

Discriminant Analysis as a Machine Learning Method for Revision of User Stereotypes of Information Retrieval Systems

Xiangmin Zhang

School of Communication, Information and Library Studies, Rutgers University
4 Huntington Street, New Brunswick, NJ 08901
xzhang@scils.rutgers.edu

Abstract. This paper proposes to use the discriminant analysis technique as a machine learning method to adjust memberships of stereotypes, based on the user's in-depth, task-related knowledge contained in the user models. The paper reports an empirical study on the user stereotypes of information retrieval (IR) systems. The participants were first assigned into stereotypes based on their self-reported characteristics. Their memberships in the stereotypes were then tested and predicted using the discriminant analysis, based on their IR knowledge. The pre-assigned membership and the predicted membership of each stereotype were compared. The study demonstrates that the discriminant analysis technique can be used to detect the conflicts between individual users' knowledge and the assumption held by stereotypes that all members in a stereotype share common knowledge. The technique can be used to revise/reclassify a person's membership of a stereotype based on the person's knowledge. Implications and future directions of the study are discussed.

1. Introduction

Stereotype technique has been widely used in user modeling systems (e.g., Rich: 1989; Chin: 1989; Brajnik, Guida, & Tasso: 1990; Kobsa, Muller, & Nill: 1994; Fernandez-Manjon, Fernandez-Valmayor, & Fernandez-Chamizo: 1998; and Paterno & Mancini, 2000). One advantage of using stereotype technique is that the knowledge about a particular user will be inferred from the related stereotype(s) as much as possible, without explicitly going through the knowledge elicitation process with each individual user. Another advantage is that the information about user groups/stereotypes can be maintained with low redundancy (Rich, 1989; Fink & Kobsa, 2000).

Nevertheless, using stereotypes is not without problems. The major problems with the stereotype-based user modeling have been related to the tasks of correctly classifying the user and of keeping the knowledge in the user model consistent (solving conflicts between assumptions in stereotypes and between the stereotypes themselves) (Bellika, Hartvigsen, & Widding: 1998). These problems have been pointed out by many researchers, e.g., Bellika, Hartvigsen, & Widding (1998), Brajnik, Guida and Tasso (1990) and Shapira, Shoval, & Hanani (1997).

The problems are largely due to the way a stereotype is formed. Most stereotypes are formed based merely on users' external characteristics and on subjective human judgment, usually of a number of users/experts (Shapira, Shoval, & Hanani: 1997). To improve the accuracy of stereotypes, various ways have been proposed to construct user classes. These include using fuzzy set theory (Mitchell, Woodbury, & Norcio, 1994) and a user questionnaire and cluster analysis for various user data (Shapira, Shoval, & Hanani, 1997).

The current paper proposes to use the discriminant analysis technique as a machine learning method to adjust stereotype memberships based on some in-depth user knowledge, after the stereotypes have been implemented. The paper demonstrates the use of discriminant analysis in an empirical study on the user stereotypes of information retrieval (IR) systems. The study used the technique to detect the conflicts between individual users' knowledge about IR systems and their assignment of stereotypes based on some explicit user characteristics. The participants were re-assigned (predicted) by the discriminant analysis into appropriate stereotype(s).

In the remainder of this paper, the discriminant analysis technique is first very briefly introduced. The stereotypes and the participants of the empirical study are then described, which is followed by a discussion of the user knowledge that served as the input for the discriminant analysis. The results of data analyses are presented in the next section. Finally, the paper concludes with a discussion of the results and implications for use of the technique in user modeling.

2. Discriminant Analysis Technique

The discriminant analysis extracts from a group members' data a group classification criterion (SAS Institute Inc., 1988) or a classification rule (Huberty, 1994) for classifying each observation in the group. Based on this classification criterion, a posterior possibility of the group membership for each member is calculated. This posterior possibility is the probability of an individual's membership in a group based on the individual's data. The individual is predicted to be in the group for which the person's posterior possibility is maximum (compared to the posterior possibilities for other groups).

