ADAPTING MULTIMEDIA DESIGN TO CONTEXT

A design framework for interactive, user context-adaptive multimodal learning environments

JENNIFER BRUNGART
School of Design, Arizona State University, USA.

and

HARI SUNDARAM, HARINI SRIDHARAN, ANKUR MANI AND DAVID BIRCHFIELD
Arts Media and Engineering Program, AME-TR-2004-08
Arizona State University, USA.

Abstract. The proposed dynamic presentation scheme for learning environments discussed within this paper serves as an investigation into how high school-level content (i.e. geography) may be presented so that it addresses and adapts to the comprehension level, existing knowledge base and worldview of individual students, and may be comprehended quickly and clearly, as well. The spatio-temporal arrangement of the content is intended to respond to the relative preparedness of students to comprehend particular content, as well as the nature and complexity of joint comprehension of multi-modal content. The richness of the learning environment is dependant upon the simultaneous presentation of associated textual, visual and auditory information. The environment offers individual students some opportunity to process information in an active manner, through analysis and synthesis of associated sets of content into a thorough and holistic conceptual understanding (i.e. of the causes and effects of population density).

1. Introduction

The proposed learning environment discussed within this paper serves as an investigation into how high school-level geography-related content (i.e. the causes and effects of population density) may be presented so that it
addresses and adapts to the comprehension level, existing knowledge base and worldview of individual geography students, and may be learned quickly, clearly and in a manner that prompts some active-as opposed to only reflective–processing of information. The learning environment is meant to provide students with opportunities for dynamic interaction with multi-modal information; to include efficient spatio-temporal arrangements of information; to adapt user context to content; to adapt the manner in which content is presented to the level of comprehension of individual users; and to addresses some preferences of active learners.

Intended as a learning environment and not as a reference source for the search and retrieval of information, the form taken by the dynamic presentation scheme discussed in this paper may be described as functioning similarly to a “chapter” within a media-rich, electronic “textbook.” The projected goal of students’ interaction with the environment is the processing and comprehension of geography-related content, through both reflective and active cognitive means, via a user context that is consistent with individual students’ past interactions with the environment.

The relationships explored and addressed within the context models described in this paper include semantic inter-relationships among concepts, linguistic and statistical relationships, and common sense rules, such as those found within the OpenCyC database (i.e. a general knowledge base and commonsense reasoning engine). The spatio-temporal arrangement of information demonstrated within the environment is based upon, and is intended to respond to, the relative preparedness of individual students to comprehend particular geography-related content, as well as the implications of necessarily joint comprehension of multi-modal content (i.e. textual, visual and auditory) as related to the informational complexity of individual media elements.

The abstract framework of the learning environment consists of a content repository, an intelligent interaction interface and a dynamic presentation engine. This framework is based upon the necessary components of an experiential learning environment which seem to include: (a) the efficient representation of content; (b) the opportunity for the optimal selection of information; (c) context-sensitive and dynamic presentation synthesis; and (d) temporal content adaptation.

A model mechanism was developed and implemented for the objective evaluation of user experience with, and comprehension of content presented within, the learning environment. This evaluation involved the assessment of user context as related to the projected goal of interaction with the environment (i.e. comprehension of causes and effects of population density) among subjects within a pilot study.
2. Content Strategy

High school-level students were selected as the target audience for this learning environment since it seems that these students—perhaps more than any other group throughout grades K-12 among Western cultures—are denied regular educational experiences that extend beyond the presentation of abstract (i.e. theoretical, conceptual) content in a linear manner. According to (McCarthy, 1987), the tradition of K-12 classroom learning materials—particularly at the high school level—has tended to benefit those learners with reflective cognitive preferences, who prefer to process information by thinking about observations. Unfortunately, this approach to the presentation of information does not seem to be compatible, for example, with the dominant preferences of active learners—approximately half of all people—who prefer to process information by making and doing (i.e. using information in order to understand it). The content presented within both print and electronic media-based learning materials has tended to be structured sequentially, placed emphasis on the classification of concepts and focused on the representation of theory in the absence of experience with concrete, sensory-based information.

There seems to be a bias among Western cultures that intellectual maturity consists of, primarily, logical and rational thought and the “ability to be abstract.” Consequently, students educated within traditional learning environments are likely groomed to “outgrow” their need to learn through experience with concrete information by the time they reach high school. Thus, it appears that the content of the majority of disciplines taught at the high school level is presented in such a way as to require that it be learned reflectively (i.e. through reading and thinking about theories and concepts). It appears that this is true for geography, as well—despite the inherent concrete nature of the physicality and consequences, for example, that may be associated with the discipline.

Geography was selected as the category of content for presentation within the learning environment—specifically, the causes and effects of population density. The opportunity to consider, for example, ways that concrete (i.e. sensory-based) information and spatial relationships associated with geography, might be represented within an electronic environment seemed to present an interesting challenge. Indeed, there seems to be some information related to the geographical causes and effects of population density that may only be fully understood through the comprehension of sensory-based information (e.g. proximity to water, noise pollution, the amount of land needed to feed a herd of cattle for one year, etc.). Also, the exploration of how information might be presented within an electronic environment—in a manner that is engaging and bears an impact—to help students actively understand how their particular ways of life have been influenced by geography and vice versa seems a worthwhile pursuit. According to
(Wiggins & McTighe, 1998), one’s capacity to empathize is indicative of a “mature understanding” of, and relationship with, the world. Perhaps the insight that students gain through an understanding of the causes and effects of population density could be a seed from which a broader understanding of relationships among the Earth and humanity might develop.

