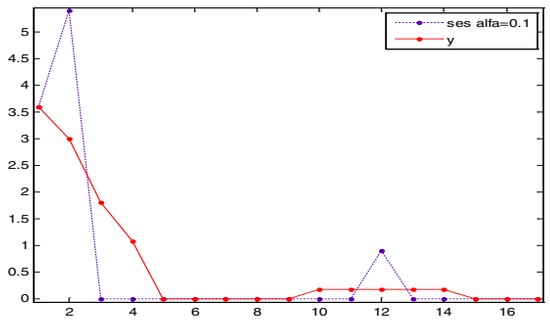

Figure 4. The Pre-Processing of msn page visits by SES.

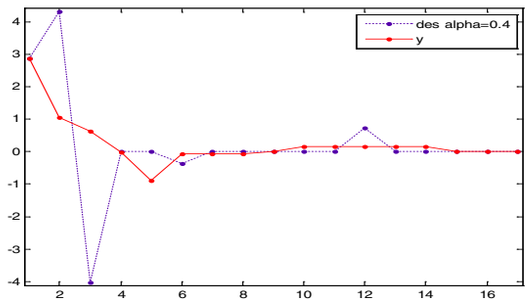

Figure 5. The Pre-Processing of msn page visits by DES.

Table 1. Methods of Pre-Processing of Visitors

Page no.	SMA	CMA	WMA	SES	DES	no. of visitor
				α=0.1	α=0.4	
1	1	0.5	1.01	3.6	2.88	16381
2	0.45	0.22	0.45	5.4	4.32	7312
3	0.21	0.1	0.21	0	-4.032	3420
4	0.36	0.18	0.35	0	0	5827
5	0.17	0.08	0.16	0	0	2715
6	0.42	0.21	0.41	0	-0.36	6810
7	0.3	0.15	0.3	0	0	4958
8	0.44	0.22	0.43	0	0	7263
9	0.2	0.1	0.2	0	0	3304
10	0.13	0.06	0.12	0	0	2119
11	0.11	0.06	0.11	0	0	1805
12	0.26	0.13	0.26	0.9	0.72	4249
13	0.22	0.11	0.21	0	0	3569
14	0.4	0.2	0.39	0	0	6487
15	0.06	0.03	0.05	0	0	930
16	0.02	0.01	0.02	0	0	266
17	0.02	0.01	0.02	0	0	304

Table 2. Basic statistics of web msnbc.com visits

Parameter	Mean	Stddev	Min	Max
SMA	0.28	0.23	0.02	1
CMA	0.14	0.12	0.01	0.5
WMA	0.28	0.24	0.02	1.01
SES	0.58	1.52	0	5.4
DES	1	2.77	-0.81	9.72

3. IF-Inference Systems Design

Let a set X be a non-empty fixed set. An IF-set A in X is an object.

$$A = \{<x, \mu_A(x), \nu_A(x) > \mid x \in X\},$$

where the function μ_A:X→ [0,1] defines the degree of membership function $\mu_A(x)$ and the function ν_A:X→ [0,1] defines the degree of non-membership function $\nu_A(x)$, respectively, of the element x∈X to the set A, which is a subset of X, more over for every x∈X, $0 \leq \mu_A(x) + \nu_A(x) \leq 1$, x∈X must hold. The amount

$$\pi_A(x) = 1 - (\mu_A(x) + \nu_A(x))$$

is called the hesitation part, which may cater to either membership value or non-membership value, or both. For each IF-set in X, we will call $\pi_A(x) = 1 - (\mu_A(x) + \nu_A(x))$ as the intuitionistic index of the element x in set A. It is obvious that $0 \leq \pi_A(x) \leq 1$ for each x ∈ X. The value indicates a measure of non-determinacy.

The existing general IF-system defined in [7]. Then it is possible to define its output

$$y^{\eta} \text{ as } y^{\eta} = (1 - \pi_A(x)) \times y^{\mu} + \pi_A(x) \times y^{\nu},$$

Where

y^{μ}- output of the FIS-μ using the membership function $\mu_A(x)$,

y^{ν}- output of the FIS-ν using the non-membership function $\nu_A(x)$.

Then, based on this equation, the possible strategy of the IF-inference system of Takagi-Sugeno type presented in Figure 6.

