ATENA: A Distributed Name Guide Server in Electronic Mail System

Lin Yang and Yoshihiko Ebihara

* University of Tsukuba, Tsukuba, Ibaraki 305, Japan
** Institute of Information Sciences and Electronics, University of Tsukuba

Abstract

ATENA (A Telephone and Electronic-mail Name guide system) is a distributed name guide system designed for developing a prototype of new E-mail system. The prototype system is based on CCITT X.500 standard. The goal of ATENA is to provide users with efficient name guide services and to experimentally find ways for constructing integrated network database system. We have provided a model of developing distributed database system and implemented ATENA as a prototype in the environment of UNIX* workstations. The rules which are the foundations of the matching algorithm have been derived to interactively find the true network mail address from ambiguous keys inputted by a user. We also have done the performance evaluations of ATENA with experiments in order to give users an object view of the system.

Key words:
distributed system, integrated directory, matching algorithm.

1. Introduction

Electronic mail systems play a very important role in our society. They are faster and more reliable than the postal mail systems in the real world. So it is possible that the most part of jobs in the postal mail systems would be replaced by E-mail systems in the future. Industry estimates show double digit growth rates for both the number of mailboxes being created and the number of messages being sent per month[1].

In 1988, CCITT and ISO have made out X.500 Standard Which is the standard of new electronic mail systems[2]. The characteristic of X.500 standard is that: Any host and any node in every domain is integrated into the whole network management mechanism which is called Global Directory Services; Not only E-mail but also telephones and facsimiles will be integrated into the global directory services. X.500 compatible backbone network will be implemented in the late of 1990s. When it were realized, we could send E-mail from one place to another in the world with the format addresses and high reliability of transmissions.

The goal for developing ATENA system is to provide an efficient name guide service to users in querying someone's login name or network mail address. The database of network mail address is the same as a telephone-number guide book, but it has been stored distributedly in the whole network and transparently accessible to any user. The first version of ATENA system has been implemented on a LAN(Local Area Network) where there are several workstations connected by Ethernet. In database reference, a flexibal matching algorithm has been implemented and a set of attributes can be applied to the algorithm. The database management was operated by the DBA(Data Base Administrator). Therefore, in application users would pay no attentions to the location of database when they make an inquiry of one's network mail address.

The most important things are global naming facilities in the development of new electronic mail system. In ATENA, we have adopted a proposal of global naming facility which was provided by D. R. Cheriton and T. P. Mann[3]. In the same way, to build up the intelligent man-machine interface, an efficient algorithm of counting the distance between two strings is needed. We will describe them in Chapter 2 and Chapter 4.

Performance evaluations have been performed with experiments as shown later in order to provide an object view of the system. We also discuss why we have selected 5 as the number of candidates which are displayed on the screen and point out the biggest factor of response time.

* UNIX is a trademark of AT&T.
2. The global directory

Designing a global directory is an important and difficult problem in distributed name guide system. User's login name is mostly used as part of a network mail address. There are at least two important problems related to login name. The first, when someone moved to another host, his login name, which would be different to his real name, would not be known by other people. The second one is, when someone uses another login name which is different to his real name, it would be difficult to correctly send him an E-mail.

At present, there are no integrated global directory services system in the world. To build up the international global directory, it needs of worldwide co-operations and organization. We provide here a real global directory services in the experimental environment and describing how to realize it in detail. We have defined three usage classes for describing global directories into three layers: global directory, regional directory and local directory. Each directory should be defined as follows.

Usage_class_0: # for global directory layer#
{ global directory name: country_num. + net_num. + domain_num.
  include: (regional directories name)
  broadcast_to: {global directories name}
  chaining_to: {global directories name}
  reference_to: {DB name} 1.}

Usage_class_1: # for regional directory layer#
{ regional directory name:
  regional directory number: regional number
  include: {local directories name}
  broadcast_to: {regional directories name}
  chaining_to: {global directory}
  reference_to: {DB name} 1.}

Usage_class_2: # for local directory layer#
{ local directory name: host's name in a LAN.
  local directory number: host's IP number
  include: {hosts and servers}
  broadcast_to: {local directories name}
  chaining_to: {regional directory}
  reference_to: {DB name} 1.}

In class 0, global directories are uniquely defined by global directory identifications or by character-string names. They are related with the network of the countries. Subnetworks, which are related with the regional directories, are included in the global directories. The global DB(Data Base) is opened to the whole directory. If there is a query that can't be satisfied in the global directory, it will be broadcasted to the other global directories. The global database is readable to the whole users but it is only modified with some proper authority.

