〔特集論文〕 「Web インテリジェンスとインタラクションの新展開」

動向に関する問いを対象とした検索エンジンの提案

Proposal of Search Engine Focusing on Trend-related Queries

高間 康史
Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University
ytakama@tmu.ac.jp, <http://krectmt3.sd.tmu.ac.jp/~takama/index.html>

加藤 優
Yu Kato

(同 上)

桑折 章吾
Shogo Kori

(同 上)

石川 博
Hiroshi Ishikawa

(同 上)

keywords: web search engine, trend information, information retrieval

Summary

This paper proposes a search engine that is designed for answering trend-related queries. Recently, people can gather huge amount of information by using existing Web search engines. However, there is a significant difference between function provided by existing search engines and users' information needs. That is, a function of existing Web search engines is just to find Web pages containing keywords specified as a query. When we want to find some information from the Web, we have to break our information need into a series of queries by ourselves. As this process is burdensome for novices and even for experienced users, advanced search engine that can solve this problem should be realized. This paper focuses on the task of answering trend-related queries. By focusing on a specific task, proposed search engine can provide advanced search functions compared with existing Web search engines. At the same time, as this task is supposed to be useful in various domains, the proposed search engine can inherit one of important characteristics of existing search engines that it can be widely used regardless of domains. Prototype search engine collects various trend-information from the Web and provides three basic functions: a function for searching time periods when a query item had a characteristic trend, that for searching items having characteristic trend during query time period, and that for searching items having similar trend with query item. These functions are derived from analysis results of search intention using existing search engines, and users can satisfy their various kinds of information needs related with trend information by combining those functions. The effectiveness of the prototype search engine is shown through experiments with test participants. Experimental results show that test participants can issue correct queries efficiently for given information needs. It is also shown that they can break their own information needs into a series of queries corresponding to various search intentions.

1. は じ め に

本論文では、次世代検索エンジンの一つとして、動向に関する情報要求に応える検索エンジンを提案する。現在、Web 上の膨大な情報にアクセスするための最も代表的な手段は検索エンジンである。タスク・ドメインを問わずに利用可能という利点があるが、効率性の面では限界を迎えつつあると考える。

既存検索エンジンの問題点として、提供する基本検索機能と、ユーザの情報要求との乖離が大きいことが挙げられる。既存検索エンジンが提供する基本検索機能は、ユーザが指定したキーワードを含む Web ページを返すといった、汎用的に利用可能なものであり、利用ユーザを限

定しないため広く一般に使われるようになったと考える。しかしその反面、検索機能としてはプリミティブなものであり、個々の情報要求に合わせ、適切なクエリ（キーワード）に分解するのに要するユーザの負担が大きい。既存検索エンジンのクエリログを分析した結果 [Kim13, Silverstein99, Spink01] では、一連の検索行為に対応するセッションあたり平均 2-3 クエリ、クエリあたり平均 2.5 単語程度の、比較的単純な検索が多いことが示されている。また、検索エンジン Excite のクエリログ約 100 万件を分析した結果 [Spink01] では、約 52% のユーザがクエリの修正を行っていることが報告されており、1 回の検索で満足いく情報が得られない場合が多いことも示唆されている。この様な現状を踏まえ、専門的なスキルを持

たない一般ユーザが精度の高い検索を行い、質の高い情報を獲得するためにはユーザ自身のスキル向上も必要との指摘がなされている [齋藤 04]。

一方、指定したキーワードを含む Web ページを検索するといっても、ある特定の情報を発見することを目的としたものや、あるトピックに関する多様な情報の収集を目的としたものなど、多様なタスクタイプに分類可能であること、必要な情報を得るために一定期間継続して行わなければならない場合もあることが Bing のクエリログ分析で示されている [Agichtein12]。また、情報検索の際に用いられる検索戦略を 12 種類に分類し、Web 検索技能によって利用する戦略に違いがある事も示されている [Thatcher08]。この様に、既存検索エンジンを使いこなしているスキルのあるユーザも存在するが、多様な検索目的を全て「キーワードを含むページの検索」に分解する現状のアプローチには効率性の面で限界があると考ええる。検索スキルによらず存在する、情報要求と基本検索機能との乖離を軽減できる、次世代検索エンジンの開発が必要と考える。

次世代検索エンジンの一つとして、ユーザの情報要求を自然言語で受け付け、答えを返すアプローチが考えられる。将来的には有望なアプローチの一つと考えるが、現状ではユーザが自然言語で自由に表現する情報要求に全て答えることは不可能といえる。それよりも、既存検索エンジンのように、ユーザ自身が情報要求をクエリに分解するアプローチを採用しつつ、検索エンジンの提供する機能（以降、基本検索機能と呼ぶ）とユーザの情報要求の乖離をできるだけ小さくすることが有望なアプローチと考える。

より高度な基本検索機能を提供するアプローチとして、対象ドメインを限定するアプローチも考えられるが [亀井 07, 小久保 02, 宮川 06, 森本 10, 西藤 03, 鈴木 05]、幅広いドメインで利用可能という既存検索エンジンとは異なる位置づけになる。

本論文ではドメインではなくタスクを限定することで、既存検索エンジンよりも高度な基本検索機能を提供し、それを組み合わせることで、ユーザの多様な情報要求に応えるアプローチを採用する。対象タスクとして、本論文では「動向に関する問い」[加藤 13]を採用する。動向に関する問いは、過去に流行したモノを調べるといった気軽な情報要求から、震災直後に需要が増すアイテムや、オリンピック開催が動向に与える影響の調査など、多様なドメインで一般に見られるものである。また、Web 上では検索ヒット数やツイート数など、動向に関する情報が日々生み出されている。さらに LOD (Linked Open Data) [大向 13] に代表されるように、各種統計データなどが公開されるようになってきており、それらを活用したデータジャーナリズム [立園 13] など注目されるようになってきている。しかし、これらの動向に関する問いに関して既存検索エンジンを用いて情報を得ようとする場