Voluntary content standards for twelve disciplines taught in K-12 education were created in the United States in recent years, under the leadership of discipline-based educational associations (Kendall & Marzano, 1996). These content standards are dedicated to learning as the development of thinking and reasoning skills, not the rote memorization of facts, and emphasize what are believed to be the most important concepts within a discipline—considered to be worthy of “enduring” understanding (Wiggins & McTighe, 1998).

The standards for geography were referenced in the development of the content presented within this learning environment (e.g. “Understands the nature, distribution and migration of human populations on the Earth’s surface,” “Understands the patterns of human settlement and their causes,” “Understands how human actions modify the physical environment,” etc.). The appropriation of content directly from existing, linear, reflection-based geography textbook sources seemed an ineffective approach in the development of an active, media-rich learning environment. Therefore, content related to the causes and effects of population density was created specifically for and tailored to the objectives of the learning environment.

The interface of the learning environment is screen-based and consists of three separate maps (i.e. of an urban location, of a suburban location and of a rural location) that represent three broad, associated sets of concepts (i.e. three areas with vastly different characteristics of population) to be explored and learned. Particular and various locations on the maps, which correspond with population density-related content, are rollover-sensitive. When one rolls over a sensitive area, a set of associated textual, visual and auditory information is revealed in a preset, timed sequence which forms a collage of text, image and sound—the text and image of which eventually nearly fill the screen (ref. Figure 4). The purpose of the preset, timed sequence is to emphasize some content as being associated, as well as ensure that some content is presented in such a manner as to maximize its potential to be comprehended clearly and quickly.

However, the learning environment is structured in a manner that allows students some opportunity for choice in regard to the order in which the categories of content (i.e. urban, suburban, rural) may be accessed. This freedom is intended to provide the students with a sense of exploration and discovery of content that, perhaps, might add a level of personal relevance to students’ relationships with a learning experience. This opportunity for a sort of personal relationship with the environment and content might then prompt
students to feel motivated, intrinsically, to engage in the experience. The
certainty afforded in allowing students to make choices in how to pursue
learning is of particular importance to active learners in terms of their
achievement of cognitive satisfaction (Brosterman & Togashi, 1997)
(McCarthy, 1987).

The richness of the learning experience is dependant upon the simultaneous
presentation of textual, visual, and auditory information and the capacity of
the electronic environment to simultaneously convey multiple levels of
information (e.g. textual, visual, auditory) strongly suggests the potential of
the electronic “textbook” as an educational tool. It is intended that the
information communicated through the three modes—text, image and sound—
be related and mutually supportive to whatever overriding concept they
refer. Preferably, however, the information presented across the three modes
should not be repetitive. For example, text should not simply describe what
might already be understood by looking at an image or listening to sounds.
Instead, the information communicated through one mode should be
composed and positioned so as to elaborate upon and enhance, rather than
illustrate, the information communicated through the other two. Repetition
of the same information through multiple modes of communication might
likely seem redundant and be a waste of media opportunity.

Nonetheless, text and images that are mutually dependant should be
positioned in close proximity to one another in order to offer a clear spatial
indication of what information should be associated together and what
should not. Attention to this sort of detail is of particular importance when
content is arranged in a collaged (i.e. overlapped) manner, since it may be
quite easy to accidentally pair information of which such immediate
association is not intended.

Keywords relevant to particular concepts appear within a dedicated section
at the right side of the interface. This is to allow students to quickly and
easily refer to the terminology that is important for them to know and
understand.

Concrete information includes that information which is perceived through
the senses (i.e. visual, auditory) or refers to things, or characteristics of
things, that are perceived through the senses (i.e. may be textual). By
contrast, abstract language refers to theories and concepts and, consequently,
is not sensory-based or particularly specific (McCarthy, 1987). Concrete
words possess the potential to add precision to communication and may be
used to cast unfamiliar information with familiar details (Davis, 2002). This
method of relating new information to existing knowledge may help to
establish the information as relevant and worthy of attention. According to
(Sperber & Wilson, 1995), as a function of preservation, the human brain is
wired to dismiss information that it does not perceive to be relevant.
Consequently, for a person to be inclined to process new information, she must consider the information to be relevant and, therefore, meaningful.

The images presented within the learning environment are in focus and full-color and the sound is clear so that the information being conveyed through these modes, as well as the nature and quality of that information, may be understood to the fullest extent. Collages of both sound (McAdams & Bigand, 1993) and image (Adams, 1986) possess the capacity to evoke a sense of time and place, the details of an environment, etc. In fact, it is possible for people to make clear inferences based upon the information communicated through both visual and auditory means. For example, a “soundscape” that conveys a passing jumble of indistinguishable footsteps and voices, along with engines humming and horns honking, might remind someone of a busy urban street. In contrast, the sound of a light wind blowing through the leaves on trees, a single pair of footsteps on a gravel drive and a bird crowing in the distance might remind someone of a rural location. Additionally, however, one may estimate, based solely upon the sensory-based details conveyed within these soundscapes, that the urban area is more densely populated than the rural area. Details pictured within an image have the capacity to suggest the relative population density of a location, as well (e.g. the number of people in the image, the height of or distance between buildings, the number of cars, waste receptacles and/or trees and grass pictured, etc.).

