分散型電子メール宛先情報探索における平均探索長の評価†

楊 林†† 海老原 義彦†††

ネットワークの規模が大きくなるほど，電子メールの利用者が増え，電子メール帳システムの果たす役割は重要となる．開発した分散型宛名システムの1つに，曖昧な検索鍵から正しいメールアドレスを探索する機能がある．ここでは，このシステムのモデル化を行い，平均探索長を定量的に求めている．特に，上位のドメイン名は正しいと仮定し，宛先人のユーザ名は正しく，ホスト名が曖昧な場合を探索するケースを対象としている．また，重み付き確率順探索アルゴリズムを提案し，これとランダム探索アルゴリズムを比較し，それぞれの平均探索長を算出した．さらに，平均探索長に与えるネットワーク内のノード数や重み付き確率分布の影響について検討している．解析の結果，重み付き確率順探索アルゴリズムは重み付き確率に片寄りがある場合，大規模ネットワークに対しても有効に機能することを示した．

1. はじめに

ネットワークを介した電子メールはその便利さから，広く利用されている．しかし，ネットワークが大規模になると，電子メールの利用者数が増加し，相手の電子メールの宛先アドレスを的確に把握するのが困難となる．ここでは電子メールの宛先アドレス情報がネットワーク内の各ホストや NIS サーバなどのノードに分散している分散環境を対象にする[1)-4)]．文献5)で開発した会話処理型宛名システム（以下，宛名システムと呼ぶ）では宛先人の曖昧な属性情報を含めて，ユーザ名，ホスト名，所属機関名や部局名などの検索鍵で電子メール宛先アドレスを検索できる．曖昧な属性情報とは，問い合わせ時のミススペリングやうろ覚えによる属性名などを指す．本論文では，相手の電子メールアドレスが登録されている目的ノードが不明な場合，ユーザ名だけでも電子メールの宛先アドレスを検索することができるという宛名システムの1つの機能について焦点を絞り，探索評価を行う．このような機能は問い合わせの頻度は少ないが，サービスの向上という観点から重要である．この場合，ネットワーク内のノードを探索する必要がある．本論文では，ランダム探索アルゴリズムとノード間の距離を考慮した重み付き確率順探索アルゴリズムについてそれぞれの平均探索長を求め，2つの探索アルゴリズムの比較評価を行っている．また，平均探索長に与える重み付き確率やノード数の影響を定量的に検討している．その結果，各ノードの重み付き確率に片寄りがある場合の電子メールの宛先アドレス探索には，ランダム探索アルゴリズムに比べて，重み付き確率順探索アルゴリズムが有効であることを示した．

以下，2章では探索モデルと仮定を述べ，3章でランダム探索と重み付き確率探索の両アルゴリズムについて説明する．4章では探索長の評価に使われるパラメータを定義し，5章で平均探索長を求めている．6章では2つの探索方法による平均探索長について比較し，評価を行っている．

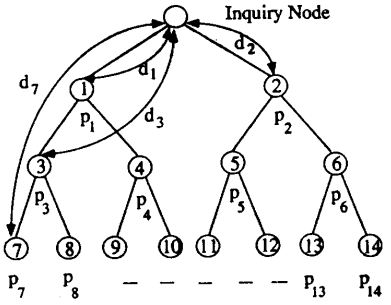
2. 探索モデルと仮定

ネットワークのトポロジーはノード間の距離も含めて与えられているものとする．電子メール利用者が宛先アドレスを問い合わせしているノードを問い合わせノードと呼ぶ．議論を簡単化するため，**図1**に示すような問い合わせノードを根とする木型構造の探索モデルに限定する．宛名システムでは，どのノードも問い合わせノードになりえるが，ここでは1つの問い合わせノードで，かつ問い合わせは一度に1つ処理され，各ノードには問い合わせの待ちキューはないものとする．また，問い合わせの宛先アドレスはある確率で各ノードに存在するものとする．ただし，問い合わせの宛先アドレスはネットワーク内にただ1つあるものとする．この条件下で，平均探索長について評価する．

