Technical Report Number: TR-2008A

PPT: A Probabilistic Approach to Extracting Preferential Probabilities from Discussion Transcripts

September 2008

Ji, Haifeng

Yang, Maria C.

Honda, Tomonori

Massachusetts Institute of Technology, Ideation Laboratory

Cambridge, MA

PPT: A Probabilistic Approach to Extracting Preferential Probabilities from Discussion Transcripts

This report presents an implicit probabilistic approach for extracting a projection of aggregated design team preference information from design team discussion as if the team is a single entity. It further takes into consideration how the design preference information of a team can evolve over time as the team changes its priorities based on new design information. Two initial models are given for representing the most probable and preferred design alternative from the transcripts of design team discussion, and for predicting how preferences might change from one time interval to the next. This section examines three aspects of preferences in design teams: design preference extraction, projecting an aggregation and understanding preference evolution over the life of a project, and presents a case example to illustrate its approaches. For the sake of compactness, the probabilistic approached in this report is named with **PPT** (Preferential Probabilities from Transcripts).

Figure 1 gives a flowchart for using PPT in studying design team discussion. The general process centers around a design team as they discuss possible design choices and make trade-offs between the choices. The design discussion is audio recorded, transcribed, and time tagged. Text analysis techniques are used to collect the design specific information, which is called *utterance data* in this research, from the transcripts. Utterance data is converted to preference data with the employment of two models – Preference Transition Model and Utterance-Preference Model. Initially both the models and the preference data are unknown, EM (Expectation Maximum) algorithm [1] is applied to seek the parameters of the two models. Finally, the evolution of design preferences is represented graphically, illustrated by preference strengths at different time intervals. In order to quantitatively evaluate the preferential probabilities extracted from the transcripts, questionnaires or surveys are used to elicit preferences from individual designers, which are then aggregated and converted to group preferential probabilities which are comparable with those from the transcripts. The details for extracting preferential probabilities from the surveys can be referred in another technical report [2].

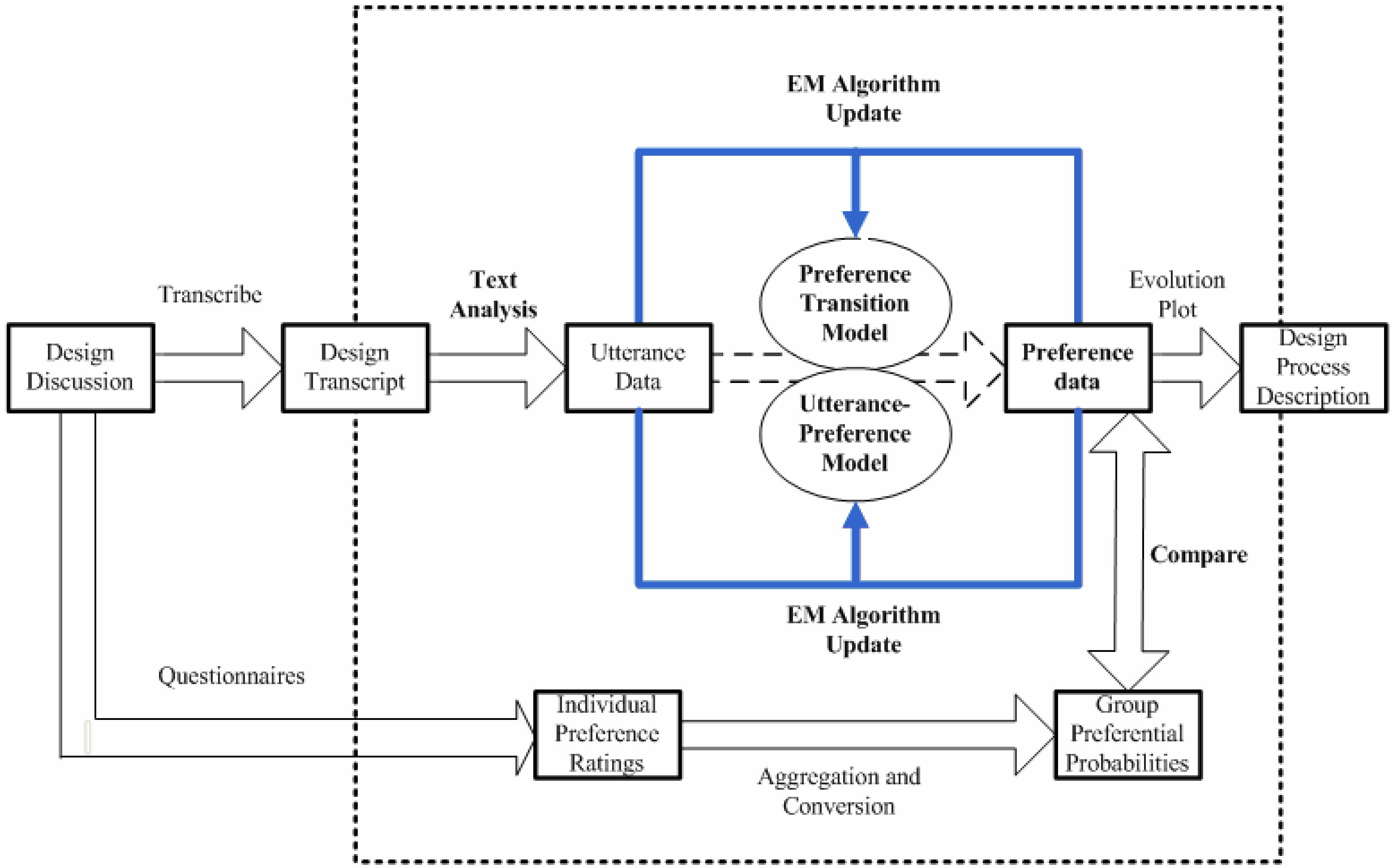

Figure 1 Flowchart of the Research Approach

Figure 1 shows that the preference data are acquired from the utterance data of the transcript. But the linkage between them is not as easy as the other steps in Figure 1 because the models and the parameters of the models linking the utterances and preferences are unknown. The basic procedures of this approach to resolve this problem are shown as follows:

(1) Collect word occurrences of all design alternatives in a transcript of a design team's discussion. The collection of word occurrences is called utterance data. In this step, variations of specific terms (synonyms) that represent the same alternative are also included as occurrences.

