Issues in Informing Science and Information Technology

Manufacturing Organizational Memory: Logged Conversation Thread Analysis

Masafumi Kotani
IBM, Tokyo, Japan

mkotani@jp.ibm.com

Abstract

Though not especially media-rich, mailing lists remain in use and retain popularity for their built-in technological controls and their capability to “reply” to a message (“continuing a thread”). The motivation for extracting knowledge fragments from the unstructured text of mailing lists is compelling, though successes doing so may be considered only partial because it requires mental processing, or a certain cognitive effort, that complicates automation. Cognitive psychology distinguishes the Long Term Memory (LTM), which may be compared to text thread storage, from the Working Memory (WM), which initiates the retrieval of knowledge fragments stored in the LTM. Searching by subject, date, and time stamp ranges, and by keyword-inclusive fragments, constitutes the commonly used methods for executing sequential LTM retrieval. Retrieval can, however, be greatly enhanced by automatically gleaning certain classes of threads from the entire structure and displaying them alongside other properties. Here, we describe automatic “class” extraction and its effect on OM manufacturing and LTM retrieval.

Keywords: Thread analysis, thread classes, OM manufacturing

Introduction

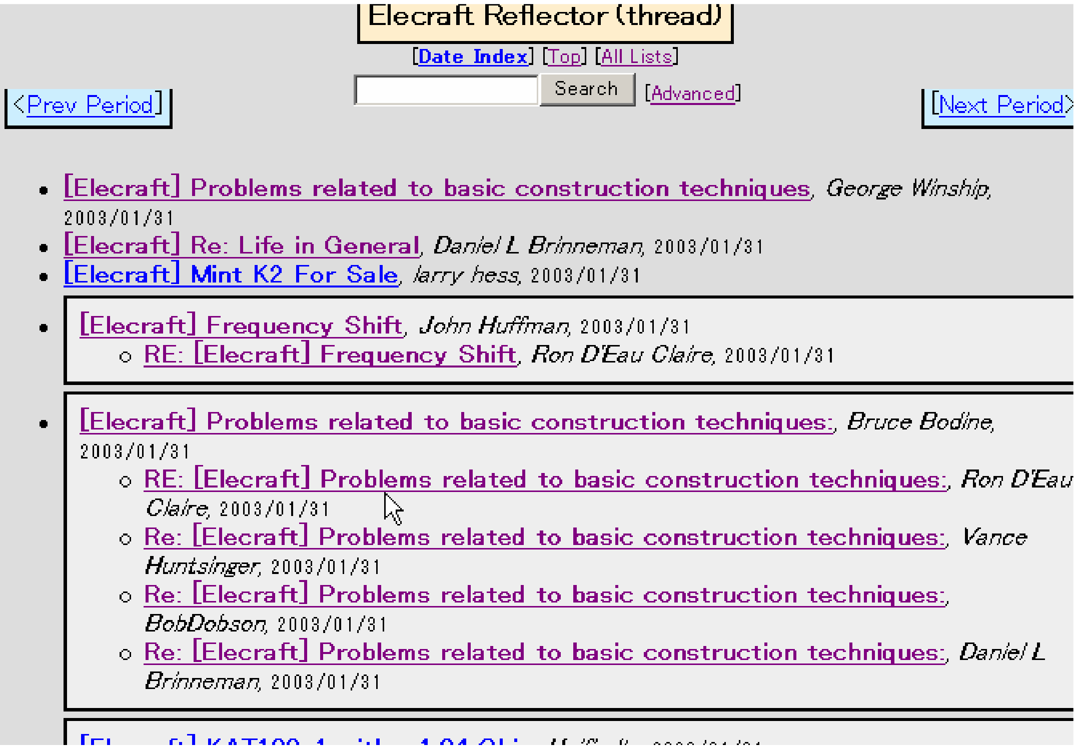
Logged conversation must be conceived of as an OM (Organizational Memory) manufacturing process acquired during active discussions among conversation participants and logged into a hierarchy of text-format data. A mailing list provides a collection of informative correspondences, and its use as a source of organizational knowledge in business is widespread and obvious. To make this use easier, however, good search tools are required. The plethora of current and widespread search technologies can be classified into four categories:

(1) *Context-extended search* using a thesaurus. In this approach, the query term is expanded in context utilizing words/clauses/phrases from a thesaurus to broaden the catchment space.

