ELSEVIER
Information Processing and Management xxx (2004) xxx–xxx
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

An empirical testing of user stereotypes of information retrieval systems

Xiangmin Zhang [a,*], Hui Han [b]

[a] *School of Communication, Information and Library Studies, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901-1071, USA*

[b] *Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA*

Received 17 July 2003; accepted 28 January 2004

Abstract

Stereotyping is a technique used in many information systems to represent user groups and/or to generate initial individual user models. However, there has been a lack of evidence on the accuracy of their use in representing users. We propose a formal evaluation method to test the accuracy or homogeneity of the stereotypes that are based on users' explicit characteristics. Using the method, the results of an empirical testing on 11 common user stereotypes of information retrieval (IR) systems are reported. The participants' memberships in the stereotypes were predicted using discriminant analysis, based on their IR knowledge. The actual membership and the predicted membership of each stereotype were compared. The data show that "librarians/IR professionals" is an accurate stereotype in representing its members, while some others, such as "undergraduate students" and "social sciences/humanities" users, are not accurate stereotypes. The data also demonstrate that based on the user's IR knowledge a stereotype can be made more accurate or homogeneous. The results show the promise that our method can help better detect the differences among stereotype members, and help with better stereotype design and user modeling. We assume that accurate stereotypes have better performance in user modeling and thus the system performance.

Limitations and future directions of the study are discussed.

Keywords: Information retrieval; Stereotype; User modeling; Statistical analyses

1. Introduction

One way to improve the performance of information systems is to build user models into systems and customize the system to a user's specific need. User models can be either *group models* built for distinctive groups of users or *individual models* built for individual users. Stereotype is a widely used technique in user modeling for group modeling as well as for creating initial individual user models.

* Corresponding author. Tel.: +1-732-932-7500x8229.
E-mail address: xzhang@scils.rutgers.edu (X. Zhang).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2004.01.005

A stereotype is a common user characteristic/trait that is shared by many users (Rich, 1979, 1989). Examples of stereotypes may be 'expert users' or 'novice users'. Stereotypes are created on the assumption that the presence of particular characteristics in one member of the stereotype would imply that of others (Harvey, Smith, & Lund, 1998). Therefore, a stereotype normally contains the common knowledge about a group of users. A new user will be assigned into related stereotype(s) if some of his/her characteristics match the ones contained in the stereotype(s).

Though stereotyping usually involves intensive knowledge engineering on the part of the system administrator (Mostafa, Quiroga, & Palakal, 1998), the advantage of using the 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 (Fink & Kobsa, 2000; Rich, 1989). Kuflik, Shapira, and Shoval (2003) have found that using a stereotype was better than using a personal-based user profile in information filtering systems.

Nevertheless, using stereotypes is not without problems. 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). The user's knowledge about the system and/or task is not involved. It is common that such stereotypes do not represent their members accurately. The issue of inaccuracy of stereotypes has been pointed out by many researchers, e.g., Beaumont (1998), Bellika, Hartvigsen, and Widding (1998), Brajnik, Guida, and Tasso (1990) and Shapira et al. (1997).

The lack of accuracy is liable to lead to conflicts between the individual models and the assignment to various stereotypes (Shapira et al., 1997), which can affect accurate construction of individual user models. Consequently, system functions adapted to individual user models will fail to achieve their goals.

A common practice for the systems using stereotypes is to continuously check user responses, detect and resolve the conflicts between stereotypes and specific user knowledge values, and then update stereotypes and user models. However, we hardly see the evidence on the improvement of user models through such conflict-resolving and stereotype/model updating. Because of the inaccuracy problem, it is important that the user classes represented by the stereotypes be as homogeneous as possible, and this homogeneity should be based on the users' knowledge of specific domain or task (Beaumont, 1998).

This paper reports the results of an empirical test on the accuracy of some common stereotypes of information retrieval (IR) systems. Since the user's IR knowledge, the knowledge about IR system components, and the relationships among them are important to information searching performance (Allen, 1996), we argue that the stereotypes of IR systems be based on the user's IR knowledge, as well as the domain knowledge that is related to the specific search task. We test the accuracy or homogeneity of 11 commonly used user stereotypes by exploring the differences among the members of these stereotypes in terms of their IR knowledge. A stereotype should include only those members with similar level of IR knowledge. We believe such a stereotype is homogeneous and is more accurate than the one based just on the user's explicit characteristics. We assume such an accurate stereotype adapts better to the individual users of IR systems/services.

The fact that different people have different levels of knowledge is obvious to human experts. However, it is hard for a computer system to intelligently identify the differences. More important, we need to know not only that there exist differences, but also how the users differ and what the exact differences are. Without systematic investigations, such facts cannot be known even to human experts. By conducting this study, we propose a formal method for exploring the differences among the members of a stereotype. This kind of formal method is necessary for IR systems to build better, accurate stereotypes.