This learning environment provides opportunities for both active and reflective information processing. The manner in which the content is presented requires that, initially, the information be ingested in a reflective manner as the content is presented on the screen. However, the cognitive processes that students must undertake in order to analyze and synthesize the new information into their own holistic understanding of the causes and effects of population density are, indeed, quite active. Analysis and synthesis are two domains of higher order thinking articulated within Bloom’s Taxonomy (Bloom, 1956) and the ability to analyze and synthesize information is considered to be a valuable skill in today’s world. According to (Davis, 2002), “Adult work in this century depends on the ability to imagine that which does not exist and to visualize meaningful patterns in complex data.” In other words, facility in the analysis and synthesis of information into fully manifested ideas–like those higher-level thinking skills fostered through activity-based learning–are lauded and highly marketable in the workplace. While the learning environment does not offer students the opportunity to make or create anything of literal substance, per se, the environment permits, if not requires, students to draw their own conclusions about the implications of population density among the information that has been presented to them.
It should be noted that it seems that it is not so much the medium or formal manner in which content is presented that determines whether an environment is equipped to support a particular learning experience but, rather, the nature of the content itself (Brungart, 2004). For example, if a body of content is written in such a way as to require that it be processed reflectively, then the experience of processing that information will be reflective—even if it is presented within an electronic environment among multiple “links.” Clicking a mouse to access information, in and of itself, does not make a learning experience cognitively active (i.e. versus reflective). Open-ended, activity-based learning experiences—which facilitate individual creativity, promote choice and allow for multiple possible outcomes—it seems, support active cognitive processing. In contrast, closed-ended, exercise-based learning experiences—which are linear, emphasize step-by-step processes and require single, correct outcomes—do not (Davis, 2002).

3. Context Model

Our previous work on context models has successfully been applied to different problems (Shevade & Sundaram, 2003; Sridharan, Sundaram, Brungart, & Birchfield, 2004). The formal model is defined using a semantic-net— a graph $G = \langle V, E, W \rangle$ where the nodes $v_i \in V$ represent the concepts, the edges $e_{ij} \in E$ represent the type of relationship (semantic, spatio-temporal, feature-level) between the nodes $i$ and $j$ and $w_{ij} \in W$, specifies the strength of the relationship between the two nodes (ref. Figure 2). In this paper, we extend the notion of a concept to multimodal concept. Thus, a concept node is associated with a specific instance that could be an image, video, an audio segment, text or a combination of one or more of these. We define the context be the union of semantic-nets:

$$C = \bigcup_{i=1}^{k} G_i$$

(1)

Figure 1: The concept waterfall being represented using text, images and sound.
where $C$ is the context, $k$ is the total number of semantic-nets and $G_i$ is the $i^{th}$ semantic-net. In the following subsections, we discuss (a) the composition of user context, (b) the construction of user context (c) the relationship between media and the user context and (d) the context evolution during the user’s interaction with the environment.

3.1 User context

As depicted in we define the user context to comprise concept nets composed from (a) the initial user profile (stating the user’s interests, background etc.), (b) the viewing history (it establishes the importance of knowledge the user has acquired while browsing the environment), (c) user behavior (locations visited and time spent on each media element) and (d) user’s learning goals (ref. Figure 3). This information is used to make concept nets that together form the user context.
3.2 Context Construction

The initial user context is modeled using the initial user profile - details like gender, ethnicity, age, profession, cultural interests and others. Concept nets are built using this data as discussed further. The user’s context is then the union of such concept nets (ref equation (1)). The user context dynamically changes with media history, behavior and goals, as the user explores the environment.

We extend our earlier framework (Sridharan et al., 2004) to incorporate multiple media (text, audio and images) to represent concepts. Thus, the initial input given by the user as a part of her profile could be in the form of text, audio clips and images. A concept cover for a set of media elements is defined as a concept that is representative of the set of concepts associated with those media elements. We now discuss how concept covers and concept nets are built for each media type.

Concept Cover for Text: The textual entities are expanded on their synonym sets (synset) and also generalized using WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1993). The entities that fall under one generalized synset (the synset for a word are a group of words which sufficiently characterize the semantics associated with the word) are grouped as belonging to one concept. The relationships between the concepts are given by generalization/specialization of WordNet. The text annotation of the image and the audio inputs are also incorporated in the context of the user in the same way as the text inputs.

Concept Cover for Images: The color histogram of the images in the user context is computed. We use the HSV color space and the histogram is calculated with 166 bins. Image histograms are clustered using the K-Means clustering algorithm (Duda, Hart, & Stork, 2001). The cluster center defines the concept cover for the images that fall into that cluster. The cluster is the counterpart of the concept net for texts. The distance between an image and its cluster center defines the image’s relationship to that concept cover.

