Contextualised Student Modelling for Enculturated Systems

Phaedra Mohammed and Permanand Mohan

Department of Computing and Information Technology, The University of the West Indies, St. Augustine, Trinidad and Tobago
{phaedra.mohammed@gmail.com, Permanand.Mohan@sta.uwi.edu}

Abstract. Contextual student modeling, also called cultural profiling or cultural modeling, refers to the process of building a computational representation of the cultural identity and background of a student. Previous works have been done that identify and use certain environmental dimensions for such a model. In this paper, a new approach is proposed that uses additional dimensions, and incorporates combinations of dimension clusters to represent and quantify a student's expression of socio-cultural group traits and preferences. The viability of this approach is demonstrated through the use of a prototype that collects dimension data and generates estimates of a student's association with particular socio-cultural groups in five categories. An evaluation of the prototype revealed that estimates were rated as reasonable and acceptable by students and confirms that the approach extends current efforts in the field of culturally-aware tutoring systems for modeling student's cultural context.

Keywords: contextual student model, cultural element, dimensions

1. Introduction

Contextual student modeling, also called cultural profiling or cultural modeling, refers to the process of building a computational representation of the cultural identity and background of a student. This identity is shaped by many dimensions that originate from an individual level such as personal demographics and from a group level such as religious or ethnic influences. The first challenge that arises in contextual student modeling is identifying which dimensions should be modeled, and determining to what extent a dimension affects a student's personality, preferences, and opinions. The second challenge that arises is whether combinations of these dimensions can be worked out such that a student's expression of particular traits and values, shared by a cultural group, are represented and measured relative the group's expression of said traits and values. The final challenge that arises in contextual student modeling is evaluating whether a computational model generated for a student is a reasonable and acceptable representation of the student's particular cultural identity and background.

This paper tackles all three challenges in a systematic manner by looking at culture as a form of context. When culture is looked at as context or rather as a focused collection of metadata, these challenges becomes more tractable and the issues that need

to be dealt with start to take on a computational form. The environmental context of an individual is therefore made up of several dimensions of metadata. These contextual dimensions fall into two groups: contextual factors and contextual influences. A contextual factor is something that brings about a particular effect on an individual and can be quantified discretely. A contextual influence is something that brings about a particular effect on an individual but whose exact nature is not readily known and can take on a range of values.

Several key ideas in this paper are based on the works of Blanchard, Mizoguchi, and Lajoie [3] who define the concept of cultural elements and cultural groups. A contextual element is considered to be a type of cultural element. It is an observable manifestation of culture and can be present in educational content expressed as different forms of media (text, pictures, videos, and audio). A contextual group on the other hand is a collection of individuals sharing similar values for contextual dimensions. Contextual groups and individuals are related by these contextual dimensions. The strength of this relationship is determined by the amount of overlap of dimension context and by the individual's expression of particular dimensions in the intersection. These definitions are central to the approach taken in this paper for dealing with the challenges outlined earlier with the goal of defining a contextual student model.

The paper is organized as follows. Section 2 identifies twenty four dimensions of context for a contextual student model (CSM) based on related research, and explains the rationale for the new dimensions identified in this paper that have not been used in culturally-aware tutoring systems (CATS) research before. Section 3 discusses how these dimensions were clustered based on relevance to particular contextual groups, for the purpose of generating estimates of a student's level of membership to five contextual groups. Section 4 then describes the ontology-based design of the CSM and the implementation of a rule-based approach for generating contextual estimates. Section 5 outlines experiments that were conducted to evaluate the CSM design and performance together with the results of these experiments. Section 6 gives an analysis of the results and the paper concludes in Section 7 with the future plans for the CSM.

2. Environmental Context: Factors and Influences

Several dimensions have been recurring in the literature as having an effect on students from a cultural perspective. The most common ones include age, gender, nationality, native languages, religion, ethnicity, emotional disposition, and locations of residence and study [4, 5, 6, 8, 9]. Of these characteristics, some are quantifiable and can be considered to be contextual factors such as age, gender, nationality, and locations of residence and study. The remaining traits and qualities such as ethnicity and religion are less easily quantified and are therefore considered to be contextual influences. A good rule of thumb for distinguishing between a factor and an influence is the answer to the following question: For a given characteristic C, how much of a C is the student in question? If the answer can be within a range of potential values then that characteristic is most likely an influence otherwise it is a factor.