The remainder of this paper is organized as follows: Section 2 reviews literatures on stereotype-based user modeling; Section 3 describes the research design; Section 4 analyzes experiment results; Sections 5 and 6 present and discuss the experiment results, and Section 7 concludes the paper and discusses future research directions.

2. Related work

Using stereotype technique for user modeling is first seen in Rich's GRUNDY (1979, 1989), where literatures are recommended to the users of specific stereotypes. However, as Allen (1990) points out, "…data on the performance of GRUNDY are sketchy. It was shown to perform better than chance, averaged across a large number of its predictions; however, there was no test of whether the adaptive mechanism improved performance" (p. 518).

Paterno and Mancini (2000) developed three stereotypes in their museum system: experts, tourists, and students of art. Different user models are constructed based on the stereotypes and on the different tasks. They observe 30 users' behaviors in initial information access, navigation, and information presentation, and find that users are able to perform desired tasks and that the system is generally easy to use. However, they do not mention how the three stereotypes fit the users, although they suggest future studies on modeling users' preference.

Art Technology Group's Personalization Server (Fink & Kobsa, 2000) manually developed group profiles and associated rules that allowed the server to assign a user to one or more user groups. These group profiles and rules resemble the stereotype approach. The group profiles contain relevant user characteristics, such as age and gender. The rules take into account not only the user characteristic, but also system usage (e.g., pages visited, products bought) and the environment information (e.g., domain name, browser type).

Brajnik et al. (1990) report the UM-tool, a generic user modeling tool based on stereotypes. Users are assigned to stereotypes based on users' background information acquired through interviews, and their computer usage history stored in the system. The authors do not explain how the stereotypes are created or how effective they are in user representation.

Fernandez-Manjon, Fernandez-Valmayor, and Fernandez-Chamizo (1998) model UNIX users in their Aran system based on the following stereotypes: intermediate-user, novice-user, editor-user, programmer-user, network user, and math-user. These domain-oriented stereotypes try to represent users' experience and interests with respect to different UNIX contexts and subdomains. They classify users to different stereotypes based on the key characteristics of users. However, there is no evaluation on the identified key characteristics, or the sufficiency and accuracy of the stereotype.

Many other systems such as the work by Chin (1989) and Kobsa, Muller, and Nill (1994) also use stereotype-based user modeling. However, we hardly see any of the above systems test the accuracy of stereotypes. As Bellika et al. (1998) point out, an important problem with stereotype-based user modeling is incorrect classification of users and inconsistent knowledge representation in user models.

Various ways are proposed to improve the accuracy of stereotypes. Mitchell, Woodbury, and Norcio (1994) use fuzzy set theory to develop fuzzy user classes. The authors claim that this method can measure the differences between users more accurately than stereotypes. Bushey, Manuney, and Deelman (1999) suggest categorizing users based on the user characteristics or behaviors that are important to the design of the related system. They propose a method to categorize users based on their performance data, rather than some explicit characteristics not necessarily related to the task performance. Shapira et al. (1997) suggest a hybrid approach to develop stereotypes. Such an approach includes the user questionnaire, cluster analysis based on the data collected from the questionnaire, and an assessment of user clusters by field experts who know the users. They also emphasize the importance of relating stereotypes with tasks, in their words, "the environment within which the system is to be applied". This approach is used in modeling the experimental participants of an information filtering system and it appears to be effective in their study (Shapira, Shoval, & Hanani, 1999).

3. Methodology

We studied the accuracy of 11 commonly used stereotypes using discriminant analysis. We modeled users' knowledge about IR systems, and analyzed if the stereotypes could represent members' IR knowledge consistently, i.e., if members of a stereotype would have the same knowledge level. We assumed a system that could adapt to users with different levels of task knowledge differently would have better performance than one that could not. The stereotypes, participants, and knowledge elicitation and modeling are introduced below.

3.1. Stereotypes investigated in the study

As shown in Fig. 1, we investigated 11 stereotypes along four dimensions of user characteristics: educational and professional status, native language, academic discipline, and level of computer experience.

Educational and professional status was the educational degree level if a user was studying at school or the working status if the person was employed by the time of this study. The reason for choosing this characteristic was that it normally reflects a person's knowledge and skills in a professional field.

A user's first language is the language the user acquired at home during his or her childhood. It is also called the user's native language. This characteristic was chosen because IR systems are closely related to languages.

Academic discipline or background refers to the major area of knowledge a user was studying. Different disciplines have different bodies of knowledge and different approaches to exploring knowledge. These differences may have an impact on users' mental models.

Computer experience is referred to as a person's experience in using any of a list of computer applications such as database management, electronic mail, information retrieval, etc. Since IR systems are presumably computerized systems, a person's computer experience was considered important in shaping the person's views of IR systems.

