Hugo Liu and Pattie Maes (2004). What Would They Think? A Computational Model of Attitudes. *Proceedings of the ACM International Conference on Intelligent User Interfaces, IUI 2004, January 13–16, 2004, Madeira, Funchal, Portugal.* ACM 2004, ISBN 1-58113-815-6, pp. 38-45.

What Would They Think? A Computational Model of Attitudes

Hugo Liu
MIT Media Laboratory
20 Ames St., Cambridge, MA, USA
hugo@media.mit.edu

Pattie Maes
MIT Media Laboratory
20 Ames St., Cambridge, MA, USA
pattie@media.mit.edu

ABSTRACT

A key to improving at any task is frequent feedback from people whose opinions we care about: our family, friends, mentors, and the experts. However, such input is not usually available from the right people at the time it is needed most, and attaining a deep understanding of someone else's perspective requires immense effort. This paper introduces a technological solution.

We present a novel method for automatically modeling a person's attitudes and opinions, and a proactive interface called "What Would They Think?" which offers the *just-in-time perspectives* of people whose opinions we care about, based on whatever the user happens to be reading or writing. In the application, each person is represented by a "digital persona," generated from an automated analysis of personal texts (*e.g.* weblogs and papers written by the person being modeled) using natural language processing and commonsense-based textual-affect sensing.

In user studies, participants using our application were able to grasp the personalities and opinions of a panel of strangers more quickly and deeply than with either of two baseline methods. We discuss the theoretical and pragmatic implications of this research to intelligent user interfaces.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: *interaction styles, natural language;* I.2.7 [**Natural Language Processing**]: *text analysis.*

General Terms

Algorithms, Design, Human Factors, Languages, Theory.

Keywords

Affective interfaces, affective memory, user modeling.

1. INTRODUCTION

Have you ever been engaged in a task – whether it's reading the news, writing a paper, or reflecting on life – where you felt uncertain about how to interpret a situation, and your family, friends, or mentors suddenly came to mind, and you thought, "what would they think?" This experience is very common because observing and modeling the attitudes and emotional reactions of others is an important aspect of how humans learn (Bandura, 1977).

If we could get frequent and timely feedback from people whose opinions we value (*e.g.* family, friends, mentors, experts), then perhaps their perspectives would enhance our ability to interpret situations and make decisions. However, we often lack access to the people whose feedback we value, so we are forced to learn about their perspectives in other ways, *e.g.* by inferring attitudes and opinions from prior conversations or from books and papers. Forming a deep understanding of a person in this manner requires immense effort, and there is no guarantee that we can recall a person's opinion on a specific topic at the time we need that feedback the most.

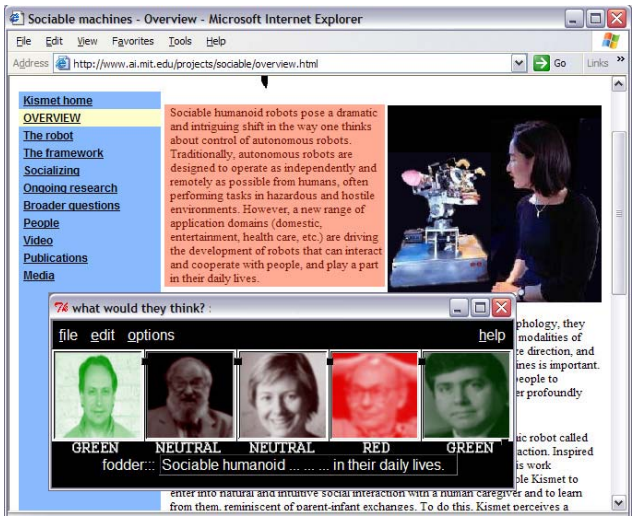

Figure 1. A panel of virtual AI researchers react affectively to a passage of text that the user is reading. A green-tinted face indicates an approving response, while a red-tint indicates disapproval. Brightness corresponds to affective arousal over a topic.

This paper introduces a technological solution to the problem of getting the *just-in-time perspectives* of people whose opinions we care about. We have built a system that can automatically generate a model of a person's attitudes from an automated analysis of a corpus of personal texts written by the person being modeled, consisting of, *inter alia*, weblogs, emails, editorial papers, and transcribed speeches. The proactive interface "What Would They Think?" (WWTT) (Fig. 1) displays a handful of these digital personas together, each reacting affectively to whatever the user happens to be reading or writing. Personas are also capable of explaining why they react as they do, with salient quotes from their personal texts to justify an affective perspective.

Hugo Liu and Pattie Maes (2004). What Would They Think? A Computational Model of Attitudes. *Proceedings of the ACM International Conference on Intelligent User Interfaces, IUI 2004, January 13–16, 2004, Madeira, Funchal, Portugal.* ACM 2004, ISBN 1-58113-815-6, pp. 38-45.

To build a digital persona, the attitudes that a person exhibits in his/her personal texts are organzed into an *affective memory system.* Personas react affectively to newly presented text by considering the affective memories it triggers. Mining attitudes from text is achieved through natural language processing, and commonsense-based textual affect sensing (Liu *et al.*, 2003). This approach to person modeling is quite novel when compared to previous work on the topic (behavior modeling, *e.g.* (Sison & Shimura, 1998), and demographic profiling, *e.g.* questionnaire-derived user profiles).

A related paper on this work (Liu, 2003b) presents an epistemological view on the work and gives a more thorough technical treatment. This paper does not dwell on the implementation-level details of the system, but rather, discusses the computational modeling of attitudes in the context of the "What Would They Think?" application.