合、他者がまとめた報告などが存在することを期待して Web ページの検索を行うか、関連する動向情報を公開している Web サイトを探して入手し、自身で分析を行う必要がある。例えば、あるアイテムと同時期に流行したアイテムを見つけたい場合 [加藤 12]、当該アイテムが流行した時期をまず同定し、その後その時期に流行したアイテムを探す必要があるが、他者が発見し、報告した情報が見つからなければその探索は困難といえる。提案する検索エンジンは、このような動向に関連する情報を探すタスクに特化することで効率的な検索の実現を目指すものである。また、既存検索エンジンと同様に対象ドメインを限定しないアプローチは、次世代検索エンジンの構築にあたり重要な視点であると考ええる。

本論文で提案する検索エンジンでは、既存検索エンジンにおけるユーザの検索意図を分析した結果などにに基づき、指定アイテムに関する動向が特徴的変動を示した期間の検索、指定期間に特徴的変動を示したアイテム・動向の検索、指定アイテムに関する動向が特徴的変動を示した期間に同様の変動を示したアイテム・動向の検索、の 3 つを基本検索機能として提案する。Web 上から収集した各種動向情報から最大値・急上昇・山形などの特徴的変動を抽出し、これらに対して基本検索機能を利用した検索を可能とする。開発したプロトタイプシステムでは、検索インタフェースだけでなく API としての機能提供も実装している。

プロトタイプシステムを用いて実験協力者による評価実験を行った結果より、実験協力者は情報要求に対して適切なクエリを表現できること、情報要求を多様な検索意図に基づく複数のクエリに分解し、検索可能であることを示す。

2. 関連研究

2.1 検索エンジンの知的化・高機能化

既存検索エンジンの知的化・高機能化を目指すアプローチとしては、自然言語によるクエリ入力を受け付ける方法 [Ferreira05, 徳永 07] や、検索対象とするドメインを限定し、専門検索エンジンを構築する方法などが提案されている [亀井 07, 小久保 02, 宮川 06, 森本 10, 西藤 03, 鈴木 05]。

前者のアプローチは、Apple の Siri などによる音声検索の普及により、将来的には有望なアプローチと考えられる。自然言語を利用するアプローチの場合、一般ユーザは人間同士のコミュニケーションの様に、情報要求を明確化したり、単純なクエリに分解しなくても、計算機側で対処してくれることを期待すると考えられる。しかし、Watson [Ferrucci10] などの事例を見ても、ある程度複雑な問いに対応したい場合にはドメインを限定し、十分な知識を事前に用意しておく必要があり、既存検索エンジンのように多様なドメインに汎用的に対応すること

は現状困難と言える [徳永 07]。現状の音声検索を検索エンジンに利用した場合には、単純なクエリにのみ対応可能であり、ユーザの期待レベルとのギャップが大きいと考える。

亀井らは、Web 上に存在するソフトウェア開発に関する知見や情報を対象とした検索エンジンを提案している [亀井 07]。Web 上には、過去のソフトウェア開発の際に得られた知見や、開発に役立つ情報が多く共有されている。しかし、それらは Web 上に点在するため、網羅的、効率的に情報収集することが困難であり、似たようなソフトウェアが様々な場所で開発されていたり、同じようなミスにより開発が滞る場合がある。そこで亀井らは、Web 上に存在するソースコードや付属するドキュメント、Tips などのソフトウェア資源をクロウリングにより収集し、索引付けを行っている。ソースコードに対しては、そのクラス名やメソッド名などを解析して索引付けすることで、名前、引数や返値の型、行数などを指定した検索を可能としている。

小久保らは、既存検索エンジンにおける情報収集では、必要な情報を得るためにある程度の知識や経験が必要となり、多くの初心者ユーザにとって使いこなすのが容易ではないとの考えに基づき「検索隠し味」と呼ぶドメインを限定した専門検索エンジンの構築手法を提案している [小久保 02]。この手法では、決定木学習を用いて Web ページ集合から抽出したブール式を、ユーザが入力したクエリに加える事で、既存検索エンジンの検索結果を特定ドメインに特化させている。

これらの検索エンジンは、検索対象ドメインをある領域に特化させることで、既存検索エンジンよりも効率的な検索の実現を目指している。これに対し、本論文で提案する検索システムでは、「動向に関する問い」という、ドメインに依存しないタスクを対象とする検索エンジンの実現を目指す点で異なっている。ドメインに依存しないタスクを対象とすることで、広く一般に利用可能という既存検索エンジンの特徴を継承するとともに、対象タスクに特化した高機能な基本検索機能の実現を図る。利用者は対象ドメインにおける知識や経験に基づき、動向に関する情報要求を基本検索機能に分解する必要があるが、既存検索エンジンよりも高度な基本検索機能を提供することで、情報要求との乖離を低減させる。

2.2 動向情報の活用

動向情報とは、ある商品の価格や売上の状況、ある会社の業績状況などの時系列データを基として、その変化を通時的にとらえつつ、総合的にまとめ上げることで得られるものであり [加藤 04]、様々なタスク・ドメインにおいて意思決定の材料として用いられている [石黒 12]。近年、LOD などに代表されるように、官公庁をはじめとする様々な企業、組織、団体が Web 上で情報公開を積極的に行う様になりつつあり、その中には製品の価格や

生産量データ、日本の人口、内閣支持率など多種多様な動向情報が含まれている。また、キーワードの検索数やブログ記事数、ツイート数といった、Web 上でのユーザ活動により生まれるデータも、動向情報の一種とみなせる。この様に、多様な動向情報が Web 上で利用可能となった結果、動向情報の利用に関する研究も多く進められている。

様々な動向情報の要約・可視化に関する技術として、新聞記事やブログなど複数文書に含まれる動向情報を抽出し、可視化する手法 [松下 06, 奥田 07, 山本 07] や、Web 上から取得した動向情報を利用した検索システム [石黒 12, 加藤 12] が提案されている。松下らは、動向情報が含まれるテキストを視覚情報として要約することを目的として、テキストに含まれる情報を用いてグラフを描画する方法を提案している [松下 06]。新聞記事などテキスト中の明示的かつ定量的な数値情報に加えて、テキスト中で暗示されている情報を比較表現や背景知識によって抽出することで、より多くのプロットを可能としている。また、テキストに出現する「安定」や「緩やかな増加」などの定性表現を用いてグラフ概形を示唆するアノテーションをグラフに貼り付けることで、動向情報の理解を支援している。山本らは、関連しあう動向情報を効率的に提示することで分析作業を支援するシステムを提案している [山本 07]。ユーザが指定した動向情報と多様な動向情報間の関連度を計算することで、関連する単語とその動向情報を獲得している。

