Table 5: Success rates for individual labelings

Success rate for labelings				
New	Same	Spec.	Gen.	*Overall*
0.74	0.86	0.64	0.64	*0.79*

Table 6: The impact of quotations on the labeling performance

	Success rate (%)		
Quot. weights	Undiff.	Low-only	Diff.
Off	72.57%	85.25%	78.90%
On	79.06%	87.31%	83.19%

Note that, since the distribution of labels provided by the users is not uniform (Table 4), the impact of different types of errors on the overall success rate varies. Table 5 presents success rates achieved by the proposed algorithm for each label. The success rate achieved for those messages labeled *new* by the users is around 74%. The success rate is as high as 86% for detecting messages that stay within the *same* topic. The fine granularity segmentation success rate in the second phase is around 64%. As can be seen from the **spec.**-u and **gen**-u columns in Table 3, most of the errors in the second phase of the algorithm are due to messages that are marked *same topic* by the algorithm but further classified into *specialization* and *generalization* categories by the users. This shows that, while the human assessors can differentiate fine topic distinctions better, the proposed algorithm may *conservatively* classify messages to be of the *same* topic to prevent over-segmentation. The overall (undifferentiated) success rate is close to 80%, as described earlier.

Effect of quotations: In order to observe the impact of the quotations on the performance of the segmentation algorithm, we calculated how the success rates changed when the context-sensitive weighting techniques proposed in this paper were turned off. When the quotations were not treated specially, the number of errors in the first step of the algorithm increased 11%, from 43 to 48 erroneous labelings. On the other hand, the total number of errors (including both phases of the algorithm) increased 30%, from 71 to 93, showing that especially the fine-granularity differentiation required in the second phase benefits significantly from the way the proposed algorithm uses quotations for context-sensitive weighting (Table 6).

In Table 3, we saw that 31% of the all errors were due to the set of *same topic* messages that were labeled as *new* by the algorithm. In order to see whether using a different impact factor formulation would improve this situation, we tried impact factors with different characteristics. A selection of the low-granularity (*same* versus *new*) labeling errors are reported in Table 7. The first row of this table corresponds to the results presented so far. The following rows shows the results obtained when the impact factors were set such that the resulting keyword vector would have a higher similarity to the ancestor from which the quotations have been taken. The results show that, indeed, the number of *same topic* errors drops when the impact of the keywords in the quotations increases. However, this is accompanied with a significant jump in the number of *new* messages that are labeled as *same*, reducing the overall success rate as shown in the last column of Table 7. In fact, between the two extremes (first and last rows) in the table, *new* message identification ($30 - 15 = 15$) is more sensitive to the weight of quotations than *same* message identification ($28 - 22 = 6$). Thus, overweighting quotations does not help the overall success rate.

Table 7: The effect of quotation impact factors on the low -granularity labeling performance

$imp(d)$ for $d = 1$	same $\xrightarrow{err}$ new	new $\xrightarrow{err}$ same	*undiff. succ.*
0.5	28	15	79.0%
1	23	18	77.6%
1.5	24	26	77.6%
2	22	30	76.4%

Table 8: Effects of different Θ_g and Θ_s thresholds

	0.0	**0.25**	**0.5**	**0.75**	**1.0**	*Exp.*
Undiff.	0.21	0.33	0.56	0.77	0.74	*0.79*
Low-only.	0.87	0.87	0.87	0.87	0.87	*0.87*
Diff.	0.51	0.57	0.69	0.82	0.77	*0.83*

The effect of threshold values: Finally, Table 8 shows the effect of various Θ_g and Θ_s values on the final success rate. As expected (since it is insensitive to the fine-granularity segmentation), the low-only success is independent of the values of Θ_g and Θ_s thresholds. Note that neither too small nor too large values are good for proper segmentation. As we mentioned earlier, threshold values need to be set through a *machine learning* process which identifies proper values based on a given training sample.

5. CONCLUSIONS

Message threads evolve with new postings as new messages may focus on or diverge from the original theme of the thread. In this paper, we presented algorithms for identifying how the hierarchical content of a discussion board grows through generalizations and specializations. This knowledge can be used in segmenting the message hierarchy into coherent units to facilitate indexing, retrieval, and ranking, as well as in guiding users through *segments* that are relevant for their navigational goals. The segmentation algorithms are being deployed in a software system, called iCare-Assistant, which aims at reducing the navigational load for blind students in accessing web-based electronic course materials through an unobtrusive, task-oriented, and individualized delivery interface. However, we note that the techniques are equally applicable for developing web summarization tools for users with sight.

6. ACKNOWLEDGMENTS

We would like to thank our students without sight, Kenneth Spector and David Paul, for their feedbacks and suggestions on the usability of iCare-Assistant; Terri Hedgpeth for her expertise in assistive technology; and Shibo Wu, Lina Peng, Yan Qi, and Ping Lin for their help in various phases of the implementation and evaluation of our work.

7. REFERENCES

[1] Blackboard. http://www.blackboard.com.

[2] Movie message board. http://www.hundland.com/movieboard.mv.

[3] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *26th annual ACM SIGIR Conference*, Toronto, Canada, July 2003.