1A3-2

人物名共起情報を利用したブログ画像検索に向けた重要度算出方法の検討

Ranking Bloggers for Web Image Retrieval Using Co-occurrence of Names of Person

植松 幸生[*1*2] 片岡 良治[*1] 大和田 勇人[*2]
Yukio Uematsu Ryoji Kataoka Hayato Ohwada

[*1]日本電信電話株式会社，NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

[*2]東京理科大学 理工学研究科
Faculty of Science and Technology, Tokyo University of Science

In this paper, we describe one of the methods concerning web image retrieval (especially the images in blogs). There are many images that reflect blogger's characteristics especially bloggers that are collectors of pictures of famous person. To treat these characteristics for blog image retrieval, we define relevancy between name of person and bloggers as the blogger's importance for the name of person. We propose a method to calculate the relevancy, using co-occurrence of name of persons.

1. はじめに

本研究は Web 上 (とくにブログ) の画像を対象とした検索に関する手法の一つである．ブログ記事は記事の投稿者 (ブロガー) の個性が反映されており，貼り付けられている画像 (特に有名人等の人物が被写体となったような画像) の傾向に特徴がある．この特徴を捉えてその人物名とブロガー間の関係性を定義し，ブログ上の画像を検索対象とした重要度として用いることを考えた．本稿ではこの人物名に対するブロガーの重要度を人物名の共起情報を利用して算出する手法を提案する．

近年 Web 上にある大量の情報から欲しい情報を得る方法として Web 検索をすることはごく一般的なこととなった．Web 検索とはユーザが入力する文字列情報に基づいて，その文字列に関連する Web ページを結果として返すシステムである．そのような大量の Web ページの中の画像のみを検索対象とした Web 検索を Web 画像検索と呼ぶ．Web 画像検索には goo 画像検索[*1]や Google 画像検索[*2]等があり，検索条件として入力される検索語のほとんどが一語で，かつ人物名であることが報告されている [3]．例えば，goo 画像検索の画像ランキング[*3]を見ると，上位のクエリはすべて人物名である．このように Web 画像検索のニーズは人物画像に偏っている事が分かる．

このような Web 上に配信される情報の中でも，CGM(Consumer Generated Media) と呼ばれるブログや SNS 上で個人が配信する情報が新しいメディアとして注目が集まりつつある．注目が集まる理由としてはこのようなメディアに投稿される情報は個人の思考や趣味が反映されており，それら情報には Web 検索の Longtail に答えるようなニッチな情報があるからだと考える．

このように個人が配信しているようなニッチな情報を手に入れたいというニーズは文字列情報だけではなく，画像もまた検索対象として注目されつつある．こうした CGM の中でもブログに限定し，ブログで配信される画像を対象とした検索をブログ画像検索と呼ぶ．ブログ画像検索には既存のシステムとして尾内等のもぶろげっと[*4]や，ライブドアの画像検索[*5]がある．これらのシステムでは，入力された文字列がブログの各記事中に存在する画像を提示している．この方法は検索対象となるメディアこそ異なるが，Web 画像検索と大きな差異は見られない．これはブログの特徴である趣味，嗜好を反映するような画像が検索結果の上位に提示できていないからである．ブログ画像検索を行うユーザの検索目的は Web 画像検索で入手困難な Longtail に応える画像を手に入れる検索目的があると考えられ現状のシステムではそれを解決するような重要度の算出が出来ていないと考える．そこで，我々は人物画像を目的とした検索において，検索対象となる人物に詳しいブロガーは Web 画像検索では入手困難な画像を持っているという仮説を立て，ある人物に詳しいブロガーを人物名の頻度情報や，共起情報を利用して特定することで上記問題を解決する．

2. 関連研究

Web 検索において検索対象の重要度算出方法に関する関連研究についてふれる．一般的に検索条件の文字列に対するある Web ページの重要度を算出する場合，ハイパーリンク関係を利用して類似度を算出する場合と検索条件と検索対象の類似度を算出する場合の 2 種類のアプローチがある．

ハイパーリンク情報を利用して重要度を算出するアプローチとしては PageRank や HITS 等がある．例えば PageRank ではハイパーリンク構造をグラフ構造とみなし，ある Web ページ閲覧者がランダムに Web ページを閲覧した際に到達する遷移確率を重要度として算出している．本研究のある人物に関して詳しいかどうかの指標に対しては直接重要度を算出する指標では無い．

