特定トピックのブログサイト検索: Wikipedia エントリとブログサイトの対応付けに向けて*

川場 真理子† 宇津呂 武仁† 福原 知宏‡
筑波大学大学院 システム情報工学研究科†，東京大学 人工物工学研究センター‡

1 はじめに

近年，ブログの爆発的普及により，多くの人が個人の関心や評判などをウェブ上で発信するようになった．それに伴い，製品やサービスに対する消費者の意見をウェブ上から取得することが可能になった．商用のブログ検索や評判の抽出分析のサービスには Kizasi.jp や Yahoo!ブログ検索，Google ブログ検索，テクノラティなどがある．また，ウェブの発展により，多くの知識がインターネットを介して得られるようになった．代表的なものとしてはウェブ百科事典である Wikipedia が挙げられる．

現在，ウェブには多くの知識と意見情報が混在しているが，それらの情報を併せ持つ体系ができあがっていないために，知識と意見の両方を同時に得ることは難しい．そこで本研究では，

Wikipedia の中のある特定のトピックから，そのトピックについての意見や評判などの情報が書かれているブログサイトを探し，対応づける

ことを目的とし，Wikipedia のエントリに対応するブログサイトを検索するというアプローチをとる．

2 Wikipedia エントリに対応するブログサイトの検索

2.1 TREC 2007 Blog Distillation タスク

TREC の 2007 年度のブログ検索のタスク [4] に

ある特定のトピック X について検索したときに，そのトピック X について詳しく書かれていて，繰り返し見たいと思うブログを検索する

というものがある．特定のトピック X を与えると，システムは X について長期的に詳しく書かれていて，そのトピック X について興味のある人に定期的に読むことを勧めることができるようなブログサイトを返す．TREC の検索トピックは番号，タイトル，説明，内容で構成されているが，[4] の報告によると，大半の参加者がタイトルのみを索引語として使用することで，各参加者の最高の性能を達成している．そこでこの結果を元に，本稿では，Wikipedia のエントリのタイトルのみでの検索実験を行った．

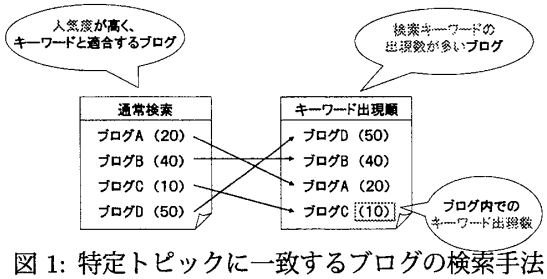


図 1: 特定トピックに一致するブログの検索手法

2.2 本研究の枠組み

本稿では検索トピックについて詳しく述べられているブログの検索を行う．つまり，**検索トピックの出現数が多いブログ**を検索する．具体的には図 1 に示すように，

通常の検索方法でブログを検索し，検索されたブログ集合を検索トピックの出現数が多い順にソートする．

本稿ではこの手順で検索実験を行った．また，本研究の発展として，Wikipedia を用いた多言語ブログの検索があげられる．そのため，本実験の検索トピックには Wikipedia から日本語ブログ，英語ブログ共に，ある程度の数のブログサイトを集められそうなトピックを 60 選んで日英のブログを検索する実験を行った．

2.3 評価手順

ブログを検索するために，本実験では日本語ブログの検索には，Yahoo!Japan 検索 API を，英語ブログの検索には米 Yahoo!検索 API を利用し，日本語ブログでは大手 11 社[1]，英語ブログ検索では大手 10 社[2]のドメインに限って検索を行った．

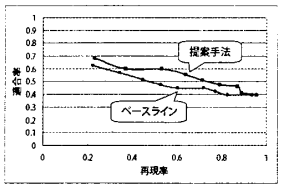
本実験では，指定したドメインにつき 100 の検索結果を取得する．API での検索の場合ブログサイト単位での検索ではなく記事単にの検索になるので、同一著者のブログは一つのブログサイトにまとめるという作業を行っ

*Blog Distillation towards Linking Wikipedia Entries to Blog Feeds

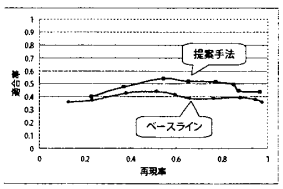
†Mariko Kawaba, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,
‡Tomohiro Fukuhara, Research into Artifacts, Center for Engineering, University of Tokyo,

[1]FC2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesa.net, jugem.jp, yaplog.jp, webry.info.jp, hatena.ne.jp

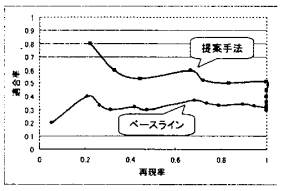
[2]blogspot.com, msnblogs.net, spaces.live.com, livejournal.com, vox.com, multiply.com, aol.com, blogsome.com, wordpress.com