Twenty four contextual dimensions have been identified for the CSM based on the works of [1, 5, 9]. The first set of dimensions for the CSM consists of personal fac-

tors: age, gender, country of birth, the locale[1] where the student lives, and the schools where the student has studied at primary, secondary and tertiary level. In order to model the historical context of a student, the CSM includes three school-related dimensions that identify locales which would have shaped a student's context over the duration of his/her time in school. The AdaptWeb project [5] uses characteristics similar to the locales of study but their work manipulates IP addresses to identify only one current locale of study for the student. The second set of CSM dimensions consists of personal influences: the student's religion, ethnicity, and native language. Religion influences have been used in [2], language influences have been used in ActiveMath [7] and ethnicity influences have been used in embodied conversational agents [10]. The CSM combines and reasons about the student's context using all three influences since the combination changes the individual impact of a particular influence and can affect the student's perception, interpretation and magnitude of response to a particular contextual element.

The third set of CSM dimensions originate from social units surrounding the student, in this case the student's parents. This is based on the work of Reinecke, Reif, and Bernstein [9] who identified that parents have an impact on users specifically through their language and nationality. The factors in this set include the parents' occupations, their occupation locales, and their ancestral home locales. This kind of context has not yet been used computationally in CATS. The reasons for including these factors stem from the assumptions that students typically visit their parents' workplaces, can be influenced educationally by the kinds of occupations that their parents have, and may frequent the locales where their parents grew up because of existing familial ties to the areas. This leads to the influences in this set which include the parents' religions, ethnicities, native languages, and level of personal influence on the student. The first three are self-explanatory but the strength of their impact depends on the fourth influence. Blanchard [1] discussed the situation of socio-cultural groups affecting the receptivity of individuals to particular cultural elements. The level of personal influence that a parent has on a child affects the child's involvement, beliefs, understanding, and behaviour regarding religion, ethnicity and language. This is therefore an example of socio-cultural group influence at a finer level of granularity and consequently, these dimensions were included in the CSM in order to separate, quantify and structure as best as possible the nature and the strength of control that a parent's context may have on shaping the student's context.

3. Contextual Student Model (CSM) Estimates

The dimensions in the CSM fall into five categories that describe particular contextual groups: geographical groups, religious groups, ethnic groups, groups that share similar education levels, and groups that are familiar with particular physical environment settings and terrains. The CSM generates estimates for each group using a

[1] A locale is considered to be a city, town, village, or hamlet that is officially recognised in a country.

combination of multiple dimensions because individual dimensions have been shown to have limited predictive capabilities when considered in isolation [4].

Geographical estimates are produced using the locale-based dimensions: the locales of the student's residence, parents' ancestral homes, parents' jobs and the student's schools. Two geographical estimates measured as ordinal and cardinal points are produced for the student: a dominant geographic region and a secondary geographic region based on which areas of the country his/her activities most frequently take place. Religious estimates are produced using the religion-based dimensions: the religion of the student, parents, and schools (if any), and the parents' level of influence on the student. Two religious estimates measured as percentages are produced for a student, a dominant religious influence and a secondary religious influence. The dominant influence would be derived from the religious group that student belongs to whereas the secondary influence would be based on the remaining dimensions. A secondary religious influence does not imply that the student belongs to that religious group but rather that the student is aware of that religious group and would have a partial membership because of that awareness. Schools in a country can have either no religious influence if they are non-religious or can influence student knowledge of the norms and practices of a particular religious group if the school is denominational.

Ethnicity estimates are produced using the ethnicity-based dimensions: the ethnicity of the student, parents, and the national ethnicity distributions for the student's residence locale. The distributions are used to approximate the influence on the student of the two largest ethnic groups in his/her locale. Two ethnicity estimates measured as percentages are produced here as well where the dominant ethnicity influence corresponds to the student's ethnicity and the secondary influence would be based on the parent's ethnicities and degree of influence that the parents have on the student. Educational estimates are produced using the schools attended by the student and the national educational statistics for the student's residence locale with the possible values of high, mid-high, mid, mid-low or low. This estimate reflects the level of education of the societal unit in the student's geographical region and does not mean that the student has a low or high level of education. This estimate allows the CSM to gauge how familiar a student would be with different levels of language. Low to mid-low educational estimates imply that more colloquial language would be commonly used by members of society in that particular area compared to more formal language for areas with mid-high to high levels. It is of note to mention that the parents' occupations are suitable factors for this estimate but were not included at this time.

