非公開データを用いた人物検索システムの開発

山元　潤† 森嶋　厚行‡

筑波大学図書館情報専門学群† 筑波大学大学院 図書館情報メディア研究科‡

1. はじめに

近年，論文などのドキュメント集合を元に，指定されたトピックに詳しい人物を検索する Expert Finding[1] の問題が注目を集めている．Expert Finding は，トピックを表す語に関連度の高い人のランキングを行うが，Web 等で公開された情報を入力ドキュメントとして用いることが多い[2]．すなわち，Web に公開するようなドキュメントをもつ専門家が主な対象である．

本稿では，Web に情報を公開するような専門家ではない人を対象とした人物検索の問題を議論する．前提として，それらの人はある組織内に属し，外に公開するような情報は持たないものの，それらの組織内でのファイル操作ログや PC に格納されたファイル，やりとりされたメールなどの，一般には公開しない情報を利用できるとする．

これらのデータは，公開を前提としたドキュメントと著者との関係と異なり，人の専門性の関係が必ずしもハッキリしているとは限らない．一方，操作ログなどの一般には公開されないより詳細な情報を利用すると，それらを用いた効果的な人の発見が出来る可能性がある．

本稿では，論文[1] で述べられた検索モデルを用いてこの問題にアプローチする．本手法ではドキュメントと人の関連度を表すテーブルを用意し，それを用いて人物検索を行う．本稿では，ナイーブな方法によりこのテーブルを作成した場合と，組織内の操作ログや ML 等の情報を考慮してテーブルを作成した場合の結果の比較を行い議論する．

2. Expert Finding

Expert Finding の問題は，ドキュメントの集合を元に，与えられた語に関連する人の名前を検索し，ランク付けした結果を返すものである．

図 1 は，Expert Finding の処理を図示したものである．まず，前処理として各ドキュメント $d \in D$ と，候補者 (人) $ca \in CA$ に対して，それらの関連 (Document-Candidate Associations) $a(d, ca)$ を算出する．つぎに，ユーザがクエリ (トピックを表す語)q を入力すると，それらの関連度を利用して人物検索とランキングを行い，結果を出力する．

2.1 人物検索モデル

本節では，ドキュメントモデルを利用した人物検索モデル[1] を説明する．一般に，人物検索モデルでは，クエリ q と関連の高い人を求めるために，クエリ q と検索対象となる人 ca $(\in CA)$ の条件付確率 $p(ca|q)$ を計算し，この確率に従って人をランキングする．この条件付確率の式をベイズの定理に基づき変換すると，次の式が得られる．

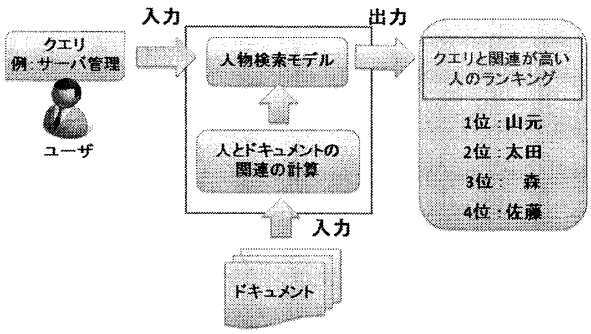

図 1　Expert Finding の処理

$$p(ca|q) = \frac{p(q|ca)p(ca)}{p(q)}$$

ここで CA のサイズは一定かつ q は与えられるため，$p(ca)$ と $p(q)$ は固定になる．したがって，$p(q|ca)$ を求めれば，$p(ca|q)$ のランキングを決めることができる．$p(q|ca)$ を求める式は，次で与えられる．

$$p(q|ca) = \sum_{d \in D} p(q|d)p(d|ca)$$

$p(q|d)$ は，ドキュメント d においてクエリ q が含まれる確率である．また $p(d|ca)$ は，ドキュメントと人の関連度を確率にしたものであり，次に説明を行う．