石黒らは、異種情報間の時間的関連性についての検索をコンテキスト検索と定義し、コンテキスト検索に基づく対話的な時系列データ分析を支援するシステムを提案している [石黒 12]。為替レートデータとニュース記事の見出しを対象データとして類似変動区間検索機能、類似イベント検索機能を基本検索機能として提供している。加藤らは、検索数やヒット数など、Web 上の動向に関連する基本情報を Web コンテキスト情報として定義し、これらに基づく同時期流行アイテムの検索手法を提案している [加藤 12]。入力されたアイテムの流行期間を Google Insights for Search^{*1} を用いて推定し、期間指定オプションを利用した検索エンジンでの検索結果も利用して入力アイテムと同時期に流行したアイテムを出力する。

ブログや Twitter、口コミサイトなどを分析し、Web 上や現実社会で流行しているアイテムや話題を抽出する研究・Web サービスも数多く存在する。Kleinberg は、キーワードの出現頻度の急激な増加により流行の話題を抽出しており、この状態を“burst(バースト)”と定義している [Kleinberg02]。マイクロブログ上での流行把握に関する研究や Web サービスでは、各単語のバースト状態を検出することで、その時期に流行している話題を発見してい

*1 2012.9.28 より Google Trends と統合：
<http://www.google.com/trends/>

る [奥村 04, 高橋 11] . kizasi.jp^{*2}では, 毎日約 25 万件のブログエントリを収集したうえで, エントリ内の全キーワードについて, 約 10 分ごとに時系列で解析している. また, 解析結果を利用し, ブログ上におけるホットな話題のランキングをきざしランキングとして公開している.

あるキーワードのバースト状態を検出した際には, その話題は社会的に既知となることが多い. これに対し高橋らは, 注目話題の早期発見を目的として, 後にバースト状態となる可能性の高い話題 (キーワード) を抽出する方法を提案している [高橋 11] .

動向情報に関わりの深いドメインとして, 金融市場が挙げられる. 近年, 機械学習を用いたテキストマイニング手法によって, 大量のテキスト情報と市場変動の関係性を発見し, 市場分析に応用する研究が増えている. 和泉らは, 金融経済月報を用いて, 日経平均株価の変動に寄与した単語の抽出を行い, 比較的長期の変動を予測する手法を提案している [和泉 09] .

分析対象は専門性を持ったテキストに限らず, ブログや Twitter など分析対象となっており, 経済に対する世の中の平均的な見方のトレンドを知ろうとする研究が行われている [Bollen10] . 灘本らは, 株銘柄に関連するキーワードをブログから収集後, 実際の株価変動と相関が高いキーワード群を抽出し, それらのキーワードの出現頻度により株価変動を予測する手法を提案している [灘本 07] .

また, 近年注目されているデータジャーナリズムという報道手法も, 動向情報と関連が深い. データジャーナリズムは「データ駆動型ジャーナリズム (データドリブンジャーナリズム)」や「データで実現するジャーナリズム」などと説明される [立園 13] が, 政府や自治体, 様々な団体などによりオープンに公開されている複数のデータをもとに分析を行うことで, 様々な領域において新しい解釈を導き出すことを目的とした報道手法である. データジャーナリズムのプロセスとしては, (1) データの発見, (2) 意味分析, (3) 可視化, (4) 混合の順に進めていく [佐々木] . このうち, データをわかりやすく可視化すること, 複数種類のデータを混合させて用いることが, 従来のジャーナリズムと特に異なる特徴であると考えられる. これまでに, 欧米で政治や経済格差, 環境問題, 人権差別など様々な分野においてデータジャーナリズムを適用した事例が多数存在する [NYT09, NY] . 日本国内においてもその取り組みは始まっており, 2013 年 12 月にはデータジャーナリズム・キャンプ&アワード 2013^{*3} が開催されている. データジャーナリズムには, 複数種類の統計データ, 様々な動向データの発見が必要不可欠である. 従って, 多種多様な動向情報を横断的に検索可能とする, 本論文で提案する検索エンジンは, これらの分野にとっても有用なサービスであると考えられる.

図1 提案システムの構成

3. 動向に関する問いを対象とした検索エンジン

3.1 システム概要

本論文では「動向に関する問い」を対象とした検索エンジンを提案する. 提案システムの構成を図1に示す. 実装には Ruby on Rails3.2, Apache2.2, MySQL5.0 を用いている. クローラー (Crawler) は, Web 上で公開されている動向情報を収集し, 検索対象とする特徴的な動向変動を計算し, データベース (DB) に格納する. なお, 既存検索エンジンの Crawler と異なり, 一度収集した情報が変更になることは通常ないため, 同一情報を再収集する必要はなく, データベースは原則として単調増加していく. Web サーバ (Web Server) はクライアント (Client) からのクエリを受け付けて DB を検索し, 検索結果を Client へ返す. Client としては通常の Web ブラウザからのアクセスを想定する他, 任意アプリケーションからの利用も可能となるように API も提供する. 以降の節において, 各部について詳説する.

3.2 対象データ

提案システムで収集対象とした動向情報は, 大きく 2 種類に分類できる. 一つは, 各アイテムの価格や販売量に関する統計データの様な, 各企業や組織・団体によりコンテンツとして公開される動向情報である. 他方は, 各アイテムをキーワードとして既存検索エンジンで検索した際のヒット数や, ブログ記事数などといった, Web 上でのユーザ活動により発生する動向情報である. 後者は Web Usage Mining に関連するデータといえる [Mobasher06, Pierrakos03, Srivastava00] .

コンテンツとして公開されている動向情報として, これまでに収集したものを表1に示す. 各サイトごとに専用のプログラムを実装し, 収集を行ったが, 近年の LOD の普及の流れを踏まえ, 今後は SPARQL エンドポイントなどからの統一的なアクセス手段による収集も検討している.