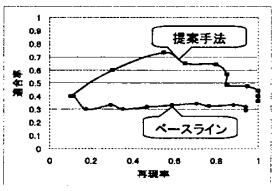
英語ブログ：4 キーワード



日本語ブログ：4 キーワード



英語ブログ：Yasukuni Shrine



日本語ブログ：靖国神社

図 2: 日英ブログ検索の評価結果

た．その結果，1 キーワードあたり得られたブログ数は日英共に約 500 ブログとなった．また，用意した約 60 キーワードの内，ドラゴンボール，Wii，新世紀エヴァンゲリオン，靖国神社の 4 キーワードを選び，それぞれ，上位 30 位と以下等間隔に 30 キーワードをサンプリングし，手動で評価した．また，手動評価の際，特定トピックについてある一定数以上のブログ記事があれば正解とし，一定期間特定トピックについて書かれているということは考慮していない．また，複数のドメインを一度に指定して検索し，API の出力順にブログサイトをランキングしたものをベースラインとした．また，ベースラインとの比較のために、API 検索の出力をさらに検索キーワードの出現数順にリランキングしたものをシステムの出力として再現率適合率のグラフにした．これらのうち，全体の評価結果と靖国神社の評価結果を図 2 に示す．

2.4 評価結果

全体的に，ベースラインと比較して，提案手法の方が精度良く検索できているが，いくつかの検索誤りが見られた．ドラゴンボール，新世紀エヴァンゲリオン，Wii などは多くの関連製品が発売されているものは，日本語のブログではアフィリエイトの文字のなかに検索トピックが多く出現するものなどが多くみられた．英語のブログでは，ブログの内容とは関係のない，自分の好きなものリスト一覧などに商品名やアニメ作品名などが記載されていたために，誤って検索してしまうという例が多くみられた [3]．またキーワードの出現数としては，現在 API が出力するヒット数をそのまま用いているが，ノイズが多いため，今後はブログ記事を直接解析して出現数を求める必要がある．

2.5 Wikipedia を用いた検索質問拡張

TREC の 2007 年度ブログ検索タスク [4] で，Wikipedia のハイパーリンクを用いた手法 [1] が最高の性能を達成している．このことをふまえ、我々はタイトルのみでのブログの検索では不十分と考え，Wikipedia の 1 つのエントリを用いて検索質問を拡張する実験を行っている．この実験はブログ検索タスクで行われた検索質問拡張 [1] の手法に加えて，検索トピック X の索引語の候補となる Y を検索エンジンのヒット数を用いた関連度 (X AND Y の検索ヒット数/X OR Y の検索ヒット数) で，ランキングするという手法を用いている．

3 日英ブログの言語対照分析

[2] では，日韓中英のブログ内で，キーワードのバーストの時系列の変化を各言語間で調べるという研究がされている．一方、本稿ではキーワードのバーストの時系列変化を見るのではなく，同一トピックにおける日英のブログの内容を分析し比較した．その結果について簡単に述べる．

捕鯨と靖国神社のような社会問題について書かれた日本語ブログでは右寄りのブログが多く，肯定的な意見が多く見られた．一方，英語ブログでは日本の捕鯨や靖国神社参拝などに否定的な意見が多く見られた．また，ドラゴンボール，Wii，新世紀エヴァンゲリオンなどのアニメ作品や商品について書かれたブログでは，社会的な興味の違いから，記事の内容に違いがみられた．具体的には，日本語のブログの場合，感想や紹介などが主で，著作権などの問題もあり，画像や動画が記載されていることは少ない．一方英語のブログの場合，動画や画像などが載せられているブログが多く見られた [5]．

4 まとめと今後の課題

本稿では Wikipedia とブログ集合の対応付けのために，トピックの出現回数の多いブログを検索することでブログ集合を検索する検索実験を行った．また，日英のブログの分析から，同一トピックに対する日本語のブログと英語のブログの違いを知ることができた．今後，TREC のブログ検索タスクの結果なども取り入れて，より精度の高いブログの検索を行うと共に，検索質問拡張の実験などを行う予定である．

参考文献

[1] J. Elsas, J. Arguello, J. Callan, and J. Carbonell. Retrieval and feedback models for blog distillation. In *Proc. TREC-2007 (Notebook)*, pp. 170–175, 2007.

[2] 福原知宏, 宇津呂武仁, 中川裕志. 複数言語間の語彙出現傾向比較による言語横断型ウェブログ関心解析システムの開発. 言語処理学会第 13 回年次大会「大規模 Web 研究基盤上での自然言語処理・情報検索研究」 ワークショップ論文集, pp. 40–43, 2007.

[3] 川場真理子, 宇津呂武仁, 福原知宏. Wikipedia エントリに対応するトピックのブログサイト検索. 言語処理学会第 14 回年次大会論文集, 2008.

[4] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2007 blog track. In *Proc. TREC-2007 (Notebook)*, pp. 31–43, 2007.

[5] 中崎寛之, 川場真理子, 宇津呂武仁, 福原知宏. 同一トピックの日英ブログサイト検索による二言語対照ブログ分析. 言語処理学会第 14 回年次大会論文集, 2008.