Measures of Threaded Discussion Properties

Ricky J. Sethi, Lorenzo A. Rossi, Yolanda Gil

USC Information Sciences Institute

Abstract. In this paper, we present a set of measures to quantify certain properties of threaded discussions, which are ubiquitous in online learning platforms. In particular, we address how to measure the redundancy of posts, the compactness of topics, and the degree of hierarchy in sub-threads. This preliminary work would very much benefit from discussion and serves as a starting point for ultimately creating optimal structures of threaded discussions depending on the context.

1 Introduction and Motivation

Threaded Discussions are widely utilized in online learning platforms like eCollege, BlackBoard, and moodle. In addition, many online forums and Social Question & Answer sites also rely on similar formats. In fact, sites like `http://www.piazza.com`, which offer modified threaded discussions, are being used as pedagogical supplements. Although our ultimate goal is to be able to restructure threaded discussions into optimal formats depending upon the context, in this paper, we present early stage work that seeks to quantify the characteristics of threaded discussions as a first step.

1.1 Overview of Related Literature

There is a consistent literature on the automatic summarization of textual documents by means of natural language processing (NLP) tools, e.g. [5,4,3,1]. Some of the proposed approaches use automatic keyword detection to then find out different key topics within the document. Summaries are subsequently built by extracting the initial sentences associated with the portion of text related to those key topics.

A subset of this literature focuses on the automatic analysis and summarization of online single- and multi-threaded discussions.[1] The focus of the application ranges from online discussions about open source software (OSS forums) to discussions between students attending a certain class and their instructors.

In [8], an approach is proposed to summarize online discussions (from the Open Source Software forum in particular). Posts are first clustered according to topics. Then the posts belonging to each topic are further categorized into

[1] A single-threaded discussion is an online discussion where each post has at most one child post. In multi-threaded discussions, a post can have more than one child post, with multiple sub-threads possible.

two classes: 'problem' and 'advice'. Note that, for the purpose of our research, we can look at online single-threaded discussions as a special case of multi-threaded discussions.

[7] study the interactions among students and teachers in threaded discussions for distance education. The posts are classified according to a different set subclasses called speech acts. According to the statistics given in the paper, the majority of speech acts turn out to be either questions or answers. The remaining ones are elaborations of certain answers, acknowledgments, announcements, corrections or objections. The features used for the classification are N-gram sequences of the terms in the post (after a preliminary word filtering stage).

Various works propose metrics to express respectively relationships between posts and topics [6], relationship among contiguous posts [6], coherence of the threaded discussions [2].

2 Preliminary Approach

In this document, we propose a set of measures to quantify properties of threaded discussions (e.g. the quota of redundant posts). In the long term, we are investigating an approach to analyze and index threaded discussions from online learning platforms by means of machine learning and crowd-sourcing tools. The final goal is to automatically break down a certain threaded discussion and then be able to automatically re-build it in ways that enhance properties valuable to a certain target user and/or purpose: e.g. creating an automatic summary for an instructor who needs to quickly address a students' discussion, or a re-arrangement of the posts to the benefit of students who are participating in the discussion. At this preliminary stage of our research, we need to define potential optimal views of restructured threaded discussions. This will then help us to define the desired atomic elements in the structure of the discussions and consequently to design the machine learning/crowd-sourcing [2] algorithms to break down such discussions. To help with the definition of views, this paper proposes a set of measures to quantify some specific properties of threaded discussions.

2.1 Metrics

We can represent threaded discussions with tree type data structures. We propose the following metrics:

- Redundancy of Posts
- Topic Coherence (Compactness)
- Degree of Hierarchy of Sub-Threads

[2] For instance, we could design interfaces to require contributors to also label their own posts in a simple way before they submit it.

We believe that these quantities can be used as simple indicators of how *good* a structure of the threaded discussion for certain purposes is and therefore be useful for definition of desired views of re-structured threaded discussions. Note that these measures can be computed only over threaded discussions whose posts have been already analyzed and classified.

Let N be the number of posts p_i in a discussion, where $i = 1, \ldots, N$ and $t(p_i)$ indicates the date and time the post was submitted.