問い合わせの探索鍵は相手の所属するホスト名とユーザ名からなるものとする．すなわち，上位のドメイン名は正しく入力されていると仮定する．このとき曖昧な検索鍵も含めて鍵の入力状態は**表1**に示されるように4つのケースが考えられる．ケース1と2の場合は，正しく入力された指定ホスト名を持つノードに直接アクセスすればよいが，ケース3と4の場合は

† Evaluation of Mean Search Length on Distributed Electronic Mail Directory Systems by LIN YANG (Program in Engineering Sciences, University of Tsukuba) and YOSHIHIKO EBIHARA (Institute of Information Sciences and Electronics, University of Tsukuba).

†† 筑波大学大学院工学研究科
††† 筑波大学電子・情報工学系



d_i: Distance of Node i from Inquiry Node

P_i: Probability of Node i

図 1 ノード探索の2進木
Fig. 1 Binary tree for node search.

表 1 入力鍵の状態
Table 1 Status of input keys.

Cases	User name	Host name
1	○	○
2	△	○
3	○	△
4	△	△

○ Correct name, △ Ambiguous name.

ネットワーク内のノード探索が必要となる．本論文では，後者のケース3のノード探索を対象とする．

次に，表1のケース3の探索基本動作を説明する．問い合わせを受けた問い合わせノードは，次章で示す探索アルゴリズムに従って，探索すべきノードを選択する．問い合わせノードは検索鍵（ユーザ名）を含む問い合わせメッセージをその選択ノードに送信する．選択されたノードは宛先アドレス情報の検索処理後，検索結果を応答メッセージで問い合わせノードに報告する．この動作を意図する宛先人が見つかるまで繰り返す．同姓同名等の場合は他の属性情報を参照しながら宛先人を特定する．この場合は問い合わせノードと選択されたノード間で何回かの会話のやり取りが行われるが，本モデルでは問い合わせメッセージに対して1回の応答メッセージで済むものとして平均探索長を求める．

3. アルゴリズム

ここでは，宛名システムがもつ代表的な2つのアルゴリズムを説明する．

1. ランダム探索アルゴリズム

2. 重み付き確率順探索アルゴリズム

アルゴリズム1は無差別にノードを選択し，問い合わせの宛先アドレスを探索する方法である．アルゴリズム2は問い合わせの宛先アドレスが存在しうるノード順，すなわち重み付き確率の高いノード順に探索する方法である．

4. パラメータの定義

平均探索長の算出のためのパラメータを定義する．

n：宛先アドレス情報を持つノード数．

d_i：問い合わせノードからノード i $(1 \leqq i \leqq n)$ までの距離．実システムでは回線の物理的距離や問い合わせメッセージ／応答メッセージなどの伝送遅延時間および検索処理時間などが対応する．

D_r：ランダム探索の平均探索長．

D_p：重み付き確率順探索の平均探索長．

p_i：問い合わせ者の意図する宛先アドレスがノード i に存在する確率．

p_i/d_i：問い合わせ者の意図する宛先アドレスがノード i に存在する重み付き確率．探索長は確率ばかりでなく，問い合わせノードからノード i への距離にも依存するので確率 p_i に距離 d_i の逆数で重みを付けている．

5. 平均探索長

重み付き確率順探索の平均探索長とランダム探索の平均探索長を求める．また，2つの探索アルゴリズムの平均探索長の大小関係を求める．

5.1 重み付き確率順探索の平均探索長

最初に，重み付き確率順探索の平均探索長 D_p について求める．重み付き確率とは，4章のパラメータ p_i/d_i の定義に従うものとする．すなわち，問い合わせ者の宛先アドレスがノード i に存在する確率 p_i を距離 d_i で割り算する．

$$p_i/d_i \qquad (1)$$

n 個のノードに対して，式(1)の値の大きい順に並び換え，値の大きい順に若い番号を付けていく．このようにノード番号を再び振り付けても一般性は失われない．この結果，それぞれが次のような関係になったものとする．

$$p_1/d_1 \geqq p_2/d_2 \geqq \cdots \geqq p_n/d_n \qquad (2)$$

重み付き確率順探索アルゴリズムはノード番号の若い順に探索する．今，順に探索していき，i 番目の

ノードに目的の宛先アドレスが存在した場合の探索距離は次式となる.

