Web 文書における効率的な情報発信源特定手法

見市 高一† 鈴木 優† 川越 恭二†
† 立命館大学 情報理工学部

1 はじめに

Web において伝播する情報の信頼度を算出する手法の一つとして，その情報の発信源を求める手法がある．本研究における発信源とは，Web における参照元となる文書である．風間ら [1] は，参照元を明記していない Web 文書の発信源を求める手法として，発信源の生成時刻を使用して発信源を特定する手法を提案している．しかし，生成時刻に注目するだけでは発信源を発見する際に手間がかかるという問題がある．ここで，利用者の閲覧 Web 文書と発信源 Web 文書は文書の内容が類似していると考えられる．そのため，発信源の候補となる対象 Web 文書 (以降対象 Web 文書) のクラスタリングを行い，対象 Web 文書数を削減することによって，発信源を特定する手間を省くことが可能となると考えられる．

そこで本研究では，対象 Web 文書のクラスタリングを行い，Web 文書の更新日時から閾値以内に更新された Web 文書のみを発信源の対象とすることによって，対象 Web 文書数を削減する手法を提案する．提案手法を用いることによって，利用者が発信源特定に要する手間を省くことが可能となる．

2 情報発信源候補対象 Web 文書の絞込み

2.1 基本的考え方

本研究では，まずクラスタリングを行う Web 文書の絞込みを行うために，検索エンジンに問い合わせるキーワードを用いる．これは，利用者がキーワードを用いて検索エンジンに問い合わせた結果に含まれない Web 文書群は，対象 Web 文書となりえないと考えたためである．また，クラスタリングを行う理由として，対象 Web 文書の絞込みを行う際に閲覧 Web 文書と発信源 Web 文書は内容が類似していると考えられる．そこで，クラスタリングを行うことによって内容の類似している Web 文書を集めておき，最も類似したクラスタ内から発信源を特定しようと考えたためである．こ

The Efficient Specific Method of Information Sending Source about Web Documents
†Koichi MIICHI, †Yu SUZUKI, †Kyoji KAWAGOE
†College of Information Science and Engineering,Ritsumeikan University
†miichi@coms.ics.ritsumei.ac.jp
†{yusuzuki,kawagoe}@is.ritsumei.ac.jp

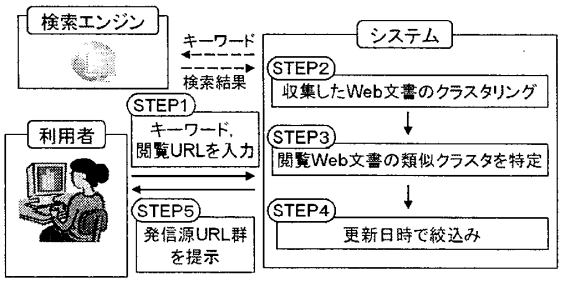

図 1: システムの概要図

のことにより，対象 Web 文書数を削減することが可能となる．そして，最も類似したクラスタ内で対象 Web 文書をさらに更新日時に着目して絞り込む．提案手法を用いることによって，対象 Web 文書数を削減し，利用者が発信源特定の際に手間を省くことが可能となる．

2.2 提案手法概要

提案手法の概要図を図 1 に示す．まず STEP1 として，利用者は閲覧 Web 文書の URL と，キーワードをシステムに入力する．次に STEP2 として，システムはキーワードを問合せとした検索結果から収集し，類似した Web 文書を分類するため，収集した Web 文書のクラスタリングを行う．そして STEP3 では，システムは閲覧 Web 文書と最も類似したクラスタを特定する．さらに STEP4 では，クラスタ内に存在する Web 文書をさらに既存手法である更新日時によって絞込む．最後に STEP5 では，システムは絞込みを行った後の対象 Web 文書を利用者に提示する．

2.3 Web 文書のクラスタリング

本研究では，Ward 法 [2] を用いて Web 文書のクラスタリングを行う．Web 文書のクラスタリングを行うために，Web 文書に対して形態素解析を行い単語を抽出する．そして各単語に対して重み付けを行い Web 文書ごとに TF-IDF 値を重みとする文書ベクトルを生成する．生成した各文書ベクトルを初期クラスタとし，クラスタ間の距離を算出する．Ward 法では，クラスタを併合したときにクラスタ内平方和の増加分が最小のクラスタ同士を併合する．平方和とは，任意のクラス

タ $G_{\mathbf{j}}$ 内の特徴ベクトル $v_{\mathbf{j}}$ に対して，クラスタ $G_{\mathbf{j}}$ の重心ベクトル $g_{\mathbf{j}}$ との距離を二乗した値の和を表す．

