情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-DBS-150 No.3
Vol.2010-IFAT-99 No.3
2010/8/4

帰属文書数に基づく Web ページ情報発信者の専門性分析

加藤義清[†1]　乾　健太郎[†1,†2]　黒橋禎夫[†1,†3]

本研究では Web ページ情報発信者の任意のトピックにおける専門性を分析する方法として，検索エンジンのヒット数に基づき発信者の専門性スコアを計算する手法と，発信者に帰属する Web ページ数に基づき発信者の専門性スコアを計算する手法を提案する．1 億 2 千万件の日本語 Web ページを検索対象とする収集検索基盤を用いた評価実験を実施した結果，帰属文書数に基づく手法が精度や処理時間について優れていることが分かった．

Expertise Analysis of Information Senders of Web Pages Based on Attribution Count

Yoshikiyo Kato,[†1] Kentaro Inui[†1,†2]
and Sadao Kurohashi[†1,†3]

In this study, we propose two methods that analyze the expertise of information senders of Web pages: 1) a method that computes expertise score based on hit count from a search engine (hit count method), and 2) a method that computes expertise score based on the number of documents that are attributed to an information sender (attribution count method). We evalated both methods using a crawl and search infrastrucure which indexes 120 million Japanese Web pages. The results show that the attribution count method outperforms hit count method in terms of precision and processing time.

†1 情報通信研究機構
National Institute of Information and Communications Technology
†2 東北大学大学院情報科学研究科
Graduate School of Information Sciences, Tohoku University
†3 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

1. はじめに

いまや Web は，様々な場面における意思決定のための重要な情報源の一つとなっている．特に，ブログ，SNS，動画投稿サイト，マイクロブログなどが普及し，いわゆる消費者発信メディア (CGM) が Web の主要な用途の一つとなっている．CGM には消費者の声が反映され，製品やサービスの選択する上で，重要な役割を果たすようになっている．Web からは様々な情報が得られる一方でその品質は必ずしも一定しない．そこで，情報の信頼性が重要となっている．情報の信頼性を判断する上で，重要な手掛かりの一つとなるのが情報の発信者である．情報の発信者の信頼性については，その意図と専門性が問題となるが，本研究では発信者の専門性に注目する．

本研究では Web ページ情報発信者の任意のトピックにおける専門性を分析する方法として，検索エンジンのヒット数に基づき発信者の専門性スコアを計算する手法と，発信者に帰属する Web ページ数に基づき発信者の専門性スコアを計算する手法を提案する．提案手法を，1 億 2 千万件の日本語 Web ページを検索対象とする収集検索基盤を用いて実験により評価した結果，帰属文書数に基づく手法が精度や処理時間について優れていることが分かった．以下，提案手法について述べ，実験について報告する．その後，現状で明らかとなっている課題について議論し，関連研究を述べた後で，本稿を締めくくる．

2. Web ページ情報発信者の専門性分析

本研究においては，Web ページ情報発信者の専門性分析を，クエリ q が与えられたときに，Web ページ情報発信者を q についての専門性の観点で順位付けをおこなう問題と捉える．

以下，Web 情報発信者の専門性分析の基本アルゴリズムを与える．その後，発信者 s のクエリ q についての専門性スコア $score(q, s)$ を計算する手法として，検索エンジンのヒット数に基づく方法と，帰属文書数に基づく方法について述べる．

2.1 基本アルゴリズム

Web ページ情報発信者の専門性分析をするに当たって，以下の前提をおく．

(1) Web ページの集合 D が与えられていて，検索エンジンにより検索可能である．
(2) Web ページ $\forall d \in D$ について，その発信者 $s = \mathrm{sender}(d)$ が与えられている．

以上の前提のもとで，以下の手順により発信者の集合及び発信者のスコアを与える．

(1) 検索エンジンにクエリ q で問い合わせ，ヒット数 $n(q)$ および検索結果の文書集合 D_q を取得する．

Vol.2010-DBS-150 No.3
Vol.2010-IFAT-99 No.3
2010/8/4

(2) D_q の発信者の集合 $S = \{s | s = \text{sender}(d) \wedge d \in D\}$ を取得する．
(3) 各発信者 s について，専門性スコア $\text{score}(q, s)$ を計算する．
(4) スコアに基づいて発信者に順位を与える．

