M.I.T. Laboratory for Computer Science

Request for Comments No. 227
June 10, 1982

An Online M.I.T. Directory and Mail Service
A Research Proposal

by Kimberle Koile

1. Introduction

In order to send electronic mail to someone, one must usually know a computer account name and the name of the computer on which that account resides. If either of these names is not known, sending mail can often be a very frustrating process. And as the number of accounts increases and the number of computers increases, the difficulty of sending mail to a person also increases. How convenient it would be if one could send electronic mail to someone simply by addressing it with that person's name!

The use of personal names in sending electronic mail, at least within an organization the size of M.I.T., is the ultimate goal of this research project. The project will be divided into several phases, the first of which will be the establishment of a directory assistance service for users that are inside and outside the M.I.T. community. In later stages, the service will be extended to include electronic mail forwarding as well. Ideally, the system will be installed on a computer whose sole purpose will be to provide these directory assistance and mail forwarding services.

The services are to be implemented in four phases. As each new phase is implemented, all features provided by earlier phases will remain available to users. Following are descriptions of the services and issues and problems involved in each phase, a discussion of the current state of the project, and a brief comparison of this project to similar projects.

2. Phase 1: Directory Assistance

Service

The directory assistance service will be available to users both inside and outside the M.I.T. community. (A user outside M.I.T. could possibly Telnet from his site to the dedicated M.I.T. computer. No login to the M.I.T. computer would be necessary.) One possible design for invoking the service would be for the user to issue a command such as "whois <name>", where <name> would be last name only, or first initial and last name, or first name and last name, etc. Ideally, the system should be able to handle as many different partial and fragmentary forms of names as possible. The system will then list the record (or records) of information that seems to match the name in the request. In phase 1, all disambiguation will be done by the user. If there are three John Smiths, for example, all three records of information will be listed, and the user will decide which record is the desired one based on other information in the record, such as department, job description, or room number. Thus, in this phase, the system would have two functions: matching exactly an entry name and retrieving information (e.g. an electronic mail address), and translating a name thought to be unique into a possibly unique name (or names) before retrieval of information.

Also in this phase, a "yellow pages" section for laboratory, department, and facility information could be implemented. A yellow pages section for computer services, information, or problems (e.g. Bug-Twenex at MIT-XX) might also prove helpful.

Issues/Problems

1. Who should be included in the directory database? A person may wish the option of not being listed. Should people who have computer accounts but who are not in the M.I.T. phone book be included?

2. What information, besides name and electronic mail address, should be included in the database? How many electronic mail addresses should be included in the database? (A person listed in the directory may wish the option of not having his mail address listed.)

3. Should a person be allowed to have more than one entry in the database? (A person, for example, with more than one job title and/or M.I.T. office is currently listed more than once in the M.I.T. phone directory.)

4. What kind of interface can be maintained between the directory database and the

M.I.T. data sources, both in setting up the database and in updating it? (The data sources include the personnel office, registrar's office, departments and labs, and individuals.) How can inconsistencies in data format be handled? How often should the database be updated and by whom?

5. How can the information in the database be verified? How often should it be verified?

6. How should the database be organized to minimize search time?

7. How much of a name is required to limit system responses to a query to a "reasonable" number? (What is a "reasonable" number?)

8. How will the system deal with nicknames and aliases?

9. How will the system deal with misspellings? Is a phonetic search reasonable?

10. The system should be simple to use. It should be self-teaching for inexperienced users, but not annoying to frequent users.

3. Phase 2: Automation of Disambiguation

Service

Rather than listing all possible responses to a query, the system will ask the user for more information in order to shorten the list to one. The system could request such information as: Is <name> on the faculty or staff? If so, in what department? If a student, what is his course or year? Thus, the user may be saved the tedious search of a (potentially long) list of records. In addition, the system may be able to provide the desired record of information faster than the user could find it by searching the list of all records that match the query. The facility for listing all entries matching the query will remain in the system (in the event that more information is not known).

