

インターネット上のテキストデータの推移に関する 研究のための計算機システムの開発

Development of Computer System for Time Trend Analysis of Text Data on the Internet

鈴木 肇

Hajime SUZUKI

1. はじめに

コンピュータはかつては電子計算機と呼ばれていたように、元来は数値データを扱うために開発された機械であった。しかしコンピュータや通信技術の発展と共に、次第に文字データを扱うことも可能になり、近年では多くの文書はワードプロセッサで作成され、電子化された文書が世界にあふれている。一方でインターネットの爆発的な普及で電子化された文字データの多くはWeb技術を使って世界で参照可能となっている。当初はそれを人が読めればそれで済んでいたが、最近ではインターネット上にあるデータはもはや一個人で読める量を遥かに超えてしまい、検索エンジンなどの助けを借りないとも扱えるものではない。そして検索できたとしても、やはりかなりの量が依然として残ってしまうのが普通である。そこで、そのような膨大なデータからどうやって自分が必要とする情報を取り出すかということに関心が集まりだした。しかしコンピュータが自然言語で書かれた文書の意味を理解するには、まだ技術が十分に達していない。コンピュータによる自動要約というものも研究は進んでいるが、¹⁾ 自動要約以外の方法でコンピュータに文章の意味を理解させることなく、大量の文書データから情報を引き出す方法がないだろうか。これが工学的な見地から見た研究動機である。このような必要なデータの収集とその処理はテキストマイニングと呼ばれ、各方面から期待が集まっている分野である。²⁾

また、インターネットが社会に及ぼす影響は近年ますます強まってきている。インターネットの技術が社会に及ぼす影響は当然として、最近ではインターネットのメディアとしての重要がますます大きくなってきている。そこでインターネット上の情報を実社会の動向と比較するということは興味深い。これが社会科学的見地からの研究動機である。コンピュータは文字データを扱えるといっても、依然としてその中心では数値データが扱われている。そこで社会情勢を解析するとしても、社会情勢を表すも

の、例えば「話題」といったものを数値化しないとコンピュータでは扱えない。このようなものをどうやって数値化していくかという問題がある。

本研究では「話題」というものを数値化し、インターネット技術、コンピュータ技術を利用して社会情勢を捕らえていく手法を開発した。

2. インターネット検索技術

Google³⁾やYahoo⁴⁾表されるインターネット上の検索をサービスしているサイトがある。すでに述べたようにコンピュータは自然言語を理解できない。そこで、まず文書を単語レベルに分解する。英語などは初めからスペースで単語ごとに区切られているが、日本語なども辞書ファイルを利用することで単語ごとに分解することが可能である。コンピュータはそれぞれの単語の意味は理解できないが、ある文書中にある単語が含まれているか否かは調べることが出来る。これが検索技術の基本である。

検索サービスを行っているサイトでは、基本的には世界中の公開されているWebサーバを定期的にスキャンして、公開されている情報を調べている。そして文章を単語レベルに分解してインデックスファイルを作成する。インデックスファイルは索引であり、単語とそれが存在するものとのHTMLやPDFファイルなどの存在場所(URL)とを対応付けている。インデックスファイルは常に更新され、最新の状態が反映するようにしている。検索サービスを利用しようとするユーザーが検索ワードを入力すると、インデックスファイルが参照され、その単語を含むファイルとそれが存在するURLがユーザーに示される。一般に検索結果のファイルは多く存在するので、結果も何行にわたって表示される。そこでこれをどのような順番で表示していくかというのが検索サービスの腕の見せ所であり、ユーザーが望む情報になるべく上のほうに来るようにさまざまな工夫をこらしている。⁵⁾ただ、この優先順序のつけ方は本研究には関係ないので、詳細については省略する。