The stereotypes based on the above characteristics have been frequently used in IR user studies. For example, librarians (search intermediaries) and graduate students are commonly studied subject types (e.g, Ma, 2002; Saracevic & Kantor, 1988; Yee, 1993). Shaw (1996) and Neuman (1995) investigate the search

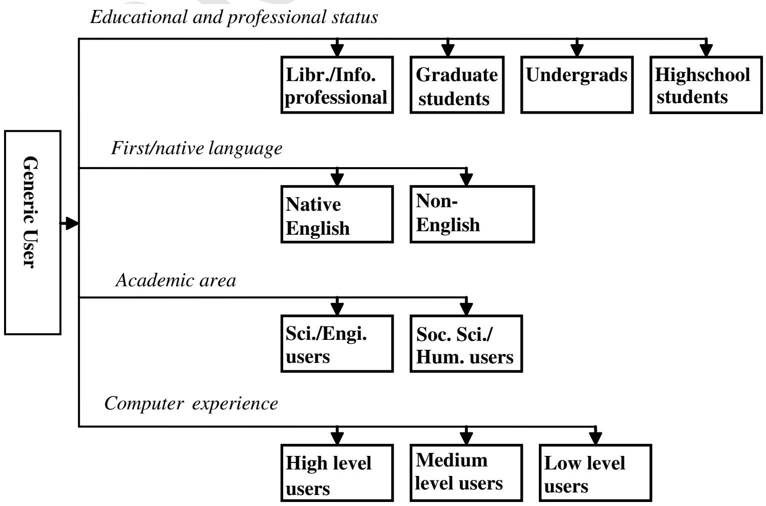

Fig. 1. User stereotypes of IR systems.

behaviors of undergraduate students and high school students. Charoenkitkarn (1996) studies the effect of native language on users' search performance. Native English speakers are also the targeted users for IR systems participated in the TREC 2003 Question Answering Track (Voorhees, 2003). Borgman (1989) examines the effects of academic orientation, among other variables, on IR performance; the subjects include undergraduates of Engineering, Psychology, and English majors. Qiu (1993) compares search patterns of different user groups in hypertext systems based on the user's academic background, as well as some other aspects. Ellis, Cox, and Hall (1993) compare researchers' information seeking patterns in Physical Sciences and Social Sciences. Su (2003) compares search performance of sciences, social sciences, and humanities.

In our study, a user can belong to different stereotypes when categorized on different dimensions. For example, a student can belong to stereotypes of undergraduate, engineering, native English speaker, and medium-level computer experience.

3.2. Participants

Our participants were mainly students, with librarians included as the benchmark group. We recruited 64 participants from four populations: professional librarians and information specialists (abbreviated as "librarians" hereafter), graduate students (both doctoral and master's), undergraduate students, and high school students. Younger students (such as elementary students) were not considered as the major type of "end users" and were therefore not included in our study.

The participants belonged to different stereotypes (as shown in Fig. 1) based on their self-reported demographic characteristics. Table 1 summarizes the distribution of participants in different stereotypes along each of the four dimensions of user characteristics, as shown column-wise. Please note that participants in the two academic disciplines were only university students (both graduate and undergraduate).

3.3. Eliciting and modeling of users' knowledge about IR systems

Various knowledge elicitation and representation techniques exist. In this study, we used the repertory grid technique (RGT). The RGT was invented as a tool for and is based on Kelly's personal construct theory (Kelly, 1955), which asserts that people understand the world (events, people, etc.) through their personal construct systems. The personal construct system that each person develops is the set of representations or model of the world that the person has developed. It is acquired through the person's social experience.

The RGT, in its simplest form, involves the generation of a list of concepts (elements) about things or events to be investigated, and the generation of attributes (constructs) based on the list of concepts.

Table 1
The distribution of experiment participants

Educational and professional status	Native language		Academic discipline		Computer experience		
	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

"Non-Engl." stands for "Non-English"; "Sci. & Engi." stands for "Science & Engineering"; "Soc. & Hum." stands for "Social & Humanities"; "Med." stands for "Medium". These abbreviations are applied to all the tables in this paper.

A concept is defined as anything that can be compared or contrasted. For example, people, vegetables or notions such as occupations, feelings, situations, events, etc. can all be elements. If the problem is to choose a future career, the concepts may be different jobs. Concepts used in a study may be elicited from the subject or provided by the tester, or both, and they need to be well known and personally meaningful to the subject (Shaw, 1980).

An attribute (construct) is a bipolar dimension that, to some degree, is a property of each concept. A construct is a way in which some things (elements) are seen as alike yet still different from others. Examples of attributes for concepts about people may be: *Don't believe in God/Very religious*; *Not athletic/Athletic*; *Understands me better/Doesn't understand at all*; *Sociable/Not sociable*, etc. (Fransella & Bannister, 1977).

