Modelling of Web Domain Visits by IF-Inference System

Vladimír OLEJ, Petr HÁJEK, Jana FILIPOVÁ
Institute of System Engineering and Informatics
Faculty of Economics and Administration
University of Pardubice, Studentská 84, 53210 Pardubice, Czech Republic
vladimir.olej@upce.cz, petr.hajek@upce.cz, jana.filipova@upce.cz

Abstract: - This paper presents basic notions of web mining and fuzzy inference systems based on the Takagi-Sugeno fuzzy model. On the basis of this fuzzy inference system and IF-sets introduced by K.T. Atanassov novel IF-inference systems can be designed. Thus, an IF-inference system is developed for time series prediction. In the next part of the paper we describe web domain visits prediction by IF-inference systems and the analysis of the results.

Key-Words: - Web mining, IF- inference system, fuzzy inference system, web domain visits, prediction, time series.

1 Introduction

Web serves, with development of e-business and e-government, as a source of information for web mining. Web mining [1,2,3] includes far more than common statistic overviews displaying for example instantaneous number of visitors or most often visited web pages. Web mining analyses:

- Where the visitors come from.
- How do various visitors behave.
- What are typical sequences of pages walkthroughs.
- Which sequence leads to a purchase or reservation.
- How long visitors stay at pages.
- How and from where they leave pages.

These are data mainly about user behaviour on the Internet. They can be created as a track of user on web page or application. That means that all steps and attributes of the user are recorded. 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 behaviour statistics.

Currently there are a number of methods which serve for modelling of data obtained by web mining. These methods can be divided to methods of modelling with learning without a teacher and learning with a teacher.

The most common method of data modelling with learning without a teacher for exploring user profiles from log files is clustering [4]. Cluster algorithms used in the area of web mining are K-means [5], Rough K-means or C-means, Fuzzy c trimmed methods and Fuzzy-c medians for fuzzy clustering [6,7], or Ant-Based Clustering (ABC) [8]. Other method of learning without teacher are associative and fuzzy associative rules [9,10].

Within data modelling methods with learning with a teacher, fuzzy logic, neural networks and machine learning are used for classification and data prediction from log files. In [11] semi-supervised learning in combination with Rough sets and Kohonen's self-organizing feature maps [12] are used for classification of web pages. Extraction of redundant attributes is solved by transformation of Singular Value Decomposition [13,14]. Prediction of trend of number of visits is stated in [15]. For cluster analysis we selected ABC and to determine trend we used genetic linear programming. For prediction in the area of web mining we use fuzzy inference system (FIS) Takagi-Sugeno [16,17], Support Vector Machines [18,19] and feed-forward neural networks [20,21].

In general, classification and prediction [16,22] can be realized by FISs. Based on general FIS structure, we can design two basic types - Mamdani type and Takagi-Sugeno type [16,22]. Both the FISs types differ by means of obtaining the output. Different output formulation results in different if-then rules construction. These rules can be designed by user (based on his experience), 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.

At this time there are several generalizations of fuzzy set theory for various objectives [23 to 31]. Intuitionistic fuzzy sets (IF-sets) theory represents one of the generalizations, the notion introduced by K.T. Atanassov [31,32]. The concept of IF-sets can be viewed as an alternative approach to define a fuzzy set in cases where available information is not sufficient for the definition of an imprecise concept by means of a conventional fuzzy set. In this paper we will present IF-sets as a tool for reasoning in the presence of imperfect fact and imprecise knowledge. The IF-sets are for example also suitable for the time series prediction of web domain visits as they provide a good description of object attributes by means of membership functions μ and non-membership functions ν. They also present a strong possibility to express uncertainty.

In the paper we present problem formulation with the aim to describe the time series of the University of Pardubice web presentation visits and possibilities of its pre-processing and basic notions of Takagi-Sugeno type FIS for time series prediction. Based on the FIS defined this way and the basic notions of IF-sets, we define a novel IF-inference system. Further, the paper includes the comparison of the prediction results obtained by the FIS characterized by membership functions μ, by the FIS characterized by non-membership functions ν, and by the IF-inference system. The comparison of designed systems is realized for the stated time series of the University of Pardubice, the Czech Republic (web upce.cz) web presentation visits from September 2008 to June 2009.

2 Problem Formulation

The data for prediction of the time series of the University of Pardubice, the Czech Republic (web upce.cz) web presentation visits over given time period was obtained from Google Analytics. This web mining tool, by means of Java Script code implemented to web presentation, offers a wide spectrum of operation characteristics (web metrics). Metrics provided by Google Analytics can be divided to following sections:

- Visits;
- Sources of access;
- Contents;
- Conversion.

In section 'Visits' we can monitor for example number of visitors, number of visits and number of pages viewed, ratio of new and returned visitors. Indicator geolocation, i.e. which country visitors are most often from, is needed to be known for example because of language mutations. In order to predict visit rate of the University of Pardubice, the Czech Republic (web upce.cz) web presentation we need to monitor indicator number of visits. The number of visits is a basic indicator which displays the number of visits within given time period. A 'Visit' is comprehended as an unrepeated combination of IP address and cookies. A sub-metrics is absolutely unique visit, which is defined by unrepeatable IP address and cookies in given time period. Basic information obtainable from Google Analytics about web upce.cz during May 2009, is following:

- Total visit rate during given month drops. Clear trend is obvious there, when Monday has the highest visit rate which keeps dropping until the end of week; Saturday has the lowest visit rate.
- Average number of pages visited is more than three.
- Visitor stays on certain page five and half a minute in average.
- Bounce rate is approximately 60 %.
- Visitors come mainly directly, which is good.
- Among the most favourite pages there is the main page, then the pages of Faculty of Economic and Administration and Faculty of Philosophy.

Measuring of visit rate of the University of Pardubice web presentation, web.upce.cz, takes place in regular equidistant time periods. It represents time series which is given by succession of material and area comparable observations, which are ordered by time [33]. Pre-processing of data is realized by means of simple mathematic-statistic methods [34]. The generally stated method is represented by a function, which for each period of time, takes certain real value [35] dependant on empirically determined values. An overview of individual methods is stated in Table 1. Primary and pre-processed time series of visit rate of the University of Pardubice web presentation, web.upce.cz, are stated in Fig. 1 to Fig. 5.