(2) Build a preference transition model to describe the relationship between preferences in two consecutive time intervals, along with an utterance-preference model to describe the relationship between what designers say and what designers prefer within the same time interval. The parameters of the two models are unknown (details in Sections 3 and 4).

(3) Assign reasonable initial values to the parameters of these two models.

(4) Apply both models to a transcript to predict preference data. The preference data will be used to describe the evolution of preference information over the design process (details in Section 5).

(5) Update the parameters of these two models using a traditional Expectation-Maximum (EM) algorithm [1] on the predicted preference data and the given utterance data (details in Section 6).

(6) Repeat steps 4 and 5 until there is convergence on the hidden parameters of the models. Parameters converge because the EM algorithm is guaranteed to improve the probability of the occurrences of the utterance data at each iteration [3].

1 Assumptions

In this section, five assumptions are made with regard to the preferential probabilities extracted from group transcripts.

Assumption 1: Embedded in group discussion is information sufficient to reflect group preference. During the design process, what designers say to each other generally corresponds with what they think. This is also an implicit assumption of protocol studies of designers .

Assumption 2: All major design alternatives for a concept selection problem are largely known a priori. While this may not be true for novel design problems, it is a reasonable assumption for incremental or re-design problems in which many design alternatives under consideration are ones that have been examined in the past.

Assumption 3: An entire discussion can be divided into time intervals during which the designers' preference are assumed to be unchanged. A change in preference can only occur between consecutive time intervals. This assumption helps to divide the whole design process into intervals, in order to describe the preference change of the design team during the design process. The ways of division are described in Section 7.

Assumption 4: what is most preferred in a time interval is related to what was most preferred in the previous time interval. This relationship can be represented probabilistically, and describes how likely the design team would change the “most preferred” alternative.

Assumption 5: Designers tend to talk positively more about the design alternative they prefer more and negatively about the design alternative they prefer less. Within the same time interval, how often an alternative is mentioned positively or negatively is linked to how much it is preferred, and this can be represented probabilistically. This assumption is an extension of Assumption 1. Probabilities are used here because of the stochastic uncertainty in the group discussion. It is possible for designers to mention the less-preferred alternative positively interlacing with the negative meaning, or mention the most-preferred alternative negatively interlacing with the positive meaning. This probabilistic relation considers that occasionally people do not mean what they say [4], but people still tend to speak out what they really mean. Probabilistic models describe how likely the team talks about the alternative they prefer the most.

2 Notation

The following explains some symbols which are used in the later mathematic formulations.

N: total number of alternatives in the studied design selection problem

T: total number of time intervals over the whole design process

i, j: the index to represent different time intervals in the design process

m, n, k: the index to represent different alternatives in the studied design selection problem

r: the index to represent the different iterations of the calculation process

a_m: the mth alternative of of the studied design selection problem in the design process

A: the set of all alternatives, i.e. A={ $a_1, a_2, \cdots, a_N$ }

π_i: the alternative which designers prefer to all other alternatives in Time Interval i, i.e., the most-preferred alternative in Time Interval i

ε_i: the alternative which designers utter at sometime during Time Interval i of the design process

σ_i: the sequence of the utterances of design alternatives in Time Interval i. e.g., if in Time Interval 2, the design alternatives are uttered as $a_2, a_2, a_1, a_1, a_1, a_2, a_3, a_1, a_1, a_3$ in the designers' transcribed discussion , then σ_2 = { $a_2, a_2, a_1, a_1, a_1, a_2, a_3, a_1, a_1, a_3$ }

$P(\pi_i = a_m)$: the probability that designers prefer Alternative a_m to all other alternatives (i.e., Alternative a_m is most preferred) in Time Interval i. If the preference value of Alternative a_m in Time Interval i is represented by $\mu_i(a_m)$ on a scale from zero to one, then $\pi_i = a_m$ is equivalent to $\mu_i(a_m) \geq \mu_i(a_n)\ for\ all\ 1 \leq n \leq N$

$P(\pi_i = a_m \mid \pi_j = a_n)$: the probability that designers prefer Alternative a_m to all other alternatives in Time Interval i, given that designers prefer Alternative a_n to all other alternatives in Time Interval j

$P(\varepsilon_i = a_m)$: the probability that Alternative a_m is uttered in Time Interval i

$P(\varepsilon_i = a_m \mid \pi_j = a_n)$: the probability that Alternative a_m is uttered in Time Interval i given the condition that Alternative a_n is preferred the most in Time Interval j

3 Preference Transition Model

This model relates the design team's preference in the current time interval to that in the next time interval. Individual designers may have different preferences, but in this model, only the accumulative group preferences are considered in a probabilistic relation.

In one time interval, it is assumed that there is an alternative which the team prefers the most, called the most-preferred alternative, and the remaining alternatives are the less-preferred alternatives. Each alternative has a probability to be the most-preferred alternative and the less-preferred one. The probability of one alternative to be most-preferred is the preferential probability of this alternative, and it describes the likelihood that a team prefers this alternative over all others.

