A System for Extracting and Ranking Name Aliases in Emails

Yin Meijuan
Zhengzhou Information Science and Technology Institute, Zhengzhou 450002, China
Email: raindot_ymj@163.com

Liu Xiaonan,Luo Junyong
Zhengzhou Information Science and Technology Institute, Zhengzhou 450002, China
Email: 97nineday@gmail.com, luojunyong@vip.371.net

Luo Xiangyang
Zhengzhou Information Science and Technology Institute, Zhengzhou 450002, China
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093
Email: xiangyangluo@126.com

***Abstract*—Mining potential information about person identity in emails is one of the popular research topics in email mining. This paper focuses on mining name aliases of a user from emails. Firstly, a system for extracting and ranking name aliases is proposed, which includes two modules: the Alias Extraction Module and the Alias Authority Ranking Module. Secondly, the methods used in the Alias Authority Ranking Module to rank the authority of name aliases of a user are presented in detail, which are based on email communication relation analysis and morphologically similar alias clustering. At last, we evaluate the proposed methods on the public subset of the Enron corpus. Experiment results show that the proposed system can efficiently extract name aliases and find the authoritative aliases of a user.**

***Index Terms*—Email mining, Name alias extraction, Alias authority ranking, Email communication relation analysis, Morphologically similar alias clustering**

I. Introduction

People usually use names different from their real-world names when communicating with each other by email. Names representing identity of email users constantly appear in the header or body of an email. The forms of these names are different, may be formal names and informal names such as anonyms, nicknames, short names and so on, which were called aliases of email users in this paper. Since people often use aliases to express a user in email communication, aliases appearing in emails can indicate users' identity to a certain extent. However, emails transmitted in the Internet may be fictitious and the social identity of an email user may be multifold, which results in that the authenticity and dependability of alias information extracted from emails can't be ensured. Therefore it is necessary to analyze and evaluate the authority of each alias extracted from emails, so as to mine the authoritative alias that is the one best to represent the identity of a user during a period of email communication and in a definite communication scope. Mining potential information about person identity in emails is one of the popular research topics in email mining. This technique can be used in many network applications, such as entity relationship discovery, identity recognition and disambiguation, important person detection, information retrieval and question answering system, and so on.

This paper focuses on the problem of extracting and ranking name aliases of a user from emails, which is a special topic of the research field on entity resolution. Entity resolution is usually to resolve the problem of matching different mentions that refers to one entity in the real world. There are many researches about entity resolution and plentiful approaches have been proposed at present [1], [2]. However, the problem of name alias extraction and alias authority ranking has not received enough attention.

D. Bollegala et al. [3], [4] studied the problem of identifying personal name aliases in web pages. They used the lexical pattern-based approach to extract a large set of candidate aliases from snippets retrieved from a web search engine for a given name, and ranked the candidate aliases by using three approaches: lexical pattern frequency, word co-occurrences in an anchor text graph, and page counts on the web. As messages in email body are unstructured, it is difficult to elicit aliases from email bodies via pattern-based methods. And the ranking scores they defined based on web pages are not fit for name aliases in emails.

As far as entity resolution in emails, several works addressed the problem of resolving name reference resolu-

This paper is based on "Ranking the Authority of Name Aliases for Email Users" by M. Yin, Q. Wang, S. Chen, X. Liu and X. Luo, which appeared in the Proceedings of the 3rd International Conference on Multimedia Information Networking and Security (MINES), Shanghai, China, Dec. 2011.

This work was supported in part by National Natural Science Foundation of China (NSFC) grants 60902102 and 60970141.

doi:10.4304/jsw.8.3.737-745

tion and entity's identity modeling for email user. C. Bird et al. [5] studied the problem of correctly relating aliases and email addresses that belong to the same entity by clustering. They extracted (alias, address) pairs only from the header of emails and clustered them by the similarity between the pairs. C. Diehl et al. [6] firstly explored the problem of resolving personal name references in the full email including the body of emails. They built email communication social network based on the email sender-recipient relationship, and resolved the personal name references by using header-based traffic analysis techniques. But they still extracted name aliases only from the header of emails. T. Elsayed [7], [8] regarded the email address as the key attribute to describe an entity identity, and elicited email address and associated name aliases from email headers and email bodies. They built email communication social network based on the email sender-recipient relationship, and resolved the personal name references by building entity identity model based on email communication relationship. The methods they used to extract aliases from email bodies are very simple and induce low precision. Besides, none of these studies mentions the important problem of evaluating the authority of name aliases for a user in emails.

Referring to the methods of ranking candidate aliases in web pages and quality estimating for web pages [4], we proposed novel methods based on email communication relation analysis and morphologically similar alias clustering to rank name aliases of a user in emails, which is presented in our former work [9]. In this paper, we expand our former work in the following main aspects. Firstly, we present our system for extracting and ranking user name aliases from emails in detail. Then, the key ideas of the ranking method based on email communication relation analysis and the improved ranking method based on morphologically similar alias clustering are described in a more detailed way. At last, we carry out an extra experiment for evaluating the performance of our system.