Concept Cover for Audio Clips: The audio clips are clustered similar to images. However, we use MFCC (Rabiner & Juang, 1993) as the features space. The distance is defined as the mean-squared cepstral distance as discussed in 5.1.2. As in the case of images, the concept cover of the set of audio clips belonging to a cluster is defined by their cluster center. The cluster corresponds to the concept net the case of text. The distance between an audio clip and its cluster center defines the clip’s relationship to the concept cover.

The relationship between the media and the user context is used to select the media to present to the user as discussed further. We now define this relationship.
3.3 Media distance to User Context

The relationship between a media element and the user context is modeled as a dissimilarity measure. We define the dissimilarities for each of the media types in this section.

3.3.1 Text Distance

For defining distance between a text and the (textual concepts in) the user profile, we use the mechanism described in our prior work (Sridharan et al., 2004). Text concept α has two entities associated with it – the parent concept, and the children concepts (as given by the WordNet hierarchy). Each entity implies the concept with a different weight – ω1 (the parent) and ω2 (all the children). Hence the implication that the concept α is true given that another concept β is true is computed as follows:

\[ I(\alpha | \beta = T) = \omega_1 I(\text{parent} | \beta = T) + \frac{\omega_2}{k} \sum_{i=1}^{k} I(c_i | \beta = T) \]

where \( I \) is the implication strength, and \( k \) is the number of children of the concept α. \( c_i \) is the \( i^{th} \) child of the concept, and where \( \omega_1 \) and \( \omega_2 \) are the weights attached to the implications of the parents and the children respectively. The weight \( \omega_1 \) is computed to be inversely proportional to the number of children of the parent, α, i.e.

\[ \omega_1 \propto 1/m \]

where \( m \) is the number of children of the parent.

The distance between the two concepts is determined as follows:

\[ d_U = 1 - I(\alpha | \beta = T) \]

\[ d(\alpha | \beta = T) = d_U / \sqrt{f_\alpha \cdot f_\beta} \]

where \( d_U \) is the un-weighted distance between the two concepts. \( f_\alpha \) and \( f_\beta \) represent normalized knowledge priors for concepts α and β. The priors are used to re-weight the distance. These could be set by the user, as part of her context model or could be determined using the normalized frequency of occurrence of the concept. The distance between a text concept and the current user profile is then the average of the distances between the given concept and all the concepts in the user profiles.

3.3.2 Image Distance
The distance between two images is defined as their low-level color histogram distance (Shevade & Sundaram, 2003). We use the HSV color space with 166 bins. The distance between an image and (the images in) the current user context is then the average distance between this image and all the cluster centers of images in the current user context.

3.3.3 Audio Distance
We define the distance between two audio clips as the mean-squared cepstral distance between them. The distance is only calculated for the first two seconds as the comprehension time for the audio clips is less than 5 seconds (refer section 6.2.2) and the first two seconds sufficiently good estimate of the cepstral distances. The audio sequence is divided into 200ms overlapping frames (100ms overlap) and the cepstrum of frames is computed. We consider only the 0th to 12th coefficients, as the estimates for the higher order coefficients are noisy. The distance between each pair of frames is given as in (Rabiner & Juang, 1993):

$$d^2(f, f') = \sum_{n=0}^{12} (c_n - c'_n)^2$$  \hspace{1cm} (5)

where $f$ and $f'$: the two frames being compared; $c_n$ and $c'_n$: the respective nth cepstral co-efficients. The distance between sequences $F = (f_1, f_2... f_m)$ and $F' = (f'_1, f'_2... f'_m)$ is given as:

$$D(F, F') = \sqrt{\frac{1}{m} \sum_{j=1}^{m} d^2(f_j, f'_j)}$$  \hspace{1cm} (6)

The distance between an audio clip and (the audio clips in) the current user context is the average distance between this clip and all the audio concept covers (cluster centers) of clips in the current user context.

A media concept is considered close to the user if its distance to the user context was less than a threshold (which is different for different media type – text, audio and images). This threshold was chosen by experimentation.

3.4 Context evolution
As the user interacts with the environment, the user context changes based upon the media consumed and the time spent. Here we discuss the process of context evolution that is analogous to human memory. Memory models are important as memory forms an essential part of communicating an experience by the way of associations.

In our system, concepts get added to the user profile as the user interacts with events. We use a leaky bucket model from (Sridharan et al., 2004) for modeling evolution of context. Concepts automatically acquired by the system into the user profile are slowly lost over time at a fixed rate, unless
they are reinforced by the user visiting related events. This concept reinforcement is analogous to forming associations of a current event with an earlier event that generated feelings of a similar experience.

For each collage visited by the user, the experiential system creates a semantic net from the media (refer section 4.2); it also measures the time spent by the user on the collage. This interaction and imbibing of information by the user would result in certain new concepts being introduced in her user context (newly gained knowledge), certain concepts getting reinforced (due to associations with similar concepts in the user’s context), and certain other concepts decaying (put behind over time). Concepts belonging to all three types of media are thus added / their weights are updated (and deleted if their weights become zero) in the user context as a result of user interaction. The weight of concepts in the user profile grow and decays exponentially with time. Concepts that don’t get reinforced periodically are lost. We used the growth and decay equations given in (Sridharan et al., 2004), to calculate concept weights based on time spent by the user in each event.