Table 1 Overview of methods of pre-processing of time series

Methods	Characteristics
Simple moving average (SMA)	5, 7, 9 days
Central moving average (CMA)	4, 6, 8 days
Moving median (MM)	5, 7, 9 days
Simple exponential smoothing (SES)	pro α = 0.1 a α = 0.2
Double exponential smoothing (DES)	pro α = 0.7 a α = 0.9

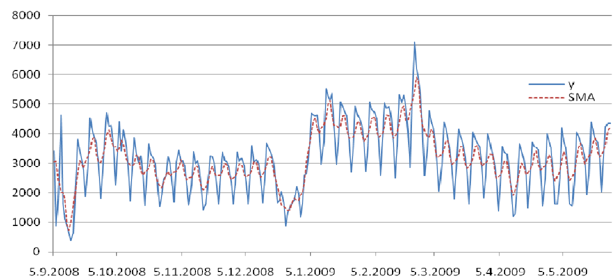

Fig. 1 The pre-processing of web upce.cz visits by SMA

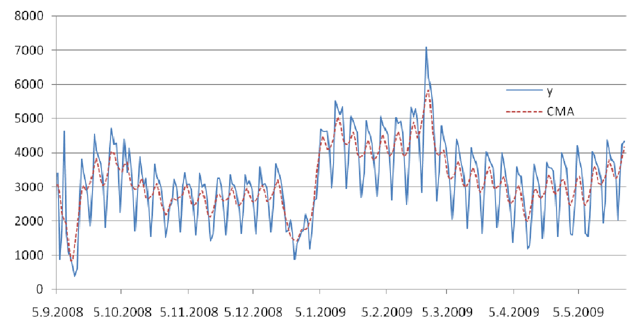

Fig. 2 The pre-processing of web upce.cz visits by CMA

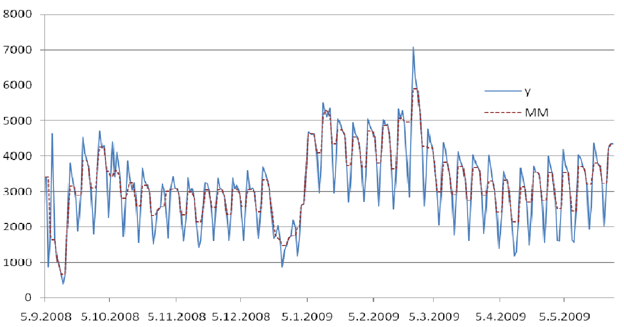

Fig. 3 The pre-processing of web upce.cz visits by MM

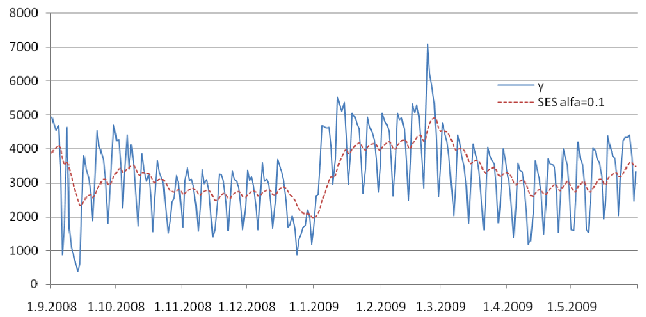

Fig. 4 The pre-processing of web upce.cz visits by SES

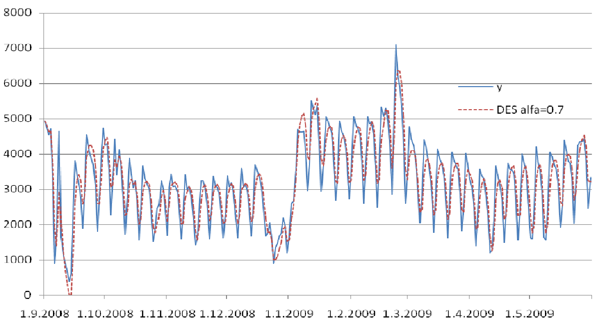

Fig. 5 The pre-processing of web upce.cz visits by DES

Basic statistics of the time series are stated in Table 2 and Table 3, where y represents the primary time series. The general formulation of the model of prediction of upce.cz web visit rate can be stated in this manner $y=f(x_1^t,x_2^t, \dots ,x_m^t)$, m=5, where y is daily web upce.cz visits in t+1, x_1^t is SMA, x_2^t is CMA, x_3^t is MM, x_4^t is SES and x_5^t is DES at time t.

Table 2 Basic statistics of web upce.cz visits

Parameter	Mean	Standard deviation	Min	Max
SMA	3173	845	739	5916
CMA	3175	835	840	5839
MM	3263	948	663	5897
SES	3192	637	1945	4902
DES	3159	1071	-11	6361
y	3168	1128	386	7081

Table 3 Basic statistics of web upce.cz visits after standardization

Parameter	Mean	Standard deviation	Min	Max
SMA	0	1	-2.88	3.25
CMA	0	1	-2.8	3.19
MM	0	1	-2.74	2.78
SES	0	1	-1.96	2.69
DES	0	1	-2.96	2.99
y	0	1	-2.47	3.47

3 Basic Notions of the Fuzzy Inference Systems

General structure of FIS [16,34,36] is shown in Fig. 6. It contains a fuzzification process of input variables by membership functions μ, the design of the base of if-then rules or automatic if-then rules extraction from input data, operators (AND,OR,NOT) application in if-then rules, implication and aggregation within these rules, and the process of defuzzification of gained values to crisp values. In the process of defuzzification, standardization of inputs and their transformation to the domain of the values of membership functions μ takes place. There is no general method for designing shape, number and parameters of input and output membership functions μ.