2.2 ヒット数に基づく専門性スコア

検索エンジンから得られる統計量を利用して 2 語間の関係を推定する方法は，これまでに様々な形で応用されてきた．例えば，同意語の抽出[14] や，ソーシャルネットワークの抽出[9] などに利用されている．そこで，ベースラインとして，検索エンジンから得られるヒット数を元に与えられたトピックに対する発信者の専門性スコアを算出する方法について述べる．

検索エンジンに語 x を問い合わせて得られるヒット数を $n(x)$ として，発信者 s のトピック q についての専門性スコアを以下のように与える．

$$\text{score}(q, s) = f(n(q), n(s), n(s \wedge q)) \quad (1)$$

ここで，$n(s)$ および $n(q)$ はそれぞれ発信者名 s およびトピック語 q のヒット数であり，$n(s \wedge q)$ は s と q を AND 検索したときのヒット数である．その際，ヒット件数が一定値以下 ($n(s) < \theta_s$ および $n(q \wedge s) < \theta_{q \wedge s}$) の発信者についてはランキングの対象外とする．

スコアの計算方法としては，以下の 6 種類を検討した．

(1) Matching 係数: $\text{score}(q, s) = n(q \wedge s)$
(2) 自己相互情報量: $\text{score}(q, s) = \log \frac{n(q \wedge s)}{n(q)n(s)}$
(3) Dice 係数: $\text{score}(q, s) = \frac{2n(q \wedge s)}{n(q)+n(s)}$
(4) Jaccard 係数: $\text{score}(q, s) = \frac{n(q \wedge s)}{n(q \vee s)}$
(5) Overlap 係数: $\text{score}(q, s) = \frac{n(q \wedge s)}{\min(n(q), n(s))}$
(6) Cosine 類似度: $\text{score}(q, s) = \frac{n(q \wedge s)}{\sqrt{n(q)n(s)}}$

予備実験の結果，自己相互情報量および Overlap 係数の性能が良かったので，以後はこの 2 つについてのみ考慮の対象とする．

2.3 帰属文書数に基づく専門性スコア

ヒット数法において用いたクエリと発信者名の AND 検索のヒット数 $n(q \wedge s)$ (hit count) の代わりに，検索結果中のページのうち，発信者に帰属するページ数 (attribution count) を利用して，専門性のスコアを計算する．

(1) 検索エンジン (TSUBAKI) にクエリ q で問い合わせ，ヒット数 $n(q)$ および検索結果の文書集合 D_q（最大で $|D_q| = 1000$）を取得
(2) D_q の発信者の集合 $S = \{s | s = \text{sender}(d) \wedge d \in D_q\}$ を取得
(3) 各発信者 s について，コーパス全体で s に帰属する文書数 $\text{df}(s)$，および D_q 中で s に帰属する文書数 $n_{D_q}(s)$ を取得
(4) 発信者のスコアを計算：$\text{score}(q, s) = f(n(q), \text{df}(s), n_{D_q}(s))$
(5) スコアに基づいて発信者に順位を与える．ただし，$\text{df}(s) < \theta_{df}$ および $n_{D_q}(s) < \theta_r$ となる発信者 s についてはランキングの対象外とした．

スコアの計算方法として，ヒット数と同じく自己相互情報量 $\log \frac{n_{D_q}(s)}{n(q)\text{df}(s)}$ および Overlap 係数 $\frac{n_{D_q}(s)}{\min(n(q), \text{df}(s))}$ を考慮した．

3. 実　　験

次に，提案手法の性能を評価する目的で実施した実験について述べる．実験には情報通信研究機構で運用する大規模 Web ページ収集・検索基盤[17]（以下，収集検索基盤）を用いた．実験を実施した 2010 年 6 月の時点において，収集検索基盤は 3 億件を超えるユニークな URL を持つ日本語ページを管理している．その中から，リンク解析，URL の特徴，ページに含まれる文数など内容に関する特徴などに基づいて，質が高いと思われる 1 億 2000 万ページが選択され，検索エンジン TSUBAKI[12] によって検索可能な状態となっている．検索対象となった Web ページに対しては，Web ページの情報発信者を同定する手法[16] を適用してサイト運営者名が抽出されているほか，人手で構築されたドメイン名からサイト運営者名を与えるサイト運営者データベースを運用しており，各ページにはいずれかの方法により情報発信者名が与えられている．