Issues/Problems

1. How much information is needed to uniquely identify a person?

4. Phase 3: Electronic Mail Forwarding

Service

A user will not have to look up a person's electronic mail address. He will simply send mail to the forwarding site (perhaps named simply "MIT") using the person's name, and the system will automatically look up the corresponding electronic mail address and forward the mail to that address. If no electronic mail address is found in the database, the mail will not be sent. It would be appropriate for the mailing system at the user's site to allow the use of local nicknames for sending mail. A reasonable way to use the system would be the following: The user queries the M.I.T. directory service for information about Kim Koile and finds that the database lists "Kimberle Koile". He can then make an entry in his local database equating "Kim" with "Kimberle Koile at MIT". (Note: equating the nickname in the local database with the name as given in the M.I.T. directory will save the mail forwarding system from having to check for misspellings, nicknames, first and/or middle initials, etc.) When the user sends mail addressed to "Kim", his local mail server will translate the address into "Kimberle Koile at MIT", then send the mail. Upon receipt of the mail, the M.I.T. system will look up "Kimberle Koile" in its database, find that the correct mailing address is "KK at MIT-XX", and will forward the mail to that address.

If a user were to send mail to a person at M.I.T. using a name that did not match exactly a name in the M.I.T. database (e.g. "Kim Koile at MIT"), the system could behave in several possible ways. It could refuse to send the mail and instead send the user a message which would include information about the closest possible match (or matches) and a request to use the exact name listed in the entry (or one of the entries) to resend the mail. If the name could be inexactly matched with one of the entries in the database, the system could send the mail to the associated electronic mail address and send the user a message giving the person's name as listed in the database entry, requesting that future mail be sent using that name.

Note that a person not listed in the database will still be able to receive mail, but the sender will have to know the recipient's computer account name and the site at which it is kept, rather than relying on the forwarding service to respond to the person's name.

Issues/Problems

1. The disambiguation stage will have to be completely reliable. In other words, the mail service will require a unique recipient name before sending mail.

2. What kind of information should the mail originator be sent by the system? Should the user, for example, be notified of the electronic mail address to which his message was sent, or just notified that it was sent?

5. Phase 4: Electronic Mail Forwarding Continued

Service

If no electronic mail address is found for the person to whom the incoming message is addressed, the message could be printed on a "mail service" printer to be delivered via interdepartmental mail.

Issues/Problems

1. How will mail privacy be guaranteed?

2. Who will pay for the printing and delivering?

3. Will the U.S. Postal Service allow this kind of service?

6. Current State of the Project

The database is currently being set up to contain information that is available in the M.I.T. phone books and electronic mail address information. An online file of faculty/staff information (about 7500 records) contains for each entry in the file: name, M.I.T. address and phone number, title, and department. (Home address and phone number were not available for this stage of the project, but they could perhaps be added later if considered necessary information.) We are in the process of putting on line a tape of student records (about 10,000), each of which will contain name, term address, M.I.T. phone number, dorm phone number, private phone number, course, and year. Files of mail addresses/user names for ITS, XX, and Multics are on line, and a simple matching scheme is being implemented to add the mail address information to the file of staff records. (The matching scheme will also be used to add mail address information to the student file.)

In an attempt to determine how much of a name will uniquely identify someone, the following statistics have been collected for the faculty/staff tape.

~ total number of records = 7541

~ total number of records with a first initial = 137

~ total number of records with a first name = 7403

~ total number of records with a middle initial = 5217

~ total number of records with a middle name = 446

~ total number of records with more than one middle name or initial = 280

Also collected were statistics for the number of records that would be ambiguous when only limited information is known. The number of records that cannot be distinguished by:

~ last name = 3380

~ first initial and last name = 639

~ first name and last name = 134

~ first initial, middle initial, and last name = 38

~ first name, middle initial, and last name = 6

~ first initial, middle name, and last name = 0

~ first name, middle name, and last name = 0

Note that the last two statistics may not be significant since only 446 of 7541 records contained middle names. (The statistics seem to indicate that perhaps people list a middle name or initial only when they have found from past experience that it is needed for disambiguation.)

