1. Introduction

E-mail is an attractive entity that encapsulates human roles and relationship (sender and receiver) together with subject and textual context, which can be used for almost any purpose of textual communication (inquiry, complaints, thanks, and so on). Personal e-mail and mailing lists can be intensively used for knowledge management, while e-mail from customers can be a primary data source for business intelligence (BI) and customer relationship management (CRM). Ducheneaut and Bellotti (2001) reports that e-mail is more like a habitat than application, implying that it is now a commodity for everyone, and therefore the research on e-mail would benefit us enormously. Knowledge discovery from e-mail (or e-mail mining) has been actively explored in a few specific areas. Machine learning and document categorization techniques (Zhang and Oles, 2001) have already made a product for quite high-quality routing of e-mail. Lotus Discovery Server1 (LDS) and KnowledgeMail2 of Tacit Knowledge Systems successfully associate users with their expertise areas by analyzing documents including the e-mail messages. E-mail mining, however, should be able to address a broader range of technologies and applications. In particular, analytic functions -- incorporating both metadata (ID, timestamp, sender, receiver, subject, quotation/reference, etc.) and textual messages of e-mail -- should provide

- context-centric e-mail digesting (threads)
- human relationship extraction, and
- trend, pattern, and regularity in e-mail

for facilitating knowledge sharing/reuse and better understanding the community (e.g., a personal e-mail universe or a newsgroup). These are the central goals of CRM and knowledge management.

2. E-mail Modelling

We will first define a schematic and topological aspect of e-mail. An e-mail message $e$ consists of values of a set of metadata $M$ -- including at least an ID, sender, receiver, date, and subject -- and body (textual content). The set $M$ can optionally include carbon-copy (cc), and reference (in a reply to a previous e-mail) ID. The body can optionally include quotations from other e-mail. For a set $U$ of users, the sender value of $e$, denoted by $e.sender$ or $sender(e)$ is an element of $U$. The receiver and cc values of $m$ are a subset of $U$. We will consider only meaningful e-mail messages such that receiver is non-empty, and is not a singleton set of the sender. We also assume that body is non-empty. In many e-mail systems, reference might be implicitly represented by quotations or explicitly represented as an attachment of $e$. The set $E$ of e-mail messages corresponds to a personal e-mail collection for the user $u$ if for every $e$ in $E$ satisfies $e.sender = u$, or $e.receiver$ or $e.cc$ includes $u$. $E$ is then said to be of type

$k$.

1. for any two non-empty sets $E_1$ and $E_2$ of e-mail messages of type kingdom, the union of $E_1$ and $E_2$ is of type kingdom iff $E_1$ and $E_2$ are sets of e-mail messages for the same user $u$. The non-empty difference and intersection of $E_1$ and $E_2$ are always of type kingdom.
2. for any two non-empty sets $E_1$ of type kingdom and $E_2$ of type republic, the union of $E_1$ and $E_2$ is always of type republic. The non-empty intersection and difference ($E_1 \cap E_2$) are always of type kingdom.

These two basic types of e-mail messages are intended to capture personal e-mail collection and mailing list (or newsgroup), respectively. Different suites of analytic functions should be defined based on individual e-mail types, but it is also important to define a common underlying model and to identify the sharable analytic functions.

3. Textual Information in E-mail Messages

The body of e-mail message is a plain text consisting of one or more sentences. Among those sentences, there are the following ones that possibly carry specific types of information:

1. imperative sentences are used for asking the receiver to do some actions
2. interrogative sentences are used for asking information from the receiver
3. sentences with 1st and 2nd person pronouns as subject and object define relationship between the sender and the receiver

If e-mail messages include reference or quotations, they will form a larger unit of coherent context, called a thread. A thread is used to summarize the textual content of these e-mail messages by removing duplicate text and visualizing 2-dimensional development of the reference/quotation links. There are six roles of quotations – advice, answer, information, opinion, question, and request – to be associated with the links between referring/referred e-mail messages. An automated assignment of roles to quotations have been investigated by applying the following method to real e-mail messages:

1. identify inline quotations marked by specific leading characters (e.g., “>”)
2. identify commentary to quotations as a paragraph immediately following the quotations
3. assign roles to sentences in the commentary paragraph based on a decision list. For example, an imperative sentence beginning with “please” is assigned a role “request”.

