

Reverse Engineering Collaboration Structures In Usenet:
An alpha-level research proposal

Philip Johnson
Collaborative Software Development Laboratory
Department of Information and Computer Science
University of Hawaii
Honolulu, HI 96822
johnson@hawaii.edu

(3500 words, 440 lines)

Introduction

A primary form of research in CSCW concerns the representation of computer-mediated collaboration. Systems such as gIBIS, SYBIL, The Coordinator, ConversationBuilder, and so forth propose mechanisms to structure and inter-relate the products of group activities. Such research typically involves the implementation of specialized environments that require participants to produce collaborative artifacts that conform to the representations proposed by the researchers.

Unlike these systems, which are typically experimental systems used by relatively few individuals in the CSCW research community, there is a magnificently successful, world-wide, on-going computer-mediated collaborative system involving literally tens of thousands of active users, the daily production of gigabytes of collaborative artifacts, and exponential growth in both participation and volume over the past several years. All this has occurred and continues to occur daily with virtually no representational support for collaboration in the sense conventionally used by CSCW researchers. This system, of course, is USENET.

USENET, technologically speaking, is nothing more than a large scale, distributed BBS accessible to virtually anyone with an internet address. It contains thousands of newsgroups (my .newsrsrc file, which contains only the subset of USENET newsgroups subscribed to at UH, has 2399 entries) arranged in tree-like hierarchy. Here is a brief excerpt from my .newsrsrc, which lists the name of the newsgroup, a ":" if I have chosen to display this newsgroup in my reader, a "!" if not, and the current number of postings.

```
comp.windows.interviews! 1-1764
rec.sport.football.college! 1-10510
rec.music.gdead: 1-73292
rec.arts.startrek! 1-104038
comp.software-eng: 1-11458
comp.sys.atari.st! 1-45198
comp.os.vms! 1-46968
alt.tasteless.pictures! 1-30
```

This should give some indication of the range of topics discussed in USENET as well as the sheer volume of material.

Various software systems, called "readers" exist to support perusal of newsgroups. In this paper, the term "reader" refers to this software, while the term "user" refers to the human using the reader.

While USENET is phenomenally successful along certain dimensions, it will soon, if it has not already, reach a size that erects a significant barrier to those attempting to use it as a mechanism for collaboration. In a manner similar to the "combinatorial explosion" problem of large search spaces, USENET is encountering the "collaborational explosion" problem.

However, USENET's size also represents a resource in collaborative diversity, geography, and immediacy that contains unparalleled and unexploited opportunities for collaboration.

This alpha-level research proposal claims that (a) the collaborative value of USENET can be increased, and (b) the "collaborational explosion" effect significantly reduced, through transforming the relatively "raw" daily global USENET data feed into a locally richer, CSCW-style collaborative representation network, supported by a new kind of reader system. Such an approach represents a "reverse engineering" for collaborative systems. It recognizes the intractability of the "forward engineering" style of CSCW research referenced at the beginning of this paper for the problems and opportunities of USENET.

The next section very briefly touches on some of the current problems in the current use of USENET for collaboration. The following sections present a thesis, method, and several evaluation experiments for this research proposal. The paper concludes by discussing some potential problems and extensions of the approach.

Problems With Usenet as a Collaborative System

Many of the problems with USENET can be categorized as: information overload, insufficient structural support, insufficient newsgroup representations, and insufficient global and local user representations.

Information Overload

The information overload problem in USENET is very real. Popular newsgroups frequently receive 50-100 messages a day. Two primary problems resulting from this concern filtering and retrieval.

First, it is difficult to effectively filter so much information, and thus coarse and error-prone decisions are made (such as to not read a newsgroup at all, since the signal-to-noise ratio is too low, or to decide to skip a posting based solely upon the Subject: line.)

Second, the volume of postings makes it virtually impossible to do "literature search" through prior postings. As a result, questions such as "What CASE tools exist for the PC?" are posted over and over again in the software engineering newsgroup, primarily because the volume of material precludes its perusal. In response to this, ad-hoc structures such as FAQ (Frequently Asked Questions) postings have arisen in some newsgroups, though they suffer from obvious

limitations.