The Preference Transition Model is the mathematical implementation of Assumption 4. At each transition, the design team can either 1) keep the most-preferred alternative unchanged from the previous time interval or 2) change from the most-preferred alternative to another alternative. The transition relationship between one interval and the next depends on the preference strengths of the most-preferred alternative and the less-preferred alternatives. In this study, a preliminary relationship is approximated in which all less-preferred alternatives in the current time interval are equally likely to become the most-preferred alternative in the next time interval.

The preferential probability for one alternative in the next time interval is cumulative, and relates both the probability this alternative is most-preferred in two consecutive intervals and the probability it transitions from less-preferred to most-preferred. Two alternatives with different preference strengths have different probabilities to be most-preferred and less-preferred, so the accumulated preferential probabilities in the next time interval may differ even with this preliminary Preference Transition Model.

In mathematical terms, the model can be expressed as in Equation (1)

$$P(\pi_{i+1} = a_n \mid \pi_i = a_m) = \begin{cases} p & when\ n = m \\ \dfrac{1-p}{N-1} & when\ n \neq m \end{cases} \tag{1}$$

where $0 \leq p \leq 1$ is an hidden parameter, which means the probability that the most-preferred alternative is kept unchanged from one time interval to the next consecutive one. The bigger p is, the more consistent the preferences are over the design process; and the smaller p is, the more frequently the preferences are changed.

4 Utterance-Preference Model

This model relates the team's preference to the utterances of the alternatives in the same time interval. In other words, it tries to approximate what designers think with what designers say. This model is the mathematical implementation of Assumption 5. In this model, a Vygotskian view [5] is adopted. This view differs between two lines of speech, including inner speech and the external speech. The inner speech reflects one's meditation and can be regarded as a self-discussion process, while external speech is used for social communications with others. Regarding the study of relationship between psychological language and thought, Vygotysky [5] said, "the area of inner speech is one of the most difficult to investigate". In engineering design, protocol analysis is widely used to investigate the inner speech for studying design activities and design decision making process. In this study, both the inner speech from individual "think aloud" and the outer speech from team conversation are considered. In the collection of verbal report data [6], it is assumed that not all thoughts which pass through attention are verbalized and some thoughts may be verbalized in variety of ways. Therefore, in the design selection process, the concept regarding design alternatives may not be uttered in deterministic patterns. Designers may utter the most-preferred alternatives in negative ways and the less-preferred alternatives in positive ways due to the possible uncertainty in the process, including the situations that designers do not know what they say, that designers do not have deterministic preferences, that designers express their thoughts wrongly, and that team dynamics influence the team work differently. Therefore, a probabilistic model can be applied to explain this random relationship.

A design alternative can be uttered in a transcript in either a positive or negative sense. When an utterance of an alternative has no negative words (e.g. “no,” “not,” “hardly”) nearby in the transcript, this utterance is regarded as positive, otherwise it is negative. The negative utterance of an alternative is counted as a positive utterance for all the other alternatives. Another strategy would be to subtract the negative utterance from the count of positive utterances, but this could lead to negative sums. Since the model is probabilistic, it inherently considers the cases when an alternative is mentioned that is not preferred the most. The model establishes the general pattern of how often a design team mentions the alternative they prefer the most and how often they mention the less-preferred alternatives. Similarly, a preliminary model is approximated in which less-preferred alternatives are equally likely to be uttered by designers in the same time interval.

The probability for one alternative is uttered is also a cumulative probability. It relates both the probability this alternative is uttered when it is most-preferred and the probability it is uttered while it is not most-preferred. For two alternatives with different preference strengths, they have different probabilities to be most-preferred and less-preferred, so the accumulated probabilities to be uttered may differ even with this preliminary Utterance-Preference Model. In equation form:

$$P(\varepsilon_i = a_n \mid \pi_i = a_m) = \begin{cases} q & when\ n = m \\ \dfrac{1-q}{N-1} & when\ n \neq m \end{cases} \tag{2}$$

where $0 \leq q \leq 1$ is an hidden parameter, which means the probability that the most-preferred alternative is uttered in the discussion. In protocol studies of designers [7], it is assumed that what designers say generally corresponds with what they think. In this study, it is assumed that designers say what they prefer in most cases. i.e., $q > \frac{1-q}{N-1}$. The reasonable value range for q is $\frac{1}{N} < q \leq 1$.

5 Preference Calculation

The preferences of the design alternatives in the design process may change over time during the discussion of the design team. One of the objectives in this research is to extract the preference evolution over the whole process. Although the preference value of each alternative cannot be acquired from the method proposed in this research, the probability of each design alternative to be most-preferred can be calculated.

The challenge: given the utterance data about the design alternatives and the preferential probabilities of the design alternatives in the current time interval, what are the preferential probabilities of the design alternatives in the next time interval?

