

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 —

Development of Data Mining System for Studying Correlation between Technology Progress and Venture Businesses

高 橋 道 郎

Michio TAKAHASHI

永 井 義 明

Yoshiaki NAGAI

1. はじめに

技術革新や政策等の外部環境の変化がベンチャー企業の消長にどのように影響を及ぼしているかを研究するに当たり、外部環境の測定手段を検討した。外部環境を定量的に表現する手段としては、事業所数、出荷額などの工業統計や、融資額等の金融統計が一般的であり、各種のアンケートも多用されている。しかし、統計値は事後の値であるとの宿命を持ち、またアンケートには多大の時間と労力を要するなど、外部環境の表現手段としては完全なものとは言えない。

そこで全く新たな計測手段として、ウェブ上の記事や論説を計数・分析するデータマイニングシステムを立案した。人文科学系の研究に威力を発揮する新たなツールに発展することを願って開発中である。このシステムの概要と試行結果を報告する。

2. データマイニング装置の構成と機能

2.1 システム概要

WEBサイト（20箇所程度）を巡回して記事を検索し、この記事をもとにテキストマイニングを行うことにより、キーワード、論説等が時間の経緯とともにどう変化するのか、ビジネスとの関わりはどうかなどを調査検討するための研究プラットフォームである。なお、情報技術の世界では、大量のデータから有益な情報を掘り起こす技術をデータマイニングと称し、この中で構造化されていない文章データを分析対象とする技術をテキストマイニングと言っている。第2章では正確を期するためにこの2つの用語を使い分ける。

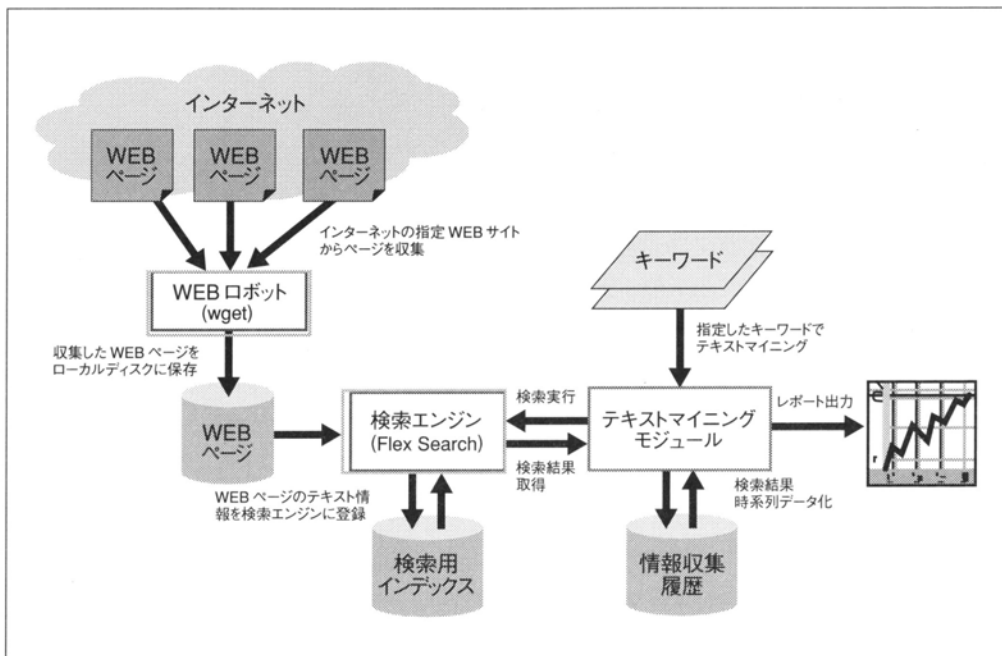
開発したシステム（Phase 1）は、基本的なプラットフォーム上に下記の機能を有する。

- ・ 指定されたWEBサイトを巡回収集するロボット機能
- ・ キーワードの表記上の振れを検索可能なあいまい検索機能
- ・ 検索にヒットした記事リストの出力，記事数の推移のグラフ出力を行うレポート機能

2.2 機能構成

本システムの機能構成を図1に示す。

図1 ITプラットフォームシステム（Phase 1）の機能構成



2.3 機能概要

2.3.1. WEB ロボット

(1) WEB ページ収集

定期的（毎週1回定時など）にWEB ページ収集を行う。

収集対象ファイルはHTML（*.html, *.htm）とする。

(2) WEB ページ収集ツール

WEB ページ収集ツールには、GNU wget（フリーソフト）を使用する。

GNU wget は、非対話型 http による WEB ページ収集ツールで、URL リストを記述したファイルから開始 URL を読み取り、リンクを辿って指定階層まで WEB ページを収集することができる。

