UDC:681.324
Original scientific paper

DECISION TREE ANALYSIS OF THE PREDICTORS OF INTERNET AFFINITY

Goran Bubaš
University of Zagreb,
Faculty of Organization and Informatics, Varaždin
gbubas@foi.hr

Božidar Kliček
University of Zagreb,
Faculty of Organization and Informatics, Varaždin
bklicek@foi.hr

Željko Hutinski
University of Zagreb,
Faculty of Organization and Informatics, Varaždin
zhutinsk@foi.hr

Abstract: *A recently developed model of Internet affinity was used for survey design and data collection on variables that have potential influence on affinity for Internet use. A total of 600 Croatian students with access to the Internet at their college participated in this survey. The collected data were used for investigation of the relation between decision tree analysis and regression analysis of predictor variables of Internet affinity. Different predictors were found to influence two distinct criteria of Internet affinity (1) frequency of Internet use and (2) desire to use the Internet. Decision tree analysis and regression analysis manifested much similarity in the uncovered lists of variables that affected the criteria. However, the results of decision tree analysis were more informative since they provided greater insight into the structural relations of important predictor variables and also into specific subgroups of subjects in the survey.*

Keywords: *Internet affinity, survey, decision tree analysis, regression analysis.*

1. INTRODUCTION

The factors of *Internet affinity* are an important subject of research in social environments with low level of Internet penetration and insufficient Internet use among those populations that have been provided with the connection to the Internet, or can afford an Internet connection for personal use. Two surveys in Croatia indicated that less than 10% of the Croatian population regularly use the Internet (IPSA, 2000) and, even more surprisingly, that only about 20% of university students use the Internet (CARNet, 2001) even though most of them have been provided with free Internet access at their college.

A model of Internet affinity was recently developed (Bubaš & Hutinski, 2001a) that outlines various factors that may affect the affinity for the use of the Internet, as well as the affinity for the use of traditional mass media like television, radio and the press. Therefore, a survey of college students was performed to test this model and the preliminary results of this survey confirmed the applicability of the model for the investigation of factors of Internet affinity (Bubaš & Hutinski, 2001b). This paper introduces *decision tree analysis* to provide additional insight into the structure of Internet affinity, and also to enable a comparison between the decision tree and regression analysis for empirical investigations in this area.

Decision tree analysis is one of the methods of *data mining* and/or *exploratory data analysis*. As a step of *knowledge discovery in databases*, data mining involves the application of particular data mining methods for *verification* of hypothesis, or *discovery* of new patterns in data for the prediction or description of phenomena (Fayyad, 1996).

It must be noted that the decision tree analysis was previously also used for the investigation of the predictors of marijuana use among adolescents (Bubaš *et al.*, 1998) and it provided additional information on the nonlinear structure of the inclination toward this specific type of substance abuse.