Like concepts, attributes can also be elicited from the subjects or provided by the tester. There are several ways to elicit attributes. The classic method used by Kelly is to consider various triads (groups of three concepts) selected successively from the whole concept list. The subject or person(s) from which attributes are to be elicited is first presented with three concepts and asked to specify some important aspects in which two of them are alike. Then the subject is asked in which aspects the third concept differs from the other two. Often the subject will indicate spontaneously which two concepts are being judged alike. The subject's description of the similarity forms one pole of the attribute and the answer to the question concerning the difference is the contrast pole. Such a process is called a sort. The examiner records this similarity and contrast as the resulting attribute dimension from the first sort, and proceeds to the second and subsequent sorts using different triads of concepts. There are no rules on how many triads of concepts should be presented to the subjects, but between 10 and 25 is a common range (Bannister, 1968; Fransella & Bannister, 1977).

Presently, the most frequently used variation of repertory grids is rating grids. In a rating grid, the subject is asked to evaluate the concepts systematically by using the attribute list to generate the grid of rating numbers. At every intersection of column and row is the subject's rating value of the concepts on the attributes. The grid form was used as both a model elicitation tool and as a formal representation of IR knowledge in this study. Readers are referred to Latta and Swigger (1992) and Zhang and Chignell (2001) for more detailed descriptions of the use of the method in information science. Discussions on the technique's validity and reliability can be found in Bannister (1968), Latta and Swigger (1992), and Shaw and Woodward (1988).

Table 2 lists the nine concepts about IR systems and the three attributes [1] generated using the RGT method.

These concepts were suggested and decided by a group of IR experts who were faculty and doctoral students in information science. The nine concepts cover important components of IR process, from users' information needs as start and documents from databases as the end. The attributes were generated by the same group of experts using the triads method. Eight attributes were generated and rated by the subjects initially. However, only the listed three attributes were able to differentiate features of mental models among the subject groups. Since the purpose of the study was to find out differences between user groups, only the concept ratings against these three were involved in the final data analyses. The concept ratings on other attributes that did not distinguish subject groups were discarded.

These concepts and attributes were then transformed into a rating form that was administered to the participants. For example, the rating form for the concept "browsing", with the three attributes, is shown in Fig. 2.

On the form, all attributes were transformed into five point scales, with "1" at the left poles and "5" at the right poles. Participants were asked to rate on the form the nine concepts (one by one) against each of the attributes. In case some participants would have difficulty in understanding an attribute or a concept, or

[1] Initial ratings are against eight attributes. However, significant differences were found on only these 3 ones.

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	

browsing

form	1	2	3	4	5	X	process
targeted	1	2	3	4	5	X	untargeted
specific to IR	1	2	3	4	5	X	applicable to all ISs

Fig. 2. Concept rating worksheet for one concept.

they would consider an attribute not applicable to a concept, a ‘‘not applicable’’ option represented by an ‘‘X’’ sign was added to the scales to account for this. Participants could simply circle this sign to rate a concept. It was completely up to the participant to interpret the meaning of the concept on the attribute dimension. Participants either individually met with the authors or they sent their ratings to the authors through surface or electronic mail.

4. Data analyses

4.1. Factor analysis (summarization) of raw concept ratings

Each individual rating of a concept on an attribute constitutes a variable. Altogether, there are nine concepts and three attributes, constituting 27 ratings or variables. A vector of 27 variables represents a participant. These ratings needed to be summarized to reveal unexpected dimensions (or factors) among the original variables and to reduce the number of original variables (Mulaik, 1972). Using the principal component approach in factor analysis, with the varimax rotation, the original 27 (9 concepts × 3 attributes) variables were transformed into principal factors. The first nine factors with the eigenvalue greater than 1 were selected, a norm used in factor analyses. These nine factors accounted for 68% of the total variations from the original ratings. Factor loadings and interpretations are summarized in Table 3.

Each of the nine factors represents certain original variables (ratings). We assigned a name to each factor to interpret the original variables. The naming was based on the interetation of the attribute dimensions of the major concept(s). For example, the first factor was named as ‘‘Purposefulness of Querying’’ because the major concepts: ‘‘Information need’’, ‘‘Query’’ and ‘‘Search’’ were all interpreted as ‘‘querying’’ and the attribute ‘‘targeted/untargeted’’ was interpreted as reflecting the purposefulness of actions (‘‘querying’’ here). A vector of nine factors represents a participant. Each factor has a score that is a weighted combination of the observed scores on original variables in the factor (Boyce, Meadow, & Kraft, 1994, p. 84). A