The IF-inference system is designed in this way, it holds:

- If intuitionistic index πA(x) is 0, then the output of IF-inference system $y^{\eta} = (1 - \pi_A(x)) \times y^{\mu}$ (Takagi-Sugeno FIS is characterized by membership function μ).
- If intuitionistic index $\pi_A(x)$ is 1, then the output of IF-inference system $y^{\eta} = \pi_A(x) \times y^{\nu}$ (Takagi-Sugeno FIS, is characterized by non-membership function ν).

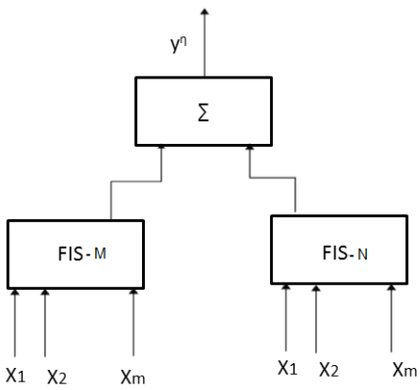

Figure 6. IF inference system.

- If intuitionistic index $\pi_A(x)$ is between 0 to 1, and the output of IF-inference system $y^{\eta} = (1- \pi_A(x)) \times y^{\mu} + \pi_A(x) \times y^{\nu}$ it is characterized by membership function μ and non-membership function ν.

4. Experimental Analysis and Results

Construct the model for prediction of msnbc.com web visit rate is formulated as follows, $y=f(x_1,x_2, \ldots ,x_m)$, m=5, where y is daily web msnbc.com visits, x_1 is SMA, x_2 is CMA, x_3 is WMA, x_4 is SES and x_5 is DES a Figure 6 it is possible to design an input membership function μ for FIS-μ and input non-membership functions ν for FIS-ν as follows. Input language variable SMA is represented by means of four membership functions. They are bell membership functions. Individual membership functions are described by means of language variable value: low SMA, med low SMA, med high SMA and high SMA. Membership functions of language variable SMA for model of upce.cz web visit rate prediction are shown in Figure 7 and non-membership functions are shown in Figure 8. Other membership and non-membership language variable functions are designed analogically (CMA, WMA, SES, and DES). Membership function μ and non-membership function ν, and if-then rules were designed using subtractive clustering algorithm [8].

To be specific, two if-then rules are designed for FIS-μ and FIS-ν correspondingly. The output level y^k of each of the k-th if-then rule R^k is weighted. The final outputs y^{μ} and y^{ν} of the FIS-μ and FIS-ν are the weighted averages of all the if-then rule R^k outputs y^k, k=1, 2… N. The output of IF-inference system is represented by the predicted value y^{η} is presented in Figure 9 to Figure 11.

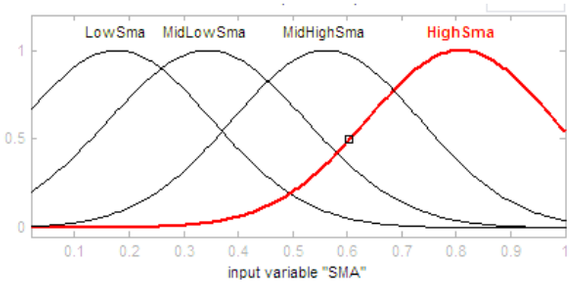

Figure 7. Input Membership function μ for SMA of FIS-μ

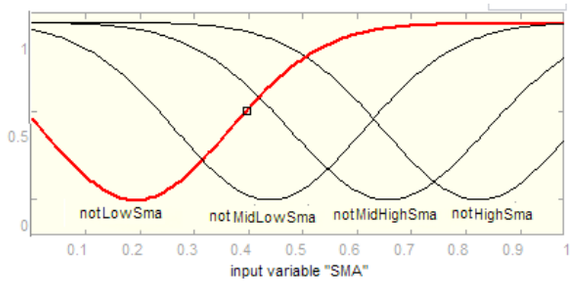

Figure 8. Input Non-Membership function ν for SMA of FIS-ν

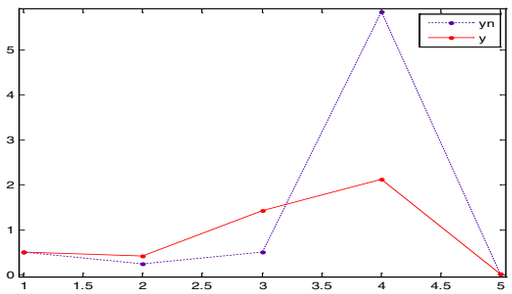

Figure 9. The result of web msnbc.com visits prediction y.