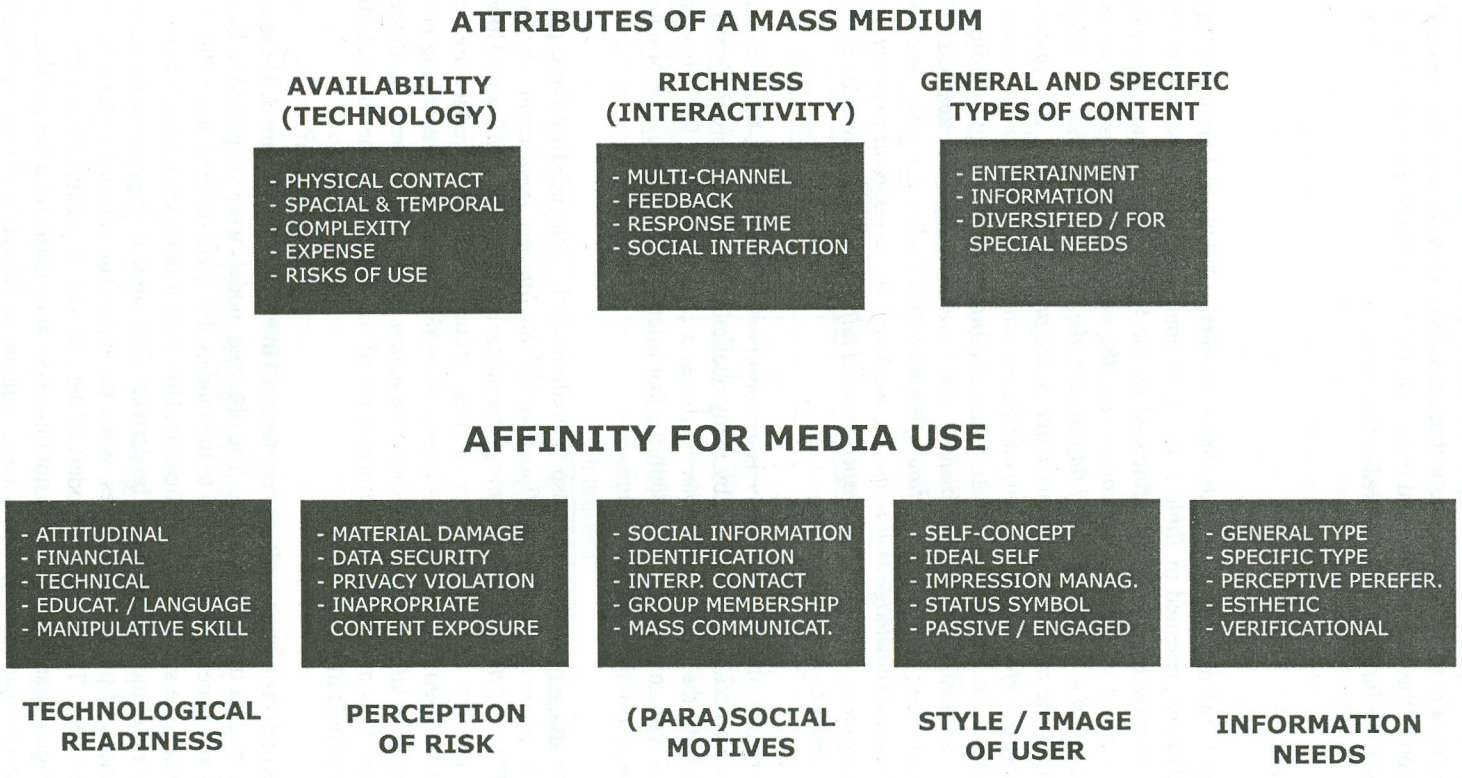
2. PROBLEM

The main goals of this paper are (1) to investigate factors that are related to Internet affinity by the method of decision tree analysis, and (2) to compare the methods of decision tree analysis and regression analysis on equal sets of predictor and criteria variables.

An important question in this research is whether or not the decision tree analysis can provide valuable additional information in comparison with standard statistical methods for prediction and description of factors that influence Internet affinity.

3. METHOD

The Internet affinity model (Bubaš & Hutinski, 2001a) that is outlined in *Figure 1* was used for the design of the items of the survey. A total of 29 statements about medium use and perceptions of a medium were formulated to represent the elements of this model as potential *predictors* of medium affinity. Also, four statements were formulated for the following *criteria* of medium affinity: the frequency of medium use, the desire to use a medium, the need or obligation to use a medium, and the expectance of the social environment about personal use of a medium. The rating was performed on a 1-5 scale (ranging from *1 - very little / poor* to *5 - very much / good.* The following mass media were rated in this survey in relation to the elements of the Internet affinity model: television, radio, Internet/Web and the press. The respondents were 600 students from four colleges in Croatia (all of the students had free access to the Internet at their college).


ATTRIBUTES OF A MASS MEDIUM
AVAILABILITY (TECHNOLOGY)
- PHYSICAL CONTACT
- SPACIAL & TEMPORAL
- COMPLEXITY
- EXPENSE
- RISKS OF USE
RICHNESS (INTERACTIVITY)
- MULTI-CHANNEL
- FEEDBACK
- RESPONSE TIME
- SOCIAL INTERACTION
GENERAL AND SPECIFIC TYPES OF CONTENT
- ENTERTAINMENT
- INFORMATION
- DIVERSIFIED / FOR SPECIAL NEEDS
AFFINITY FOR MEDIA USE
- ATTITUDINAL
- FINANCIAL
- TECHNICAL
- EDUCAT. / LANGUAGE
- MANIPULATIVE SKILL
TECHNOLOGICAL READINESS
- MATERIAL DAMAGE
- DATA SECURITY
- PRIVACY VIOLATION
- INAPROPRIATE CONTENT EXPOSURE
PERCEPTION OF RISK
- SOCIAL INFORMATION
- IDENTIFICATION
- INTERP. CONTACT
- GROUP MEMBERSHIP
- MASS COMMUNICAT.
(PARA)SOCIAL MOTIVES
- SELF-CONCEPT
- IDEAL SELF
- IMPRESSION MANAG.
- STATUS SYMBOL
- PASSIVE / ENGAGED
STYLE / IMAGE OF USER
- GENERAL TYPE
- SPECIFIC TYPE
- PERCEPTIVE PEREFER.
- ESTHETIC
- VERIFICATIONAL
INFORMATION NEEDS
ATTRIBUTES OF THE USER