$$(d_1+d_2+\cdots+d_i)=\sum_{k=1}^{i} d_k \qquad (3)$$

ノード番号 i にある確率は p_i であるので,

$$(d_1+d_2+\cdots+d_i)p_i=p_i\sum_{k=1}^{i} d_k \qquad (4)$$

となる. したがって, 重み付き確率順探索の平均探索長は次のようになる.

$$D_p=\sum_{i=1}^{n} p_i\sum_{k=1}^{i} d_k \qquad (5)$$

5.2 ランダム探索の平均探索長

ランダム探索の平均探索長 D_r を求める. ランダム探索の場合はノード数 n のとき, n までの探索順序数は $n!$ の順序組み合わせ数がある. その組み合わせの1つを次のように表す.

$$(j_1, j_2, \cdots, j_k, \cdots, j_n) \qquad (6)$$

ただし, 上の組み合わせの1つは重複なしの 1, 2, 3, …, n の順列を表す.

このとき, j_1 から探索を始めていって, j_k に意図する目的宛先アドレスが存在したとすると, j_k に至る探索距離は次式で表される.

$$\sum_{t=1}^{k} d_{jt} \qquad (7)$$

n まで考慮すると, 次式となる.

$$\sum_{k=1}^{n} p_{jk}\sum_{t=1}^{k} d_{jt} \qquad (8)$$

さらに, すべての組み合わせを考慮して, 平均すると次式を得る.

$$D_r=1/n!\left(\sum_{\text{すべての組み合わせ}}\sum_{k=1}^{n} p_{jk}\sum_{t=1}^{k} d_{jt}\right) \qquad (9)$$

まとめると, D_r は次式となる (付録参照).

$$D_r=\sum_{i=1}^{n}(1+p_i)d_i/2 \qquad (10)$$

5.3 平均探索長の大小関係

ここでは, 重み付き確率順探索の平均探索長は常にランダム探索の平均探索長より大きくならないことを示す. まず, 2つの平均探索長の差を取る. (5)式と(10)式より,

$$\begin{aligned} D_r-D_p&=\sum_{i=1}^{n}(1+p_i)d_i/2-\sum_{i=1}^{n} p_i\sum_{j=1}^{i} d_j \\ &=1/2[\{(d_1+d_2+\cdots+d_n)p_1+(d_2+d_3+\cdots \\ &\quad +d_n)p_2+\cdots+d_np_n\} \\ &\quad -\{d_1p_1+(d_1+d_2)p_2+\cdots \\ &\quad +(d_1+d_2+\cdots+d_n)p_n\}] \\ &=1/2\sum_{k=1}^{n-1}\sum_{j=k+1}^{n} d_kd_j(p_k/d_k-p_j/d_j) \end{aligned} \qquad (11)$$

任意の k に対して $k<j$ であり, かつ

$$p_k/d_k\geqq p_j/d_j \qquad (12)$$

が成り立つので, 式(11)は負にならない.

6. 比較評価

ランダム探索の平均探索長 D_r と重み付き確率順探索の平均探索長 D_p について比較評価を行う. ここでは, ネットワークトポロジーの1例として図1に示すような2進木を選択する.