検索サービスは基本的にはこのような仕組みで動いており、大変有用なものであるが、いくつか問題点はある。まず世界中から情報を収集しているため、検索結果後もかなりのデータが表示される。検索サイトの並び替えの工夫の結果、初めのページに表示されるファイルは比較的信頼がおけるものであることが多いが、基本的には玉石混交と言ってよい。その結果、検索結果のファイルを開き、人間が内容を読んで内容が信頼できるかどうか判断しなくてはならない。これは、なるべく人間がいちいち文書を読まずに情報を取り出したいという本研究の趣旨に反する。またインデックスファイルは常に最新の情報に更新されているので、ある話題の時間的变化というのが解析しにくい。今現在の状態はわかるが、過去にさかのぼって解析するのが難しいのであ

る。⁶⁾ また、インデックスファイル自体は新しくても、元のファイルが古い情報のまま更新されていないということもある。そのため、時間変化がそのようなファイルのためにぼけてしまうということもある。そのようなことを踏まえ、新しい検索システムを開発した。

3. FAMAシステム

FAMA⁷⁾ システムはLinux上で開発されたシステムである。開発に当たり、中部大学の山口作太郎教授、武者小路公秀教授、山口貴行氏の御意見も参考にさせていただいた。基本的にはGoogleなどの検索サービスと同様のことを行っているが、いくつか改良を加えている。まず全世界のWebサーバを対象にすると、すでに述べたように信頼できる情報と出来ない情報が混ざってしまうので、信頼できるサイトのみを対象とする。具体的には日本語で情報を公開している朝日新聞、読売新聞、毎日新聞、日本経済新聞、朝鮮日報(日本語版)、人民日報(日本語版)のサイトを捜査対象とした。これらのサイトは内容に信頼が置け、情報の更新も速やかに行われるので常に最新の話題が提供され、今回の研究に都合が良いと判断した。これらのサイトは一日一回訪問し、`wget` コマンドでファイルを獲得し、`chasen`(茶筌)⁸⁾ というシステムで単語(形態素)に分解して、`namazu`⁹⁾ というシステムでインデックスファイルを作成している。インデックスファイルは作成するごとに名前を変更し、過去のインデックスファイルも全て保存している。元のHTMLファイルも保存しているが、量が多いため`tar` コマンドでアーカイブを作り、`gzip` コマンドで圧縮して保存している。以上の操作は、`wget` から `gzip` までを一括処理するスクリプトを作成して、`crontab` を使って一日一回夜間のネットが比較的空いている時間に自動運転している。一日にインデックスファイルとその関連ファイルで約25MByte、元のHTMLファイルで圧縮後に約20MByte程度のデータが蓄積される。

検索をかけるときに、どのインデックスファイルを参照するかを指定できるため、ある時点で、ある単語が上記のサイトのどのファイルに存在していたかを調べることが可能である。¹⁰⁾ 検索につかうキーワードは複数指定することが可能であり、Aという単語を含むファイル、Bという単語を含むファイル、その両方を含むファイルを検索することが可能である。さらに検索キーワード、調査開始日と終了日を指定することで、その間の一日ごとのヒットしたファイル数を出力することが可能となっている。これはCSV形式で書かれているのでExcelなどでグラフ化することが可能である。開発当初はFAMAシステムがLinux上で動いていることもあり、漢字コードにEUCを使っていたため、データをWindows機にダウンロードしてExcelで処理しようとする

ると漢字コードを変換しなくてはならなかったが、これも結果の出力の漢字コードを shift-JIS に変換するルーチンを組み込み、そのまま Excel で処理できるように改良した。結果の出力は 1 年分を調査しても数十秒で完了する。