Insufficient Structure

Beyond the hierarchical newsgroup organization, there is almost no explicit structure in USENET. Individual postings appear at a site in almost, but not exactly, the order they were posted in, and are presented to users organized by newsgroup. A set of ad-hoc conventions have developed within the USENET community to overlay a logical structure on top of this chronological structure. For example, the information stream of a newsgroup is typically composed of a set of conversational threads, interconnected through the use of "Re: <prior subject line>" in the Subject: line and/or through the use of context lines in the body of the message, where a portion of a previous posting is copied and then responded to with new material. Thus, while there is conceptually a logical network of inter-related material, typical USENET readers.

Inefficient Collaborative Search Support

The tens of thousands of regular USENET participants form an incredible resource of human knowledge. Collaboration occurs passively, through simple reading of postings, or actively, by posting requests and responding to previous postings.

Unfortunately, active searching for collaboration typically involves "blind" search: one selects one or more newsgroups, posts the request, and hopes that someone with relevant will happen to retrieve that posting and respond to it. Given the ubiquitous need to perform coarse and error-prone filtering based upon newsgroup name or Subject line, requests are often reposted two or more times, or to several overlapping groups, hoping that the redundancy will lead to success. This strategy exacerbates the problems of USENET for all participants by further decreasing the signal to noise ratio and increasing the overall volume of information.

A not uncommon phenomenon that results from blind search is to post a request for information to a world-wide newsgroup, only to have it responded to by a colleague down the hall.

Insufficient Newsgroup Representation

Given the thousands of newsgroups available, most users have only time to reference a fraction of them, and thus a wealth of potentially useful information is lost. Further, the sheer effort of keeping up with newsgroups that one has a continuing interest in precludes any serious attempt at browsing through others, since such browsing activity can only take the form of downloading the most recent set of postings and sampling them. This activity is highly error-prone, since the topics may not be representative of the group, and because such a "snapshot" does not provide answers to more important characteristics of the newsgroup useful to determining its potential interest and quality. A sampling of such characteristics might include: What are the current set of topics? How many topics are discussed concurrently? How long is the typical discussion? What is the ratio of requests to responses? How detailed are the responses? How many currently active participants are there?

Note that these three problems have a profound, self-perpetuating effect on USENET: the volume of postings results in behaviors by users that decrease the signal to noise ratio of individual postings, leading to coarse filtering strategies, which lead to an increased volume of postings, and so forth.

Thesis

Due to the variety of interfaces and computational facilities, it is impractical to improve the collaborative representational mechanisms of USENET at the "source", where postings originate. This proposal claims that significant improvement to the collaborative capabilities of USENET can be accomplished through a process termed "collaborative reverse engineering". This consists of analyzing the "raw" USENET information stream and re-representing it into a richer representational framework. This richer framework forms the basis for improved USENET reader software. We claim that this framework and supporting software will ameliorate the three problems noted above, and propose quantitative experiments designed to test this claim.

Method

The USENET feed (actually, only a subset of the feed consisting of perhaps a half dozen related newsgroups) will be automatically processed by a front-end collaborative system called EGRET developed by the Collaborative Software Development Laboratory in the Department of Information and Computer Sciences at the University of Hawaii. (For the purposes of this discussion, EGRET can be viewed simply as a Unix and X-window-based, distributed, client-server hypertext system with both graphical and Emacs front-ends.) The preliminary analysis will perform the following collaborative reverse engineering activities:

- (1) Each posting will be saved in an individual node, with author, subject, date, contents, and so forth represented in distinct fields.
- (2) Cross-references within the content field to prior postings will be re-represented as hypertext links. A variety of ad-hoc conventions exist for referring to prior postings, such as reproducing relevant lines, preceding each with a ">". Those cross-references that can be recognized by a simple pattern-matching parser will be automatically converted into hypertext links. While raw USENET postings consists solely of "backlinks" (one can never refer to an as yet unposted article), a significant effect of this stage of processing is to replace the awkward backlink representation with a more natural forward-link representation, where each posting contains within it hypertext links to the future postings that refer to it.

Although some cross-references cannot be extracted automatically, user facilities exist to manually add them. However, we believe that the basic utility of the system must not require high levels of user encoding of cross-references.

- (3) An initial "type" for the node is heuristically derived from examining the contents of the Subject line and the actual contents of

the posting. Such types are typically composite and hierarchical. An example might be: [Reply, CASE tool list needed]. Note that, following the philosophy of EGRET, users may later extend or alter this initial type, and different users may maintain different types for the same node, although a progression toward a common consensual definition is encouraged by the design of the system.