タイムスタンプ機能により未更新のファイルを重複してダウンロードしないことにより、転送時間を節約することができる。

ダウンロードしたファイルは一旦ローカルファイルに蓄える。

(3) WEB 同期

GNU wget は蓄積型の WEB ページ収集ツールのため、WEB ページの削除、リンク切れを検出できない。したがって、ローカルディスクの WEB ページとインターネットの指定 WEB サイトの状態の同期を取るため、削除、リンク切れのあったローカルファイルを削除する WEB 同期機能を作成する。

2.3.2. 検索エンジン

(1) キーワード検索

WEB 収集ツールにより収集した WEB ページファイルに対し、指定キーワード毎に検索を行い、検索ヒット件数およびヒットした記事の出典 (URL)、ファイル日付を出力する。

(2) キーワード検索エンジン

検索エンジンには、FlexSearch for Web (アドイン研究所)^{4) 5)}を使用する。

FlexSearch for Web (以下 FlexSearch) は、キーワードおよびテキストコンテンツの表示上の振れを吸収した照合・検索が可能な検索エンジンである。

2.3.3. テキストマイニングモジュール

(1) 記事抽出

キーワードリストファイルに記述されたすべてのキーワードについて、FlexSearch による検索を行い、得られた検索ヒット件数、記事の出典 (URL)、ファイル日付をキーワードと共にアーカイブ (記事収集履歴) に格納する。

(2) 記事収集履歴管理

検索に使用したキーワード、検索された記事の数、出典 (URL)、記事ファイル日付およびそれら元データから算出される評価指標等情報を管理する。

評価指標には次のものを想定する。

- ・ あるキーワードにおける検索ヒット数の時系列推移
- ・ キーワード間での時系列推移の相関

(3) レポート出力

…検索結果記事リスト

検索された記事の結果（記事の数や出典）をリスト形式に出力する。

…キーワード別記事数時系列グラフ

縦軸に週当たりの記事の数，横軸に月日を表し，記事の数の推移をグラフに出力する。

…キーワード間相関グラフ

キーワード毎に，記事数の増減の変化量に基づく指標値を算出し，各キーワードの指標値の分布をグラフに出力する。

2.4 稼働環境

(1) サーバ仕様

本システムは，下記の稼働環境を想定する。

- ・ C P U Pentium 3 1.0 GHz 以上
- ・ メモリ 1.0 GB 以上
- ・ H D D 40 GB 以上
- ・ O S Windows 2000 professional

(2) セキュリティ

本システムは，インターネットからの情報収集を行うシステムであり，インターネットへの情報（サービス）配信は行わない。本システムのセキュリティは，本サーバが設置されるネットワークのファイアウォール等セキュリティ環境に依存する。

本サーバ単体でのセキュリティ対策は行わない。

3. 運用

2003年6月より11月までの5ヶ月間にわたり，キーワードとして「特許，産学連携，SARS，リストラ，ベンチャー，起業」の6つを選定し，そのキーワードの現れる記事数を計数した。対象としたサイトは以下に示す新聞社等である。

表 1. 収集対象としたサイトの一覧

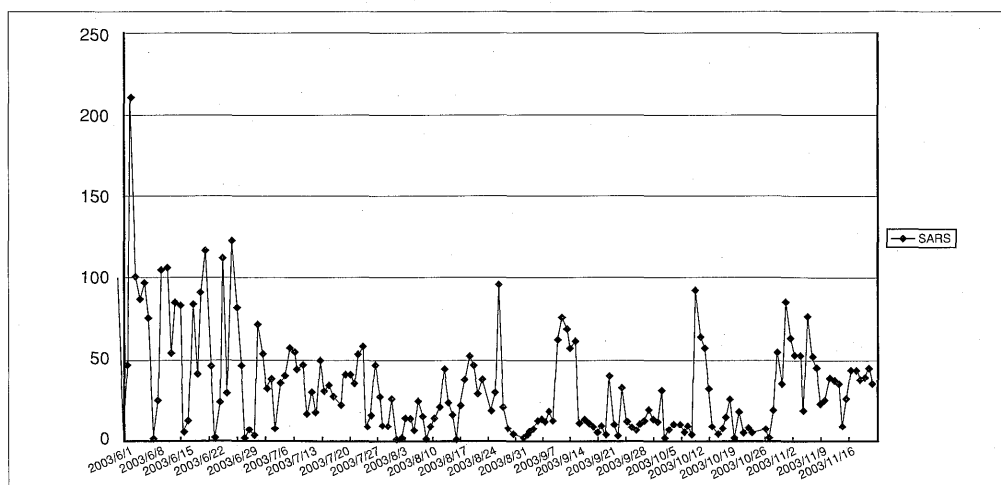
対象サイト名	URL
日本経済新聞社	http://www.nikkei.co.jp/
朝日新聞社	http://www.asahi.com/
読売新聞社	http://www.yomiuri.co.jp/
共同通信社	http://www.kyodo.co.jp
時事通信社	http://www.jiji.com/
ロイター通信社	http://www.reuters.co.jp/
ニューズウィーク日本	http://www.nwi.ne.jp/
日経産業新聞	http://ss.nikkei.co.jp/ss/
ニフティ	http://www.nifty.com/category/news/
ヤフー	http://www.headlines.yahoo.co.jp/hl
産経新聞社	http://www.sankei.co.jp/
東洋経済新報社	http://www.toyokezai.co.jp/