By Law of total probability,

$$P(\pi_i = a_k \mid \sigma_i, \sigma_{i-1}, \sigma_{i-2}, ..., \sigma_1) =$$
$$\sum_{1 \le m \le N} P(\pi_i = a_k \mid \sigma_i, \sigma_{i-1}, ..., \sigma_1,\ \pi_{i-1} = a_m) P(\pi_{i-1} = a_m \mid \sigma_{i-1}, \sigma_{i-2}, ..., \sigma_1) \quad (3)$$

By Bayes Theorem,

$$P(\pi_i = a_k \mid \sigma_i, \sigma_{i-1}, ..., \sigma_1,\ \pi_{i-1} = a_m) =$$
$$\frac{P(\sigma_i \mid \pi_i = a_K, \sigma_{i-1}, ..., \sigma_1, \pi_{i-1} = a_m) P(\pi_i = a_K \mid \sigma_{i-1}, ..., \sigma_1, \pi_{i-1} = a_m)}{\sum_{1 \le n \le N} P(\sigma_i \mid \pi_i = a_n, \sigma_{i-1}, ..., \sigma_1,\ \pi_{i-1} = a_m) P(\pi_i = a_n \mid \sigma_{i-1}, ..., \sigma_1, \pi_{i-1} = a_m)} \quad (4)$$

Equation (4) can be simplified into Equation (5) because the utterance data in the current time interval are independent of the utterances in the historical time intervals while given the preference in the current time interval, and the preference in the current time interval is independent of the utterance data in the historical time intervals while given the preference in the latest previous time interval.

$$P(\pi_i = a_k \mid \sigma_i, \sigma_{i-1}, \ldots, \sigma_1, \ \pi_{i-1} = a_m) =$$

$$\frac{P(\sigma_i \mid \pi_i = a_k) P(\pi_i = a_k \mid \pi_{i-1} = a_m)}{\sum_{1 \le n \le N} P(\sigma_i \mid \pi_i = a_n) P(\pi_i = a_n \mid \pi_{i-1} = a_m)} \tag{5}$$

Substituting Equation (5) back into Equation (3) gives the following two equations ((6) and (7)).

When $i \ge 2$,

$$P(\pi_i = a_k \mid \sigma_i, \sigma_{i-1}, \sigma_{i-2}, \ldots, \sigma_1) =$$

$$\sum_{1 \le m \le N} \frac{P(\sigma_i \mid \pi_i = a_k) P(\pi_i = a_k \mid \pi_{i-1} = a_m)}{\sum_{1 \le n \le N} P(\sigma_i \mid \pi_i = a_n) P(\pi_i = a_n \mid \pi_{i-1} = a_m)} P(\pi_{i-1} = a_m \mid \sigma_{i-1}, \sigma_{i-2}, \ldots, \sigma_1) \tag{6}$$

When $i = 1$,

$$P(\pi_1 = a_k \mid \sigma_1) =$$

$$\sum_{1 \le m \le N} \frac{P(\sigma_1 \mid \pi_1 = a_k) P(\pi_1 = a_k \mid \pi_0 = a_m)}{\sum_{1 \le n \le N} P(\sigma_1 \mid \pi_1 = a_n) P(\pi_i = a_n \mid \pi_0 = a_m)} P(\pi_0 = a_m) \tag{7}$$

Suppose the design alternatives are uttered w_i times in the ith time interval, as $a_i^{(1)}, a_i^{(2)}, a_i^{(3)}, \ldots, a_i^{(w_i)}$, which are all in Alternative Set A. Assume that the utterances of alternatives in one time interval strongly depend on the designers' preference, and the utterances of design alternatives are indifferent of each other given the strong dependence on preferences, then

$$P(\sigma_i \mid \pi_i = a_k) = \prod_{u=1}^{w_i} P(\varepsilon_i = a_i^{(u)} \mid \pi_i = a_k) \tag{8}$$

Equations (6), (7) and (8) recursively calculate the preference in the next time interval from the preference in the current time interval. In order to make the recursion work, two pieces of information should be given.

- The initial preferential probabilities of all alternatives before the first time interval.
- The parameters of Preference Transition Model and Utterance-Preference Model. i.e. Parameters *p* and *q*.

The first one can be resolved by several ways:

(1) Conducting surveys of designers before the start of the design process;

(2) Collecting preference information from an earlier design process;

(3) Analyzing preferences from the design of similar products;

(4) Establishing an unbiased starting point which assumes a uniform alternative distribution.

In the case example in this study, because of the unknown initial preferences, all alternatives are initiated with uniform alternative distribution, which gives an unbiased starting point. i.e.,

$$P(\pi_0 = a_k) = \frac{1}{N},\ k = 1,2,...,N \tag{9}$$

Equation (9) means that before the discussion of the design process, all alternatives have equal probability of being most preferred.

The parameters of two key models are not as easy to acquire as the initial preferential probabilities of alternatives. In this research, an EM (Expectation-Maximization) algorithm [1] is applied to searching the parameters of Preference Transition Model and Utterance-Preference Model.

6 Estimation of Hidden Parameters

If the Preference Transition Model and the Utterance-Preference Model are given, starting from the initial preferential probabilities at the beginning of the design discussion, it is feasible to calculate the preferential probabilities for each design alternative in each time interval and then to plot the preference evolution over the whole design process. But the problem is that initially the parameters of these two models are unknown. In this situation, utterance data are observable but preference data are unobservable, and the models are incomplete because of the hidden parameters. An EM (Expectation-Maximization) algorithm [1] is often used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved hidden variables. In this study, it can be applied to seek the values of the two hidden parameters of the two models.

An EM algorithm has two steps, the E-step and the M-step. The E-step estimates the unobservable data. It can be accomplished by Equations (6) and (7). The M-Step computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found on the E-step. In this study, it corresponds to estimating the values of p and q which make the utterance sample of the design alternatives occur in the discussion with the maximal likelihood.

From Equation (1), it is known that $P(\pi_{i+1} = a_k \mid \pi_i = a_k)$ is independent of *k* and *i*. It means that no matter what time interval it is in, no matter which alternative designers prefer the most in the previous time interval, designers have a fixed probability to keep the most-preferred alternative unchanged.

