A clustering-based semi-automated technique to build cultural ontologies

Ramesh Srinivasan
Department of Information Studies, University of California, Los Angeles
Email: srinivasan@ucla.edu

Alberto Pepe
Department of Information Studies, University of California, Los Angeles
Email: apepe@ucla.edu

Marko A. Rodriguez
Center for Non-linear Studies, Los Alamos National Laboratory
Email: marko@lanl.gov

This article presents and validates a clustering-based method for creating cultural ontologies for community-oriented information systems. The introduced semi-automated approach merges distributed annotation techniques, or subjective assessments of similarities between cultural categories, with established clustering methods to produce "cognate" ontologies. This approach is validated against a locally-authentic ethnographic method, involving direct work with communities for the design of "fluid" ontologies. The evaluation is conducted with of a set of Native American communities located in San Diego County (CA, USA). The principal aim of this research is to discover whether distributing the annotation process among isolated respondents would enable ontology hierarchies to be created that are similar to those that are crafted according to collaborative ethnographic processes, found to be effective in generating continuous usage across several studies. Our findings suggest that the proposed semi-automated solution best optimizes between issues of interoperability and scaleability, de-emphasized in the fluid ontology approach, and sustainable usage.

Section 1. Introduction

The ubiquity of digital portals for cultures and communities has begun to add a "cultural" question to the usability debate. This question focuses on the idea that usability cannot simply be reduced to the interaction a "neutral" individual has with a system, but that the system's interface, ontologies, and deployment all must consider the larger social and cultural context within which the user is embedded (Kling, 2005; Suchman, 1995; Srinivasan, 2006a; Gurstein, 2000a/b; Schuler, 1994; Loader and Keeble, 2002; Pinkett, 2003; Quaan-Haase et al, 2002). E-governance systems, digital library projects (Bishop et al, 2000; Borgman, 1989; Marchionini, 2000), and localized information kiosks in rural and international environments, are but a few instantiations of larger attempts to create an information society (Webster, Lievrouw, WSIS), where users and technologies can be networked across geographic and cultural bounds. When considering the cross-cultural impact of information technologies, a key question emerges around the topic of localization: How can information portals be created that serve diverse communities and cultures? Remarkable projects addressing this issue emerge in social informatics (Kling et al, 2005), fluid ontologies (Srinivasan and Huang, 2005), ethnomethodology (Srinivasan, 2007; Crabtree, 1998; Crabtree et al., 2000; Garfinkel, 1967), and participatory design (Gregory, 2003; Bodker and Pedersen, 1991; Bodker and Gronbak, 1991).

The consensus amongst researchers has been that the evolution of information systems involves a focus on the role of semantics (Berners-Lee and Miller, n.d), the belief that meaning can be contextualized into information and networks, based on the categories by which information objects are represented. This revolution, described as Web 3.0, is focused on extending the mashups of Web 2.0 (Vanderwal, 2007; O'Reilly, 2005; Shirky, 2002; Duval et al, 2002) to mine and draw connections between previously distributed knowledge objects lying on different web servers. These dynamics: of an evolving web meeting more networked communities and cultural groups – raise an important question: What methods can researchers, engineers, and policymakers alike embrace in creating digital systems that categorize, share, and transfer information meaningfully for diverse local communities? Moreover, can solutions for designing information systems for diverse publics be found that resonate across diverse cultural contexts? And can these solutions still work in relatively automated and decentralized manners, so that scaleable yet locally cognizant systems can be deployed? (Srinivasan, 2006a/b; Menou, 1989; Schware, 2005).

This article presents and validates a semi-automated, scalable and economical technique of combining distributed annotation of ontology topics with clustering methods to model cultural and community-oriented information systems. This technique is compared relative to an ethnographic "fluid ontology" generation method that has been found to significantly engage users across diverse cultural communities (Srinivasan and Huang, 2005; Crabtree, 1998; Srinivasan, 2004, 2006a). In the next section (2), we provide a literature review of the impact of Information and Communication Technologies (ICTs) on diverse social, cultural and ethnic environments. Section 3 presents our study and methodology, introducing several existing methods for representing and sharing cultural knowledge as well as our experiment design and dataset employed. Section 4 follows this by presenting the results and analyses of our comparative study. Finally, in the conclusion, we speculate on the potential of future studies involving other modeling techniques for cross-cultural information systems.

Section 2. Culture and Systems

Scholars have repeatedly discovered that ICTs, engineered and designed in research and development centers, often do not perform in the diverse ethnic and cultural environments in which they attempt to embed themselves (Warschauer, 2002; Srinivasan, 2006a/b, 2007; Ginsburg, 2005; Suchman, 1995). This has revealed itself even when explicit attention has been paid to predictions of user profiles in market research (Jain and Krishnapuram, 2001; Vertommen et al., 2004; Belkin and Croft, 1992). Traditional information retrieval systems have been focused on analyzing documents, rather than on the diverse needs or seeking behaviors of users (Vakkari, 1999; Cole, 1993; Vakkari et al, 1999; Wilson, 1998; Dervin, 1998/2001/2003; Dervin and Nilan, 1986). Information scientists have thus repeatedly called for a greater consideration of users (Dervin and Nilan, 1986), and the stages by which they interact with systems and access information (Borgman, 1989; Kuhlthau, 1991/2004; Bates, 1989; Fisher et al, 2004). It is from these lessons that the movement of participatory design has been aided (Crabtree, 1998; Crabtree et al., 2000; Jacob, 1999; Gregory, 2003; Bodker and Gronbak, 1991), based around the philosophy that the direct involvement of end-users in system design and development can have promising aims in terms of its sustainable and productive usage. This transcends the historical approach toward usability that only involves the user in evaluation rather than design. Participatory design, in contrast, argues that users could become integral to the design of the system's interfaces, information architectures, and more. Moreover, work has emerged that attempts to move past the homogenized category of "user" to consider cultural and social difference in ICT design (Bishop et al, 2000; Srinivasan, 2006a/2007; Suchman, 1995).

As an example, Mark Warschauer has written extensively about studies of information kiosks set up in rural areas within India (2002). One example he writes about is the Hole in The Wall, a study of whether a university-designed information kiosk could be sustainably used by villagers to assist locally constructive aims around education, development, and so on. He has argued that the data has conclusively revealed that such efforts can only engage communities when they are designed around the localized

knowledge within. This means that the technology-design and introduction may benefit from a cognizance of the sociocultural context of the village. Similar results have been found in systems designed for universal access, such as digital libraries. Caidi and Komlodi (2003a/b), for example, have concluded in their study of the International Children's Digital Library that users from different nations browse, share, and utilize the digital library knowledge in different manners. While they advocate for sensible improvements in interfaces for diverse cultural users, one could envision further adaptations around how the digital library content is categorized and retrieved based on culturally-specific categories and topics.

A great deal of research focused around these topics has revealed that culturally specific priorities and attitudes must be built into the ICT-community discussion, when working with diverse cultures and communities. Existing literature focuses on the different approaches toward information seeking of digital resources (Fisher et al., 2004; Flythe, 2001; Su and Connaway, 1995), belief systems that indicate different priorities and systems of knowledge/cognition (Srinivasan, 2004, 2006a/b), goals and activities with respect to their uses of the web (Hiller and Franz, 2004; Srinivasan, 2004/2005; Srinivasan and Huang, 2005; Matei and Ball-Rockeach, 2005), and linguistic-cognitive patterns (Lakoff, 1987; Whorf, 1956). Additionally, cultural information systems must work to acknowledge "information-grounds" (Fisher et al, 2004; Pettigrew, 1998), understood as particular social and cultural contexts wherein communities share information. This relates to an understanding of how informational behavior (Wilson, 1998; Dervin and Nilan, 1986) may vary based on the socially-patterned processes as well as belief systems held by a particular cultural group. These can be detected via the methods by which the group makes sense of the new possibilities that it is presented with, as well as the methods by which it is able to reconcile its own boundaries with the content presented by the system (Star, 1989). Diverse cultural groups therefore maintain particular information 'flow' patterns (Geertz, 1978; Fisher et. al., 2004), and tend to hold distinct ontological priorities and representations of information (Srinivasan and Huang, 2005). The interactions between the cultural beliefs, social structures, and physical environments are unique to the performance of any information system within the sociocultural environment (Nardi and O'Day, 1999). Therefore, while diverse cultures and communities can indeed benefit from information systems (Srinivasan, 2004; Srinivasan and Huang, 2004; Sy, 2001; Hillier and Franz, 2004; Matei and Ball-Rokeach, 2001; Ess and Sudweeks, 2001), the development of these systems must still acknowledge community-specific realities. Yet, as mentioned earlier, unless such "local" systems can communicate and harmonize with global issues of scale and interoperability, they run the risk of being rendered obsolete and overly expensive.

