画像類似度の有向グラフを用いたCGM上の画像時系列ランキング手法

山根　遥香†　豊田　正史‡

東京大学情報理工学系研究科†　東京大学生産技術研究所‡

1. はじめに

web 上では，膨大な量のリッチコンテンツが用いられており，製品，キャラクタ，テレビ CM，イベントの画像・動画など，広報，広告，キャンペーンに呼応して多くのコンテンツが出現する．特に，ブログなどの CGM は流行に敏感であり，その時の話題の製品や CM などの記事が数多く書かれている．広報，広告等の影響を，Web データを用いて観測しその変遷を調査することは，マーケティング，リスク管理などに有用であると看做されている．

本研究は，ブログなどの CGM で公開されている画像に着目し，指定された話題に関連する画像の変遷を調査可能にすることを目的としている．本論文では指定されたクエリを含むブログ記事の時系列から画像を抽出し，テキスト情報，画像の特徴量を複合的に用いて，時期ごとの画像のランキングを行い，提示する手法を提案する．

まず，第 2 章では今回提案する手法についてのべる．第 3 章では実験について述べる．最後の 4 章では本稿のまとめについて述べる．

2. 提案手法

本節では画像間の類似度に基づくグラフを用いた Visual Rank[1] を改良した directed Visual Rank [2]を改良し，画像が属するブログ記事の特徴量，及び時間による類似度の減衰を考慮した画像時系列のランキング手法を提案する．本手法は，まずユーザから与えられたクエリqを入力とし，qを含むブログの記事集合B(q)を，後述するブログデータから取得する．次にB(q)中の記事に含まれる全画像のランキングを行い，時間毎(例えば 1 月毎)の画像のランキングを提示する．以下に，ランキング手法の詳細を述べる．

2.1 テキスト情報を用いたランキング

ここでは画像が属するブログ記事のテキスト情報を用いて画像を重み付けする手法を述べる．まず，各ブログ記事$b \in B(q)$を，出現単語の tf-idf を用いて重み付けする．記事bに含まれる各単語wの重みS_wは以下のように定義される．

$$S_w = \sum_{b \in B(q)} f(w,b) \times idf(w)$$

ただしf(w,b)は，記事b中のwの出現頻度，idf(w)はブログデータ全体におけるwの idf である．

各記事bの重みS_bは，S_wを記事ごとの合計し記事中の単語数で正規化したものとして定義される．

$$S_b = \frac{\sum_{w \in b} S_w}{|b|}$$

ただし|b|は記事bに含まれる単語の数である．

記事bに含まれる画像pの重みS_pは，その記事の重みと同じとみなす．

$$S_p = S_b (p \in b)$$

2.2 画像特徴量を用いたランキング

画像特徴量を用いたランキングには画像間の類似度に基づくグラフに PageRank[3]を適用したランキング手法である，Visual Rank[1]を改良した directed Visual Rank[2]を用いる．

Visual Rank では，画像の類似度に SIFT 特徴量[4]を用い，2 つの画像間でマッチする特徴点の数を類似度としているが，directed Visual Rank では，この類似度に以下のように方向性を持たせ，画像x，yの類似度S(x,y)を以下のように定義している．

$$S(x,y) = \frac{|X \cap Y|}{|X|}$$

ただし，X，Yは画像x，yの特徴点の集合を表わし，$X \cap Y$はマッチする特徴点の集合を表わす．

Ranking Time Series of Images on CGM based on a Directed Graph of Visual Similarity

†Haruka Yamane, University of Tokyo

‡Masashi Toyoda, IIS, University of Tokyo

表 1　評価に用いたクエリの詳細

	クエリ	記事数	画像数	期間
A	おとなグリコ OR オトナグリコ OR “otona grico”	862	428	2008/9～2009/10
B	ロッテ AND フィッツ	481	434	2009/3～2009/12
C	ソフトバンク AND お父さん AND CM	3504	2127	2007/6～2009/12