Definition 1. ***Duplicate Post:*** *Given two posts p_i and p_j where $t(p_j) > t(p_i)$, we say that p_j is a duplicate post of p_i when the content of p_j is so similar to the content of p_i that p_j could be removed without relevant loss of information for the readers.*

Redundancy Factor: *Given a certain discussion, where N_d is the total number of duplicate posts, we define the redundancy factor as $r = N_d/(N - N_d)$, where N is the total number of posts in the discussion.*

In online student forums or threaded discussions, we may have participants making posts that are duplicates of pre-existing posts in the discussion. Under some circumstances it may be desirable to remove duplicate posts and reduce redundancy. E.g., if we consider the original threaded discussion in Table 1, we can see its redundancy with 20 total posts and 2 duplicate posts of $r = 0.11$ vs a redundancy of $r = 0$ for the re-structured threaded discussion in Table 2. In addition, sometimes, retaining redundancy can be useful for instructors and users and may indicate popular topics or topic clouds. The automatic assessment of the redundancy of a post requires a high semantic analysis and therefore is a very challenging natural language processing task. In our future work, we will investigate ways to infer this metric, including using summarization methods like MEAD [3].

Definition 2. ***Topic Coherence:*** *Let's assume that the posts of a certain single-threaded discussion can be classified into a certain set of topics (or stances) s_j, $j = 1, \ldots, M$, where M is the total number of topics addressed in the discussion. Let N_j the number of posts on a topic s_j. We can map each post in the thread to a one dimensional space, where the location in the space is simply given by the number of parents of the post. Therefore we can measure how dispersed (or **compact**) a certain topic is in the discussion by measuring the standard deviation or spatial dispersion of posts when projected to a single dimension, or possibly multiple dimensions.*

The automatic estimation of coherence requires a preliminary classification of the topics addressed by the posts. A possible approach may consist of clustering the posts based on sets of keywords. This problem will also be a subject of our future investigations.

It may be useful in some cases to re-structure the discussion by aggregating posts belonging to the same topic, hence the need for defining a measure of compactness. Similarly, if we have a set of posts $p_k(t)$, $k = 1, ...N_j$ belonging to a certain topic s_j, we could introduce a measure of the distance between the

[3] `http://www.summarization.com/mead/`

chronological sequence of posts and the sorting of posts that is most effective for user understanding (measure of **chronological coherence**).

Definition 3. ***Degree of Sub-Thread Hierarchy:*** *We can define the Degree of Hierarchy of a sub-thread in terms of b (breadth) and d (depth) as $h = d/b$. Given M (number of proto-topics/stances, defined as each first-level sub-thread) and N_i (Number of posts in each topic/stance, $i = 1, \ldots, M$), a Flat sub-thread is when $d = 1$ and $b = N$ giving a degree of hierarchy of $h = 1/N$, whereas a Hierarchical sub-thread is when $d = max_i|N_i|$ and $b = M$, giving a degree of hierarchy of $h = \max_i |N_i|/M$.*

Thus, if we consider the original threaded discussion in Table 1, we can see that it has a degree of hierarchy of $h = 0.36$ vs a degree of hierarchy of $h = 0.80$ for the re-structured threaded discussion in Table 2.

3 Conclusion

In this paper, we have presented some initial attempts to quantify various characteristics of threaded discussions with the eventual goal of re-structuring threaded discussions into optimal structures depending on context. Although we have collected some threaded discussions from online classes (examples shown in Figure 1), this work is in a very early stage and we would welcome any comments and discussion.