This paper is structured as follows. First, we introduce a computational model of a person's attitudes, a system for automatically acquiring this model from personal texts, and methods for applying this model to predict a person's affective reaction to new text. Second, we discuss how a collection of digital personas can provide real-time perspectival feedback in "What Would They Think?" and present two evaluations of our approach. Third, we situate our work in the literature. The paper concludes with further discussion and presents directions for future work.

2. COMPUTING A PERSON'S ATTITUDES

First-person texts such as, *inter alia,* weblog diaries, emails, editorial papers, and transcribed speeches and interviews, are rich sources of attitudes and opinions. People are good at inferring attitudes from text and compiling them into a mental model of the author, but computationally the problem is more challenging.

Our approach to the problem can be summarized as follows. We implement a computer reader to skim a corpus of personal texts and appraise the affect of the text at the sentence and concept level. For this task, we use commonsense-based textual affect sensing (Liu *et al.*, 2003). Concepts, topics, and "episodes" are extracted from text and associated with their respective affective valence scores; each (concept, affective valence score) pair constitutes a single exposure of an *attitude.* The analysis of each personal text yields many attitude exposures, which accumulate in an affective memory system. The affective memory system has a reflexive component, in which repeated attitude exposures are required to form a stable attitude, a method akin to *classical conditioning* in psychology. Conditioning helps to make the attitudes model more robust to errors in the affective appraisal of text.

In this section, we first present a bipartite model of the affective memory system. Second, we discuss the mechanism for mining attitudes from personal texts. Third, we describe how an affective memory system can be applied to predict a person's affective reaction to new texts. Fourth, we present some advanced features that enrich our basic person modeling approach.

2.1 A Bipartite Affective Memory System

A person's affective reaction to a concept, topic, or situation can be thought of as either instinctive, due to attitudes and opinions conditioned over time, or reasoned, due to the effect of a particularly vivid recalled memory. Borrowing from cognitive models of human memory function, attitudes that are conditioned over time can be best seen as a reflexive memory, while attitudes resulting from the recall of a past event can be represented as a long-term episodic memory (LTEM). Memory psychologist Endel Tulving equates LTEM with "remembering" and reflexive memory with "knowing" and describes their functions as complementary (Tulving, 1983). We combine the strengths of these two types of memory to form a bipartite affective memory system.

2.1.1 Affective long-term episodic memory

Long-term episodic memory (LTEM) is a stable memory capturing significant events. The basic unit of memory, called an episode, captures a coherent series of sequential events. Episodes are *content-addressable*, meaning, they can be retrieved through a variety of cues encoded in the episode, such as a person, location, or action. With LTEM, even events that happen only once can become salient memories and can recurrently influence a person's future thinking. In modeling attitudes, we must account for the influence of these particularly powerful one-time events.

In our affective memory system, we compute an affective LTEM as an *episode frame*, coupled with an affect valence score that best characterizes that episode. In Fig. 2, we show an episode frame for the following example episode: "John and I were at the park. John was eating an ice cream. I asked him for a taste but he refused. I thought he was selfish for doing that."

::: EPISODE FRAME :::

SUBEVENTS: (eat John "ice cream"), (ask I John "for taste"), (refuse John)	MORAL: (selfish John) CONTEXTS: (date), (park), () EPISODE-IMPORTANCE: 0.8 EPISODE-AFFECT: (-0.8,0.7,0)

Figure 2. An episode frame in affective LTEM.

As illustrated in Fig. 2, an episode frame decomposes the text of an identified and parsed episode into simple verb-subject-argument propositions like (eat John "ice cream"). Together, these constitute the subevents of the episode. The "moral," or root cause, of an episode is important because the episode-affect can be most directly attributed to it. The details of extracting morals are presented elsewhere (Liu, 2003b).

The affect valence score in the above example is a numeric triple representing valences in the three nearly independent affective dimensions of Pleasure-Displeasure, *e.g.*, feeling happy or unhappy; Arousal-Nonarousal, *i.e.*, heightening one's feelings; and Dominance-Submissiveness, *i.e.*, the amount of confidence/lack-of-confidence felt. This is known as the PAD model (Mehrabian, 1995) for short. Each dimension can assume values from –100% to +100%, and a PAD valence score is a 3-tuple of these values (*e.g.* [-.51, .59, .25] might represent anger). The robustness implications of PAD's continuous account of affect makes it preferable to finite repetoire models of affect such as Manfred Clyne's "sentics" schema (1977) because in PAD, the valences of different discrete emotions can be unified along one of the three PAD dimensions (*e.g.,* fear, anger, surprise can be unified along the Arousal dimension).

2.1.2 Affective reflexive memory

While long-term episodic memory deals in salient, one-time events and must generally be consciously recalled, reflexive memory is full of automatic, instant, almost instinctive associations. Reflexive memories are formed through the conditioning of repeated *exposures* rather than one-time events. The conditioning process also acts as a noise filter against any incorrect textual affect classifications.

The affective reflexive memory is represented by a lookup-table. The lookup-keys are concepts which can be semantically recognized as a person, action, object, activity, or topic. Associated with each key is a list of *exposures*, where each exposure describes a distinct instance of the concept appearing in the personal texts. An exposure, E, is represented by the triple: (date, affect valence score V, saliency S). At runtime, the affect valence score associated with a given conceptual cue can be computed using the formula given in Eq. (1).

$$\frac{1}{n}\left[\log_b(\max(n,b))\right]\left[\sum_{t=startdate}^{enddate} S(E_t)V(E_t)\right] \quad (1)$$

where n = the number of exposures of the concept; $b = 2$

This formula gives the valence of a conceptual cue averaged over a particular time period. The term, $[\log_b(\max(n,b))]$, rewards frequency of exposures, while the term, $S(E_t)$, rewards the saliency of an exposure. In this simple model of an affective reflexive memory, we do not consider phenomena such as belief revision, reflexes conditioned over contexts, or forgetting.