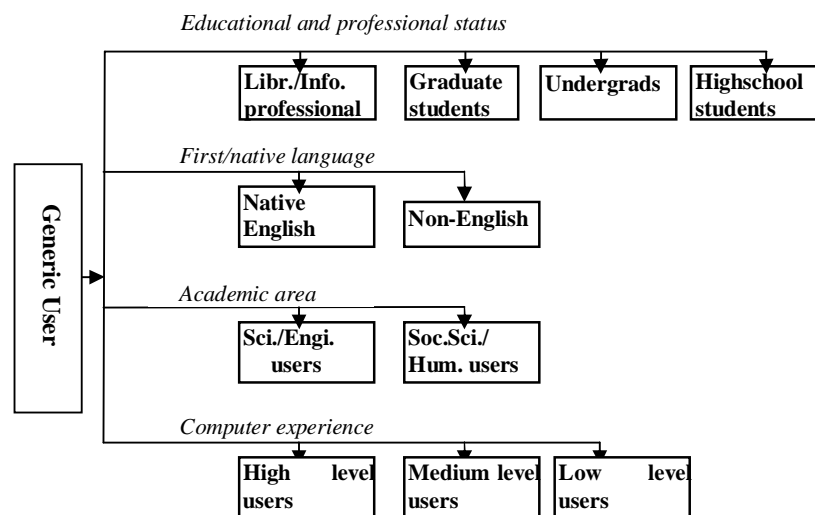
If an individual's calculated posterior possibility of belonging to a pre-determined group is not the maximum one, the individual's initial assignment will be judged as being "misclassified" into that group. An error rate (probabilities of misclassification) estimate for a whole group will be calculated.

The reader is referred to Huberty (1994) for a detailed technical description of the technique. This technique can be used as machine learning in determining the memberships of stereotypes. The re-assignment of stereotype members based on the new knowledge the system has obtained about individual users. It is particularly suitable for maintaining implemented stereotypes because it does not break up the initially defined stereotypes.

3. Stereotypes and Participants in the Study

The stereotypes that were evaluated in this study are described in Figure 1. These stereotypes are categorized in a hierarchical structure along four different dimensions: educational and professional status, first (or native) language, academic discipline, and level of computer experience. A user belongs to different stereotypes when categorized on different dimensions. Uses of these stereotypes are common in IR user studies. For example, academic discipline has been found having an impact on IR performance. Engineering and science major students performed better than those in social sciences and humanities, particularly on complex search tasks (e.g., Borgman, 1989; Zhang, 2002). If a user were a social sciences student and the IR system the user would be using were able to identify the person's academic discipline background, the system could provide tailored help on constructing a powerful search query, and also provide the results to the user with appropriate technical details.

Fig. 1. User Stereotypes of IR Systems



Sixty-four people participated in the study. They were recruited from four populations: professional librarians and information specialists, graduate students (both doctoral and master's), undergraduate students and high school students. The participants were pre-assigned into the corresponding stereotypes based on their self-reported explicit characteristics. The distribution of participants in different stereotypes is summarized in Table 1.

In Table 1, the first column lists the numbers of participants on educational and professional status. The second column lists the numbers of participants according to their first language: native English and non-native English (Non-Engl). In the third column, the distribution of the participants in two academic disciplines is presented: Science/Engineering (abbreviated as Sci./Engi.) and Social Sciences/Humanities (abbreviated as Soc./Hum.). It should be noted that the participants in these categories

4 Xiangmin Zhang

were only university students (both graduate and undergraduate). Librarians and high school students were not included. The last column, column 4, presents the distribution of participants in terms of computing experience.

Table 1. Distribution of Experiment Participants

<i>Educational & Professional Status</i>	<i>First Language</i>		<i>Academic Discipline</i>		<i>Computer Experience</i>		
	English	Non-Engl.	Sci.& Engi.	Soc. & Hum.	High	Med.	Low
Librarian	4	4	N/A ¹	N/A	8	0	0
Graduate	10	8	9	9	9	9	0
Undergraduate	4	10	6	8	7	4	3
High School	7	17	N/A ²	N/A	2	9	13
Total	25	39	15	17	26	22	16

¹ "N/A": not applicable to librarians. They were recruited as a group who were particularly trained for information retrieval tasks and they were not asked to state their undergraduate or graduate disciplines other than library and information science.

²The high school students did not yet have a formal academic orientation.

4. User Knowledge about IR Systems

When considering user modeling, the first question to be asked should be "what is being modeled" (Spark Jones, 1989), i.e., what types of user properties should be modeled. Bushey, Manuney, & Deelman (1999) suggest that users be categorized based upon the characteristics/behavior that are important to the design of the related system.

In this study, users' knowledge about IR systems was modeled and analyzed to see if members of a stereotype would have the same knowledge level. This knowledge is considered as an important factor affecting a user's search performance (Allen, 1996). An assumption is made that a member of a stereotype has the knowledge that other members within this group commonly have.

