Project idea: mining crisis googlegroups

The Problem

Let's start with "what is the problem I'm trying to solve here".

I'm responsible for designing Hunchworks. I'm also responsible for making it what the users need, and for understanding and fitting into the systems that they already use to do hunch-style collaborative inference. Now we know a bit about business analysis, so we've spent a while looking for these existing systems, and found them mainly in two places: closed discussion groups (e.g. Googlegroups), and Skype group chats.

I've traced (and anonymised) a couple of threads by hand - the main point is the type of information that people post rather than who and what it is - but whilst I was doing it I realised there were mining tools that could help with this, and my usual problem of not being able to find information on X in a Googlegroup after it's posted (no, the search doesn't work for this). Also, any visualisation or compression could be extremely useful to people trying to navigate the thousands of emails and hundreds of topics that thread their way through the crisismapping googlegroups.

Getting the Data

So to work. The first thing I needed was googlegroups data into my machine. There's an RSS feed for group NameHere at http://groups.google.com/group/namehere/feed/rss_v2_0_topics.xml but that appears to be only for new posts. A Google search on “mining google groups –site:groups.google.com” didn't show anything, but I did get a useful journalists guide to Google’s search operators as commiseration prize. And then paydirt: a PHP Googlegroup scraper that produces xml, that I've modified to dump xml into a file.

The data is about a group, so it has people (in username and cc fields), topics (in subject field), areas of interest (in the subject and text, so I’ll have to do some text analysis to get these) and times/dates.

As an aside, it would be really cool if Google allowed tagging on these posts. That way, people could come in later and add tags to each post, to make searching by area or subjects much easier.

Knowing which questions to ask

What I'm trying to understand at first is how people communicate data around an emerging crisis. Which people get involved, do they bring each other into the group to help, what types of evidence do they provide and which external links do they point at and when.

Another question is “are there cliques in this group” – not, “are there groups that don’t talk to

each other", but "are there natural groups of people who cluster around specific topic types or areas". And "who are the natural leaders – who starts the conversation on each topic, who keeps the conversations going".

More subtle things to find are 'hubs' (people who connect other people together), attachments/links (so we can see what's being linked to outside the group pages) and information flows between people and affiliations.

So first, what else can I do with the raw data? Track threads? Track related subjects? Pick out geographical terms, for instance?

Let's start with the people, and how they connect together. Let's assume that all people on the same thread are linked, or possibly that people who reply to each other are linked. A google search on "mining relationships in email" proves useful here… turns out there's something called "relationship mining" in data warehousing already. And the really simplest thing to do with this is a histogram of people who posted, or who sent the first post in a thread. But… but… most of the literature on mining emails assumes that there are 1:1 links between people, i.e. each email has a sender and 1 or more recipients. What we're looking at here has a sender, all the group as recipients, and maybe a couple of other outside-the-group people added as recipients too. So we don't have direct links between people: we have the group as an intermediary. So how do we know when two people are engaged with each other via the group? One answer is "reply-to" – i.e. when one user replies to another user's posting. This will take a little work in the scraper to do, but could be a useful way of establishing chains of people through the group.

This is just the start of a chain of thought that is currently moving at the speed of code. There will be more, just not for a little while...',