解説論文

# 世界へのインタフェースとしての検索エンジン

松尾　豊[†*a)]

Yutaka MATSUO[†*a)]

**あらまし**　ここ数年の Google や Yahoo!など商用検索エンジンの発展は目覚しい。日本でも、次世代の Web の検索・解析技術を研究開発するプロジェクトが始まっているが、検索エンジンはなぜそれほど大事なのだろうか？すでに多くの書籍が検索エンジンの重要性については述べているが、一般的なユーザから見た重要性と、学術的にみた（すなわち長期的な研究開発のための）検索エンジンの重要性は全く異なる。端的に言うと、計算機から見たときに検索エンジンは実世界への窓、インタフェースである。膨大な社会現象、言語現象を、検索エンジンを通じて取得することができ、それは Web からの大規模知識の抽出や社会動向の測定・分析など、今後重要な技術へつながる。以前から行われてはいたが、検索エンジンを使ってその先の新しい技術を探る研究が最近ではますます活発になっている。本稿では、検索エンジンを取り巻く研究の流れと今後の見通しについて解説する。

**キーワード**　検索エンジン、世界知識、社会学、言語学

## 1. インタフェースとしての検索エンジン

Google や Yahoo!などの検索エンジンは、ここ数年、急速に進展し、今では人々の日常の情報収集の手段として定着した感がある。「Web 進化論」[1] や「グーグル—Google 既存のビジネスを破壊する」[2] など、検索エンジンに関する本はたくさん出ており、一般の人々の検索エンジンに対する理解は数年前と比べものにならない。

検索エンジンは、データベースや情報検索、言語処理、分散処理、ユーザインタフェースなど、さまざまな情報処理技術の上に成り立っているものであるが、一般的なユーザから見た重要性と、学術的にみた検索エンジンの重要性は全く異なる。一般的な人が気にするのは、検索エンジンの利便性やビジネスや法制度に与える影響であるが、学術的に見ても中長期的に検索エンジンの占める位置は大変重要である。

単純に考えれば、Web 上のコンテンツの処理は、大規模な自然言語文書（または多種なメディアのデータ）の処理と変わらない。ただ量が極端に多いだけである。実際、自然言語処理の研究コミュニティでは、新聞記事などの大規模コーパスの延長として、Web のアーカイブを扱う方向で研究が進展してきた。しかし、Web を（超）大規模な文書コーパスとみる見方は、Web の持つ重要な側面を見落としているように思う。それは、Web 上のデータが、我々の社会、日常生活に密接に直結したものであり、その分析結果は、我々の社会や日常生活を映し出したそのものであるということである。

新聞記事のコーパスとの比較を考えてみよう。これまで、新聞記事の検索、要約、分類などの研究が行われてきた。検索や要約の結果をうまくユーザに提供することで、有用なシステムとなる。多くの場合、新聞記事に特有の性質があり、それをうまく組み込んで活用することが精度を上げるために重要である。しかし、Web の場合はそうではない。検索、要約、分類した結果は、そのまま我々の社会活動の検索、要約、分類になっている。その精度を上げるための「特有の性質」は、コーパスの性質というよりは、我々自身の社会的性質である。

端的に言うと、検索エンジンは、計算機から見たときの実世界への窓、インタフェースである。膨大な社会現象、言語現象を、検索エンジンを通じて取得することができ、それは Web からの大規模知識の抽出や社会動向の測定・分析など、今後重要な技術へつながる必要不可欠なものである。実際、検索エンジン自体

† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
† スタンフォード大学
Stanford University
a) E-mail: y.matsuo@aist.go.jp

をひとつのモジュールとして使う研究が、国際的な学術コミュニティでは顕著に増えている。著者は、2002年から検索エンジンをモジュールとして捉える研究を行ってきたが、近年の WWW やデータマイニングの国際会議(注1)の発表を見ると、確実にその傾向が強まっていることを感じている。