2.2 ドキュメントと人の関連

この人物検索モデルにおいては，前処理としてドキュメント d と人 ca の関連度 $a(d, ca) > 0$ を，ドキュメント d と人 ca に関連があることを示す値として与えられていることを前提とする．例えば，ドキュメント $d1$ の作成者が山元であれば，$a(d1, \text{山元}) = 1$ とするといったことである．この $a(d, ca)$ から $p(d|ca)$ を次のようにして計算する．

$$p(d|ca) = \frac{a(d, ca)}{\sum_{d' \in D} a(d', ca)'}$$

3. 組織内のログやファイルを用いた人物検索

前述したとおり，本稿では，通常は公開されていないような組織内の操作ログやメール，PC 内のファイルを利用して Expert Finding を行おうとするものである．したがって，これらの情報から $a(d, ca)$ のテーブルを作成することが問題となる．

3.1 利用する情報

(1) ファイル操作ログとファイル． 組織内における PC のファイル操作ログを利用する．現在，そのようなログを管理するためのツールは各種入手可能であるが，本稿では，InfoSpace Plug[3] を用いて入手したログを利用する．ただし，本稿の議論はツールと独立して成立する．次の操作がログに存在すると仮定する．(1) ファイルへの書き込み (2) ファイルの生成 (3) ファイルからの読み出し (4) ファイルへのパス (5) ファイルへ書き込みがあったアカウントと日時．

また，組織内のルールの範囲で，ログから参照されてい

Development of a people search system using unpublished data
Jun Yamamoto† Atsuyuki Morishima‡
Sch. of Library and Information Science, Univ. of Tsukuba.†
Grad. Sch. of Library Information and Media Studies, Univ. of Tsukuba.‡

るPC上のファイルは全て利用可能とする．そのファイルの作成者，更新日などの情報等も入手し利用する．

(2) メールログとメール． 組織内におけるメールのやりとりが，特定のメールサーバを通じて行われている状況を想定する．したがって，メールのやりとりのログの他，メールの本文も参照することが可能である．

3.2 ドキュメントと人の関連の計算

本章では，3.1節で述べたような情報から，ドキュメントと人の関連度 $a(d, ca)$ を算出する方法について議論する．具体的には，(1) ナイーブな方法 (手法 A) と (2) 各種ログやファイル，メールの性質を考慮した方法 (手法 B) の2つを検討する．

(1) 手法 A：ナイーブな方法

ドキュメント集合 D を，ファイル操作ログから参照された各ファイルとメールサーバにあるメールの集合とし，$d \in D$ に対して下記のように $a(d, ca)$ を定義する．すなわち，d がファイルの場合，人 ca がファイル d にアクセスした記録がファイル操作ログに存在する場合，$a(d, ca) = 1$ とし，メールの場合，メールの to，from，cc に含まれる ca に対して $a(d, ca) = 1$ とする (表1)．メーリングリストの場合は，組織内では ML のメンバが既知であることから，その情報を利用する．

表1 手法 A における $a(d, ca)$ の設定

ファイル

関連のある人	$a(d, ca)$
アクセス者	1

メール

関連のある人	$a(d, ca)$
送信者	1
受信者	1

(2) 手法 B：各種ログ・ファイル・メールの性質を利用

手法 A と同じ $d \in D$ を利用し，$a(d, ca)$ を定義する．まず，d がファイルの場合，ファイルの種類によっては，ファイルのメタデータにファイルの作成者情報が入手出来ることに着目し，作成者の方がアクセス者より関連が高いと見なす．また，アクセス者に関しては，ファイルへのアクセス回数を考慮し，回数が多くなるにつれてファイルとの関連を高くする (表2(左))．