Web 上でのユーザ活動により発生する動向情報に関しては, 表2に示す情報を収集している. 急上昇ランキン

*2 <http://kizasi.jp>

*3 <http://jcej.info/djc2013/>

グやきざしランキングなど、ランキング上位の単語が定期的に公開されるサービスでは、それらの単語を収集対象としている。急上昇ランキングのようにポイントなどの数値が与えられているものについてはその数値を、きざしランキングのように順位しか公開されていない場合は順位を動向情報として扱う。

ヒット数、ブログ記事数、検索数等の場合は、対象単語のリストを事前に用意し、それらについて情報を収集した。単語リストは、Yahoo! Japan が提供する年間検索ワードランキングで上位に含まれる人名や企業名、サービス・製品名などを中心に、約 400 語を選択したものを利用した。

表 1 コンテンツとしての動向情報の収集対象

団体名	動向情報	収集期間	URL
総務省統計局	人口、雇用者など	1953.1–2013.8	http://www.stat.go.jp/
日本即席食品工業協会	生産数量(袋めんなど)	1958.1–2011.12	http://www.instantramen.or.jp/
日本ミネラルウォーター協会	ミネラルウォーター消費量	1986.1–2011.12	http://minekyo.net/
NHK 放送文化研究所	内閣政党支持率	1998.4–2013.7	http://www.nhk.or.jp/bunken/
日本エネルギー経済研究所	ガソリン卸価格	2000.7–2011.3	http://oil-info.ieej.or.jp/
内閣府	景気ウォッチャー調査(百貨店、コンビニなど)	2000.10–2013.4	http://www.cao.go.jp/
農畜産業振興機構	価格(にんじん、キャベツなど) 消費者物価指数(にんじん)	2005.1–2012.8 2008.1–2010.11	http://vegetan.alic.go.jp/
自転車産業振興協会	自転車販売台数	2009.1–2011.12	http://www.jbpi.or.jp/

表 2 ユーザ活動による動向情報の収集対象

団体名	動向情報	収集期間	URL
Yahoo! JAPAN	急上昇ワードランキング(ポイント)	2007.11–2012.8	http://searchranking.yahoo.co.jp/
	ヒット数	2003.1–2012.12	http://search.yahoo.co.jp/
	ブログ記事数	2005.1–2012.12	http://search.blogs.yahoo.co.jp/
Kizasi.jp	きざしランキング(順位)	2006.4–2012.7	http://kizasi.jp/
ついつぶるトレンド	HOT ワード(盛り上がり度、ツイート数)	2010.3–2012.7	http://tr.twipple.jp/
Google	Google Trends (検索数)	2004.1–2012.12	http://www.google.com/trends/

提案する検索エンジンでは、収集した動向情報について、特徴的な変動を示した期間に基づく検索を可能とす

る。提供する基本検索機能については次節に述べるが、ここでは検索対象とする特徴的変動について説明する。今後種類を増やすことを検討しているが、本論文では文献[蓮井 13]を参考に、以下の 3 組 6 種類の特徴的変動を定義し、検索対象とする。なお、収集した動向情報の粒度は月であるものがほとんどであるため、特徴的変動も月を基本単位として定義する。最大値 / 最小値の検索結果は月となるが、他は条件を満たす任意長の期間となる。

- 最大値 (MAX) / 最小値 (MIN): 各動向情報が最大値 / 最小値を取る月
- 急上昇 (SI) / 急下降 (SD): 3ヶ月以内に、その動向情報の | 最大値 – 最小値 | の 1/5 以上の単調増加 / 減少が見られる期間
- 山形 (PEAK) / 谷形 (BOTTOM): その動向情報の | 最大値 – 最小値 | の 1/10 以上の単調増加 / 減少が見られた後、減少 / 増加に転じた期間

3.3 基本検索機能

1 章でも述べた通り、本論文では「既存検索エンジンよりも動向に関して高度な検索が可能であること」、「組み合わせることで、動向に関するユーザの多様な問いに答えられること」を満たすことを条件として、提案する検索エンジンが提供する基本検索機能およびクエリ構文を定める。検討に際し、ある画像が撮影された場所を既存検索エンジンのキーワード検索のみで探してもらう課題を予備実験として行った[桑折 13]。課題遂行中に見られる多様な検索意図を分析した結果を参考にし、提案する検索エンジンでカバーする検索意図を図 2 の様に定める。検索意図には、新たな情報を得るために行う Discover と、情報の確認を目的とする Verify があるとし、それらはさらに Navigational と Informational 以下に細分化される。一般に、Navigational は特定 Web ページを期待して行う検索、Informational は目的とする情報を所在を問わず探す検索を意味するが[Manning08]、提案システムの検索対象は Web ページではないため、特定アイテムを指定して関連する動向情報を検索する場合を Navigational、アイテムを限定しない検索を Informational と定める。さらに、特徴的変動を指定して行う検索を Pinpoint、指定しない場合を Broad とし、異なる複数の状態が検索条件に該当しうる場合を Multiple、一意に定まる場合を Single とする。Multiple か Single かは特徴的変動に対応して決まり、急上昇 / 急下降や山形 / 谷形は Multiple、最大値 / 最小値は Single となる^{*4}

これらの検索意図に対応するため、以下の 3 つの基本検索機能を提案する。基本検索機能 (1) は Navigational に対応し、(2) は Informational に対応する。Pinpoint, Broad は変動タイプを指定するか否かに対応する。また、動向

^{*4} 最大値 / 最小値も複数期間が検索結果として返される可能性があるが、値 (状態) としては同一となるため Single とみなす。

図2 検索意図の分類

に関する問いの中で、同時期に流行したアイテムの検索は需要が多いと考え、基本検索機能(3)を定める。なお、これは検索結果として任意アイテムを期待するため Informational に対応するとみなす。