もちろん、ここで述べたことはやや極端であり、Web上にない情報もたくさんあるし、偏りもある。しかし、Web は現時点で人間が入手できる最も大量かつ多様な人間の社会的活動、言語的活動のデータであることは間違いないだろう。そして、このデータの「量」は、情報処理のパラダイムを大きく変える。非常に単純なアルゴリズムが驚くほど有効に働く。データベースの問い合わせ言語を用いて、データを活用する情報処理システムを作るのと同じように、これからは、検索エンジンのクエリーを用いて、Web 全体の情報を活用する技術がますます重要になっていくだろう。

以下、本稿では、このような見方の背景となる研究の動向について概説する。

## 2. 言語現象を捉えるマイニング

Web には膨大な量のテキストデータがあるので、それを使った言語処理が可能である。特に、新聞記事や論文といったタイプの文書集合にはない、さまざまな特徴がある。例えば、口語の表現を含んでいる、ユーザが非常に多様である（研究者や記者だけが書いたものではない）、ユーザ間のインタラクションのデータ（掲示板や Blog でのやりとり）がある、更新が早くリアルタイムに更新される、などである。Web は、無数の目的・文脈における数多くの文があり、特に近年の Blog の普及によって、多くのユーザの日々の生活が日常的な言葉でつづられるようになった。

Web のデータを使うと、人々が日常的な感覚に近い処理が簡単に実現可能である。例えば、「犬も歩けば」何だろうか?すぐに思い浮かぶのは「棒にあたる」だろう。「犬も歩けば」で検索して次に来るものを列挙するという簡単なアイディアで見つけることができる。これを多言語で実現した研究が[3] であり、例えば、「東京」と言えば何かなどを、検索したあとにその語を含む前後のテキストを取り、統計処理を行う（頻度を数える）ことで抽出できる。こういった用例の検索は日常的に行っている人も多いのではないだろうか。この仕組みのシンプルさと結果の有用性は、まさに Web における「量」の力を示唆している。

Web から語の関連性（関連語、同義語、類義語）を捉えることもできる。その中で有名な研究は、Turney らによる検索エンジンを用いた同義語の把握の研究[4] である．TOEFL のシソーラスの同定問題（「次の中から，A と同義である語を選びなさい。選択肢：$C_1$, $C_2$, $C_3$, $C_4$」）を，検索エンジンを使って解くものであるが、アルゴリズムは驚くほどシンプルである。各選択肢のスコア $score(C_i)$ を次のように求める(注2)。

$$score(C_i) = hits(A \text{ AND } C_i)/hit(C_i)$$

ここで、$hits(A \text{ AND } C_i)$ は、語 $A$ と語 $C_i$ を AND でつないで検索エンジンに入れたときのヒット件数、$hit(C_i)$ は語 $C_i$ を入れたときのヒット件数である。AND 検索を行うだけでも正答率が 62.5%（より工夫をすれば 73.75%）の精度であると報告している。この研究で印象的なのは、この単純なアルゴリズムで、母国語が英語でない平均的な学生のスコアを上回る結果を示したことである。

また、同義語だけでなく、上位語や下位語を見つけることもできる。例えば、「うどん・そばなどの麺類」という句は、うどんやそばが麺類というクラスに属していることを意味している。例えば「うどん・そばなどの麺類の販売を行っております」などの文が多くあれば、うどんが麺類であり、そばと対比されることが多いなどのことが分かる。英語では Hearst のパターンが有名で、例えば、次のようなものである。

- $NP_0$ such as $NP_1, NP_2, \ldots, (\text{and}|\text{or}) NP_n$
- $NP_1, \ldots, NP_n$ (and|or) other $NP_0$
- $NP_1$ is a (kind of) $NP_0$

これを使うと、$NP_1, \ldots, NP_n$ が $NP_0$ のクラスに属する（$NP_0$ と is-a 関係がある）ことが分かる。すなわち、is-a 関係にあるかどうか調べたい２つの語 $X$ と $Y$ があれば “X such as Y” というクエリーを検索エンジンに入れ、それに該当する十分な数の文書があるかどうかを調べれば良い。

以上のような処理は、基本的に大規模なコーパスであればある程度可能であるが、Web というデータの量があるからこそ、簡単なアルゴリズムでも良い精度が

(注1)：International WWW conference, SIGKDD conference, International Semantic Web conference など