Table 2. Concepts and Attributes Used in Final Data Analysis

Concepts	Attributes
1. browsing	1. form/process
2. classification	
3. data structure	2. targeted/untargeted
4. document content	
5. feedback	3. specific to IR systems /applicable to all information systems (ISs)
6. information need	
7. interface	
8. query	
9. search	

A participant's IR knowledge is represented by the person's ratings on 9 concepts about IR systems against three attributes* (Zhang & Chignell, 2001). These concepts

* Initial ratings are against 8 attributes. However, significant differences were found on only the 3 ones.

and attributes are listed in Table 2, and the sample worksheet that was used by the participants to rate the concepts against the attributes is illustrated by Fig. 2.

Fig. 2. Sample Concept Rating Worksheet

Concept (1)	
Attribute(1)(left pole)	1 2 3 4 5 X Attribute(1)(right pole)
Attribute(2)(left pole)	1 2 3 4 5 X Attribute(2)(right pole)
...	...
Attribute(n)(left pole)	1 2 3 4 5 X Attribute(n)(right pole)

On the worksheet, all attributes were transformed into five point scales, with "1" at the left poles and "5" at the right poles of the attributes. In case some participants had difficulty in understanding an attribute or a concept, or they thought an attribute was not applicable to a concept, a "not applicable" option was added to the scales, which was represented by an "X" sign.

To reveal unexpected dimensions (or factors) among the original variables and to reduce the number of original variables to fewer ones (Mulaik, 1972), the original ratings were summarized using the factor analysis technique. Using the principal components approach in factor analysis, with the varimax rotation, the original 27 (9 concepts x 3 attributes) variables were transformed into principal factors. The first nine factors were selected to use because their eigenvalues were greater than 1, which is a norm used in factor analyses. These 9 factors accounted for 68% of the total variations from the original ratings.

Each of the 9 factors represented certain original variables (ratings). Factor loadings and interpretations of the factors are summarized in Table 3.

On each factor, a participant had a factor score. A high factor score means the concepts were rated on the high value end of the attribute scale in the factor. A low score means the concepts were rated on the low value end of the scale.

Table 3. Rotated Factor Structure for Concept Ratings

Factor and Variable ¹	Factor Loadings ²
Factor 1: Purposefulness of Querying	
Information need: <i>targeted/untargeted</i>	0.84
Query: <i>targeted/untargeted</i>	0.82
Search: <i>targeted/untargeted</i>	0.69
Document content: <i>targeted/untargeted</i>	0.57
Factor 2: Applicability of Data Organization	
Data structure: <i>specific to IR systems/applicable to all ISs</i>	0.77
Doc. Content: <i>specific to IR systems/applicable to all ISs</i>	0.72
Feedback: <i>specific to IR systems/applicable to all ISs</i>	0.65
Interface: <i>specific to IR systems/applicable to all ISs</i>	0.65
Classification: <i>specific to IR systems/applicable to all ISs</i>	0.50
Factor 3: Function of Querying	
Information need: <i>form/process</i>	0.73

Query: <i>form/process</i>	0.70
Search: <i>form/process</i>	0.60
Factor 4: Applicability of Querying	
Query: <i>specific to IR systems/applicable to all ISs</i>	0.82
Information need: <i>specific to IR systems/applicable to all ISs</i>	0.66
Search: <i>specific to IR systems/applicable to all ISs</i>	0.53
Factor 5: Applicability of Browsing	
Browsing: <i>specific to IR systems/applicable to all ISs</i>	0.85
Feedback: <i>form/process</i>	0.52
Factor 6: Function of Data Structure	
Data structure: <i>form/process</i>	0.74
Interface: <i>targeted/untargeted</i>	0.72
Factor 7: Purposefulness of Browsing	
Browsing: <i>targeted/untargeted</i>	0.86
Factor 8: Function of document	
Document content: <i>form/process</i>	0.84
Factor 9: Purposefulness of data structure	
Data structure: <i>targeted/untargeted</i>	0.79

¹Variables (ratings) within each factor are presented in descending order of their factor loadings. Each variable consists of a concept and an attribute that is italicized.

²Factor loadings are sorted by factor and only those loadings greater than .50 are shown.

5. Results of Discriminant Analysis on Factor Scores

Using the four user characteristics as the grouping variables, and the 9 factors as the predicting variables, the discriminant analysis technique was used to examine if an individual's predicted membership of a stereotype, based on the factor scores, was consistent with the person's pre-assigned membership of a stereotype. If not consistent, which stereotype this person should be assigned into based on the individual's factor scores.