(2) *Query-by-example.* Here, the user selects relevant document snippets, which are then used for the query base.

(3) *Keyword Search* on full/partial text contents.

(4) *List **Scan***, by which the user sorts by date, author, and subject by thread listing within the structured list (See Figure 1).



- [Elecraft] Problems related to basic construction techniques, *George Winship*, 2003/01/31
- [Elecraft] Re: Life in General, *Daniel L Brinneman*, 2003/01/31
- [Elecraft] Mint K2 For Sale, *larry hess*, 2003/01/31
- [Elecraft] Frequency Shift, *John Huffman*, 2003/01/31
  - RE: [Elecraft] Frequency Shift, *Ron D'Eau Claire*, 2003/01/31
- [Elecraft] Problems related to basic construction techniques:, *Bruce Bodine*, 2003/01/31
  - RE: [Elecraft] Problems related to basic construction techniques:, *Ron D'Eau Claire*, 2003/01/31
  - Re: [Elecraft] Problems related to basic construction techniques:, *Vance Huntsinger*, 2003/01/31
  - Re: [Elecraft] Problems related to basic construction techniques:, *BobDobson*, 2003/01/31
  - Re: [Elecraft] Problems related to basic construction techniques:, *Daniel L Brinneman*, 2003/01/31

Figure 1. Typical Thread list (levels are compressed and distorted by the print width limitations)

Experimental Setting and Results

Corpus Selection

From the collection of 20 mailing lists gathered for this study, we found on analysis that they consist of three primary types:

1. In "strongly-typed" mailing lists, each reply or thread-initiating message can only start after the completion of a "category" input. A typical example of this type of mailing list is Organizational Memory Information Systems (OMIS), which force users to input a category type into the system, often presenting an unnecessary entry barrier. Such "category type" information can significantly enhance later OMIS search usability and effectiveness, but our observation has been that the threads do not extend to more than 10 messages. Here, we suspect the negative effect of the "entry barrier" outweighs usability, and we cannot argue that this type of mailing list is good for our thread classification (Mark & Bordetsky, 1998).

2. A typical "expert locating" system allows each member to register and modify the index list pertaining to his or her specialized knowledge or experiences so that later users can locate subject matter experts by a simple keyword query. Each expert consultant will be registered, and the collected expert profiles form a "thread" in the mailing list. Users can add comments and feedback, rating the services received. Such registration data also constitutes a mailing list. It is rare here, as in type (1), to see sufficiently long threads. Another drawback is that OM manufacturing does not happen in the "expert locating" system per se. Usually, the OM is created while messages are exchanged in a "weakly-

typed" message swap (Fisher, 2002). For these reasons, it is inadvisable to choose the "expert locating" system for thread analysis and classification.

3. The "plain-vanilla free format" or "weakly-type-forced" mailing lists is often used for hobby and special interest groups. In enterprise, employees form "intra-net" mailing list to assist OM creation and utilization (Hood, 2003).

The mailing list we selected for analysis was borrowed from the amateur radio kit building community (Elecraft, 2003), where OM processes are actively performed and databases frequently accessed. Our list displays distinctive aspects arising from the nature of this community:

1. The list has been used since 1998 and the total number of messages per month increased constantly from the start.
2. The January 2003 list consists of 1406 messages by 311 members. As shown in Figure 2, 156 authors (50%) account for 86% of the messages, forming the list's active core. Forty-seven authors (15%) account for 54% of the messages.
3. Visitors and infrequent guests welcomed by "expert" advisers constitute 63% of the posters, posting only once or twice in the course of a typical month.