3.1 方　　法

評価用データとして，**表 1** に挙げた 16 のトピックについて収集検索基盤で $\max |D_q| = 1000$ として，評価対象とした各手法で検索された上位 10 件の発信者を評価対象としてプールした．その際，ヒット数法の場合は $\theta_{q \wedge s}$ を，帰属文書数法の場合は θ_r をそれぞれ，1 から 10 まで変えて検索をおこなった．この際，ヒット数法の θ_s および帰属文書数法の θ_{df} はそれぞれ 10 とした．

こうして得られた評価対象の発信者集合について，評価者 1 名が専門性について判定をおこなった．判定の基準を**表 2** に示す．この中で，not a sender というカテゴリは，発信者名を抽出する際に，誤った名前が発信者名が抽出される場合があり，その場合には専門性は判定せずに not a sender に分類した．

評価者による評価に基づき，各条件について，上位 10 件の適合率 (P@10, precision at

Vol.2010-DBS-150 No.3
Vol.2010-IFAT-99 No.3
2010/8/4

表 1 評価に用いたトピックの一覧.
Table 1 Topics used for evaluation.

ゆとり教育	イソフラボン	クールビズ	セカンドライフ
ら抜き言葉	カテキン	サマータイム制	ダイエット食品
アガリクス	キシリトール	ジェネリック医薬品	ドラフト制度
アンチエイジング	クローン技術	ステロイド剤	ネットオークション

表 2 発信者の専門性の判定基準
Table 2 Criteria for expertise judgment.

カテゴリー	判定基準
Expert	(1) 発信者が人物であって，クエリの分野に関する著書がある，専門分野に関係の深いの肩書きを有するなど，クエリについての専門家だと判断できる．(2) 発信者が組織であって，業務内容などからクエリについての専門家だと判断できる
Semi-expert	(1) クエリに関して，発信者が一定の影響力を有していると考えられる (例: クエリが「年金制度」で，発信者が政治家である) (2) 発信者が自身の Web サイト，ブログ等においてクエリに関連する分野に関して継続的に情報発信をしている．
Not an expert	クエリの専門家ではない．
Not a sender	発信者ではない．(発信者名抽出誤り等)

10)，平均適合率の平均 (MAP, mean average precision)，および平均逆順位 (MRR, mean reciprocal rank) を求めて比較をおこなった．その際，expert のみを適合とする場合 (strict 条件) と expert および semi-expert ともに適合とする場合 (loose 条件) に分けて評価をおこなった．

3.2 結　　果

各条件における P@10，MAP，MRR を表にまとめたものをそれぞれ**表 3**，**表 4**，および**表 5** に示す．AC は帰属文書数法を，HC はヒット数法を，OL はスコア計算に Overlap 係数を用いた場合，PMI はスコア計算に自己相互情報量を用いた場合を表す．各行は，$\theta_{q\wedge s}$ および θ_r が特定の値の場合に相当する．この結果より，loose 条件，strict 条件いずれの場合においてもヒット数法に比べて帰属文書数法が高い性能示すことが分かる．

$\theta_{q\wedge s}$ および θ_r を変化させたときの影響を見るために，同じデータを $\theta q \wedge s$ あるいは θ_r を横軸にしてグラフにした者を**図 1**〜**図 6** に示す．これらの図より，以下のことが分かる．

(1) $\theta = 1$ の条件では，ヒット数法の性能は帰属文書数法のそれと同じか上回る．
(2) $\theta > 1$ では，帰属文書数法の性能がヒット数法の性能を上回る．
(3) ヒット数法は $\theta_{q\wedge s}$ の値に対して安定的な性能を示す．
(4) 帰属文書数法は θ_r の値に大きく影響を受け，ピーク値が存在する．

表 3 各手法における上位 10 件の適合率の比較.
Table 3 The precision at 10 of each method.