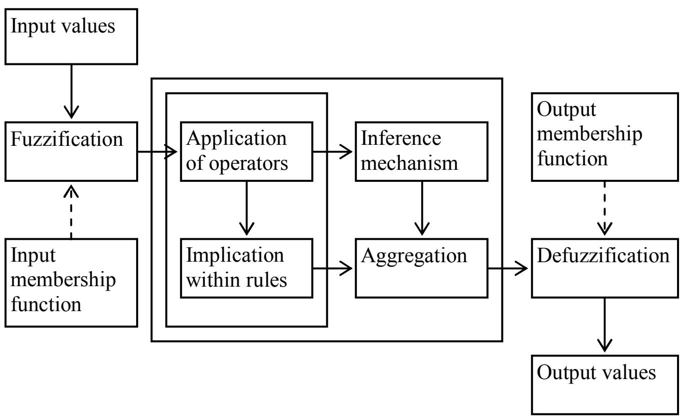

Fig. 6 General FIS structure

The input to fuzzification process is a crisp value given by the universum (reference set). The output of fuzzification process is the membership function μ value. The design of the base of if-then rules can be realized by extraction of if-then rules from historical data, provided that they are available. In [16,34,36] there are mentioned optimization methods of the number of if-then rules. The best results are obtained by the method based on the optimization of the form of output membership functions μ based on input data. The membership functions μ of output variables are assigned to input data so that the membership functions μ of the i-th input and output are equal. If input data do not form a convex function, a new membership function μ is created for the data. Each additional membership function μ causes addition of a new if-then rule to base rules. Advantage of this method is that it generates only a small number of if-then rules. Their number reduces the computing intensity of FIS, which enables us to design FIS capable of working in real time.

Another option of deriving base rules from historical data is the so-called Adaptive Neuro-Fuzzy Inference System method [16]. The principle of this method is a neuro-adaptive learning process, based on which it is possible to derive membership function μ parameters from historical data and to extract base rules. This method helps to assign input data to output data, while in the process of learning the parameters of individual membership functions μ are gradually changed in order to characterize the relations between the range of input variables and the range of output variables in the best way. In the process of learning, it is possible to use the method of regressive error spread, the combination of the method of regressive error spread and minimal squares method, or evolutionary stochastic optimization algorithms. The base of rules consists of if-then rules. These rules are used for creating predicate clauses representing the base of FIS.

Let $x_1,x_2, \dots ,x_j, \dots ,x_m$ be input variables defined on reference sets $X_1,X_2, \dots ,X_j, \dots ,X_m$ and let y be an output variable defined on reference set Y. Then FIS has m input variables and one output variable. Further, each set X_j, j=1,2, ... ,m, can be divided into i=1,2, ... ,n fuzzy sets

$$\mu_{j,1}(x),\mu_{j,2}(x), \dots ,\mu_{j,i}(x), \dots ,\mu_{j,n}(x). \quad (1)$$

Individual fuzzy sets represent assignment of linguistic variables values, which are related to sets X_j. Then the k-th if-then rule R_k in FIS Takagi-Sugeno type can be written down in the following form

R_k: if x_1 is $A_{1,i(1,k)}$ AND x_2 is $A_{2,i(2,k)}$ AND ... AND x_j is $A_{j,i(j,k)}$ AND ... AND x_m is $A_{m,i(m,k)}$ then y=h, (2)
j=1,2, ... ,m; i=1,2, ... ,n,

where $A_{1,i(1,k)},A_{2,i(2,k)}, \dots ,A_{j,i(j,k)}, \dots ,A_{m,i(m,k)}$, represent the values of linguistic variable and h is constant. Fuzzy inference system composed of if-then rules defined by relation (2) is called a zero order Takagi-Sugeno type FIS. If the k-th if-then rule R_k in Takagi-Sugeno type FIS is in form

R_k: if x_1 is $A_{1,i(1,k)}$ AND x_2 is $A_{2,i(2,k)}$ AND ... AND x_j is $A_{j,i(j,k)}$ AND ... AND x_m is $A_{m,i(m,k)}$
then $y=f(x_1,x_2, \dots ,x_m)$, (3)

where $f(x_1,x_2, \dots ,x_m)$ is a linear function, then the FIS consisting of if-then rules R_k, k=1,2, … ,N, defined by relation (3) is called a first order Takagi-Sugeno type FIS. In case that $f(x_1,x_2, \dots ,x_m)$ is a polynomial function, it is called a second order Takagi-Sugeno type FIS.

Operator AND between elements of two fuzzy sets (A_1 AND A_2) can be generalized by t-norm [37]. It implies from the interpretation of conditions satisfied by t-norm, i.e. the intersection of two fuzzy sets can only lead to an exclusion of elements (values of membership function μ will be zero) or to a preservation of current membership function μ value. Further, decreasing the value of membership function μ in set A_1 or A_2 cannot lead to an increase of the value of membership function μ of intersection ($A_1 \cap A_2$), t-norm is not dependent on the order of combined fuzzy sets and it allows combining arbitrary pairs of elements of fuzzy sets in arbitrary order. Analogous to operator AND, operator OR between elements of two fuzzy sets (A_1 OR A_2) can be generalized by s-norm [37].

The Takagi-Sugeno type FIS was designed in order to achieve higher computational effectiveness. This is possible as the defuzzification of outputs is not necessary. Its advantage lies also in involving the functional dependencies of output variable on input variables. The output level y_k of each the k-th if-then rule R_k is weighted by the $w_k=\mu(x_1)$ AND $\mu(x_2)$ AND … AND $\mu(x_m)$. The final output y of the Takagi-Sugeno type FIS is the weighted average of all N if-then rule R_k outputs y_k, k=1,2, … ,N, computed as follows

$$y = \frac{\sum_{k=1}^{N} y_k \times w_k}{\sum_{k=1}^{N} w_k}. \quad (4)$$

4 IF-Inference Systems Design

The concept of IF-sets is the generalization of the concept of fuzzy sets, the notion introduced by L.A. Zadeh [37]. The theory of IF-sets is well suited to deal with vagueness. Recently, in this context, IF-sets have been used for intuitionistic classification [38,39,40, 41,42] and prediction [43] models which can accommodate imprecise information.

Let a set X be a non-empty fixed set. An IF-set A in X is an object having the form [31,32]

$$A = \{ \langle x, \mu_A(x), \nu_A(x) \rangle \mid x \in X\}, \quad (5)$$

where the function $\mu_A:X\rightarrow [0,1]$ defines the degree of membership function $\mu_A(x)$ and the function $\nu_A:X\rightarrow [0,1]$ defines the degree of non-membership function $\nu_A(x)$, respectively, of the element $x\in X$ to the set A, which is a subset of X, and $A\subset X$, respectively; moreover for every $x\in X$, $0 \leq \mu_A(x) + \nu_A(x) \leq 1$, $\forall x\in X$ must hold. The amount

$$\pi_A(x) = 1 - (\mu_A(x) + \nu_A(x)) \quad (6)$$

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 a hesitancy degree of x to A. It is obvious that $0 \leq \pi_A(x) \leq 1$ for each $x\in X$. The value denotes a measure of non-determinancy. The intuitionistic indices $\pi_A(x)$ are such that the larger $\pi_A(x)$ the higher a hesitation margin of the decision maker. Intuitionistic indices allow us to calculate the best final results (and the worst one) we can expect in a process leading to a final optimal decision.