また，記事そのものではなく記事を書くブロガーに重要度を付与する試みとしては，藤村等の研究がある [1]．藤村等は EigenRumor と呼ばれるコメントやトラックバックの情報から，ブロガーの重要度を算出する手法を提案している．しかしながら，対象をブログにすると，16.5%のブログ記事にのみ 1

A: 植松幸生，NTT サイバーソリューション研究所，神奈川県横須賀市光の丘 1-1，046 - 859 - 4923，046 - 855 - 1730，uematsu.yukio@lab.ntt.co.jp

∗1 http://bsearch.goo.ne.jp/
∗2 http://images.google.co.jp/
∗3 http://bsearch.goo.ne.jp/rss/mmm_imgrank.rdf
∗4 http://mobloget.jp/
∗5 http://blogimage.livedoor.com/

つ以上のリンクが存在することが藤村等により指摘されており，重要度を算出する対象が非常にスパースであり適用が難しい．さらに画像を持つエントリは限られてしまう．

また，検索条件と検索対象との関連度を算出する方法として $TFIDF$ がある．$TFIDF$ は文書中に存在する単語の網羅性 TF と特定性 IDF から単語の重み付けを行い，コサイン類似度等で文書間の関連度を算出する．後述する我々の手法は，TF の代わりにブログ記事頻度を利用している．

上記した重要度算出方法は検索対象となる情報がブログ記事すべてであるため，ユーザの検索目的が多様なため，ブロガーなどの重要度を汎用的に求める必要があった．しかしながら，検索対象を画像に限定した場合，人物名で検索されることがほとんどであるために，重要度を汎用的に求めるよりも，人物に特化した重要度算出手法がユーザニーズを反映するために効果的であると考えた．

3. 重要度算出方法

本節では画像を投稿したブロガーの重要度算出方法について述べる．ここで言う重要度とは前述した通り，あるブロガーがその人物に対してどの程度詳しいのかを数値化したものである．人物名 $P = (p_1, p_2, \cdots p_j)$ に対してのブロガー B の重要度を $importance_P(B)$ とし，下記のように算出する．

$$importance_P(B) = EF(P)(Entry \in B) \tag{1}$$

$$EF(P) = \sum_{i=1}^{N_B} EO(P) \tag{2}$$

$$EO(p_j) = score \tag{3}$$

ここで，$EF(P)$ (Entry Frequency) はブロガー B が投稿した記事 (Entry) の中で，人物名 P を持つ記事の数であり，$EO(P)$ (Entry Occurrence) はそのブロガーが書いた記事の中で人物名 p_j が出現する場合にスコア $score$ を付与する．与えるスコアに関しては後述する．

これを頻度情報のみを使ったベースラインと提案手法である人物の共起情報を利用した方法との比較検討を行う．まず，ベースラインとして単純頻度を用いる．これは前述した $score$ に対して，下式のように 1 を付与する．

$$\begin{array}{ll} EO(p_j) = 1 & \text{if } p_j \text{ occurs in Entry } i \\ EO(p_j) = 0 & \text{otherwise} \end{array}$$

このベースラインとなる指標は，ある人物名 p_j を多く言及した頻度を重要度としており，これを EO とする．

また，単純出現頻度では，例えばエントリの中に多くの異なり人物名を記述することで簡単に $importance_P(B)$ を上げる事が出来るので，各記事の重要度が 1 になるように正規化する．よって $EO(P)$ を下記の式に置き換え算出する．

$$\begin{array}{ll} EO2(p_j) = \frac{|p_j|}{N_P} & \text{if } p_j \text{ occurs in Entry } i \\ EO2(p_j) = 0 & \text{otherwise} \end{array}$$

ここで N_P は記事 i に出現する人物名の総数で，$|p_j|$ は人物名 p_j が記事 i 中に出現する頻度である．よって人物名 p_j がすべての異なり人物名の中でどの程度の割合を占めているのかを利用する事で $EO(p_j)$ を正規化し，これを $EO2$ とする．

前述したベースラインでは同一人物の異なり表記や，有名人が所属するようなグループ名等が検索条件として入力された場合そのグループに所属する人物名等を異なる人物として算出している．そこで提案手法では入力された人物名と異なり人物名の関連度を共起情報から求めることで上記問題を解決する．

$$EO3(p_j) = \sum_{m \in P} ECO(p_j, p_m) \tag{4}$$

$$ECO(p_j, p_m) = \frac{EF(p_m \cap p_j)}{EF(p_j)} \tag{5}$$

ここで $ECO(p_j, p_m)$ (Entry Co-Occurrence) は記事での共起頻度情報から求められる p_j,$p_,$ 間類似度の事で，$EF(p)$ は文書集合全体での人物名 p の頻度で，$EF(p_m \cap p_j)$ は人物名 p_m と p_j が共起するような文書数である．

