Modeling Identity in Archival Collections of Email: A Preliminary Study

Tamer Elsayed
Department of Computer Science and UMIACS
University of Maryland, College Park, MD 20742
telsayed@cs.umd.edu

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742
oard@umd.edu

ABSTRACT

Access to historically significant email archives poses challenges that arise less often in personal collections. Most notably, searchers may need help making sense of the identities, roles, and relationships of individuals that participated in archived email exchanges. This paper describes an exploratory study of identity resolution in the public subset of the Enron collection. Address-name and address-address associations in explicit, embedded and implied email headers are augmented with name and nickname associations discovered from consistent use in salutations and signatures. Limited transitive closure heuristics are employed to extend pair-wise associations to richer representations of identity. Assessment of sampled results indicates that many potentially useful nontrivial associations can be detected.

1. INTRODUCTION

Those who seek to understand the distant past rely on two types of primary sources; documentary evidence that is (generally) intentionally retained by some official body (e.g., the National Archives) and informal communications (e.g., personal letters) that are (typically) serendipitously preserved. Preservation of persistent informal communications declined markedly with the advent of the telephone. The widespread use of email and the declining cost of long-term storage have the potential to dramatically reverse that trend, however, opening an important new window on our society that future scholars will surely wish to exploit. For example, the National Archive has 32 million Clinton White House emails, and they expect to receive more than 150 million from the present administration in 2009. Making sense of collections at this scale will require new types of tools; in this paper, we explore one important capability that such tools will need: computational models of identity.

Research on email access has traditionally focused on tools for managing personal collections, in part because large and diverse collections were not available for research use. That is starting to change, most notably with the introduction of the Enron collection [8]. Email is a conversational medium in which individual identity can play a key role for tasks such as exploratory search and social network analysis. Scholars working with more formal media have long relied on source characteristics (e.g., journal reputation or Web page-rank) but the central role of individuals in the construction of email conversations results in finer-grained distinctions, and thus an explosive proliferation of sources that searchers new to the collection could have great difficulty comprehending. Fortunately, email provides a substantial amount of evidence that opens up opportunities that are not fully exploited by the existing tools.

By a computational model of identity we mean a system that seeks to infer which entity (person, group, or machine) sent, received or was mentioned in an email. Names and email addresses serve as references to identities that must be resolved, and entities must each be associated with at least one identity. All real models are imperfect, of course; in general, ambiguous and imprecise references will prevent us from determining with certainty even how many entities should be represented in such a model. In this paper, we describe the construction of a simple computational model of identity based on exploiting both system-encoded metadata and regularities that result from habit and social conventions. We begin with a brief survey of prior work on related problems and a description of the salient aspects of the Enron collection. Section 4 then introduces the structure of our model, and Section 5 presents the implementation details. Section 6 is the heart of the paper, assessing the potential utility of the relationships that were actually discovered by our system. Much remains to be done to extend this preliminary exploration, so Section 7 concludes the paper with a brief recap of some of the remaining open issues.

2. RELATED WORK

Research on modeling identity in email collections can draw on a substantial amount of prior work.

Attribute/Association Extraction: Carvalho and Cohen [1] applied machine learning methods to effectively detect signature blocks and quoted text in the Enron collection [8]. We approach this problem using a simpler unsupervised technique and extend it by detecting salutations and nicknames as well. Studies have also used the Web as an external source. Culotta, Bekkerman, and McCallum [4] used the Web to extract contact information for people whose names and email addresses were extracted from email headers. Holzer, Malin and Sweeney [6] proposed a graph-proximity-based technique to determine which email addresses correspond to the same entity. This is exactly the problem of extracting "address-address associations" that we discuss in section 5.2.1. They approached the problem by analyzing the relational network of addresses extracted from Web pages. We restricted our work to the email collection.

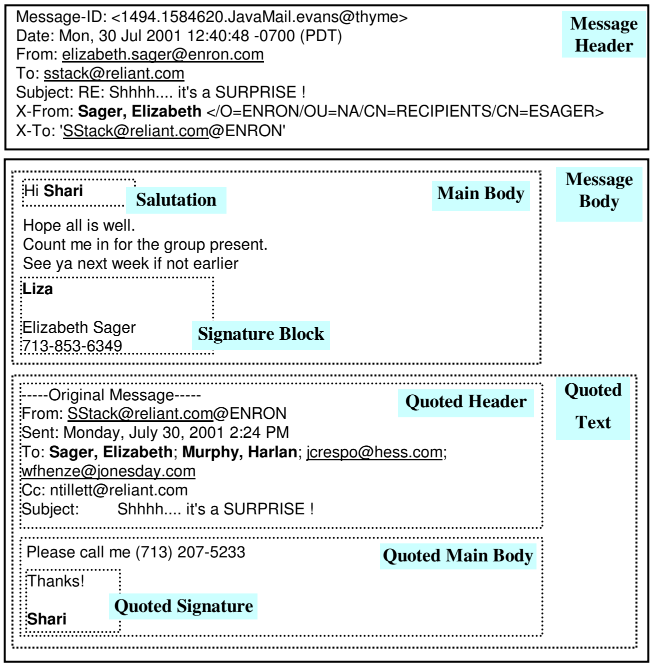
Name recognition and reference resolution: Diehl, Getoor, and Namata [5] used temporal models of email traffic to resolve email name references in a subset of Enron collection. In contrast to our work with the entire collection, they focused only on Enron-domain email addresses. Exploiting name repetition in the email collection, Minkov, Wang, and Cohen [12] proposed a recall enhancing technique for name recognition in email collections.

CEAS 2006 – Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California USA

They used name dictionaries to help train their models. Malin [9] proposed using community similarity rather than exact name similarity to resolve name references in relational networks in which an entity's name can be listed in multiple sources, each with a number of related entity's names. This corresponds to one of the assumptions that underlie our model.