Data analysis was performed by *stepwise linear regression* and also by *decision tree analysis* (based on a minimal entropy algorithm). However, only the data that was collected on the evaluation of the *Internet/Web* and in relation to the criteria *frequency of Internet use* and *desire to use the Internet* are presented in this paper.

4. RESULTS

The results of the regression analysis that were obtained in an earlier study (Bubaš & Hutinski, 2001b) are presented in *Table 1*. It can be concluded from these results that the ***frequency of Internet use*** is mostly influenced by the following predictors: *the physical availability (location) of the Internet connection, the needed knowledge/skill to use the medium*, the potential for specific need fulfillment*, the characteristic that the Internet is not rich with text, the possibility for interactive redesign of the medium, the concordance with the perceptive preference of the user, and the presence of information on daily/weekly news*. Furthermore, for the criterion ***desire to use the Internet*** a substantially different list of predictors was found to be most influential: *the creation of favorable impressions, the amount of entertaining content, the feeling of social contact when using the medium, the possession of needed knowledge/skill to use the medium*, the assessment (amount/quality) of professional content, the potential for specific need fulfillment*, and possible exposure to inappropriate content.*

When the results of regression analysis are observed in more detail, the *desire to use the Internet* as a criterion is influenced by many qualitatively different predictor variables in comparison with the *frequency of Internet use* as a criterion (except for two variables that are marked with an asterisk). This confirms that Internet affinity should be observed as *composed of these two quite distinct components.*

In fact, the frequency of Internet use is predominantly influenced by more practical factors like *the physical availability (location) of the Internet connection, the needed knowledge/skill to use the medium,* and *the potential for specific need fulfillment.* However, the desire to use the Internet is under greatest influence of the social/leisure related predictors like *the creation of favorable impressions, the amount of entertaining content,* and *the feeling of social contact when using the medium.* It can be observed that the results of regression analysis at least partly confirm some of the elements of the Internet affinity model that is outlined in *Figure 1*.

The results of decision tree analyses are presented in *Figure 2* and *Figure 3*. These results were obtained on data from 550 subjects. The obtained models were simplified to facilitate understanding and generally correspond to the models that were derived on smaller sets of data/records that were subjected to verification. However, it is important to emphasize that a number of parallel models could also be derived, with some divergence in the generated structural relations of predictor variables, when the input data sets and criteria for model generation are varied. Therefore, the models of the relations of predictor variables that are presented in *Figure 2* and *Figure 3* are not universal since there are other parallel models with additional informative value. For instance, with more subjects in the survey there could

more *nods* and *leaves* in the decision tree graph, i.e. more predictor variables would probably be included/represented in the model, with equivalent or even greater precision.

Most of the variables that were identified as the *predictors* of Internet affinity by regression analysis (that are presented in *Table 1*) were also found to affect the two criteria variables according to the results of decision tree analysis in *Figure 2* and *Figure 3*. Still, the results of *decision tree analysis* have some obvious comparable advantage:

a) the relations of predictor variables are much more structured and informative, with greater heuristic value;

b) it is more convenient to select the *crucial* or *key* predictor variables for low and high average values of the criteria variables.

Table 1. Predictors of *frequency* of Internet/Web use and of the *desire* to use the Internet/Web

PREDICTORS OF FREQUENCY OF USE	β (R=0.69)	PREDICTORS OF DESIRE OF USE	β (R=0.42)
Physical availability (location)	.30	Creation of favorable impressions	.15
Have needed knowledge / skill*	.28	Amount of entertaining content	.13
Enables specific need fulfillment*	.19	Feeling of social contact	.13
Medium is rich with text	-.10	Have needed knowledge / skill*	.12
Interactive redesign is possible	.09	Assessment of professional content	.10
Is according to perceptive preference	.08	Enables specific need fulfillment*	.08
Information on daily/weekly news	.08	No exposure to inappropriate content	-.08

β - beta coefficients; **R** - multiple regression; * - predictor variables that are related to both criterion variables

The results of decision tree analysis that are presented in *Figure 2* indicate that the *lowest average frequency of Internet use* is attributed to the 87 subjects that are represented by Leaf 1 in the graph of the decision tree. These subjects are characterized by:

- lower self-rating of needed knowledge for Internet use (ranging from *1 - very little/poor* to *3 - average*) and

- low rating of physical availability of the Internet as a medium (ranging from *1 - very little/poor* to *2 - little/poor*).