The remainder of this paper is organized as follows. Section 2 presents our system for extracting and ranking user name aliases from emails and gives the process flows in detail. Section 3 describes the ranking method based on email communication relation analysis, including the definition of several authority indexes for a single alias and the methods to calculate the indexes based on email communication relation analysis. In Section 4, the improved ranking approach to obtain the authoritative aliases of a user by clustering aliases and estimating the authority of an alias cluster is proposed in detail. In Section 5, the proposed methods are evaluated on the public subset of the Enron collection. Results of our approach are concluded in the last section.

II. The System for Extracting and Ranking Name Aliases in Emails

In this paper, a user is represented by the email address extracted from emails, just like in [8], and we ignore the case that a user may have several email addresses in reality to simplify the problem to research. Our alias extracting and ranking system includes two primary modules: the Alias Extraction Module and the Alias Authority Ranking Module, as shown in Fig.1. The framework and elaboration of these two modules are respectively presented in the next two parts.

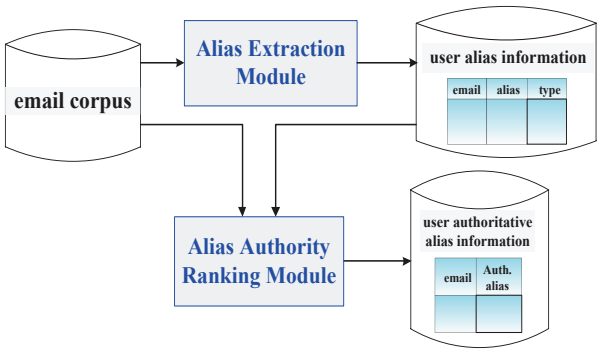

Fig. 1. The Framework of our system.

A. The Alias Extraction Module

Before ranking the authority of aliases of a user, we need first to extract all aliases of the user. The aim of the Alias Extraction Module is extracting all aliases of each user from the given email corpus.

A typical email message is shown in Fig. 2. Aliases of a user mostly appear in the header or body of an email. In email headers, an alias can be elicited from address fields, including "From", "To", "Cc" and "Bcc", and directly associated with the corresponding email address in the same address field. In email bodies, only aliases appearing in the salutation and signature blocks can be directly related to the corresponding email addresses elicited from email headers. So we extract aliases of a user from both email headers and the salutation and signature blocks in email bodies in our system.

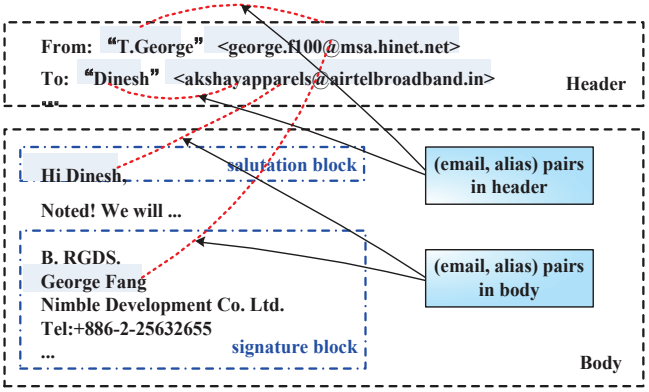

Fig2. A typical email message in the Enron corpus.

The principal process steps of the Alias Extraction Module are illustrated in Fig. 3. For each email in the corpus, we firstly preprocess the email and respectively

get the header part and the body part. Then for the header part, as the email address and its corresponding alias are respectively in the specific marks, the double quotation marks and the brackets, we can easily extract the email address and its corresponding alias by tag matching from the email address field, such as “From”, “To”,“Cc” and so on. And then for the body part, we first locate salutation and signature blocks from the bodies, then extract aliases from texts of these blocks and respectively associate each alias extracted from the salutation block with the email address in the “To” field and each alias extracted from the signature block with the email address in the “From” field. To exactly locate and elicit block texts from bodies of plain-text emails, we proposed Salutation and Signature Block Locating Algorithm based on statistical and rules restricted methods (SSBLA) in our former work [10]. And to efficiently extract aliases in the salutation and signature blocks, we proposed Name Boundary Word Template based Alias Extracting Algorithm (NBWT_AEA) in our former work [11].

B. The Alias Authority Ranking Module

The purpose of the Alias Authority Ranking Module is to rank the authority of different aliases of the same user, which are extracted from the email corpus by the Alias Extraction Module, and to find the most authority aliases to represent the user. An authoritative alias of a user is defined as the alias that is fittest for representing the user’s identity in a period time and a certain scope of email communication, and there can are more than one authoritative aliases for a user.

Our key idea for ranking the authority of user aliases is based on the fact that different aliases of the same user take different roles in the communications with other users by emails. By the relationship of users sending and receiving emails, we can build the email communication networks, which can reflect social relations among peoples. As aliases in emails represent the identities of the corresponding email senders or receivers, the breadth and frequency of an alias used in email communication of a user can directly decide the authority of the alias. Obviously, the bigger the breadth is or the higher the frequency is, the greater the authority of an alias is. Besides, the importance of correspondence object users who communicate with the user using one of his aliases has also important effect on the authority ranking score of the alias. Apparently, the higher importance of the correspondence object user communicating with a user, the greater the authority of the user’s alias used in the communication is. According to above analysis, we can rank the authority of user aliases based on the integrated factor of three types of roles reflected in the user’s email communication relationships, including the breadth, frequency of an alias being used, and the importance of correspondence object users communicating with the user.