Terrain or setting estimates are produced using the locales of the student's residence, student's secondary school, parents' ancestral homes, parents' jobs, and the parents' level of influence on the student. Three terrain/setting estimates are produced and each estimate may contain one or more categories with percentages of membership. Economic activity context captures whether a student's locale is influenced by industrial, residential, commercial, agricultural or sporting activities. Terrain context captures the type of physical environment the student may be familiar with such as coastal, desert, grassland, mountainous, forested, tundra or wetland terrains. These are based on his/her dominant geographic influences in the country. Urban/rural/semi-rural context deals with the population density of the student's locale. Together, these three areas contribute towards the terrain/setting estimates for a student. Overall, the

five categories of estimates are related to the student's contextual identity through specific combinations of contextual dimensions in the CSM and model the degree of a student's membership to a particular contextual group.

4. CSM Design and Implementation

The CSM was implemented using Java and JESS (Java Expert System Shell) and has an ontological design but was implemented using a rule-based approach for prototyping. Figure 1 below shows the main concepts and relationships in the CSM.

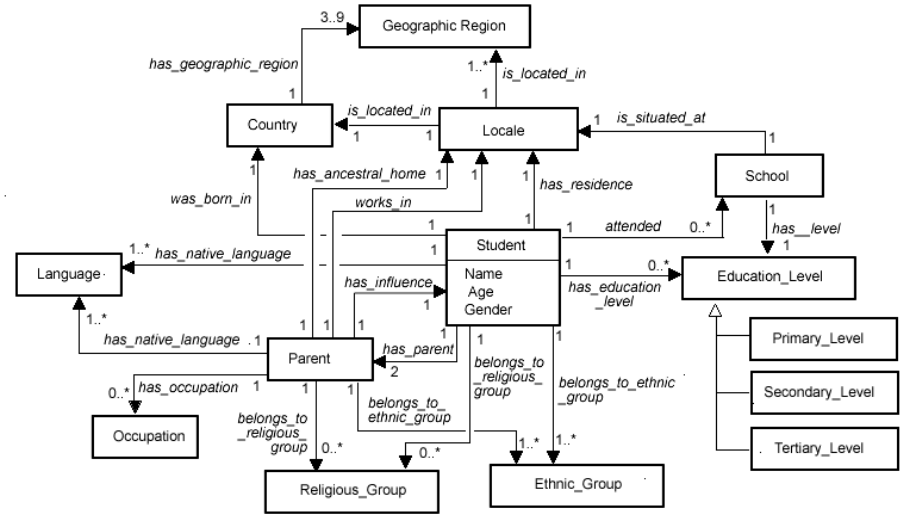

Fig.1. Metadata Structure of the Contextual Student Model

All of the concepts are not shown in the diagram due to space constraints. Each of the twenty four contextual dimensions described in Section 2 are included in the CSM and are supplemented with statistical data from the target country's national statistical office. Data on schools, locales, ethnic groups and their distributions, religious groups and their distributions, population distribution, economic activities across locales, terrain and physical data for locales were loaded into the CSM and used to generate the estimates described in Section 3. Values for the dimensions are sourced from either the student or from the target country's national statistical office. For example, the values for locale would be selected from the list of locales situated in the target country recorded by the national statistical office for the country. Similarly, the value for religion would be selected from the list of religious groups common in the target country as recorded by the statistical office. The use of country-level data to define the value spaces for some of the dimensions allows subtle nuances and variations in naming conventions for these values to be considered. Furthermore, compared to asking the opinion of a few members of a target country, the national records provide a more comprehensive, objective snapshot of the possible values that a dimension can take.