To give an example of how affective reflexive memories are acquired from personal texts, consider Fig. 3, which shows two excerpts of text from a weblog and a snapshot sketch of a portion of the resulting reflexive memory.

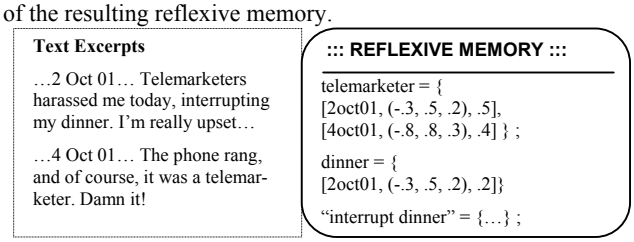

Figure 3. How reflexive memories get recorded from excerpts.

In the above example, two text excerpts are processed with textual affect sensing, and concepts both simple (*e.g.* telemarketer, dinner, phone) and compound (*e.g.* telemarketer::call, interrupt::dinner, phone::ring) are extracted. The saliency of each exposure is determined by heuristics such as the degree to which a particular concept in topicalized in a paragraph. The resulting reflexive memory can be queried using Eq. (1). Note that while a query on 3 Oct 01 for "telemarketer" returns an affect valence score of (-.15, .25, .1), a query on 5 Oct 01 for the same concept returns a score of (-.24, .29, .11). Recalling that this valence triple corresponds to (pleasure, arousal, dominance), we can interpret the second annoying intrusion of a telemarketer's call as having conditioned a further displeasure and a further arousal to the word "telemarketer".

How does conditioning help the system cope with noise? In Fig. 3, "phone" also inadvertently inherits some negative affect. However, unless "phone" consistently appears in a negative affective context in the long run, Eq. (1) will tend to cancel out inconsistent affect valence scores, resulting in a more neutral valence.

In summary, we have motivated and characterized the two components of the affective memory system: an episodic component emphasizing the affect of one-time salient memories, and a reflexive component, emphasizing instinctive reactions to conceptual cues that are conditioned over time. In the following subsection, we propose how this bipartite affective memory system can be acquired automatically from personal texts.

2.2 Mining Attitudes from Personal Texts

The bipartite affective memory system presented above is the framework we use to represent a person's attitudes. We acquire a person's affective memory system by analyzing personal texts, and use this memory to affectively appraise new textual episodes.

In contrast to this simple memory-based affective appraisal model, other affective modeling frameworks in the literature such as (Gratch & Marsella, 2001) have proposed deeper models of affective appraisal by considering beliefs, desires, and goals. Admittedly, these models capture the dynamism of human personality better than a deterministic memory-based model can. However, whereas other model can be acquired automatically from existing personal texts, these deeper models need to be handcrafted. Robustly inferring beliefs, desires, and goals from text remains an unsolved problem in story understanding. While the model presented in this paper needs to be developed further to produce more realistic and believable reactions, our thesis is that even this simple memory-based model can provide perspectival feedback that is helpful to a user.

The process of mining attitudes from personal texts to populate the affective memory system involves natural language processing to deconstruct text into concepts (for the reflexive memory) and episodes (for the episodic memory), and textual affect sensing to assign affect valence scores to each concept and episode. In this paper, we restrict discussion to the textual affect sensing component, leaving details of natural language processing to be presented in (Liu, 2003b). This subsection presents a commonsense-based approach to affective appraisal of personal texts. We also discuss fitness criteria for personal texts, and present some limitations of our affective appraisal mechanism.

2.2.1 Affective Appraisal of Personal Text

Appraising the affect of personal text is a difficult task. The affective classification method needs to judge affect at the sentence-level with good accuracy. Several common approaches fail to meet the criteria. Naïve keyword spotting, which looks for mood keywords, is not robust as a stand-alone method because affect is often conveyed without mood keywords. Statistical affective classification using statistical learning models such as latent semantic analysis (Deerwester *et al.*, 1990) generally require large inputs and thus, cannot appraise texts with satisfactory granularity.

To analyze personal text with the desired robustness and granularity, we employ a model of textual affect sensing using commonsense knowledge, as proposed by (Liu *et al.*, 2003). In this model, defeasible knowledge of everyday people, things, places, events, and situations from the Open Mind Commonsense (OMCS) corpus (Singh *et al.*, 2002) is leveraged to classify the affect of text by evaluating the affective implications of each event or situation. For example, to affectively classify "I got fired today," this model evaluates the consequences of this situation and characterizes it using negative emotions such as fear, sadness, and anger. This model, coupled with a naïve keyword spotting approach, provides rather comprehensive and robust affective classification. The output of the textual affect sensing subsystem is a PAD score.

One of the most interesting issues is learning *personal affect* using a person-neutral affect sensing mechanism. Because the OMCS corpus was built collaboratively by 11,000 web teachers, the affective appraisal made by such a model represents the judgment

of a *typical person,* which can sometimes differ from the affective judgment of the *particular person* being modeled. However, we can assume that although a personal affect judgment may deviate from that of a typical person on small particulars, it is less likely to deviate *on average*, when examining a larger textual context. The implication of this is that moving from the sentence-level to the paragraph- or document-level, the accuracy of the affective appraisal should increase. The evaluation of Liu *et al.*'s affective navigation system (2003b) yields indirect support for the idea that the accuracy increases with the size of the textual context. In that user study, users found affective categorizations of textual units on the order of chapters to be more accurate and useful to information navigation than affective categorizations of small textual units such as paragraphs.