In this section we discussed the multimodal context model, its creation, concept relationships and context evolution. This model forms the foundation of the dynamic context-adaptive nature of the presentation.

Figure 4: The graphs ((a), (c), (f)) represent the context for the same user at different times, the map (b) shows the specific location used by the student to learn of the geographical concept of population density. The student does a mouse-over over the “fields” location on the map, and the system responds by dynamically creating a collage (ref. (c) ) of images, text and sounds. As the user stays on, the system refines the presentation since the user is assumed to acquired additional information to view this information. Finally, the user context has been updated to reflect the increase in the knowledge.
The Gobi Desert in northeastern China is a windswept, nearly treeless wasteland.

4. **Optimal media selection**

In our multimodal environment, a concept is represented by a set of media elements and their interrelationships. In this section, we discuss the criteria for selecting media elements for the next collage. The media set is given by the domain expert (ref. Figure 5).

A singular concept could be represented using various media elements. These elements have inter-relationships that represent a single concept and

*Figure 5:* The domain expert will associate with each concept, a set of images, sounds and text (each media element is represented as a filled circle).

In our multimodal environment, a concept is represented by a set of media elements and their interrelationships. In this section, we discuss the criteria for selecting media elements for the next collage. The media set is given by the domain expert (ref. Figure 5).

*Figure 6:* The user context and the media elements are mapped to the same semantic space.
should co-occur – spatially and temporally – to be able to convey the concept. Hence, the media elements are grouped together into co-occurring media sets such that each set represents a particular concept. This is done by an expert in the current system; however it may well be automated based upon the media annotations.

Each concept can be represented by a large number of media sets. From these, those subsets of media elements are selected that maximize the coherence with respect to the user context and the location on the map. Hence, media elements are chosen such that they

- They match the user’s current context or their distance from the user context is less than a certain threshold. (ref. Figure 6)
- They are relevant to the location chosen on the map

5. Dynamic presentation synthesis

The most important and distinguishing feature of the system is the dynamic presentation synthesis that optimizes learning and coherence. In this section we discuss the process of presentation synthesis and the dynamic presentation evolution with user interaction. For any given media set and a presentation scheme (collages, in this case), there are various ways in which the collection can be organized and structured - spatially and temporally. For example, a collection of images that are to be presented using a storyboard can have arbitrary arrangements on the image storyboard. The presentation synthesis problem examines: how do we choose an optimal arrangement of images and an optimal display time? The process of presentation synthesis determines (a) presentation duration for the media set for a given concept and (b) the order of concepts (presentation path) to be presented to maximize learning speed given the current user context and constrained by the knowledge flow graph (to aid progressive knowledge gain). The rest of the section discusses the solutions to these problems.

5.1 Media presentation duration

Once a media set is selected to represent a particular concept, the system estimates the optimal time for presentation such that the media elements are well comprehended. We discuss the estimation procedure for comprehension time for this media set in this section. First we discuss the estimation of comprehension time for individual media elements. We use these estimates to find the estimate of the joint comprehension time for the media set.

The effective duration required for presenting a media element depends upon the time the user takes to comprehend its information content. The comprehension time of a media element in turn depends upon its complexity. There is empirical as well as experimental evidence that suggests a relationship between the comprehension time of a media element and its
complexity. As for example, it is well known that in films the close-ups and less complex shots seem to last longer than shots with much detail (Sharff, 1982). The above idea has been successfully used to calculate the comprehension time and generate the visual skims in (Sundaram & Chang, 2001).

Similar ideas exist for audio comprehension. In (Scheirer, 1999), the author says that the perception is related to the identification of structure from the stimuli. This notion forms the foundation of the MPEG-4 structural audio coding and the structural audio orchestra language (SAOL). Experiments in cognitive neuro-psychology show that sentence comprehension is related to the structural complexity of the sentence (Caplan & Waters, 1999; Gibson, 1998).

A fundamental study on the relationship between the subjective difficulty in learning a concept and its Boolean complexity was done in (Feldman, 2000). Boolean complexity of a concept is defined as the number of literals, ‘n’ in its irreducible form (the length of its shortest logically equivalent formula). Feldman showed through his experiments that the human concept learning is dependent on the Boolean complexity of the concept or its logical incompressibility. Feldman’s work suggests the relationship between the compressibility and the concept learning that greatly influences our work on comprehension time of media based upon media compressibility.

In this work, we have attempted to create a model extending this idea and through our experiments found that the joint comprehension time for a set of media elements (visual, audio and textual) depends upon the compressibility of the individual media elements. The experiments had the assumption that all media elements are uncorrelated and thus there comprehension is not conditioned by other elements. This is not the case with produced media but this assumption of independence gives us a conservative estimate of the comprehension time. We now discuss the experiments on the relationship between the comprehension time and complexity of the media elements that help us develop a joint complexity model for a set of media elements.