The decision list at this moment is incomplete, and a machine learning approach is considered as an alternative to assign the roles to quotations.

In many e-mail systems, it is now a common practice to copy an entire e-mail message as an attachment to a new e-mail message. In this case, there is no clear markup of a quotation, and the link

4. $E$ can be of type kingdom simultaneously for more than one user. For example, if $U = \{\text{Tom, Jerry}\}$, any set $E$ of e-mail messages over $U$ is of type kingdom for Tom as well as Jeffry.
between the referred/referring e-mail messages is implicit, but a thread is constrained to be a tree, which makes analytic functions more efficient than those for general graph structures. For assigning roles to these implicit links, we focus on the use of 1st and 2nd person pronouns and imperative/interrogative forms of sentences in the body of e-mail messages as follows.

A) Wh-questions, and Yes/No-questions form a "question".
B) Imperative sentences and pragmatically imperative sentences (e.g., "Would you mind …?") form a "request".
C) Modal VP expressions ("should", "must", and "have to") and a verb phrase), with 2nd person pronoun as a subject, form an "advice".
D) Lexicalized patterns of subject-verb-object, where subject and object are 1st and 2nd person pronouns, are used to define individual roles.

After applying the above rules, we can characterize the body in terms of weighed vector of six roles. Again, machine learning approach can be further applied to determine a dominant role of individual e-mail message. Intention analysis (Nasukawa and Nagano 2001) can also be used to associate broader expressions with the roles of sentences.

4. Analytic Functions for E-mail Messages

Basic functions for e-mail mining are two-fold: organizing e-mail messages for facilitating search and knowledge sharing/reuse and analyzing patterns of e-mail messages for better understanding of the nature of a given collection of e-mail messages and the sender and receiver relationship.

4.1 Thread analyzer

Thread analyzer identifies the following chains of e-mail messages included in a thread. Here, a thread is regarded as a rooted, directed acyclic graph with a set of nodes for e-mail messages and a set of arcs for quotation links.

1. Open Thread: a chain of nodes starting with a query arc from the first node. A chain might be just a singleton node.
2. Debate: a chain of more than five nodes between two users with alternating sender/receiver roles.
3. Big Bang: more than three chains ending with the same node. The ending node is a source of all the development of the independent chains.
4. Cue Ball: a chain of nodes such that the receiver of a node is identical to the sender of the linked node. This is typical of successive mail forwarding.

4.2 Human relationship miner

For a set E of e-mail messages and a set of users U, we can define a directed graph G, called a user graph, as follows:

- each node corresponds to a user u in U.
- each arc from the node for u1 to the node for u2 exists if there is an e-mail message e in E with u1 as a sender and u2 as a receiver or cc. The arc has a weight w if there are exactly w such distinct e-mail messages in E.

Now, the human relationship miner finds the following types of users from the graph G.

1. Consumer: nodes with incoming arcs only.
2. Provider: nodes with outgoing arcs only.

Partner(u, k): a collection of node v such that, for a user u, and a constant k, the sum of weights of the two arcs from u to v and from v to u is larger or equal to k. It is often the case when E is a kingdom for u.

5. Implementation and Experimental Result

Graph construction and analytic functions in the previous sections have been implemented in our E-mail mining project. The figures 1 are obtained from currently implemented system output. Preliminary experiment shows the use of pronouns for role identification works fine for English e-mail messages. It is interesting to note, however, that since pronouns in Japanese are often omitted (zero pronouns) from surface expressions, modal expressions and pragmatic expressions (including honorific expressions) have to be more extensively investigated. This will lead us to the issue of language dependence of role identification techniques for e-mail messages.

We are working on acquisition of more lexicalized patterns for identifying roles of sentence and quotation link. More experiments will be conducted when we can associate roles with most of the quotation links.

References

Ducheneaut, N. and Bellotti, V. (2001) E-mail as Habitat. ACM Interaction 8/5, pp.30-38