Applications: There are a wide range of research problems that take some form of an identity model as a starting point, the most deeply explored of which is social network analysis. McArthur and Bruza [10] found explicit and implicit connections between people by mining semantic associations inferred from their email communications. McCallum, Corrada-Emmanuel, and Wang [11] proposed a Bayesian network that learns a topic distribution for communication between two entities based on the content of the messages sent between them. Finding experts in social networks has also been studied using a variety of techniques [13, 14]. The Text Retrieval Conference (TREC) also recently introduced an expert finding task using mailing list emails as part of the Enterprise Search track [2]. Finally, Keila, and Skillicorn [7] applied a model based on patterns of word usage to detect deceptive emails in Enron collection.

3. THE ENRON COLLECTION



Message-ID: <1494.1584620.JavaMail.evans@thyme>
Date: Mon, 30 Jul 2001 12:40:48 -0700 (PDT)
From: elizabeth.sager@enron.com
To: sstack@reliant.com
Subject: RE: Shhhh.... it's a SURPRISE !
X-From: **Sager, Elizabeth** </O=ENRON/OU=NA/CN=RECIPIENTS/CN=ESAGER>
X-To: 'SStack@reliant.com@ENRON'

Hi **Shari**

Hope all is well.
Count me in for the group present.
See ya next week if not earlier

Liza

Elizabeth Sager
713-853-6349

-----Original Message-----
From: SStack@reliant.com@ENRON
Sent: Monday, July 30, 2001 2:24 PM
To: **Sager, Elizabeth**; **Murphy, Harlan**; jcrespo@hess.com; wfhenze@jonesday.com
Cc: ntillett@reliant.com
Subject: Shhhh.... it's a SURPRISE !

Please call me (713) 207-5233

Thanks!

Shari

Figure 1. A typical Enron email message.

We consider the Enron collection as a representative collection of large informal communication media with huge number of communicating parties. This collection was first released by the Federal Energy Regulatory Commission (FERC) during the investigation of the former energy trading company. The version of the collection we used[1] includes 517,431 messages in 150 top-level directories; each of which contains the retained emails of a former Enron employee on the date that the collection was obtained.

A typical email message from the Enron collection is shown in figure 1. The message header comprises the basic metadata part of the email and consists of a set of RFC-822 header fields (from, to, cc, bcc, date, and subject) along with special fields that are specific to this collection. In this collection, the typical fields that represent the parties communicating in the email consist only of email address without names. The names then come in separate fields, sometimes unsynchronized with the email addresses in the corresponding RFC-822 fields.

The rest of the message is considered the message body, which may include quoted text if the original was a reply or forward of another. The quoted text may start with a system generated quoted header that acts as the metadata part of a quoted message. The body of the quoted message, "the quoted body," may in turn include another quoted message and so on.

We can further classify the lines of the message main body (excluding any quoted text) into salutation, free text, and signature block. The salutation could appear inline or in a separate line. The free text is considered the actual message that the sender intends to express to the recipients. The signature block may consist of a manually-typed signature (which we call a "free signature") and a relatively static set of system-generated signature lines.

4. IDENTITY FRAMEWORK

In order to be precise about what we mean by a model of identity, we must distinguish between three foundational concepts: (1) a person, (2) an identity, and (3) an entity. By a person, we mean a human who acts in some way that we can observe. Our notion of identity, the focus of this paper, is somewhat more fine-grained. We allow for the fact that one person may construct more than one identity (e.g., striving to completely separate their persona at work from a business that they run from their home trading merchandise on eBay). There may also be some identities that might be adopted by different people at different times (e.g. an assistant sending email on behalf of a manager). It is also possible that an identity might be more closely associated with a machine than any identifiable person (e.g., a periodic message reporting on stock prices). In this paper, we model identities rather than people because that's typically as far as the observable data will take us. Identity modeling is an iterative process, at each stage of which we seek to merge (or split) candidates. For convenience, we refer to these candidate identities as entities.

Identity is at best imperfectly observable in informal communication. We therefore must combine evidence from the available sources and reason based on that evidence if we wish to construct model identity with the greatest possible degree of confidence. We can identify three types of evidence:

1. **Attributes** that characterize observable distinguishing features of that identity. There are two types of these attributes:
   a) ***Personal attributes*** that represent relatively stable explicit features of the identity such as name, email address, and contact information.

[1] Available at http://www.cs.cmu.edu/~enron/

b) ***Behavioral attributes*** that represent key communication features of the identity, such as patterns of communication or selection of discussion topics.

2. **Associations** that can serve as a basis for linking entities together. These associations provide greater support for inference when observed more often.

4.1 Personal Attributes

The most common personal attribute that is available in email collections is the *email address*. Email addresses are particularly useful because they control the routing of email and are thus strongly bound to identity. A second common personal attribute in email collection is the person's *name*. Names are often found in locations (e.g., headers or signature blocks) that make it relatively easy to associate them with email addresses. Names for the same identity can, of course, appear in different forms (e.g. full name, first name, or nickname).

A machine-generated inline *signature block* or vCard attachment can also be considered a personal attribute, acting like a business card of the sender of the email. This can be further decomposed to obtain additional personal attributes such as title, full name, and contact information (e.g., preferred email address, office location, and phone number). Some of these attributes may be time-dependant (e.g., a change in job responsibilities may result in a new title and office location, and perhaps a new email address).