By the maximum likelihood [8], $P(\pi_{i+1} = a_k \mid \pi_i = a_k)$ can be estimated as

$$P(\pi_{i+1} = a_k \mid \pi_i = a_k) = \frac{\sum_{1 \le i \le T-1} \sum_{1 \le m \le N} C(\pi_{i+1} = a_m, \pi_i = a_m)}{\sum_{1 \le i \le T-1} \sum_{1 \le n \le N} \sum_{1 \le m \le N} C(\pi_{i+1} = a_n, \pi_i = a_m)} \tag{10}$$

where $C(\pi_{i+1} = a_n, \pi_i = a_m)$ is a fractional count that counts the cases that a_n is most-preferred in the current time interval while a_m is most-preferred in the previous time interval. $C(\pi_{i+1} = a_n, \pi_i = a_m)$ can be calculated as follows.

$$C(\pi_{i+1} = a_n, \pi_i = a_m) = P(\pi_{i+1} = a_n \mid \sigma_{i+1}, \sigma_i, ..., \sigma_1) P(\pi_i = a_m \mid \sigma_i, \sigma_{i-1}, ..., \sigma_1) \tag{11}$$

The fractional counts are fractional numbers, and they are not the same as the normal counting numbers, which are integers. But the values of fractional counts have the proportional relationship with the integral numbers which count the cases when $\pi_{i+1} = a_n$ and $\pi_i = a_m$, so fractional counts can be used in Equation (10) to estimate of Parameter p.

Similarly, from Equation (2), $P(\varepsilon_i = a_k \mid \pi_i = a_k)$ is independent of *i* and *k*. It means that in a certain time interval, designers have a fixed probability to utter the same alternative as the one they prefer the most.

By the maximum likelihood, $P(\varepsilon_i = a_k \mid \pi_i = a_k)$ can be estimated as

$$P(\varepsilon_i = a_k \mid \pi_i = a_k) = \frac{\sum_{1 \le i \le T} \sum_{1 \le m \le N} C(\varepsilon_i = a_m, \pi_i = a_m)}{\sum_{1 \le i \le T} \sum_{1 \le n \le N} \sum_{1 \le m \le N} C(\varepsilon_i = a_n, \pi_i = a_m)} \tag{12}$$

where $C(\varepsilon_i = a_n, \pi_i = a_m)$ is also a fractional count, which counts the number of cases that a_n is uttered while a_m is most-preferred in the same time interval. It can be calculated as follows.

$$C(\varepsilon_i = a_n, \pi_i = a_m) = C(\varepsilon_i = a_n) P(\pi_i = a_m \mid \sigma_i, \sigma_{i-1}, ..., \sigma_1) \tag{13}$$

where $C(\varepsilon_i = a_n)$ is the number of utterances of Alternative a_n in the time interval i.

Equations (12) and (13) calculate *q* based on the samples of alternative utterances and preferences. When using the above procedure to calculate q, it should be noted that the value of q should be more than 1/N.

Because the EM algorithm is guaranteed to improve the probability of the sample of alternative occurrences at each iteration, *p* and *q* will converge to values which try to maximize this probability [3]. These converged values can be regarded as the parameters for Preference Transition Model and Utterance-Preference Model. The shortcoming of EM algorithm is that it may converge to a local optimum. Multiple initial estimates can be used to avoid being trapped in a local optimum. Simulated annealing can be combine with EM algorithm to overcome the local optima problem [9, 10].

7 Time Intervals

In this research, it is assumed that designers do not change their preferences on design alternatives in one time interval. Preferences can only be changed at the transitions between time intervals. Based on the designers' transcribed discussion, there are several ways to determine time intervals.

1. Collect all transitional words (e.g., "but", "however," "while") in the transcript, and divide the transcript into varying time intervals with these transition words.

2. Collect all key design alternative occurrences from the transcript of designers' discussion, and time-stamp all collected words. Big time gaps between the key alternative words mark the separations of time intervals.

3. Set up a fixed word frequency. The process is divided into time intervals in which there are equal numbers of word utterances of key alternatives.

4. Make a fixed time interval. The length of each interval depends on the desired granularity of preference evolution.

Although Methods 3 and 4 are not as accurate as Methods 1 and 2, they are more direct to implement. In this research, Method 4 is chosen and modified to specify the time intervals in the design process. The time intervals are nearly of the same lengths but not exactly equal because the divisions occurred only after one finished his/her conversations and there was no immediate following-ups. If another designer was ready to talk while one was still talking, divisions of time intervals would wait until both finished. Even in this way, the real preference of the team may change inside the interval as well. For the sake of simplicity, it is assumed that designers do not change their preferences for design alternatives within a time interval in this study. The preferences are considered accumulatively for each interval and the preference changes are only considered between the intervals. The precise granularity of the changes inside the time interval could be studied in future research.

8 Case Study

8.1 Case Background

The case used for PPT in this report is a real-world design team working on the design of a large-scale space system architecture. This design team was composed of 17 experienced scientists and engineers of different disciplines working together in a co-located, highly concurrent setting. The team had collaborated on several similar projects in the past. The project took place over three 3-hour sessions spaced out over several weeks. This research focuses on the audio-recorded utterances of one member of the team as he explained his design decision-making process in detail to a novice member of his team. This recording was transcribed into a text document of approximately 28,000 words. All data was time coded. In the transcript, the primary team member talked nearly 85% of the time, and four other members made up the remainder. The case study in this section is focused on the first session of the second component selection problem.

8.2 Data Collection and Method Implementation

The transcript in the first session was input as the raw data in this case study. The utterances of the three alternatives (represented by a_1, a_2, and a_3) for the second selection problem were collected in intervals of 10 minutes, as shown in Table 1.

