Combining stereotypes and a cognitive theory for human-like advice in an intelligent GUI

Maria Virvou
Department of Informatics, University of Piraeus
Piraeus, Greece
mvirvou@unipi.gr

Katerina Kabassi
Department of Informatics, University of Piraeus
Piraeus, Greece
kkabassi@unipi.gr

Abstract.
IFM is a Graphical User Interface (GUI) that incorporates intelligence. The research described in this paper focuses on the new User Modeling (UM) component of IFM. This component employs a novel combination of user stereotypes with a cognitive theory in order to generate hypotheses about users’ beliefs and intentions. IFM uses this reasoning while monitoring every user action so that it can provide spontaneous assistance to users who are having problems with their interaction. The stereotypes have been built based on the results of an empirical study that involved a variety of users. As a result, a two-dimensional stereotype has been identified classifying users into novice, intermediate and expert in one dimension and careful or careless in the second dimension. These stereotypes affect the values of the certainty parameters attached to hypotheses about users that are generated using the cognitive theory. Finally, an evaluation of IFM revealed that the UM was quite successful at generating hypotheses concerning users in comparison to the hypotheses of human experts.

Keywords: User Modelling, Intelligent User Interfaces, Human Computer Interaction.

1. Introduction
As the number of software users increases rapidly, the need for more efficient on-line help becomes more apparent. However, traditional on-line help is not always sufficiently helpful. For example, as Matthews et al. (2000) point out, on-line manuals must explain everything and novices find them confusing, while more experienced users find it quite annoying to have to browse through a lot of irrelevant material. Furthermore, one important problem that has been revealed by empirical studies (Virvou et al. 2000, Virvou & Kabassi 2000) is that users do not always realize that they have made an error immediately after they have made it. Therefore, they may lose valuable time and energy to recover from their errors.

In order to deal with these problems we have developed a Graphical User Interface that monitors users’ actions, reasons about them and provides spontaneous assistance in case it diagnoses a problem. The GUI is called IFM (Intelligent File Manipulator) and is meant to operate for the management of a user’s file store, like the Microsoft Windows Explorer (Microsoft Corporation 1998). IFM keeps an individual, long-term user model that is updated every time the user interacts with the system. The use of new commands by the user triggers the update of the user model.
The system uses Human Plausible Reasoning (Collins & Michalski 1989) (henceforth referred to as HPR) and its certainty parameters in order to make inferences about possible users’ errors based on evidence from users’ interaction with the system. HPR is a cognitive theory about human reasoning and has been used in IFM to simulate the users’ reasoning, which may be correct or incorrect (but still plausible) and thus may lead to “plausible” user errors. The adaptation and incorporation of HPR into an intelligent GUI is quite important since it aims at rendering the user interface more human-like.

However, in previous versions of IFM one problem that came up was that no sufficient information about the user could be obtained before s/he had interacted with the system for quite a long period of time. A solution to this problem was the incorporation of stereotypes to provide default assumptions about users until the user model acquired sufficient information about each individual user. Indeed as Rich (1989) points out a stereotype represents information that enables the system to make a large number of plausible inferences on the basis of a substantially smaller number of observations; these inferences must, however, be treated as defaults, which can be overridden by specific observations.

Stereotypes constitute a powerful mechanism for building user models (Kay 2000). Therefore, stereotypes have been widely used in advisory software (Ardisson & Goy 1999, Rich 1999, Chin 1989, Wilensky et al. 2000). IFM is primarily concerned with the integration of stereotypes with HPR, which is an inference mechanism for plausible reasoning. This novel combination aims to simulate the reasoning used by human experts who act as advisors while watching the user over the shoulder.

2. Human Plausible Reasoning

Human Plausible Reasoning theory (Collins & Michalski 1989) is a cognitive theory that attempts to formalize plausible inferences that occur in people’s responses to different questions when the answers are not directly known. The theory is grounded on an analysis of people’s answers to everyday questions about the world, a set of parameters that affect the certainty of such answers and a system relating the different plausible inference patterns and the different certainty parameters. For example, if the question asked was whether Greece produces oil and the answer was not directly known to the person asked then the answer would depend on the knowledge retrieved from memory. If the subject knew that the climate of Greece is similar to another country where olive trees grow, this would trigger an inductive, analogical inference, and generate the answer yes.