Notable information science research has focused on ontological modeling, the formal representation of a body of knowledge. Research of this kind is situated at the convergence of a large body of literature that addresses issues related to document retrieval, indexing and classification (Svenonius, 2000) via the construction of ad hoc thesauri (Aitchison et al., 2000; Tudhope, 2001), taxonomies, controlled vocabularies (NISO, 2005), and topic maps (Garshol, 2004). A number of different methods to construct, adapt and validate ontologies have appeared in the literature. It is beyond the scope of this article to provide a detailed review of research emerged in this broad field. Yet, it is worth pointing the reader to a number of studies in this domain that are particularly relevant to the present article. In particular, it is interesting to note that the methods employed for the construction of knowledge constructs (from ontologies to taxonomies to controlled vocabularies) include automatic, manual techniques and a blend of them (the latter being the focus of this article). Whereas traditional language and domain thesauri were built manually, in the past two decades, the bulk of research has shifted towards automated and semi-automated techniques.

Recent research employing fully automated techniques includes keyphrase extraction using semantically-rich taxonomy descriptions over domain-specific thesauri (Medelyan and Witten, 2008), taxonomy population using association between words/phrases extracted from bibliographic titles and subject

descriptors in metadata records (Wang, 2005) and ontology-based automatic extraction of knowledge from the web to generate personalized narrative biographies of artists (Alani et al., 2003). Work employing semi-automated methods often involves a blend of automated construction techniques and the engagement of end users for control and evaluation; examples are the use of relevance feedback to improve the quality of web-extracted taxonomies (Kumar, 2001), task-based evaluations to improve development and design of retrieval associative thesauri (Nielsen, 2005), and empirical evaluations of how searchers use implicit feedback and automated assistance while searching (Jansen and McNeese, 2005). Many scholarly taxonomies and domain ontologies, have been developed using a top-down approach, in which catalogers and domain experts carve out a domain's fundamental concepts and relationships (e.g. Gil, 2004). However, a number of recent research efforts are increasingly relying on user engagement not simply to provide evaluation and quality control; employing large scale collaborative platforms, many recent endeavors have utilized bottom-up approaches, in which users contribute and merge free-form labels and categories (Shirky, 2000; Golder and Huberman, 2006; Van Der Wal, 2007). In turn, ontological databases produced in the context of large-scale user-driven collaborative endeavors are being employed for a number of computational applications on the Semantic Web (Legg, 2007).

Section 3. Study and Methodology

Drawing from the rich theoretical framework presented in the previous section, this article focuses specifically on the underpinnings of existing methods by which cultural ontologies have been modeled to accommodate and engage different communities. Ontologies are often created for multiple purposes. First, they allow information objects to be classified, providing users with basic metadata with which they can be associated. Second, their structure describes not just the information in the database but the semantic relationship presumed between categories. The ontology, in other words, presents the world-view of the information system, allowing information to be presented within a set of classifications and categories presented alongside one another. Finally and similarly, ontologies present expertise around a particular domain or subject. In creating rich databases and systems that represent a topical or in this case, community domain, the ontology reveals rich diversity and specificity.

This article extends literature that has argued that an effective community ontology serves multiple important purposes. First, it presents a body of information from a diverse community with the integrity and respect within that group’s worldview. Second, it engages community members to actively design their own systems, enabling participation and more effective long-term usage of the system. Finally, it presents an appropriate basis by which users can retrieve and browse objects stored in the database. The power of classification, well understood in the information sciences, looms large when creating a system that is supposed to both serve and represent a particular community.

The hypothesis investigated in this article is that the use of a semi-automated approach integrating distributed annotation of ontology topics and hierarchical clustering could enable ontologies to be created that are similar in structure to ethnographically-derived "fluid ontologies". Via semi-automated methods, communities would be actively involved in the system design but instead of engaging in the intensive process of designing their own fluid ontologies, they are asked to collectively agree on a set of ontology topics, and then individually be presented with randomized pairs of topics to evaluate in terms of cultural similarity. Instead of basing the ontology design process around the presence of an external researcher/ethnographer, this process could allow communities to semi-automatedly design their own ontologies. In that regard, the hypothesized method would function as both a scaleable and decentralized solution. Larger numbers of community members could participate and on their own time. We were curious as to whether the ethnographically derived "fluid ontology", given its hierarchical shape (as discussed in the coming sections), could be approximated by a method that places communities themselves in complete control of the system design. Therefore, an experiment was designed to study the

potential of the presented semi-automated approach. The optimal solution uncovered would balance the needs of local authenticity (embodied by the ethnographic method) with scalability (embodied by the distributed annotation method) and interoperability (embodied by the cluster analytical techniques). Both the ethnographic and the semi-automated ontology generation procedures shall be explained in the following sub-sections.

3.1 Generating cultural ontologies using ethnographic techniques

"Fluid ontologies" are representations of knowledge that are based on community-articulated categories and interrelations (Srinivasan and Huang, 2005). Fluid ontologies may take several forms by engaging in the following steps: (a) involvement of content creators: the ontology can become richer and more contextual if the content creator is directly involved in its definition; (b) sharing metaviews: the browsing and learning experience can be significantly enriched if the way the participant makes sense, browses, rearranges and visually displays the knowledge are made explicit and are made available to others; (c) adaptiveness: ontologies that adapt and are continually redesigned in time can be more useful than static ontologies; (d) bots and personalization: the use of artificial bots and proactive agents that track and analyze interaction histories can effectively help adapt and fine-tune the dynamic evolution of ontologies. (Srinivasan and Huang, 2005).

Notable for this specific study are the first three criteria from above. They present the possibility that creating systems around locally-created, ever-adaptive, and collectively-authored/shared ontologies presents a possibility to enable communities to meaningfully own and sustainably utilize ICTs. These methods have demonstrated significant resonance when placed in contrast to systems that solely engage collaborative filtering agents or standard corpus-generated keywords (Srinivasan, 2005/2007). These "fluid" techniques, drawing from principles of participatory design (Gregory, 2003; Crabtree, 1998; Harrison and Zappen, 2004; Srinivasan and Shilton, 2006), engage communities to not just author content for localized ICTs but also serve as the information architects of their systems. These 'cultural ontology' findings continue to be meaningfully applied in ongoing efforts related to cultural tagging in digital museums (e.g. Boast, Bravo, and Srinivasan, 2007).

The fluid ontology that we adopt in our comparative analysis comes from previous work within the "Tribal Peace" research initiative, a multi-year project focused on the study of methods by which an information system could be developed to bridge the significant fragmentation faced by 19 Native American reservations in San Diego County (Srinivasan, 2005). The Tribal peace project involved the presence of the ethnographic researcher working with the tribal communities for close to two years. This time involved a variety of research activities; including participant observation (Spradley, 1990; Hammersley and Atkinson, 1983), working with different tribal members to develop technological literacy, outreach, and gathering a group of individuals relatively representative of the different reservations and their social strata to serve as the project leadership committee. This group (over several meetings) would view all the digital content submitted by different community members and by reflecting on, and discussing the themes, concepts, and topics around this content, worked to design an iterative and adaptive fluid ontology around which the system content could be organized. The fluid ontology that resulted from this work is presented in Figure 1. The end nodes of the tree structure represent specific concepts, topics, themes, and categories that tribal members deemed relevant. These were listed during the brainstorming process by the focus group as they viewed and reflected upon different submitted media pieces.