- (1) 指定したアイテムに関する動向が特徴的変動を示した期間の検索
 - (2) 指定した期間に特徴的変動を示したアイテム・動向の検索
 - (3) 指定したアイテムに関する動向が特徴的変動を示した期間に同様の変動を示したアイテム・動向の検索
- 以上を踏まえ、クエリ構文を以下のように定める。クエリはアイテム名([Item])、期間([Period])、特徴的変動名([Variation])、および検索結果のタイプ(アイテム(@item)あるいは期間(@period))、リソースの5要素から構成される。検索において指定しない要素は空欄としてよい。基本検索機能(1)は[Item]と@periodを、検索機能(2)は[Period]と@itemを必須項目としてそれぞれ指定する。基本検索機能(3)の場合は[Item]と@itemを必須項目として指定する。[Variation]は、前節で述べた6種類の特徴的変動(MAX, MIN, SI, SD, PEAK, BOTTOM)から一つを指定する。また、全ての特徴的変動を対象とする場合にはBROADを指定する。リソースは、検索対象を指定した動向情報に限定することで検索効率を上げるために指定する。指定は、具体的な統計データ名、公開団体名、あるいはコンテンツとして公開されているものかユーザ活動によるものかなど、複数の粒度で指定することが可能である。

Query → [Item][Period][Variation]
[@item@period][Resource]

3.4 インタフェース

提案する検索エンジンは、Web ブラウザから、および API を利用して任意のアプリケーションから利用可能である。図3は、Web ブラウザでアクセスした場合のプロトタイプインタフェースのスクリーンショットである。図の上部は入力フォーム部分に「2011/03-2011/12の間に最大値をとったアイテムの検索」(基本検索機能(2))をクエリとして入力した例であり、その検索結果を図の下部

図3 プロトタイプインタフェースのスクリーンショット:(上)クエリ入力フォーム(下)検索結果画面

図4 検索結果の動向を示すグラフ

に示している。検索結果として、アイテム名、統計データ名、クエリを満たす期間、該当地域についての情報が返される。検索結果画面において、表示されるアイテム名(左端)をクリックすることで、その動向の折れ線グラフが表示される(図4)。折れ線グラフはJavaScriptのHighchartsを用いて実装している。また、検索結果画面右端のGoogle検索へのリンクをクリックすることで、その動向名(アイテム名+動向情報名)をクエリとし、期間をオプションとして指定したGoogle検索結果が表示され、その当時のWeb上での話題などを知ることができる。

4. 評価実験

4.1 実験概要

本論文で提案する検索エンジンは、「基本検索機能の組み合わせで、多様な情報要求に対応することができる」という既存検索エンジンの利点を継承しつつ、タスクを動向に関する問いに限定することでより高度な基本検索機能を提供する点に特徴がある。この点について有効性を示すために、以下の二点に関して評価実験を行う。

- 情報要求に対して適切なクエリを表現できるか
- 多様な検索意図に基づく検索が実行されるか

(1)に関しては、あらかじめ用意した問題に関して検索を行ってもらった課題を行い、適切なクエリを効率的に生

成できるか検証する。(2)に関しては、基本検索機能を高度化することでその用途が限定され、結果として多様な検索意図に基づく検索が行われなくなる可能性があるため、評価実験での検証目的の一つとした。これに関しては、各自で検索目的を設定してもらい、自由に検索エンジンを利用してもらった結果について分析を行う。以降、前者の実験を課題 1、後者を課題 2 と呼ぶ。

実験協力者は工学系大学生および大学院生 12 名である。各実験には著者の一人が立ち合い、問題解答前に、用意した操作マニュアルにより提案システムの操作について説明を行った。この際、実験協力者は提案システムでの検索練習を行っていない。また、問題解答中には、実験協力者から問題文の解釈について質問を受けた場合のみ、それについて回答した。

課題 1 では、以下に示す 8 つの問題について検索を行い、解答してもらった。以下では、解答を求めるために行う必要がある検索が対応する検索意図(図 2)、および正解が得られるクエリの例も示している。ここで、問題 7 と問題 8 は正解を得るために最低 2 回の検索を必要とする問題であり、検索意図、クエリをそれぞれセミコロンで区切って記載している。また、問題 4 のクエリ例における !S は検索対象リソースをユーザ活動に由来するものに、問題 5 の S+ヒット数はユーザ活動由来の一つであるヒット数に限定することをそれぞれ表している。

- (1) 自転車に関する動向がピーク(山形変動)を迎えた期間は？

検索意図: Navigational-Pinpoint-Multiple

クエリの例: [自転車 PEAK @period]

- (2) 2008 年 5 月～2008 年 12 月に谷を迎えたアイテムは？

検索意図: Informational-Pinpoint-Multiple

クエリの例: [2008/05-12 BOTTOM @item]

- (3) 自転車に関する動向が 2008 年～2010 年の中で特徴的な変動を迎えた期間は？

検索意図: Navigational-Broad

クエリの例: [自転車 2008/01-2010/12 BROAD @period]

- (4) iPad の(Web 上でのユーザ活動に関連する)動向が最小値を迎えた期間は？

検索意図: Navigational-Pinpoint-Single

クエリの例: [iPad !S MIN @period]

- (5) iPad のヒット数が最大である期間に、同じ変動をしたアイテムは？

検索意図: Informational-Pinpoint-Single

クエリの例: [iPad S+ヒット数 MAX @item]

- (6) 2005 年中に特徴的な変動をしたアイテムは？

検索意図: Informational-Broad

クエリの例: [2005 BROAD @item]

- (7) 安倍内閣の内閣支持率が最大である期間に、最小値を迎えたアイテムは？

検索意図: Navigational-Pinpoint-Single; Informational-Pinpoint-Single

クエリの例: [安倍内閣 MAX @period]; [2007/08-09 MIN @item]

- (8) 2004 年に急上昇を迎えたアイテムの中で、2008 年にも急上昇したアイテムは？

検索意図: Informational-Pinpoint-Multiple; Informational-Pinpoint-Multiple

クエリの例: [2004 SI @item]; [2008 SI @item]

課題 2 に関しては、実験協力者各自が任意に選択した期間の中でインパクトが大きかったモノや出来事ベスト 3 を選び、順位付けはせずに回答してもらった。これは課題 1 の実行後に続けて行ってもらい、提案する検索エンジンの他、既存検索エンジンを補足的に利用しても構わないとした。

