ブログを用いた「なんでも早期発見システム」の開発
― 細かい流行も捉えるブログ分析システム ―

1. 背景

現在、日本で100～200万人が定期的ブログを書いているといわれており、コンピュータに詳しい人によるブログだけでなく、一般的な主婦や学生によるブログも多い。これらは一般の人達の率直な声を反映しているといわれており、マーケティングのための市場調査・効果測定や製品開発計画のための安価かつ迅速に利用できる情報源として有望視されている。実際にビジネスに応用する試みもはじまっている。

ブログに書かれた情報を利用して社会動向や流行を捉えようという試みは既に多数あり、有用なものとしては、例えば入力したキーワードに対してブログ上での登場頻度を返すようなサービスや、ブログ全体の動向を分析して流行語を抽出する試みがある。

既存のシステムでは、ブログを収集して単語ごとに出現頻度を数えて流行を分析するといった、ブログ全体をマクロな視点でみる分析法が主流であった。このような手法ではブログ界全体に影響を与えるような大きな変化を捉えることはできるが、ブログの書き手には偏りがあるため（ 都市部の20,30代が多い）、流行分析の結果も自ずと偏りがあるものになる。また、実際のマーケティング戦略の企画では、ターゲットを絞った戦略をとることが多いが、既存のブログ分析法ではこのような目的を達成できない。

2. 目的

今回、ブログをその著者の属性を推定することにより、男女別などのターゲットを絞った分析を可能にするシステムの実現を目的とした。細分化して数が少なくなったデータからも流行検出などの時系列分析ができるようになれば、既存のシステムでは発見できなかった「ちょっとした」変化も発見することができる。例えば既存のブログ全体を分析する手法では、10代で流行しているアニメの名前などは20代， 30代の流行に埋もれてしまい、知ることはできない。一方今回目的としたシステムでは、10代のみに対象を限った流行の分析できるため、このような細かい流行も検知することができる。

3. 開発の内容

本プロジェクトでは、個々のブログ著者の属性を推定することにより、より細かい流行の分析ができるような手法の確立とシステム構築を行い、この機能を一般ユーザーが利用できるようにすることを目指した。

この目的を達成するため、下記の各項目について研究・開発した。

- 低コストで、大量データを処理でき、障害耐性の高い計算機システムの構築
- 大量のブログを安定して収集し、後処理しやすい形で保管するブログクローラの開発
- 高速に分析できる全文検索インデックスの構築
- ブログのテキストから、高速に著者属性を汎用的に推定する手法の確立
- 上記手法で性別・年齢層・居住域を推定するために必要なデータの収集
- 400万ブログの著者属性を推定する分散処理プログラムの開発
- 確率分布を与えられた時系列イベントから流行を検知するアルゴリズムの研究
- 単語ごとに性別・年代・地域別に流行度スコアを計算するプログラムの開発
- Webフロントエンド