The ranking score of the alias authority based on email communication relation analysis can reflect the authority of each alias from the point of email communication, and the alias with the greatest ranking score is regard as the authoritative alias. However, in the ranking results, there is usually a case that some aliases with low ranking scores have similar morphology in linguistics with the aliases of high ranking scores for a user, which is especially common for English names. The reason for this is that these variation aliases are sometimes shortened names, nicknames, or other spelling forms of the authoritative alias. So it’s greatly possible that morphologically similar aliases are affinal names. In practical applications, it is necessary to find various forms of the authoritative alias of a user, for fear of missing important information. Therefore, after having ranked the authority of each alias based on email communication relation analysis, we should also find aliases having similar morphology in linguistics with the authoritative alias, and add them to the authoritative alias set.

Based on above analysis, the principal process flows of the Alias Extraction Module in our system are illustrated in Fig. 4.

In the Alias Authority Ranking Module, we firstly analyze email communicating relationship to count the breadth and frequency of each alias used in email communication of a user. To calculate the importance of correspondence object users communicating with a user, the basic idea of PageRank [12] are introduced into our alias authority ranking method. Then the authority of each alias for the user is calculated according to the alias authority ranking approach, which is presented in the next section. The aliases with high score in the ranking result are only the candidate authoritative aliases of the user. At the same time, we cluster morphologically similar aliases into groups and evaluate the authority of each alias cluster by two factors of aliases in a cluster. At last, we can find the ultimate authoritative alias and authoritative alias cluster by the improved alias authority ranking algorithm based on morphologically similar alias clustering, which is presented in detail in Section 4.

III. Definition and Calculation of Alias Authority Indexes

To rank the authority of an alias, three basic indexes based on the email communication relation analysis are defined: alias using breadth authority index, alias using frequency authority index, and correspondence object importance authority index. By integrating the three basic indexes, the authority index of an alias can be calculated.

A. Alias Using Breadth Authority Index

As the bigger the breadth of an alias N_i used in email communication of a user is, the greater the authority of the alias N_i is, we extend the index of degree measuring the centrality of nodes in social network analysis and define a basic index of ranking alias authority for email users, which is named as alias using breadth authority index (abbreviated as *BA*). The larger the value of index *BA* of N_i is, the greater the authority of N_i is.

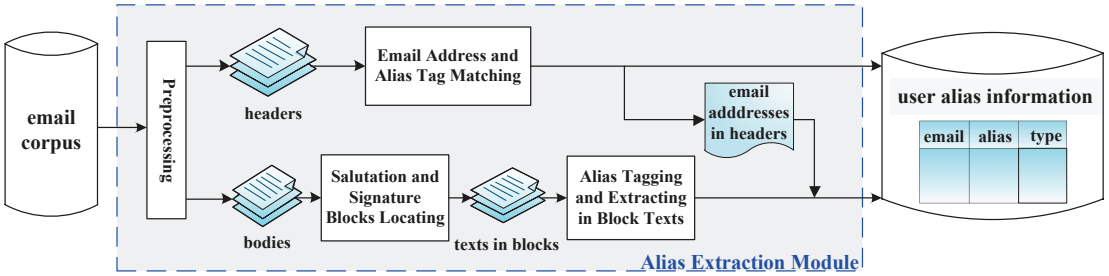

Fig. 3. Process steps of the Alias Extraction Module.

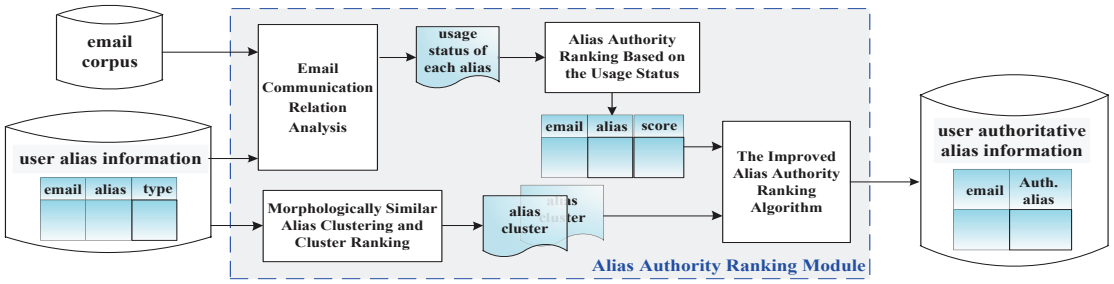

Fig 4. Process steps of the Alias Authority Ranking Module.