To assess the affect of a sentence, we factor in the affective assessment of not only the sentence itself, but also of the paragraph, section, and whole journal entry or episode. Because so much context is factored into the affect judgment, only a modest amount of affective information can be gleaned for any given sentence. Thus we rely on the confirming effects of being able to encounter an attitude multiple times (*i.e.* conditioning the reflexive memory). In exchange for only being able to learn a modest amount from a sentence, we reduce the impact of erroneous judgments.

2.2.2 What Personal Texts are Suitable?

Suitable personal texts satisfy the following criteria. 1) Texts should be first-person, because having to attribute opinions to multiple sources requires more in-depth story understanding. 2) Texts should be rich sources of candid opinion, so an editorial paper is better than a dispassionate paper. 3) If the digital persona is intended to represent the whole of a person's personality and attitude, the selection of personal texts should cover a good breadth of topics, not covering one or two topics disproportionately. Unbalanced collections of personal texts will generate digital personas skewed toward particular discourses, and texts that are too esoteric will hurt the accuracy of the affective appraisal mechanism because they may be out of the evaluative scope of commonsense knowledge. 4) A text source like a weblog diary is preferred because it covers attitudes and opinions on day-to-day life, and has an explicit episodic organization. Texts not already organized by episodes are heuristically segmented using natural language processing.

Like in any machine learning problem, the quality of the attitudes model varies greatly with the size and quantity of personal texts fed to it. A large corpus of highly suitable personal texts covering a large, well-balanced range of topics will yield the best results.

2.2.3 Limitations

Even if these fitness criteria for personal texts are met, there are still inherent limitations to the affective appraisal mechanism. 1) Since sentences (actually, independent clauses, to be specific) are the largest units of context addressed by the commonsense-based textual affect sensing engine, simple declarative assertions can be appraised much more accurately than complex arguments. The appraisal of "Mr. X" in the following passage would be erroneously positive: "Mr. X is such a nice guy. Everyone loves Mr. X. Gimme a break!" 2) The affective appraisal mechanism cannot recognize humor and sarcasm because these phenomena require a more subtle understanding of text. It does not know that if you say "I hope you die a horrible death" to a friend who has just played a minor joke on you, the statement should not be taken literally.

The system has two strategies for coping with erroneous affective appraisals. First, as discussed in Section 2.2.1, the affective classification of a sentence is tempered by the affect of the paragraph, section, and document in which it is contained, since larger contexts of texts can be appraised with less noise. Second, the process of conditioning in the formation of reflexive memories will tend to cancel out inconsistent instances of affective appraisal.

In summary, digital personas can be automatically acquired from suitable personal texts using natural language processing and textual affect sensing. Suitable texts meet certain fitness criteria such as being first-person, opinion-rich, well-balanced, and explicitly episodic. The proposed affective appraisal mechanism employs coping strategies for dealing with erroneous appraisals, especially over sarcastic, humorous, or argument-based text.

2.3 Predicting Attitudes using the Model

Having acquired the model, the digital persona attempts to predict the attitudes of the person being modeled by offering some affective reaction when it is fed some new text. Both the reflexive and episodic memories contribute to the affective reaction.

In the reflexive memory, the *point-of-view* of the new text is jisted, and using the attitudes model, the system tries to infer whether or not the person would agree or disagree. Point-of-view is extracted as follows. 1) The new text is parsed into objects and associated attributes. For example, "Computers are dumb" will return the object-attribute pair, "(computers, dumb)". Each attribute is affectively appraised using commonsense-based textual affect sensing and a back-off mood keyword spotter, and the "P" dimension of the resulting PAD score gives us a point-of-view instance, e.g. "(computers, dumb)". These point-of-view instances are compared against the person's attitudes model. If the person's point-of-view concurs with that of the new text, the digital persona will predict an agreeing/approving reaction. In addition, for every concept in the new text that triggers an affective memory, the associated arousal and dominance valence scores contribute to the overall arousal and dominance of the outputted affective reaction. Because reflex memories are not usually triggered in the same contexts that they were recorded, the synthesis of many such triggered reflex memories is required for a successful prediction. The premise is that the contexts of many triggered reflex memories will have some commonality and overlap, and this contextual intersection will lead to a better prediction.

The triggering process for episodic memory is somewhat more complex. Episodes are parsed from the new text, and heuristic pattern matching maps this new episode frame into the LTEM's library of episode frames. The attitude predicted by the episodic memory is the sum of the affect valence scores of all the triggered episodes. The weighted sum of the reactions predicted by the reflexive and episodic memories gives us the actual outputted affective reaction (episodic memory is weighed more heavily because it is more contextually precise).

There are some limitations to the proposed attitude prediction mechanism. Just as attitude extraction from personal texts is vulnerable to phenomena like complex argument constructions and sarcasm, so is the affective appraisal of new text. Also, because affective memories are indexed by natural language phrases, there may be many missed opportunities to trigger memories because of vocabulary differences. In (Liu, 2003b), we describe the addition of conceptual analogy to help to bridge slight linguistic differ-

ences. Finally, predicting a person's affective reaction to new text is likely to be error prone if only one or two triggered memories account for the prediction. Longer textual inputs will trigger more affective memories and lead to more successful predictions.

