5N-3

CGM画像の時系列トレンド可視化システム

上條　哲也† 伊藤　正彦‡ 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

ブログなどのCGM(Consumer Generated Media)上には、商品やCMなどその時々の話題が数多く取り上げられている。CGMを用いることでユーザが自身の意見を自由に記述できることから、マーケティング調査のための情報源として注目を集めている。ブログなどの画像は、ブロガーが愛着を持つものを撮影したものであることが多く、商品の普及や、広報の効果などを測定する重要な情報となりうる。

本論文ではこうした調査を可能にするため、与えられた話題に関連した記事と画像を抽出した上で、画像をクラスタリングして適合度の高いクラスターを抽出し、各クラスターの画像の遷移を時間軸上で可視化するシステムを提案する。提案手法により、クラスターに属する画像の時系列変化を観測することで、実世界の商品、CM、イベントなどの話題の変遷が閲覧可能となる。

2　CGM画像のトレンド時系列可視化

本手法ではCGM画像のトレンド時系列可視化システムを提案する。提案システムは検索クエリに関連する記事と画像を抽出した上で画像をクラスタリングし、ランキングした上位のクラスターを抽出し、各クラスターの内容を表すラベルを生成し、各クラスターの画像とラベルを月ごとの時系列上に可視化する。画像のクラスタリング手法、クラスターのランキング手法は[1]で提案した手法を用いる。本稿ではクラスターの内容をラベル付けする手法および画像時系列の可視化手法を提案する。

2.1　クラスターラベリング手法

クラスターラベリング手法ではCGM画像クラスタリング手法[1]を用い、得られたクラスターの内容を表すラベルとして文を抽出する。[1]で提案したCGM画像クラスタリング手法では、クラスタリング手法として階層クラスタリングを用い、ランダムウォークを用いて重みづけした画像を利用してクラスターに属する画像の重みの総和をクラスターのスコアとしてクラスターのランキングを行っている。

本手法は、各画像の周辺の文に対して単語の重みを用いたスコアを算出しランキングを行う。最も単純な重み付けとしては、クラスターの中の単語の出現頻度をtfとし、idfを考慮した手法が考えられるが、提案手法ではさらに画像の特徴量を利用して単語の重みを算出するタグランキング手法[2]を拡張して用いている。

以降では、分類したクラスターを$C=(c_1,...,c_k)$、クエリをqとした時の画像集合を$X(q)=(x_1,...,x_j)$とし、画像x周辺の文の集合を$L(x)=(l_1,...,l_n)$とする。

本手法ではクラスターに属する各画像の前後3つの文を収集し、各文についてスコアを計算し、ランキングを行う。画像xに存在する文lのスコアを以下のように定義する。

$$score(l,x)=\sum_{w\in l}\frac{tf_w(c_x)\times idf(w)\times TR(w,x)}{|l|} \quad (1)$$

ここで単語wは文lを形態素解析して生成された単語であり、$tf_w(c_x)$は画像xが属するクラスターc_xにおける単語wの出現頻度を表す。$|l|$は文lに含まれる単語の数を表し、$idf(w)$は画像集合$X(q)$における画像を単位としたidfを表す。$TR(w,x)$はタグランキング手法[2]で算出した単語wのスコアであり、以下のように定義される。

$$TR(w_i,x_n)=\frac{1}{|X_i|}\sum_{x_m\in X_i}sim(x_n,x_m) \quad (2)$$

X_iは単語w_iを含む画像集合であり、$sim(x_m,x_n)$は[1]で定義した画像類似度を使用しており、以下のように定義した。

$$sim(x_n,x_m)=\alpha\cdot sift(x_n,x_m)+(1-\alpha)\cdot text(x_n,x_m) \quad (3)$$

$sift(x_n,x_m)$はSIFT特徴量を用い、時間減衰を考慮して算出した画像類似度であり、$text(x_n,x_m)$は画像周辺のテキストのtfidfをスコアとしたコサイン類似度を用い時間減衰を考慮して算出した画像類似度である。αは画像類似度の精度が最大となった0.3を使用した。

2.2　クラスターラベリング手法の評価

本手法の有効性を示すため、テレビCM、商品、観光名所、社会問題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5年間にわたって収集したブログアーカイブを用いたもので、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。評価に用いたクエリ、記事数、画像数および収集した期間を表1に示す。

本評価では、検索クエリに関連した画像をサンプリングし、人手でラベル付けしたクラスターの正解データを用いて評価を行った。評価に用いたラベル数および画像数を表1に示す。ラベル付けした画像を用いて、各ラベルごとに文を抽出し、上位3つまでの文をラベルに属する画像との関連性を考慮して、2(非常に関連している)、1(関連している)、0(関連していない)の3段階評価を行った。

ベースラインとしては式(1)においてタグランキングを考慮しない手法を用い、式(1)で定義した手法と適合率を比較した。

表2はクラスターラベリング手法の適合率の結果である。評価1および評価2を正解とした場合、提案手法ではベースラインと比較して、P@1では11％、P@3では15％精度が上昇した。評価2のみを正解とした場合、提案手法ではベースラインと比較して、P@1では12％、P@3では9％精度が上昇した。

2.3　可視化システム

[1]で提案したクラスタリング手法、クラスターランキング手法、画像ランキング手法[3]および2.1節で提案したクラスターラベリング手法を用いてブログ画像を時系列上に可視化するシステムを構築した。提案手法では3次元空間の3軸を時間、クラスター、画像数とし、各クラスターの画像が時間軸上で遷移していく様子を閲覧することができる。