A different way of looking at the same statistics is to calculate what percentage of the time a system response list will be less than, say, 10 entries long using the various naming schemes. This representation is an attempt to give the user an idea of how often he will receive unwanted information ("false alarms") and how often he will receive information ("missed hit").

~ Using last name only:

* 93% of the time the list will be 10 or shorter

* 84% of the time the list will be 5 or shorter

* 70% of the time the list will be 2 or shorter

* 55% of the time the list will be 1

~ Using first initial and last name:

* 100% of the time the list will be 5 or shorter

* 97% of the time the list will be 2 or shorter

* 92% of the time the list will be 1

~ Using first name and last name:

* 99% of the time the list will be 1

Complete statistics collection is proving more difficult than first expected because the statistics are very dependent on the database search algorithm. For example, if a user queries the system for "Frank Stephens" and the database entry is "James F. Stephens", the search will actually resolve to a search on a middle initial and a last name, rather than on a first name and last name. Therefore, expecting the response list to include one entry 99.5% of the time, as in a first name and last name search, may be completely invalid (since all the records that contain last name "Stephens" that don't have a middle initial listed may be potential matches).

In addition, calculation of statistics for the inclusion of a middle name or initial in a query is difficult because not all records in the database contain middle name or middle initial information. For example, if a user queries the system for "David J. Andrews" and the database entry is simply "David Andrews", then the user's added middle initial information does not shorten the search. The search will resolve to one involving only a first name and a last name.

In the "near" future:

~ More complete statistics will be collected for the faculty/staff tape, and statistics will be collected for the student tape.

~ The electronic mail addresses will be correlated with the file of faculty/staff and student records of information.

~ A "simple" search routine will be implemented.

7. Comparison to Similar Work

Other work done on name and mail servers differs from this project in both scale and design goals. The INQUIR system on the XX machine at M.I.T., for example, deals with a database of about 350 entries and only provides a directory assistance service. The Xerox Clearinghouse [1] is a system for naming and locating objects in a distributed environment. Grapevine [2], also developed at Xerox, is primarily a mail system, but also provides for naming, authenticating, and locating people, machines, and services in a distributed environment. Both of these systems are designed to operate in a multi-network environment with databases and services distributed among different machines on different networks. (In our project, a centralized database in a known location seems to be more appropriate.) In addition, both Xerox databases are an order of magnitude smaller than the M.I.T. database (on the order of 10^3 rather than on the order of 10^4). Finally, the Xerox Clearinghouse and Grapevine do not provide the user any assistance in locating a database entry, nor do they attempt to allow communication using a personal name rather than an account name.

The CSNET [3], designed and implemented at the University of Wisconsin--Madison, is a system that provides services such as directory assistance and mail forwarding to "member" institutions, i.e. institutions that are connected to CSNET. It is designed to have a centralized database, as in our project, but one that contains information from all the institutions that use the service. (So if M.I.T. were a member institution, an M.I.T. person's account name could be found by "looking it up" in the CSNET database--at the University of Wisconsin.) CSNET also does not attempt to allow electronic mail to be sent using a personal name rather than an account name.

References

1. D.C. Oppen and Y.K. Dalal, "The Clearinghouse: A Decentralized Agent for Locating Named Objects in a Distributed Environment," Technical Report OPD-T8103, Xerox Office Products Division (October, 1981).

2. A. Birrell, R. Levin, R. Needham, and M.D. Schroeder, "Grapevine: An Exercise in Distributed Computing," Presented at the 8th ACM Symposium on Operating Systems Principles (to appear in CACM).

3. M. Solomon, L.H. Landweber, and D. Neuhengen, "The Design of the CSNET Name Server," Preliminary Report, CSNET Design Note DN-2 (November, 1981).