最初に, ノード数が両者の平均探索長に与える影響を検討する. 図1の問い合わせノードを除く14個のノードを対象にする. 問い合わせにおいて, まるっきりでたらめなホスト名を入力してきた場合は, 重み付き確率順探索の探索効率はランダム探索と同程度にすぎなくなる. しかし, 曖昧なホスト名の中にはホスト名のミススペリングが1文字または2文字程度の誤りである場合が多く見られるのも事実である. 後者の場合は誤りホスト名によく似たホスト名から探索するのは順当である. 本論文では, 後者の良くありうるケースを考慮して, 目的の宛先アドレスがノード i に存在する確率を, 1例として次のようにする. ただし, λ は $0<\lambda<1$ の係数とする.

$$p_1=1-\sum_{i=2}^{n}\lambda e^{\lambda(1-i)} \qquad (13)$$

$$p_i=\lambda e^{\lambda(1-i)},\quad i=2, 3, \cdots, n \qquad (14)$$

さらに, 問い合わせノードから各ノードまでの距離を次のように定める.

$$\begin{aligned} &(d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, \\ &\ d_{12}, d_{13}, d_{14}) \\ &\quad=(1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3) \end{aligned} \qquad (15)$$

上記のトポロジーをもつ例として, それぞれの階層がゲートウェイで接続された3階層のネットワークが考えられる. 例えば, 図1に問い合わせノードとノード1-2は最上位の光リングに, ノード3-6は中位のイーサネットのセグメントに, ノード7-14はサブセグメント上にある場合などである. また, 上記の例では隣接ノード間の距離が1で構成されている. 4章で距離の定義は伝送時間と検索処理時間の和としている. いま, 伝送時間に比して検索処理時間のほうがかかり, かつ, 各ノードの探索処理時間が同じと仮定すると, 隣接ノード間の距離はほぼ同距離と考えられる.

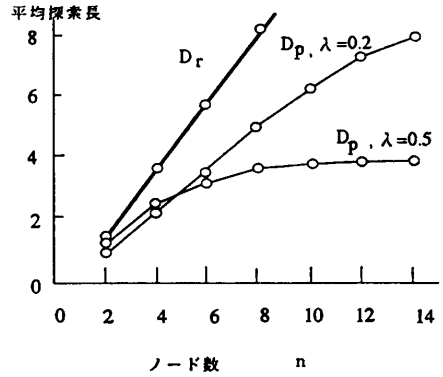


図 2 平均探索長とノード数
Fig. 2 Mean search length v. s. number of nodes.

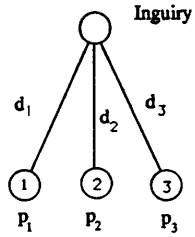


図 3 3ノード構成
Fig. 3 3-Node structure.

このような条件のときの両者のアルゴリズムの平均探索長とノード数 n の関係を**図 2** に示す．D_r はノード数の増加に伴い，平均探索長がノード数 n に比例して線形的に増加する．しかし，D_p はノード数の増加にも係わらずある範囲から飽和状態になる．このことは，式(13)，(14)と(15)の確率と距離の条件を満たす場合，重み付き確率順探索アルゴリズムは多くのノードをもつネットワークであっても有効な探索アルゴリズムであると言える．

次に，距離が両者の平均探索長に与える影響を検討する．**図 3** は宛名アドレス情報をもつ3つのノードかりなり，若いノード番号順に高い確率が順に割り付けてある．すなわち，図3の各確率の大小関係は次のようになる．

$$p_1 \geqq p_2 \geqq p_3 \tag{16}$$

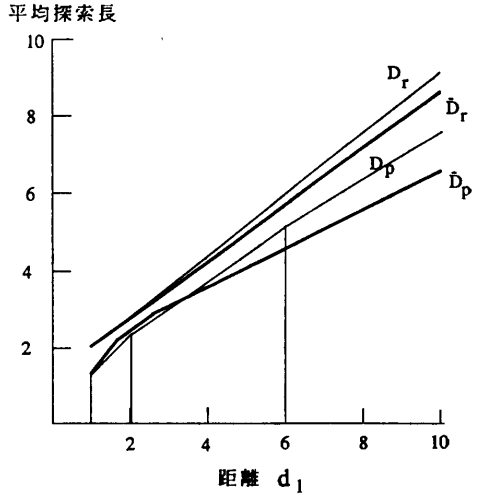
ここで，確率値の変化に対する影響も同時に見るため，2つのケースについて考える．

$$(p_1, p_2, p_3) = (0.6, 0.3, 0.1) \tag{17}$$

$$(p_1, p_2, p_3) = (0.5, 0.3, 0.2) \tag{18}$$

次に問い合わせノードから各ノードまでの距離を1例として下記のように定める．

$$(d_1, 1, 1) \tag{19}$$



細線：$(p_1, p_2, p_3) = (0.6, 0.3, 0.1)$, D_r, D_p
太線：$(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$, $\tilde{D}_r$, $\tilde{D}_p$