2.4 Enriching the Basic Model

The basic model of a person's attitudes gleans these attitudes from an automated analysis of personal texts. While this basic model is sufficient to produce reactions to text for which there exists some relevant passages in the personal texts, the generated digital personas are still often quite sparse in what they can react to. We have proposed and evaluated some advancements to the basic model. In particular, we have looked at how a person's attitudes model can be enriched by the attitudes models of people whom the modeled person wants to fashion himself/herself after – perhaps a good friend or mentor. More technically, we mean an *imprimer*. Marvin Minsky describes an imprimer as someone whose goals and attitudes we admire and hope to emulate (Minsky, forthcoming). Imprimers can also be any non-person that can be personified to have desires and goals, e.g. the stereotype of a dog-lover, or a religious faith.

From the supposition that we aspire to many of the attitudes of our imprimers, we hypothesize that affective memory models of these imprimers, if known, can complement the person's own affective memory model in helping to predict a person's attitudes. This hypothesis is supported by work in psychoanalysis on attitude introjection (Freud, 1991). Based on Minsky's suggestion that imprimers evoke self-conscious emotions like pride and embarrassment, we developed and implemented a heuristic approach to automatically identifying imprimers from a person's affective memory. Once identified, the system searches for text on the imprimer and attaches the imprimer's affective memory to supplement the person's own affective memory when appraising new textual episodes. (Liu, 2003b) provides a more complete account of imprimers in attitude modeling. In indicative trials, the addition of imprimers enhanced the success of predictions.

In summary, we have presented a reflex-episode model of affective memory as a memory-based representation of a person's attitudes. The model can be acquired automatically from personal texts using natural language processing and textual affect analysis. The model can be applied over new textual episodes to produce affective reactions that aim to emulate the actual reactions of the person being modeled. We have also discussed how the basic attitudes model can be enriched with added information about the attitudes of imprimers of the person being modeled.

In the following section, we describe how digital personas are composed to create the What Would They Think? application.

3. WHAT WOULD THEY THINK?

What Would They Think? (WWTT) is a proactive interface which offers the *just-in-time perspectives* of people whose opinions we care about, based on whatever the user happens to be reading or writing (Fig. 1). The application consists of a panel of "advisors" who sit on the desktop. The system observes a user as he/she browses a webpage, writes an essay, or replies to an email, and the advisors constantly react to the current text being read or written. In this section, we present elements of the interface design, followed by discussions of two evaluations, one for the underlying attitudes model and one for the application.

3.1 Interface Design

Digital personas acquired from an automated analysis of personal text are represented visually with pictures of faces, which occupy a panel (or *n* x *n* matrix, to accommodate more personas). Given a new textual episode, each persona expresses an affective reaction by modulating the graphical elements of its icon. Each digital persona is also capable of explaining what motivated its reaction by displaying salient quotes from its repository of personal texts.

The iconographic face. A virtual representation of a person is given as a normalized, gray-scaled image of that person's face. Affective reactions are conveyed through modulations in the color, brightness, and sharpness of the face image. From early experimentation, we found that faces are a far more convincing visual metaphor for a digital persona than something textual or abstract. People are pre-wired with the ability to quickly recognize and remember faces, and to use a face as a cognitive container for an individual's unique identity and personality.

We deliberately chose to convey affect by modulating the image rather than trying to manipulate facial expression and gaze. It is important to not portray more detail in the face than our attitude model is capable of elucidating, for the face is fraught with social cues, and unjustified cues could do more harm than good. Scott McCloud has explored extensively the representational-vs.-realistic tradeoff of face drawing in comics (1993).

Visualizing an affective reaction. We employ a rather straightforward scheme to map the three PAD dimensions of an affective reaction (pleasure, arousal, dominance) onto the three graphical dimensions of color, brightness, and sharpness, respectively. The baseline image being modulated is gray-scaled and its brightness and contrast are equalized to be uniform across all images. Using a traffic light metaphor, a pleasurable or approving reaction tints a face green, while an unpleasurable or disapproving reaction tints a face red. An affectively aroused reaction results in a brightly lit icon, while a non-aroused reaction results in a dimly lit icon. A dominant (confident) reaction maps to a sharp, crisp image, while a submissive (unconfident) reaction maps to a blurry image. While better visual interfaces may exist, our experience with users who have worked with this interface tells us that the current scheme conveys the affect reaction quite intuitively.

Configuring the panel. Presently, WWTT is configured for several tasks. The user can use WWTT to get the just-in-time perspectives of a panel of advisors reacting to text read and typed. To add a persona to the panel, a user specifies a face icon, and a url to a weblog or to a corpus of texts, which obey the suitability criteria given in section 2.2.2 (although this is not explicitly enforced). In addition, WWTT can be used to visualize the personalities and strong opinions of an online community. The application automatically analyzes any one of several online communities – including a blog ring, a circle of friends on friendster.com, and a usenet newsgroup) – and generates an appropriate matrix of personas. WWTT has also been implemented to react to the content of conversations using speech recognition.

Explanation. A digital persona is capable of some limited explanation. Clicking on a persona's reaction will display a collection of salient quotes from that persona's text. These quotes are generated by backpointers to the text associated with each affective memory. For episodic memory, a particularly salient episode is quoted, while there are many quotes given to support a triggered reflex memory.

The presentation of the quotes is rank-ordered by saliency and relevance. Quotes which make the largest contribution or best exemplify the resulting affective reaction are promoted to the top of the explanation page.

In most cases, triggered quotes can only offer *indirect and partial justification* for a persona's reaction because the context of the quotes does not match the context of the new episode. However, exercising some critical thinking and synthetic reasoning, a user should be able to verify from the indirect explanation whether or not an affective reaction is indeed justified. This lends the interface some fail-softness, as a user can recover if the system erroneously represents a person's reaction.