For this study, equal prior probability was used. That is, no assumption was made about the size of the various populations covered in the study.

The analysis was performed using SAS Windows version 8.2.

The discriminant analysis generated two results for each pre-defined stereotype: the misclassification or error rate for the whole stereotype and the predicted (or corrected) members of the stereotype. The latter for all stereotypes by the 4 grouping variables are summarized in Tables 4a to 4d, respectively. The error rates are illustrated in Figures 3a through 3d.

In Tables 4a to 4d, the results are presented in two dimensions for each stereotype. A row shows the classification results for a pre-defined stereotype and the column displays the result for a predicted stereotype. For a pre-defined stereotype, the results show which stereotypes the original members of this stereotype should be classified into based on the factor scores, including the current stereotype itself. For a predicted stereotype, the results show from which original pre-defined stereotypes the members of the current predicted stereotype come from. Both numbers and percentages of participants are displayed. Italic percentages in the tables are the percentages that the discriminant analysis predicted.

Table 4a. Classification Results for User Stereotypes on Educational & Professional Status Variable

Predetermined User Stereotypes	Predicted User Stereotypes				
	<i>Librarian</i>	<i>Graduate</i>	<i>Undergraduate</i>	<i>High school</i>	Total¹
<i>Librarian</i>	8(100%),(89%) ³	0 (0%), (0%)	0 (0%), (0%)	0 (0%), (0%)	8 (100%)
<i>Graduate</i>	0 (0%), (0%)	11 (61%), (65%)	4 (22%), (24%)	3 (17%), (14%)	18 (100%)
<i>Undergrad</i>	0 (0%), (0%)	2 (14%), (12%)	6 (43%), (35%)	6 (43%), (29%)	14 (100%)
<i>High school</i>	1 (4%), (11%)	4 (17%), (24%)	7 (29%), (41%)	12 (50%), (57%)	24 (100%)
Total²	9 (100%)	17 (100%)	17 (100%)	21 (100%)	64 (100%)

Table 4b. Classification Results for User Stereotypes on Native Language Variable

Predetermined User Stereotypes	Predicted User Stereotypes		
	<i>Native English</i>	<i>Non-native English</i>	Total¹
<i>Native English</i>	25(100%),(50%) ³	0 (0%), (0%)	25 (100%)
<i>Non-native English</i>	25(64%),(50%)	14 (36%), (100%)	39 (100%)
Total²	50 (100%)	14 (100%)	64 (100%)

Table 4c. Classification Results for User Stereotypes on the Academic Background Variable

Predetermined User Stereotypes	Predicted User Stereotypes				
	<i>No Major</i>	<i>Sci./Engi.</i>	<i>Soc.Sci./Hum.</i>	<i>Prof.</i>	Total¹
<i>No Major</i>	13(54%),(65%) ³	3 (13%), (18%)	7 (29%), (39%)	1 (4%), (11%)	24 (100%)
<i>Sci./Engi.</i>	3 (20%), (15%)	11 (73%), (65%)	1 (7%), (5%)	0 (0%), (0%)	15 (100%)
<i>Soc.Sci.Hum.</i>	4 (24%), (20%)	3 (18%), (18%)	10 (59%), (56%)	0 (0%), (0%)	17 (100%)
<i>Professional</i>	0 (0%), (0%)	0 (0%), (0%)	0 (0%), (0%)	8 (100%), (89%)	8 (100%)
Total²	20 (100%)	17 (100%)	18 (100%)	9 (100%)	64 (100%)

Table 4d. Classification Results for User Stereotypes on the Computer Experience Variable

Predetermined User Stereotypes	Predicted User Stereotypes			
	<i>High</i>	<i>Medium</i>	<i>Low</i>	Total¹
<i>High</i>	12(46%),(71%) ³	7 (27%), (41%)	7 (27%), (23%)	26 (100%)
<i>Medium</i>	3 (14%), (18%)	9 (41%), (53%)	10 (45%), (33%)	22 (100%)
<i>Low</i>	2 (13%), (12%)	1 (6%), (6%)	13 (81%), (43%)	16 (100%)
Total²	17 (100%)	17 (100%)	30 (100%)	64 (100%)