For a user with email address E, all of the email addresses having communicated with $E$from the email dataset are elicited and form the set of email addresses $V = \{V_k\}(k = 1, 2, ..., n)$, and then all aliases of E are extracted from emails and construct the set of aliases $N = \{N_i\}(i = 1, 2, ..., m)$. Then, for each alias N_i, the formula to measure the value of index *BA* of N_i is defined as in (1).

$$BA(N_i) = \varphi \times \frac{DegIn(N_i)}{DegIn(E)} + \sigma \times \frac{DegOut(N_i)}{DegOut(E)} \quad (1)$$

Where $DegIn(N_i)$ and $DegOut(N_i)$ are separately the number of N_i that N_i is used in the communication relation of $< V_k, E >$ and $< E, V_k >$, $DegOut(E)$ and $DegIn(E)$ are separately the number of communication relation of $< E, V_k >$ (E sends emails to V_k) and $< V_k, E >$ (V_k sends emails to E), so as to normalize the value of index *BA*. φ and σ are weighting factors, $\varphi + \sigma = 1$ and $\varphi > \sigma$ as the creditability of evaluation from others is usually higher than that of evaluation from itself. In the experiments, the default values are $\varphi = 0.6$ and $\sigma = 0.4$.

B. Alias Using Frequency Authority Index

As an alias with low using breadth but high using frequency should also be assigned a high value of authority, another basic authority index named alias using frequency authority index (abbreviated as *FA*) is defined, which can indicate the using frequency of an alias. The higher the frequency of alias N_i used in email communication of a user is, the larger the value of index *FA* of N_i is, and the greater the authority of N_i is.

$FrqSnd(N_i)$ and $FrqRcv(N_i)$ were separately denoted as the number of emails that N_i is used for E in all emails that including communication relation $< E, V_k >$ and $< V_k, E >$. Since different positions of an alias appearing in emails result in different reliabilities for the same alias, we should treat the frequency of the alias N_i appearing in different position of emails in different weight when computing the value of $FrqSnd(N_i)$ and $FrqRcv(N_i)$. It's well known that the authority of an alias appearing in the salutation or signature of email bodies is higher than that of the same alias appearing in email headers, and that the authority is much higher when the alias appears in both the body and the header of an email. So we can compute the value of $FrqSnd(N_i)$ and $FrqRcv(N_i)$ by more accurate means as follows.

$$FrqSnd(N_i) = \alpha n_{s-only} + \beta n_{sign-only} + \gamma n_{s-both} \quad (2)$$

$$FrqRcv(N_i) = \alpha n_{r-only} + \beta n_{salu-only} + \gamma n_{r-both} \quad (3)$$

In above formulas, n_{s-only} and n_{r-only} are separately the number of emails that appears in "From" Header and the number of emails that N_i appears in "To" Header, n_{sign} and n_{salu} are separately the number of emails that N_i appears in signature block and the number of emails that N_i appears in salutation block, n_{s-both} is the number of emails that N_i co-occurs in both "From" Header and signature block, n_{r-both} is the number of emails that N_i co-occurs in both "To" Header and salutation block. The weighting factor α, β and γ are separately for alias appearing in email header, body, and co-occurrence in both positions, $\alpha + \beta + \gamma = 1$, and can be set different values according to different applications. In our experiments, we set the default values as $\alpha = \beta = 1/4$ and $\gamma = 1/2$, for

we believe that the authority of alias appearing in email header is close to that of alias appearing in email body and the authority of alias co-occurring in both the header and body of an email is higher than that of the former two cases.

Based on the above two formulas, the value of index *FA* for the alias N_i can be calculated by the following expression.

$$FA(N_i) = \varphi \times \frac{FrqSnd(N_i)}{n_s} + \sigma \times \frac{FrqRcv(N_i)}{n_r} \quad (4)$$

where n_s and n_r are separately the number of emails including communication relation $< E, V_k >$ and $< V_k, E >$, so as to normalize the value of index *FA*. The meanings of φ and σ are the same as in (1). As the credibility of aliases of a user as a sender is no difference with that of aliases of the same user as a receiver, the default values are $\varphi = \sigma = 0.5$ in the experiments.

C. Correspondence Object Importance Authority Index

According to the basic theory of the well-known web page quality estimating algorithm PageRank [12] that the authority of a web page depends on the authorities of those web pages linking to this page, there is also some similar dependence between the authority of an alias used by a user and the importance of correspondence object users in email communication of the user. The higher the importance of correspondence object users who communicate with a user using his alias N_i is, the greater the authority of alias N_i of the user is. So to evaluate the authority of an alias, the third basic index is defined, that is the importance of correspondence object users based alias authority index, called object importance authority index (abbreviated as *OA*).

Based on the definition in the previous section, the set of correspondence object users $M = \{B_j\}(j = 1, 2, ..., n)$ using the alias N_i of E is defined when communicating with E. The higher the importance of B_j in set M is, the authority of N_i is. And the steps to compute the value of index *OA* for alias N_i are as follows.

1) *Evaluating the Importance of Each Correspondence Object User* B_j

For each correspondence object user B_j in set M, the importance of B_j can be evaluated by two factors: reply quantity and correspondence frequency. The more emails$E$has replied to B_j and the more frequent Ehave sent to B_j, the higher the importance of B_j is for E. So the method to compute the value of importance of correspondence object B_j (Object Importance, OI) is as follows.

$$OI(B_j) = \omega_1 \cdot \frac{re_{B_j}}{n_{B_j}} + \omega_2 \cdot \frac{f_{B_j}}{f} \quad (5)$$

where n_{B_j} is the quantity of emails that B_j has sent to$E$, and re_{B_j} is the quantity of emails that E has replied to B_j. f_{B_j} is the frequency of$E$ having sent emails to B_j, and f is the total number of emails that E has sent emails to all the users in set M.ω_1 and ω_2 are weighting factor, $\omega_1 + \omega_2 = 1$, and can be set different values according to different applications. In the experiments, the default values are $\omega_1 = \omega_2 = 0.5$.