図 4 平均探索長と距離の影響
Fig. 4 Effect of node distance on mean search length.

今，距離 d_1 の変化に対して，両アルゴリズムの平均探索長に与える影響を考察する．

このときの両アルゴリズムの平均探索長を**図 4** に示す．図4中の D_p は式(17)の場合であり，このとき式(2)の関係から次のようになる．

$$(0.6/d_1, 0.3, 0.1) \tag{20}$$

この場合は，

$1 \leqq d_1 < 2$ のとき，ノード1→2→3の順，

$2 \leqq d_1 < 6$ のとき，ノード2→1→3の順，

$d_1 \geqq 6$ のときは，ノード2→3→1の順

の探索が良いことを示している．

結論として，図2と図4から次のことが言える．存在する確率が大きく，かつ，問い合わせノードからの距離が短いノードから探すと探索効率が良い．また，式(2)がすべて等号のとき，2つのアルゴリズムの平均探索長の値は等しくなる．

7. ま と め

ユーザ名を探索鍵として目的の宛先アドレスを探索するランダム探索と重み付き確率探索アルゴリズムの平均探索長を算出した．算出結果より，重み付き確率探索アルゴリズムの平均探索長は，常にランダム探索アルゴリズムの平均探索長より大きくならないことを示した．また，ノード数や距離が両者の平均探索長に与える影響を定量的に求めた．特に，重み付き確率間

に片寄りがあるときは大規模なネットワークであっても，重み付き確率探索アルゴリズムは有効に機能することを示した．

次に，重み付き確率探索を現実システムに適用するには，少なくとも次の手順が必要と思われる．ただし，距離 d_i は実システムの測定から得られているものとする．

1. 問い合わせノードは曖昧なホスト名から有限個の候補ノードを抽出する．たとえば，レーベンシュタインの文字列間距離[6]などの手法により，曖昧なホスト名と候補ノード i のホスト名の距離 l_i を算出する．

2. 距離 l_i の短い候補ノードほど，そのノードに高い確率を与えるよう，ある関数 $p_i=f(l_i)$ を施す．

3. p_i/d_i を計算し，候補ノードの探索順序を決める．

関数 f の求め方は種々考えられるが，例としては，次式があげられる．

$$p_i=l_{n+1-i}\Big/\sum_{i=1}^{n} l_i,\quad i=1,2,\cdots,n$$

または，問い合わせノードに学習機能を持たせ，ある期間距離 l_i に対する発生確率 p_i を測定し，関数 f を動的に求める方法も考えられる．

参考文献

1) Kerr, S.: The Coming Global Directory for E-mail, *Datamation*, Vol. 35, No. 22, pp. 105-108 (1989).
2) 宮内直人，中川路哲男，勝山光太郎，水野忠則：OSI ディレクトリの実現，情報処理学会マルチメディアと分散処理研究会資料，40-3 (1989).
3) 田中啓介，井田昌之：Apostle の名前管理機構，情報処理学会マルチメディアと分散処理研究会資料，40-4 (1989).
4) 古宇田フミ子，田中英彦：分散環境における名前管理システム，情報処理学会論文誌，Vol. 29, No. 10, pp. 975-984 (1988).
5) Yang, L. and Ebihara, Y.: ATENA: A Distributed Name Guide Server in Electronic Mail System, *Proc. of International Symposium on Database Systems for Advanced Applications*, pp. 386-390 (1991).
6) Okuda, T., Tanaka, E. and Kasai, T.: A Method for the Correction of Garbled Words Based on the Levenshtein Metric, *IEEE Trans. Comp.*, Vol. C-25, No. 2, pp. 172-178 (1976).