4.2 Behavioral Attributes

A substantial number of behavioral attributes have been proposed, including topic choice, lexical features (e.g., characteristic misspellings or use of emoticons), stylistic features (e.g., whether new and quoted text are typically interleaved, or tendencies towards terseness or verbosity), frequent correspondents (both individually and as groups), conversational initiative, temporal rhythms (e.g., at what times on which days is email being sent), and response times. For example, if email sent to a mailing list generally receives responses from certain set of email addresses, that could provide evidence regarding mailing list membership. With sufficiently large and densely sampled collections, combinations of these attributes can productively be analyzed to identify weak signals among what is otherwise random variation (e.g., administrative assistants may be more likely to initiate meeting scheduling, and more likely to respond to requests for assistance with travel arrangements).

4.3 Associations

Two types of evidence can be used to reasoning about associating two entities, each of which has modeled attributes. Perhaps the most obvious is attribute similarity. For example, if we have "joe.engle@enron.com" as an email address and "Joe Engle" as a name, perhaps this similarity in their personal attributes would support an inference that they refer to the same identity. We can extend this notion of similarity using side information (e.g., our system might know that "Bill" is a common nickname for "William"). Behavioral attributes can be used in a similar way; for example, observing similar topic choices, similar lexical and stylistic features, differing temporal rhythms (e.g., one by day and the other by night) and no direct responses may lead us to believe that two email addresses might be associated with the same identity.

Attribute co-occurrence offers a complementary source of evidence. At its most basic level, association of an email address with any observed personal or behavioral attribute within an email is an example of attribute co-occurrence. Co-occurrence (or omission) of email addresses with the to, cc and bcc header fields is another type of co-occurrence evidence that can sometimes be useful.

In general, the degree to which evidence of association is useful depends on the degree to which it is surprising (e.g., it might not be surprising to find two people names Smith), the degree to which it is reinforced (since random variation will naturally produce some surprising but meaningless coincidences). It is this tension between surprise and reinforcement that makes it necessary to work with large collections if we are to discover interesting associations. Small collections simply lack the potential for multiple instances of rare (and thus surprising events).

4.4 Constructing Entities from Associations

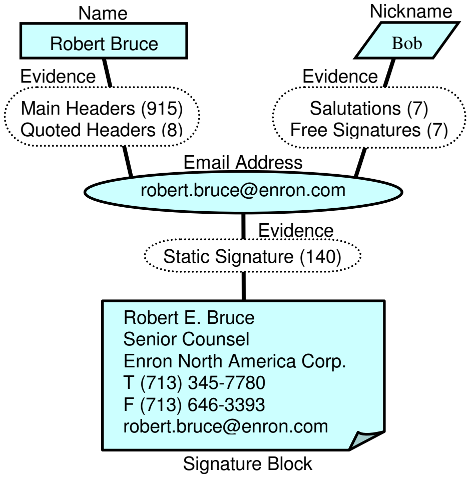


Figure 2. An entity example.

We model an identity as a set of attributes, much in the same way that WordNet models meaning (i.e. a word sense) as a set of words that express that meaning. We can think of any entity as an undirected graph in which the nodes represent attributes and the edges represent associations that are weighted in a way that reflects the strength of the available evidence. In other words, an entity is a component in the attribute graph.

Figure 2 depicts an example from the Enron collection. For this simple example, two attributes were linked whenever a strong co-occurrence association with "robert.bruce@enron.com" is observed. The process could be made iterative, perhaps later merging this entity with others in which a Robert Bruce is also referred to as Bob (but not those in which a Robert Bruce is also referred to as "Rob"). For cases in which sharp associations must be drawn from weak evidence, an interactive process in which automated techniques are used to nominate possible associations for decision by the operator might be a suitable approach. Retaining weak associations without human judgment might, however, be equally useful for cases in which downstream applications are designed to reason effectively under uncertainty.

5. A Simple Identity Resolution Architecture

Figure 3 shows our basic data flow for the experiments reported in this paper. In order to limit the complexity of this first instance of our identity model, we made the following simplifications:

- We treat each email address as if it is associated with a single identity (although we allow an identity to have multiple addresses). This condition is actually violated in the Enron collection (notably with the use of one executive's email address by multiple members of that executive's staff), but such exceptions are rare.
- Heuristics were hand-tuned for the Enron collection. This allowed us to rapidly explore the potential utility of feature sets that might later be used more broadly with machine learning techniques, but without the up front investment in human annotation that supervised learning techniques require.
- Our reliance on hand-coded heuristics led us to focus exclusively on personal rather than behavioral attributes because that is where our intuition was strongest.
- We did not attempt to reconcile multiple identities for a single person, nor did we try to classify identities as machines (which are indeed present in the collection) or people. Our study is therefore focused on the lower levels of identity modeling from which higher-level abstractions might ultimately be built.

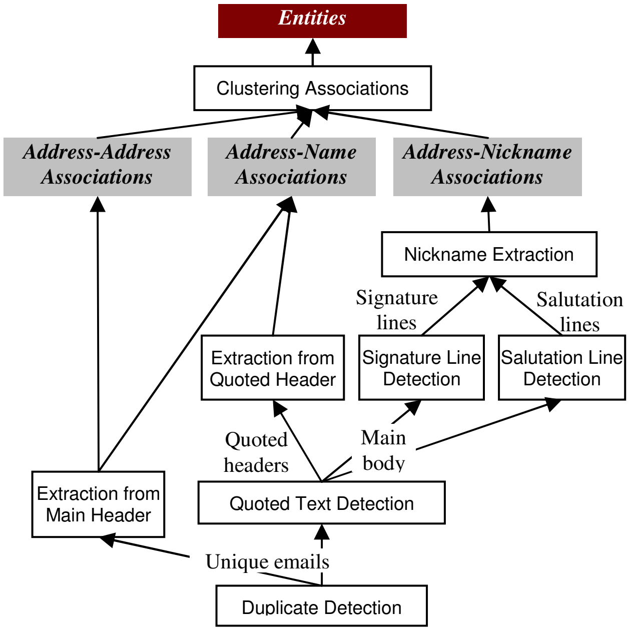


Figure 3. Data flow for identity resolution.