4. 実験結果

下記はこれらのサイトにキーワード「特許，産学連携，SARS，リストラ，ベンチャー，起業」を含む記事が合計何個あるかを日毎に計測した値である。横軸は日，縦軸はこれらのキーワードを含む記事の数である。

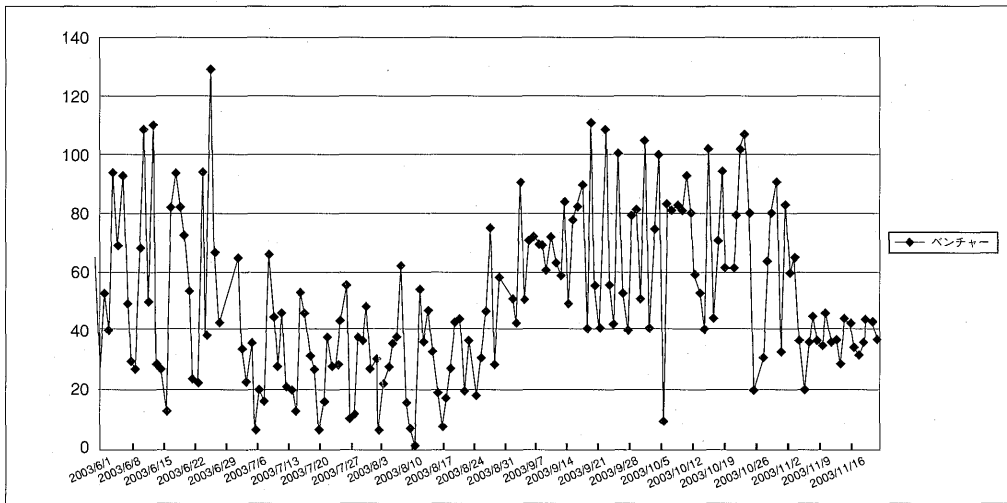
まず始めに新型肺炎 SARS をキーワードとした計測結果を図 2 に示す。

図 2 SARS をキーワードとした計測結果



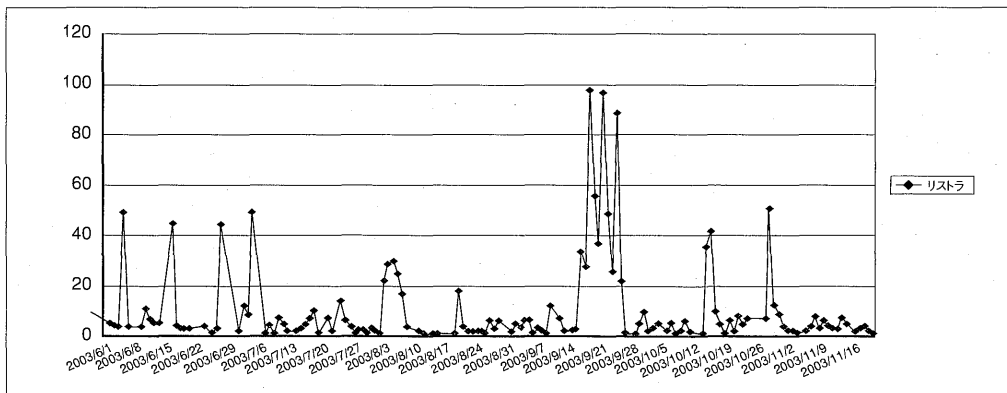
SARSは2003年正月から春先にかけて猛威をふるっていたが夏には沈静化した。キーワードヒット数をみると6月と7, 8月の間には明らかな差があり, キーワードヒット数から記事も沈静化していることがよく分かる。11月になり寒気をむかえると記事が増えているのも興味深い。また, 6月に顕著であるが曜日によって大きな差があるのは, 曜日と記事の編集方針に大きな相関があることを表していると考えられる。ちなみに6月の記事数の谷はいずれも週末である。