次に d がメーリングリスト宛以外のメールの場合，送信者の方がメールにおける関連が高いと見なし，表2(右) のように $a(d, ca)$ を設定する．

さらに，メールの宛先がメーリングリストの場合，通常のメールと比べ，受信者は関連が低いと考えられる．したがって，表3のように $a(d, ca)$ を設定する．ただし，メール本文の先頭 (今回は先頭5行) に，メーリングリストに登録している人の名前やアドレスなどの情報が含まれている場合にはその関連が高いとみなし，通常のメールの受信者と同一の扱いとする．

4. 予備実験

手法 A と手法 B を比較するための予備実験を行った．実験対象は，筑波大学のある研究室で記録していた各 PC のファイル操作ログとメールログ，それらに関連するファイルとメールである．ファイル数は 914，メール数は 415 である．

実験の対象とするクエリ Q は次のように作成した．すなわち，まず研究室内の構成員 $ca \in CA$(11名) に対し，彼らに

表2 手法 B における $a(d, ca)$ の設定

ファイル

関連のある人	$a(d, ca)$
作成者	1
アクセス者	$1 - \frac{1}{n+1}$

※ n:ファイルにアクセスした回数

メール

関連のある人	$a(d, ca)$
送信者	1
受信者	0.5

表3 手法 B におけるメーリングリストの $a(d, ca)$ の設定

関連のある人	$a(d, ca)$
送信者	1
受信者	0.1
メールの内容に識別できる情報がある受信者	0.5

関連が高いと予想される用語の集合 Q_1 を作成した．つぎに，Q_1 と，D 内に11個以上出現した単語の集合 Q_2 の積集合 $Q' = Q_1 \cap Q_2$ を求め，Q' から無作為に10個を選択し Q とした．また，正解ランキングを次のように作成した．すなわち，研究室内の構成員のうち8名に $q \in Q$ における候補者のランク付けを行ってもらったものを正解のランキングとした．つまり，8つの正解ランキングがあることになる．各正解ランキングでは同順位に複数の候補者を許した．次に，8つの正解ランキングからそれぞれ $AnsPairs_i = \{(ca_i, ca_j, point), \ldots\}$ $(1 \leq i \leq 8)$ を作成した．これは，各正解ランキングに含まれる人の全ての順序関係 $ca_i < ca_j$ に対して $(ca_i, ca_j, point)$ を作成したものである．$point$ は，正解ランキング中で ca_i が k 位の時 $point = 6 - k$ とする．これは上位のランキングが重要であるためである．ca_i が6位以降は全て $point = 0$ とする．

手法 A もしくは B で作成したランキングからも $AnsPairs$ と同様に $Pairs$ を作成し，$AnsPairs_i (1 \leq i \leq 8)$ とどれぐらい似ているかを評価するため，次の式により値を計算した．

$$\sum_i \left(\frac{\sum_{p \in AnsPairs_i \cap Pairs} point(p)}{\sum_{p \in AnsPairs_i} point(p)} \right) \cdot \frac{1}{8}$$

その結果，手法 A と B に関してそれぞれ61.8%，70.5%という値となり，8.7%の差があった．

5. おわりに

本稿では，組織内のログやファイルなどの非公開データを用いた人物検索システムの開発ついて述べた．今後の課題として，既存のドキュメントを用いるだけではなく，様々なログから疑似ドキュメントを作るなどのアプローチの検討等があげられる．

謝辞

本研究の一部は科学研究費補助金若手研究 (B)(#20700076) による．

参考文献

1) K.Balog, L.Azzopardi, M.de Rijke. Formal models for expert finding in enterprise corpora, In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp.43-50, 2006.
2) K.Balog, I.Soboroff, P.Thomas, P.Bailey, N.Craswell, A.de Vries. Overview of the TREC 2008 Enterprise Track, The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings, http://trec.nist.gov/pubs/trec17/papers/ENTERPRISE.OVERVIEW.pdf
3) 三森，森嶋．分散ファイル群高度管理のためのミドルウェアの開発，DEIM2009，pp.9-11，2009.