According to the theory a large part of human knowledge is represented in “dynamic hierarchies”, which are always being updated, modified or expanded. This way, the reasoning of people with patchy knowledge can be modelled. Statement transforms, the simplest class of inference patterns, exploit the 4 possible relations among arguments and among referents to yield 8 types of statement transform. Statement transforms can be affected by certainty parameters. The degree of similarity ($\sigma$) signifies the degree of resemblance of a set to another. The degree of typicality ($\tau$) represents how typical a subset is within a set (for example, cow is a typical mammal). The degree of frequency ($\varphi$) count how frequent a referent is in the domain of the descriptor. Dominance ($\delta$) indicates how dominant a subset is in a set (for example, elephants are not a large percentage of mammals). Finally the only certainty parameter applicable to any expression is the degree of certainty ($\gamma$) or belief that an expression is true.

3. Intelligent File Manipulator

Intelligent File Manipulator (IFM) is an Intelligent Graphical User Interface that works in a similar way as a standard file manipulation program, such as Windows 98/NT Explorer but it also incorporates intelligence. IFM is meant to help users during their navigation and manipulation of the file store and provides advice in case this is considered necessary. In general, IFM tries to act as a human expert who watches the user over the shoulder and offers advice spontaneously. The way IFM works is described as follows:

The user issues a command and the system reasons about it in order to categorize it in one of four categories, namely “expected”, “neutral”, “suspect”, “erroneous”. A command is categorised as expected if it is compatible with the user’s hypothesized goals. It is considered suspect if it contradicts the system’s hypotheses about the user’s goals and erroneous if the command is wrong with respect to the user interface formalities. The command is
considered neutral if it cannot be assigned to one of the former categories. If the action is categorised as expected or neutral, it is executed normally. However, if the action is categorised as suspect or erroneous then the system tries to generate alternative actions that the user may have meant to issue instead of the one issued. Therefore, the action issued is transformed so that similar alternatives can be found, which would not be suspect or erroneous. The transformations are based on HPR. However, since the transformed action has to fit better in the context of the user’s goals, the system reasons about every alternative action generated from the transforms. As a result, each transformed action is categorized in one of the four categories in a similar way as the actual command issued by the user.

As Mitrovic et al. (1996) point out, there may be different explanations of observed incorrect user’s actions. Therefore, there is a need to attach priorities to different explanations so that some may be preferred over others. For example, expected actions have priority over neutral ones. Additionally, IFM uses the certainty parameters of HPR to determine the priority among actions belonging to the same category.

Certainty parameters of HPR have been adapted to fit IFM’s requirements. In IFM, the degree of similarity ($\sigma$) is used to calculate the resemblance of two commands or two objects. The typicality ($\tau$) of a command represents the estimated frequency of execution of the command by the particular user. The degree of frequency ($\phi$) of an error represents how often a specific error is made by a particular user or the users of a particular stereotype. The particular user’s weaknesses can be recognized by the dominance ($\delta$) of an error in the set of all errors. Finally, a degree of likelihood is related to every alternative command generated by IFM. This is called the degree of certainty ($\gamma$) and represents the system’s certainty that the user intended the alternative command generated. This degree is calculated as a sum of all certainty parameters, with each parameter being multiplied to a weight, which is determined with respect to how important the particular certainty parameter is. However, the way of calculation of the degree of certainty is not fully specified in HPR. In order to specify the exact way of calculation of the degree of certainty we conducted an empirical study.

The empirical study involved both expert and novice users interacting with a standard file manipulation program and human experts acting as advisors (Virvou & Kabassi 2000, Virvou & Kabassi 2001). During the empirical study users were asked to work as they would normally do and their actions were recorded. Then the protocols of user interactions with a standard file manipulation program were given to human experts to comment on them. The study of these comments revealed the important aspects that human experts were taking into account when they reasoned about users’ actions in order to give advice. Such advice was the suggestion of an alternative action to users who had issued an erroneous action with respect to their hypothesized intentions. However, for an expert to suggest an alternative action, s/he had to evaluate candidate alternative actions in order to select the most appropriate one.

The most important criterion of a human expert when evaluating an alternative action, which was going to be proposed to the user, was the similarity of that action to the one issued by the user, because users usually tend to tangle up actions or objects that are very similar. Thus, the weight of the degree of similarity is estimated to 0.4, which is the largest weight of all. The second most important criterion that human experts used was whether a particular user’s error was the most frequent error of all errors that this user made. However, the human experts pay more attention whether a particular error is the most important error of the user or not. So the weight of dominance of the particular error in the set of all errors is 0.3. Furthermore, a very important criterion when evaluating an alternative action was the frequency the user makes such an error. The degree of frequency of the particular error is multiplied to 0.2.