5.1.1 Visual complexity and comprehension time

From our previous work, the visual complexity of an image is defined to be its Kolmogorov complexity (Sundaram & Chang, 2001). Thus the complexity is:

\[ K_u(x/n) = \min_{p:U(p)=x} \ell(p) \]  

(7)

where, \( U(p) \) denotes the output of the program \( p \) on an universal Turing machine, \( x \) is the string of length \( n \) and \( KU(x/n) \) is the Kolmogorov complexity of the string \( x \) given the length \( n \).
Further the Kolmogorov complexity of any string is shown to be asymptotically upper-bounded by the compression ratio provided by any universal lossless image coding such as the Lempel-Ziv coding (Sundaram & Chang, 2001).

\[
\lim_{n \to \infty} \frac{1}{n} I_{LZ}(X) \to \frac{1}{n} K_n(X \mid n) \\
I_{LZ}(X) + c \geq K_n(X / n)
\] (8)

An experiment on the relationship between this compression ratio and the comprehension time is discussed in (Sundaram & Chang, 2001). Based upon the experiment, the upper bound on the comprehension time for an image with complexity \( c \) is given as follows:

\[ U_b(c) = 2.40c + 1.11 \] (9)

Here the upper bound is the 95th percentile bound. This means that 95% of the time, the images with the complexity \( c \) can be comprehended within this time. The author in (Sundaram & Chang, 2001) assumed that the different images viewed by the user in a produced scene will be uncorrelated and thus the images selected for the experiment were uncorrelated. This gives a conservative upper bound on the comprehension time as correlation among consecutive images only makes comprehension easier and faster.

### 5.1.2 Audio complexity and comprehension time

![Figure 7: Comprehension time plotted against audio complexity and the upper-bound](image)

We define the audio complexity of an audio clip as its Kolmogorov complexity. As shown by eq.(8) and Figure 7 the complexity is asymptotically upper bounded by the compression ratio of the sound clip when compressed by a universal lossless coding scheme.
To determine the relationship between the comprehension time and the compressibility of audio, a simple experiment was conducted as for the visual case. A corpus of 300 sound clips was arranged with compression ratios ranging from 0.4 to 1.0. Each sound clip was 20 seconds long and was sampled from the user’s personal music store to make sure that the user had heard them earlier. The compression ratio for each sound clip was computed when compressed with the lossless audio codec FLAC. Most of the clips had a compression ratio ranging between 0.4-0.8. The audio sequences with a higher compression ratio were generated by adding Gaussian noise to the original audio sequences such that the SNR was between 0db to 1 db. During the experiment, a sound clip was chosen at random and presented to the user. The user was asked to identify the clip in terms of answering the question, ‘what is causing the sound?’ This was done in multiple sessions of five minutes each to avoid tiredness and boredom for the user that affects the data.

The response time for each audio clip was plotted against its complexity. The complexity axis was divided into bins and the histogram of the response times for each bin was plotted. For bins with sufficient number of samples, the histogram showed similarity to the Rayleigh distribution as for the visual case in (Sundaram & Chang, 2001). By using the 95th percentile cut-off for each histogram we get an upper bound on the comprehension time for each bin. The upper bound for the comprehension time for each value of complexity was then estimated by the least squares fit to the upper bound in each bin. The equation of the lines generated is as follows:

\[ U_b(c) = 4.62c + 1.90 \]  

where \( c \) is the normalized complexity and \( U_b \) is the upper bound and \( L_b \) is the lower-bound (calculated at 10th percentile) on the comprehension time. The upper bound signifies that 95% of the time, the audio clip can be comprehended in this time.

### 5.1.3 Text complexity and comprehension time

Prior experiments on sentence complexity show the dependence of comprehension on sentence structure. In (Caplan & Waters, 1999; Gibson, 1998; Miyake, Carpenter, & Just, 1995) the authors show the dependence of comprehension on the working memory usage. These works classify the sentences based upon their complexity. Some principles were suggested regarding certain sentences, such as Center embedded sentences are more complex than right branching sentences or Object relative sentences are more complex than Subject relative sentences. However, there is no relationship shown between the sentences that differ in two of these properties or two degrees of freedom (thus the Right branching object relative may not be compared to center embedded subject relative). Hence
we don’t provide an ordering of complexity over the different categories instead treat each category individually.

For our purposes, we conducted a simple experiment to quantify the time required for comprehension for different categories of sentences. A corpus of sentences was arranged where each sentence belongs to one of these classes with a maximum of two clauses. This is sufficient for our purposes as the environment is created with sentences with no more than two clauses. For a maximum of two clauses, we defined eleven categories as ‘Subject relative (SR)’, ‘Object relative (OR)’, ‘Conjoined subject relative (CSR)’, ‘Conjoined object relative (COR)’, ‘Conjoined role changing (CRC)’, Right branching subject relative (RBSR)’, ‘Right branching object relative (RBOR)’, ‘Right branching role changing (RBRC)’, ‘Center embedded subject relative (CESR)’, Center embedded object relative (CEOR)’, Center embedded role changing (CERC)’.

Sentences were presented to a set of 6 users in a random sequence and the time taken by the user to comprehend the sentence was noted. The 95th percentile of the comprehension time for each category was calculated and was fixed as the upper-bound on the comprehension time for the sentence class. The comprehension times and their upper (95th percentile) and lower (5th percentile) bounds for each class, normalized by the length of the sentence, are as shown in Figure 8.

The upper-bound gives us a conservative estimate of the comprehension time and can still be followed in our system. All the sentences in the database are classified in one of the given categories and a comprehension time equal to the upper bound for that class is allocated to them.