(注2)：確率的な根拠に基づいており、$A$ との相互情報量を最大にする $C_i$ を見つける操作に相当する。

得られる。例えば、Apple はりんごであるが、会社名でもある。自然言語処理でよく用いられるシソーラスの WordNet には、Apple が会社名であるという情報は載っていない。しかし、Web を調べれば、“Apple is a great company” “Apple is a consumer company” などの記述がすぐに見つかる。これを Web ではなく、論文集合から取り出そうとしても難しいであろう。

このように語の関連性や語に関する知識を得る研究はさまざまな形で行われている。例えば、Staab らは、PANKOW (Pattern based ANnotation through Knowledge On the Web) というシステムで、検索エンジンを使ってオントロジの抽出を試みている。Lapata らは、Web の検索ヒット数をさまざまな自然言語処理のタスクに適用し，タスクによっては既存のコーパスの結果を凌ぐことを報告している [5]．これまでの研究は、基本的に、Turney らのようなヒット件数を用いる処理、そして Hearst のパターンに代表されるパターンを用いる処理に二分されるが、この両者を融合させる試みもある [6]。

さて、言語現象という点からは、Web2.0 とよばれる多くのユーザが関与するサービスが興味深い。例えば、ソーシャルブックマーキングサービス（SB）は，Web ページや写真，動画、論文などさまざまなもの（インスタンス）を、ユーザがタグ（キーワード）をつけてブックマークできるサービスである．CiteUlike というサービスでは，自分が要チェックだと思った論文を，タグをつけて管理することができる．ユーザはそれぞれ、自分が好きなタグをつけてよい。そこに制約はない。しかし、他の人がよく使っているタグをつけると、関連するものを検索しやすくなる。そうすると、同じことを指すときには同じタグを使う方向に力が働き始める。こういった多くの人によって作られた語彙は、Folksonomy（民衆の語彙）と呼ばれる。「専門家」が領域知識を形式化して作るオントロジと対比的に語られることもある。（両者は目的が異なるので単純な対比は危険である。）SB ではまさに，個人にとっての意味がコミュニティで共有されるに至る過程を見ることができる [7]。言語学者のソシュールはその著書「一般言語学講義」の中でラングとパロールという２つの概念を対立させたが、こういった現象が実際に観測できるようになったわけである。そこでもやはり、検索という機能がこのプロセスに重要な役割を担っている。

## 3. 社会現象を捉えるマイニング

一般的な語に関する用例や上位・下位関係を得るのであれば、それは言語知識の獲得であるが、もっと具体的な会社名、製品名、人名など（named entity と呼ばれる）に適用すると、それは、言語知識の獲得というよりは、社会的な知識の獲得という側面が強くなる。

最もシンプルな例では、Google のヒット件数と研究者の名声（fame）を調べた研究がある [8]。Web でのヒット件数というのは、ある意味でそのエンティティの社会的な有名さを表している。

検索エンジンのヒット件数を、エンティティのペアに適用すると、エンティティ間の関係性を把握することができる。（基本的なアイディアは Turney らの同義語の把握と同じで、AND 検索をする。）MIT の A. McCallum らを中心としたグループでは，e-mail のメッセージの中から名前を見つけ，対応するホームページを見つけ，コンタクトアドレスを埋めるシステムを作っている [9]．この中で，Web 上で共起する名前はその人と関係があるとしてネットワークを抽出する。P. Mika らは，Web 上の名前の共起関係や FOAF ファイルから社会ネットワークを抽出し図示する Flink（www.semanticweb.org）というシステムをつくっている．著者らの研究グループでは、人工知能学会を中心とする研究者のネットワークを Web の共起関係で取り出しており、さらにテキスト処理を組み合せることで、その関係の種類やキーワード等を把握している [10]。基本的には、例えば「石原慎太郎 田中康夫」などのクエリーで検索して、そのヒット件数が統計的にどのくらい有意に多いか、またどういった文脈で共起しているかを調べる。時系列にその関係がどう変わっていくかを分析することも可能である [11]。これは、そのエンティティ同士の社会的な関係性が変化するために、それを反映した Web の情報も変化し、それを捉えることができるわけである。