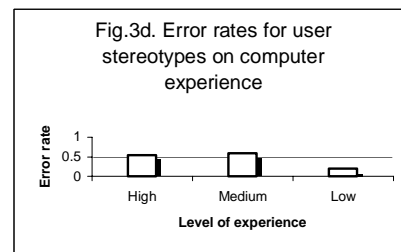
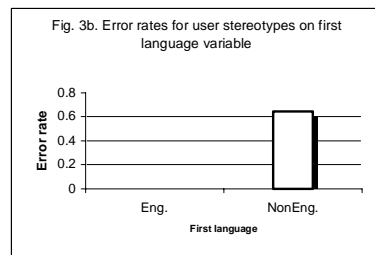
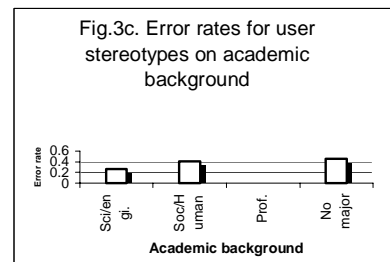
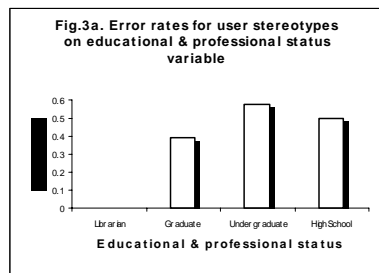
(Notes for Tables 4a to 4d:

¹Number of subjects who were pre-assigned in the group.

²Number of subjects who were predicted to be in the group by the discriminant analysis.

³Italic percentages are the percentages of the predicted membership in stereotypes.)

For example, Table 4a describes the results for the stereotypes on the educational & professional status variable. Four stereotypes were defined along this dimension: librarians/information professionals, graduate students, undergraduate students, and high school students. For the librarians/information professionals stereotype, which is displayed on the first row and in the first column, the number of pre-determined members is 8, with its total number and percentage appear in the last cell of the row. None of these 8 members was re-assigned into other types by the discriminant analysis, as in the cells under the other 3 types the numbers and percentages are all zeroes. However, these 8 original members account only 89% of the predicted membership by the data analysis, as indicated by the italic (89%) in the first cell. Go down along the column, the other 11% (one member) of the predicted group membership came from the original high school type. This pre-defined high school student was judged as “mistakenly” classified and the student should be classified in the librarians/professionals group. So the total for the predicted librarians/information professionals type is 9, as indicated by the italic total at the bottom of the same column. Tables 4b to 4d display the results for the stereotypes on first/native language, academic background, and computer experience, respectively.



Figures 3a to 3d complement Tables 4a to 4d. They show the error rates for stereotypes in each category. For example, Figure 3a depicts the error rates for the stereotypes along the educational and professional status dimension.

As shown in the figure, the error rate for librarians/information professionals is zero, since none of the 8 pre-determined members of this type was re-assigned into other types. Graduate students type had an error rate of 0.39, because 0.39 percent of its pre-defined members were re-assigned into other groups: 22% (n=4) were re-assigned into the undergraduate students group and 17% (n=3) were re-assigned as

high school students. Undergraduate students group had the highest error rate: 0.5714. Of 14 pre-assigned undergraduate participants, only 6 (43%) were consistent with their pre-assigned "undergraduate" status. The other 8 (57%) were misclassified ones: 2 (14%) were predicted as graduate students and another 6 (43%) were predicted as high school students. The high school student stereotype had an error rate of 0.5. Of its pre-assigned 24 members, 12 (50%) were predicted as "high school" participants by the analysis. Another 12 were misclassified ones. One (4%) of them was predicted as "librarian", 4 (17%) as "graduate" participants, and 7 (29%) as "undergraduate" participants.

6. Discussion and Conclusions

The results of the empirical study show that some stereotypes are reliable and accurate ones, such as librarians, while some others are not good or accurate in representing the members based on the task related knowledge. Uses of these inaccurate stereotypes will bring performance issues to a user modeling system. The results also show that the medium type along a dimension (such as undergraduate between graduate and high school students, and the medium level of computer experience between the high and low) suffer the highest error rates. This probably implies that it is not a good idea to have medium types because there are no very clear features/attributes to separate the medium type from the others at the two ends of the dimension. Perhaps a good practice would be to construct just a few very distinctive stereotypes, such as librarians and students, rather than to have many for a given system.

The empirical study described in this paper demonstrated that the discriminant analysis technique can detect the conflicts of individual knowledge and the assignment of stereotypes based on some explicit user characteristics, and can correct inaccurate assignments that are not match an individual's knowledge. As the system learns from the users and accumulates sufficient user data (can be performance data or the users' task-related knowledge) that is related to the task, the technique can be used as a machine learning method to automatically adjust the memberships of stereotypes and thus resolve the conflicts between individuals' knowledge and the knowledge held by stereotypes. One advantage to use the discriminant analysis is that the pre-defined stereotypes do not need to be changed. Only the memberships of the stereotypes will be re-assigned.