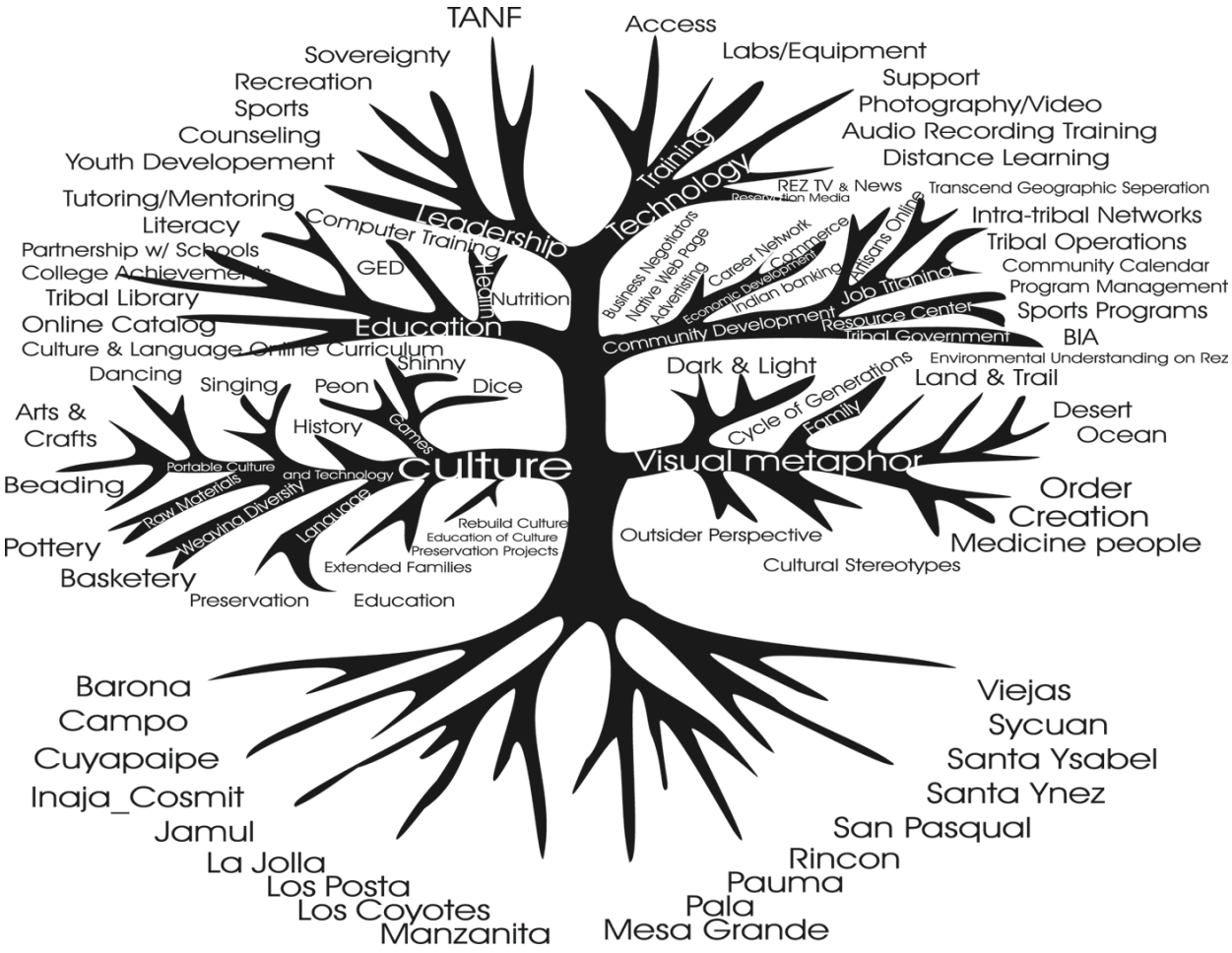

FIGURE 1: Community Fluid Ontology (Srinivasan, 2007)

The relative impact of this system has been described in several other publications (Srinivasan, 2005, 2006a, 2007), yet the results have largely validated previous findings that such locally-designed ontologies correlate to more sustained and diverse forms of usage, as well as greater adoption by community institutions. It has presented promise that ICTs can be designed to embed themselves effectively in diverse cultural environments, yet also has raised the relative alarm that an overly localized set of systems run the risk of being unable to communicate with important issues of interoperability and scaleability. In particular, we identify two limitations associated with the fluid ontology method. First, creating fluid ontologies is intensely time-consuming and requires the presence of an ethnographer, thus the method can be considered largely unscaleable. Second, although being efficient and locally authentic, fluid ontology building engages only a small subset of the community and is thus inappropriate for the study of large heterogeneous groups and domains.

3.2 Generating cultural ontologies using semi-automated techniques

We have noted, in this and other work, that fluid ontologies demonstrate hierarchical form (Srinivasan, 2005/2007). That is, we see a tree-type structure with top-level nodes that are conceived of as meta-topics, and lower-level nodes as instantiations of these concepts, as in Figure 1. Such hierarchical form suggests that other means of semi-automatedly eliciting cultural ontologies may present reasonable alternatives to ethnographic methods. We introduce here a method that blends manual annotation of ontology topics, via distributed subjective assessment of similarity, with algorithm-driven clustering techniques.

Similar to the 'fluid ontology' method presented above, the semi-automated approach introduced here involves direct involvement with the community studied. However, instead of relying on the ethnographer and selected community members (e.g. a focus group) to derive together a definitive ontological model, this method relies on aggregated subjective judgments of similarity provided by a portion of the population. This method consists of the following steps: (a) generating ontology topics, by asking respondents to identify ontological concepts that are closely related and authentic to their position and status in the community, (b) scoring the similarity between pairs of topics, by randomly distributing ontological concepts among respondents and asking the following question "How similar are these concepts?", and (c) aggregating similarity scores into an ontological model, by using a similarity based aggregator. By decentralizing subjective decision (ontology topics) and judgment (similarity scores) over a set of community members, we build on the positive attributes of participatory design and successively exploit algorithmic, automated techniques to assist the creation of the actual ontology.

Each one of the steps described above is crucial to the generation of the cultural ontology. The first step, generating the initial set of ontology topics, might be performed both by individually asking respondents to propose topics that are related to their community experience or by organizing collective groups in which respondents collectively construct the set of ontology topics. In our study, in order to ease the comparative analysis among the techniques, we borrow the ontology topics generated via the ethnographic method presented above; using exactly the same ontological concepts for comparison yields a more distilled picture. This constitutes a set of 60 ontology topics (the top-level nodes of the tree structure of Figure 1).

The second step consists of presenting respondents with a random set of ontology topic pairs. Respondents are then asked to score the similarity between these topics in the cultural context of their communities. Similarity is only one way to assess the relationship among ontology topics, yet an important one. Goldstone and Yun Son (2005, p. 13) note: “human assessments of similarity are fundamental to cognition because similarities in the world are revealing.” The concept of similarity is central to the domain of cognitive psychology and its fuzziness is extensively discussed by the authors (Goldstone and Yun Son, 2005). Another obvious dimension that could have been used to assess the relationship among ontology topics is the inverse of similarity, dissimilarity. The use of dissimilarity, or other types of measures, would have possibly yielded very different results, but, in some cases, it would have been arduous to convey the true notion of the desired measure to the community members involved with the scoring procedure. For the purpose of our study, we are interested in finding a definition of similarity that the respondents would be naturally familiar with. In our case, the specific question that we asked every respondent was: “How similar are these two themes/concepts in the life of individuals across the tribal reservations?”. For example, respondents indicated a strong relationship between the terms "Crafts" and "Basketry" and a weak relationship between the terms "Dancing" and "Computer Training". We speculate that the use of a familiar notion of similarity, situated in the natural context of the community under study, provides more accurate, locally-authentic results.

The third step of the ontology generation mechanism involves the aggregation of similarity judgments into a condensed ontological format. A large number of algorithmic models of similarity have emerged in many diverse academic fields, notably in psychology and in the cognitive and computer sciences. Approaches range in nature and application: from geometric and spatial models, such as Multi-Dimensional Scaling (Shepard, 1962), to feature models based on weighting of matching features (Tversky,1977). In this article, we use cluster analytical methods to aggregate the collected similarity scores. The choice of the clustering method is particularly important, as different clustering methods yield different results (Murtagh, 1984). Cluster analytical techniques categorize different objects into groups maximizing the degree of association between them. Although several approaches to clustering exist, the categorization of objects into groups broadly falls into two paradigms: agglomerative and partitioning

methods based on whether they progressively merge or separate objects from the preceding configuration (Kaufman, 1990). The aim of both methods, however, is the same: to uncover cluster structures from related pairs of objects. In our case, the distance among term objects is represented by the similarity scores between ontology topics manually given (annotated) by the respondents.

For the purpose of our study, we present two popular clustering methods to aggregate the similarity scores among ontology topics: Ward's hierarchical clustering (agglomerative) and K-means clustering (partitioning). Our choice of methods follows exploratory work in which we assessed the relative performance of a number of clustering techniques. Besides K-means and Ward's, we tested five other hierarchical clustering methods, namely single link, complete link, group average, centroid, and median. All these methods performed considerably worse than the Ward's and K-means. In particular, all these hierarchical clustering methods, with the exception of complete link and group average, resulted in dendrograms with high skewness and were thus excluded immediately. Complete link and group average yielded more balanced trees with similar cluster sizes to Ward's and the fluid ontology. Yet, later analyses (chi-square test and qualitative inspection) revealed that the content of the clusters generated by complete link and group average were significantly more discrepant than Ward's from those emerged in the locally-authentic fluid method. Thus, we limit our discussion to a comparative analysis of the best performing hierarchical clustering method (Ward's) and K-means.