さて、実世界と Web の世界の対応を考えたときに、実世界のエンティティが Web ではどう表現されるのかという対応は興味深い。人名の場合には、同姓同名の解消（name disambiguation）というタスクになり、ここ２、３年、研究が増えているトピックである。例えば、「松尾」で検索エンジンを引くと松尾電機株式会社、松尾スズキ（俳優、脚本家）、シェ松尾（レストラン）が出てくる。「松尾豊」で引くと私がトップに出てくる。（他に化学研究者の「松尾豊」や肉屋の社長の松

尾豊も出てくる。）つまり、乱暴に言えば、Web 上で松尾というと著者のことではないが、松尾豊というと著者のことである。Web でそのエンティティ（その人自身）をどう同定しなければならないかというのは、実は、個体を認識するにはどうすれば良いか、そもそも同一であるとはどういうことかという問題を含む。

我々のイメージは、実世界の対象それぞれに Web での検索クエリーがついているというものである。最近、「〜で検索してください」というだけで詳細を書かない広告や名刺などを目にすることがある。これは、まさにこのイメージの通りである。すなわち、我々は日常的に、誤る可能性のない程度に曖昧にエンティティを指定しており、誤る可能性のある場合にはもっと詳細に指定する。（田中さん、佐藤さんは、下の名前もつけて呼ばれることが多いのではないだろうか。）実は、エンティティが同一かどうかは、基本的には多くの手がかりから推測するしかないものであって、そのときにエンティティの同一性とその同一性の表現（クエリー）はセットになっている。同一であるとは何かという問題は、これまでも議論されてきたが[12]、検索エンジンはこの問題を具体例として取り扱う手段を提供している。

Web からの社会現象の分析という点から言えば、Blog の分析を抜きに語ることはできない。Web 上のデータをクチコミの分析、マーケティングに用いようという動きも以前から続いている。最近では、Blog の情報を収集分析する実際のサービスがいくつか立ち上がっている。例えば、「シリコンバレー」という語が最近、よく出現するようになったとしよう。これは、シリコンバレーという語の言語的な性質が変わったのではなくて、社会的な使用が変わっている（盛り上がっている）わけである。こういったことを利用して、Blog からこういった語のバーストやトレンドを抽出する研究 [13], [14] などが行われている。さらに、ある製品の評判情報（ポジティブかネガティブか）やどういった製品と比較されているかなどを Blog から取り出す研究もある。Blog や Web 上での言及と Amazon の売り上げの直接の関係を調べた研究もある [15]。この研究では、Blog や Web のコメントの数を使って、本の売り上げを予測できるかを調べており、その結果、本の売り上げの数自体を予測することは難しいが、売り上げの急上昇（スパイク）を数日から数週間前に予測することができると報告されている。こういった研究は、まさに Web を社会を映す鏡として、社会動向の調査に用いているわけである。そして、これらの研究では、Technorati や Blog Wathcer などの検索システムを利用して分析を行っており、Blog に限ってもなお、エントリの収集と索引付けという検索システムは基盤となる。

## 4. マイニングによる知識化

Web から社会現象、言語現象に関する知識をマイニングしたとして、それをどう処理するかはさまざまな方法がある。基本的には、人工知能の分野で長く研究されてきた知識表現や推論の枠組みが役立つ。Semantic Web はこういった知識表現、そして推論を Web 上で実現しようとするものである。RSS や FOAF など、一部の技術は十分に広まって活用されているが、Semantic Web が当初から想定していたような情報の統合や推論はまだ十分に行われていない。しかし、検索エンジンを突破口に、知識を抽出し、統合する研究が徐々に広がりつつある。