Next we define an accuracy function H to evaluate the degree of accuracy of IF-set by the form $H(A) = \mu_A(x) + \nu_A(x)$, where $H(A)\in[0,1]$. From the definition H, it can be also expressed as follows $H(A) = \mu_A(x) + \nu_A(x) = 1-\pi_A(x)$. The larger value of H(A), the more the degree of accuracy of the IF-set A.

If A and B are two IF-sets of the set X, then

- $A\cap B = \{\langle x, \min(\mu_A(x), \mu_B(x)), \max(\nu_A(x), \nu_B(x))\rangle \mid x\in X\}$,
- $A\cup B = \{\langle x, \max(\mu_A(x), \mu_B(x)), \min(\nu_A(x), \nu_B(x))\rangle \mid x\in X\}$,
- $A\subset B$ iff $\forall x\in X$, $(\mu_A(x) \leq \mu_B(x))$ and $(\nu_A(x) \geq \nu_B(x))$,
- $A\supset B$ iff $B\subset A$,
- $A=B$ iff $\forall x\in X$, $(\mu_A(x) = \mu_B(x)$ and $\nu_A(x) = \nu_B(x))$,
- $\overline{A}=\{\langle x, \nu_A(x), \mu_A(x)\rangle \mid x\in X\}$.

Let there exists a general IF-system defined in [44]. Then it is possible to define its output y^η as

$$y^\eta = (1- \pi_A(x)) \times y^\mu + \pi_A(x) \times y^\nu, \quad (7)$$

where y^μ is the output of the FIS^μ using the membership function $\mu_A(x)$, y^ν is the output of the FIS^ν using the non-membership function $\nu_A(x)$.

Then, based on equation (7), it is possible to design the IF-inference system of Takagi-Sugeno type presented in Fig. 7.

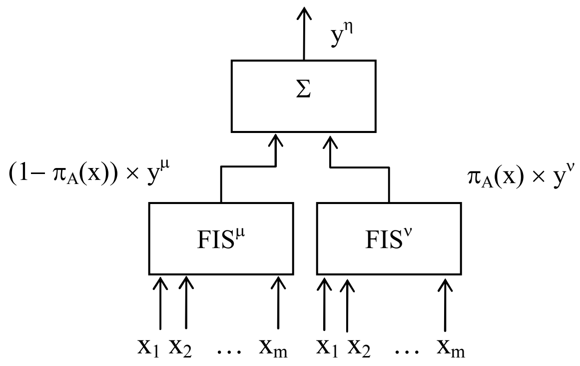

Fig. 7 IF-inference system

For the IF-inference system designed in this way, it holds:

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

Let $x_1,x_2, \ldots ,x_j, \ldots ,x_m$ be input variables FIS^η defined on reference sets $X_1,X_2, \ldots ,X_j, \ldots ,X_m$ and let y^η be an output variable defined on reference set Y. Then FIS^η has m input variables $x_1,x_2, \ldots ,x_j, \ldots ,x_m$ and one output variable y^η, where $\eta=\mu$ are membership functions ($\eta=\nu$ are non-membership functions). Further, each set X_j, $j=1,2, \ldots ,m$, can be divided into $i=1,2, \ldots ,n$ fuzzy sets which are represented by following way

$$\eta_{j,1}(x),\eta_{j,2}(x), \ldots ,\eta_{j,i}(x), \ldots ,\eta_{j,n}(x). \quad (8)$$

Individual fuzzy sets, where $\eta=\mu$ are membership functions ($\eta=\nu$ are non-membership functions) represent mapping of linguistic variables values, which are related to sets X_j. Then the k-th if-then rule R_k in FIS^η can be defined as follows

R_k: if x_1 is $A_{1,i(1,k)}{}^\eta$ AND x_2 is $A_{2,i(2,k)}{}^\eta$ AND ... AND x_j is $A_{j,i(j,k)}{}^\eta$ AND ... AND x_m is $A_{m,i(m,k)}{}^\eta$ then $y^\eta=h$,
or $y^\eta=f(x_1,x_2, \ldots ,x_m)$, $j=1,2, \ldots ,m$; $i=1,2, \ldots ,n$, (9)

where $A_{1,i(1,k)}{}^\eta,A_{2,i(2,k)}{}^\eta, \ldots ,A_{j,i(j,k)}{}^\eta, \ldots ,A_{m,i(m,k)}{}^\eta$ represent the values of linguistic variable for FIS^μ and FIS^ν, h is constant, $f(x_1,x_2, \ldots ,x_m)$ is a linear or polynomial function. The output y^{μ} of FIS^{μ} (the output y^{ν} of FIS^{ν}) is defined in the same way as presented in equation (4).

5 Inference Mechanism of IF-Inference Systems

The results of the designed IF-inference system of Takagi-Sugeno type can be interpreted by means of the following rationale. Fuzzy sets, the notion introduced by L.A. Zadeh, are represented by membership function $\mu_A(x)$ degree, that is

$$A=\{\langle x, \mu_A(x)\rangle \mid x\in X\}. \quad (10)$$

Fuzzy sets have associated a non-membership function $\nu_A(x)$ degree

$$A=\{\langle x, \mu_A(x),\nu_A(x)\rangle \mid x\in X\}= =\{\langle x,\mu_A(x),1-\mu_A(x)\rangle \mid x\in X\}. \quad (11)$$

Since $\mu_A(x) + \nu_A(x) = \mu_A(x) + 1- \mu_A(x) = 1$, fuzzy sets are considered as a particular case IF-sets. Let be given an automorphism [45] of the unit interval, i.e. every function $\psi : [0,1] \rightarrow [0,1]$, that is continuous and strictly increasing such that $\psi(0)=0$ and $\psi(1)=1$. Further, let be given a function n: $[0,1] \rightarrow [0,1]$ in such a way that it holds $n(0) =1$ and $n(1) =0$. It is called a strong negation and it is always strictly decreasing, continuous and involutive. Then, as proved by [45], n: $[0,1] \rightarrow [0,1]$ is a strong negation if and only if there exists an automorphism ψ of the unit interval such that $n(x) = \psi^{-1} (1 - \psi(x))$. Let L^* be a set for which

$$L^*=\{(x,y)|(x,y)\in[0,1] \times [0,1] \text{ and } x+y\leq 1\} \quad (12)$$

and the elements $0_{L^*}=(0,1)$ and $1_{L^*}=(1,0)$. Then $\forall((x,y),(z,t))\in L^*$ it holds:

- $(x,y) \leq_{L^*}(z,t)$ iff $x\leq z$ and $y\geq t$. This relation is transitive, reflexive and antisymmetric.
- $(x,y) = (z,t)$ iff $(x,y) \leq_{L^*}(z,t)$ and $(z,t) \leq_{L^*}(x,y)$.
- $(x,y)\preccurlyeq (z,t)$ iff $x\leq z$ and $y\leq t$.

