Table 1: Weights used to measure undifferentiated, low-only, and differentiated errors

	Undiff.				Low-only				Diff.			
	N	S	SF	SG	N	S	SF	SG	N	S	SF	SG
N	0	1	1	1	0	1	1	1	0	1	1	1
S	1	0	1	1	1	0	0	0	1	0	0.5	0.5
SF	1	1	0	1	1	0	0	0	1	0.5	0	0.5
SG	1	1	1	0	1	0	0	0	1	0.5	0.5	0

Table 2: The weighted success rate for the proposed algorithm is greater than 79%, even for the undifferentiated scheme, where all errors are counted

Success rate (%)		
Undiff.	Low-only	Diff.
79.06%	87.31%	83.19%

4. EXPERIMENTAL EVALUATION

In order to evaluate the effectiveness of the segmentation techniques presented in this paper, we performed a user study and compared the segmentation feedback provided by assessors of a discussion board with the segmentation results obtained by the proposed algorithm.

Setup: For the evaluations presented here, due to the diversity of its postings and message hierarchies, we used the movie message board available at [2] as the message data source. We randomly selected

- 20 discussion threads, with
- a total of 368 messages,
- average thread depth of 12.45,
- average quotation depth of 1.3 (86% of the total of 5241 quotations are from the parent)

from this source and asked 5 users to assess each message to label it with N for new topic, S for same topic as the parent, SF for specialization (or focussing), or SG for generalization. Given all manual labelings from multiple assessors, we took the majority label to denote the page's relationship with its parent. We then compared these manual labeling results with the labels assigned by the proposed automated segmentation algorithm (which took only **560ms** to segment the given 20 threads). In this section, we report the results when the threshold for detecting *new* segment boundaries is set to 0.35, generalization threshold, Θ_g, is set to 0.6, and the specialization threshold, Θ_s, is set to 0.8 (we discuss the effects of varying these thresholds later in the section). Also, for the results presented here, the impact factor for the parent quotations ($d = 1$) is $imp(d) = \frac{1}{2}$ (we discuss the effect of different impact factor values in the later section).

Evaluation criteria: In order to observe the effectiveness of the proposed algorithms, we computed a labeling *success rate* (or precision),

$$success\_rate = \frac{\sum_{m \in messages} 1 - error\_weight(m)}{number\ of\ messages} \times 100,$$

where error weights are used to account for *gravity* of the error in the computed success rate. We experimented with three different schemes as shown in Table 1:

- *Undifferentiated weights:* Weights in first partition of the table mark all errors with the same (maximum) error weight, independent of the type of error.
- *Low-only weights:* Weights in second partition in the table only count errors in the first, low-granularity, step of the algorithm; i.e., only
  - those pages that are marked erroneously as being of a *new topic* or
  - those that should have been marked as a *new topic*, but not marked as such

  count towards the error rate.
- *Differentiated weights:* Weights in third partition in the table penalize different error types differently. More specifically, errors within the high-granularity group (S, SF, and SG) are marked half as costly as errors across the low-granularity segmentation.

Table 3: Distribution of various types of errors

Alg.\ **User**	**New**-u	**Same**-u	**Spec.**-u	**Gen.**-u	*Tot.*
New-a	–	31.0	1.4	7.0	*39.4*
Same-a	14.1	–	16.9	11.3	*42.3*
Spec.-a	0.0	4.2	–	0.0	*4.2*
Gen.-a	7.0	5.6	1.4	–	*14.1*
Total	*21.1*	*40.9*	*19.7*	*18.3*	*100*

Table 4: User labelings for the 368 messages in the randomly selected 20 threads

New	Same	Spec.	Gen.	No Majority (unlabeled)	Tot.
58	206	39	36	29	368

Table 2 shows the weighted success rates observed in the experiments.

Undifferentiated success rate: Based on the user study, we observed that the undifferentiated success rate, where all errors are penalized with the maximum weight without distinguishing between the types of errors, was around 79.06% (first column in Table 2).

Low-only success rate: On the other hand, when we focus on only the errors in the first, low-granularity, step of the algorithm, we observed that the success rate jumped to 87.31% (second column in Table 2).

Differentiated success rate: When a differential penalty scheme (where errors within the high-granularity group – S, SF, and SG – are marked half as costly as errors across the low-granularity segmentation between same and new topics) is used, the success rate was 83.19% (last column in Table 2).

Distribution of the errors: Table 3 provides a detailed tally of the types of errors (around 20% of all labelings as described above) observed during the user study. In this table, the columns correspond to the labelings chosen by the users, while the rows correspond to those assigned by the proposed algorithm.

As can be seen by studying the last row of the table, which shows the aggregate number of the errors made by the proposed algorithm for each user labeling, the greatest percentage (40.9%) of labeling errors is due to messages that are marked *same* by the users. In fact, the biggest single contributor to the number of errors is the set of *same topic* messages that are labeled as *new* by the algorithm (31% of all errors). In the last column of the table, which shows how the errors are distributed among labeling of the algorithm, we see that 42% of all errors are due to messages that are incorrectly marked *same*, whereas around 40% of the errors are due to those that are incorrectly marked as *new*. The total contribution of specialization and generalization errors to the overall rate of the error is less than 20%.