4.2 実験結果

課題 1 の各問題について、各実験協力者の解答時間を記録し、定量的ユーザビリティ評価手法 NEM (Novice Expert ratio Method) [鱗原 99] により評価する。NEM とは、開発者とユーザの操作時間の比率 (NE 比) に基づき、両者間の操作モデルのギャップを客観的・定量的に見積もる評価手法である。開発者はシステムの利用方法を熟知しているため、その解答時間は提案システムを最も効率的に利用できた場合の解答時間と考える。そこで、実験協力者の解答時間をこれと比較することで、情報要求からクエリへの変換がスムーズに行われるかを定量的に評価し、変換が困難な問題などについて考察する。

実験では、各問題を読む前に [start+問題番号] をクエリとする検索を行い、クエリがサーバに送信された時刻を、その問題における解答開始時刻として扱う。解答終了時刻については、検索結果から解答用紙への記入作業が終了した後で、[end+問題番号] をクエリとする検索を行ってもらい記録する。図 5 に、問題毎の実験協力者の平均解答時間、開発者の解答時間、および NEM 比を示す。解答時間を棒グラフ(左側の軸)、NEM 比を折れ線グラフ(右側の軸)で示している。NEM 比は最小で 1.61 (問題 8)、最大で 2.47 (問題 1) となった。また、いくつかの問題においては、設計者と同等の時間で解答クエリ送信を行う実験協力者も複数名存在した。また、前述したクエリの例とは異なるクエリで解答を得た実験協力者もいたが、開発者と同じクエリを用いた実験協力者と比較して、解答時間に大きな差は生じなかった。

既存検索エンジンの利用ログなどを分析した研究や、実験参加者に既存検索エンジンを利用して検索課題を行ってもらった結果を分析したいくつかの研究において、クエリあたりに費やす時間は 4 分程度、セッションや検索課題あたり 7～10 分前後であることが報告されている [Kim13, Thatcher08, White07]。図 5 に示した実験協力者の解答時間も 5 分以内(問題 8 は 301 秒)に収まって

図5 課題1の解答時間とNEM比

おり、既存検索エンジンを利用した先行研究と同等の結果が得られていると考える。NE比が相対的に高くなっているのは問題1, 4, 7であるが、問題1は最初の問題である事、問題7から初めて複数のクエリが解答に必要となることが原因と考える。問題4に関しては、実験協力者の解答時間は問題1-3とほとんど変わっておらず、設計者の解答時間が短くなった事が原因といえる。

また、入力されたクエリの数、および正しい構文に従ったものの数・割合を表3に示す。実験協力者Cが、半数のクエリにおいて構文が間違っていたが、12名中3名(F, G, I)は全て正しい構文に従ったクエリのみであり、全実験協力者の平均でも79.2%となっている。1章で述べたとおり、既存検索エンジンにおいてセッションあたり2, 3件のクエリが入力されているが、提案する検索エンジンを用いて2回続けて構文を間違える確率は、上述の平均値(79.2%)を用いれば4.3%に過ぎない。また、Exciteのクエリログ約100万件を分析した結果、+オペレータ(必須単語の指定)を含むクエリ44,320件の内、正しく利用されているのは17,028件(38%)に過ぎないことが示されている[Spink01]。さらに、代表的な商用検索サービスであるSiri, Google Nowで正しい回答が得られる確率が79%という調査結果も報告されている[Jones13]。これらより、システム利用経験のないユーザであっても、提案する検索エンジンを用いて「動向に関する問い」を効率的に検索可能と考える。なお、クエリ入力に関して支援を行わない、シンプルなインタフェースで実験を行ったが、実用にあたっては要素ごとに指定可能な検索フォームなどを用意することで、入力精度の改善が期待できる。

課題2に関して、回答例を以下に示す。括弧内は各実験協力者が設定した期間である。各自、自身の高校時代や学部時代などを振り返り、当時関心のあったことを検索を通じて確認し、回答していた。

- 「ニコニコ動画開始、ニンテンドーDS発売、初音ミクがニコニコ動画から有名に」(2006–2008)
- 「民主党政権誕生、ルンパ発売、ユニクロ海外出店」(2008–2010)
- 「東日本大震災、なでしこジャパンW杯優勝、新型

表3 クエリ数と正しい構文によるクエリの割合

実験協力者	クエリ数	正しい構文	割合
A	15	11	0.73
B	11	10	0.91
C	20	10	0.5
D	14	12	0.86
E	14	11	0.79
F	19	19	1.0
G	10	10	1.0
H	15	12	0.8
I	10	10	1.0
J	12	10	0.83
K	22	16	0.73
L	16	10	0.63
平均	14.8	11.8	0.79

インフルエンザ流行」(2009–2011)

- 「ドラゴンクエストIX 発売、ガソリン価格高騰、サンドウィッチマンM1優勝」(2007–2009)

また、図2に示した検索意図毎のクエリ数を表4にまとめる。実験協力者Lには事情により課題2を依頼できず、それ以外の11名に回答してもらった。なお、表中ではMultiple, SingleをそれぞれM, Sと略記している。3.3節で述べたとおり、検索意図は入力されたクエリの形式により形式的に判断している。ただし、情報の発見を目的としたもの(Discovery)か、確認目的(Verify)なのかはクエリのみからでは判断できないため、ここでは区別していない。

表4より、協力者C, E, G, Hの4名はアイテムを指定したNavigationIを一切使用していないことがわかる。Informationalを使用しなかった協力者はいないが、D, I, Kの3名はNavigationalの方を多く利用しており、特に協力者Dは相対的にNavigationalを多用していた。変動タイプを指定しないBroadは6名が利用しており、Navigationalでの利用が多いが、協力者D, J, Kの3名が突出して多く利用している。InformationalでBroadの利用が少ないのは、アイテム、変動タイプ両方を指定しないとあまりに絞り込めていないのが原因と考えられる。以上より、既存検索エンジンを用いた場合と同様に、利用者毎に異なる検索行動が見られ、個々の協力者においても複数の検索意図を組み合わせで提案システムを利用していることがわかる。

5. まとめ

本論文では、動向に関する問いを対象とした検索エンジンを提案した。タスクを動向に関する問いに限定し、指定アイテムに関する動向が特徴的変動を示した期間の検索、指定期間に特徴的変動を示したアイテム・動向の検