3.2 Evaluation of Underlying Attitude Models

The quality of attitude prediction was evaluated experimentally, working with four subjects. Subjects were between the ages of 18 and 28, and have kept diary-style weblogs for at least 2 years, with an average entry interval of three-to-four days. A digital persona was automatically generated for each subject from their weblog. The generated digital personas all had a reflexive memory, episodic memory, and an imprimer memory (cf. section 2.4).

In the interview, subjects and their corresponding generated models evaluated 12 short news snippets taken from Yahoo! News. The snippets are each approximately 150 words long, and 4 snippets were selected from each of three genres: social, business, and domestic. The same set of texts was presented to each participant and the examiner chose texts that were generally evocative. The subjects were asked to summarize their reaction by rating three factors on Likert-5 scales.

→ Feel negative about it (1)…. Feel positive about it (5)
→ Feel indifferent about it (1) … Feel intensely about it (5)
→ Don't feel control over it (1)… Feel control over it (5)

These factors are mapped onto the PAD valence format, assuming the following correspondences: 1→-1.0, 2→ -0.5, 3→0.0, 4→ +0.5, and 5→ +1.0. Subjects' responses were not normalized. To assess the quality of attitude prediction, we recorded the spread between the human-assessed and computer-assessed valences,

$$V_{spread} = \left| V_{human} - V_{computer} \right| \quad \textbf{(2)}$$

We computed the mean spread and standard deviation across all episodes along each PAD dimension. On the –1.0 to +1.0 valence scale, the maximum spread is 2.0. Table 1 summarizes the results. Note that smaller spreads correspond to higher accuracy, and smaller standard deviation correspond to higher precision.

Table 1. Performance of attitude prediction.

	Pleasure		**Arousal**		**Dominance**	
	mean spread	std. dev.	mean spread	std. dev.	mean spread	std. dev.
SUBJECT 1	0.39	0.38	0.27	0.24	0.44	0.35
SUBJECT 2	0.42	0.47	0.21	0.23	0.48	0.31
SUBJECT 3	0.22	0.21	0.16	0.14	0.38	0.38
SUBJECT 4	0.38	0.33	0.22	0.20	0.41	0.32
$\text{Baseline}_{\text{static}}$	0.50		$\text{Basline}_{\text{uniform}}$		0.67	

We give two baselines. $\text{Baseline}_{\text{static}}$ always gives a neutral reaction, so the mean spread will be 0.50 on average. In the context of an interactive interface, $\text{Baseline}_{\text{static}}$ is not a fair comparison because it would never produce any behavior. $\text{Baseline}_{\text{uniform}}$ gives a random reaction from –1.0 to +1.0 assuming a uniform distribution, so the mean spread will be 0.67. The most realistic baseline would probably follow a Gaussian distribution, implying a mean spread whose lower bound is 0.50 and upper bound is 0.67.

On average, our approach performed noticeably better than both baselines, excelling particularly in predicting arousal, and having the most difficulty predicting dominance. The standard deviations were very high, reflecting the observation that predictions were often either very close to the actual valence, or very far. The results along the arousal dimension recorded a mean spread of 0.22, and mean standard deviation of 0.20. This suggests that our attitude prediction models confidently outperform baselines in predicting arousal.

For each news snippet, reflexive memory was triggered an average of 21.5 times, episodic memory 0.8 times, and imprimer reflexive memory 4.2 times. We also re-ran the experiment to measure the effectiveness of each type of memory (for details, see (Liu, 2003b)). We found that episodic memory did not contribute much to attitude prediction because of its low rates of triggering (it was hard to map personal episodes to news story episodes). A pleasant surprise was that imprimers seemed to measurably improve performance, which is a promising result.

Overall, the evaluation demonstrates that the attitude prediction approach presented in this paper is promising, but needs further refinement. The highest accuracy and precision is demonstrated along the arousal dimension. The approach does quite well against the active $\text{Baseline}_{\text{uniform}}$, putting it within the performance range of entertainment applications. But the alarmingly poor precision along the pleasure and dominance dimensions throw caution on other possible applications. Taking into account that reactions are often erroneous, we were careful to design What Would They Think? as a fail-soft interface. The reacting faces are *evocative*, and encourage the user to click on a face for further explanation. Used in this manner, the application is fail-soft and users can decide on the merits of the explanations whether the reaction is justified or mistaken. We do not suggest that the approach is yet ready for *fail-hard* applications, such as deployment as a sociable agent, because fallout (bad predictions) can be very costly in the realm of affective communication (Nass *et al.*, 1994).

3.3 Evaluation of the Application

In addition to evaluating the underlying attitude prediction model, we also performed a user study to test the hypothesis that WWTT can help someone grasp the personalities and opinions of a panel of strangers more quickly and deeply than with baseline methods.

The subjects of the user study were 36 college students, formed into three comparable test groups. The examiner used WWTT to created a panel of four individuals who are extensive webloggers (at least two years of regular blogging), using their blogs as the corpora of personal texts. The study was posed as a "game" and the objective is for each subject to do their best to answer questions regarding the general personalities and specific attitudes of the panel of four strangers, who are previously unbeknownst to the subjects. They are allotted 20 minutes to answer 15 questions.

Subjects in Group 1 were allowed to read through the weblogs of the four individuals with a basic text browser as their information interface. Group 2 used a textual version of WWTT. Group 3 used the real WWTT interface. The Group 1 baseline represents how a user typically learns about the perspectives of people without the

assistance of technology. The Group 2 baseline provides a keyword-retrievable textual memory and uses the WWTT interface, controlling for all the non-affective elements of the application.