3.3 - Research questions

Our study focuses on the assessment of the relative performance of the semi-automated technique (section 3.2) to approximate the ethnographic fluid ontology (section 3.1). We summarize here our motivations for the present study. These are focused around the main research question of: How can information systems represent and categorize knowledge for community access without being subject to the expense of the ethnographer's presence, and the lack of scaleability and interoperability of the fluid ontology? Sub-questions include:

- How can an ontology be created that is more inclusive of larger numbers of community members? The reported focus group numbers range from 15-20 members, and are compromised by which reservations the researchers had greater access to.
- Similarly, how can an ontology be created without asking community members to overcome the infrastructural and distance-based barriers they currently face? Is it possible to generate an inclusive ontology by collecting decentralized input from community members in their at-home or at-work environments?
- How can an ontology be derived that is adaptive of the priorities held by different subsections of the community?
- How can the process of engaging communities to develop their own ontologies serve as a scaleable methodology that may be applied in ICT projects with other communities?
- Finally, how can this process of local ontology development work with the need for interoperable systems? How can communities design local ontology systems that still communicate with other systems?

3.4 - Methodology

To begin to answer the above questions, we performed a comparative analysis of the ontologies resulting from the ethnographic fluid ontology and the semi-automated methods. Our study was conducted between January and March of 2006 with 20 randomly selected tribal members who represented 15 of the 19 reservations with a relatively equal distribution of men, women, youth, and elders. These were members who had not previously used the Tribal Peace system nor were part of the previous focus groups, so they would not be biased by the pre-existing ontology design, yet were as representative a population sample as the fluid ontology focus groups. As topics were already created for the Tribal Peace ontology, we provided these 20 members with random pairs of ontology nodes, and asked them to score the similarity between these topics in the cultural context of their communities. The specific question we asked was:

"How similar are these two concepts in your community life?". Respondents were allowed to give a similarity score between 1 (not similar) and 5 (very similar). For example, most respondents indicated a strong similarity between terms "Crafts" and "Basketry" (score: 5) and a weak similarity between terms "Dancing" and "Computer Training" (score: 1). A sample of the collected similarity scores is shown in Table 1.

Ontology topic 1	Ontology topic 2	Similarity score
Dancing	Computer Training	1
Crafts	Basketry	5
Darkness&Light	Sovereignty	2
Dancing	Pottery	4
History	Literacy	4
Tribal Government	Support	2
Medicine people	Outsider perspective	3
Order&Creation	Job training	2

TABLE 1. SAMPLE SIMILARITY SCORES

A total of 60 ontology topics (extracted from the fluid ontology depicted in Figure 1) resulted in 1770 (i.e. (60 * 60) / 2) topic combinations that were provided to the community respondents. By reducing the number of topic arrangements to a combination, rather than a permutation, we made the assumption that ontology topics were symmetric in terms of similarity. In other words, we assumed that respondents would score ontology topics identically, regardless of their ordering, so that, for example "Dancing - Computer Training" and "Computer Training - Dancing" (from Table 1) would both yield a similarity assessment of 1. This assumption, typical of many standard geometric models of similarity, is considered a limitation of experimental studies in cognitive psychology (Tversky, 1977). In the study presented here, the symmetry problem was mitigated by prompting respondents with randomly ordered ontology pairs and subsequently averaging results. Also, the same ontology pairs were assigned them to more than one respondent for cross-validation: each respondent "annotated" an average of approximately 500 random pairs.

Data gathered was used to construct a 60 by 60 symmetric similarity matrix with a diagonal of 5 (i.e. every topic is completely similar to itself) that listed out the averaged similarity score provided by community members between every pair of ontology topics. In order to perform a quantitative comparison between this ontology and the fluid ontology in matrix terms, the latter also was converted to a similarity matrix. This was done by reinterpreting the distance among topics in the tree-like configuration of Figure 1 in terms of similarity. Ontology topics lying close to each other on the same tree branch were considered more "similar" than topics lying on different branches. Figure 1 depicts the fluid ontology as a tree, composed of six main branches (e.g., “Culture” branch in the lower left portion). In turn, these branches can have annotated sub-branches (e.g., “Games” sub-branch in the “Culture” branch). In particular, the following scheme was employed:

- Topic pairs lying on the same branch or sub-branch of the tree (e.g., “Dancing” and “Singing” in the “Culture” branch) were given a similarity score of 5.
- Topics lying on sub-branches (e.g. “Dice” in the “Games” sub-branch) were given a similarity score of 4 with topics present in the main branch (e.g., “Dancing” in the main “Culture” branch).
- Topic pairs lying on different, but adjacent branches (both vertically and horizontally) were given a similarity score of 3. For example, “Dancing” and “Literacy” are on vertically adjacent branches (“culture” and “education”, respectively) and were given a score of 3.
- Topics pairs lying on different non-adjacent branches were given a score of 2 if both branches are on the same side of the tree, or 1 otherwise. For example, topics in “Culture” and “Leadership” were given a score of 2, whereas topics in “Culture” and “Technology” were scored 1.

By associating values in this fashion, a similarity matrix for the ethnographic ontology was generated. We refer to this similarity matrix (and the corresponding ontology) as "cognate".

Two distinct similarity matrices were thus available to us, one resulting from the distributed annotation technique and the other from the reinterpretation of the fluid ontology as a similarity matrix. As explained, these both describe the extent of similarity between any two ontology topics. The analytical comparison among these matrices and the corresponding ontologies shall be presented and discussed in the next section.

Section 4. Analysis and results

As mentioned, the goal of this article is to introduce a semi-automated approach for the modeling of cultural ontologies and evaluate it against locally-authentic ethnographic methods. In the previous section we have discussed how we generated a hierarchical ontological construct (shown in Figure 1) and a corresponding similarity matrix using ethnographic "fluid" techniques. We also discussed how we gathered similarity assessment annotations from respondents and used them to construct a "cognate" similarity matrix. In this section we perform the following analyses: (i) in section 4.1, we conduct a crude discrepancy test among the matrices, (ii) in section 4.2, we generate and present ontologies from the cognate similarity matrix using Ward's and K-means clustering techniques, (iii) in section 4.3, we perform a cluster analysis of the terms in the fluid and cognate ontologies and determine if there exists a statistical relationship between the clusters in each one of them.

4.1 Matrix comparisons

A preliminary technique to compare two similarity matrices is to subtract them and analyze the distance (dissimilarity) matrix obtained. We generated a 60 by 60 matrix containing the distances between the ontologies obtained via the fluid ontology relative to the cognate ontology. The matrix consisted of 1770 distance values, ranging between 0 (when terms were scored identically via both methods) and 4 (when term scores were entirely different). In the dissimilarity matrix, the mean was found to be -0.50 with a standard deviation of 1.83 from the mean. A histogram depicting the results is shown in Figure 2. It is important to note that values represent discrepancies among matrices, thus positive and negative distances can be collapsed. A total of 355 pairs (out of 1770) were scored equally via both methodologies and over 600 within a distance of 1 (the sum of +1 and -1 scores). This result indicates similarity between the two matrices, as demonstrated in Figure 2.

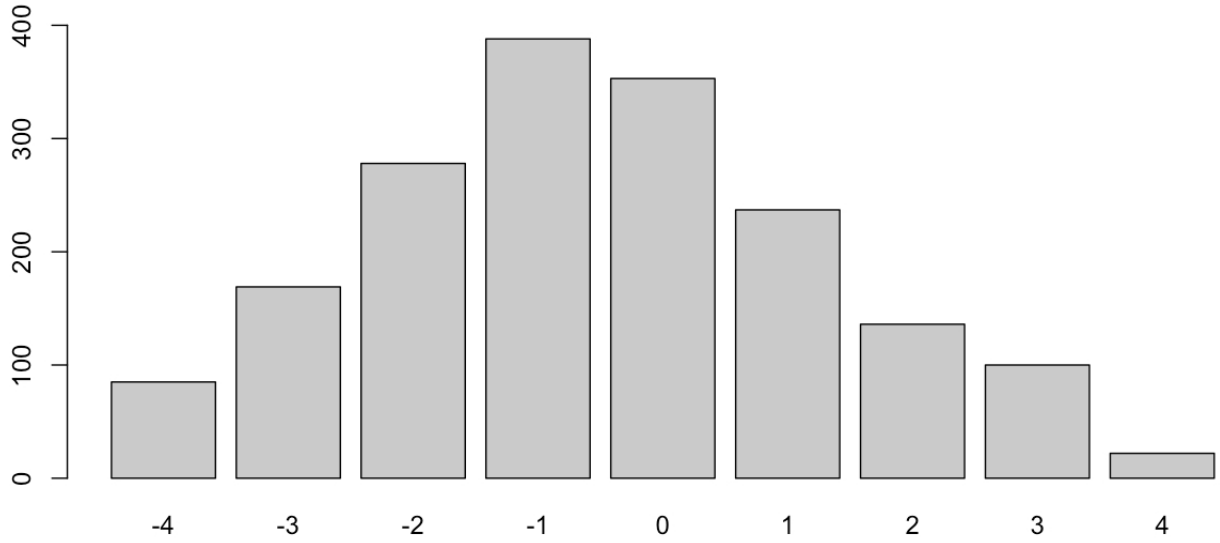

FIGURE 2. DISCREPANCY AMONG FLUID AND COGNATE SIMILARITY MATRICES

4.2 Ontology generation via clustering techniques

As discussed in the previous section, we adopted cluster analysis to convert the cognate similarity matrix into an ontology construct. The choice of a clustering method affects very much the resulting ontological configuration. In this study, we used an agglomerative hierarchical clustering method, Ward's, whose performance has been evaluated in a number of studies (Lin & Hsueh, 2006) and a partitioning method, K-means. The ontological configuration generated via the Ward method resulted in 6 clusters, depicted in Figure 3.