表 4 課題 2 における検索意図毎のクエリ数

実験 協力者	Navigational		Broad	Informational		
	Pinpoint			Pinpoint		Broad
	M	S		M	S	
A	0	0	3	1	6	4
B	0	4	0	1	6	0
C	0	0	0	6	7	0
D	0	3	6	0	3	0
E	0	0	0	13	15	3
F	1	2	0	4	6	1
G	0	0	0	3	4	0
H	0	0	0	4	2	0
I	1	9	0	0	9	0
J	5	2	9	9	7	0
K	2	0	7	3	0	2
合計	9	20	25	44	65	10

索, 指定アイテムに関する動向が特徴的変動を示した期間に同様の変動を示したアイテム・動向の検索, の3つを基本検索機能として提案した. 検索インタフェースだけでなく API も備えたプロトタイプシステムを実装し, 実験協力者による評価実験を行った. 指定した問題に対してクエリを生成し, 解を求める課題では, 実験協力者は適切なクエリを十分効率的に生成可能であることを示した. また, 各自で検索目的を設定し, 自由に検索を行ってもらった課題では, 協力者毎に多様な検索行動が見られたこと, 各協力者は自身の情報要求を意図の異なる複数のクエリに分解し, 組み合わせることが可能であることを示した. これより, 多様な情報検索要求を基本検索機能に分解して満たすことができるという, 既存検索エンジンの重要な特徴を継承していると考えられる.

今後の課題として, プロトタイプシステムの機能向上, データベースの拡充, 運用を通じた検証実験などが挙げられる. 機能向上に関しては, 周期的変動など, 検索可能な特徴的変動の種類を増やすこと, 検索フォームの整備などが考えられる. データベースの拡充に関しては, 総務省統計局による政府統計データ API や SPARQL エンドポイントなど, LOD の活用を進める予定である. また, 既存検索エンジンの様に, 検索スキルに応じ異なる検索戦略などが採用されるのか, あるいは既存検索エンジンでのスキルが提案する検索エンジンの利用方法に影響を与えるのかについても評価実験を行いたいと考えている.

謝 辞

本研究は JSPS 科研費 24650040 の助成を受けたものです.

◇ 参 考 文 献 ◇

- [Agichtein12] E. Agichtein, R. W. White, S. T. Dumais, P. N. Bennett, Search, Interrupted: Understanding and Predicting Search Task Continuation, SIGIR'12, pp. 315–324, 2012.
- [Bollen10] J. Bollen, H. Mao, X.-J. Zeng, Twitter Mood Predicts the Stock Market, CoRR abs/1010.3003, 2010.
- [Ferreira05] A. Ferreira, J. Atkinson, Intelligent Search Agents Using Web-Driven Natural-Language Explanatory Dialogs, IEEE Computer, Vol. 38, No. 10, pp. 44–52, 2005.
- [Ferrucci10] D. Ferrucci et al., Building Watson: An Overview of the DeepQA Project, AI Magazine, Vol. 31, No. 3, pp. 59–79, 2010.
- [蓮井 13] 蓮井, 松下, 言語表現による時系列データ検索システムの提案, 人工知能学会第 3 回インタラクティブ情報アクセスと可視化マイニング研究会, pp.58–62, 2013.
- [石黒 12] 石黒, 高間, 為替レートを対象としたコンテキスト検索システムの提案, FSS2012, TF3–3, pp. 199–202, 2012.
- [和泉 09] 和泉, 後藤, 松井, テキスト情報による金融市場の逐次外挿予測, 人工知能学会研究会資料, SIG-FIN-003-02, 2009.
- [Jones13] Apple's Siri And Google Both Get C+ Grades From Gene Munster, Forbes.com, <http://www.forbes.com/sites/chuckjones/2013/12/10/apples-siri-and-google-both-get-c-grades-from-gene-munster/>, 2013. (2014.7.13 現在)
- [亀井 07] 亀井, 門田, 松本, WWW を対象としたソフトウェア検索エンジンの構築, 電子情報通信学会技術研究報告ソフトウェアサイエンス, Vol.102, No617, pp.59–64, 2007.
- [加藤 13] 加藤, 桑折, 高間, 「動向に関する問い」を対象タスクとしたコンテキスト検索の提案, 人工知能学会第 3 回インタラクティブ情報アクセスと可視化マイニング研究会, pp.7–12, 2013.
- [加藤 04] 加藤, 松下, 平尾, 動向情報の要約と可視化に関するワークショップの提案, 情報処理学会研究報告 / 自然言語処理研究会報告, 2004(108), pp.88–94, 2004.
- [加藤 12] 加藤, 高間, Web コンテキスト情報に基づく同時期流行アイテム検索手法の提案, FSS2012, pp. 115–118, 2012.
- [Kim13] J. Kim, M. Cramer, J. Teevan, D. Lagun, Understanding How People Interact with Web Search Results that Change in Real-Time Using Implicit Feedback, CIKM'13, pp. 2321–2326, 2013.
- [Kleinberg02] J. Kleinberg, Bursty and Hierarchical Structure in Streams, Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1–25, 2002.
- [小久保 02] 小久保, 小山, 山田, 北村, 石田, 検索隠し味を用いた専門検索エンジンの構築, 情報処理学会論文誌, Vol.43, No.6, pp. 1804–1813, 2002.
- [桑折 13] 桑折, 加藤, 高間, 検索エンジンを用いた情報検索におけるユーザ行動の分析, 人工知能学会第 4 回インタラクティブ情報アクセスと可視化マイニング研究会, pp.9–14, 2013.
- [Manning08] C. D. Manning, P. Raghavan, H. Schütze, Chapter 19: Web Search Basics, Introduction to Information Retrieval, Cambridge University Press, 2008.
- [松下 06] 松下, 加藤, 数値情報の補填とグラフ概形の示唆による複数文書からの統計グラフ生成, 知能と情報, Vol. 18, No. 5, pp.721–734, 2006.
- [宮川 06] 宮川, 鈴木, 鍋島, 岩沼, 専門検索エンジン的高速半自動生成法, 第 5 回情報科学レクターズ LL-005, pp.355–358, 2006.
- [Mobasher06] B. Mobasher, Web Usage Mining, Encyclopedia of Data Warehousing and Mining, Idea Group, 2006.
- [森本 10] 森本, 辻, 山田, 学習コンテンツのメタデータ検索エンジンの開発: 電子情報通信学会ライフインテリジェンスとオフィス情報システム研究会, 110(42), pp.13–17, 2010.
- [灘本 07] 灘本, 堀内, ブロガーの注目情報を用いた株価変動予測の試み, 情報科学技術フォーラム一般講演論文集 6(2), pp.369–370, 2007.
- [NY] The New Yorker, Inequality and New York's Subway, <http://www.newyorker.com/sandbox/business/subway.html> (2014/5/6 現在)
- [NYT09] The New York Times, Clean Water Laws Are Neglected, at a Cost in Suffering, http://www.nytimes.com/2009/09/13/us/13water.html?_r=0 (2014/5/6 現在)
- [西藤 03] 西藤, 喘息用語に特化した検索エンジンの構築と運用