	loose				strict			
$\theta_{q\wedge s},\theta_r$	AC-OL	AC-PMI	HC-OL	HC-PMI	AC-OL	AC-PMI	HC-OL	HC-PMI
1	0.138	0.138	0.150	0.175	0.081	0.081	0.069	0.088
2	0.275	0.263	0.144	0.194	0.169	0.169	0.063	0.088
3	0.288	0.313	0.150	0.194	0.156	**0.175**	0.069	0.100
4	0.269	0.294	0.150	0.194	0.138	0.156	0.069	0.094
5	0.278	**0.316**	0.150	0.194	0.127	0.145	0.069	0.094
6	0.262	0.306	0.150	0.213	0.094	0.113	0.063	0.100
7	0.275	0.306	0.150	0.206	0.094	0.101	0.063	0.094
8	0.284	0.315	0.144	0.194	0.079	0.079	0.069	0.100
9	0.278	0.297	0.144	0.188	0.078	0.078	0.069	0.094
10	0.252	0.265	0.144	0.188	0.062	0.062	0.069	0.100

表 4 各手法における平均適合率の平均の比較.
Table 4 The mean average precision of each method.

	loose				strict			
$\theta_{q\wedge s},\theta_r$	AC-OL	AC-PMI	HC-OL	HC-PMI	AC-OL	AC-PMI	HC-OL	HC-PMI
1	0.191	0.191	0.274	0.232	0.147	0.147	0.145	0.159
2	0.316	0.309	0.265	0.212	0.210	0.210	0.083	0.114
3	0.358	0.355	0.266	0.227	0.250	0.255	0.089	0.124
4	0.414	0.411	0.266	0.232	**0.266**	0.253	0.089	0.119
5	0.406	0.420	0.265	0.236	0.247	0.253	0.090	0.124
6	0.375	0.393	0.266	0.252	0.217	0.226	0.097	0.129
7	0.363	0.391	0.266	0.249	0.182	0.219	0.097	0.124
8	0.368	**0.431**	0.266	0.253	0.176	0.225	0.092	0.128
9	0.356	0.423	0.266	0.259	0.191	0.233	0.092	0.121
10	0.325	0.408	0.266	0.253	0.181	0.231	0.092	0.126

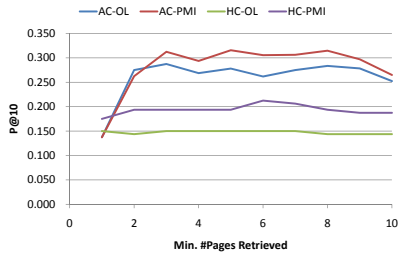

図 1 loose 条件における上位 10 件の適合率.
Fig. 1 Precision at 10 with loose condition.

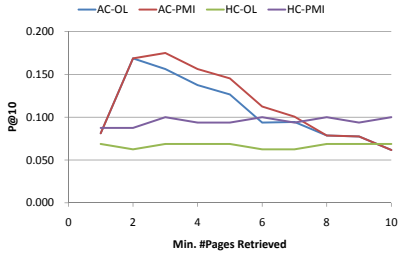

図 2 strict 条件における上位 10 件の適合率.
Fig. 2 Precision at 10 with strict condition.

Vol.2010-DBS-150 No.3
Vol.2010-IFAT-99 No.3
2010/8/4

表 5 各手法における平均逆順位の比較.
Table 5 The mean reciprocal rank (MRR) of each method.

	loose				strict			
$\theta_{q\wedge s},\theta_r$	AC-OL	AC-PMI	HC-OL	HC-PMI	AC-OL	AC-PMI	HC-OL	HC-PMI
1	0.241	0.241	0.328	0.268	0.205	0.205	0.168	0.198
2	0.328	0.307	0.328	0.253	0.227	0.227	0.105	0.129
3	0.382	0.358	0.328	0.256	0.280	0.284	0.111	0.135
4	**0.467**	0.422	0.328	0.254	**0.299**	0.292	0.111	0.130
5	0.451	0.418	0.328	0.257	0.281	0.285	0.112	0.135
6	0.427	0.403	0.328	0.267	0.252	0.260	0.113	0.144
7	0.385	0.378	0.328	0.267	0.200	0.231	0.113	0.144
8	0.390	0.458	0.328	0.275	0.202	0.252	0.114	0.150
9	0.383	0.452	0.328	0.285	0.202	0.252	0.114	0.141
10	0.339	0.442	0.328	0.285	0.181	0.231	0.114	0.144

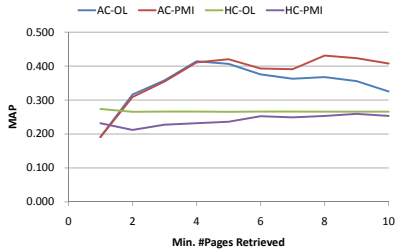

図 3 loose 条件における平均適合率の平均.
Fig. 3 Mean average precision with loose condition.