Table 3
Rotated factor structure for concept ratings

Factor and variable[a]	Factor loadings[b]
Factor 1: Purposefulness of querying	
Information need: *targeted/untargeted*	0.84
Query: *targeted/untargeted*	0.82
Search: *targeted/untargeted*	0.69
Document content: *targeted/untargeted*	0.57
Factor 2: Applicability of data organization	
Data structure: *specific to IR systems/applicable to all ISs*	0.77
Doc. content: *specific to IR systems/applicable to all ISs*	0.72
Feedback: *specific to IR systems/applicable to all ISs*	0.65
Interface: *specific to IR systems/applicable to all ISs*	0.65
Classification: *specific to IR systems/applicable to all ISs*	0.50
Factor 3: Function of querying	
Information need: *form/process*	0.73
Query: *form/process*	0.70
Search: *form/process*	0.60
Factor 4: Applicability of querying	
Query: *specific to IR systems/applicable to all ISs*	0.82
Information need: *specific to IR systems/applicable to all ISs*	0.66
Search: *specific to IR systems/applicable to all ISs*	0.53
Factor 5: Applicability of browsing	
Browsing: *specific to IR systems/applicable to all ISs*	0.85
Feedback: *form/process*	0.52
Factor 6: Function of data structure	
Data structure: *form/process*	0.74
Interface: *targeted/untargeted*	0.72
Factor 7: Purposefulness of browsing	
Browsing: *targeted/untargeted*	0.86
Factor 8: Function of document	
Document content: *form/process*	0.84
Factor 9: Purposefulness of data structure	
Data structure: *targeted/untargeted*	0.79

[a] Variables (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.
[b] Factor loadings are sorted by factor and only those loadings greater than 0.50 are shown.

high factor score means that the concepts were rated on the high value end of the attribute scale in the factor. A low score means that the concepts were rated on the low value end of the scale.

4.2. Discriminant analysis on factor scores

With the four user characteristics as grouping variables, and the nine factors as predicting variables, we used the discriminant analysis technique to examine if an individual's predicted membership of a stereotype is consistent with the person's actual membership of a stereotype. The purpose was to detect the differences among the members of the same stereotype in terms of their IR knowledge.

Discriminant analysis extracts from the participants' factor scores a stereotype classification criterion/rule for classifying each observation in a user stereotype (Huberty, 1994; SAS Institute Inc., 1988). Specifically, we employed the k-nearest-neighbor method of non-parametric discriminant analysis. The participants' data set was used as both the training set and the test set to generate and evaluate the classification criterion.

Based on the classification criterion, we calculated the posterior probability of the stereotype membership for each participant based on the participant's factor scores. We used equal prior probability, assuming the equal size of various populations in the study. The participant was predicted to be in the stereotype where the participant's posterior probability was maximal (compared to the posterior probabilities for other stereotypes). If a participant's posterior probability of belonging to the actual stereotype was not the maximum one, the participant was judged as being "misclassified" into that actual stereotype. The misclassification rate, referred to as "error rate" in the remainder of the paper, was estimated for each stereotype. The analysis was performed using Windows version of SAS 8.2.

5. Results

For each actual stereotype, we generated two types of results based on discriminant analysis: the predicted stereotype membership for each person and the error rate for the whole actual stereotype.

Table 4 (panels A–D) summarize the results of stereotypes on each of the four user characteristic dimensions respectively. Fig. 3a–d show the corresponding error rates.

Each table presents the result of a stereotype in two dimensions: The row shows the classification results for the members of an actual stereotype, i.e., how many members of an actual stereotype are classified to which stereotypes. The ratio following a number on the same row is the percentage of that number in the total number of actual members. Each column displays the results for a predicted stereotype, i.e., from which actual stereotype(s) and how many consist of the current predicted stereotype membership. The ratio below a number is the percentage of that number in the total number of predicted members.

For example, Table 4(panel A) describes the results for the four stereotypes of the educational and professional status. Row 3 shows that none of the eight actual librarians is predicted into other stereotypes. However, these eight actual librarians account for only 89% of the predicted librarians, as shown in Row 5 of Column 3. The other 11% (1) predicted librarian is from the actual high school student group, as shown in Rows 10 and 11 (We evaluate this high school student as being "misclassified".). Therefore, the total number of the predicted librarians is 9, as shown in the last row of Column 3, with the total percentage beneath the number.

6. Discussion

Our empirical study shows that the librarian is the most accurate stereotype among all 11 stereotypes. None of the actual librarians was misclassified, and only one of the 56 actual student participants was predicted into the librarian stereotype. Other stereotypes have varying degrees of consistencies in representing members with respect to their IR knowledge. Along the educational and professional status dimension, the stereotype error rate tends to increase as the education level decreases. Undergraduate stereotype has the highest error rate, and therefore is the least accurate stereotype. Most of the pre-assigned undergraduates who are misclassified are predicted as the high school stereotype. A reason may be that these wrongly classified undergraduates have a low level of IR knowledge. High school student stereotype has a high error rate too. The graduate student stereotype is better than undergraduate and high school stereotypes, but still worse than the librarian stereotype.