The designed IF-inference system of Takagi-Sugeno type works with the inference mechanism, based on Atanassov's intuitionistic fuzzy t-norm and t-conorm, by means of t-norm and t-conorm [46] on interval [0,1]. A function T: $(L^*)^2 \rightarrow L^*$ is called Atanassov's intuitionistic fuzzy t-norm if it is commutative, associative, and increasing in both arguments with respect to the order $\leq_{L^*}$ and with neutral element 1_{L^*}. Similarly, a function S: $(L^*)^2 \rightarrow L^*$ is called Atanassov's intuitionistic fuzzy t-conorm if it is commutative, associative, and increasing with neutral element 0_{L^*}. Atanassov's intuitionistic fuzzy t-norm T is called t-representable if and only if there exists a t-norm T and t-conorm S on interval [0,1] such that $\forall(x,y),(z,t)\in L^*$ it holds.

$$T((x,y),(z,t)) =(T(x,z),S(y,t)) \in L^*. \quad (13)$$

If T=min on interval [0,1] then $\min((x,y),(z,t))= = (\min(x,z),\max(y,t))$. Accordingly, Atanassov's intuitionistic fuzzy t-conorm S can be defined, and it is called t-representable if and only if there exist a t-norm T and t-conorm S on interval [0,1] such that $\forall(x,y),(z,t)\in L^*$ it holds

$$S((x,y),(z,t)) =(S(x,z),T(y,t)) \in L^* \quad (14)$$

If S=max on interval [0,1] then $\max((x,y),(z,t))= = (\max(x,z),\min(y,t))$. An Atanassov's intuitionistic fuzzy negation [46] is a function n: $L^* \rightarrow L^*$ such that it is decreasing with respect to the $\leq_{L^*}$ and $n(0_{L^*})= 1_{L^*}$.and $n(1_{L^*})= 0_{L^*}$. Then if $\forall((x,y),(z,t))\in L^*$ $n(n((x,y))) = (x,y)$, it is said that n is involutive. Function n: $L^* \rightarrow L^*$ is an involutive Atanassov's intuitionistic fuzzy negation if there exists an involutive fuzzy negation n such that

$$n((x,y)) =(n(1-y), 1- n(x)). \quad (15)$$

Based on presented facts a generalized Atanassov's IF-index can be defined as a function π_G: $L^* \rightarrow [0,1]$ associated with the strong negation n if it satisfies the following conditions:

- $\pi_G(x,y)= 1$ iff $x=0$ and $y=0$.
- $\pi_G(x,y)= 0$ iff $x+y=1$.
- If $(z,t)\preccurlyeq(x,y)$, then $\pi_G(x,y) \leq \pi_G(z,t)$.
- $\pi_G(x,y) = \pi_G(n((x,y)))$ $\forall(x,y)\in L^*$ such that n is generated from an involutive negation n.

6 Modelling and Analysis of the Results

Based on the fact that the model for prediction of upce.cz web visit rate is formulated as follows $y=f(x_1^t,x_2^t, \dots ,x_m^t)$, m=5, where y is daily web upce.cz visits in t+1, x_1^t is SMA, x_2^t is CMA, x_3^t is MM, x_4^t is SES and x_5^t is DES at time t a Fig. 7 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 presented in Fig. 8 and non-membership functions are presented in Fig. 9. Other membership and non-membership language variable functions are designed analogically (CMA, MM, SES, DES). Membership function μ and non-membership function ν, and if-then rules were designed using substractive clustering algorithm [47].

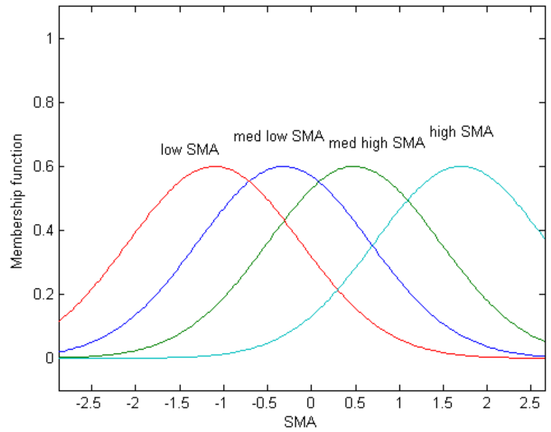

Fig. 8 Input membership functions μ for SMA of FIS$^{\mu}$

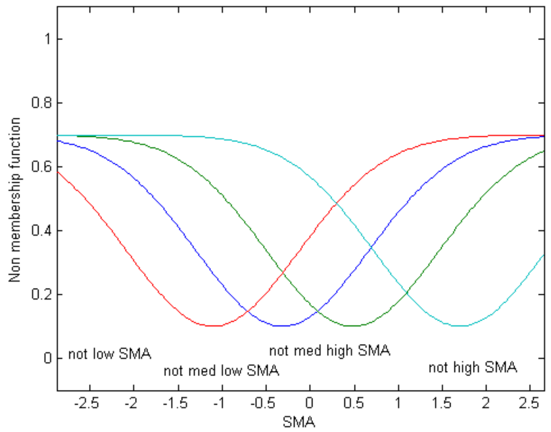

Fig. 9 Input non-membership functions ν for SMA of FIS$^{\nu}$

To be specific, two if-then rules are designed for FIS$^{\mu}$ and FIS$^{\nu}$ respectively. 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$^{\mu}$ and FIS$^{\nu}$ 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} in time t+1. The results of visit rate of the web upce.cz visits on training data O_{train} and testing data O_{test} for μ_{max}=0.6 for intuitionistic index π=0.3 are presented in Fig. 10 to Fig. 13, where μ_{max} represents the maximum value of input membership functions μ.