また，本手法ではさらに，時間が離れている画像は関連性が低くなると見なし，類似度に時間的な減衰要素を取り入れる．x，yの時間差を$t(x,y)$とすると，時間減衰を考慮した類似度S_dは以下のように表わされる．

$$S_d(x,y) = S(x,y) \times e^{-\lambda \cdot t(x,y)}$$

こうして定義した類似度を用い，以下のようにPageRankを用いたランキングを行う．

$$dVR = dA^* + (1-d)p, \qquad p = [p_1, p_2, \cdots]^T$$

ただしpはランダムジャンプの際に用いるベクトルであり，2.1で定義した画像特徴スコアに基づき以下のように定義する．

$$p_i = \frac{S_{p_i}}{\sum_j S_{p_j}}$$

3.　実験

提案手法の有用性を示すため，3種類のテレビCMに関するクエリを用いた実験を行った．

ブログデータとしては我々が2006年より収集しているブログアーカイブを利用する．ブログ記事集合は2006年2月から約100万のRSS，ATOMフィードを毎日収集したもので，その中から検索クエリに対応する記事を取得し，記事中に含まれる画像を抽出した．

用いたクエリに対する，記事数，画像数，出現期間を表1に示す．取得した画像を月ごとにランクングし，ランクの上位10枚について画像が適切であるか評価し，適合率を算出した．評価は各画像について，2:適合，1:関連あり，0:不適合の三段階のスコアを付与する．適合率は(上位10件のスコアの合計)/20で計算した．

テキスト特徴量のみのランキング:T，画像類似度のみのランキング:V，テキスト・画像類似度・類似度の減衰を用いたランキング:T+V+Dの3手法についてランキングを行い，適合率を計算した．結果を図1に示す．全ての特徴を考慮したT+V+Dの平均適合率が，各特徴を個別に考慮したT，Vを上回っていることがわかる．なお，ランダムジャンプのパラメタdの最適な値は0.5～0.7，類似度の時間減衰における半減期の最適な値は半年～1ヶ月と，クエリごとに異なっており，現状では個別に調整する必要がある．

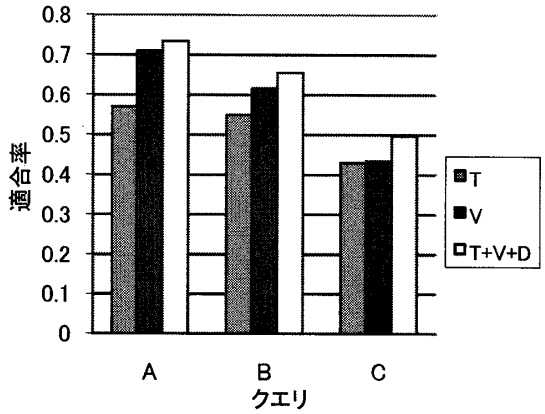

図 1　適合率の比較

4.　まとめと今後の課題

本稿ではブログから時系列的に話題の画像をランキングするため，テキスト情報のみでなく画像情報および画像の類似度の減衰を用いる手法を提案した．また，テキスト情報のみの場合，画像情報のみの場合，テキスト情報と画像情報，減衰を用いた場合について比較実験を行い，すべての特徴量を用いた手法の優位性を示した．今後の課題としてより詳細な特徴量を用いた適合率の改善，および大規模実験，アプリケーションの開発などが挙げられる．

参考文献.

[1] Y. Jing and S.Baluja, “VisualRank Applying PageRank to Large-Scale Image Search,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30, no.11, pp.1877-1890, 2008.

[2] 山根遥香，豊田正史“画像類似度に基づくリンク解析を用いた画像ランキング手法の比較検討”第二回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，2010.

[3] D. G. Lowe, “Object recognition from local scale-invariant features,” Proc. of the 7th international Conference on Computer Vision, pp.1150-1157, 1999

[4] L awrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”, 1998.