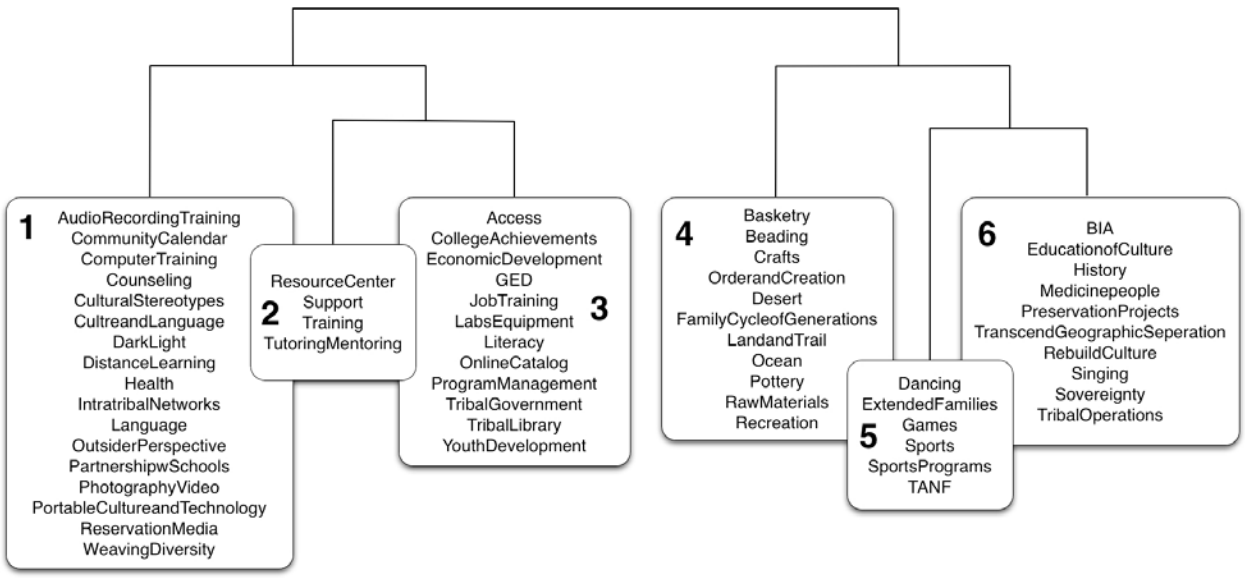

FIGURE 3. COGNATE ONTOLOGY: CLUSTERS GENERATED VIA THE WARD HIERARCHICAL CLUSTERING METHOD

There are various K-Means algorithms that are used to derive object clusters within an *n*-dimensional space. The K-Means algorithm that was used in this study was developed by Hartigan and Wong (Hartigan & Wong, 1979). In brief summary, the algorithm randomly assigns *k* cluster centroids to an *n*-dimensional space that contains a collection of objects whose object properties can also be represented in the same *n*-dimensional space. Next, the algorithm attempts to minimize the distance that a cluster centroid has from the objects in the space by relocating the centroid as need be. This process iterates until a desired minimization is achieved. For the purposes of this study, a *k*=6 was used to be able to compare the results of the K-Means clustering with both the Ward and diagram-based ontologies. The cognate ontology that was generated using the K-means algorithm is displayed in Figure 4.

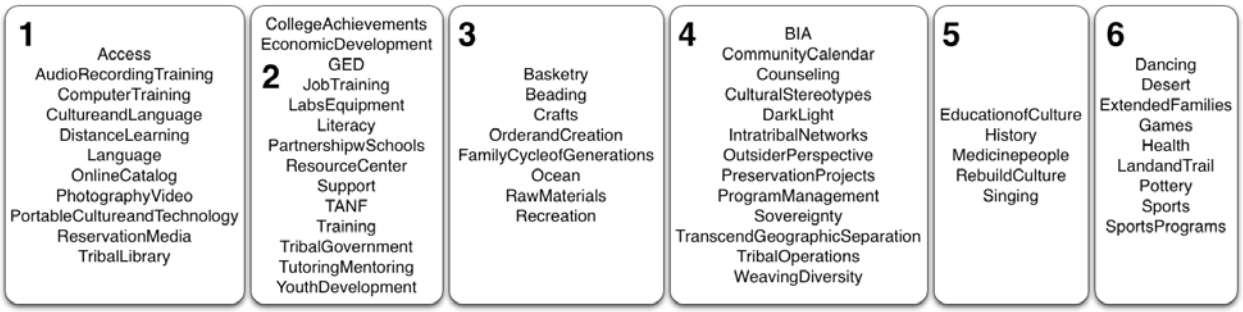

FIGURE 4. COGNATE ONTOLOGY: CLUSTERS GENERATED VIA K-MEANS CLUSTERING

4.3 Ontology comparisons: Quantitative and qualitative analyses

The previous section outlines the creation of two ontologies from the cognate similarity matrix: one obtained via Ward clustering and the other via K-Means partitioning. In this section we discuss how these ontologies differ quantitatively and qualitatively from the locally authentic ontology generated via fluid ethnographic methods. To aid visual analysis and comparison among the ontologies, we present in Figure 5, the cluster configuration of the fluid ontology. The ontology topics have been arranged to exactly correspond to the end nodes in the branches of the hierarchical representation of Figure 1. Also, rather than assigning numbers to the clusters, the original branch names have been indicated.

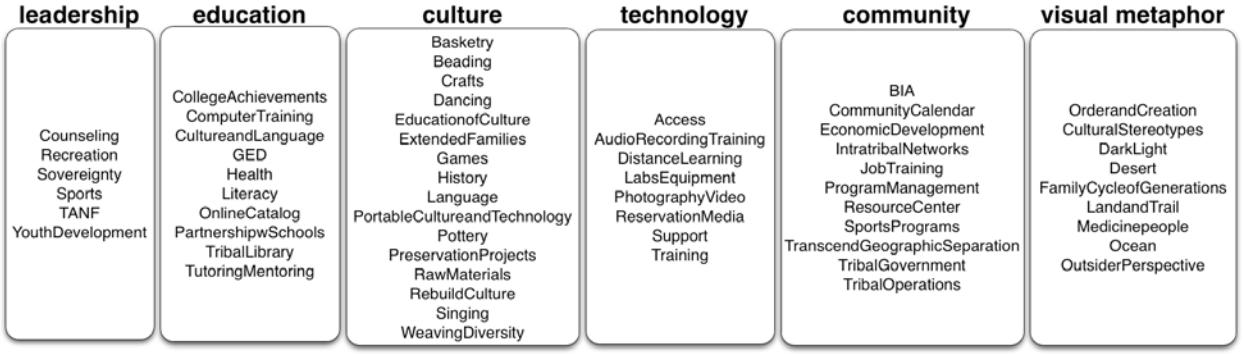

FIGURE 5. FLUID ONTOLOGY: CLUSTERS GENERATED VIA THE ETHNOGRAPHIC METHOD

Now that the cognate ontologies (Figures 3 and 4) have been presented and displayed alongside the fluid ontology (Figure 5), we can perform a quantitative as well as a qualitative comparative analyses of our results. Overall, both Ward and K-means were successful in generating a balanced ontological configuration with cluster arrangements similar to that of the fluid ontology; cluster sizes for Ward (17 4

12 11 6 10) and K-means (14 8 13 11 5 9) resemble very much those of the fluid ontology (16 8 9 11 6 10). The fluid ontology, however, being generated via ethnographic participatory methods, has labels associated with clusters (tree branches): "leadership", "education", "culture", "training and technology", "community development" and "visual metaphor". How does the distributed annotation of ontology topics coupled with the different clustering methods (Ward and K-means) succeed in delineating these categories and their content?