One limitation of this research is that the system performance data with different user stereotypes, both before and after the revision of the memberships of the stereotypes, was not available. Future research is needed to test the effectiveness of the method in real IR tasks with real users, to see if the system and user performance will improve after stereotype memberships are adjusted. It is currently not clear how seriously the errors in the membership assignments of stereotypes would affect the IR system performance. The assignment of participants into various stereotypes could also be done manually by some human experts based on the participants' concept ratings, so the results could be compared with that from the discriminant analysis.

References

- Allen, B.L. (1996). *Information Tasks: Toward a User-Centered Approach to Information Systems*. New York: Academic Press.
- Bellika, J.G., Hartvigsen, G., & Widding, R.A. (1998). The virtual library secretary: A user model-based software agent. *Personal Technologies*, 2:162-187.
- Borgman, C. L. (1989). All users of information retrieval systems are not created equal: An exploration into individual Differences. *Information Processing & Management*, 25(3):237-251
- Brajnik, G., Guida, G., & Tasso, C. (1990). User modeling in expert man-machine interfaces: A case study in intelligent information retrieval. *IEEE Transactions on Systems, Man and Cybernetics*. 20(1):166-185.
- Bushey, R., Manuney, J.M., & Deelman T. (1999). The development of behavior-based user models for a computer system, In: Proceedings of the Seventh International Conference on User Modeling, Banff, June 20-24, 1999; SpringerWien New York, pp.109-118.
- Chin, D.N. (1989). KNOME: Modeling what the user knows in UC. In A. Kobsa, W. Wahlster (Eds.): *User Models in Dialog Systems*, Springer-Verlag, Berlin. pp.35-51.
- Fernandez-Manjon, B., Fernandez-Valmayor, A., & Fernandez-Chamizo, C. (1998). Pragmatic user model implementation in an intelligent help system. *British Journal of Educational Technology*, 29(2): 113-123.
- Fink, J. & Kobsa, A. (2000). A review and analysis of commercial user modeling servers for personalization on the World Wide Web. *User Modeling and User-Adapted Interaction*, 10:209-249.
- Harvey, C. F., Smith, P. & Lund P. (1998). Providing a networked future for interpersonal information retrieval: InfoVine and user modeling. *Interacting with Computers*, 10: 195-212.
- Huberty, C. (1994). *Applied Discriminant Analysis*. John Wiley & Sons, New York.
- Kobsa, A., Muller, D. & Nill, A. (1994). KN-AHS: An adaptive hypertext client of the user modeling system BGP-MS. In: *Proceedings of the Fourth International Conference on User Modeling*. Hyannis (MA), pp.99-105.
- Mitchell, K., Woodbury, M.A. & Norcio, A.F. (1994). Individualizing user interfaces: Application of the grade of membership (GoM) model for development of fuzzy user classes. *Information Science*, 1:9-29.
- Mulaik, S.A. (1972). *The Foundations of factor analysis*. McGraw-Hill Inc. p.174.
- Paterno, F. & Mancini, C. (2000). Effective levels of adaptation to different types of users in interactive museum systems. *Journal of the American Society for Information Science*. 51(1):5-13.
- Rich, E. (1989). Stereotypes and user modeling. In A. Kobsa, W. Wahlster (Eds.): *User Models in Dialog Systems*, Springer-Verlag, Berlin. pp.35-51.
- SAS/STAT User's Guide, Release 6.03 Ed.*, SAS Institute Inc., Cary, NC, USA, 1988.
- Shapira, B., Shoval, P. & Hanani, U., (1997). Stereotypes in information filtering systems. *Information Processing and Management*, 33(3):273-287.
- Sparck Jones, K. (1989). Realism about user modeling. In: A. Kobsa & W. Wahlster (Eds.): *User Models in Dialog Systems*, Springer-Verlag, Berlin
- Zhang, X. & Chignell, M. (2001). Assessment of the effects of user characteristics on mental models of information retrieval systems. *Journal of the American Society for Information Science & Technology*, 52(6): 445-459
- Zhang, X. (2002). Collaborative relevance judgment: A group consensus method for evaluating user search performance. *Journal of the American Society for Information Science and Technology*, 53(3):220-235.