In fact, 90% of the subjects in this research, that are characterized by the former

condition/rule, reported low frequency of Internet use.

Furthermore, according to the results of decision tree analysis in *Figure 2* the *highest average frequency of Internet use* was reported by 83 subjects that are represented by Leaf 6 in the graph of the decision tree and were characterized by the following attributes:

- high self-rating of needed knowledge for Internet use (ranging from *4 - much/good* to *5 - very much/good*),
- higher rating of the Internet as a source of daily/weekly news (ranging from *3 - average* to *5 - very much/good*),
- very high rating of the Internet as a medium that can fulfill specific needs (using the highest point on the rating scale *5 - very much/good*), and
- high rating of the medium considering the possibility for interactive redesign (ranging from *4 - much/good* to *5 - very much/good*).

It was found that 85% of the subjects in this research, that are characterized by all of the preceding attributes (or condition/rule), reported a frequency of Internet use that is above average

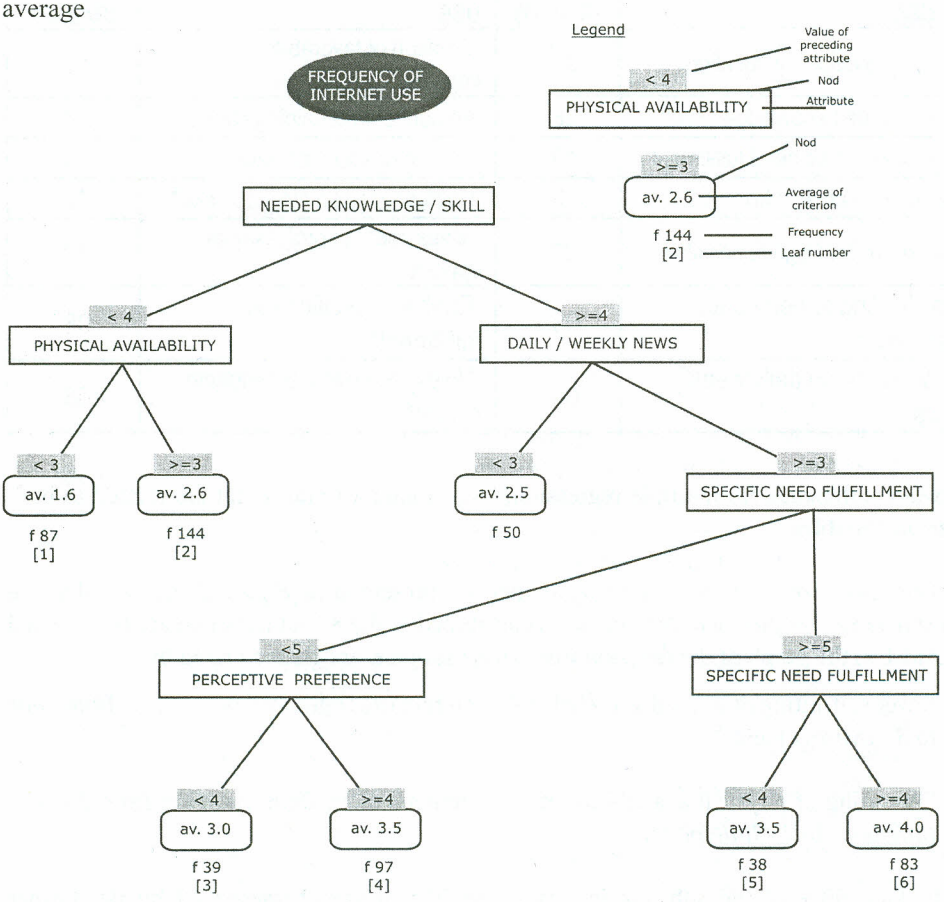


Figure 2. Results of decision tree analysis of predictors of *frequency of Internet use* (N=550; standard deviation σ is in the range 0.6-1.1 for all leafs in the graph)

It can be concluded that, according to the decision tree analysis and for the subjects in this research, the low frequency of Internet use is most clearly associated with *low level of knowledge/skill for Internet use* and *poor physical availability of this medium*. However, high frequency of Internet use is associated to a somewhat different set of variables, i.e. to (1) higher ratings of this medium for the *possibility to fulfill specific needs* of the users and (2) average or high rating of the Internet as a *source of daily/weekly news*.