(4) If one or more links result from from step (2), each is heuristically assigned an initial type through another simple pattern-matching process. Such link types represent relationships such as "agrees", "disagrees", "elaborates", "summarizes", etc. Obviously, this initial, automated assignment is highly error-prone, and we believe that the basic utility of the system depends upon user verification of these values. We believe that such verification can occur as a normal result of using the system: if a user traverses a "agrees" link and finds a posting that disagrees, the reader will provide the user with simple and fast mechanisms to correct the link type so that other users will not encounter the error.

(5) A "user profile" of the poster of the node is augmented with information extracted from this posting. Such information includes structural information derived from steps (2)-(4). This user profile indicates not only the interest areas of the user but also their style of interaction with USENET.

(6) A "newsgroup profile" is augmented with information from steps (2)-(4). The newsgroup profile summarizes the content of information contained within the newsgroup, dynamic aspects of the flow of information, and characteristics of the active contributors to the newsgroup.

(7) An EGRET user interface to USENET is provided consisting of three readers, one for postings, one for users, and one for newsgroups. The Postings reader displays the set of nodes posted for a particular newsgroup, with links between them where cross-references appear. The Users reader displays the set of active users in the system, organized according to interests and behavioral characteristics. The Newsgroups reader displays a summarized status or historical perspective on the newsgroup concerning the topics discussed and their structural characteristics, the user community, and the current "state of play".

(8) Even with a small subset of newsgroups, the size of the hypertext database will quickly become unmanagable, and so some policy for "expiring" posts must be implemented. Initially, a mechanism that establishes a moving window of the most recent thousand or so postings seems reasonable. Interestingly, the width of this window can differ dramatically for the postings, users, and newsgroups profiles. For a set of active newsgroups, a thousand nodes might be posted every week or two. However, the size of the users and newsgroups profiles grows much more slowly, and thus the window of retained information for profiles might be months or years wide.

Discussion

The reverse-engineered posting, the newsgroup profile and the user profile together present a radically different image of USENET to the user.

The Postings reader presents individual postings as nodes in a browsable hypertext network that makes structurally explicit the implicit contextual structure present in the postings. In so doing, it re-engineers backlinks into forward links. This facilitates navigation and raises the signal to noise ratio by eliminating redundant copies of text. It allows users to physically browse a newsgroup in a manner corresponding to the logical structure of the conversation, rather than via the arrival time of postings at the local site.

It is the Users and Newsgroups readers, however, that fundamentally change the style of interaction with USENET. They do this by making visible to users characteristics of other users and newsgroups that provide novel support for collaboration. For example, if a user has a question regarding a certain topic, the Users reader can identify a set of posters that appear likely to be able to answer the question, based upon the fact that they have posted related topics in the past. This enables the user to send the question via e-mail directly to a set of people relatively likely to have an answer, and relatively likely to respond. (Those who take the time to post to USENET are probably also those who will take the time to respond to a request for information directly related to their areas of interest.) Only if this initial request fails will the user need to revert to the error-prone, global and blind search strategy of posting to entire newsgroups.

The Users reader also provides useful behavioral data about individual users, by summarizing the structural characteristics of the postings made by the user. First, the Users reader can provide a profile of a user's interests by listing the conversational threads this user has participated in and their types. Second, for each of these threads (and in aggregate form), the Users reader can provide insight into the collaborative character of the user: did the user originate the post? How many follow-ups were made by the user? Were the follow-ups agreements, disagreements, elaborations, etc.? Third, the Users reader can construct a network that displays the set (or subset) of users, along with the kind of collaboration that has occurred between them. This could potentially expose interesting collaborative subgroups, such as cliques of self-supporting users, or warring factions.

The Newsgroup reader allows users to obtain a high-level perspective on other newsgroups without actually having to read sample articles, and thus aid them in determining when the conversational topics or collaborative style of a newsgroup appeals to them. The newsgroup reader displays the typed, hypertext node and link structure of the conversations as a graphical network. Such a network allows the user to browse a newsgroup in a more selective and efficient fashion, for example, by reading a few postings that originate conversational threads, instead of merely reading the most recently arriving threads and attempting to reconstruct from that what the conversation is about. The structure of the graphical network displayed by the Newsgroup reader can also supply valuable, "gestalt" information about the current character of the collaboration within the newsgroup. Does the newsgroup consist mainly of unanswered requests for information? Is there one, a few, or many simultaneous on-going conversations? Are conversational threads long and complex, or simple and short?