In class 1, regional directories are defined. They are the subnetworks under the global directories, and network servers are generally installed in the regional directories. ATENA server has also been installed here. If there is an inquiry to the ATENA server that can't be satisfied in the regional directory, it will be broadcasted to the other regional directories. If the inquiry is still unsatisfied, then the chaining relation between the regional and global directory will be established, in such a way that the query requirement spread all over the country. The regional DB is accessed by a group of hosts that are geographically or administratively close to each other.

The local directories are defined in class 2. They are geographically close to the users. Local directory's identification is host's identification in a LAN, that is represented by an IP address. The local DB is frequently accessed by users at Local Area Network. Local DB is part of regional database, sometimes it just looks like an aliases list defined by user themselves. Through this distributed database design, the response time will be shortened, compared with that of a centrally controled database management system. The user-interface of ATENA system is installed here. The global directory is shown in Fig. 1.
3. Distributed Database Management System

ATENA system uses a DDBMS (Distributed Database Management System) for local and remote query processings. In our experimental environment, we have build up the distributed database on three local directories of which name are "utissq", "coins" and "gama", and we also have formed a database system on regional directory layer of which is "@is.tsukuba.ac.jp". Fig.2 shows the system configuration. Many commands have been prepared for users to communicate interactively with the system. The query requirements are transferred to a formal message, then the message will be sent to the closest ATENA server going through LAN, at where database reference will be performed. If the requirement was not satisfied at this ATENA server, it will be broadcasted to the other ATENA servers. At there, remote database reference will be performed. The result message will be returned back to the message sender. The name guide facility has been realized based on the global directory which was described above and it is consistent with the requirements of distributed system environment.

Fig.2 System Configuration

Conventional database is centrally managed by DBMS. In application the DB will become enormously large in number of its entries to serve user's needs and queries. All user requirements are arranged to be answered in queue, so response time tends to become larger than that of a distributed database system. On the other hand, if we divided the database into small amount of databases and distributed them to hosts, the quantum of information on each host will become too small to satisfy user's needs. We have provided a proposal which will solve the problem.

We define two types of data. Distributed data is defined as the data that is stored in local database. They are generated by users. They are managed by users, or by host administrators. They have the lowest read and write priority. Local database will be frequently modified. One more type is migrated data which is stored in regional and global layer database. They are migrated from local database to the upper layer database. They are readable to all users, but only writable with some proper authority.

If the query will not be satisfied at one regional directories, it will be broadcasted to other regional directories. If all the regional directories could not satisfy the requirement, the query would be send to the top layer directory—the global directory. Users themselves can define an alias for one person and assign it into a cache buffer named aliases_list. Through the design, ATENA server can shorten the response time of queries.

ATENA server has been currently installed on the three hosts of which each name is "utissq", "coins" and "gama" as a prototype of newly developed E-mail system. In the future, ATENA server will include telephone number guide books and facsimile guide books into database, then various services will be supported for users.

4. Ambiguous Key Reference

In today's database technologies, how to find out correct information quickly is a problem to be solved. In E-mail system, there is an ambiguous situation which is named as "can't send E-mail out until the address is perfectly known", and "with the error address the E-mail can't be sent." We describe here the key reference technique of ATENA name guide server and give you the algorithm we have used in database reference mechanism.

Many matching algorithms have been developed according to reference[4]. We pay attention to the ambiguous character string matching algorithms. We have designed a new matching algorithm based on Non-linear Elastic matching method. Here we introduced rules into database reference mechanism and from doing that we found this methodology is correct and the algorithm is efficient in finding out candidates according to the input string key.

When there are several candidates of network mail address according to the input string key, the appropriate candidates less than 5 will be displayed on the screen and selected by the user himself. We also describe how to calculate the distance between input character-string and reference string registered at the database. The distance is automatically calculated out during the database searching. The rules are described as follows and they are changeable, so it has the

388
characteristic of flexibility.

We define the distance and describe the rules used in calculating distance as follows.