5.1.4 Joint Comprehension time

The experiments discussed above gave an estimate of the comprehension time for the media elements assuming all media elements to be uncorrelated. We now present a model for estimating the joint comprehension time for the
set of media elements representing a particular concept. Since the elements in the media set represent the same concept, they are correlated. However, we take a conservative estimate of the comprehension time by considering the media elements in the co-occurring media set to be mutually uncorrelated. This set may consist of images, audio and text. The estimate of the joint comprehension time of the three elements is given as the maximum of the comprehension time for the audio and the comprehension time for the text and image:

$$t_j = \max(t_A, t_I)$$  \hspace{1cm} (11)

where \(t_I\) = image comprehension time + text comprehension time and \(t_A\) = audio comprehension time (as detailed in section 5.1.2). \(t_J\) known as the joint visual comprehension time is considered to be the sum of image comprehension time and text comprehension time because the user can only process one of it at a time. The restriction is because the information from both the media elements is consumed through the same visual sensory organ and thus can’t be parallel. On the other hand audio can be consumed in parallel with the visual media. The estimate of the joint comprehension time is plotted against the audio and visual complexity axes for a given textual complexity. It can be easily shown that increasing the textual complexity results in the upward shift of the visual comprehension time and the intersecting line until the text comprehension time reaches a value such that the visual comprehension time dominates over the audio comprehension time and \(t_J = t_I\).

In this section, we discussed the estimation procedure for the comprehension time for media (images, audio and text) and also presented a model for estimating the joint comprehension time for a set of media elements. We now discuss the scheme for obtaining the optimum presentation path that maximizes learning speed and coherence of the presentation.

### 5.2 Presentation order

Presentation order of the various information units that comprise the goal concepts to be learnt is crucial to understanding those concepts. A good presentation scheme for a learning environment should help the user learn the required concepts fast, while allowing her to interact freely with the environment. Also, the information presented should be coherent as coherence is very important for maintaining user attention and increasing the learning capability. In this section, we discuss a presentation scheme that addresses these issues.
The goal of the user is to learn a set of concepts related to causes and effects of population density. These goal concepts have scores indicating their importance in learning about population density. Each of these goal concepts also has an ‘associated set’ of concepts, some of which are critical for its understanding (ref. Figure 9). Hence, in the process of finding an optimal presentation order, we try to maximize the presentation of the critical subsets of the associated set of concepts. This has to be done within the constraint of the knowledge flow dependency between the concepts represented by a Knowledge Flow Graph (KFG). A KFG is a directed acyclic graph with vertices representing concepts and the edges representing the progressive knowledge flow. Thus if understanding concept B is predicated on understanding concept A, then concept A should be presented before concept B.

The initial importance scores, associated set for goal concepts and the KFG are given by an expert. Given that the system finds an optimal sequence of concepts to be shown that maximizes the slope of the learning curve. The following steps are performed to select the next concept to be chosen:

1. Select the most important goal concept.
2. For this goal concept, choose the critical concept, \( C_1 \) closest to the user’s current context and then the KFG that \( C_1 \) is a part of.
3. Present to the user, media (from the selected media as described in section 4) representing the concepts in the KFG in the order of knowledge progression.

Figure 9: Associated Set for a goal concept with knowledge flow dependencies within it.
4. It is possible that there is no media in the selected media set that can represent the chosen concept \( C_1 \). In such cases, provide the user an indication to go to another location to find the related media. If the user chooses to stay on in the same location, present to the user the non-critical concept closest to her context.

5. As the presentation progresses reevaluate the importance scores of the goal concepts dynamically. The new importance score is given by:

\[
I_{\text{new}} = I_{\text{old}} \times \frac{C_{\text{seen}}}{C_{\text{present}}}
\]

where \( I_{\text{old}} \) is the initial importance score of the goal concept, \( C_{\text{seen}} \) is the number of critical concepts seen and \( C_{\text{present}} \) is the total number of critical concepts for that goal concept. The above procedure maximizes user learning and presentation coherence. This is corroborated by our user study.

6. **Experimentation and Evaluation**

<table>
<thead>
<tr>
<th>The environment is coherent.</th>
<th>6.2 / 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>The media concepts shown in increasing knowledge progression.</td>
<td>5.0 / 7</td>
</tr>
<tr>
<td>The speed of the presentation is just right.</td>
<td>7.0 / 7</td>
</tr>
<tr>
<td>In the environment, the text, audio and images appeared correlated.</td>
<td>6.8 / 7</td>
</tr>
</tbody>
</table>

In this section we discuss our experiments with real users and the experimental results. We conducted experiments to evaluate two aspects of our presentation system: a) coherence and b) effectiveness for learning.

We evaluated our models through a pilot user study with five users. Users are asked to interact with the system by freely navigating through the different locations on the map. Two different presentation systems were created, the first one with the dynamic presentation synthesis we discussed above and the second with concepts presented in a random manner. They were asked to compare the two presentations with respect to coherence and comprehension of concepts presented. All users found our dynamic presentation synthesis to be more coherent and comprehensible.