5.1 Duplicate Detection

In our approach, two emails are considered duplicate if they have exactly the same: (1) email addresses of sender and receivers, (2) subject, and (3) body (after being normalized by Lucene's[2] standard tokenizer). This process resulted in detection of 268,980 duplicate emails, about 52% of the whole collection. Subsequent processing was therefore restricted to the remaining 248,451 unique emails. This is just slightly fewer than the 250,484 unique emails found by Corrada-Emmanuel by using an MD5 hash function for duplicate detection [3]. The additional duplicates we discovered could result from inconsequential differences such as date formats, layout differences, or optional header fields to which MD5 is sensitive.

5.2 Identifying Attributes and Associations

5.2.1 Extraction from Main Headers

We have developed a parser for the fields of the main headers of Enron emails. For the sender and recipients headers, we used a set of regular expressions to match names and email addresses. Different forms of names are extracted such as "First Last", "Last, First", and "Last, First MI" sometimes surrounded by single or double quotes. All of these forms are normalized to the first form with optional middle-initial and without any quotation marks.

In the CMU version of Enron collection, the email addresses and the associated full names are separated into two different sets of headers, as shown in figure 1. Email addresses are included in "From, To, Cc and Bcc" headers while corresponding names appear (when they are present at all) in the "X-From, X-To, X-Cc, and X-Bcc" headers. Surprisingly, the number of entries in the address header does not always match the number in the corresponding name header. We therefore rely on attribute similarity to map a name (if possible) to the appropriate email address based on their orthographic similarity.

"Address-name" associations are constructed by mapping a name in a name header to an address in the corresponding address header. Email addresses are sometimes found together with names in the name headers; such cases are also extracted. If an address found in this way in the name header differs from the address in the corresponding address header, an "address-address" association is also constructed. This process results in identification of 70,214 address-name associations and 10,708 address-address associations. Of course, some of these associations could be incorrect, since both attribute detection and orthographic matching employ imperfect heuristics.

5.2.2 Extraction from Quoted Headers

In order to identify the main (original) body of an email message, we have first to detect any quoted text included in that message. The quoted text generally appears in two forms. The first, used generally for forwarded messages (and for replied messages by some email clients), typically consists of an entire message (i.e., the header section of the quoted message in RFC-822 format, followed by the main body of that quoted message) in its original format. In the other common style, used generally for messages that are being replied to, just the body of the quoted message appears (usually with each line marked on the left by some special character such as ">" or "|"). In both cases, the quoted text is normally introduced by a single line that may include information extracted from the header of the message being quoted (e.g., the sender and the date). Quoted messages can themselves be quoted, and the two forms can be interleaved (as could happen if a forwarded message is replied to, for example).

We detect the quotation format normally associated with forwarding by using a set of regular expressions that first match

[2] A retrieval engine available at http://jakarta.apache.org/lucene.

the common forms of the system-generated head line then match the different headers that come next. Because quoted body text is not generally marked in this format, we treat all of the following lines as the body of the quoted message. The process of quoted text detection is then repeated on that body. Detection of quoted body text in the form associated with replies is also performed.

Once all quoted headers have been identified, address-name associations are then extracted from those headers in which email addresses are associated with names (From, To, Cc and Bcc). This process results in the extraction of 11,870 additional address-name associations that were not extracted before from main headers, increasing the relative recall by about 17%. An additional 9,289 address-name associations that were previously extracted from main headers were observed again in the quoted headers. Quoted messages in the CMU Enron collection do not use separate header fields for names, so no address-address associations are extracted from quoted headers.

5.2.3 Signature Line Detection

We define signature lines as lines that often appear similarly formatted near the end of email messages sent from s particular address. Two types of signature lines are of interest: (1) machine-generated static signature lines, from which we can detect signature blocks, and (2) lines containing manually typed free-form signatures lines from which we can later detect nicknames. We identify candidate regions for signature lines by identifying blank lines in the body text in a (possibly quoted) email, using those blank lines to separate block of continuous text, and then focusing on the last two blocks of contiguous text.

For each line in these two blocks, we tokenize the text in that line using Lucene's standard tokenizer and then count the number of lines on which the same pattern of tokens are found. Use of a standard tokenized suppresses some variations that are observed in hand-typed signatures (e.g., "-Dave", "dave", "--Dave", and "Dave"). Tokenized lines with a message count meeting a pre-established threshold (in our experiments, exactly 2 for the "weak evidence" threshold and at least 3 for the "stronger evidence" threshold) are passed to further processing as signature lines. Signature blocks are reconstructed in order from the original (untokenized) lines, while nickname detection (described below) is performed using tokenized lines.

5.2.4 Salutation Line Detection

The process of detecting salutations (brief greetings at the start of a message) is similar to that used to detect signature blocks. Our initial implementation of salutation line detection is quite rudimentary, focusing solely on lines that contain nothing but a salutation. We start by identifying the first line of the main body of every email message in which the recipient's email address is alone in the "To" header (cc to other people are allowed) and in which the message body contains at least two lines. We then filter out lines that start with "fyi" or that end with "?", "!", "." and any line with a length exceeding two words. We normalize what remains using Lucene's standard analyzer and then consider each normalized line that appears exactly three times (for the weakest evidence condition) or at least four times (for the stronger evidence condition) to be a salutation line. Limitations of this process are: (1) salutations embedded in the start of a longer line will not be detected, (2) complex salutation forms (e.g., "hey Bobbie Rae!!") are not detected, and (3) no use is made of evidence found in messages addressed to more than one primary recipient.