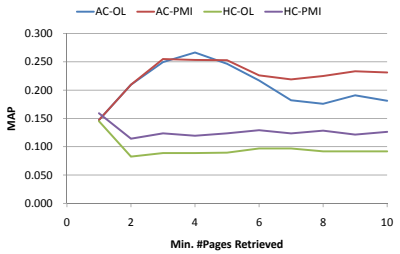

図 4 strict 条件における平均適合率の平均.
Fig. 4 Mean average precision with strict condition.

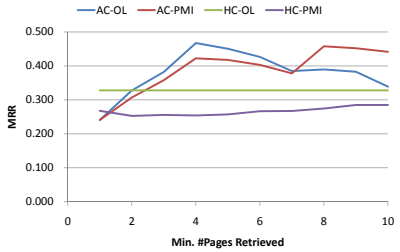

図 5 loose 条件における平均逆順位.
Fig. 5 Mean reciprocal rank with loose condition.

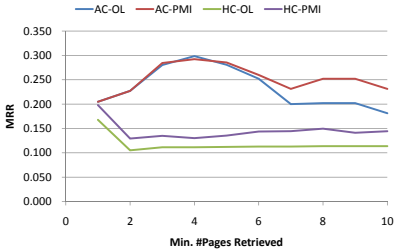

図 6 strict 条件における平均逆順位.
Fig. 6 Mean reciprocal rank with strict condition.

4. 議　　論

4.1 誤り分析

帰属文書数法について，発信者名の抽出誤りの影響による誤りと，検索結果中の関連性の低い文書の影響による誤りが観察された．以下，それぞれの場合について議論する．

4.1.1 発信者名抽出誤りの影響

帰属文書数法では，自動抽出された発信者名を元に発信者に帰属する文書数を求めているため，発信者名の表記揺れや発信者抽出の誤りにより影響を受ける．影響の受け方として，大きく 2 通りに分けられる．1 つ目は異なる発信者のページに同じ名前を付与されたことにより同一の発信者として扱われる場合である．例えば,「赤ちゃんポスト」というトピックの場合，トピックに関連する組織として「慈恵病院」がページ中に良く出現する．多くのページで誤って同じ名前が抽出されると，$n_{D_q}(s)$ が大きくなり，専門性の高い発信者として上位に出現しやすくなるため，特にこの種の誤りは目立つことになる．

2 つ目は同じ発信者のページから異なる発信者名が抽出されたことによりそれぞれ異なる発信者に帰属するページとして扱われる場合である．具体的には，日本語名と英語名など表記の異なる名前が抽出されたり（例えば「情報通信研究機構」と「NICT」），そもそも誤った名前が抽出されたりする場合などが挙げられる．これらの問題に対しては，発信者名抽出の精度の向上を図るとともに，実体解決[4] により表記揺れの問題へ対応することが必要となる．

4.1.2 関連性の低い文書の影響

検索エンジンの通常の使用では，ユーザは検索結果の上位のページしか見ないので，下位のページの質についてはそれほど問題とはならないが，帰属文書数法では $|D_q| = 1000$ と下位の結果も利用するため，下位のページの質にも影響を受けることになる．特に次のようなページが目立った．

(1) コンテンツに直接関係のないテキストの影響
(2) 商品販売サイト
(3) SEO やアフィリエイト目的の自動生成サイト（スパム）

(1) はタイトルやリンクにトピック語が含まれているが，ページの内容がトピックと必ずしも関係しない場合である．このようなケースは，メインテキストのみを検索対象とするなどの対策によりある程度の対応は可能だと考えられる．(2) や (3) については，スパムフィルタや検索対象の選択の段階で対応が必要となる．これらの点については収集検索基盤でも継