Because of these well-behaved community aspects, OM created in the above list qualified as a relevant analysis corpus for our thread classification.

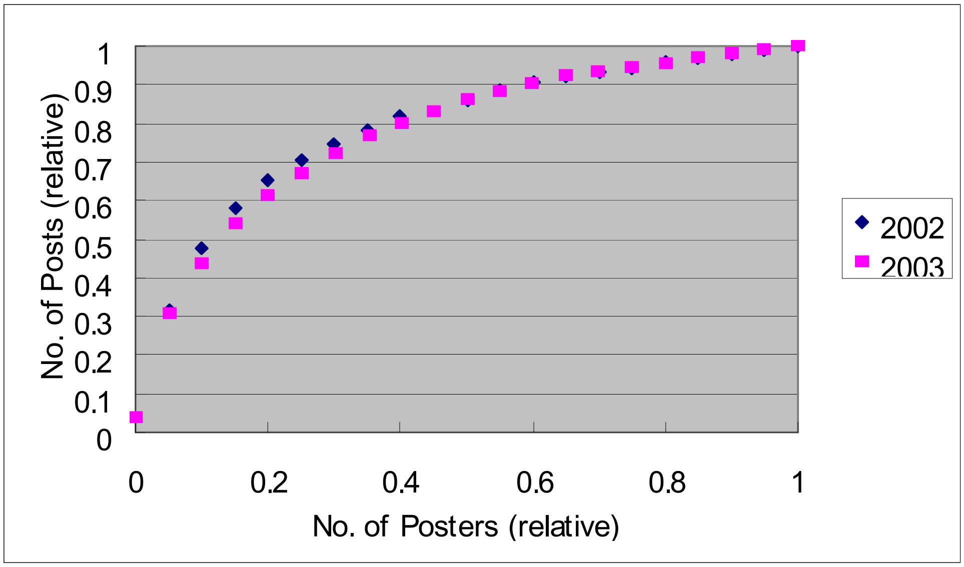
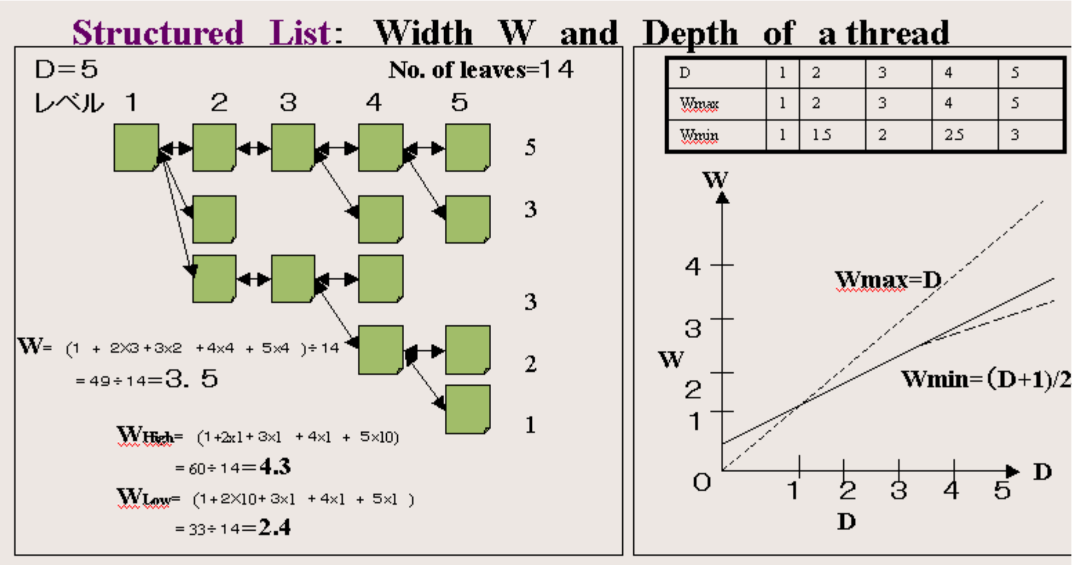