図3 キーワード「ベンチャー」を持つ記事の増減



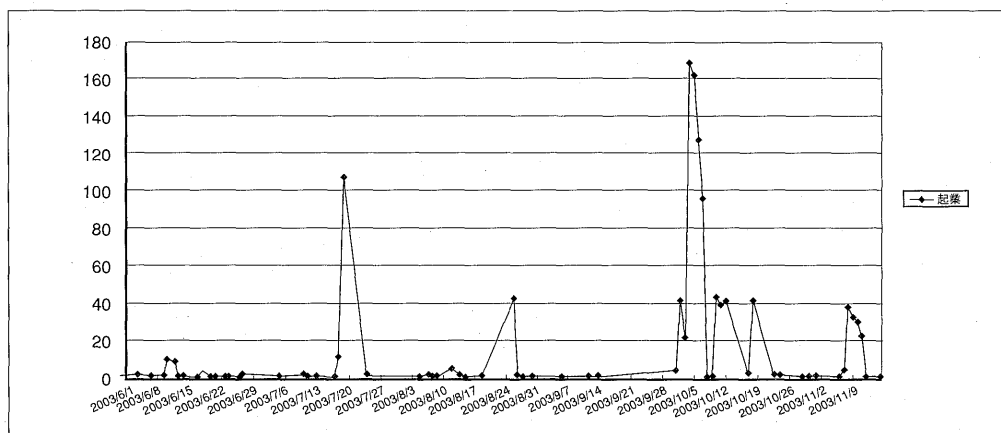
キーワード「ベンチャー」を持つ記事の増減は興味深い。半年の間に2つの山谷がある。6月は多く, 7月8月は少なく, 9月10月が多い。11月にはまた低下する。これをどのように解釈すべきかは今後の課題である。

図4 「リストラ」をキーワードにする記事の増減



「リストラ」は「ベンチャー」に比べ新しい言葉であるせいか出現頻度に偏りがある。6月中に何度か多く現れる日があり、7月と8月には少なく、9月末に著しく多い日が続く。企業の業績が良くないので「リストラ」をして対策します、業績が良くなったのは「リストラ」の効果です、といった文脈で語られることが多いと理解すれば、株主総会の時期や中間発表の時期に出現頻度が増加することもうなずける。

図5 「起業」をキーワードにする記事の増減



「起業」や「産学連携」は言葉としては定着途上といえる。何度か出現頻度の多い日があるが世論や流行と言えるものではなく、ある新聞が特集記事を組む、何かの行事が行われた、という散発的な事象に連動したものと考えられる。