図 I 新潟中越地震の話題の推移

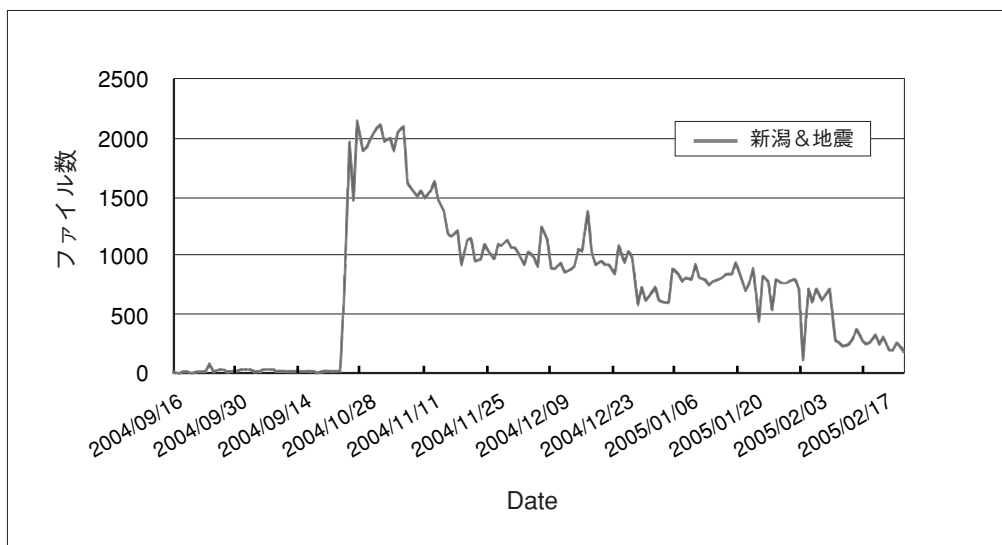
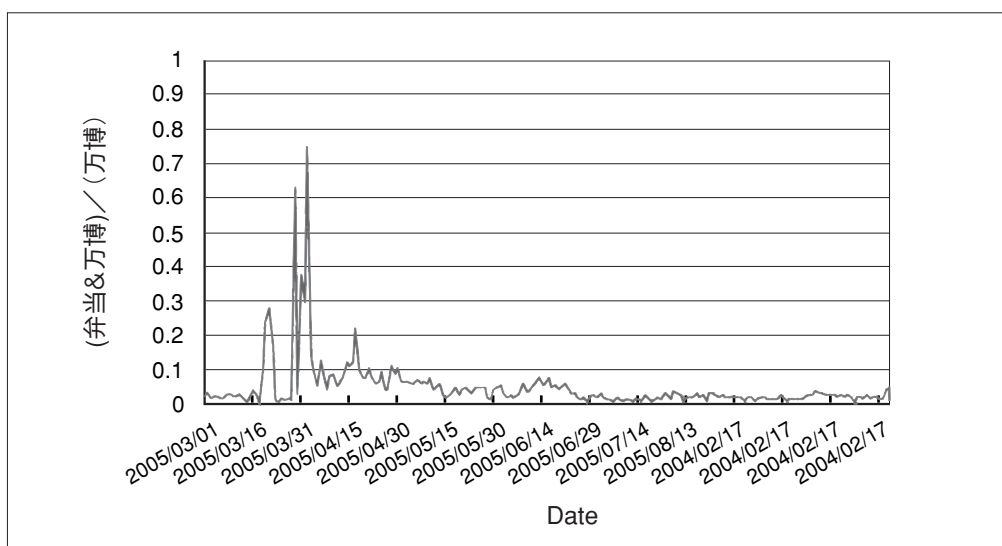


図 II 愛知万博の弁当持込問題が占めた比率



図Ⅰは2004年10月23日に発生した新潟中越地震前後の「新潟」と「地震」を含むファイル数の時間変化である。「新潟」だけだと地震に関係ない記事もあり、同様に「地震」だけでも新潟と関係ない記事があるが、「新潟」「地震」を両方含む記事は新潟中越地震関連の記事と見てよいだろう。「新潟」「地震」を両方含む記事は、10月23日以前はほぼ0であったのに対して、10月23日の地震発生を同時にページ数が増加したことがわかる。このようにある単語の出現頻度という形で話題を数値化し、その時間的変化を追跡することが可能である。

また図Ⅱは愛知万博開催当初の「万博」という語を含むファイルの中で、「万博」と「弁当」の両方を含むファイル数を解析した結果である。2005年3月25日に開幕した愛知万博は、開催当初は弁当の外部からの持込が禁止されており、そのことが問題になった。¹¹⁾ この措置は4月1日から解除されたわけであるが、多いときは万博関連の記事のうち、実に6割以上が弁当問題を扱っていたということがこの解析結果からわかる。このように、FAMAシステムを使うことで、社会の動向を新しい切り口で解析できると考えられる。