2) *Computing the Value of Index OA of* N_i

The set of correspondence object users M has influence on the index *OA* of alias N_i in the following three aspects.

- The quantity of elements in M.

The bigger the quantity is, the higher the value of *OA*(N_i) is.

- The importance of each element B_j in $M$for E, that is $OI(B_j)$.

The greater $OI(B_j)$ is, the higher the value of $OA(N_i)$ is.

- The rate of the number of emails communicated between $B_j$and E using alias N_i of E in the number of all emails communicated between B_j and E using any alias of E, defined as $f(B_j)$.

The bigger $f(B_j)$ is, the higher the value of $OA(N_i)$ is. And the method to evaluate is $f(B_j)$as follows.

$$f(B_j) = n_{EB_j}(i) / \sum_{k=1}^{m} n_{EB_j}(k) \quad (6)$$

where $n_{EB_j}(i)$ is the number of emails communicated between B_j and Eusing alias N_i of E, and m is the count of different aliases used in the emails communicated between B_j and E.

So the formula to measure the value of index *OA* for alias N_i is defined as in (7).

$$OA(N_i) = \frac{1}{n} \sum_{j=1}^{n} OI(B_j) \cdot f(B_j) \quad (7)$$

D. Measuring the Authority Index of an Alias

The value of the ultimate Authority Index of alias N_i (abbreviated as *AI*) can be calculated by the weighted average value of above three indexes, as shown in (8).

$$AI(N_i) = \lambda_1 \cdot BA(N_i) + \lambda_2 \cdot FA(N_i) + \lambda_3 \cdot OA(N_i) \quad (8)$$

where λ_1, λ_2 and λ_3 are influencing coefficients, $\lambda_1 + \lambda_2 + \lambda_3 = 1$, can be set proper values according to the demands of different applications, and the default values are $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$.

IV. Improved Alias Authority Ranking Algorithm Based on Morphologically Similar Alias Clustering

The value of authority index of an alias calculated by email communication relation analysis can elementarily judge the authority of the alias. However, the authoritative alias representing the identity of a user in the space of email usually has some variant aliases with relatively high authority. To find the really authoritative alias and its variant aliases, we cluster the aliases by the morphological similarity of affinal aliases with different forms, estimate

the authority of each alias cluster, and evaluate the ultimately authoritative alias and its affinal variant aliases based on the authority index of aliases and the authority of alias clusters.

The following sections describe the methods to cluster affinal variant aliases and evaluate the authority of each cluster, and our alias authority ranking algorithm.

A. Clustering Aliases of a User

The method to cluster aliases is similar to text clustering methods, but as the characteristic of alias is different from that of plain text, text clustering methods are not quite fit for clustering alias. Consequently, a novel agglomerative alias hierarchical clustering algorithm is proposed after analyzing the characteristic of aliases and the performances of various text clustering methods.

Our alias clustering algorithm includes three primary steps: initially, each alias is regarded as a cluster, then repeat finding and merging two most similar clusters of aliases, until all the similarities of each pair of clusters are less than the given threshold, the proper value of which can be trained according to the different application and is assigned as 0.3 in our experiments. As aliases can't be expressed in the form of vectors, which is different from the texts, it's difficult to amend the center of each cluster after merging two clusters during each step of iteration. So the average similarity of each alias pairs between two clusters is taken as the similarity of the centers of the two clusters during each iteration step of clustering.

1) *Computing the Similarity of Two Aliases*

As computing the similarity of two aliases is essentially to evaluate the similarity of two alias strings, the method of Longest Common Subsequence (LCS) is selected, which is fitter for comparing variation alias and the complexity of which is lower in numerous pattern matching methods [13].

The problem of LCS is to find the longest common subsequence of two string sequences. As for alias similarity computing, the algorithm of LCS is iteratively used to find and remove the LCS of two alias strings s and t, until the length of LCS reaches the minimum value of 2 or 3, then add the length of each removed LCS to the total length $\sum LLCS(s,t)$in each iteration, and the similarity of s and t can be calculated by the following formula.

$$sim(s,t) = \frac{\sum LLCS(s,t)}{\max(|s|,|t|)} \tag{9}$$

where $\max(|s|,|t|)$ is the bigger one of the lengths of $s$and t. The larger the total length of removed LCS is, the higher the similarity is. The value of the similarity is from 0 to 1, and the value is 1 when two aliases are just the same and is 0 when there is nothing in common between two alias strings.

2) *Computing Similarity of the alias cluster pair*

The formula to compute the similarity of two clusters is in (10).

$$d = \frac{1}{n_i n_j} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} sim(s_i, s_j) \tag{10}$$

Where s_i is an alias in cluster C_i and s_j is an alias in cluster C_j, n_i and n_j is separately the number of elements in cluster C_i and C_j, $sim(s_i, s_j)$ is the similarity of alias s_i in cluster C_i and alias s_i in cluster C_j calculated by formula (9).