Finally, when proposing an alternative action to a user, the system must know if a user uses that particular action quite often or not. It would not be likely that the user had made a mistake in the execution of a command that s/he uses quite often and thus probably knew how to use correctly. However, still there is a possibility that the user may have made a carelessness mistake in such a command. Therefore, the typicality of a certain command for the particular user is taken into account but is not considered of primary importance. Hence, it is multiplied to just 0.1.
In view of the above the proposed formula for the calculation of the degree of certainty is presented below.

\[ \gamma = 0.4 * \sigma + 0.3 * \delta + 0.2 * \varphi + 0.1 * \tau \] (1)

An example of interaction of a user with IFM is described below. The user’s initial file store state is shown in figure 1. The user intends to delete A:\conferences\conference1\. However, s/he accidentally attempts to delete A:\conferences\conference2\. In this case, s/he runs the risk of losing the information stored in the two files that the folder A:\conferences\conference2\ contains.

Figure 1. The user’s initial file store state

IFM would suggest the user to delete the folder A:\conferences\conference1\ because the folder A:\conferences\conference1\ is empty whereas the folder A:\conferences\conference2\ is not and the two folders have very similar names therefore one could have been mistaken for the other.

4. User Stereotypes for Initializing Certainty Parameters

Stereotypes may serve as a tool to model the beliefs and preferences that the users of a system may have. The main reason for the application of stereotypes is that they provide a set of default assumptions, which can be very useful during hypotheses generation about the user. Generation of default assumptions can prove very effective for modeling a large proportion of users. However, this approach has many problems as well. For example, despite the similar behavior that users of the same group may have, every one is an individual that differentiates in many aspects from all the others. Therefore, stereotypes should be used for initializing the user model, until there is more information.

All default assumptions in stereotypes used in IFM, give information about the errors that users belonging to this category usually make. Information about each error is expressed by using the certainty parameters of HPR theory. So, we used frequency to show how often users belonging to a certain group make a particular error. Another piece of information that can be derived from a stereotype is the weak point of a user that belongs to this particular stereotype. The weak point is expressed as a number representing the dominance of a particular error within the set of all errors for users belonging to the particular stereotype. Finally, the typicality shows how typical a command is in the set of all the commands that a user has issued.

The empirical study that we conducted revealed that users could be classified into three major classes according to their level of expertise, namely, novice, intermediate and expert. Each one of these classes represents an increasing mastery in the use of the particular file manipulation system. Such a classification was considered crucial because it would enable the system to have a first view of the usual errors and misconceptions of a user, belonging to a group. One might not expect expert users to make mistakes but this does not correspond to reality. There are some experts that are very prone to mistakes because of their carelessness. As a result, another classification that was considered rather important was dividing users into two groups, careless and careful.

A stereotype usually has a set of trigger events. Trigger events are boolean expressions that once one of them becomes true, the corresponding stereotype must be activated for a particular user. IFM infers information about the user by watching him/her during his/her interaction with the system. However, the system cannot decide where to categorize a user before s/he has executed a satisfactory number of commands. The empirical study revealed that an early conclusion could be drawn only after the execution of thirty commands. A novice user will have made a large number of errors after the execution of thirty commands, whereas an expert in the same amount of commands will have made only a few errors that can be regarded as accidental slips. By command errors, we mean cases where the user has selected the wrong command with respect to his/her hypothesized intentions or cases where a command has failed to be executed. By structure errors, we mean errors that reveal the user’s unawareness of the structure of a standard file manipulation program. For example, when the user confuses
the parent folder at the left part of the Explorer with the folder shown at the right part of the program. Triggers for categorizing the users according to their carelessness when executing certain tasks are constructed similarly to the triggers of the stereotypes that correspond to the user’s proficiency.

<table>
<thead>
<tr>
<th>Stereotype</th>
<th>Frequency</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOVICE</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>StructureError</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>INTERMEDIATE</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>StructureError</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>EXPERT</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>StructureError</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Default assumptions of stereotypes concerning the frequency and dominance of types of error relating to the level of expertise of a user.

The default assumptions for users of the novice, intermediate and expert stereotype are presented in Table 2. We can see that novice users are more prone to command errors, which are considered to be their weak point, rather than structure errors. Intermediate users also commit such errors, although these are not usual errors for them. On the other hand, expert users do not make such errors.

<table>
<thead>
<tr>
<th>Stereotype</th>
<th>Frequency</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARELESS</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>SpellingErrors</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>MouseError</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>IdenticalNamesError</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>CAREFUL</td>
<td>0.1</td>
<td>0.45</td>
</tr>
<tr>
<td>SpellingErrors</td>
<td>0.1</td>
<td>0.45</td>
</tr>
<tr>
<td>MouseError</td>
<td>0.1</td>
<td>0.45</td>
</tr>
<tr>
<td>IdenticalNamesError</td>
<td>0.025</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 2. Default assumptions relating to representing the carelessness of a user.