In order to answer this question, we perform a cluster analysis of the terms in each of the respective ontologies and then determine if there exists a statistical relationship between the clusters in each ontology. Thus, for both the fluid ontology and the hybrid approach, each of the 60 term objects are placed within one of 6 different clusters. Two contingency tables were created from the ontologies presented above: the first compares the fluid ontology with the cognate ontology generated via the Ward method; the second compares the fluid ontology with the other cognate ontology, obtained via the K-Means method. For example, the first contingency table identifies how many times a term for cluster *x* in the fluid ontology was found in cluster *y* of the cognate ontology (Ward). A chi-square test determines if the values of the contingency table (i.e. observed values) differ significantly from a contingency table that is created given a non-related sample (i.e. expected values). When a chi-square was applied to the observed contingency tables, the p-value that resulted from the calculations were 0.0044 for the first chi-square test (Ward cognate and fluid ontology) and 0.0032 for the second test (K-Means cognate and fluid ontology). Thus, both cognate ontologies were found to be statistically related (both < 0.05) at the level of a term clustering to the fluid ontology, with the K-Means cognate ontology being slightly more representative than the Ward cognate ontology. To check our results, we ran a chi-square test on a cognate ontology obtained using a clustering method (complete linkage) whose poor performance was clear to the eye (cluster sizes: 27 6 6 7 7 7). As expected, the chi-square test returned a much higher value of 0.3675 --- well above the 0.05 significance level. A summary of the chi-square results is presented in Table 2.

	Fluid	**Ward's**	**K-Means**
Fluid	-	0.0044	0.0032
Ward's	0.0044	-	0.001
K-Means	0.0032	0.001	-

TABLE 2. THE P-VALUES OF THE CHI-SQUARE TEST FOR FLUID, WARD'S, AND K-MEANS

In the context of information system design, the generation of ontologies is important not solely for their detailed configuration, but also for the overall knowledge representation that they convey. We further our comparative quantitative analysis with a qualitative study, presented below. The aim of the qualitative analysis is to investigate whether K-means and Ward clustering methods result in ontological differences that considerably affect the knowledge organization of the resulting ontologies.

Starting with the cognate ontology generated by the Ward method, shown in Figure 3, we notice the net separation of the cluster into two broad sections. The right section (sub clusters 4, 5 and 6) contains elements drawing from "culture", "community development" and "visual metaphor" groups. The left section (subclusters 1, 2 and 3) contains mostly elements from "education" and "training and technology". The category "leadership" is the only one that does not fall into a specific portion of the dendrogram. "Leadership" concepts are highly diffused across the entire hierarchy. This is probably due to the fact that it is a more abstract, less tangible concept, and thus inadequate to be captured by freelisting mechanisms. The "culture" topic permeates the entire right portion of the diagram. In particular, we notice a high concentration of "culture" and "community development" topics in the cluster 6and "culture" and "visual metaphor" in cluster 4. This reflects well the considerations posited above concerning the ability of this method to uncover previously unnoticed ontological constructs. In the left portion of the dendrogram,

"education" topics prevail. Cluster 1 is the least focused, but clusters 2 and 3 bring together highly related topics from both "education" and "training and technology". A conspicuous example is found in cluster 2, composed of "Tutoring/Mentoring", "Support", "Resource Center" and "Training". These topics belong to four distinct categories of the ethnographic ontology, yet they are clearly ontologically related.

The cognate ontology generated via the K-Means method (Figure 4) provides a slightly different picture. Subclusters 2 and 6 cannot be easily located within the classification provided by the fluid ontology. However, certain clusters, namely 1 and 4, can immediately be associated with their respective ontology branches: "training and technology" and "community development", respectively. Clusters 3 and 5 contain a strong presence of both "culture" and "education" components. This suggests that the fluid ontology was probably generated following a predefined cognitive structure, even though the creation process was collective and participatory. At the microlevel, the K-means cognate ontology also highlights certain ontological constructs that the typically "top-level" hierarchical approaches have missed, such as the grouping of "Sports", Sport programs" and "Games" (otherwise in three distinct categories). Taken altogether, the clusters identified by K-Means display solid ontological coherence internally and, also approximate very well the themes posited by the locally authentic fluid ontology method.

What can we learn from our results? We have demonstrated that the generation of this cultural ontology can be decentralized by asking respondents in the community to assess similarity among ontology topics and then using clustering techniques to construct the ontology. The clustering methods evaluated, Ward and K-Means, are successful in identifying the high-level classes present in the ethnographic fluid ontology. However, the two methods display different types of strengths. The Ward method provides a hierarchical structure that might prove extremely useful in approximating the typical tree-like structure found in many cultural ontologies. The K-Means method, on the other hand, captures very well categories established in the ethnographic process and is able to uncover new associations between specific topics. In this study, we have used the fluid ontology as our reference ontology to perform quantitative discrepancy tests. Our qualitative study, however, has revealed that the presented cognate ontology generation method has the potential to delineate knowledge representations that ethnographic methods might fail to recognize. We would like to build on these results to argue that the very first step of the semi-automated method could be also decentralized, for example, by asking community respondents to "freelist" ontology topics (Robbins and Nolan, 1997). Distributed freelisting, coupled with distributed similarity assessment as presented here, might prove a solid basis to generate scaleable, economical and locally-authentic ontologies without being subject to the burden of in-depth ethnography.

5. Conclusions and Future Work

In this article we have presented a semi-automated approach to model cultural ontologies for community-oriented information systems. The approach leads to ontology generation via merging subjective assessments of similarities between cultural categories (distributed annotations) with well-established clustering methods. We have evaluated our semi-automated approach against a locally-authentic ethnographic "fluid" method, generated via direct participatory interaction with communities. The scope of our work is to discover whether manually-performed decentralized similarity assessments coupled with purely algorithmic clustering techniques can produce reliable cultural ontologies without being subject to the expense of the ethnographer's presence, and the lack of scaleability and interoperability of fluid ontology methods. Our quantitative analysis indicates that clustering distributed annotations via both Ward and K-means methods yields to ontologies that approximate very well high-level structure structure of fluid ontologies. Our qualitative analysis confirms this results and also reveals that the similarity assessments clustered with K-means methods produce cognate ontologies that are capable of uncovering novel associations between ontology topics. We thus speculate that it is possible to design locally sustainable and culturally representative ontologies without a multi-year ethnographic process and a greater scale of interaction with different community members.

While this article is focused on the structural aspects of fluid ontologies, and how they can be modeled using clustering techniques, we note that fluid ontologies have been actually used as the basis for user interfaces in systems. The Tribal Peace example that this article focuses on (presented in Section 3.1) engages users to select one or multiple nodes of the ontology and retrieves information objects based upon their choices. In this process, users view the documents they submit relative to the collectively designed ontology and make decisions (that they can change at anytime) as to which nodes in the ontology shall the object correspond through a web form that asks users to select the association. Thus, in multiple manners, the ontology is practically used, in terms of retrieval, uploading/annotation, and browsing.

While this study has shown the potential of modeling a fluid ontology through the use of clustering techniques, it also only touches the surface of several other important areas for future research. First, it assesses the quality of the approximation via the initial community ontology. Future research could further empirically verify whether community members actually perceive a tangible difference in terms of level of satisfaction with the semi-automated ontology relative to the ethnographic fluid version. Moreover, it would be interesting to analyze whether the semi-automated technique works more effectively in situations where the community is non-local, and geographically unbounded, such as in the case of a fragmented diasporic community (Srinivasan & Pyati, 2007). A future study in this vein would comparatively assess the merit of the two approaches in this scenario. Finally, future work could also analyze whether this approach functions effectively not just in ethnic, cultural scenarios but also with domains of knowledge, such as with a group of scientists or medical professionals.

In this article, we have hypothesized that the presented semi-automated method is more scaleable than purely ethnographic techniques, yet we have evaluated it on a cultural ontology of fairly limited dimension (60 ontology topics). In future work, it would be interesting to assess its efficiency and validity on a much larger ontological set. Also, we wish to investigate the potential of accessing many more community members from diverse social categories, and enabling them to participate in the decentralized creation of ontologies. By increasing the population of respondents we envision to produce more reliable results, reducing statistical errors linked to the inherent subjectivity of similarity judgments.

As discussed above, we also plan to engage community members in the very initial stage of the ontology generation, by evaluating freelisting methods for the generation of shared cognitive and cultural representations of community knowledge (Romney et al, 1996). We believe that from this we may learn which types of ontologies are most representative for different pockets within a community, and move past a paradigm that sees communities as largely homogenous entities.