The amount of education, and thus the IR task knowledge, may be the key to interpret the different accuracies among these stereotypes. Librarians represent those who have completed their formal profes-

Table 4
Classification results for user stereotypes on educational & professional status variable (Panel A), native language variable (Panel B), the academic background variable (Panel C), the computer experience variable (Panel D)

Actual user stereotypes		Predicted user stereotypes									
Panel A Educational and professional status		Librarian		Graduate		Undergraduate		High school		Total of actual	
		Num	Ratio	Num	Ratio	Num	Ratio	Num	Ratio	Num	Ratio
Librarian	Num	8	100%	0	0%	0	0%	0	0%	8	100%
	Ratio	89%		0%		0%		0%			
Graduate	Num	0	0%	11	61%	4	22%	3	17%	18	100%
	Ratio	0%		65%		24%		14%			
Undergraduate	Num	0	0%	2	14%	6	43%	6	43%	14	100%
	Ratio	0%		12%		35%		29%			
High school	Num	1	4%	4	17%	7	29%	12	50%	24	100%
	Ratio	11%		24%		41%		57%			
Total of predicted	Num	9		17		17		21		64	100%
	Ratio	100%		100%		100%		100%		100%	

Panel B Native Language		Native English		Non-native English		Total of actual	
		Num	Ratio	Num	Ratio	Num	Ratio
Native English	Num	25	100%	0	0%	25	100%
	Ratio	50%					
Non-native English	Num	25	64%	14	36%	39	100%
	Ratio	50%		100%			
Total of predicted	Num	50		14		64	100%
	Ratio	100%		100%		100%	

Panel C Academic background		No Major		Sci. & Engi.		Soc. Sci. & Hum.		Professional		Total of actual	
		Num	Ratio	Num	Ratio	Num	Ratio	Num	Ratio	Num	Ratio
No major	Num	13	54%	3	13%	7	29%	1	4%	24	100%
	Ratio	65%		18%		39%		11%			
Sci. & Engi.	Num	3	20%	11	73%	1	7%	0	0%	15	100%
	Ratio	15%		65%		5%		0%			
Soc. Sci. & Hum.	Num	4	24%	3	18%	10	59%	0	0%	17	100%
	Ratio	20%		18%		56%		0%			
Professional	Num	0	0%	0	0%	0	0%	8	100%	8	100%
	Ratio	0%		0%		0%		89%			
Total of predicted	Num	20		17		18		9		64	100%
	Ratio	100%		100%		100%		100%		100%	

IPM 706
24 February 2004 Disk used ARTICLE IN PRESS No. of Pages 14, DTD = 4.3.1

Table 4 (*continued*)

Actual user stereotypes		Predicted user stereotypes							
Panel D Computer experience		High		Medium		Low		Total of actual	
		Num	Ratio	Num	Ratio	Num	Ratio	Num	Ratio
High	Num	12	46%	7	27%	7	27%	26	100%
	Ratio	71%		41%		23%			
Medium	Num	3	14%	9	41%	10	45%	22	100%
	Ratio	18%		53%		33%			
Low	Num	2	13%	1	6%	13	81%	16	100%
	Ratio	12%		6%		43%			
Total of predicted	Num	17		17		30		64	100%
	Ratio	100%		100%		100%		100%	

Notes: "Num" stands for Number and "Ratio" indicates the percentages: the ratios following numbers in the same row (cross the row) are for pre-assigned stereotypes and the ratios beneath numbers (down the column) are for predicted stereotypes.

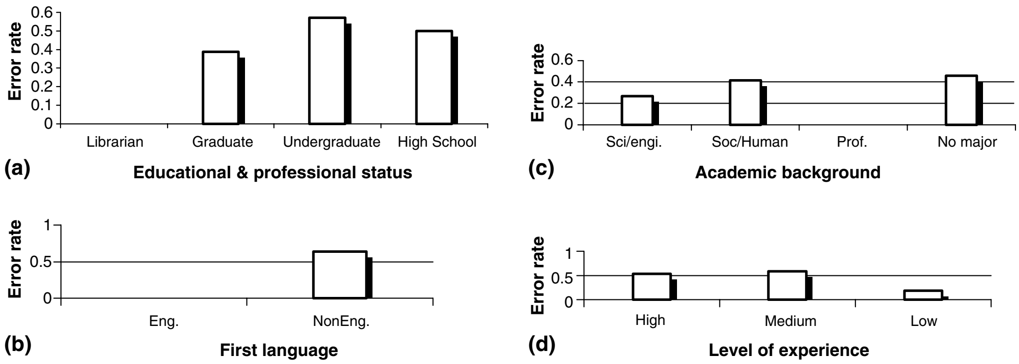

Fig. 3. (a) Error rates for user stereotypes on educational and professional status variable. (b) Error rates for user stereotypes on first language variable. (c) Error rates for user stereotypes on academic background. (d) Error rates for user stereotypes on computer experience.

sional education/training and have obtained practical experience in their field. They apparently have a high level of IR knowledge and this knowledge level seems to be the same among the members. The librarian stereotype thus consistently represents all librarians. This result matches well with the reality that librarians are trained professionals for IR tasks.