The Table 4 and Table 5 (on trainig data O_{train}) and Table 6 and Table 7 (on testing data O_{test}) shows the quality of web upce.cz visits prediction represented by Root Mean Squared Error (RMSE) for different values of μ_{max} and different values of intuitionistic index π.

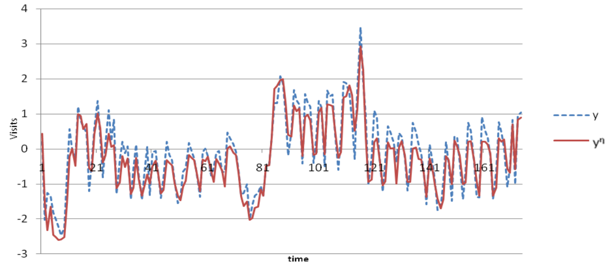

Fig. 10 The results of web upce.cz visits prediction for training data O_{train}

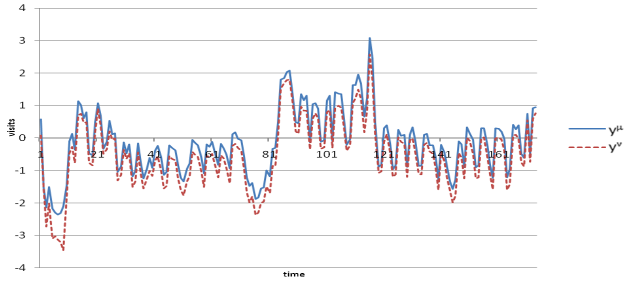

Fig. 11 The results of of web upce.cz visits prediction for training data O_{train}

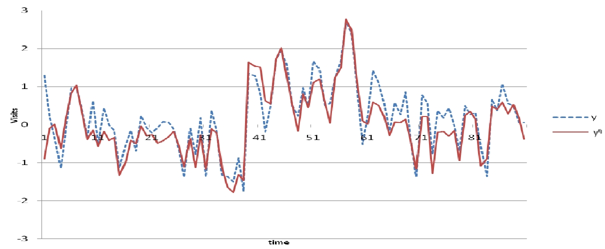

Fig. 12 The results of web upce.cz visits prediction for testing data O_{test}

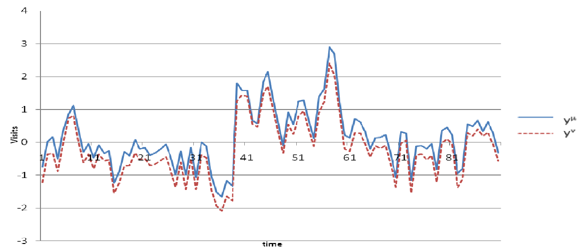

Fig. 13 The results of web upce.cz visits prediction for testing data O_{test}

The results show that the RMSE$^{\mu}$ is for FIS$^{\mu}$ constant. The size of μ_{max} does not affect the resulting error of FIS$^{\mu}$. This results from the fact that the output y^{μ} is a weighted average of outputs y_k from the single if-then rules R_k. Relative weights w_k remains the same for different values of μ_{max}. Maximum RMSE is obtained for the FIS$^{\nu}$ and the FIS$^{\eta}$, for which ν_{min}=0 holds, i.e. $\mu_{max}+\pi=1$. Therefore, non-membership functions ν limited in this way are not suitable for the used data.

Table 4 RMSE on training data O_{train} for different values of μ_{max} and intuitionistic index π

π=0.1					
μ_{max}	0.5	0.6	0.7	0.8	0.9
$RMSE^{\mu}$	0.221	0.221	0.221	0.221	0.221
$RMSE^{\nu}$	0.307	0.319	0.336	0.362	0.410
$RMSE^{\eta}$	0.224	0.225	0.225	0.226	0.228
π=0.2					
μ_{max}	0.5	0.6	0.7	0.8	
$RMSE^{\mu}$	0.221	0.221	0.221	0.221	
$RMSE^{\nu}$	0.314	0.331	0.358	0.410	
$RMSE^{\eta}$	0.230	0.231	0.234	0.239	
π=0.3					
μ_{max}	0.5	0.6	0.7		
$RMSE^{\mu}$	0.221	0.221	0.221		
$RMSE^{\nu}$	0.325	0.353	0.410		
$RMSE^{\eta}$	0.238	0.243	0.253		
π=0.4					
μ_{max}	0.5	0.6			
$RMSE^{\mu}$	0.221	0.221			
$RMSE^{\nu}$	0.347	0.410			
$RMSE^{\eta}$	0.253	0.270			

Table 5 RMSE on training data O_{train} for different values of μ_{max} and intuitionistic index π

π=0.5					
μ_{max}	0.1	0.2	0.3	0.4	0.5
$RMSE^{\mu}$	0.221	0.221	0.221	0.221	0.221
$RMSE^{\nu}$	0.284	0.294	0.311	0.340	0.410
$RMSE^{\eta}$	0.243	0.246	0.252	0.263	0.290
π=0.6					
μ_{max}	0.1	0.2	0.3	0.4	
$RMSE^{\mu}$	0.221	0.221	0.221	0.221	
$RMSE^{\nu}$	0.286	0.302	0.331	0.410	
$RMSE^{\eta}$	0.251	0.258	0.272	0.312	
π=0.7					
μ_{max}	0.1	0.2	0.3		
$RMSE^{\mu}$	0.221	0.221	0.221		
$RMSE^{\nu}$	0.291	0.319	0.410		
$RMSE^{\eta}$	0.261	0.278	0.335		
π=0.8					
μ_{max}	0.1	0.2			
$RMSE^{\mu}$	0.221	0.221			
$RMSE^{\nu}$	0.302	0.410			
$RMSE^{\eta}$	0.278	0.359			
π=0.9					
μ_{max}	0.1				
$RMSE^{\mu}$	0.221				
$RMSE^{\nu}$	0.410				
$RMSE^{\eta}$	0.384				

Table 6 RMSE on testing data O_{test} for different values of μ_{max} and intuitionistic index π