5.2.5 Nickname Extraction

We use a precision-oriented approach to detect nicknames from signature and salutation lines. For each known email address, we apply a set of filtering rules to each detected unique signature and salutation line. We first remove a set of hand-selected stop words (e.g., "Hi" or "Dear" for salutations or "Thanks" or "Regards" for signatures). We then filter out any line that includes one or more non-alphabetic characters (after Lucene's normalization, this mainly removes digits), and any line found in a signature block after the fourth line in that block. Lines with more than one word are then compared with the first part of the email address (before the "@") using an edit distance measure, and those with little similarity are discarded. The entirety of any remaining line is then considered to be a nickname, with its frequency in both salutation and signature lines serving as a measure of the strength of evidence for that assignment. A total of 3,151 address-nickname associations were discovered in this way. There are, of course, many ways in which the exhaustivity (i.e., recall) of this process could be improved. For example, cue words (e.g., "Mr.") and name lists are often used in named entity recognition systems.

5.3 Identifying Entities

The foregoing processes resulted in extraction of a total of 82,084 address-name associations, 19,708 address-address associations, and 3,151 address-nickname associations. These associations form links in the undirected graph on which we perform agglomerative clustering to identify components, each of which represents an entity. Because we treat addresses as unique pivot points, address-name and address-nickname links result in no reduction in the number of entities. Address-address associations can, however, connect two disconnected entities (although the accuracy of those associations will naturally depend on the strength of evidence). The complete process resulted in 66,715 entities that together cover 77,420 unique email addresses (58% of 133,581 unique email addresses identified in the collection).

6. EVALUATION

We don't yet have the downstream systems that would make use of a computational model of identity, so constructing an extrinsic evaluation of adequacy for a specific purpose is not presently possible. We therefore chose to perform an intrinsic evaluation of perceived accuracy and utility. There are three levels of extraction to evaluate: attributes, associations, and entities. We focus our evaluation on the associations because we can extrapolate from association accuracy to estimate the accuracy of entity extraction, and because we can incidentally detect errors in attribute extraction when assessing the relationships that an attribute participates in.

The total number of extracted associations (95,943) is far larger than we could assess, so we have to sample that set in some way. We adopted the 12-cell stratified sampling strategy shown in Table 1 to characterize the results based on association type, source of evidence, and strength of evidence. We defined the weakest evidence as the minimum absolute detected strength of evidence (indicated by 1 observation in headers and the threshold values in salutation and signature detectors), and the stronger evidence as all other conditions. For address-address associations, we have only one source of evidence, so we stratify based only on the strength of evidence in that case.

Table 1. Stratified samples and population sizes.

	Weakest Evidence	Stronger Evidence
Address-Name Assoc.		
Main headers only	50 / 29677	50 / 31248
Quoted headers only	50 / 8042	50 / 3828
Both headers	50 / 9289	
Address-Nickname Assoc.		
Salutations only	50 / 272	50 / 465
Signatures only	50 / 172	50 / 1754
Both headers	50/490	
Address-Address Assoc.	50 / 6514	50 / 4194

6.1 Judgment Process

We recruited one independent assessor who has experience with email search system design to judge the accuracy and potential utility of the sampled associations based on the following criteria:

An address-name or address-nickname association is considered *incorrect* if either of the two attributes is incorrectly extracted or if both of them are correctly extracted but linking them is incorrect. Otherwise, the association is considered *correct*. For address-address associations, only the correctness of the linking is assessed.

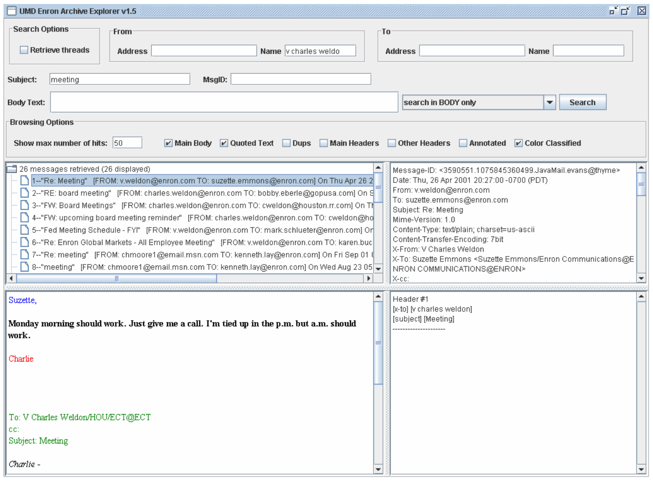
For correct associations, the assessor was asked to further distinguish among three cases (for which artificial examples can be found in Table 2).

1) "*not informative*": if a simple and obvious rule could have been used to construct a name from an email address, or if simple string matching would have indicated that two email addresses were likely for the same person.
2) "*somewhat informative*": if recognizing an association would have been possible, but only with some side knowledge (such as a list of common names).
3) "*very informative*": if the information contained in the name and/or address(es) was not sufficient to reliably infer the association.

Table 2. Examples of different types of correct associations.

Judgment	Association Examples
Correct but not informative	williams.john@enron.com ⇔ "john williams"
	Williams_john@enron.com ⇔ "john"
Correct and somewhat informative	johnwilliams@enron.com ⇔ "john williams"
	williamsjohn@enron.com ⇔ "john"
Correct and very informative	jw@enron.com ⇔ "john"
	happy@enron.com ⇔ "john williams"

To simplify the judgment process, we have developed a java GUI tool that specifically designed to search the Enron collection using Lucene. The interface enables the assessor to search in headers (by either email address and/or name), subject, or body text. She can also restrict the search in the main body, quoted text, or both. The tool displays a ranked list of emails that matches the query, and gives the user a chance to see each result in both raw text or html that can be customized to display any of the different email parts shown earlier in Figure 1. For each email listed in the search results, a list of the extracted attributes and associations is presented as well. The assessor then can use the attributes to generate search queries that may help her finding evidences for the current judged association. A screen shot of the interface is shown in figure 4.