Google の研究者らが昨年発表した研究は、エンティティ間の関係を事実に関する知識として取り出すものである [16]。基本的には、3 章で述べたパターンに基づくエンティティ間の関係の認識が基盤となる。しかし、エンティティの関係や、それを取り出せるパターンは無数にあるので、それを自動的に学習する技術が重要になる。この研究では、例えば、10 個の人名と生まれた年のペアが与えられると、その組み合わせが現れるパターンと具体例をブートストラップ的に学習していく。例えば、George W. Bush と 1946 年の関係は次のような文で現れる。

- **George Walker Bush** (born 6 July **1946**) is the 43rd and current President of ...
- President **Bush** was born on July 6, **1946**, in New Haven, Connecticut, and grew up ...
- Born July 6, **1946**, in New Haven, Connecticut. **Bush** — often referred to as simply "W" — is the eldest son of former President ...

これらの例の共通性を抽象化し、パターンとする。それを使って新たな例を見つけ、さらにパターンを得る。ここで重要なのは、パターンをいかに抽象化するかである。うまく抽象化することで、多様なエンティティ間の関係を学習できる。こういった技術は実は Google の創業者である Brin が 1998 年の論文でも扱っているが [17]、この研究では、Web 上の 1 億の文書を対象に 100 万の事実を取り出すという数値目標を掲げ、それ

が次世代の検索エンジンの核となると述べている。

一般常識の知識ベースを作るために 1980 年代から延々と続けられている Cyc プロジェクトは有名であるが、Cyc の研究者も知識ベースを Web 上の情報で増強しようとしている。増強したい知識（例えば、パレスチナ・イスラム・ジハードを作った人は誰か。(founderAgent PalestinaIslamicJihad ?WHO) と表される）から Google へのクエリーを生成する。例えば、”Palestine Islamic Jihad founder *”などである。結果のなかから Cyc の型制約に合うものを見つけ、知識ベースに加える。

さらに進んで、知識の収集だけでなく知識の統合・推論まで踏み込んだ研究も行われ始めた。Sheth らはバイオインフォマティクスのコーパスを対象として、エンティティとその関係を RDF（Resource Description Framework）を用いて記述し、その上で、自明でない関係性を導き出すこと研究を行っている [18]。例えば、「片頭痛」と「マグネシウム」の関係を調べたいとき、RDF を検索することで「片頭痛」が「血小板の振る舞いの異常」で起こり,「コラーゲン」が「血小板」を刺激し,「マグネシウム」が「コラーゲンが引き起こした血小板凝集」を抑制することが分かる.（「」がここでのエンティティである。）これらはそれぞれ別々の論文の別々の文から得られたものである．これは、論文を対象とした知識の集約の例であるが、このグループでは Web を対象にした研究も行っており、研究者の協働関係（共著等）のデータをマイニングし、それによって利益相反がないかどうかを調べるシステムを構築している [19]。

このように、Web の情報を、検索エンジンを使って収集し集約する、知識化する、その知識を利用するという一連の流れが、次世代の Web 上の情報処理の基盤となるのではないだろうか。そして、検索エンジンは、その処理において決定的に重要な働きを果たすのではないかというのが筆者の考えである。

## 5. ま と め

本稿では、検索エンジンについて、その研究における重要性を述べた。以上で紹介した方法は驚くほど簡単なアルゴリズムである。もちろん、シンプルなアルゴリズムをベースにさまざまな工夫を凝らすことで精度を上げることができ、実際にはそのような工夫が行われている。しかし、ここで最も述べたいことは、アルゴリズムが優れていることではなく、Web のもつ情報の量、そしてそれが我々の実社会を反映していることが、圧倒的に大きな力を生み出しているということである。

例えば、文書分類を行うことを考えると、適度な大きさのコーパスではサポートベクターマシン（SVM）が良い性能を発揮する。しかし、1 億、10 億といった規模の文書になると、SVM は適用可能ではない。一般的に機械学習の精度は、学習データの量とともに向上するので、大量のデータを扱えることはアルゴリズムの多少の優劣を跳ね返してしまう。結果的に、SVM よりもっと単純な、例えば Naive Bays の分類器の方が実は大規模データには適している [20]。このように、Web を対象とするときに最も重要な点は、アルゴリズムがスケールするかどうかであり、一般的な（小規模な）研究グループで学術的な研究を行うには、検索エンジンを利用することは最も簡便で有効な手段である。