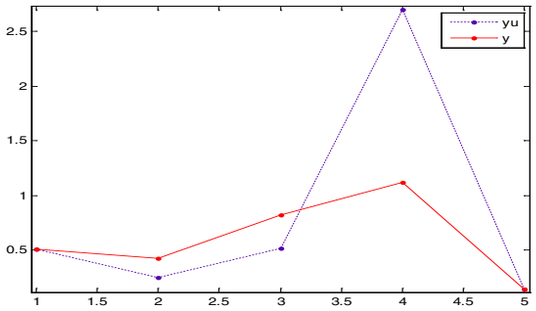

Figure 10. The result of web msnbc.com visits of y^{μ}.

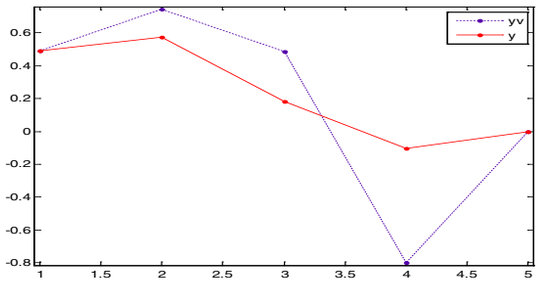

Figure 11. The result of web msnbc.com visits of y^{ν}.

The result from the fact that the predicted value y^{η} is weighted that the output value of y^{μ} . Therefore, non-membership functions ν limited in this way are not suitable for the used data.

5. Conclusions

The design based on IF-sets as a model for web mining is modeled in the paper as they allow processing. IF-sets can be viewed in the context as a proper tool for representing both membership and non-membership of data sequence. The IF-inference system FIS-η defined this way works more effective than the standard of Takagi-Sugeno type FIS-μ as it provides stronger possibility to accommodate in definite information and better model imperfect fact and imprecise knowledge.

In this study we present a novel approach to web domain visit prediction based on the extension of Takagi-Sugeno type FIS-μ which is characterized by membership function μ with Takagi-Sugeno type, FIS-ν which is characterized by non-membership function ν. The designed IF-inference system represents an efficient tool for modeling of web domain visits, which is demonstrated on the prediction of the msn web page visit rate prediction. Data for web mining needs were obtained from log files of msnbc.com. The model design was carried out in Mat lab [10] in MS Windows XP operation system.

6. Acknowledgements

The authors would like to thank the management of Bharath University for the support and encouragement for this research work.

7. References

1. Cooley R, Mobasher B et al. (1997). Web mining: information and pattern discovery on the world wide web, Proceedings of the 9th IEEE International Conference on Tools With Artificial Intelligence, (ICTAI '97), Newport Beach, CA.
2. Zaine O, and Han J (1998). WebML: Querying the world wide web for resources and knowledge, Proceedings of the International workshop on web information and data management, WIDM´98, Bethesda.
3. Kuncheva L I (2000). Fuzzy classifier design, A Springer Verlag Company, Germany.
4. Bandemer H, and Gottwald S (1995). Fuzzy sets fuzzy logic, fuzzy methods, John Wiley and Sons Inc., New York.
5. Olej V (2003). Modeling of economics processes by computational intelligence, Hradec Kralove, (in Slovak).
6. Trešl J (1999). Statistical methods and capital market, 1st edition, Prague, (in Czech).
7. Montiel O et al. (2008). Mediative fuzzy logic: A new approach for contradictory knowledge management, Soft Computing, vol 20, No.3, 251–256.
8. Guillaume S (2001). Designing fuzzy inference systems from Data: An Interpretability-oriented review, IEEE Transactions on Fuzzy Systems, vol 9, 426–442.
9. Available From: http://www.wseas.us/e-library/transactions/computers/2010/88-321.pdf
10. Fuzzy Logic Toolbox™ User's Guide© Copyright 1995–2012, The Math Works, Inc.
11. Available From: http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data.
12. Availbale From: http://www.mathworks.in/help/matlab/ref/sparse.html
13. Kumaravel A, and Pradeepa R (2012). On constructing regular expression of web page traversals for efficient filtering, IEEE Conference Publications, 156-160.