□ Retrieve threads

From — Address: Name: v charles weldo

To — Address: Name:

Subject: meeting MsgID:

Body Text: search in BODY only

Show max number of hits: 50 ☑ Main Body ☑ Quoted Text □ Dups □ Main Headers □ Other Headers □ Annotated ☑ Color Classified

26 messages retrieved (26 displayed)
1--"Re: Meeting" [FROM: v.weldon@enron.com TO: suzette.emmons@enron.com] On Thu Apr 26 2
2--"RE: board meeting" [FROM: charles.weldon@enron.com TO: bobby.eberle@gopusa.com] On S
3--"FW: Board Meetings" [FROM: charles.weldon@enron.com TO: cweldon@houston.rr.com] On Th
4--"FW: upcoming board meeting reminder" [FROM: charles.weldon@enron.com TO: cweldon@ho
5--"Fed Meeting Schedule - FYI" [FROM: v.weldon@enron.com TO: mark.schlueter@enron.com] On
6--"Re: Enron Global Markets - All Employee Meeting" [FROM: v.weldon@enron.com TO: karen.buc
7--"Re: meeting" [FROM: chmoore1@email.msn.com TO: kenneth.lay@enron.com] On Fri Sep 01 (
8--"meeting" [FROM: chmoore1@email.msn.com TO: kenneth.lay@enron.com] On Wed Aug 23 05

Message-ID: <3590551.1075845360499.JavaMail.evans@thyme>
Date: Thu, 26 Apr 2001 20:27:00 -0700 (PDT)
From: v.weldon@enron.com
To: suzette.emmons@enron.com
Subject: Re: Meeting
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: V Charles Weldon
X-To: Suzette Emmons <Suzette Emmons/Enron Communications@E
NRON COMMUNICATIONS@ENRON>
X-cc:

Suzette,

Monday morning should work. Just give me a call. I'm tied up in the p.m. but a.m. should work.

Charlie

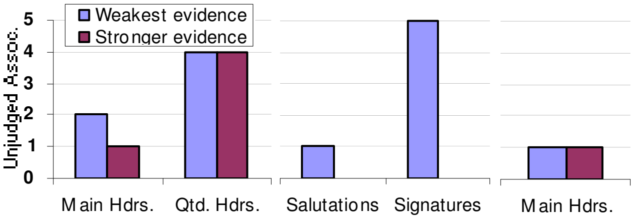
To: V Charles Weldon/HOU/ECT@ECT
cc:
Subject: Meeting

Charlie -

Header #1
[x-to] [v charles weldon]
[subject] [Meeting]

Figure 4. The GUI used in the judgment process

6.2 Results



(a) Address-name (b) Address-nickname (c) Address-address

Figure 5. "Can't Tell" judgments.

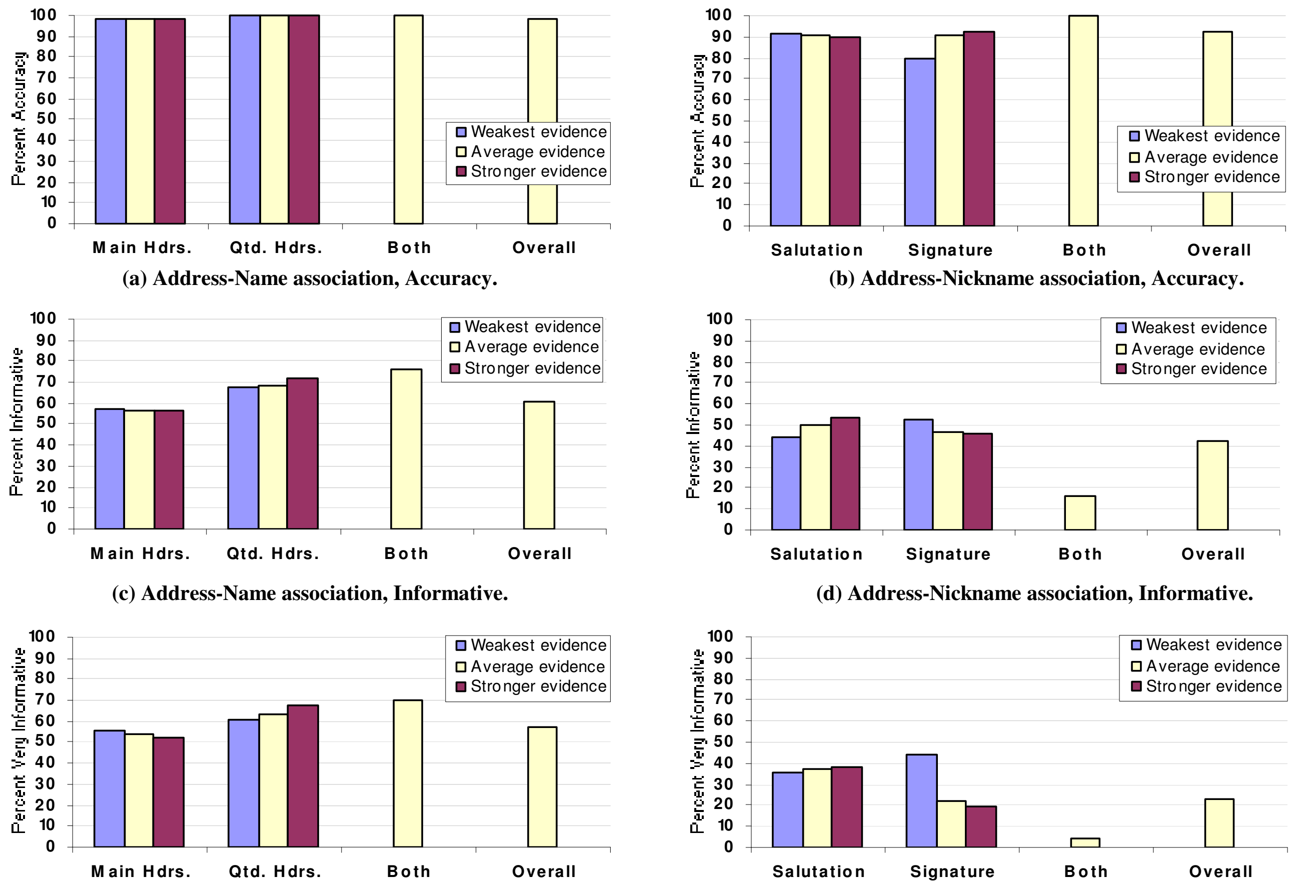
If the assessor found after a while that she could not judge a specific association from the available evidence (including counter-evidence), then they could choose a fifth option: "*Can't tell.*" This occurred in only 19 of 600 cases. As Figure 5 illustrates, stronger evidence never hurt and sometimes helped, and associations found in salutations were surprisingly reliable.