The textual version of WWTT warrants some further description. Textual WWTT does not use the color dimension (for expression of approval/disapproval) or the focus dimension (for expression of dominance/submission). Only arousal is expressed (through brightness). Concepts are extracted using identical natural language processing mechanisms as in the real WWTT. Whereas arousal is affective in the real WWTT, arousal in the textual version is set proportional to the number of textual memories triggered by the new textual episode. For example, suppose a new textual episode contained the concepts X and Y. X occurs 10 times in the personal texts, and Y occurs 19 times. Thus the total number of textual instances is 29, and the extent of the arousal reaction is proportional to this score. In the explanation mechanism of Textual WWTT, the quotes are rank-ordered to promote quotes which have the greatest number of concepts in common with the new textual episode being evaluated.

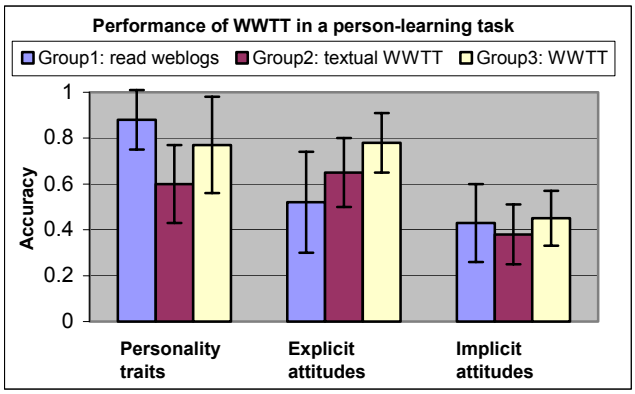

Figure 4. WWTT and two baselines in a person-learning task.

The 15 test questions, given in random orders, fall into three categories of knowledge: general personality traits, specific attitudes explicitly contained in the weblogs, and specific attitudes not contained in, but implied by the weblogs. The examiner is careful to ensure one clear answer for each multiple choice question. Answers to questions on implied attitudes not explicit in the weblogs were verified with the relevant individual. Questions on personality traits are of the vein, "who is the most shy?" Questions on explicit attitudes (*e.g.* "how does Sally feel about religion?") are designed to test the information retrieval capabilities of each tested interface. Questions on implicit attitudes tell of how well each tested interface enables its user to project how a panelist might react to something novel, *e.g.* "what would Sally think of Jim, given the bio on his web page?" The test results are summarized in Fig. 4.

The results are promising. Group 3 consistently outperformed Group 2, came close to Group 1 in "personality traits," and clearly outperformed Group 1 in "explicit attitudes." All three groups struggled with "implicit attitudes" and performed comparably. It was observed that on average, Group 3 subjects spent less time answering each question than Group 1 and Group 2 subjects, and also had to make fewer last-minute guesses on unanswered questions than subjects in the other two groups.

Subjects in Group 1 reported that it felt easy to build an overall picture of a person by skimming an extended sample of their writing as in a weblog. However, searching for "explicit attitudes" handicapped subjects in Group 1, who had to use the search feature in the text editor, but could not query all four panelists in parallel as the WWTT interface enables. At first, subjects in Groups 2 and 3 struggled to come up with text to pose to the application. Many people in both groups came up with a surprisingly efficient strategy of passing in a string of keywords which would define the linguistic context probable to contain the information they wanted. For instance, to answer the question, "who loves to party the most?" a subject in Group 3 typed something like, "party clubbing booze drinking drinks threw up" and then clicked on each face, reading salient quotes in the explanation to verify the attitude. When faced with a choice of which face to click first, subjects in Group 3 usually clicked the one showing the highest arousal. Subjects in Group 3 spent less time sifting through explanation quotes to find a satisfactory answer than subjects in Group 2, suggesting that affective saliency is a useful way to order quotes.

The results of this study support the idea that WWTT allows a user to more quickly and deeply grasp the personalities and specific attitudes of a panel of strangers than either of two baseline approaches. The results suggest that an affective memory can in many cases be a more useful way of organizing and presenting information than a purely textual memory. Despite the poor precision of attitude prediction as suggested in the evaluation of the underlying attitudes model, WWTT's fail-soft explanation mechanism bolstered the usefulness of the attitudes prediction. This study does not directly examine the usefulness of real-time feedback to a user engaged in some task, which would require a study of longer-term interactions.

4. RELATED WORK

The panel of personalities metaphor has been previously explored with Guides (Oren *et al.*, 1990), a multi-character interface that assisted users in browsing a hypermedia database. Each guide embodied a specific character (e.g. preacher, miner, settler) with a unique "life story." Presented with the current document that a user is browsing, each guide suggested a recommended follow-up document, motivated by the guide's own point-of-view. Each guide's recommendations were based on a manually constructed bag of "interests" keywords.

Our affective memory-based approach to modeling a person's attitudes appears to be unique in the literature. Existing approaches to person modeling are of two kinds: behavior modeling, and demographic profiling. The former approach models the actions that users take within the context of an application domain. For example, intelligent tutoring systems track a person's test performance (Sison & Shimura, 1998), while collaborative filtering systems track user purchasing and browsing habits and compare them with those of like-minded people to make predictions about the user's attitudes (Shardanand & Maes, 1995). The demographic approach uses gathered demographic information about a user to draw generalized conclusions about user preferences and behavior.

Neither of these two approaches are appropriate to the modeling of "digital personas." In behavior modeling, knowledge of user action sequences is generally only meaningful in the context of a particular application and does not significantly contribute to a picture of a person's attitudes and opinions. Demographic profiling tends to overgeneralize people by the categories they fit into,

is not motivated by personal experience, and often requires additional user action such as filling out a user profile.