B. Estimating the Authority of Alias Clusters

The authority of an alias cluster can be evaluated by the two factors without regard to the *AI* value of each alias measured in (8).

- *The number of aliases in a cluster.*

If the number of aliases in a cluster is large, then it tells that the original alias of this cluster is extensively used in many different forms. So the authority of the cluster should be high.

- *The frequency of co-occurrence in the same email for aliases in a cluster.*

If different aliases in a cluster appear simultaneously in the header and body of the same email (abbreviated as "co-occurrence"), then it means that the authority of the original alias of the cluster is high. So the larger the co-occurrence frequency of aliases in the cluster is, the higher the authority of the cluster is.

So the formula to evaluate the value of authority $A(x)$ of an alias cluster x is defined in (11).

$$\begin{aligned} A(x) &= \alpha \times \frac{n_{alias}}{\sum_{l} n_{alias}} + (1-\alpha) \\ &\times \Big(\sum_{1 \le i < n_{alias}} \; \sum_{i+1 \le j \le n_{alias}} f_{ij} \Big) / n_{email} \end{aligned} \tag{11}$$

where n_{alias} is the total number of aliases in the cluster x, and l is the number of alias clusters for the same user, f_{ij} is the co-occurrence frequency of alias N_i and N_j in emails related to the same user, n_{email} is the total number of emails that including any alias in the cluster x in emails related to the same user. The weight factor α can be set a proper value by the characteristic of the email corpus and demands of different practical applications, or set an experiential value based on abundant experiments.

C. Algorithm Description of Alias Authority Ranking Based on morphologically similar alias clustering

The primary idea of our alias authority ranking algorithm is that the ultimate authoritative alias and authoritative cluster are evaluated by the results of the authority index of each alias and the authority of each cluster evaluated by formula (8) and (11). If the alias with the biggest authority index value is in the cluster with the highest authority, then the cluster is the ultimate authoritative cluster and the alias with the biggest authority index value is the ultimate authoritative alias. Otherwise, the cluster with the highest authority and the cluster including the alias with the biggest authority index value are compared, and the one in which the average authority of all aliases is larger is the ultimate authoritative cluster and the alias with biggest authority index value in this cluster is the ultimate authoritative alias.

To describe the algorithm, firstly some related definitions are given: the sequence number of alias cluster, x; the sequence number of an alias in a cluster, i and j, A_i and A_j are the corresponding aliases in the cluster x; the frequency of co-occurrence alias pairs in a cluster, f_{cooc}; the authority array of alias clusters $A_C[l]$. The steps of the improved alias authority ranking algorithm based on morphologically similar alias clustering are as follows.

Input: Set C of alias clusters $C = \{C_1, C_2, \cdots, C_l\}$, the alias N_a with the highest authority index value.

Output: Set A of authoritative aliases, $A = \{A_1, A_2, \cdots, A_m\}$, and the authoritative alias A_a.

Begin:

1) for ($x = 1$,$x \leq l$,$x++$)

2) $\{n_x \leftarrow |C_x|$, $|C_x|$ is the number of aliases in C_x;
for($i = 1$, $i \leq n_x$, $i++$)
for($j = x + 1$, $j \leq n_x$, $j++$)
{
for each email in the dataset, if alias A_i and $A_j (A_i, A_j \in C_x)$ are co-occurrence in the same one, then $f_{Cooc}++$;
}
compute $A(x)$ by (8), and$A_C[x-1] \leftarrow A(x)$;
}

3) if $A_C[k] = \max\{A_C[i], 0 \leq i < l\}$ and $N_a \in C_k$, then $\{A_a \leftarrow N_a$; $A \leftarrow C_k$; $\}$

4) else if $N_a \in C_p$, then
{
if $\frac{\sum_{i=1}^{|C_p|} Auth(N_i)}{|C_p|} > \frac{\sum_{i=1}^{|C_k|} Auth(N_i)}{|C_k|}$, then $k \leftarrow p$;
$A \leftarrow C_k$;
$N_i \leftarrow \max\{Auth(N_i), N_i \in C_k\}$; $A_a \leftarrow N_i$;
}

5) return A, A_a;

end.

V. Experimental Results and Analysis

The experiments are carried on the public Enron collection [14] published by Federal Energy Regulatory Commission (FERC) in 2003, and many approaches about email mining are testified on the real email dataset. We select a subset of the Enron dataset which only includes emails the date of which are in half a month from Oct. 7, 2001 to Oct. 22, 2001. The alias ranking algorithm proposed in this paper are implemented in Visual C++, and the MySQL database is used, which was created by J. Shetty to store all important data extracted from the email dataset in 4 tables, including employeelist, message, recipientinfo and referenceinfo table [15].

In our experiments, a user is represented by his email address extracted from emails, and the case that a user may have several email addresses in reality to simplify the problem to research is ignored. We extract the aliases of each user by the methods of our former works [10], [11]. And as the "From" and "To" header fields of each email in Enron dataset don't include any names before email addresses, we extract aliases from the header of the quotation messages, take them as aliases in headers of original messages and associated them with corresponding email addresses in the header of the original messages.