Moreover, we intend to employ a broader set of clustering tools to condense distributed annotation into formal ontologies. In particular, we consider to move beyond hierarchically-focused models and towards distributed network paradigms by adopting novel divisive algorithms for community detection. Community detection methods based on random walks (Pons & Latapy, 2006), resistor networks (Wu & Huberman, 2004), modularity (Newman, 2006b), and eigenvector centrality (Newman, 2006a) have proved particularly efficient (Costa et al, 2007) to describe community formation in highly distributed networks, such as biological networks (Wilkinson & Huberman, 2004) as well as social and political networks (Porter et al, 2007). These detection methods might prove convenient to discern the specific semantics of generated ontologies, rather than simply the topological configuration conveyed by partitioning-based and hierarchical clustering techniques. A method for studying the relationship between nominal, or categorical, object properties and a network of object relationships has recently been applied to identifying the semantics of scholarly collaboration (Rodriguez & Pepe, 2008). Such a framework may prove useful in the future when studying network-based, as opposed to hierarchical-based, ontologies.

This article has represented an initial attempt to optimize between important notions of sustainability and scalability within digital community and cultural information systems research. While much work still remains to be done to validate other hybrid methodologies with further diverse community subjects, it is important for scholars of information science to consider both quantitative and qualitative questions in ontology construction and deployment within emergent digital systems. We anticipate important future work being done to extend these research findings and further close the gap between qualitative and quantitative community data modeling.

References

Aitchison, J, Bawden, D. and Alan Gilchrist (2000). Thesaurus Construction and Use: A Practical Manual. Routledge, UK

Alani H., Kim S., Millard D. E., Weal, M. J., Hall W., Lewis P.H. , Shadbolt, N. R. (2003) "Automatic Ontology-Based Knowledge Extraction from Web Documents". IEEE Intelligent Systems, 18(1):14-21.

Appadurai, Arjun (1998). "Disjuncture and Difference in the Global Cultural Economy," pp. 27-47 and "The Production of Locality," pp. 178-200. In: Modernity at Large Cultural Dimensions of Globalization. Minneapolis: University of Minnesota Press.

Bates, Marcia J. (1989). "The Design of Browsing and Berrypicking Techniques for the Online Search Interface" Online Review 13 (October 1989): 407-424.

Belkin, N. and Croft, B. (1992). “Information Filtering and Information Retrieval,” Comm. ACM, vol. 35, no. 12, pp. 29-37.

Berners-Lee, T. and Miller, E. (n.d) “The Semantic Web Lifts Off” http://www.ercim.org/publication/Ercim_News/enw51/berners-lee.html

Bishop, A. P., Mehra, B., Bazzell, I., & Smith, C. (2000) "Socially Grounded User Studies in Digital Library Development." First Monday, 5 (6) June 2000.

K. Bodker and J. Pedersen, (1991). "Workplace Cultures: Looking at Artifacts, Symbols and Practices," in J. Greenbaum and M. Kyng (eds.), Design at Work: Cooperative Design of Computer Systems, Lawrence Erlbaum Associates, Hillsdale, NJ, 1991, pp. 121-136.

S. Bodker and K. Gronbaek, (1991). "Cooperative Prototyping: Users and Designers in Mutual Activity," International Journal of Man-Machine Studies, Vol. 34, No. 1991, pp. 453-478.

Borgman, C.L. (1989). All users of information retrieval systems are not created equal: An exploration into individual differences. Inf. Process. Manage. 25(3): 237-251.

Caidi, N. and Komlodi, A. (2003a). "Digital libraries across cultures: Design and usability issues" SIGIR Forum, 37(2), Fall issue.

Caidi, N. and Komlodi, A. (2003b). “Cross-cultural considerations in Digital Library Research” D-Lib Magazine 9(7/8). Report of the Workshop. In Brief section.

Cole, C. (1993). Shannon revisited: Information in terms of uncertainty. Journal of the American Society for Information Science, 44(), 204-211.

Costa, L. da, Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. Advances In Physics, 56, 167.

Crabtree, A. (1998) Ethnography in Participatory Design, Proceedings of the 1998 Participatory Design Conference, 93-105, Seattle: Computer Professionals for Social Responsibility.

Crabtree, A. et al. (2000) Ethnomethodologically informed ethnography and IS design. Journal of the American Society for Information Science and Technology, 51 (7) 666-682.

Dervin, B. (1998). Sense-Making theory and practice: An overview of user interests in knowledge seeking and use. Journal of Knowledge Management, 2 (2), 36-46.

Dervin, B. (2001). What we know about information seeking and use and how research discourse community makes a difference in our knowing.

Dervin, B. (2003). Human studies and user studies: A call for methodological interdisciplinarity. Information Research [On-line serial] 9 (1), paper 166.

Dervin, B., & Nilan, M. (1986). Information needs and uses. Annual review of information science and technology (Vol. 21, pp 3-33). White Plains, NY: Knowledge Industry Publications.

Duval, Erik, Wayne Hodgins, Stuart Sutton and Stuart L. Weibel (2002) Metadata Principles and Practicalities. D-Lib Magazine, 8(4). Retrieved: 15 Oct 2005. (http://www.dlib.org/dlib/april02/weibel/04weibel.html)

Ess, C. and Sudweeks, F (2001). Introduction , New Media & Society 2001 3: 259-269.

Feld, Scott (1982). Social Structural Determinants of Similarity among Associates. American Sociological Review 47: 797-801.

Fisher, Karen E., Joan C. Durrance, Marian Bouch Hinton (2004). Information grounds and the use of need-based services by immigrants in Queens, New York: A context-based, outcome evaluation approach. Journal of the American Society for Information Science and Technology 55(8): 754-766.

Flythe, F. (2001). Identification of the information needs of the newly arrived Hispano/Latino immigrants in Durham county, North Carolina. Master Thesis. University of North Carolina Chapel Hill.

Frank, R. n.d. The Tribal Digital Village: Sovereignty, Technology, and Collaboration in Indian Southern California.

Garfinkel, H. (1967) Studies in Ethnomethodology, Englewood Cliffs, New Jersey: Prentice Hall.

Garshol, L. (2004) Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all. Journal of Information Science. 30:378-391.

Geertz, C. (1978), The bazaar economy: Information and search in peasant marketing. American Economic Review 68(2): 28-32.

Gil, T. (2004). Building semantic bridges between museums, libraries and archives: The CIDOC

Conceptual Reference Model. First Monday. volume 9, number 5.

Ginsburg, F., (2005). Re-thinking the Digital Age. http://flowtv.org/?p=651 and www.media-anthropology.net/ginsburg_digital_age.pdf

Golder, S. A. and Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science 32: 198-208.

Goldstone, R. and Yun Son, J. (2005). Similarity. In The Cambridge Handbook of Thinking and Reasoning. Chapter 2. Cambridge University Press.

Granovetter, Mark. (1973). The Strength of Weak Ties. American Journal of Sociology 78: 1360-80.

Gregory, J., (2003). Scandinavian Approaches to Participatory Design. International Journal of Engineering Education, 2003. 19(1).

Gurstein, Michael (2000a). Community informatics: enabling communities with information and communications technologies. Hershey, PA, Idea Group Pub.

Gurstein, M. B. (2000b). "Effective use: A community informatics strategy beyond the Digital Divide" First Monday, volume 8, number 12. (http://firstmonday.org/issues/issue8_12/gurstein/index.html)

Hammersley, M. and P. Atkinson (1983) Ethnography: Principles and Practice. London: Tavistock.

Harrison, T.M. and J.P. Zappen, (2004). Methodological and Theoretical Frameworks for the Design of Community Information Systems. Journal of Computer-Mediated Communication, 2004. 8(3).

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100–108.

Hiller, H.H and Tara M. Franz, (2004). New ties, old ties and lost ties: the use of the internet in diaspora , New Media & Society 2004 6: 731-752.

Jacob, E. (1999) The everyday world of work: two approaches to the classification in context. Journal of Documentation, 57(1), 76-99.

Jain, V.; Krishnapuram, R., (2001). Applications of fuzzy sets in personalization for e-commerce. IFSA World Congress and 20th NAFIPS International Conference. Volume 1, Issue , 25-28 July 2001 Page(s):263 - 268.

Jansen, B. and McNeese, M. (2005). Proceedings of the American Society for Information Science and Technology. 41(1):280-286.

Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

Kling, R., Rosenbaum, H., & Sawyer, S. (2005). Understanding and Communicating Social Informatics: A Framework for Studying and Teaching the Human Contexts of Information and Communications Technologies. Medford, New Jersey: Information Today, Inc.