付　録

(10)式を次のように表す．

$$D_{r,n}=\sum_{i=1}^{n}(1+p_i)d_i/2 \tag{21}$$

上式を帰納法により証明する．

$n=2$ のとき，すなわち，図3のノード3がない場合の平均探索長を求める．探索パスはノード1→2の順で探すやり方とノード2→1とする2通りある．

まず，ノード1→2の探索パスを考える．このときノード1に目的のアドレスがある場合は探索距離は d_1 である．ノード2にある場合は探索距離 (d_1+d_2) である．したがって，ノード1→2の平均探索距離は，

$$p_1d_1+p_2(d_1+d_2) \tag{22}$$

となる．

同様に，ノード2→1の平均探索距離は，

$$p_2d_2+p_1(d_1+d_2) \tag{23}$$

である．ランダム探索は2つの探索パスを同確率で選択するので，$D_{r,2}$ は次式となる．

$$D_{r,2}=\{(22)+(23)\}/2$$
$$=(1+p_1)d_1/2+(1+p_2)d_2/2$$

この結果は(21)式に $n=2$ を代入した結果と一致する．

$n=(k-1)$ のとき，次式が成り立つものと仮定する．

$$D_{r,(k-1)}=\sum_{i=1}^{(k-1)}(1+p_i)d_i/2$$

ただし，$k\geqq 3$ の整数とする．

$n=k$ のとき，$(k-1)$ 個のノード群の平均探索距離は $D_{r,(k-1)}$ である．ここで，$(k-1)$ 個のノード群をまとめて，1つのノード a で代表する．$n=k$ のときは $D_{r,(k-1)}$ の距離をもつノード a に，確率 p_k と距離 d_k をもつ k 番目のノード b が新たにネットワークに加わったことと同じである．$n=2$ のときと同様に，探索パスはノード a→b の順で探すやり方とノード b→a とする2通りある．

まず，ノード a→b の探索パスを考える．このときノード a に目的のアドレスがある場合は探索距離は $D_{r,(k-1)}$ である．ノード b にある場合は探索距離 $(d_k+D_{r,(k-1)})$ である．したがって，ノード a→b の平均探索距離は，

$$(1-p_k)D_{r,(k-1)}+p_k(d_k+D_{r,(k-1)}) \tag{24}$$

となる．

同様に，ノード b→a の平均探索距離は，

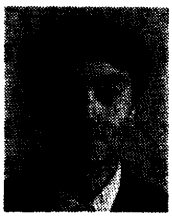
$$p_kd_k+(1-p_k)(d_k+D_{r,(k-1)}) \tag{25}$$

である．ランダム探索は2つの探索パスを同確率で選択するので，$D_{r,k}$ は次式となる．

$$D_{r,k}=\{(24)+(25)\}/2$$
$$=\sum_{i=1}^{k}(1+p_i)d_i/2$$

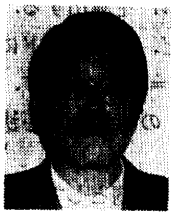
結果は(21)式と一致する．証明終わり．

(平成4年2月 10 日受付)
(平成4年9月 10 日採録)



楊　　林（正会員）

昭和37年生．昭和59年中国北京計算機学院軟件工程系卒業．平成3年筑波大学大学院理工学研究科修士課程修了．現在，同大学院工学研究科博士課程在学．計算機網のネットワーク・アーキテクチャ，データベースシステム，デジタル通信システムおよびシステム性能評価の研究に興味がある．



海老原義彦（正会員）

昭和22年生．昭和45年東北大学工学部電子工学科卒業．昭和50年同大学院工学研究科博士課程修了．工学博士．現在，筑波大学電子・情報工学系助教授．主たる研究分野は，計算機網ネットワーク・アーキテクチャ，デジタル通信システムおよびシステム性能評価など．