A and B are defined as character-strings, i.e.

\[ A = a_1 a_2 \cdots a_n \text{ and } B = b_1 b_2 \cdots b_n. \]

The distance between A and B is defined as

\[ D(A, B) = d(m, n), \text{ such that} \]

\[ d(i, j) = \min\{ d(i-1, j-1)+\delta, d(i, j-1)+p, d(i-1, j)+q \} + \delta \]

\[ \delta, p, q \text{ and } r \text{ are determined by the following rules:} \]

\[
\begin{align*}
\text{rule 0:} & \quad (\max(i, j) > 2*\min(i, j)) \quad \delta = 2; \quad \delta = 0 \\
\text{rule 1:} & \quad a_a \rightarrow b_{b_i} : (a_i = b_i) \quad p = 0; \quad (a_i \neq b_i) \quad p = 1 \\
\text{rule 2:} & \quad a_i \rightarrow \emptyset : (a_i = b_i) \quad q = 0; \quad (a_i \neq b_i) \quad q = 1 \\
\text{rule 3:} & \quad \emptyset \rightarrow b_i : (a_i = b_i) \quad r = 0; \quad (a_i \neq b_i) \quad r = 1
\end{align*}
\]

\( \emptyset \) is an empty string.

The process of query is executed interactively between user and ATENA user interface. One user can use the "atena" command to enter the dialogue process, input some attributes of someone, then wait for the answer. When there are several candidates are displayed, the user himself can decide to which one the mail is sent.

The procedures used in implementation of the algorithm is as follows.

# This is an algorithm for the evaluation of distance # between two character strings. It can be realized as a # executive program in application #

# as[CHARMAX] is the input string, and # bs[CHARMAX] is the database item string.

\textbf{procedure} dis(strlen(as), strlen(bs)) > 0)

\textbf{begin}

i := strlen(as); j := strlen(bs);
while (i != 1 or j != 1)

\textbf{begin}

del := cal_delta(i, j);
\textbf{end}

\textbf{begin}

\textbf{end}

\textbf{if} (i = 1 \&\& j = 1) then

\begin{verbatim}
\textbf{begin}
p := cal_p(i, j);
q := cal_q(i, j);
r := cal_r(i, j);
dis := \min\{ dis(i-1, j-1) + p, dis(i, j-1) + q, dis(i-1, j) + r \} + del;
\textbf{end}
\textbf{end}
\end{verbatim}

\textbf{end}

5. Experiments

In order to evaluate the efficiency of ATENA system, we have performed several experiments on the number of candidates and the response time. There will be several candidates when a user inquires a network mail address by inputting ambiguous key words. What is the relationship between the number of candidates and distance? How can we fastly calculate out distance? How many candidates have to be find out for display? These are the key problems of a system.

We describe here the environment and conditions of our experiment. There are three hosts of which name are "utissq", "coins" and "gama" as a LAN environment. The database has been made by ourselves and has been distributed over the hosts. At each host, there is about 100 people's, 200 people's and 800 people's information in the database respectively. About 50 times of query experiments have been performed. The distance is calculated in integer. Fig. 3 shows the experimental results.

The average number of candidates of the distance 0,
Fig. 3 Candidates and Distance

The sum of them becomes 3.9. So we have decided that 5 candidates at most are enough for users to select the correct network mail address. It means that as one user uses ATENA system to query for a network mail address, it can be expected that there are 5 candidates at most for selection.

The hit ratio is defined as the number of times that the candidates contain the inquired network mail address. Candidates are searched out in accordance with distance. For distance 0, 1, 2 and above 3, many times of experiments have been performed and Fig. 4 shows the result. We notice that if we calculate the candidates until distance 2, the sum of hit ratio becomes 96%. That means we only have to calculate the distance until 2 in order to decrease response time.

Fig. 4 Hit Ratio for Distance

Now the ATENA server can find out candidates in the order of distance 0, 1 and 2 with the average elapsed time of 0.24 seconds. One part of the time is spent on the jobs of transmitting and receiving message. The other part of it is spent on database reference. The ratio of elapsed time to the calculating distance time is about 1.7:1. If ATENA server are installed on remote host machines, this ratio will become greatly larger than that.

6. Conclusion

We have provided a fundamental idea of constructing new global directory services and we have made out a prototype of it. The system model is described in detail, and functions and algorithms have been discussed. In order to construct a well user-server interface, the ambiguous character-strings matching algorithm is needed. It can be said that the proposed matching algorithm is working well. A user can make clear of unknown network mail address and login name by the dialogue process of ATENA system, and he can correctly send out an E-mail. Measurement results of system performance show that ATENA system works efficiently and successfully.

Acknowledgements

The authors would like to acknowledge the contributions of Associate Professor Sakaguti of Tsukuba University and Associate Professor Tomoo Nakamura of Ryukoku University for their cooperation and encouragement to carry out this research project. Thanks to Mr. Koji Satoh of Sakaguti laboratory in Tsukuba University for his discussions and helps.

References