References

1. J Chu-Carroll. Leveraging Wikipedia Characteristics for Search and Candidate Generation in Question Answering. *Twenty-Fifth AAAI Conference on Artificial*, pages 872–877, 2011.
2. Donghui Feng, Jihie Kim, Erin Shaw, and Eduard Hovy. Towards modeling threaded discussion using induced ontology knowledge. *AAAI*, 2006.
3. David Ferrucci, Eric Brown, Jennifer Chu-carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, and John Prager. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, pages 59–79, 2010.
4. N. Kwon and E.H. Hovy. Information acquisition using multiple classifications. In *Proceedings of the K-CAP07 conference*, 2007.
5. N. Kwon, S. Shulman, and E.H. Hovy. Multidimensional text analysis for erulemaking. In *Proceedings of the National Conference on Digital Government (dg.o 2006)*, 2006.
6. Chen Lin, Jiang-Ming Yang, Rui Cai, Xin-Jing Wang, Wei Wang, and Lei Zhang. Simultaneusly modeling semantics and structure of threaded discussions: A sparse coding approach and its applications. *SIGIR*, pages 131–138, 07 2009.
7. Sujith Ravi and Jihie Kim. Profiling student interactions in threaded discussions with speech act classifiers. In *Proceedings of the AI in Education Conference*, 2007.
8. L. Zhou and E.H. Hovy. On the summarization of dynamically introduced information: Online discussions and blogs. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

A Appendix of Figures and Tables

Post	Topic
Abdul: To the Bonus question: C++ design aspect is very limited...	T1
Brian: The syntax of Java is closely based on the syntax of the C++ programming language...	T2
Brian: Although there are many similarities I believe using the Java language is a bit easier.	T3
Abdul: There are so many differences and similarities, it sometimes is hard to mistake...	T2
Connie: When Sun Microsystems came out with Java, it like an answer to MOST of their prayers...	T3
Kerry: I prefer the one that's right for the job. For a high-level user-facing application...	T2
Kristine: three things that come to mind on what is different between Java & C++ are...	T3
Julius: I do find JAVA to be a bit less combersum when putting together your methods...	T3
Ken: Why might it be said that Java is an object-oriented language while C++ is a procedural...	T4
Bernard: I think that both have there place.	T2

- Jacob: ... With Java, the programming is more user friendly...
  - Luke: I love the analogy, Jacob...
    - Vu: Hi Luke–I had...
    - Jacob: ... I took the same classes...
      - Kristine: Java has both kinds...
  - Deborah: I don't think Java...
    - EXPERT: Sounds like Java has 1 and C++...
      - Deborah: The STL is the Standard Template...
      - Ajay: What is STL? What is C++'s STL...
    - Jody: I agree Deborah, and I may be because I...

- T3
  - T3
    - T3
    - T2
      - T3
  - T3
    - T5
      - T5
      - T5
    - T3

Table 1. Detailed posts from Example, un-structured Threaded Discussion in Figure 1. Here, we see the hierarchical posts on the left and the corresponding Topic for each post on the right, such that all posts belong to one of five topics, T1 - T5.