π=0.1					
μ_{max}	0.5	0.6	0.7	0.8	0.9
$RMSE^{\mu}$	0.237	0.237	0.237	0.237	0.237
$RMSE^{\nu}$	0.297	0.309	0.325	0.352	0.407
$RMSE^{\eta}$	0.239	0.240	0.240	0.242	0.245
π=0.2					
μ_{max}	0.5	0.6	0.7	0.8	
$RMSE^{\mu}$	0.237	0.237	0.237	0.237	
$RMSE^{\nu}$	0.304	0.320	0.348	0.407	
$RMSE^{\eta}$	0.243	0.245	0.248	0.256	
π=0.3					
μ_{max}	0.5	0.6	0.7		
$RMSE^{\mu}$	0.237	0.237	0.237		
$RMSE^{\nu}$	0.314	0.343	0.407		
$RMSE^{\eta}$	0.250	0.255	0.269		
π=0.4					
μ_{max}	0.5	0.6			
$RMSE^{\mu}$	0.237	0.237			
$RMSE^{\nu}$	0.337	0.407			
$RMSE^{\eta}$	0.263	0.285			

Table 7 RMSE on testing data O_{test} for different values of μ_{max} and intuitionistic index π

π=0.5					
μ_{max}	0.1	0.2	0.3	0.4	0.5
$RMSE^{\mu}$	0.237	0.237	0.237	0.237	0.237
$RMSE^{\nu}$	0.278	0.286	0.301	0.329	0.407
$RMSE^{\eta}$	0.250	0.253	0.259	0.270	0.302
π=0.6					
μ_{max}	0.1	0.2	0.3	0.4	
$RMSE^{\mu}$	0.237	0.237	0.237	0.237	
$RMSE^{\nu}$	0.279	0.292	0.320	0.407	
$RMSE^{\eta}$	0.255	0.262	0.275	0.321	
π=0.7					
μ_{max}	0.1	0.2	0.3		
$RMSE^{\mu}$	0.237	0.237	0.237		
$RMSE^{\nu}$	0.282	0.308	0.407		
$RMSE^{\eta}$	0.262	0.278	0.341		
π=0.8					
μ_{max}	0.1	0.2			
$RMSE^{\mu}$	0.237	0.237			
$RMSE^{\nu}$	0.292	0.407			
$RMSE^{\eta}$	0.276	0.362			
π=0.9					
μ_{max}	0.1				
$RMSE^{\mu}$	0.237				
$RMSE^{\nu}$	0.407				
$RMSE^{\eta}$	0.384				

The RMSE for FIS^{ν} and FIS^{η} increases with a higher intuitionistic index π, i.e. with a higher uncertainty. The results, however, show that it is possible to achieve a relatively low level of RMSE (on training and testing data) even in the cases where it is not possible to determine the membership functions μ and non-membership functions ν unambiguously (i.e. for high intuitionistic index π).

7 Conclusion

The model based on IF-sets as a model for web mining is designed in the paper as they allow processing uncertainty and the expert knowledge. IF-sets can be viewed in the context as a proper tool for representing hesitancy concerning both membership and non-membership of an element to a set. 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 imprecise information and better model imperfect fact and imprecise knowledge.

In this study we present a novel approach to times series prediction [43,48] 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 central point in the design of IF-inference system lies in the intuitionistic index π expressing the level of uncertainty. The designed IF-inference system represents an efficient tool for modelling of time series, which is demonstrated on the prediction of the University of Pardubice visit rate prediction. Data for web mining needs were obtained from log files of upce.cz web. The model design was carried out in Matlab in MS Windows XP operation system.

ACKNOWLEDGMENTS

This work was supported by the scientific research projects Ministry of Environment, the Czech Republic under Grant No: SP/4i2/60/07 with title Indicators for Valuation and Modelling of Interactions among Environment, Economics and Social Relations, Czech Science Foundation under Grant No: 402/09/P090 with title Modelling of Municipal Finance by Computational Intelligence Methods, and Czech Science Foundation under Grant No: 402/08/0849 with title Model of Sustainable Regional Development Management.

References:

[1] R. Cooley, B. Mobasher, J. Srivistava, Web Mining: Information and Pattern Discovery on the World Wide Web, *Proc. of the 9th IEEE International Conference on Tools With Artificial Intelligence, (ICTAI '97)*, Newport Beach, CA, 1997.

[2] O. Zaine, J. Han, WebML: Querzing the World Wide Web for Resources and Knowledge, *Proc. of the Int. Workshop on Web Information and Data Management, WIDM´98*, Bethesda, 1998.

[3] J. Srivistava, P. Desikan, V. Kumar, Web Mining-Accomplishments, [online], 2004, [cit. 2009-05-04].

[4] O. Nasraoui, *World Wide Web Personalization, Encyclopedia of Data Mining and Data Warehousing*, J. Wang, (Ed.), Idea Group, 2005.

[5] J. Mac Queen, Some Methods for Classification and Analysis of Multivariate Observations, *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1, Berkeley, University of California Press, 1967.

[6] R. Krishnapuram, A. Joshi, L. Yi, A Fuzzy Relative of the K-medoids Algorithm with Application to Document and Snippet Clustering, *Proc. of IEEE International Conference on Fuzzy Systems*, Korea, 1999, pp.1281-1286.

[7] W. Pedrycz, *Conditional Fuzzy C-means*, Pattern Recognition Lett, Vol.17, 1996, pp. 625-632.

[8] N. Monmarché, Algorithmes de Fourmist Artificielles: Applications a la Classification et a l´optimisation, Ph.D. Thesis, Tours, Univeristé Francois Rabelais, France, 2000.

[9] B. Mobasher, et al., Web Mining: Pattern Discovery from World Wide Web Transactions, [online], 1997, [cit. 2009-05-05].

[10] M. N. Garofalakis, et al., Data Mining and the Web: Past, Present and Future, *Workshop on Web Information and Data Managament*, 1999.

[11] B. Yang, W. Song, Z. Xu, Semi-Supervised Learning for Web Text Clustering, [online], 2000, [cit. 2009-05-05].

[12] T. Kohonen, *Self-Organizing Maps*, 3rd edition, New York, Springer-Verlag Berlin Heidelberg, 2001.

[13] S. A. Alvarez, T. Kawato, C. Ruiz, Mining over Loosely Coupled Data Sources Using Neural Experts, [online], 2003, [cit. 2008-02-20], pp.1-7.

[14] A. Asuncion, D. J. Newman, UCI Machine Learning Repository, Irvine, CA, University of California, School of Information and Computer Science, [online], 2007, [cit. 2008-02-20].