The research in this paper builds upon the approach elaborated in [2] for quantifying a student's membership to a contextual group. Blanchard [2] measured this relationship as a membership score dynamically calculated as the weighted difference between the student's characteristics and those of a contextual group. Our approach also uses weighted values but differs in the calculation of the membership score and the determination of weights. The weights in our approach are applied to contextual influences and are based on two sources of data: parent's level of influence and country level statistical data. This improves upon the approach in [2] by using weights directly related to the student's context. This means that the CSM would strengthen one student's contextual group membership for a particular category and weaken the same membership for another student as their weights change based on the significance of a dimension for their particular cluster context. If two students have similar contexts but different parental influences for example then their estimates would vary. The same holds true for different statistical distributions for the dominant influences in their contextual categories. In this case, further information is derived from a dimension using statistical data from the central or national statistical office in the country where the students reside for the course of their studies. In doing so, the socio-cultural group contexts of the social units relevant for the students are factored into the estimates. These two features advance the calculation described in [2]. Furthermore, the definition of groups that relate to contextual dimensions and elements in this paper extend content manipulation beyond the educational dimensions used in [2].

There are several potential uses envisioned for the CSM, and these hinge on adaptation at the application layer of CATS environments. One use could involve the dynamic selection of contextual elements deemed suitable for adapting learning content based on the values and estimates in the CSM. Here, the contextual elements that appeal most to students could be inserted into educational content thereby producing contextualised content. Another use of the CSM could involve the generation of contextualised instructional feedback with emotive qualities. Affective feedback generated using casual or formal varieties of language as defined by the CSM could be used to elicit different emotive responses in students in accordance with instructional goals.

5. CSM Evaluation and Results

Two studies were conducted in response to the research challenges posed at the beginning of the paper using the CSM. The first study evaluated the likelihood that the data required for generating a contextual student model will be readily supplied by users. The second study evaluated the acceptability of the estimates produced by a CSM application, built for the context of Trinidad and Tobago, based on student ratings of the estimates. This section describes the methods and results of each study.

5.1 Likelihood of Data Collection for the CSM

An online questionnaire was administered to thirty six participants (36) from a cross section of the population in Trinidad. It consisted of questions dealing with a participant's willingness to supply information on a contextual dimension. Participants

were asked to answer whether they would be willing to supply information, uncomfortable but willing to supply information, or unwilling to supply information for each of the twenty four dimensions in the CSM. Figure 2 shows the number of responses categorised by user willingness and comfort to supply contextual data. Out of 864 responses, 786 responses were classified as willing and comfortable (91%), 49 responses were classified as willing but uncomfortable (5.7%) and 29 responses were classified as unwilling (3.3%). Overall, the majority of users were willing and comfortable to supply contextual data on themselves and their social units (parents).

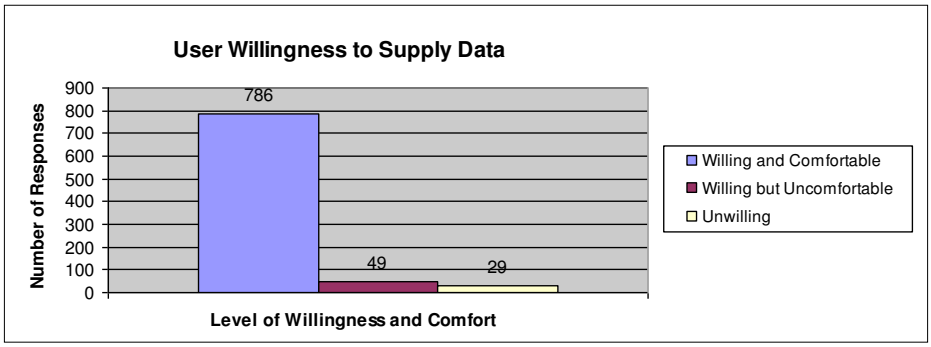

Fig.2. User Willingness to Supply Data for a Contextual Student Model

5.2 Acceptability of Contextual Estimates Generated by the CSM

Thirty (30) undergraduate students enrolled in a programming course at UWI voluntarily participated in the experiment. The students ran the CSM application which prompted for data for each of the twenty four factors. Using this data, the CSM application produced estimates of contextual influences in the following areas: geography, religion, ethnicity, education, and physical setting. Students were asked to rate the estimates for correctness using a four point Likert scale rating. Usage logs were stored and retrieved from a server for analysis.

```
GEOGRAPHY
Dominant  Geographic Region: South
Secondary Geographic Region: North

SETTING
You are familiar with the following settings:
Urban/Rural/Semi-Rural Settings: URBAN
Economic Activity Settings: INDUSTRIAL,RESIDENTIAL,COMMERCIAL
Terrain Settings: MOUNTAINOUS
```