Figure 2. Posters and Number of posts (2002-2003 January comparison graph)

Thread Metrics

Every thread adheres to the general pattern of a life-cycle curve – messages increase to reach a peak and then trail off. The width and depth metrics of a thread are defined in Figure 3. Width W is a weighted average of depth reflecting the 2-dimensional characteristics of a thread. If the thread is wider in a later part of the entire structure, W is larger and nearer to the Wmax. When the thread is wider at an early stage of the life cycle, and not wide in the later part, W gets closer



Structured List: Width W and Depth of a thread

D	1	2	3	4	5
Wmax	1	2	3	4	5
Wmin	1	1.5	2	2.5	3

W= (1 + 2×3 + 3×2 + 4×4 + 5×4)÷14
= 49÷14=3. 5

WHigh= (1+2x1+3×1 + 4×1 + 5×10)
= 60÷14=4.3

WLow= (1+2×10+3×1 + 4×1 + 5×1)
= 33÷14=2.4

Figure 3. Definition of thread depth D and width W

to, or smaller than, Wmin. We will concentrate on developing a classification of the threads using both the D and W measures nescience, agnostic method with minimum detention.

Our standard for analysis was that the ideal classification model should be applicable to lists with both small and large message counts. Since we examined a small mailing list (of 200 messages and 35 threads), convenient classification types—such as "announcement/acknowledgement"—may be observed (Gushiken, 2002). For larger-sized mailing lists (with messages counts of 1400 and 400 threads containing serious opinion exchanges), we are no longer able to identify such classifications, since relevant information becomes buried under the sample sizes. Commensurate with the scattered plot of large mailing lists, we were able to identify 3 primary thread clusters: (A) threads which continue to grow because of continued active discussion among posters, (B) threads containing mainly one-to-one discussions, and (C) threads attracting numerous postings at an early stage of the life cycle, then decreasing in frequency of use. These characteristics form the basis for automatic classification, and color-coding threads by these cluster characteristics assisted in the identification of posting activity type. The contents of (A)-type clusters, for example, are discussions of daily used radio components like antenna, and posting continues by gathering participants. (C)-type clusters discuss new features, events, components, parts, and programs, gathering early experiences and then trailing off with follow-through postings. One message thread is a barren D=W=1 tree and is not colored.

Results

Color-coding the life cycle of a given posting tree helps to identify, explicitly and implicitly, its textual content, differentiating 50% of entire threads. Red (category A) applies to 17%, Green (category B) 6%, and Yellow (category C) 27%. Figure 4 represents how the color code applies to the thread population.