Kuhlthau, C.C. (2004). Seeking meaning: a process approach to library and information services. (2nd. ed.). Westport, CT: Libraries Unlimited.

Kuhlthau, C.C. (1991). Inside the search process: information seeking from the user's perspective. Journal of the American Society for Information Science, 42(5), 361-371.

Kumar, R., Raghavan, P. Rajagopalan, S. and Tomkins, A. (2001). On semi-automated web taxonomy construction. Proceedings of the Fourth International Workshop on the Web and Databases.

Lakoff, G. (1987). Women, Fire, and Dangerous Things: What Categories Reveal About the Mind University of Chicago Press.

Legg, C. (2007). “Ontologies on the Semantic Web”. In Cronin, B. (Eds), Annual Review of Information Science and Technology. Volume 41.

Lin, F. and Hsueh, C. (2006). Knowledge map creation and maintenance for virtual communities of practice, Information Processing & Management. Volume 42, Issue 2, Pages 551-568.

Loader, B. and L. Keeble (2002). Community informatics : shaping computer-mediated social relations. New York, Routledge.

Marchionini, Gary (2000). Evaluating Digital Libraries: A Longitudinal and Multifaceted View. Preprint from Library Trends, Fall 2000 vol 49(2), p. 304-333.

Matei, S., and Ball-Rokeach, S. (2001). Real and virtual social ties: Connections in the everyday lives of seven ethnic neighborhoods. American Behavioral Scientist, 45(3), 550-564.

Medelyan, O. and Witten, I. (2008). Domain-independent automatic keyphrase indexing with small training sets. Journal of the American Society for Information Science and Technology 59(7):1026:1040.

Menou, M.J. (1983). Cultural barriers to the international transfer of information. Information Processing and Management. 19(3). 121-29.

Merton, Robert and Elinor Barber. 2004. The Travels and Adventures of Serendipity. Princeton: Princeton University Press.

Murtagh, F. (1984) Structure of hierarchic clusterings: implications for information retrieval and for multivariate data analysis, Information Processing & Management. Volume 20, Issues 5-6, Pages 611-617.

Newman, M. E. J. (2006a). Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74.

Newman, M. E. J. (2006b). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103, 8577.

Nielsen, M. L. (2005). Task-based evaluation of associative thesaurus in real-life environment. Proceedings of the American Society for Information Science and Technology. 41(1):437-447

NISO - National Information Standards Organization (2005). Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Bethesda, MD: NISO Press

O'Reilly, T. (2005). What is Web 2.0.

http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Pettigrew, K. E. (1999). Waiting for chiropody: contextual results from an ethnographic study of the information behavior among attendees at community clinics. Information Processing & Management, 35(6), 801-817.

Putnam, Robert (2000). Bowling Alone: The Collapse and Revival of American Community. NewYork: Simon and Schuster.

Pinkett, R. (2003). "Community technology and community building: Early results from the Creating Community Connections Project." The Information Society, 19(1), p. 365-379.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications, 10 (2).

Porter, M. A., Mucha, P. J., Newman, M. E. J., & Friend, A. J. (2007). Community structure in the United States House of Representatives. Physica A, 386.

Putnam, R. D. (2000). Bowling Alone. The collapse and revival of American community, New York: Simon and Schuster.

Quan-Haase, Anabel, Barry Wellman, James C. Witte, and Keith N. Hampton (2002). Capitalizing on the Net: Social Contact, Civic Engagement, and Sense of Community. pp. 291-324 in The Internet in Everyday Life, edited by Barry Wellman and Caroline Haythornthwaite. Oxford: Blackwell

Robbins, M. C., and Nolan, J. M. (1997). A Measure of Dichotomous Category Bias in Free Listing Tasks' Field Methods. Cultural Anthropology Methods 9(3):8-12

Rodriguez, M.A. and Pepe, A. (2008). On the relationship between the structural and socioacademic communities of a coauthorship network, Journal of Informetrics. 2(3):195-201.

Romney, A., Boyd, J., Moore, C., Batchelder, W. and Brazill, T. (1996). Culture as shared cognitive representations. Proceedings of the National Academy of Sciences 93: 4699-470

Schuler, D. (1994). Community Networks: Building a New Participatory Medium. Communications of the ACM, 37(1).

Schware, R. (2005). E-Development From Excitement to Effectiveness. World Bank Publications.

Shepard, R. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Psychometrika. Volume 27, Issue 3.

Shirky, Clay (2002) Communities, Audiences, and Scale. In Clay Shirky, Writings About the Internet: Networks, Economics, and Culture. Written: 6 April 2002. Retrieved: 6 May 2005. (http://shirky.com/writings/community_scale.html)

Spradley, James P. (1980) Participant observation / James P. Spradley. Published : New York : Holt, Rinehart and Winston, c1980.

Srinivasan, R., and Huang, J., (2005), “Fluid Ontologies for Digital Museums” Journal for Digital Libraries, Springer Verlag, Vol 5, No.3, pp.193-204.

Srinivasan, R. (2004). "Reconstituting the Urban through community-articulated digital environments", Journal of Urban Technology, Taylor and Francis, New York/London.

Srinivasan, R. (2005). Weaving Spatial, Digital and Ethnographic Processes in Community-Driven Media Design. Doctoral Dissertation, Graduate School of Design. Harvard University.

Srinivasan, Ramesh, (2006a). "Indigenous, Ethnic, and Cultural Articulations of New Media", International Journal of Cultural Studies, 9(4).

Srinivasan, Ramesh, (2006b). "Where Information Society and Community Voice Intersect", The Information Society, 22(5).

Srinivasan, R., and Shilton, K., (2006). "'The South Asian Web': An Emerging Community Information System in the South Asian Diaspora" ACM Proceedings of the Participatory Design Conference (2006), ACM Press, http://pdc2006.org/.

Srinivasan, R, (2007). "Ethnomethodological Architectures - The Convergence Between an Information System and the Cultural Landscape", Journal of the American Society of Information Science and Technology, 58(5)

Srinivasan, R. and Pyati, A. (2007)., "Diasporic Information Environments: Re-framing Information Behavior Research"., Journal of the American Society of Information Science and Technology, in press.

Star, Susan Leigh (1989). The Structure of ill-structured solutions: boundary objects and heterogeneous distributed problem solving. Distributed artificial intelligence, (eds. L. Gasser and MN Huhns). London: Pitman

Su, S.S., Conaway, C.W. (1995), "Information and a forgotten minority-elderly Chinese immigrants", Library and Information Science Research, Vol. 17(1).

Suchman, L. (1995). Making Work Visible. Communications of the ACM, 38(9).

Svenonius, E. (2000). The Intellectual Foundation of Information Organization. MIT Press.

Tudhope D., Alani H., Jones C. (2001) "Augmenting Thesaurus Relationships: Possibilities for Retrieval." Journal of Digital Information. 1:8

Tversky, A. (1977). Features of similarity. Psychological Review. Vol 84(4) 327-352.

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. Information Processing & Management, 35(6), 819-837

Vakkari, P. & Savolainen, R. & Dervin, B. (Eds) (1997). Information Seeking in Context. Taylor Graham: London & Los Angeles 1997.

Vanderwal, Thomas. (2007). Folksonomy Coinage and Definition. http://vanderwal.net/folksonomy.html

Vertommen J., Janssens F., De Moor B., Duflou J. (2004). "Advanced Personalization and Document Retrieval Techniques in Support of Efficient Knowledge Management". To be published in Proceedings of the Digital Enterprise Technology Conference, Seattle, 2004.

Wang, J. (2005). Automatic thesaurus development: Term extraction from title metadata. Journal of the American Society for Information Science and Technology. 57(7):907-920.

Warschauer, M. (2002). Reconceptualizing the digital divide. First Monday 7(7).

Wasserman, Stanley and Katherine Faust. (1993). Social Network Analysis: Methods and Applications. Cambridge, MA: Cambridge University Press.

Whorf, Benjamin (John Carroll, Editor) (1956). Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf. MIT Press.

Wilkinson, D. M., & Huberman, B. A. (2004). A method for finding communities of related genes. Proceedings of the National Academy of Sciences, 101 Suppl 1, 5241–5248.

Wilson, T.D. (1999). Models in information behaviour research. Journal of Documentation, 55(3), 249-270.

Woolcock, Michael. (1998). Social Capital and Economic Development: Toward a Theoretical Synthesis and Policy Framework. Theory and Society 27: 151-208.

Wu, F., & Huberman, B. A. (2004). Finding communities in linear time: A physics approach. The European Physical Journal B, 38 (2).