Fig.3. Sample of CSM Estimates Generated for a Student

Figure 3 shows a sample of the geographic and the terrain/setting estimates generated for a student who lives in an industrialized hilly city in the southern part of Trinidad. The student rated the setting estimate as correct but rated the geographical esti-

mate as mostly wrong even though one of his parents' ancestral homes and work location were situated in the north of the country. The graph in Figure 4 below shows the relative differences in student ratings of the accuracy of the contextual student model estimates that were produced. When ranked in order of increasing accuracy as being either correct or mostly correct the categories are as follows: setting (80%), religion (87%), geography (90%), ethnicity (93.3%), and education (96.7%). The most inaccurate estimates (wrong and mostly wrong) were in the setting category (20%) followed by the religion category (13.3%), and then the geography category (10%). All categories of estimates were rated on average as correct or mostly correct by over 80% of the students. Collectively the estimates were rated as being 89.3% accurate and 10.7% inaccurate.

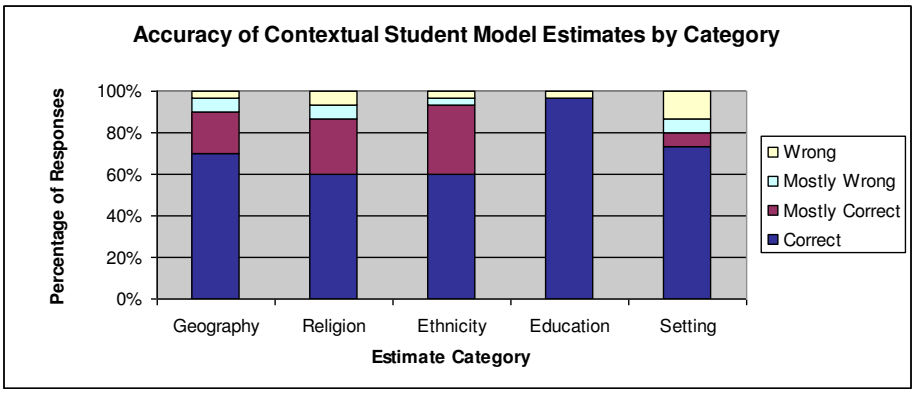

Fig.4. Accuracy of Contextual Student Model Estimates by Category

6. Analysis and Discussion of Results

The first experiment aimed to evaluate the likelihood that the data required for generating a contextual student model will be readily supplied by users. The results showed that the majority of users polled for this experiment were willing and comfortable to supply contextual data about themselves and their parents. Closer examination revealed that all of the users were willing and comfortable to give information about their schools, and languages spoken by themselves and their parents. There were differences in the number of users (ranging from 100% to 77.8%) who were willing and comfortable to supply data for the remainder of dimensions. Users were the least comfortable to give information about their parents compared to themselves but were willing to give levels of influence. Overall, the experiment indicates that users would readily supply information for the majority of dimensions that are used in the CSM to generate the estimates. In the cases where information would not be supplied for particular dimensions such as for parental social units, the information provided for personal dimensions seem to be sufficient for estimating missing data through averages using country-level statistics. It is therefore not unreasonable to conclude that data

collection for the CSM is viable for the country investigated. Further study is required to determine whether data collection is viable for other countries since there are differences across countries with respect to what users may be willing to divulge about themselves along with legal and ethical issues as evidenced by the study in [1].

Given that data can be collected for the CSM from users in general, the second experiment aimed to evaluate the acceptability of the estimates produced by the CSM based on student ratings of the estimates. Students were used in this experiment since the intended use of the CSM is for educational purposes. The results showed there were variations in the accuracy ratings for each category of estimates but overall more than 80% of students rated the estimates generated as correct or mostly correct. The setting category was rated as least accurate. This happened possibly because of the limited metadata on the country locations which did not sufficiently distinguish cities or towns as rural compared to semi-rural or even urban for the students. This highlights one limitation of the CSM in depending on statistical data from a country's central statistical office or department. Errors can be introduced into the estimates if the data is incomplete or not specific enough. Nonetheless, the estimate was still reasonably accurate since it was rated as wrong by 13.3% of the students but only mostly wrong by 6.7% of the students. Estimates in the religion category may have fallen short by not assigning a larger weight to the student's religion since a few estimates recorded a different dominant religious factor for students whose religions differed from their parents. Even so, the estimate was still reasonably accurate since it was rated as wrong by 6.7% of the students but only mostly wrong by 6.7% of the students. The estimates for geography, ethnicity and education were rated as over 90% accurate and this shows that these estimates were on point for the students. Despite the accuracy of the estimates, there were cases of students rating the estimates as inaccurate as shown in Figure 5 even though the reasoning for the estimate was logical and made sense for the student's context. Overall the CSM rules, dimension combinations and weightings were reasonable for estimating the student's membership to various contextual groups as indicated by the favorable accuracy ratings.