After omitting the unjudged ("can't tell") associations, three measures of performance can be defined:

$$Accuracy = \frac{\mathrm{N}(correct)}{\mathrm{N}(\text{judged associations})} * 100$$

$$Percent\ Informative = \frac{\mathrm{N}(somewhat\ informative) + N(very\ informative)}{\mathrm{N}(\text{judged associations})} * 100$$

$$Percent\ Very\ Informative = \frac{\mathrm{N}(very\ informative)}{\mathrm{N}(\text{judged associations})} * 100$$



(a) Address-Name association, Accuracy.

(b) Address-Nickname association, Accuracy.

(c) Address-Name association, Informative.

(d) Address-Nickname association, Informative.

(e) Address-Name association, Very Informative.

(f) Address-Nickname association, Very Informative.

Figure 6. Evaluation results.

Accuracy is an estimate of the probability that an extracted association will be correct (i.e., precision), regardless of whether the assessor thinks it would be useful for any purpose. Percent Informative is an estimate of the probability that an extracted association will be non-trivial (and thus perhaps useful for subsequent processing by automated techniques). Percent Very Informative is an estimate of the probability that an extracted association would provide new information to a human user.

Figure 6 shows the evaluation results for address-name and address-nickname associations. Each graph in that figure shows one type of association, one metric (on the vertical axis) and each source of evidence (along the horizontal axis). For ease of comparison with the "both" condition (where weakest and stronger evidence are combined) a weighted average is shown between the weakest and stronger evidence bars for the single sources of evidence. Figure 7 shows the results for address-address associations. The three performance measures are plotted side by side in that case since they are all based on just one source of evidence (main headers).

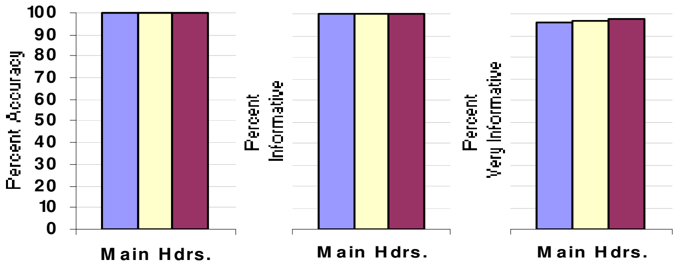


Figure 7. Evaluation results, Address-Address association.

6.2.1 Accuracy

As Figures 6(a) and 6(b) illustrate, 100% accuracy was achieved whenever multiple sources of evidence supported an extracted association, regardless of the strength of each component of that evidence. Address-name association was nearly perfect in every case; while the minimum accuracy in any single source of evidence was 80% (appeared in the case of weakest evidence of signature-based address-nickname associations). This is not surprising, since our approach to address-name association exploits regularities in system behavior, while address-nickname association is based on more variable human behavior. Nonetheless, we can probably improve our address-nickname association accuracy by using supervised machine learning to optimize the extraction process.

As Figure 6(b) illustrates, increasing the strength of evidence improved the accuracy of address-nickname association extraction from signatures, but a similar effect was not evident for salutations. This may result from deficiencies in our process for determining the start of signature blocks. As the weighted average indicates, however, there were relatively few cases with the weakest evidence.

As Figure 7 shows, address-address associations are almost always accurate. If we factor in the overall accuracy of address-name association extraction (squared), single address-address associations would be expected to be completely correct 97% of the time. Inspection of the sample indicates that most address-address associations are for Enron employees. The Microsoft Outlook email system was apparently widely used at Enron, and the routine use of multiple email aliases is characteristic of Outlook.

The average entity includes 1.23 address-name associations, 0.16 address-address associations, and 0.05 address-nickname associations, so we can estimate the overall accuracy of an entity as $(0.98)^{1.23} * (0.97)^{0.16} * (0.92)^{0.05} * 100 = 96.7\%$ (assuming independence between extraction of different associations).

6.2.2 Informativeness

As figures 6(c)–6(f) illustrate, address-nickname associations are generally less informative than address-name associations. This makes sense, since nicknames are usually just one word, while full names typically include two words. Interestingly, nicknames that appear in both salutations and signatures are almost uniformly uninformative. Most nicknames in that category belong to Enron employees. Enron email addresses were usually constructed from their first and last names separated by dot, thus leaving little opportunity for surprise. The opposite is true for address-name associations: when observed in both main and quoted headers, informativeness is higher than when the same type of association is inferred from a single source of evidence. In that case, it turns out that most of the email addresses were from non-Enron domains, for which it is less common to find the full name embedded within the email address.