In the experiments we choose emails from the dataset in different periods, such as 2 days, 5 days, 10 days and so on, to test the basic authority indexes, including *BA* index, *FA* index, and *OA* index, and the authority index *AI* of each alias, and our alias authority ranking algorithm. Each of above ranking methods can be solely used according to different applications. The value of those indexes and the cluster of each alias of an example user evaluated by above methods are listed in Table 1, and Table 2 lists names of five users and their authoritative alias and cluster obtained by above methods, In the column of "Authoritative alias cluster ranked by proposed algorithm" of Table 2, the first name is the ultimate authoritative alias.

Precisions of those methods are evaluated by comparing the authoritative alias obtained based on those alias ranking methods with the data of table employeelist. While the precision of alias clusters is evaluated by manually judge whether all morphologically similar alias are grouped into the same cluster and the ultimate authoritative alias is agree with the one in table employeelist. Precisions of ranking methods on different size of datasets are listed in Table 3.

Table 3 shows that precisions of each method are different when carrying out those methods on different size of email subset, and the larger the size of email subset is, the higher the precision is. For the three basic authority index *BA*, *FA* and *OA*, the average precision of the method only based on *BA* index is almost the same as that based on *FA* index, which is higher than that based on *OA* index. And the average precision of the authority index *AI* composed of the three basic indexes is more than 96%, which is higher than that of each basic authority index. As taking into account authority of the alias clusters composed of affinal variant aliases based on alias authority index evaluated by email communication relation analysis, the alias authority ranking algorithm can efficiently find the most authoritative alias and its variant aliases, and its average precision is relatively higher than that of the *AI* index.

Besides, with the continuously increasing of the size of email subset, the average process time of those methods keeps rising. The change trend of the relation between the size of email subset and the average process time of the improved ranking method is illustrated Fig. 5.

Fig. 5 shows that when the methods are applied to email dataset of a large scale, the precision is high but the time consumed by the process can't be endured. So the future work must try to improve the efficiency of the methods. And as the experiments were carried out only on Enron email corpus, in which the formats of emails are relatively

TABLE I
AUTHORITY INDEX VALUE AND CLUSTER NUMBER OF ALIASES OF AN EXAMPLE USER.

Alias	Value of alias authority indexes defined in this paper				Cluster number
	BA	*FA*	*OA*	*AI*	
Rick Buy	0.9404255	0.9629630	0.2900276	0.7311387	1
Rick	0.5787234	0.6666667	0.2459374	0.4971091	1
Rita	0.0595745	0.0833333	0.0173768	0.0534282	2
R. Buy	0.0042553	0.0092593	0.5021277	0.1718807	1
Authoritative alias (or cluster)	Rick Buy	Rick Buy	R. Buy	Rick Buy	1

TABLE II
LISTED NAMES AND AUTHORITATIVE ALIASES AND ALIAS CLUSTERES OF FIVE USERS.

Email address	First name	Last name	Auth. alias by *AI* index	Auth. alias cluster by proposed algorithm
rick.buy@enron.com	Rick	Buy	Rich Buy	Rich Buy; Rick; R. Buy;
rod.hayslett@enron.com	Rod	Hayslett	Rod Hayslett	Rod Hayslett; Rod;
tracy.geaccone@enron.com	Tracy	Geaccone	Tracy Geaccone	Tracy Geaccone;
d.steffes@enron.com	James	Steffes	James D. Steffes	James Steffes; James; James D. Steffes;
lynn.blair@enron.com	Lynn	Blair	Lynn Blair	Lynn Blair; Lynn;

TABLE III
PRECISIONS OF DIFFERENT RANKING METHODS.

Periods (day)	Email quantity	Precision of different ranking methods (%)				
		BA	*FA*	*OA*	*AI*	Proposed algorithm
1	102	90.48	89.93	85.62	91.53	92.07
2	967	93.16	93.84	91.40	95.24	95.48
5	1657	95.28	95.10	92.02	96.16	96.52
7	4276	96.07	95.82	93.89	96.87	97.66
10	6897	96.20	96.36	94.52	97.30	97.93
15	11478	96.61	96.75	94.75	97.94	98.15
Average precision(%)		94.64	94.63	92.03	95.84	96.30

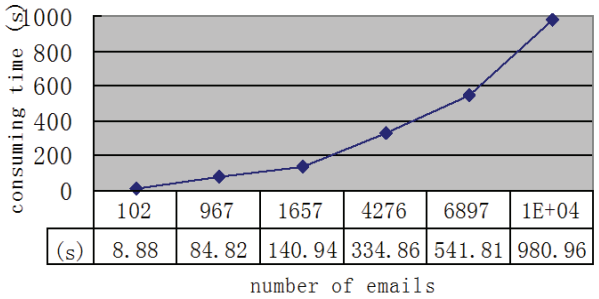

Fig. 5. The relation between the size of email subset and the average process time of the improved ranking method.

normal and aliases embodied in emails are mostly formal names, we should do more experiments to validate the applicability of our methods after we can collect adequate more extensive and ordinary email datasets.

VI. CONCLUSION

In this paper, we addressed the problem of ranking the authority of name aliases of the same user based on emails. The framework and the process flows of our alias extracting and ranking system are described in detail. A novel ranking method based on email communication relation analysis and morphologically similar alias clustering is proposed. The email communication includes three aspects, which are alias using breadth, alias using frequency and the importance of correspondence object users. Experimental results on the public subset of the Enron corpus show that the proposed algorithm can efficiently find the authoritative aliases for an email user, and that the authority index method can also make good results in applications ignoring variant aliases.