Stereotypes that classify users according to their degree of carelessness include default assumptions about the errors made due to carelessness. For example, a user, who is considered by the system as careless, usually makes 30% mouse errors, 30% spelling errors and only a few errors are due to confusion between objects with similar names. In Table 5, one can see the default assumptions for careless and careful stereotype. By mouse errors we mean that the user has tangled up neighboring objects in the graphical representation and by confusion between objects with similar names we mean that the user confused objects with exactly the same name that were situated in different places in the file store, etc.

5. User Modeling by Combining Stereotypes and Individual History of Users

As mentioned in the introduction, stereotypes are used in IFM only for capturing the initial impression of a user. A stereotype is activated after implicitly acquiring information by observing the user while interacting with the system. The system cannot generate hypotheses before the user has executed a satisfactory number of commands. After a stereotype has been activated, the system makes some default assumptions about users’ possible errors and can provide some kind of advice.

In the beginning, information is acquired only by the stereotype. However, the system is also constantly collecting information about a particular user’s behavior and errors and informs the individual user model of the user. As the system collects more and more evidence about a user, information is acquired in part by the stereotype and in part from the individual user model. The percentage of information acquired by the stereotype diminishes as the percentage of acquisition by the individual user model increases.

In case a conflict appears, the system always lays more weight on the information acquired from the individual user model. For example, if the stereotype supports that a user’s weak point is in command errors and the individual user model supports that the particular user’s weak point is in structure errors, then the system will consider that the user’s weak point is the one proposed by the individual user model.

6. Evaluation

One of the most important challenges for intelligent user interfaces is to prove that their adaptive behavior does in fact improve the interaction with the user. Only through designing useful adaptations and then evaluating them with users can we be sure that we are solving the right problem (Höök 2000). Therefore, an important aim of the evaluation of IFM was to find out how successful the system was at initializing the user model similarly to human experts who observed the interaction.
For the above purposes, we selected 30 users, who had diverse backgrounds and, therefore, constituted a representative sample. The users were asked to work with IFM as they would normally do with a standard file manipulation program. IFM constantly monitored every user's action and after the users had executed 30 commands, the system had categorized the user into a stereotype concerning his/her level of expertise and another stereotype related to his/her degree of carelessness. The users were also video-recorded and the videos were given to 10 human experts. All human experts were asked to watch all videos and then decide what the corresponding stereotype for each user was.

There was unanimity of human experts' opinion in 30% and 40% of the cases, concerning the level of expertise and degree of carelessness, respectively. IFM had the same opinion in 24% and 33% of those cases, respectively. However, there were cases where human experts had conflicting opinions. In those cases IFM's selection was compared to that of the majority of human experts. However, human experts could not make a decision in 16.67% of the cases where they were asked to comment the user's level of expertise and in 6% of the cases concerning the degree users' carelessness. On a balance, concerning the user's level of expertise, the degree of compatibility between the majority of human experts and IFM was 63.33%, which is quite satisfactory. This percentage was even greater (66.67%) when human experts and IFM's categorized the user accordingly to his/her level of expertise.

6. Conclusions

In this paper we described an intelligent graphical user interface for a program that manipulates files, called IFM. The main focus of this paper has been on the incorporation of stereotypes of users in user modeling and the combination of these stereotypes with principles of a cognitive theory called Human Plausible Reasoning. In particular the inferences made by the stereotypes concern some certainty parameters that are defined by the theory and are adapted to be used in IFM. Depending on the values of these certainty parameters IFM selects the most appropriate alternative command to be suggested to a user in case s/he has issued a problematic one. Stereotypes in IFM have been built based on an empirical study so that the inferences which would be based on them could be similar to human experts. However, IFM keeps a detailed history of each user and when information collected is enough, it makes inferences based on this individual long term user model.

The results of the evaluation of IFM in comparison with human experts in terms of the initialization of the user model were quite encouraging. Indeed, IFM's reasoning in the categorization of users in stereotypes was very similar to that of the human experts. Furthermore, the combination of Human Plausible Reasoning theory with stereotypical knowledge is quite important for the adaptation and application of the theory to individual user modeling. In this way, the interaction is made more human-like in the sense that the system may provide to a large extent automatic assistance similar to a human advisor who watches a user over the shoulder.

References