について 検索エンジンを用いたインターネットで求められている喘息に関する情報の調査, アレルギー, 52(12), pp.1122-1131, 2003.

- [奥田 07] 奥田, 難波, 奥村, 新聞記事と blog からの動向情報の抽出と可視化, 言語処理学会 第 13 回年次大会, C5-4, 2007.
- [奥村 04] 奥村, 南野, 藤木, 鈴木, blog ページの自動収集と監視に基づくテキストマイニング, 人工知能学会セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01, 2004.
- [大向 13] オープンデータと Linked Open Data, 情報処理, Vol. 54, No. 12, pp. 1204-1210, 2013.
- [Pierrakos03] D. Pierrakos, et al., Web Usage Mining as a Tool for Personalization: A survey, User Modeling and User-Adapted Interaction 13, pp.311-372, 2003.
- [齋藤 04] 齋藤, 三輪, Web 情報検索におけるリフレクションの支援, 人工知能学会論文誌, Vol. 19, No. 4, pp. 214-224, 2004.
- [佐々木] 佐々木俊尚公式サイト, データジャーナリズムとは何か?, <http://pressa.jugem.jp/?eid=200> (2014/5/6 現在)
- [Silverstein99] C. Silverstein, M. Henzinger, H. Marais, M. Moricz, Analysis of a Very Large Web Search Engine Query Log, SIGIR'99, pp. 6-12, 1999.
- [Spink01] A. Spink, D. Wolfram, M. B. J. Jansen, T. Saracevic, Searching the Web: The Public and Their Queries, Journal of the American Society for Information Science and Technology, Vol. 52, Issue 3, pp. 226-234, 2001.
- [Srivastava00] J. Srivastava, et al., Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD Explorations Newsletter 1.2, pp.12-23, 2000.
- [鈴木 05] 鈴木, 鍋島, 岩沼, 精錬手法に基づく検索隠し味型専門検索エンジンの半自動構築, 情報科学技術フォーラム一般講演論文集, 4(4), pp.199-202, 2005.
- [高橋 11] 高橋, 安村, 上原, 半教師付き学習を用いた話題出現頻度推移からの注目話題の早期発見, 電子情報通信学会知能ソフトウェア工学研究会, 110(386), pp. 31-36, 2011.
- [立園 13] 立園, 世界のデータジャーナリズム最前線, Web インテリジェンスとインタラクション研究会特別講演発表スライド <http://www.slideshare.net/tsudamag/vol105-ddj> (2014/5/6 現在)
- [Thatcher08] A. Thatcher, Web search strategies: The influence of Web experience and task type, Information Processing and Management, Vol. 44, pp. 1308-1329, 2008.
- [徳永 07] 徳永, 言語処理を利用した知的情報アクセス 検索, 抽出, 要約, 分類, QA, オペレーションズ・リサーチ 経営の科学, 52(11), pp.713-718, 2007.
- [鱗原 99] 鱗原, 古田, 田中, 黒須, 設計者と初心者ユーザーの操作時間比較によるユーザビリティ評価手法, ヒューマンインタフェースシンポジウム '99 論文集, pp.537-542, 1999.
- [White07] R. W. White, S. M. Drucker, Investigating Behavioral Variability in Web Search, WWW'07, pp. 21-30, 2007.
- [山本 07] 山本, 谷岡, 殿井, 動向情報の検索による情報編纂, 第 21 回人工知能学会全国大会, 3H9-3, 2007.

[担当委員: 庄司 裕子]

2014 年 5 月 8 日 受理

著者紹介

高間康史(正会員)

1994 年 東京大学工学部電子工学科卒業. 1999 年同大学院博士課程修了. 1999-2002 東京工業大学大学院総合理工学研究科助手, 2002-2005 東京都立科学技術大学助教授, 2005-2014 首都大学東京システムデザイン学部准教授, 2014 年より同教授. 博士(工学). Web Intelligence や情報可視化, 知的インタフェースの研究に従事. 主要著書は「インテリジェントネットワークシステム入門」(コロナ社). IEEE, 日本知能情報ファジィ学会, 情報処理学会, 電子情報通信

学会各会員.

加藤 優(学生会員)

2012 年首都大学東京システムデザイン学部卒業. 2014 年同大学大学院システムデザイン研究科博士前期課程修了. 在学中は検索エンジンの研究に従事.

桑折章吾(学生会員)

2013 年首都大学東京システムデザイン学部卒業. 2013 年より同大学大学院システムデザイン研究科博士前期課程在籍. 検索エンジンの研究に従事.

石川博

首都大学東京システムデザイン学部情報通信システムコース教授. 東京大学理学部情報科学科卒業. 東京都立大, 静岡大を経て 2013 年より現職. 東京大学博士(理学). 著書に『データマイニングと集合知』(共立出版)等, ACM TODS, IEEE TKDE, 国際学会 VLDB, IEEE ICDE 等を含め学術論文多数. 1994 年情報処理学会坂井記念特別賞, 1997 年科学技術庁長官賞(研究功績者)受賞. 情報処理学会データベースシステム研究会主査, 情報処理学会論文誌(データベース)共同編集委員長, International Journal Very Large Data Bases Editorial Board, 日本データベース学会理事歴任. 情報処理学会フェロー, 電子情報通信学会フェロー. ACM, IEEE 各会員.