Vol.2010-DBS-150 No.3
Vol.2010-IFAT-99 No.3
2010/8/4

続的に取り組んでいる[17]．

4.2 処理時間

ヒット数法では，D_q から得られる発信者数 $|S|$ に対して $n(s)$ および $n(q \wedge s)$ と問い合わせるために $2|S|$ 回の検索エンジンへの問い合わせが必要となる．実験に用いた収集検索基盤では 1 回の問い合わせに平均で 5 秒程度要し，$|S|$ は 1 トピックにつき平均で 300 あるので，1 トピックに付き 3000 秒かかることになる．トピックとは独立した $n(s)$ については事前に求めておくことができるが，$n(q \wedge s)$ はトピックのクエリ q が与えられて初めて計算ができるので，問い合わせ時間が大幅に改善しない限り，ヒット数法はオンライン処理には不利である．一方，帰属文書数法で用いる $n_{D_q}(s)$ は D_q 内のページ数を数えるだけで求まり，df(s) も予め求めておくことができるのでヒット数法のように時間はかからず，オンライン処理に向いている．ヒット数法でも n_{D_q} を $n(q \wedge s)$ の近似として用いれば高速化を図れるが，その精度の評価は今後の課題である．

5. 関連研究

専門家検索の必要性について，これまでは知識マネジメントや企業内検索の文脈で語られることが多かった[5]．当初は，従業員の技能や知識をデータベースで管理するアプローチが取られたが，そのようなデータベースの構築や保守は費用と時間がかかり，イントラネットの情報から従業員の技能や知識を自動抽出する試みもなされている[3]．その後，2005 年の Text REtrieval Conference (TREC)[⋆1]で企業内検索タスクの一つとして専門家検索が取り上げられるようになる．そこでの課題は電子メールや Web ページを含むコーパスから，特定のトピックについての専門家を検索するというものである．それをきっかけに，主に情報検索的なアプローチの手法が提案された．Balog らは確率的言語モデルに基づく専門家検索手法を提案している[2]．そこでは，コーパス中の文書と専門家候補の関連性に基づいてクエリと専門家候補の関連性を求めるモデルを提案している．また，Macdonald と Ounis は各専門家候補毎に関連文書集合をプロファイルとして与えた上で，トピックの検索結果とプロファイルの共通集合を求め，そこに含まれる文書のスコアを集約して候補のスコアを与える方法を提案している[8]．これらの手法と比較して，提案手法は企業内文書ではなく一般の Web ページを対象としている点と，専門家候補として Web ページの発信者を用いている点で異なっている．

⋆1 http://trec.nist.gov/.

学術論文を対象に著者の専門性をモデル化する研究もおこなわれている．トピックモデルを用いるアプローチとして，Rosen-Zvi や Steyvers の研究グループが Author-Topic モデルの提案と，それを用いた応用を示した[11),13)]．その後，Mimno と McCallum は Author-Topic モデルを拡張した Author-Persona-Topic モデルを提案している[10]．異なるアプローチとして，Zhou らは論文の引用ネットワークおよび著者の共著ネットワークに基づく著者のランキング手法を提案している[15]．これらの研究は，学術論文を対象としている点で，本研究とは異なっている．学術論文は文書と著者の関係を所与として扱うことができる一方で，Web ページの場合には発信者の抽出が必要となる．ただ，発信者の抽出さえできれば，これらの手法を Web ページに適用することは可能である．これらの手法の Web ページ情報発信者の専門性分析への適用は今後の課題である．

Web を対象とした専門家検索の研究としては，Wikipedia を利用した専門家検索の提案[6]や，セマンティック Web 的なアプローチの提案[7] などがあるが，十分な評価はまだなされていない．専門性分析に関連する研究として，中島らは特定の分野の熟知度に基づいてブログをランキングする手法を提案している[18]．この手法では分野ごとに関連の深いキーワードの辞書を予め構築しておき，一定期間継続的に同じ話題の投稿を続けているなどの基準に達したブロガーに対して，エントリに含まれるキーワードの頻度に基づいてスコアを計算する．本研究で提案する手法は，Web ページ一般を対象としている点と，分野ごとに辞書を用意する必要がない点で，中島らの手法とは異なっている．他にブログを対象とした研究として，Agrawal らは影響力のあるブロガーを発見する手法を提案している[1]．専門性分析そのものではないが，専門性分析と組み合わせることにより，より関連性の高い発信者をユーザに提示できる可能性がある．