Surprisingly, Figure 6(f) shows that in the case of evidence from signatures, the most informative associations resulted from the weakest evidence. Since this was exactly the case in which the accuracy was lowest, this points up the importance of considering both accuracy and informativeness. A focused effort to improve the accuracy of address-nickname association extraction from signatures could therefore have a high payoff.

Overall address-address associations were almost always very informative (97%), address-name associations were very informative in more than half the cases (57%), and address-nickname associations were very informative in about a quarter of the cases (23%). This is considerably higher than we had initially expected, suggesting that further work on increasing the accuracy of our extraction, and extending the range of evidence that we can productively exploit would be a good investment.

7. CONCLUSION AND FUTURE WORK

We have described a computational model of identity and assessed its potential utility in the context of one fairly complex email collection. Among the novel features of this technique are automatic detection of nicknames in salutations and signatures. Our approach is relatively simple, and obvious next steps would be to incorporate prior knowledge (e.g. from organization charts), to extend the model to exploit temporal features and behavioral evidence, to integrate (weaker) evidence from name co-reference in addition to address co-reference, and to implement machine learning techniques that optimize a well-defined objective function and that yield confidence measures. We will also need to perform some additional ablation studies to characterize the contributions of each feature to the performance of the overall system. Another thing that we need to do is to characterize the coverage of our methods in more detail. At present, many of our methods (e.g., nickname detection) are intentionally precision-oriented because our heuristics are far from perfect. Adding coverage measures to our analysis would give us a basis for beginning to explore precision-recall tradeoffs.

As additional collections become available, it will be important to replicate this work in other contexts since any single collection can yield only limited insight. The Enron collection was rescued rather than systematically archived, it is from a particular type of organization, from a specific and relatively narrow period of time, and some material has been removed. Larger collections may provide additional evidence, but larger collections are also likely to be more diverse, and thus more challenging. The results we have reported for accuracy and informativeness provide a useful baseline to which future implementations can be compared, and our evaluation design illustrates one way in which such comparisons can be made.

Finally, we ultimately need to broaden our focus to include additional genres (e.g., mailing lists and Usenet) and to integrate these techniques with the ultimate applications for which computational models of identity are needed (e.g., social network analysis). Some obvious next steps are to focus on the more extended conversational interactions that are the natural unit of content analysis, characterizing the needs of real users of these collections, and development of test collections that reflect those needs.

8. ACKNOWLEDGMENTS

The authors are grateful to Yejun Wu for his valuable comments and his help with the judgment process, to Jason Baron, Nick Craswell, Susan Davis, Jim Gallagher, Dan Gallivan, Mara Hemminger, David Kirsch, Anthony Ramirez, Susan Schreibman, Ben Shneiderman and Lee Strickland for the insights that they have shared into user needs, and to Peter Bruza, Lise Getoor, Anton Leuski, Rob McArthur, Adam Perer, Roberta Sabin, Jade Goldstein Stewart, and Yejun Wu for helping to explore technical opportunities to address those needs. This work has been supported in part by the Joint Institute for Knowledge Discovery.

REFERENCES

[1] Carvalho, V. R., and Cohen, W. W. Learning to Extract Signature and Reply Lines from Email. In *Proceedings of the Conference on Email and Anti-Spam 2004*, 2004.

[2] Craswell, N., P. De Vries, A., and Soboroff, I. Overview of the TREC-2005 Enterprise Track. In *Working Notes of Text Retrieval Conference TREC-2005,* 2005.

[3] Corrada-Emmanuel , A. Enron Email Dataset Research. http://ciir.cs.umass.edu/~corrada/enron/index.html

[4] Culotta, A., Bekkerman, R., and McCallum, A. Extracting social networks and contact information from email and the web. In *Proceedings of the Conference on Email and Anti-Spam*, 2004.

[5] Diehl, P.C., Getoor, L., and Namata, G. Name reference resolution in organizational Email Archives. In *SIAM International Conference on Data Mining*, Bethesda, MD , USA, April 20-22, 2006.

[6] Holzer, R., Malin, B., and Sweeney, L. Email alias detection using social network analysis. In *Proceedings of the ACM SIGKDD Workshop on Link Discovery: Issues, Approaches, and Applications*, Chicago, Illinois, USA, August 2005.

[7] Keila, P.S., and Skillicorn, D. Detecting unusual and deceptive communication in email. In *Centers for Advanced Studies Conference*, Richmond Hill, Ontario, Canada, October 17-20 2005.

[8] Klimt, B., and Yang, Y. Introducing the Enron corpus. In *Conference on Email and Anti-Spam*, Mountain view, CA, USA, July 30-31 2004.

[9] Malin, B. Unsupervised name disambiguation via social network similarity. In *SIAM International Conference on Data Mining*, Newport Beach, CA, USA, April 21-23 2005.

[10] McArthur, R., and Bruza, P. Discovery of implicit and explicit connections between people using email utterance. In *Proceedings of the Eighth European Conference of Computer-supported Cooperative Work, Helsinki*, Helsinki, Finland, September 14-17 2003.

[11] McCallum, A., Corrada-Emmanuel, A. and Wang, X. Topic and Role Discovery in Social Networks. IJCAI, 2005.

[12] Minkov, E., Wang, R.C., and Cohen, W.W. Extracting Personal Names from Emails: Applying Named Entity Recognition to Informal Text. *Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing*, October 6-8, 2005, Vancouver, B.C., Canada, 2005.

[13] Schwartz, M., and Wood, D. Discovering shared interests among people using graph analysis of global electronic mail traffic. *Communications of the ACM*, 36:78-89, 1992.

[14] Zhang, J., and Ackerman, M.S. Searching For Expertise in Social Networks: A Simulation of Potential Strategies. In *Proceedings of the 2005 international ACM SIGGROUP*, November 6-9, 2005, Sanibel Island, Florida, USA, 2005.