4. この後の課題

今回の研究は2年計画であり、一応の区切りはついたが、やり残した部分、今後さらに発展できるであろうという部分がある。本研究の目的は、インターネット上の話題と社会情勢との関わりを調査するというものであった。そのために今回は「話題」というものを数値化して解析する手段を開発したが、今後は社会情勢を表すものを数値化してコンピュータの解析に載せられるようにしなくてはならない。比較的簡単なものとしては、例えば株価の変動などが挙げられる、これはすでに数値化されたデータなので株価変動と、FAMAシステムが出力する話題の時間変化に相関があるかどうかを解析するのはそれほど難しくないだろう。一方で社会が抱える漠然とした不安などは数値化しにくい。インターネットの犯罪報道が社会不安とどのように結びつくかといったことは興味深い研究であろうが、これにはまず社会不安をどうやって数値化していくかといった解析手法の開発が必要となる。

また、今回のシステム開発では社会で話題性のある地震や万博などを使って正しくグラフ化できるかを検討しながらシステム開発を行った。これらの話題は確かに瞬間的にファイル数が増加するので、結果がクリアに見えた。しかし、それほどファイル数は増加しないが重要な変化というものもあるはずである。こういうものはファイル数の絶対数だけでなく相対的な変化を捉えることも必要だろう。たとえば普段はめったに登場しない単語がある期間だけ集中して登場したとすると、仮にファイルの絶対数が少なくてもその単語はその期間の重要な単語であったはずである。ある単語の

平均的登場頻度というようなものがわかっていると、このような解析も容易に出来るはずである。このようにもう一段高度な文字データ解析を持ち込むことで、さらに見えてくる情報というものがあると考えられる。高度な文字データ解析を組み合わせたデータ解析手法を開発していきたい。

注

- 1) コンピュータによる自動要約の解説は、例えば「知の科学 テキスト自動要約」(奥村学・難波英嗣著 オーム社)、あるいは「自動要約」(Inderjeet Mani 著, 奥村他 訳, 共立出版)がある。
- 2) テキストマイニングの参考書としては「Excelで学ぶテキストマイニング入門」(林 俊克著, オーム社)などがある。
- 3) <http://www.google.co.jp> あるいは <http://www.google.com>
- 4) <http://www.yahoo.co.jp> あるいは <http://www.yahoo.com>
- 5) Google がどのような順序で検索結果を表示しているかは「GOOGLE HACKS」(Tara Calishain, Rael Dornfest 著, 山名他 訳, オライリージャパン) に解説がある。
- 6) これまで行われたインターネット上の単語の出現頻度を調査したものとして「技術革新や政策等の外部環境とベンチャー企業の消長の相関—データマイニングシステムの構想と運用試験—」(高橋道郎, 永井義明, 「中部大学産業経済研究所起用」第14号) などがあるが、これはあらかじめ調査語を決めておいて、毎日その推移を調査した。
- 7) Fama はギリシャ・ローマ神話の噂を司る神である。fame (名声) の語源となっている。
- 8) chasen (茶筌) は奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室)によって開発された日本語の形態素解析を行うフリーソフトである。
詳細についてはたとえば <http://chasen.naist.jp/hiki/ChaSen/> を参照。
- 9) namazu は Namazu project の有志が開発を進めているフリーの日本語全文検索システムである。詳細についてはたとえば <http://www.namazu.org/> を参照
- 10) 実は Google でも過去のインデックスファイルの参照や表示サイトの限定をすることが可能である。詳細については注 3) で挙げた参考文献を参照。但しこの方法は、結果が出るまで FAMA システムよりかなり時間がかかる。
- 11) 2006年3月25日に開幕した愛知万博は、開催当初は食中毒対策及びごみ対策で弁当の外部からの持込は禁止され、会場内のレストランの使用が奨励されたが、この措置に入場者から不満の声が爆発して新聞等でも大きく取り上げられた。小泉首相がこれに対して疑問を述べた

インターネット上のテキストデータの推移に関する研究のための計算機システムの開発

ことがきっかけとなり、問題は急速に解決に向かい2005年4月1日から弁当の持込は可能となった。