6. おわりに

本稿では，Web ページ情報発信者の専門性について，あるトピックにおける発信者の専門性をトピックと発信者の関連性の強さと捉え，検索エンジンのヒット数に基づき専門性スコアを求めるヒット数法と，発信者に帰属する文書数に基づいて専門性スコアを求める帰属文書数法を提案した．評価の結果，帰属文書数法が精度および処理速度の点で優れていることが分かった．

発信者名抽出の誤りに起因する問題，発信者名の表記揺れの問題，および検索結果中の関連性の低い文書の問題などが課題として明らかになった．言語モデルによるアプローチや，リンク解析に基づくアプローチとの比較は今後の検討課題である．

Vol.2010-DBS-150 No.3
Vol.2010-IFAT-99 No.3
2010/8/4

参 考 文 献

1) Agrawal, N., Liu, H., Tang, L. and Yu, P.S.: Identifying the Influential Bloggers in a Community, *Proceedings of the First ACM International Conference on Web Search and Data Mining (WSDM'08)*, pp.207–217 (2008).
2) Balog, K., Azzopardi, L. and de Rijke, M.: Formal Models for Expert Finidng in Enterprise Corpora, *Proceedings of SIGIR'06*, pp.43–50 (2006).
3) Becerra-Fernandez, I.: Searching for Experts on the Web: A Review of Contemporary Expertise Locator Systems, *ACM Transactions on Internet Technology*, Vol.6, No.4, pp.333–355 (2006).
4) Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E. and Widom, J.: Swoosh: a generic approach to entity resolution, *The VLDB Journal*, Vol.18, No.1, pp.255–276 (2009).
5) Davenport, T. and Prusak, L.: *Working Knowledge: How Organizations Manage What They Know*, Harvard Business Press (1998).
6) Demartini, G.: Finding Experts Using Wikipedia, *Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS'07)*, pp. 33–41 (2007).
7) Jung, H., Lee, M., Kang, I.-S., Lee, S.-W. and Sung, W.-K.: Finidng Topic-centric Identified Experts based on Full Text Analysis, *Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS'07)* (2007).
8) Macdonald, C. and Ounis, I.: Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task, *Proceedings of CIKM'06* (2006).
9) Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K. and Ishizuka, M.: POLYPHONET: an advanced social network extraction system from the web, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, ACM, pp.397–406 (2006).
10) Mimno, D. and McCallum, A.: Expertise Modeling for Matching Papers with Reviewers, *Proceedings of KDD'07*, pp.500–509 (2007).
11) Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P.: The Author-Topic Model for Authors and Documents, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp.487–494 (2004).
12) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S.: TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology, *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp.189–196 (2008).
13) Steyvers, M., Smyth, P., Rosen-Zvi, M. and Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery, *Proceedings of KDD'04*, pp. 306–315 (2004).
14) Turney, P. D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, Lecture Notes in Computer Science, Vol. 2167, London, UK, Springer-Verlag, pp. 491–502 (2001).
15) Zhou, D., Orshanskiy, S. A., Zha, H. and Giles, C. L.: Co-Ranking Authors and Documents in a Heterogeneous Network, *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pp.739–744 (2007).
16) 加藤義清，河原大輔，乾健太郎，黒橋禎夫，柴田知秀：Web ページの情報発信者の同定，人工知能学会論文誌， Vol.25, pp.90–103 (2010).
17) 赤峯　享，加藤義清，河原大輔，レオン末松豊インティ，新里圭司，乾健太郎，黒橋禎夫，木俵　豊：Web 情報分析のための大規模 Web ページの収集・選択・検索，言語処理学会第 16 回年次大会論文集，pp.238–241 (2010).
18) 中島伸介，稲垣陽一，草野奉章：高信頼性情報の提示を目指した熟知度に基づくブログランキング方式の提案，日本データベース学会論文誌， Vol. 7, No. 1, pp. 257–262 (2008).