[15] A. Abraham, V. Ramos, Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming, [online], 2003, [cit. 2009-05-07].

[16] L. I. Kuncheva, *Fuzzy Classifier Design*, A Springer Verlag Company, Germany, 2000.

[17] G. J. Deboeck, *Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial*

Markets, New York, John Wiley and Sons Inc., pp. 189-263, 1994.

[18] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd edition, Prentice-Hall Inc., New Jersey, 1999.

[19] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag New York, 1995.

[20] V. Kvasnička, et al., *Introduction to Neural Networks Theory*, Bratislava, IRIS, 1997, p.285, (in Slovak).

[21] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Learning Representations by Back-propagating Errors*, Nature, 1986, pp.533-536.

[22] H. Bandemer, S. Gottwald, *Fuzzy Sets, Fuzzy Logic, Fuzzy Methods*, John Wiley and Sons Inc., New York, 1995.

[23] G. Wang, Y. He, Intuitionistic Fuzzy Sets and L-Fuzzy Sets, *Fuzzy Sets and Systems*, Vol.110, 2000, pp.271-274.

[24] P. Burillo, H. Bustince, Construction Theorems for Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, Vol. 84,1996, pp.271-281.

[25] P. Burillo, H. Bustince, Entropy on Intuitionistic Fuzzy Sets and on Interval-Valued Fuzzy Sets, *Fuzzy Sets and Systems*, Vol.78, 1996, pp.305-316.

[26] D. Coker, Fuzzy Rough Sets are Intuitionistic L-Fuzzy Sets, *Fuzzy Sets and Systems*, Vol.96, 1998, pp.381-383.

[27] G. Deschrijver, E. E. Kerre, On the Relationship between some Extension of Fuzzy Set Theory, *Fuzzy Sets and Systems*, Vol.132/2, 2004, pp.227-235.

[28] H. Bustince, P. Burillo, Vague Sets are Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, Vol.79, 1996, pp.403-405.

[29] D. Dubois, S. Gottwald, P. Hajek, J. Kacprzyk, H. Prade, Terminological Difficulties in Fuzzy Set Theory-the Case of Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, Vol.156, 2005, pp.485-491.

[30] K. T. Atanassov, Answer to D. Dubois, S. Gottwald, P. Hajek, J. Kacprzyk, H. Prade's Terminological Difficulties in Fuzzy Set Theory-the case of Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, Vol.156, 2005, pp.496-499.

[31] K. T. Atanassov, Intuitionistic Fuzzy Sets, *Fuzzy Sets and Systems*, Vol. 20,1986, pp.87-96.

[32] K. T. Atanassov, *Intuitionistic Fuzzy Sets*, Springer, Berlin Heidelberg, Germany, 1999.

[33] R. Hindls, S. Hronová, L. Novák, *Methods of Statistical Analysis for Economics*, 2nd edition, Prague, Management Press, 2000, (in Czech).

[34] V. Olej, *Modelling of Economics Processes by Computational Intelligence*, Hradec Králové, 2003, (in Slovak).

[35] J. Trešl, *Statistical Methods and Capital Market*, 1st edition, Prague, 1999, (in Czech).

[36] W. Pedrycz, *Fuzzy Control and Fuzzy Systems*, 2nd ed., John Wiley and Sons Inc., New York, 1993.

[37] L. A. Zadeh, Fuzzy Sets, *Inform. and Control*, Vol. 8, 1965, pp.338-353.

[38] V. Olej, P. Hájek, *Air Quality Modelling by Kohonen's self-organizing feature maps and Intuitionistic Fuzzy Sets*, A. P. Pobil, (Ed.), Calgary, Canada, ACTA Press, 2008, pp.22-27.

[39] P. Hájek, V. Olej, Air Quality Indices and their Modelling by Hierarchical Fuzzy Inference Systems, *WSEAS Transactions on Environment and Development*, WSEAS Press, 2010, Issue 10, Vol.5, 2009, pp.661-672.

[40] P. Hájek, V. Olej, An Air Pollution Index based on Hierarchical Fuzzy Inference Systems. *Proc. of the 5th WSEAS International Conference on Energy, Environment, Ecosystems and Sustainable Development*, (EEESD'09), N. Mastorakis, and all. (Eds.), Vouliagmeni Beach, Greece, 2009, pp.83-88.

[41] V. Olej, P. Hájek, Modelling Municipal Creditworthiness by Hierarchical Structures of Fuzzy Inference Systems. *Proc. of the 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases,* (AIKED '07), Corfu, Greece, 2007, pp.234-242.

[42] V. Olej, P. Hájek, Intuitionistic Hierarchical Fuzzy Inference Systems Design for Air Quality Modelling, *Proc. of the 5th WSEAS International Conference on Energy, Environment, Ecosystems and Sustainable Development*, (EEESD'09), N. Mastorakis, and all. (Eds.), Vouliagmeni Beach, Greece, 2009, pp.89-94.

[43] V. Olej, P. Hájek, IF-Inference Systems Design for Prediction of Ozone Time Series: The Case of Pardubice Micro-Region, *Proc of the 20th International Conference on Artificial Neural Networks*, ICANN 10, Thessaloniki, Greece, 2010, (in Press).

[44] O. Montiel, et al., Mediative Fuzzy Logic: A new Approach for Contradictory Knowledge Management, *Soft Computing*, Vol.20, No.3, 2008, pp.251-256.

[45] E. Barrenechea, Generalized Atanassov's Intuitionistic Fuzzy Index, Construction Method, *IFSA-EUSFLAT*, Lisbon, 2009, pp.478-482.

[46] G. Deschrijver, C. Cornelis, E. Kerre, On the Representation of Intuitionistic Fuzzy t-norm and t-conorm, *IEEE Transactions on Fuzzy Systems*, Vol.12, 2004, pp.45-61.

[47] S. Guillaume, Designing Fuzzy Inference Systems from Data: An Interpretability-Oriented Review,

IEEE Transactions on Fuzzy Systems, Vol.9, 2001, pp.426-442.

[48] V. Olej, J. Filipová, P. Hájek, Time Series Prediction of Web Domain Visits by IF-Inference System. *Proc of the 14th WSEAS International Conference on Systems*, Latest Trends on Computers, N. Mastorakis at all. (Eds.), Vol.1, Corfu, Greece, 2010, pp.156-161.