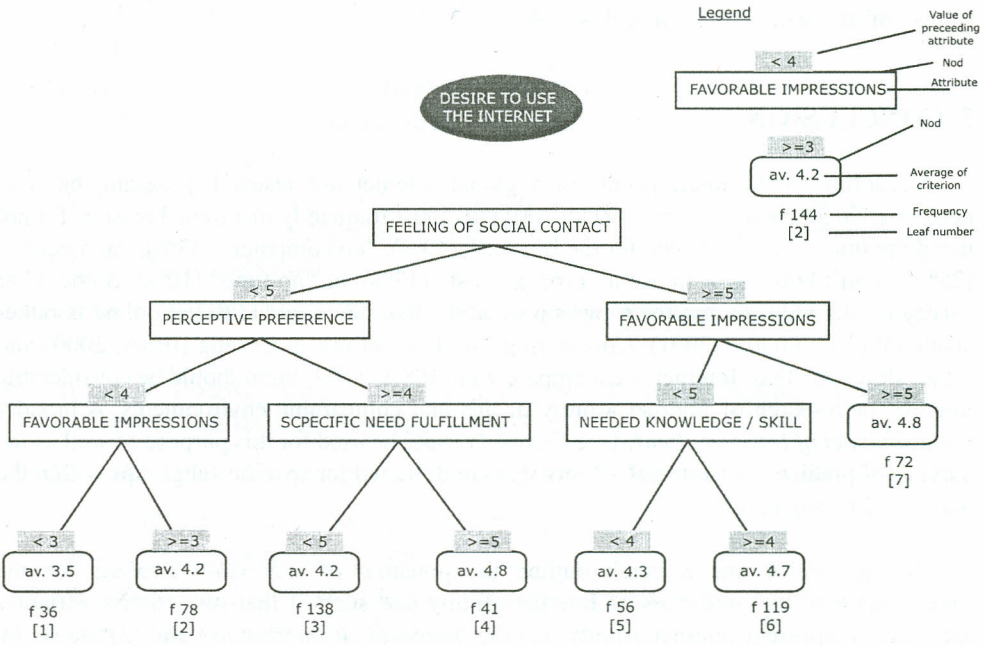


Figure 3. Results of decision tree analysis of predictors of *desire to use the Internet* (N=550; standard deviation σ is in the range 0.4-1.2 for all leafs in the graph)

The results of the decision tree analysis that are presented in *Figure 3* indicate that the *highest level of average rating of the desire to use the Internet* is associated with subjects that are denoted by Leaf 4, Leaf 6, and by Leaf 7. Only 2-3 rules/attributes are needed to characterize the subjects denoted by these leafs in the graph of the decision tree that is presented in *Figure 3*. For instance, in case of Leaf 7 the subjects are characterized by highest ratings of the *feeling of social contact when using the Internet* and *possibility to create favorable impressions* because of Internet use. Other important predictors are that the *Internet is according to perceptive preference of the user(s)* and that it has the potential for *specific need fulfillment* (Leaf 4), and also the *possession of needed knowledge/skill to use the Internet* (Leaf 6).

It must be noted that rating of the desire to use the Internet was high for most of the subjects in this research (87% of the subjects rated their desire to use the Internet using points 4 or 5 on a 1-5 scale).

Finally, as was mentioned earlier, researchers should have in mind that the potential of the decision tree analysis is related to the number of subjects in the survey. At least

several hundred subjects are needed to enable more detailed structures of decision trees that are produced by this method of pattern discovery in data. Therefore, the plans for future investigation in this area include at least twofold of subjects included in a survey. Also, parallel models of investigated phenomena can often be derived that should not be observed as an inconvenience but as a source of additional information about the subject of research and as a possibility for verification of previous results of data analysis and theoretical implications. Correspondingly, such parallel models and their compatibility should be a subject of future investigation of this topic.