7. Conclusion and Future Research

The contributions of this paper are the identification of the main contextual dimensions of a student's cultural background that are important for adaptation at the application layer in CATS together with the dimension combinations that work to generate reasonable estimates of a student's membership to various cultural groups. Rules were developed to estimate a student's degree of membership to these contextual groups. Results from the evaluations of the CSM revealed that the model was accurate in assigning contextual group membership scores to students. The techniques described in this paper are non-trivial and harness many pieces of metadata in order to create a reasonable computational representation of a student's contextual background. In doing so, this research has revealed that a considerable amount of effort will be required by practitioners seeking to create contextual student models due to the heavy reliance on model values at a student level, resource level and country level. The CSM approach was developed with generalization at the core since it is important for others

to be able to replicate these results in their own country and context in order for CATS research to continue to move ahead. Strategies for building models of student context would be worth very little if the students agree with the model but do not wish to have their cultural context factored into their learning experience.

Future research includes the transition of the CSM prototype to an ontological representation to facilitate reuse and better context matching through ontological alignment and merging with resource contexts. Additional dimensions of personal student contexts will be included in the CSM together with more integrated learner context in order to fine-tune the estimates generated. More importantly, work is planned for the investigation of techniques that allow students to accept, adjust or even turn off contextualisations in culturally-aware tutoring systems.

References

1. Blanchard, E.G.: Is it adequate to model the socio-cultural dimension of e-learners by informing a fixed set of personal criteria? In Proc. 12th IEEE International Conference on Advanced Learning Technologies. 388-392. USA: IEEE Computer Society. (2012)
2. Blanchard, E.G.: Adaptation-oriented culturally-aware tutoring systems: When adaptive instructional technologies meet intercultural education. In Song, H., Kidd, T. (Eds.): Handbook of Research on Human Performance and Instructional Technology. Hershey, PA: IGI Global. 413-430. (2009)
3. Blanchard, E.G., Mizoguchi, R., Lajoie, S. P.: Structuring the cultural domain with an upper ontology of culture. In Blanchard, E., Allard, D. (Eds.): The Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. 179-212. Hershey, PA: IGI Global. (2011)
4. Blanchard, E. G., Roy, M., Lajoie, S. P., Frasson, C.: An evaluation of socio-cultural data for predicting attitudinal tendencies**.** In Proc. 14th International Conference on Artificial Intelligence in Education, Brighton, UK. 399-406. Amsterdam: IOS Press. (2009)
5. Gasparini, I., Pimenta, M.S., de Oliveira, J.P.: How to apply context-awareness in an adaptive e-learning environment to improve personalization capabilities? In Proc. 30th International Conference of the Chilean Computer Society, SCCC 2011, Chile. 161-170. (2011)
6. Horton, W.: Graphics: The not quite universal language. In Aykin, N. (Ed.): Usability and Internationalisation of Information Technology. 157-187. Mahwah, NJ: Lawrence Erlbaum Associates. (2005)
7. Melis, E., Goguadze, G., Libbrecht, P., Ullrich, C.: Culturally-aware mathematics education technology. In Blanchard, E., Allard, D. (Eds.): The Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. 543-557. Hershey, PA: IGI Global. (2011)
8. Rehm, M.: Developing enculturated agents: Pitfalls and strategies. In Blanchard, E., Allard, D. (Eds.): The Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. 362-386. Hershey, PA: IGI Global. (2011)
9. Reinecke, K., Reif, G., Bernstein, A.: Cultural user modeling with CUMO: An approach to overcome the personalization bootstrapping problem. In Proc. First International Workshop on Cultural Heritage on the Semantic Web at the 6th ISWC 2007. 83-90. (2007)
10. Cassell, J.:Social practice: Becoming enculturated in human-computer interaction. In Stephanidis ,C. (Ed.): Universal Access in HCI, Part III, HCI 2009, LNCS 5616. 303-313. (2009)