これらのコンポーネントを組み合わせ、 ブログを収集し、リアルタイムに分析を行うシステムを構成した。システムの構成は図1の通りである。

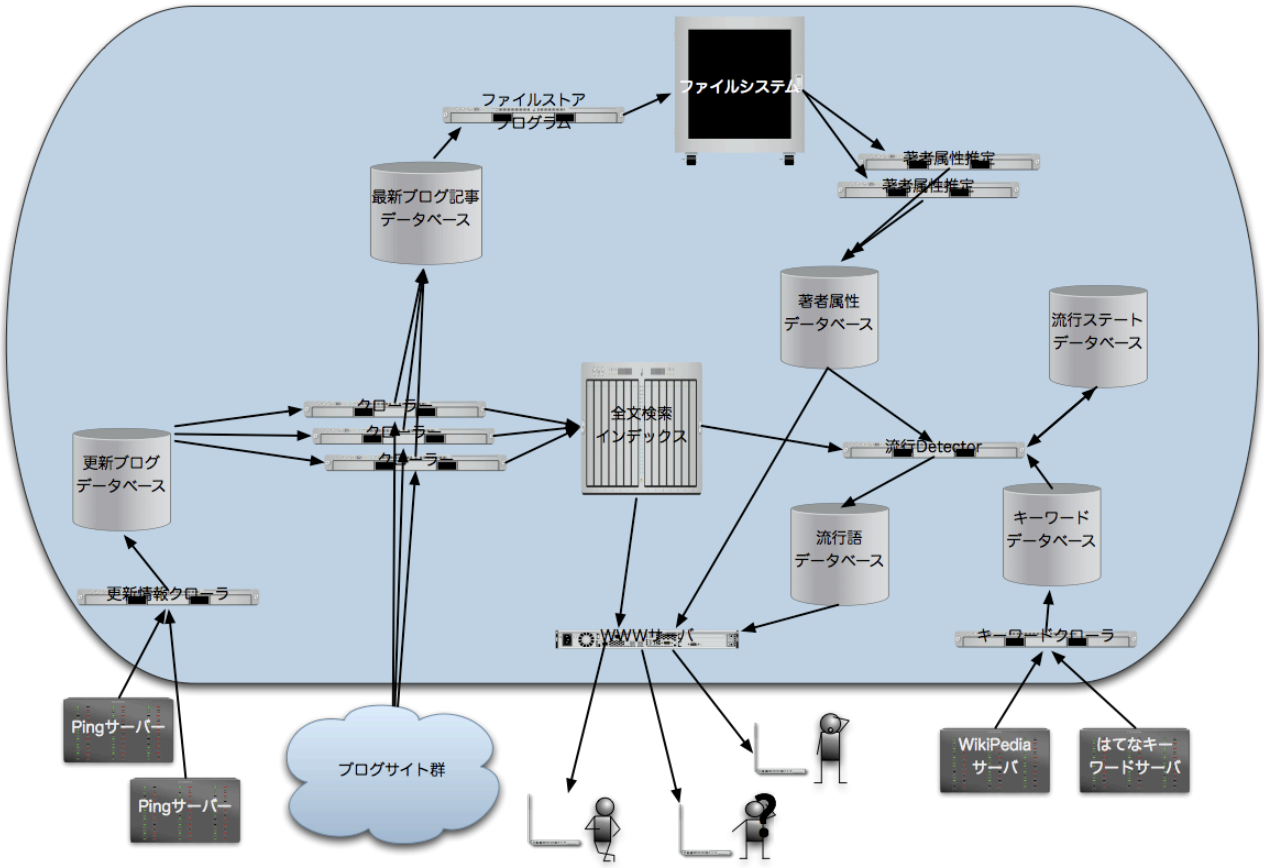

図1　システム全体の構成

このシステムによって、下記2点の機能をwebから利用できるサービスとして提供した。

- 性別・年齢層・居住域ごとの流行語ランキングをリアルタイムで表示する機能
- ランキングで表示された語（もしくはユーザーが検索ボックスに入力した任意のキーワード）について、日付別の使用頻度、性別による差異、年齢層による差異、都道府県による差異を分析し、グラフで表示する機能

サービスのトップ画面を図2に示す。また、30代でのランキングを表示した場合を図3に示す。これらのページを用いて、ユーザーは現在話題になっているものを属性ごとに知ることができる。

図2　サービスのトップ画面

図3　属性別ランキングの例（30代）

ランキング中のキーワードをクリックがされた場合、図4のようなキーワードの詳細ページを表示する。この画面ではこのキーワードが使われたブログ記事をリストアップすることでこのキーワードが話題になった原因を推察できるようにすると共に、このキーワードがどの程度急に話題になったのか、話題にしている人の男女比、年齢構成、主な居住域はどこなのかを表示することで、ユーザーがマーケティング等の際に必要となるセグメントに関する情報を提供する。

これらの機能により、本プロジェクトの目的としていた「細かい流行分析」が実現することができた。

図4　キーワード詳細画面

4. 従来の技術（または機能）との相違

既存のブログ分析システムとの大きな違いは、ブログの著者について属性分析を行えるという点である。これにより、話題になっているキーワードを「どんな人が話題にしているのか」を知ることができる。さらに、新しく開発した属性推定結果をうまく利用する流行分析アルゴリズムを用いることで、10代などブログの書き手が少ない属性の間での流行も適切に検出できるようにした。

また、他のサービスのアルゴリズムが不明なため厳密ではないが、他の類似サービスに比べてより迅速に新しい話題を捉えられているという意見もいただいている。また、ブログスパムが少ないという評価もいただいている。

5. 期待される効果

消費者の要望を正しく把握できれば、それを次の製品に生かすことができ、企業はより正しい方向を目指して製品・サービスの開発にしのぎを削ることができる。 このように、消費者の動向をより的確に調査できるということは、サービスを提供している企業のみならず、消費者、そして社会全体にとって有用といえる。ブログを用いたマーケティング調査は安価かつ迅速に実現可能であるため、 今回のプロジェクトで開発したような技術が普及すれば、消費者の声が製品に反映される機会は大幅に増える。これにより、消費者の求めるものを頻繁に調査し、迅速に開発に生かすことが可能となる。今回のプロジェクトは、長期的な視点でみれば生活を豊にしていくスピードを向上させることができるものだと考えている。

また、今回開発した著者属性推定技術はブログへの広告貼り付けに応用できるものである。Googleなどが提供するコンテンツマッチ広告はwebページに出現する単語を元に似合いそうな広告を配信する仕組みであるが、著者属性推定技術を用いれば20代女性のみに広告を出向すると言ったことが可能になり、より広告効果の高いインターネット広告を実現することができると考えられる。

6. 普及（または活用）の見通し

12月13日にblogeyeという名称のサービスとして公開し、1月中旬までに20,000ページが閲覧された。数多くのニュースサイトでも取り上げられ、またブログでの感想も多く得られた。

本プロジェクトの成果はシステムを実現したことよりも、むしろこのシステムの実現を可能にした新開発の著者属性推定技術、流行分析技術であると考えている。ブログ分析はビジネス目的で実利用が始まっており、このようなB2Bサービスを提供している企業への導入を目指している。

7. 開発者名（所属）

大倉　務（東京大学大学院　情報理工学系研究科　修士課程）

（参考）

サービスを公開しているサイト

http://blogeye.jp/