このような意味での検索エンジンの学術的な重要性を鑑みると、現状の検索エンジンの学術的な研究環境は良いとは言いがたい。例えば、Google では API を配布し、検索を 1 日 1000 件まで可能にしている。Yahoo!では、API で 1 日 5000 件までの検索が可能である。しかし、このような制限は、学術的な研究には大きな制約になる。また、一般ユーザに利用できる検索機能だけが、検索エンジンに提供できる検索機能の全てではない。（処理コストが高いために開放していない機能もあるだろう。）したがって、研究目的に利用のしやすい検索エンジンの環境を構築していくことは、学術コミュニティ全体にとって重要である。検索エンジンのサービスを提供する企業と連携して、こういった環境を整備していくことが必要であろう。また、国内外でも検索エンジンに関連したさまざまな活動が試みられている(注3)。こういった試みが、今後の研究を加速し、日本の情報処理に関する学術コミュニティの競争力となっていくことを期待して、本解説の結びとしたい。

## 文　献

[1] 梅田：“ウェブ進化論”, 筑摩書房 (2006).
[2] 佐々木：“グーグル –Google 既存のビジネスを破壊する”, 文藝春秋 (2006).
[3] K. Tanaka-Ishii and H. Nakagawa: “A multilingual usage consultation tool based on internet searching –more than a search engine, less than qa–” (2005).
[4] P. Turney: “Mining the web for synonyms: PMI-IR

(注3)：国内では、情報大航海プロジェクトや東京大学で開発されている検索エンジン基盤 Tsubaki など。

versus LSA on TOEFL", Proc. ECML-2001, pp. 491–502 (2001).
[5] M. Lapata and F. Keller: "The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of nlp tasks", Proc. HLT-NAACL 2004, pp. 121–128 (2004).
[6] D. Bollegala, Y. Matsuo and M. Isizuka: "Measuring semantic similarity between words using web search engines", Proc. WWW2007 (2007).
[7] S. Golder and B. A. Huberman: "The structure of collaborative tagging systems", Journal of Information Science (2006).
[8] J. Bagrow and D. ben Avraham: "On the google-fame of scientists and other populations" (2005).
[9] A. Culotta, R. Bekkerman and A. McCallum: "Extracting social networks and contact information from email and the web", CEAS-1 (2004).
[10] Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida and M. Ishizuka: "POLYPHONET: An advanced social network extraction system", Proc. WWW 2006 (2006).
[11] 安田, 松尾, 武田："人工知能学会におけるネットワーク構造と変化", 人工知能学会全国大会 (2006).
[12] ロバート："考えることを考える", 青土社 (1997).
[13] N. S. Glance, M. Hurst and T. Tomokiyo: "Blogpulse: Automated trend discovery for weblogs", Proc. WWW2004 (2004).
[14] 藤木, 奥村："周期的に発生する burst の予測と抑制", 人工知能学会第 73 回知識ベースシステム研究会 (2006).
[15] D. Gruhl, R. Guha, R. Kumar, J. Novak and A. Tomkins: "The predictive power of online chatter", Proc. KDD 2006 (2006).
[16] M. Pasca, D. Lin, J. Bigham, A. Lifchits and A. Jain: "Organizing and searching theworldwideweb of facts - step one: the one-million fact extraction challenge", Proc. AAAI2006 (2006).
[17] S. Brin: "Extracting patterns and relations from the world wide web", the International Workshop on the Web and Databases (1998).
[18] C. Ramakrishnan, K. Kochut and A. Sheth: "A framework for schema-driven relationship discovery from unstructured text", Proc. ISWC2006 (2006).
[19] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, A. Sheth, I. Arpinar, L. Ding, P. Kolari, A. Joshi and T. Finin: "Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection", Proc. WWW2006 (2006).
[20] C. Manning, P. Raghavan and H. Schutze: "Introduction to Information Retrieval", Cambridge University Press (2007). online version.

(平成 xx 年 xx 月 xx 日受付)

**Abstract**

**Key words**