The alias authority ranking algorithm is greatly fit for emails with plenty affinal variant aliases, such as emails in Europe language (e.g. English). However, for applications ignoring variant aliases, the authority index method based on email communication relation analysis is the best choice.

Beside, in the period of computing the authority index of each alias and estimating the authority of alias clusters in the proposed method, it is must to scan each email in the dataset, which is time consuming. So when the proposed approach is used for a large-scale email corpus, the high time complexity will become a problem. How to improve the efficiency of the approach will be one of the future researches on alias authority ranking.

REFERENCES

[1] S. Sarawagi and A. Bhamidipaty, "Interactive deduplication using active learning." *In ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2002.
[2] I. Bhattacharya and L. Getoor, "A latent dirichlet model for unsupervised entity resolution." in *In The SIAM International Conference on Data Mining (SIAM-SDM)*, 2006.
[3] D. Bollegala, T. Honma, Y. Matsuo, and M. Ishizuka, "Mining for personal name aliases on the web," in *Proc. of the 17th international conference on World Wide Web*, April 2004.
[4] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Automatic discovery of personal name aliases from the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 831–844, June 2011.
[5] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proc. of the 2006 international workshop on mining software repositories*, 2006, pp. 137–143.

[6] C. Diehl, L. Getoor, and G. Namata, "Name reference resolution in organizational email archives," in *Proc. of SIAM International Conference on Data Mining*, 2006, pp. 77–81.
[7] T. Elsayed, D. W. Oard, and G. Namata, "Resolving personal names in email using context expansion," in *Proc. of Association for Computational Linguistics(ACL)*, June 2008, pp. 941–949.
[8] T. Elsayed, G. Namata, L. Getoor, and D. Oard, "Personal name resolution in email: A heuristic approach," University of Maryland, Tech. Rep. UMIACS LAMP-TR-150, Mar. 2008.
[9] M. Yin, Q. Wang, S. Chen, X. Liu, and X. Luo, "Ranking the authority of name aliases for email users," in *Proc. of the third International Conference on Multimedia Information Networking and Security (MINES)*, Dec. 2011, pp. 425–430.
[10] M. Yin, J. Luo, D. Cao, X. Liu, and M. Li, "Automatically locating salutation and signature blocks in emails," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, vol. 3. IEEE, 2011, pp. 1783–1787.
[11] M. Yin, X. Li, J. Luo, X. Liu, and Y. Tan, "Automatically extracting name alias of user form email," in *The 2nd International Conference on Biomedical Engineering and Computer Science*, April 2011, pp. 315–320.
[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Tech. Rep., 1998.
[13] P. Christen, "A comparison of personal name matching: Techniques and practical issues," in *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, 2006, pp. 290–294.
[14] "The email collection of enron corporation," 2003, http://www.cs.cmu.edu/ enron/.
[15] J. Shetty and J. Adibi, "Enron email dataset," USC Information Sciences Institute, Tech. Rep., 2004, available from http://www.isi.edu/ adibi/Enron/Enron.htm.

Meijuan Yin was born in Anhui Province, China at November, 1977. She was conferred a M.Sc. in computer science by the university of Zhengzhou Information Science and Technology Institute at Zhengzhou, China, in 2003. She is working on the Ph.D. in computer software and academic of the same University. After graduating from the University of Zhengzhou Information Science and Technology Institute at Zhengzhou, China, she became an assistant of the University in 2003 and turned to a lecturer in 2005. Her current research interests include data mining, social network analysis, and information security. Ms. Yin joined China Computer Federation (CCF) as a common member in 2007 and received the IEEE membership in 2010.

Xiaonan Liu was born in Liaoning Province, China. He was conferred a M.Sc. in computer science by the University of Zhengzhou Information Science and Technology Institute at Zhengzhou, China, in 2006. He is working on the Ph.D. in computer software and academic of the same University. He graduated from the University of Zhengzhou Information Science and Technology Institute at Zhengzhou, China, and became an assistant of the University in 2000 and turned to a lecturer in 2006. His research interests include binary translation, compile, and decompile.

Junyong Luo received a M.Sc. in computer science and engineering from the university of Zhengzhou Information Science and Technology Institute at Zhengzhou, China, and became a teacher of the university in 1992. He was developed into a professor and doctoral supervisor of computer science and engineering in 2005. His research has covered many areas, including database, network security, data mining, and information security. His current research projects are on knowledge discovering, social network analysis and parallel computing.

Xiangyang Luo was born in 1978. He received his B. S. degree in computer science, M. S. degree in operational research and Ph. D. degree in computer application technology from Zhengzhou Information Science and Technology Institute in 2001, 2004 and 2010, respectively. He has been with Zhengzhou Information Science and Technology Institute since 2004. From 2006 to 2007, he was a Visiting Scholar of the Department of Computer Science and Technology of Tsinghua University. From 2011, he is a Postdoctoral of Institute of China Electronic System Equipment Engineering Co., Ltd. He is the author or co-author of more than 50 refereed international journal and conference papers. His research interest includes image steganography and steganalysis.