The Newsgroup reader can be simply extended using EGRET "agents" to provide each user with a personal monitor that watches the flow of subject matter in all newsgroups, for example, and informs the user when topics pre-selected as interesting arise within a newsgroup.

Evaluation

This research claims that a reverse engineering approach to constructing collaborative representations can improve the process of collaboration in USENET. How can such a claim be empirically verified? We outline below three experiments as representative of the many that can provide evidence to confirm or disconfirm this thesis.

(1) Measurement of the reduction in textual redundancy.

The raw USENET feed contains a significant amount of textual redundancy that occurs in two forms: (a) replicated postings to multiple newsgroups, and (b) replicated lines of a posting inserted into a new posting in order to provide context for the new information. The EGRET interface to USENET can eliminate both of these forms of redundancy, replacing the latter with hypertext links to the related information. One quantitative measure of the support for collaboration consists in simply determining the amount of reduction in textual information presented to the user, which corresponds to an increase in the signal to noise ratio. Newsreaders such as GNUS already support elimination of redundant nodes when cross-posted, so the elimination of this form of redundancy will not be factored into the comparison.

The next two experiments compare the actions of users employing the EGRET reader to some other newsreader. We propose to compare our reader with the GNUS reader, because GNUS is relatively sophisticated among USENET readers, and because a modified version of GNUS can be easily built with instrumentation that collects data on user actions for our experimental purposes.

(2) Measurement of navigation sequence of conversational threads.

As noted above, our proposed technology provides structural support for the logical structure of the conversational thread. This should make an impact on the sequencing of post retrieval by users away from a strictly chronological form and toward a more logical orientation. For example, we expect that GNUS reader users have a significant tendency to simply some subset of the recent unread postings, in whatever sequence they arrived, then "catch up" to eliminate the remainder. Using the EGRET reader, we expect users to read postings in a more logical manner, perhaps waiting until a thread has developed for some time before starting with the original post and then sequentially reading through the conversational thread in its logical order.

(3) Measurement of newsgroup browsing.

The sheer volume of USENET means that any individual user can only regularly read a small fraction of the total information available.

Current reader technology, by providing little support for summarizing the current state of the conversational structure of a newsgroup, leads to many users to subscribe to only a handful of newsgroups on a regular basis. However, with the EGRET newsgroup reader and associated structural information, it becomes possible to "follow" a newsgroup at its structural level while retrieving actual postings relatively rarely. We expect that these retrieved postings will represent a logical choice (such as the originating posting for a conversational thread) in EGRET, as opposed to a chronological choice in GNUS. We further expect that with the EGRET interface, users will subscribe to a significantly larger number of newsgroups than with GNUS, while doing approximately the same amount of actual node retrieval.

Technical Feasibility

For prototyping and experimental purposes, the EGRET environment provides an almost ideal platform for the implementation of this reverse engineering approach to USENET collaboration. The use of an hypertext-enriched Emacs as its primary front-end provides a high-level environment for restructuring USENET posts if simple pattern-matching techniques prove sufficient. The agent mechanism in EGRET, which allows autonomous clients to interact with the database, further simplifies the implementation: a "downloading" Emacs/EGRET client can be automatically awoken each night to perform the restructuring on all postings received during the previous day. Similarly, an "expiration" client can check the database periodically and remove old postings. However, we can already discern several potential problems with our approach, which we discuss below.

One potential technical problem is if the "simple pattern matching" techniques to discover links and assign types turns out to be not so simple, and in fact requires AI natural language understanding techniques in order to function effectively.

A second technical problem is with the graphical browser for newsgroups: such a browser may require a unexpectedly sophisticated layout algorithm in order to accomodate the complexity of interactions in certain newsgroups.

A third technical problem concerns scale. While we believe our choice of implementation environment is well-suited to the research nature of this proposal, we do not believe EGRET can currently handle the volume of postings occurring across the entire USENET and maintain adequate responsiveness.

Finally, we note that this approach deals exclusively with gathering information from the postings, and does not gather any information from the user as reader. For example, it is clearly possible to improve user profile information by combining information from those postings made by a user with information about those postings read by the user. We leave the exploration of this as a future extension.