The users were also asked two sets of questions regarding their interaction with our learning environment. The first set aimed at evaluating the coherence and comprehensibility of the presentation, and the correlation between different media elements. The users were asked to rate these aspects of the system on a scale of 1-7, 1 representing strongly disagree and 7,
strongly agree. The results obtained are presented in Table 1. Using the standard t-test we can show that these results are statistically significant at confidence value 0.99 or better.

The second set of questions evaluated the user learning. The users were presented with a set of questions and asked to rate their confidence (on a scale of 1 to 7) in answering each of the questions based on their interaction. The questions asked were chosen such that they involved the understanding of relationships between two or more concepts in the presentation. The average score for the questions related to learning was 5.4 (77%).

To evaluate the dynamic nature of the presentation and its influence on learning we define a measure of learning in terms on the distance between the goal concepts (along with its associated set) and the concepts in the user profile. This distance measure is given by

\[ d = W^T S W \]  

where \( W \) is a vector \([1-w_1, 1-w_2, ..., 1-w_n]\) where \( w_i \) is the weight of \( i^{th} \) goal concept in the user profile (the weight is taken to be 0 if the concept is not present in the user profile). \( S \) is a \( m \times m \) matrix where \( m \) is the number of concepts in our system and \( S[i,j] \) is the WordNet implication between the concepts \( i \) and \( j \). As the presentation progresses the distance decreases if the focus on the critical concepts is maintained and the user learns more and more critical concepts.

This distance was regularly computed while the user interacted with the system. The plot of the distances against time for two users is shown in Figure 10. The plot shows a sharper learning curve and larger learning for User 1 as compared to User 2. This was corroborated by the answers to the learning questions from the two users.

![Figure 10: Graph depicting the speed of learning for two users](image-url)
7. Limitations and Future Work

The experimental results are encouraging; however there are limitations of our approach towards modeling the learning environment that we plan to address in our future work. The main limitations of our work are as follows:

- **Similarity measure:** As suggested by Tversky (Tversky, 1977), the metric measures of similarity are inappropriate in certain cases, esp. when it comes to measuring the human notion of similarity between concepts. We plan to investigate into building a similarity measure that can model the “non-geometric” measures of human-similarity.

- **Joint Comprehension Measure:** The joint comprehension model for the media assumes the media to be uncorrelated giving a higher estimate on the comprehension time. We plan to investigate into the joint comprehension of correlated media to fine tune the estimate of the comprehension time.

It is possible that this work may evolve into the pursuit of the development of adaptable, dynamic presentation schemes that are capable of facilitating and supporting opportunities for real “making and doing” (i.e. activity). In other words, the design of an activity-based learning environment may be a next step. Activities may support the creation of visual (e.g. sketches, maps, charts) and/or verbal (e.g. written descriptions of visual representations, stories, lists of ideas) representations. The electronic environment would, of course, afford students the opportunity to learn and create directly within the learning space. An assignment for a creation-type activity that supports the consideration of causes and effects of population density might be, for example: “Develop a story or cartoon that tells of the meeting of three people from three differently populated places (i.e. urban, suburban and rural locations). What would the similarities and differences among the details of their everyday lives be like? What could the three people learn from one another?” Other possible activities might include a charted comparisons of the population density of locations with different geographic profiles or, perhaps, drafting a plan for how some negative effects of population density might be alleviated. It is likely that traditional, objective methods of evaluation might not well matched with the evaluation of activity-based learning experiences. However, the development of a rubric-based system may be an initial option that is explored.

Additionally, future work may include the implementation of content from academic disciplines other than geography into the learning environment. Designed as a model for adaptable, dynamic learning and not for the demonstration of content alone, perhaps the learning environment may be adjusted to support a wide variety of content.

We are also planning to work on the creation on an organic electronic / paper cycle of learning. The key issue is that the current environment creates a
disconnect between the electronic world and the paper interfaces that we are culturally trained to use. *We are looking at the problem of how we can dynamically create context adaptive electronic books* as a consequence of interaction with the electronic world. The key idea is that we shall create visual puzzles / activities based on the interaction with the electronic world. This will actively make use of the cognitive skills of the child to parse the information presented on paper. The child is then expected to naturally interact with paper, making drawings, marking text. The pieces of paper that comprise the electronic book, shall then be scanned in. The scanned data shall be interpreted to determine changes in the user context, in turn affecting the encounter with the electronic world. We would have thus closed the loop, thus creating a rich mediated learning environment.

8. Conclusions

The user study evaluation of our investigation into how high school-level geography-related content may be presented so that it addresses and adapts to the comprehension level, existing knowledge base and worldview of individual geography students, and may be learned quickly and clearly, elicited some encouraging results. For example, it seems that a multi-modal user context may be a novel approach after which to model optimal dynamic user context for electronic media-based applications. Also, such a context may be applied to the temporal adaptation of information in order to create information summaries.

Additionally, the development process and evaluation of the proposed learning environment has prompted new questions in regard to the very nature of the context and phenomena of comprehension that is being explored. As an early attempt, there are indeed many more refinements to be made to and studies to be done of the dynamic presentation scheme demonstrated within this learning environment.

Future work may involve the implementation of activity-based, creation-type learning opportunities and the adaptation of content from other academic disciplines (i.e. other than geography) into the presentation scheme.

9. References