Undergraduate and high school participants have lower levels of education, and normally fewer IR skills. For example, many of them never did online searches before participating in this study. These participants appear to have various levels of IR knowledge. Graduate students seem to have more opportunities to use IR systems, and have higher level of IR knowledge.

Stereotypes on the native language dimension do not show satisfying results. Although the native English speaker stereotype has zero error rate, the predicted members for this type actually include many non-native English speakers. An explanation is that many non-native English speakers are as IR knowledgeable as native English speakers. This indicates that the stereotype based on native language may be inappropriate for user modeling.

The science/engineering user stereotype appears to be more homogeneous than the social sciences/humanities stereotype. The number of the actual social sciences/humanities participants who are predicted as other types of participants is higher than that of sciences/engineering students. This indicates that the social sciences/humanities stereotype may not be accurate for IR user modeling. Two reasons may account for the above phenomenon. One is that students in science/engineering majors may have more opportunities to access information systems and obtain more IR knowledge. They have similar levels of IR knowledge. The other reason may be that the training in science/engineering majors is more standardized than that in social sciences/humanities majors. Standardized training can decrease the difference of IR knowledge levels among students in science/engineering majors compared with those in social sciences/humanities majors.

Users' prior computer experience does not seem to help with accurate stereotyping. Both high-level and medium-level actual stereotypes have over 50% of their members predicted as members in other stereotypes on this dimension. The low-level stereotype has a high proportion of its members predicted to be in the same type. These members, however, account for only 43% of the predicted stereotype. 57% of its members are from other actual groups. A reason might be that it is hard to measure "computer experience". Different measures may generate different results.

We also observe that the medium type along a dimension (such as the medium level of computer experience between the high and low) suffers the highest error rates. This may be because that the medium type is hard to quantify. There are no very clear features/attributes to separate the medium type from the others at the ends of the dimension, which suggests that just a few very distinctive stereotypes be constructed.

It should be pointed out that when the total 64 participants were divided into different groups, the sample size was really small. Therefore, the results should be interpreted as exploratory. Larger scale studies should be conducted to confirm the findings.

7. Conclusion

This paper contributes in two ways to stereotype modeling in IR systems. First, we suggest that users' IR knowledge should be considered for stereotype modeling for IR systems, and we argue the stereotypes based on the IR knowledge can be more homogeneous and accurate than the ones that are based just on the user's explicit characteristics. We assume the system or service built for stereotypes based on users' IR knowledge would perform better than those based just on the user's demographic characteristics. This is based on the assumption that IR knowledge is important to the user's IR performance (Allen, 1996).

Second, we propose a formal method to empirically distinguish individual users in terms of their IR knowledge, and to evaluate the accuracy of a stereotype. Our method categorizes users' membership according to the demographic characteristics, and then generates users' new membership according to their IR knowledge using discriminant analysis. We compare the degree of similarity between the actual and predicted assignment to see if members of a actual stereotype have the same knowledge level, so as to study the accuracy of the stereotypes.

We conducted empirical tests on the accuracy of 11 commonly used stereotypes. The results show that our method well captures the difference among members of a stereotype, e.g., librarians and information specialists is the most accurate stereotype; other stereotypes such as high-school students have varying degree of IR knowledge among their members. The results reflect the fact that librarians and information specialists are professionally trained for IR tasks and this training makes them different from other types of users. Our method shows promise in detecting the differences among stereotype members and thus in helping developing more accurate stereotypes and individual user models. The method can be used in the design of stereotype-based systems, such as many recommender systems (Motaner, Lopez, & Lluis De La Rosa, 2003; Sollenborn & Funk, 2002), in two aspects: (a) Detecting the accuracy or homogeneity of the stereotypes that have already been developed for the system, and (b) revising and/or forming new memberships for stereotypes that can be more accurate or homogeneous (Zhang, 2003).

This study has limitations in stereotype evaluation in real applications. We would like to implement the stereotypes in a system and to evaluate the reliability of stereotypes according to the goal of the stereotypes, e.g., to empirically compare the performance of a system that uses stereotypes with and without the use of IR knowledge, or stereotypes with and without consistent representation of members' IR knowledge. It may also help to invite human experts to assign participants into various stereotypes based on the participants' concept ratings. The assignments by the experts and by the discriminant analysis can then be compared to better evaluate the stereotypes. In addition, we would like to invite more participants and experiment on a larger scale of data.

Another interesting research direction for our future work is to consider the situation when a user belongs to multiple stereotypes. This is an important research issue, since the IR knowledge a user has may relate him/her perfectly to one of the actual stereotypes based on one categorization dimension, but may raise the error rate of another stereotype on another dimension.