CGM画像可視化システムでのクラスター軸では、画像を各クラスタごとに色分けし、ランキングの高いクラスターを手前から順に描画した。時間軸では画像が出現した月ごとに画像を描画した。またクラスターのラベルにおいては、各クラスター内の画像がはじめて出現した月にクラスターラベリング手法で算出した上位2つのラベルを描画した。それらにより、各クラスターの話題や画像の出現時期、出現期間が観測可能と

System Visualization for Trend of Time-series of CGM image
†Tetsuya KAMIJO ‡Masahiko ITOH ‡Masashi TOYODA
†University of TOKYO
‡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	評価に用いたラベル数	ラベル付けした画像数
ソフトバンク AND お父さん AND CM	3504	2591	2007/6~ 2009/12	12	91
アフラック AND CM OR まねきねこダック	1909	1324	2009/8~ 2009/12	10	136
おとなグリコ OR オトナグリコ OR "otona grico"	862	487	2008/9~ 2009/10	9	77
マスキングテープ	456	980	2006/4~ 2010/06	8	64
ヒルサイドテラス	1502	3670	2006/4~ 2010/06	9	65
派遣村	16497	6542	2006/4~ 2010/06	11	95

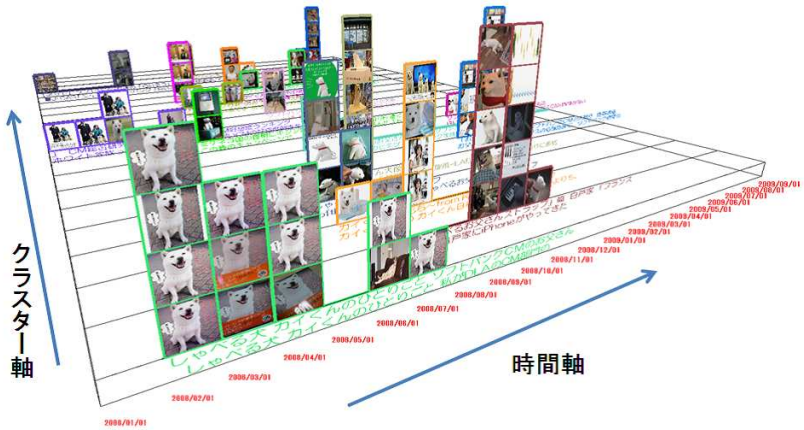

図 1: CGM 画像クラスタリングの可視化例

表 2: クラスターラベリング手法の適合率

	評価 1 および 2 を正解				評価 2 を正解			
	ベースライン		提案手法		ベースライン		提案手法	
	P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3
afrac	1	1	1	1	0.7	0.57	0.6	0.6
softbnak	0.92	0.83	0.92	0.89	0.92	0.64	0.92	0.72
otona-grico	1	0.78	1	0.89	0.56	0.49	0.7	0.59
mt	0.5	0.375	0.625	0.67	0.125	0.125	0.125	0.21
hillside	0.11	0.3	0.67	0.56	0	0.19	0.33	0.33
hakenmura	0.82	0.64	0.82	0.85	0.55	0.33	0.64	0.4
Average	0.72	0.65	0.84	0.81	0.44	0.39	0.55	0.48

なる。また同じクラスターで同じ月に画像が複数出現した場合、画像ランキングの高いものを上から順に描画した。それにより各画像の出現頻度が観測可能となる。

図１にソフトバンクをクエリとした時の画像の可視化例を載せた。クラスターは２０個表示しており、商品や CM、イベントといったクラスターの画像が時系列上で遷移していく様子や、画像数からイベントや商品が話題になった時期が閲覧することができる。例えば、「カイ君のひとりごと」という本や「お父さんストラップ」といった景品のクラスターでは、発売した時期に画像が多数出現し時間が経つにつれ画像が減っていく様子が閲覧できる。また、ソフトバンクのお父さんがホークスの応援隊長となったイベントなどでは、そのイベントが話題になった特定の時期に画像が出現している様子が分かる。

3 まとめ

本稿では [1] で提案したクラスタリング手法、クラスターランキング手法を用いて CGM 画像の時系列トレンド可視化システムを提案した。また各クラスターにおいてクラスターの内容を表すラベルを抽出するクラスターラベリング手法を提案した。

クラスターラベリング手法ではクラスター内に存在する画像に属するブログ記事の周辺テキストおよびタグランキング手法 [2] を単語のスコアとして用いた。タグランキング手法を考慮することにより、考慮しない場合と比較して適合率が上昇することが分かった。

時系列トレンド可視化システムにおいては各クラスターごとの画像の時系列変化を閲覧することにより、イベントや商品、景品などの画像の出現時期、出現頻度を観測することが可能となった。

参考文献

[1] 上條哲也, 豊田正史, "社会分析を目的とした CGM 画像クラスタリング手法に関する一検討", DEIM(2011)

[2] D.Liu,X.Hua,L.Yang,M.Wang,H.Zhang, "Tag Ranking", WWW 2009 MADRID!, Track: Rich Media / Session: Tagging and Clustering

[3] Y.Jing, S.Baluja, "PageRank for Product Image Search", WWW 2008 / Refereed Track: Rich Media, April 21-25, 2008, Beijing, China