Table 1 Sample Data: Utterances of Alternatives

Interval \ Alternative	a_1	a_2	a_3
1	9	0	3
2	9	2	7
3	1	0	0
4	4	0	9
5	0	0	6
6	3	1	0
7	1	0	2
8	0	1	2
9	1	0	8
10	0	0	1
11	0	1	2
12	0	0	5

Initially, we can give any values to p and q if $0<p<1$ and $1/3<q<1$ are met. And the values will be updated in the later iterations. To distinguish p and q in different iterations, let p_r, q_r be the variables of p and q in the r^{th} iteration. In this example, initial values are randomly chosen as $p_1=0.5$ and $q_1=0.4$. And the initial preferential probabilities of the alternatives all equal 1/3:

$$P(\pi_0 = a_1) = P(\pi_0 = a_2) = P(\pi_0 = a_3) = \frac{1}{3} \tag{14}$$

In the first time interval of the transcript, there were no utterances of a_2, while 9 times of a_1, and 3 times of a_3. By Equation (8)

$$P(\sigma_1 \mid \pi_1 = a_1) = (\frac{1-q_1}{2})^0 (q_1)^9 (\frac{1-q_1}{2})^3 = 7.0779\text{-}06 \tag{15}$$

$$P(\sigma_1 \mid \pi_1 = a_2) = (q_1)^0 (\frac{1-q_1}{2})^9 (\frac{1-q_1}{2})^3 = 5.3144\text{-}07 \tag{16}$$

$$P(\sigma_1 \mid \pi_1 = a_3) = (\frac{1-q_1}{2})^0 (\frac{1-q_1}{2})^9 (q_1)^3 = 1.2597\text{-}06 \tag{17}$$

These three values are not the normalized probabilities. Every value only has meanings when comparing with each other.

Substituting values from Equations (15), (16) and (17) into Equation (7) gives

$P(\pi_1 = a_1 \mid \sigma_1)$ =0.7798

$P(\pi_1 = a_2 \mid \sigma_1)$ =0. 0663

$P(\pi_1 = a_3 \mid \sigma_1)$ =0. 1539

The preferential probabilities in later time intervals can be calculated recursively by Equations (6) and (8). The probability values in different time intervals are listed in Table 2, in the first iteration with p_1=0.5 and q_1=0.4.

Table 2 Preferential Probabilities of Design Alternatives (the First Iteration)

Interval \ Alternative	a_1	a_2	a_3
1	0.7798	0.0663	0.1539
2	0.6829	0.0574	0.2597
3	0.4878	0.2328	0.2795
4	0.2179	0.0579	0.7242
5	0.112	0.0934	0.7946
6	0.4419	0.2481	0.3099
7	0.3514	0.2309	0.4177
8	0.2508	0.2993	0.4498
9	0.1058	0.0835	0.8107
10	0.2418	0.2367	0.5215
11	0.2267	0.2971	0.4763
12	0.1497	0.1595	0.6909

Parameters p_r and q_r would be updated by Equations (10) and (12) with the fractional counts calculated from Equations (11) and (13) in the previous iteration.

After updating the parameters, the new preferential probabilities of alternatives are re-calculated according to Equations (6), (7) and (8). The above procure is iterated until converged.

8.3 Results

Table 3 shows the iterative results of p_r and q_r. It shows that p_r converges to 0.716 after 5 iterations, while q_r converges to 0.672 after 7 iterations. In the experiment, several initial estimates for p and q are tried, and all of them are converged to the same values.

Table 3 Iterative Values of Parameters

Parameter / Iteration	p_r	q_r
1	0.4	0.5
2	0.564	0.418
3	0.694	0.546
4	0.714	0.644
5	0.716	0.668
6	0.716	0.671
7	0.716	0.672
8	0.716	0.672

Figure 2 shows the preferential probability evolution of three alternatives when the converged parameters are applied to the Preference Transition Model and the Utterance-Preference Model. The solid line with square dots stands for the evolution of probabilities that Alternative a_1 is most preferred, the broken line with circle dots stands for the evolution of probabilities that Alternative a_2 is most preferred, and the dotted line with triangle dots stands for the evolution of probabilities that Alternative a_3 is most preferred. In terms of preferences, Figure 2 suggests that Alternative a_1 and a_3 dominate and that these two alternatives alternate with each other during the design process. This conjecture from the chart is validated by a qualitative reading of the original transcript.