4. 比較実験

ベースラインと提案システムについて実験を行う．実験データとしては 2005/12/27 から 12/31 に投稿されたブログ記事 286315 件で実験を行った．また全記事中 83269 件 (約 30%) の記事に画像が含まれていることが分かった．人物名を特定する際には磯崎等 [2] の固有表現抽出を利用した．固有表現抽出を利用する理由は例えば人物名 $p =$ さくら だったとすると，人物名の “さくら” と一般名詞の “さくら” を区別する事が出来るからである．上記データを利用して，前述したベースライン $EO, EO2$ ならびに提案手法である $EO3$ の検索結果の比較を行う．

4.1 結果比較

ここで手法の出力例として goo 画像検索で人気の高かった人物 A，グループ名 B を検索語として入力し，どのようなブロガーの重要度が高かったかを報告する．ベースラインシステムではいずれも人物名 A を含むようなエントリを持つブロガーが検索結果の上位に来たが，そもそも人物名 A について上記期間内に投稿した数が少なく，頻度 3 が最高であった．一方提案する共起情報を用いる場合は，登場する頻度が少ない場合でも人物 A に関連するような人物を収集しているようなブロガーの重要度が高く判定された．また，グループ名のように人物をメンバにするような場合は本手法の効果が高いことが分かった．

4.2 提案手法の効果と限界

ベースラインである人物の頻度情報を用いた手法は，ブログ上で頻繁に参照，投稿される有名人に対して重要度を算出する場合はトピックドリフトがおきにくいベースラインの方が効果が高かった．しかしながら，そのような日常的に言及されるような人物はごくわずかであり，普段あまり言及されないような人物を検索対象とする場合，抽象的な概念で重要度を算出する本手法が有効であった．また，人物名 A について当時言及していたブロガーも時間が経つにつれ興味が変化し，当時とは全く違う人物に関してウォッチしているような例も散見された．そこで，実用的に本手法を用いる場合は時間減衰も考慮する必要があると言える．

また，本手法は多くの記事を投稿するようなブロガーに対しては何も対策を施していないため，多くの無駄な人物名の入った記事と画像を投稿するようなスパムに対して脆弱であると言える．しかしながら，こうしたスパムもある人物名に特化して重要度を上げる事しかできないため，汎用的に importance(B) を上げることは出来ない．

今回の手法は重要度として検索対象全体で $P = (p_1 \cdots p_j)$ の人物名が含まれるとすると，各記事に対して j 個の重要度が求める必要があり，さらにその計算は $j \times j$ の行列計算を必

要とする．実用的に本手法を適用する際は j を圧縮するか，共起を求める集合を限定する必要がある．

5. まとめと今後の課題

本稿ではブログ記事中にある人物名とそれを投稿したブロガーに着目し，ある人物名に対するブロガーの重要度を人物名の共起頻度から算出する方法について述べた．本手法を利用することで，検索条件となる人物名が 1 度しか無い場合でも，他の人物名との共起頻度を用いることで，Web では手に入りにくいような画像を手に入れることが出来る．本手法はブログ特有の記事を投稿したブロガーが特定出来るという性質を利用したものだったが，一般的な Web 上においても “サイト” が定義できる場合はそのサイトに対して本手法を適用することが出来る．

今後の課題としては今回のシステムの定性的な評価方法などを策定し評価する必要がある．また，今回実験した記事数は非常に少ないため，必然的に人物名に対する頻度がスパースになってしまった．そこで実験に使用するデータセットを数千万記事にして，実験することで本手法の有効性を検証できると考えている．

参考文献

[1] Ko Fujimura, Takafumi Inoue, and Masayuki Sugisaki. ”the eigenrumor alogorithm for ranking blogs”, May 2005.

[2] Hideki Isozaki and Hideto Kazawa. Efficient support vector classifiers for named entity recognition. In *COLING*, 2002.

[3] 植松幸生, 片岡良治, 大和田勇人. Web 人物画像検索における周辺テキストからの重要固有表現特定手法の提案. 電子情報通信学会 Web インテリジェンスとインタラクション研究会, March 2006.