5. CONCLUSION

According to the latest results of a global Internet use research program that was performed in 30 countries (Ipsos-Reid, 2001) the most frequently mentioned reasons for not using the Internet are "no need for the Internet" (40%), "no computer" (33%), "no interest" (25%), "don't know how to use it" (16%), "cost" (12%), or "no time" (10%). Some other survey results indicate that the subgroup of adults that never intend to get online is rather stable (Which? Online, 2001). Considering low Internet use in Croatia (IPSA, 2000) and rather slow growth of Internet use (compare with: IPSA, 1999) there should be considerable interest for research of Internet affinity in this and comparable environments. A broadly defined *model of Internet affinity* (see *Figure 1*) could be used for this purpose as well as the surveys of predictors of Internet affinity that are designed for specific subgroups within the general adult population.

The results of this research outline the potential of *regression analysis* for the investigation of the predictors of Internet affinity and suggest that two criteria variables are used to represent Internet affinity: (1) the *frequency of Internet use* and (2) the *desire to use the Internet*. The frequency of Internet use was predominantly influenced by *the physical availability (location) of the Internet connection, the needed knowledge/skill to use the medium,* and *the potential for specific need fulfillment*. However, the desire to use the Internet was mostly affected by *the potential for the creation of favorable impressions, the amount of entertaining content,* and *the feeling of social contact when using the medium*. Such findings, that correspond to the elements of the Internet affinity model that is outlined in *Figure 1*, could be utilized in campaigns aimed at increasing Internet affinity in certain subgroups of a general population with low average Internet use.

Furthermore, it was found that the *decision tree analysis* could prove more predictive and informative than regression analysis for groups that are characterized by very low or very high Internet affinity. Still, to use this method appropriately, quality predictor variables should be selected, a large number of respondents should be included in the survey, and implications should be reserved for groups that the subjects in the survey can really represent. Problems of uncertainty of the results of such data mining methods should definitely be taken into account (Fayyad, 1998), and the patterns that are discovered by decision tree analysis and other data mining techniques should be verified and evaluated for relevance before making empirical implications or theoretical inference (Hand, 1998).

REFERENCES

[1] Bubaš, G., Hutinski, Ž. (2001a). Činitelji sklonosti uporabi interneta kao masovnog medija. Treći znanstveni skup "Novi mediji", Zagreb, 2001.

[2] Bubaš, G., Hutinski, Ž. (2001b). Preliminary evaluation of a model of Internet affinity: Comparisons with traditional mass media. *International conference "Information T echnology and Journalism"*, Dubrovnik, Croatia. http://www.foi.hr/itn/en/itn01/bubhut01.htm

[3] Bubaš, G., Kliček, B., Čolović Rodik, Ž., Fulir, Z. (1998). Application of knowledge discovery in databases techniques to an analysis of survey data on substance abuse. *Zbornik radova*, 22(2), 79-96.

[4] CARNet (2001). CARNet započeo promotivnu akciju "Moja veza". Croatian Academic and Research Network, Zagreb, Croatia. http://cn.carnet.hr/arhiva/2001/20010327veza/index.html

[5] Fayyad, U. (1998). Diving into databases. *Database Programming & Design*, 11(3), 24-31.

[6] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

[7] Hand, D.J. (1998). Statistics and more? *The American Statistician*, 52, 112-118.

[8] IPSA (1999). *Istraživanje o korisnicima i korištenju Interneta u Hrvatskoj*. IPSA, Zagreb, Croatia, http://www.ipsa.hr/b1.htm

[9] IPSA (2000). *Istraživanje o korisnicima i korištenju Interneta u Hrvatskoj*. IPSA, Zagreb, Croatia, http://www.ipsa.hr/b2.htm

[10]Ipsos-Reid (2001). *Why aren't more people online?* Ipsos-Reid Corporation, Canada. http://www.ipsos-reid.com/media/content/displaypr.cfm?id_to_view=1244

[11] StatSoft, Inc. (2001). *Electronic Statistics Textbook*. StatSoft Inc., Tulsa, OK. http://www.statsoft.com/textbook/stathome.html

[12]Which? Online (2001). *'The Net Result: evolution not revolution' - Which? Online annual internet survey*. Which? Online, Hertford, UK. http://www.which.net/whatsnew/pr/jun01/general/internet.html

Received: 15 December 2001
Accepted: 1 July 2003