Acknowledgements

We wish to thank the volunteers who participated in this study, and we acknowledge the valuable comments from Paul Kantor on an earlier draft of this paper.

References

Allen, R. B. (1990). User models, theory, method, and practice. *International Journal of Man–Machine Studies, 32*, 511–543.

Allen, B. L. (1996). *Information tasks: toward a user-centered approach to information systems*. New York: Academic Press.

Bannister, D. (1968). *The Evaluation of Personal Constructs*. London and New York: Academic Press.

Beaumont, I. H. (1998). User modelling in the interactive anatomy tutoring system ANATOM-TUTOR. In P. Brusilovsky, A. Kobsa, J. Vassileva, (Eds.), *Adaptive hypertext and hypermedia* (pp. 91–115).

Bellika, J. G., Hartvigsen, G., & Widding, R. A. (1998). The virtual library secretary: a user model-based software agent. *Personal Technologies, 2*, 162–187.

Boyce, B. R., Meadow, C. T., & Kraft, D. H. (1994). *Measurement in information science*. San Diego: Academic Press.

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* (pp. 109–118).

Charoenkitkarn, N. (1996). The effect of markup-querying on search pattern and performance in large-scale text retrieval. *Ph.D. Dissertation*. Department of Industrial Engineering, University of Toronto.

Chin, D. N. (1989). KNOME: modeling what the user knows in UC. In A. Kobsa, W. Wahlster, (Eds.), *User models in dialog systems* (pp. 35–51).

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation, 49*(4), 356–369.

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.

Fransella, F., & Bannister, D. (1977). *A manual for repertory grid technique*. London: Academic Press.

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*. New York: John Wiley & Sons.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

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* (pp. 99–105).

Kuflik, T., Shapira, B., & Shoval, P. (2003). Stereotype-based versus personal-based filtering rules in information filtering systems. *Journal of the American Society for Information Science and Technology, 54*(3), 243–250.

Latta, G. F., & Swigger, K. (1992). Validation of the repertory grid for use in modelling knowledge. *Journal of the American Society for Information Science, 43*(2), 115–129.

Ma, W. (2002). A database selection expert system based on reference librarian's database selection strategy: a usability and empirical evaluation. *Journal of the American Society for Information Science and Technology, 53*(7), 567–580.

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.

Motaner, M., Lopez, B., & Lluis De La Rosa, J. (2003). A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review, 19*, 285–330.

Mostafa, J., Quiroga, L. M., & Palakal, M. (1998). Filtering medical documents using automated and human classification methods. *Journal of the American Society for Information Science, 49*(14), 1304–1318.

Mulaik, S. A. (1972). *The Foundations of factor analysis* (p. 174). McGraw-Hill.

Neuman, D. (1995). High school students' use of databases: results of a national Delphi study. *Journal of the American Society for Information Science, 46*(4), 284–298.

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.

Qiu, L. (1993). Markov models of search state patterns in a hypertext information retrieval system. *Journal of the American Society for Information Science, 44*(7), 413–427.

Rich, E. (1979). Users modeling via stereotypes. *Cognitive Science, 3*, 329–354.

Rich, E. (1989). Stereotypes and user modeling. In A. Kobsa & W. Wahlster (Eds.), *User models in dialog systems* (pp. 35–51). Berlin: Springer-Verlag.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving: I. Background and methodology; III. Searchers, searches, overlap. *Journal of the American Society for Information Science, 39*(3), 161–176, 197–216.

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.

Shapira, B., Shoval, P., & Hanani, U. (1999). Experimentation with an information filtering system that combines cognitive and sociological filtering integrated with user stereotypes. *Decision Support Systems, 27*, 5–24.

Shaw, D. (1996). Undergraduate use of CD-ROM databases: human–computer interaction and relevance judgments. *LISR* (18), 261–274.

Shaw, M. L. G. (1980). *On becoming a personal scientist*. London: Academic Press.

Shaw, M. L. G., & Woodward, J. B. (1988). Validation in a knowledge support system: construing and consistency with multiple experts. *International Journal of Man–Machine Studies, 29*, 329–350.

Sollenborn, M., & Funk, P., (2002). Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems. In *Proceedings of the 6th European conference on case based reasoning* (pp. 395–405).

Su, L. T. (2003). A comprehensive and systematic model of user evaluation of web search engines: II. An evaluation by undergraduates. *Journal of the American Society for Information Science and Technology, 54*(13), 1193–1223.

Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. In *Proceedings of TREC 2003* (pp. 14–27).

Yee, H. I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161–174.

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 and Technology, 52*(6), 445–459.

Zhang, X. (2003). Discriminant analysis as a machine learning method for revision of user stereotypes of information retrieval systems. In S. A. Macskassy, et al. (Eds.), *Workshop proceedings of machine learning, information retrieval and user modeling, 9th international conference on user modeling* (pp. 1–10).