クラスタ $G_{\mathbf{j}}$ に u 個のベクトルが含まれているとすると，クラスタ $G_{\mathbf{j}}$ での平方和 $E(G_{\mathbf{j}})$ は (1) 式に示す．

$$E(G_j) = \sum_{l=1}^{u} \sum_{h=1}^{M} (d_h(w_h^l) - d'_h(w_h^j))^2 \qquad (1)$$

ここで，$d_i(w_i^k)$ は単語 $w_i^k (1 \leq i \leq M)$ における重み，$d'_b(w_b^j)(1 \leq b \leq M)$ は，単語 $w_{\mathbf{b}}$ における重みの平均，M は単語数，k は任意の Web 文書を表す．(1) 式ので算出された平方和を用いて (2) 式よりクラスタ G_α とクラスタ G_β 間の距離 $D(G_\alpha, G_\beta)$ を計算する．

$$D(G_\alpha, G_\beta) = E(G_\alpha \cup G_\beta) - E(G_\alpha) - E(G_\beta) \qquad (2)$$

(2) 式により算出された距離が最小のクラスタ同士を併合する．そして新しく出来たクラスタと他のクラスタとの距離を再度計算し，併合を繰り返す．クラスタ数が一定の個数以下になるとクラスタリングを終了する．

2.4 利用者の閲覧 Web 文書とクラスタとの類似度算出

2.3 節の結果を用いて利用者の閲覧 Web 文書がどのクラスタに属するのかを決定する．そこで，利用者の閲覧 Web 文書の特徴ベクトル $V_{\mathbf{r}}$ とクラスタ $G_{\mathbf{j}}$ の重心ベクトルとの類似度 $sim(V_{\mathbf{r}}, G_{\mathbf{j}})$ を (3) 式によって求める．

$$sim(V_{\mathbf{r}}, G_{\mathbf{j}}) \quad = \quad E(V_{\mathbf{r}} \cup G_{\mathbf{j}}) - E(G_{\mathbf{j}}) \qquad (3)$$

(3) 式の値が最も小さいクラスタに利用者の閲覧 Web 文書が属することを決定する．

2.5 更新日時による Web 文書の絞込み

利用者の閲覧 Web 文書の更新日時より一定時間以上前に更新された Web 文書については閲覧 Web 文書との関連性が低いと考えられる．そこで，利用者の閲覧 Web 文書の更新日時と比較し，一定時間以内に更新された Web 文書のみ対象 Web 文書とする．また，利用者の閲覧 Web 文書の更新日時より前に更新された Web 文書のみ対象 Web 文書とする．

表 1: 実験結果

検索語	手法	対象 Web 文書数	絞込み率
内定　取り消し	提案手法	82	0.164
	従来手法	311	0.622
高校サッカー　注目	提案手法	114	0.228
	従来手法	306	0.612
日本　景気	提案手法	196	0.392
	従来手法	274	0.548

3 評価実験

3.1 実験環境

本提案手法が従来手法と比較して利用者が閲覧 Web 文書の発信源を特定する際の手間を省くことが可能になることを証明するために評価実験を行った．従来手法として更新日時のみから対象 Web 文書の特定を行う手法を用いた．実験環境として，Yahoo!JAPAN より本研究における検索語を問合せとした検索結果上位 500 件を対象とした．また，クラスタリングを行う際のクラスタ数は 10 個，更新日時の閾値を 3 日と設定した．

3.2 実験結果と考察

各検索語に対する利用者の閲覧 Web 文書の発信源 Web 文書数を検索語実験結果を表 1 に示す．絞込み率を収集した Web 文書数に対する特定した発信源 Web 文書数の割合と定義すると，表 1 より従来手法と比較して，提案手法は絞込み率の値が小さくなっている．これは，提案手法は従来手法に加えクラスタリングを用いることによってさらに対象 Web 文書数を削減することが可能となったためであると考えられる．また，提案手法と従来手法の双方において，発信源特定精度 (適合率) の変化はほぼ見られなかった．このことにより，提案手法を用いた方がより利用者の手間を省くことが可能になったことが考えられる．しかし，対象 Web 文書のクラスタリングに関して，鎖効果が起こり特定のクラスタに対象 Web 文書が集中したので，クラスタリングの手法とクラスタ数について検討する必要があると考える．

4 おわりに

本稿では，Web 文書の絞込みを行うことにより，閲覧 Web 文書の発信源特定に要する手間を減らす手法を提案した．今後は，Web 文書のクラスタリング手法およびクラスタ数の設定，更新日時の閾値について検討を行う．また，収集する Web 文書数や収集方法についても検討を行う．さらに，利用者のキーワード入力を不要とするための検討も行う．

参考文献

[1] 風間一洋, 今田美幸, 柏木啓一郎：“ブログ空間における情報伝播ネットワークの抽出と分析”, Web とデータベースに関するフォーラム (2008).

[2] 神嶌敏弘：“データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう！－”, 人工知能学会誌, **Vol.18**, No.1, pp. 59–65 (2003).