図6 「産学連携」をキーワードにする記事の増減

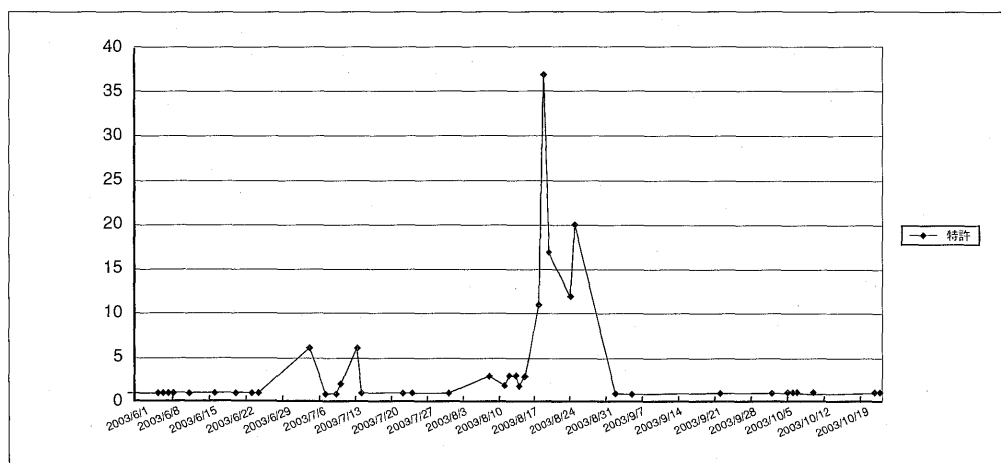
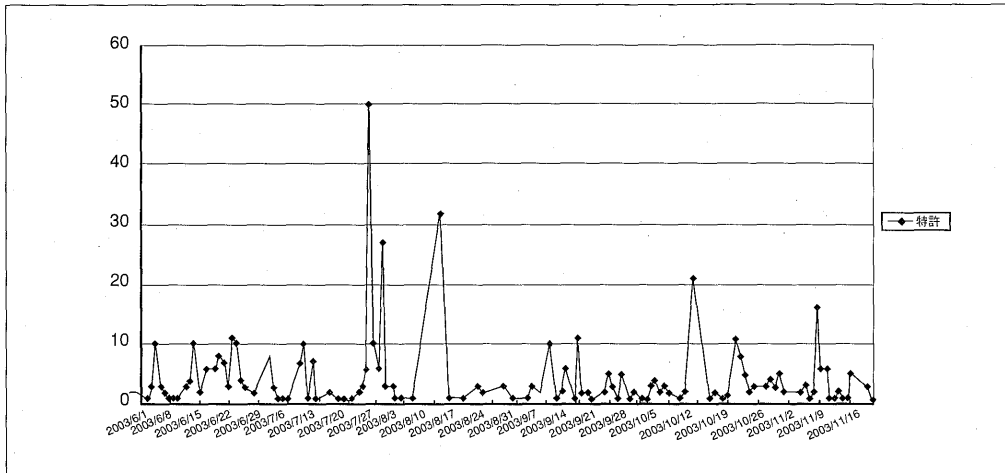


図7 「特許」をキーワードとする記事の増減



「特許」は一般用語と考えていたが、記事中への出現は少ない。知的財産との用語を用いることが一般化するに影響されている可能性も強い。この測定結果から何か結論づけることは困難である。

5. 考 察

開発したシステムは2003年5月以来稼働を続けている。しかし、収集した記事件数の集計方法に誤りがあり、すでに記事が削除されているにも拘わらず計数値にそれが反映されていないなどの不具合があった。図2から図7までは収集した計数値を手作業にて再加工して求めたものである。他の不具合は生じていず、この問題はすでに対策を施したので今後は手作業を介さずとも所望のグラフが得られる見込みである。

キーワードとして「特許、産学連携、SARS、リストラ、ベンチャー、起業」の6つを対象に実験を行ったが、すでに述べたようにいくつかの興味深いことが見いだされた。用語と対象サイトを適切に選定することと、数年にわたる継続的な変化を追跡することが重要と考えられる。

6. 結 論

人文科学系の研究に役立つ新たなツール開発を目指し、データマイニングシステムを制作した。WEBサイト(20箇所程度)を巡回して記事を検索し、この記事をもとにテキストマイニングを行うことにより、キーワード、論説等が時間の経緯とともにどう変化するのか、ビジネスとの関わりはどうかなどを調査検討するための研究プラットフォームである。

キーワードとして「特許，産学連携，SARS，リストラ，ベンチャー，起業」の6つを対象に5ヶ月間の運用実験を行い基本性能を確認した。今後はデータの積み上げと平行し，用語間の相関把握，キーワードの自動更新，テーマでの検索可能性，英文対応，等につき検討を進めていく。

ツールとしては合格点に達するものが完成したので，当初の目的であった技術とベンチャービジネスとの関連についての研究へも進めていく予定である。

謝 辞

この研究のアイディアは電気工学科の山口作太郎氏によってもたらされた。彼の卓抜な着想がなければこの研究はあり得なかった。システムそれ自体は（株）アドイン研究所の徳永雅彦氏の作品である。高機能なシステムが短期間で開発できたのは徳永氏をはじめとする同研究所の人工知能に関する高い技術力と経験の賜である。また，このシステムは中部大学アントレプレナーシップ研究センターのスポンサーシップにより開発された。ここに記して謝意を表する。

参考文献

- 1) 「ITプラットフォームP1基本設計書」，2003.3
- 2) 「ITプラットフォームP1詳細設計書」，2003.3
- 3) 「ITプラットフォームP1取扱説明書」，2003.3
- 4) CBI学会 2002年大会，2002.9『ポストゲノム研究に対応したCBIポータルの構築実験』
徳永雅彦（（株）アドイン研究所）
- 5) 情報処理学会 2001年全国大会『構造適応型RBMアルゴリズムの非正規化文書検索への応用』 徳永雅彦，早川光春（アドイン研），植村俊亮（奈良先端大）