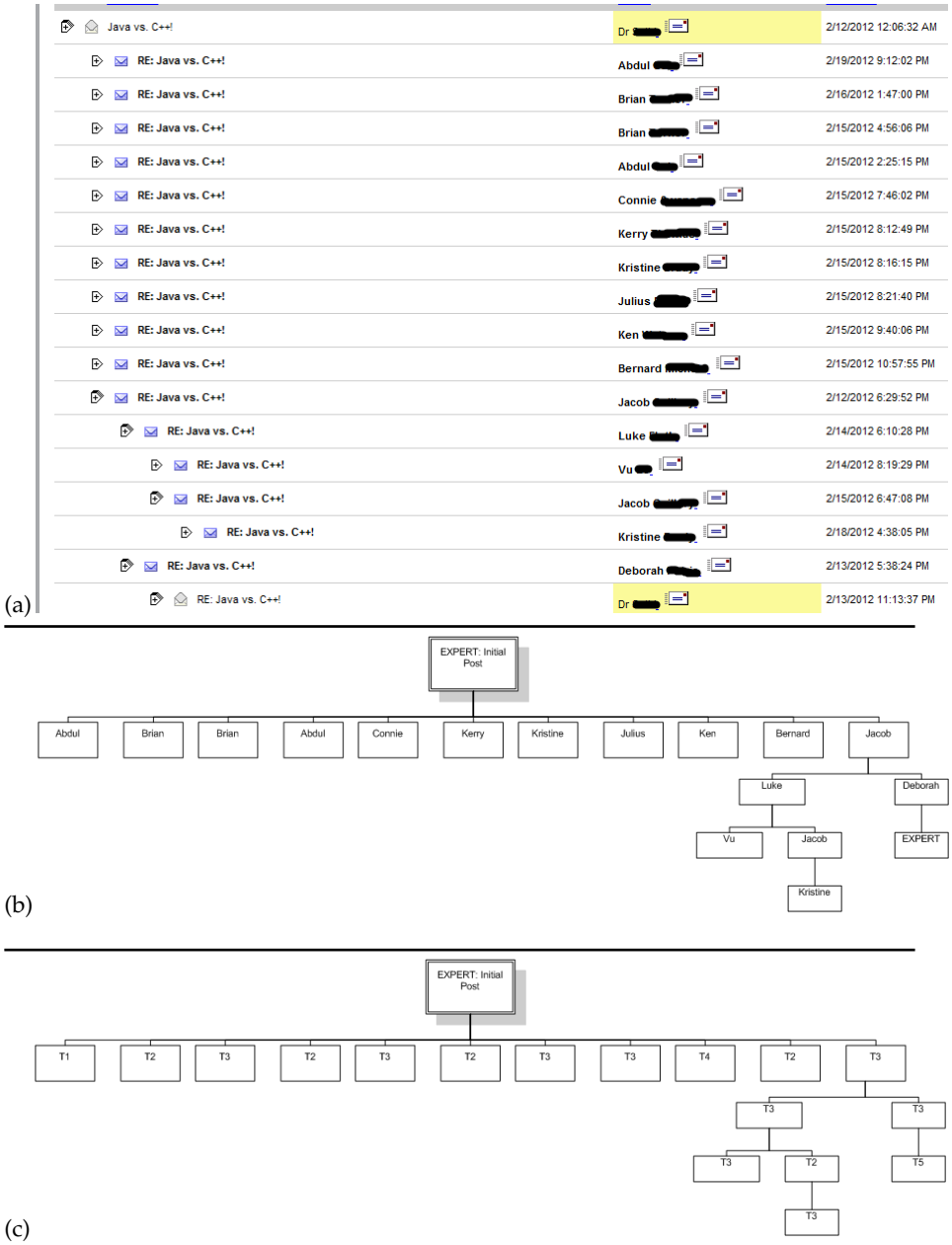

Fig. 1. Example Threaded Discussion: (a) original posts, (b) tree-view of posts, and (c) tree-view of topics.

Post	Topic
– Abdul: To the Bonus question: C++ design aspect is very limited...	– T1
– Brian: The syntax of Java is closely based on the syntax of the C++ programming language... • Abdul: There are so many differences and similarities, it sometimes is hard to mistake... • Kerry: I prefer the one that's right for the job. For a high-level user-facing application... • Bernard: I think that both have there place. ∗ Jacob: ... I took the same classes...	– T2 • T2 • T2 • T2 ∗ T2
– Julius: I do find JAVA to be a bit less combersum when putting together your methods... • Connie: When Sun Microsystems came out with Java, it like an answer to MOST of their prayers... • Kristine: three things that come to mind on what is different between Java & C++ are... • Jacob: ... With Java, the programming is more user friendly... ∗ Luke: I love the analogy, Jacob... · Vu: Hi Luke–I had... · Kristine: Java has both kinds... ∗ Deborah: I don't think Java... · Jody: I agree Deborah, and I may be because I...	– T3 • T3 • T3 • T3 ∗ T3 · T3 · T3 ∗ T3 · T3
– Ken: Why might it be said that Java is an object-oriented language while C++ is a procedural...	– T4
– EXPERT: Sounds like Java has 1 and C++... • Deborah: The STL is the Standard Template...	– T5 • T5

Table 2. Restructured representation of the Threaded Disucssion from Table 1, after removing redundant posts and minimizing the dispersion of topics. Here, T1-T5 again represent five different topics (stances) inferred from the posts. Different levels of minimization of these properties would be desirable under different conditions. E.g., this can be useful to students who might have a hard time distinguishing important topics or components when they're dispersed or serve as a reference which summarizes the content instead of having dispersed content that makes it difficult to understand the central ideas.