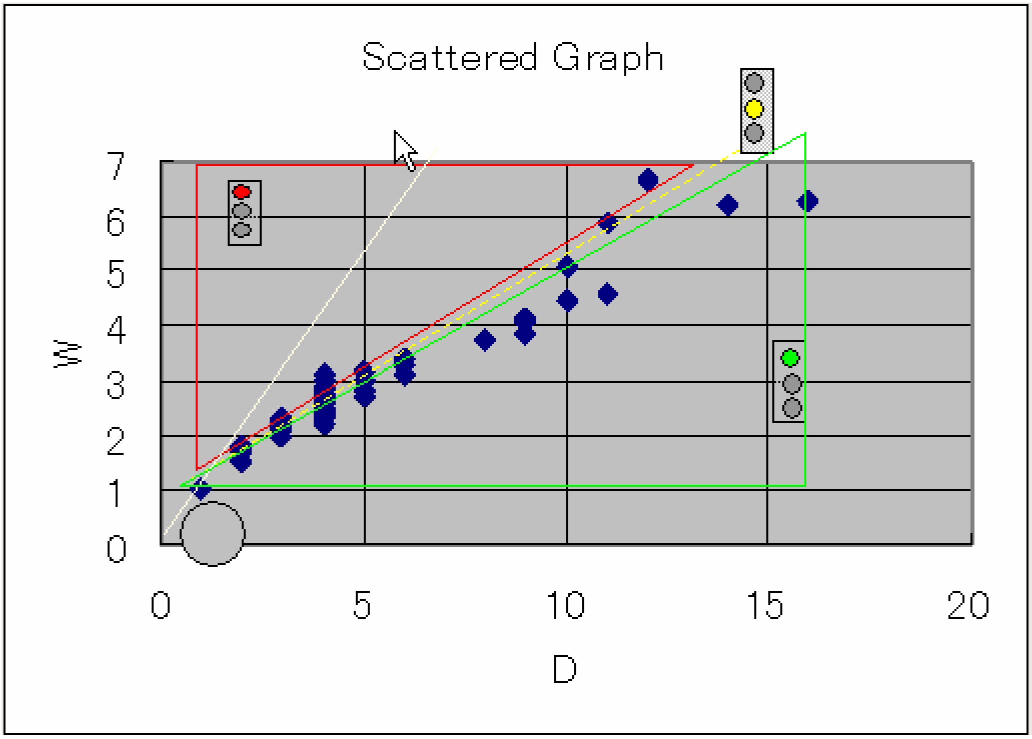


Figure 4. Thread classification Color –coding type A, B, C counter-clockwise

Summary, Conclusions, and Future Work

This paper offers a new approach to applying the automatic classification of threads using an automatic classifying and analysis algorithm. We have shown a systematic recipe for improving efficiency in knowledge manufacturing at creation time. Our contributions are three fold:

1. Thread lists integrate additional indicators (color coding) to present the posting pattern (posting frequency and life-cycle shape) for particular threads.
2. An enhanced search method having new criteria
3. Authors will become motivated, at the time of entry, to base decisions on the life-cycle position of the posting, thus OM manufacturing is enhanced.

In addition, there are a few relevant research topics introduced by the proposed approach (Iske & Boekhoff, 2002). We applied our method to calendar-month-blocked lists, but should we increase the length of blocking time, the size of threads may yield different properties than those presently revealed. Future work could include refinement of the model by applying it to larger mailing lists from much "heterogeneous" membership communities (Borgatti, 1995). We are currently applying the described approach to implement a new mailing list archival/retrieval software package.

Acknowledgements

The research presented here was made possible by the generous support by the OM Study Group at Tokyo Institute of Technology under guidance of Dr. Kunihiko Higa. The author would like to thank Ms. Yuriko Satoh and Mr. Daisuke Sahori for their insightful influence on this research.

References

Borgatti, S. P. (1995). Social network analysis - Centrality and aid. *INSNA*. University of South Carolina.

Elecraft Mailing List archive server information. (2003). Elecraft Corporation www.elecraft.com and Doug Netherton ve3mcf@rac.ca Thomas Martin martin@ac6rm.net http://www.ac6rm.net/mailarchive/html/elecraft-list/2003-01/threads.html

Fisher, D. (2002, March). Newsgroups and Mailing List Analysis: A few notes and preliminary results. UC Berkeley Computer Science. (Danyelf@cs.berkeley.edu).

Gushiken, E. (2002). Mailing List Analysis. University of Shizuoka.

Hood, E. (2003). MHonArc A Mail-to-HTML converter. (mhjonarc@mhonarc.org).

Iske, P. & Boekhoff, T. (2002). The value of knowledge doesn't exist: A framework for valuing the potential of knowledge. *Lecture Notes in Artificial Intelligence 2569*, p. 634. Springer-Verlag.

Mark, G. & Bordetsky, A. (1998). Structuring feedback for groupware use: Memory-based awareness. *IEEE*, California State University (bord@csuhayward.edu).

Biography

Masafumi Kotani is Marketing Executive of Industrial Sector of IBM Asia Pacific Service Corporation. His field of specialization is integrated risk management in business process reengineering.