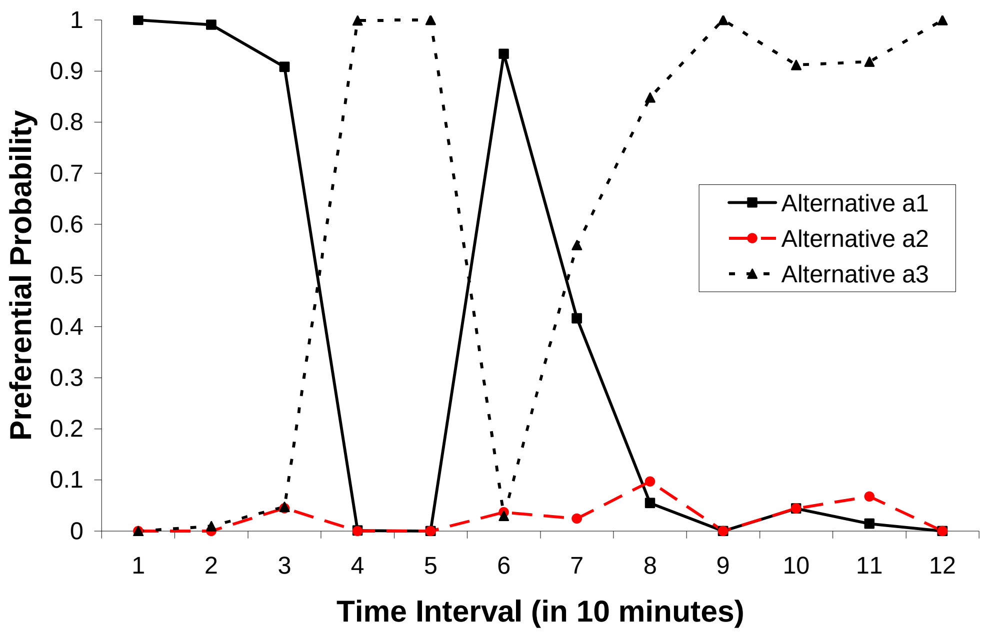

Figure 2 Design Process Evolution: Preferential Probabilities of the Three Alternatives

Design sensitivity. Since the converged values of p and q are estimated with EM algorithm, they might vary from the true values. Let $\hat{p}, \hat{q}$ be the converged values of p and q, and suppose p, q are in the range with $\hat{p}, \hat{q}$ shifting 10%. The evolutions of the preferential probabilities for the cases when p or/and q are underestimated or overestimated are plotted as shown in Figures 3-5. The gaps between the uppermost lines and the lowermost lines give the true value ranges that alternatives are most preferred. In this example, qualitatively speaking, the true values of p and q are in close ranges of the converged p and q. Therefore, the preferential probabilities of design alternatives calculated based on $\hat{p}$, $\hat{q}$ can approximately describe the preferences over the design process.