Memory-based modeling approaches have also been tried in related work on assistive agents. The Remembrance Agent (Rhodes & Starner, 1996) uses an associative memory to proactively suggest relevant information. Sunil Vemuri's project, "What Was I Thinking?" (2004) is a memory prosthesis that records audio from a wearable device, and intelligently segments the audio into episodes, allowing the "audio memory" to be more easily browsed.

5. CONCLUSION

Understanding the perspectives of people we care about and having those perspectives available to us *just-in-time* during a task has been up to now a difficult problem with no good technological solutions. In this paper, we present a novel intelligent user interface called "What Would They Think?" that observes what a user reads and writes and proactively shows the affective reactions of a panel of people whose opinions are valued. Each person's attitudes model is built automatically by mining attitudes from their corpus of personal texts. A commonsense-based textual affect sensing engine is adapted to extract personal attitudes.

Both the underlying attitude prediction model and the application were evaluated in user studies. The results suggest that a person's affective arousal is the dimension that can be most accurately and precisely predicted. Arousal was also perceived to be the most useful visual cue in user studies with WWTT. Using WWTT, subjects of a user study were able to more quickly and deeply grasp the personality and specific attitudes of a panel of strangers. We learned that an affective memory is a more helpful organization of a person's attitudes than a purely textual memory.

The automated, memory-based personality modeling approach introduced in this paper represents a new direction in person modeling. Whereas behavior modeling only yields information about a person within some narrow application context, and whereas demographic profiling paints an overly generalized picture of a person and often requires a profile to be filled out, our modeling of a person's attitudes from a "memory" of personal texts paints a richer, better-motivated picture about a person that has a wide range of potential applications. Our user studies illustrate that the model for attitude prediction need not be perfect and free from erroneous predictions to usefully improve a user task. By offering an explanation mechanism, users can independently verify the validity of an affective reaction, and this lends all-important *failsoftness* to our interface.

In future work, we intend to give more prominence to the affective arousal dimension in the interface, as it is the component that can be most accurately predicted by the model. We would also like to investigate further how a persona can be supported by the personas of other people, of social identities, *etc.* For example, a particularly strong belief such as "I love dogs" can cause the persona of the "dog-lover" identity to be attached to one's own existing persona. Finally, we are working on modeling personal attitudes from non-first-person texts, and investigating other applications for our person modeling approach, such as virtual mentors and guides, marketing research, and document recommendation.

6. ACKNOWLEDGMENTS

The authors would like to thank Deb Roy, Roz Picard, Marvin Minsky, Cynthia Brezeal, Bruce Blumberg, Henry Lieberman, Ted Selker, and the blind reviewers for their helpful feedback.

7. REFERENCES

[1] Bandura, A. (1977). Social Learning Theory. New York: General Learning Press.

[2] Clynes, M. (1977). *Sentics: The Touch of Emotions*. Garden City: Anchor Press.

[3] Deerwester, S. *et al.* (1990). Indexing by latent semantic anlysis. *Journal of Am. Soc. of Info. Sci.:416(6)*, pp 391-407.

[4] Freud, S. (1991). *The essentials of psycho-analysis* by Anna Freud. London: Penguin.

[5] Gratch, J., and Marsella, S. (2001). Tears and fears: modeling emotions and emotional behaviors in synthetic agents. *Proceedings of Agents 2001*: 278-285.

[6] Liu, H. (2003b). A Computational Model of Human Affective Memory. *MIT Media Lab Technical Report.* At: web.media.mit.edu/~hugo/publications/papers/ham-tr.doc

[7] Liu, H., Lieberman, H., Selker, T. (2003). A Model of Textual Affect Sensing using Real-World Knowledge. *Proceedings of IUI 2003*, pp. 125-132.

[8] Liu, H., Selker, T., Lieberman, H. (2003b). Visualizing the Affective Structure of a Text Document. *Proceedings of CHI 2003*, pp. 740-741.

[9] McCloud, S. (1993). *Understanding Comics*, Kitchen Sink Press, Northhampton, Maine,

[10] Mehrabian, A. (1995b). Framework for a comprehensive description and measurement of emotional states. *Genetic, Social, and General Psychology Monographs, 121,* 339-361.

[11] Minsky, M., (forthcoming). The Emotion Machine, Pantheon, New York. Several chapters are available at: http://web.media.mit.edu/~minsky.

[12] Nass, C.I., Stener, J.S., and Tanber, E. (1994) Computers are social actors. *Proceedings of CHI '94*, pp. 72-78,

[13] Oren, T., *et al.* (1990). Guides: characterizing the interface. In Laurel, B. (Eds.) *The art of human-computer interface design*. Addison-Wesley.

[14] Rhodes, B. and Starner, T. (1996). The Remembrance Agent: A continuously running automated information retrieval system. *Proceedings of PAAM '96*, pp. 487-495.

[15] Shardanand, U. and Maes, P. (1995). Social information filtering: Algorithms for automating "word of mouth", *Proceedings of CHI'95*, 210-217.

[16] Singh, P., (2002). The public acquisition of commonsense knowledge. *Proceedings of AAAI 2002 Spring Symposium.*

[17] Sison, R. and Shimura, M. (1998). Student modeling and machine learning. *International Journal of Artificial Intelligence in Education, 9*:128-158.

[18] Tulving, E. (1983). *Elements of episodic memory.* Oxford: New York

[19] Vemuri, S., *et al.* (2004). The Design of an Audio-Based Personal Memory Aid. *Submitted to CHI '2004*