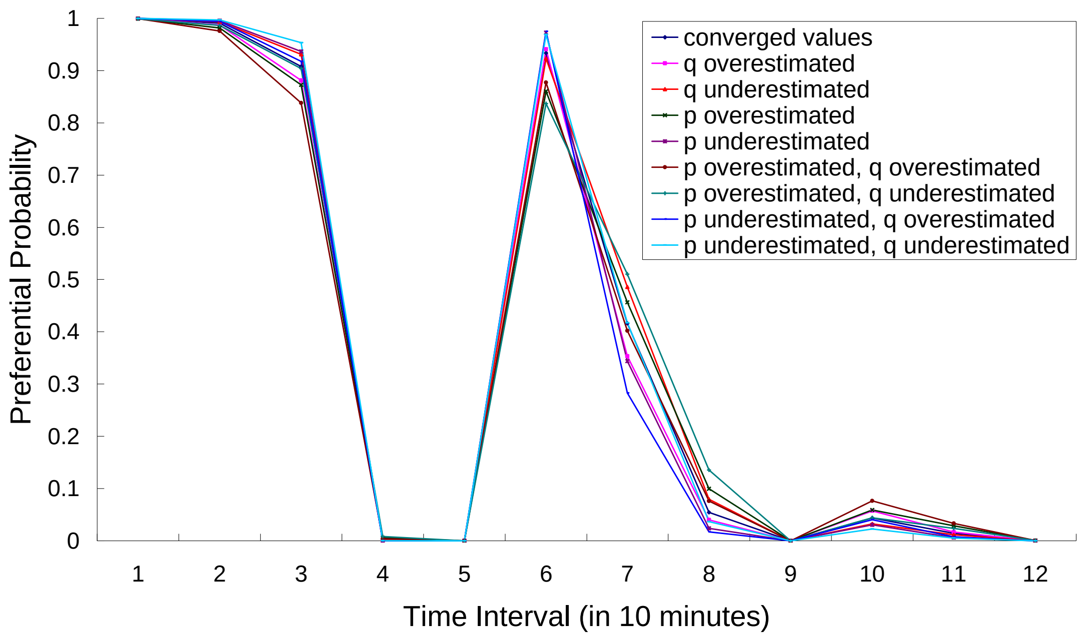

Figure 3 Preferential Probability Ranges of Alternatives a_1

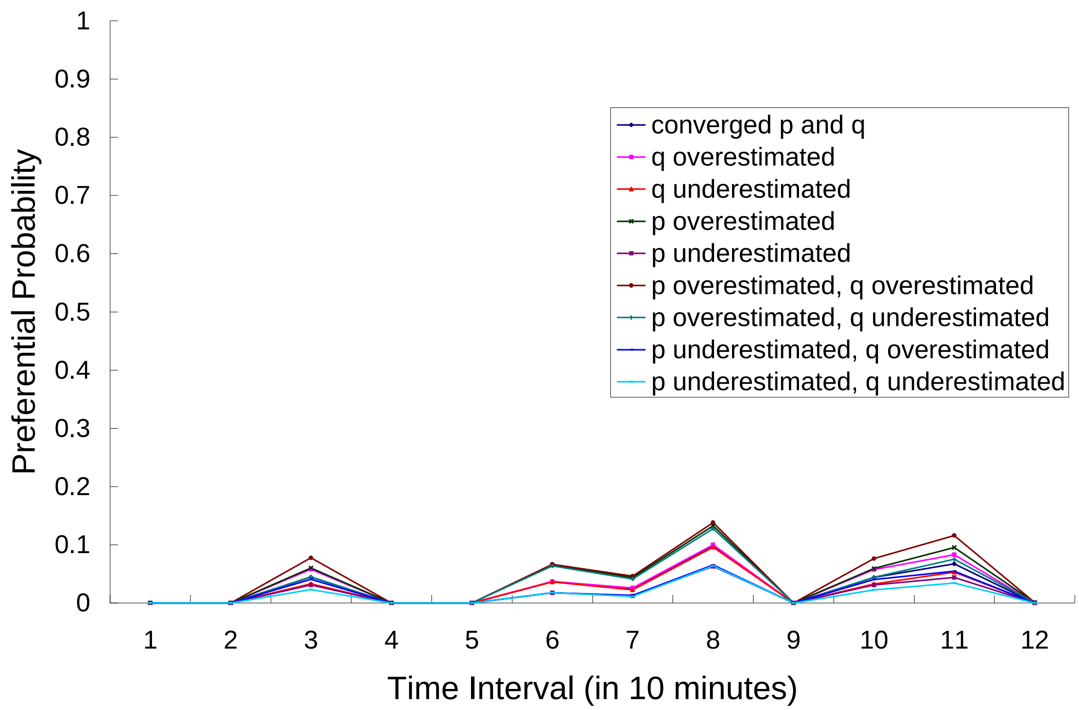

Figure 4 Preferential Probability Ranges of Alternatives a_2

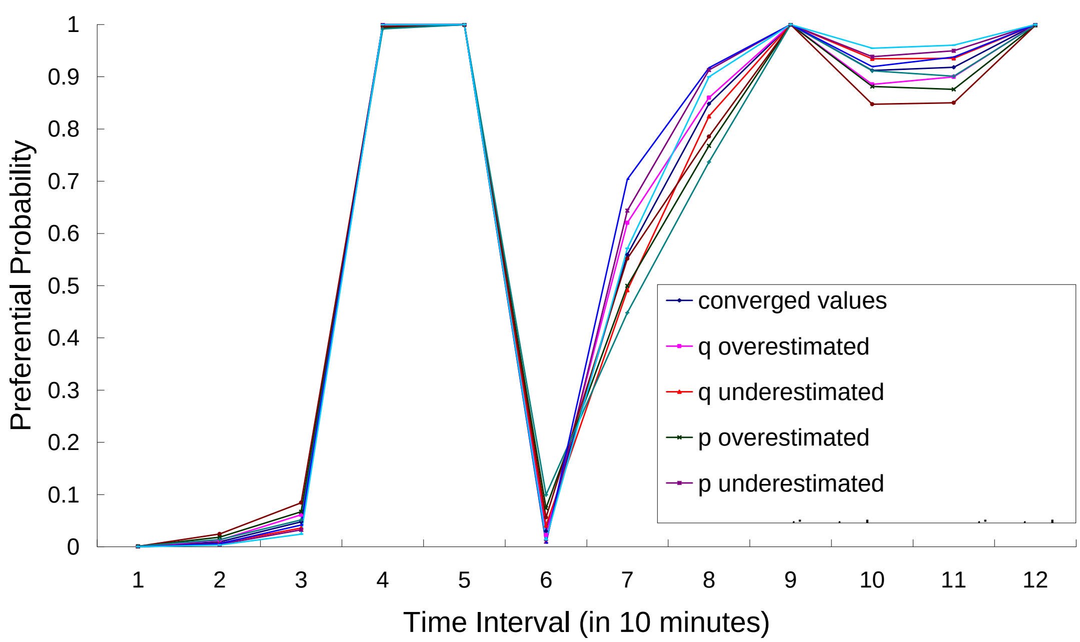

Figure 5 Preferential Probability Ranges of Alternatives a_3

9 Remarks and Discussions

The methodology presented suggests a probabilistic way to describe the preference information in the design process and model the design selection process through the evolution of preferential probabilities. The results of applying the methodology to the case study are consistent with the qualitative reading of the design transcript. It was expected that design alternative choices would oscillate in a large-scale system design problem, and this was true for the component selection problem which was chosen for the case study in this section.

This methodology can also work as a way to implicitly extract the preference information which may be used in the further engineering decision making process. Comparing with the traditional preference elicitation method, this method is simply based on what designers say in the design discussion. It does not depend on the questionnaires or surveys of all team designers after finishing the design project.

This study models the preference evolution using probabilistic approaches. The methodology in this section shows the main idea of modeling the relationship between the preferences of two consecutive time intervals and the relationship between the utterances and the preferences. The work developed in this report may lead to a novel way to understand the evolving nature of a team's preferences over the life of a project.

References

[1] Dempster, A., N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, Series B, 1977. **39**(1): p. 1-38.

[2] Ji, H., M.C. Yang, and T. Honda, *Example for Implementing PPS: Preferential Probabilities Translated from Survey under Principle of Maximum Entropy*. Technical Report No. TR-2008B. 2008, Massachusetts Institute of Technology: Cambridge, MA.

[3] Bilmes, J.A., *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report*. 1998, International Computer Science Institute: Berkeley, CA, USA.

[4] Bertrand, M. and S. Mullainathan, *Do People Mean What They Say? Implications for Subjective Survey Data.* The American Economic Review, 2001. **91**(2): p. 67-72.

[5] Vygotsky, L., *Thought and Speech*. 1986, Cambridge, MA: MIT Press.

[6] Ericsson, K.A. and H.A. Simon, *Protocol Analysis: Verbal Reports as Data*. 1993, Cambridge, MA: MIT Press.

[7] Cross, N., H. Christiaans, and K. Dorst, *Analysing Design Activity*. 1996, Chichester: Wiley.

[8] Fisher, R.A., *On the Mathematical Foundations of Theoretical Statistics.* Philosophical Transactions of the Royal Society, 1922. **222**: p. 309-368.

[9] Ueda, N., et al., *SMEM Algorithm for Mixture Models.* Neural Computation, 2000. **12**(9): p. 2109-2128.

[10] Ueda, N. and R. Nakano, *Deterministic annealing EM algorithm.* Neural Networks, 1998. **11**(2): p. 271-282.