「動向に関する問い」を対象タスクとした コンテクスト検索の提案

Proposal of Context Search Engine Focusing on Trend-related Queries

加藤 優[1]　桑折 章吾[2]　高間 康史[1,2*]

Yu Kato[1], Shogo Kori[2], Yasufumi Takama[1,2]

[1] 首都大学東京大学院システムデザイン研究科
[1]Graduate School of System Design, Tokyo Metropolitan University
[2] 首都大学東京システムデザイン学部
[2]Faculty of System Design, Tokyo Metropolitan University

Abstract: 本稿では，「動向に関する問い」を対象タスクとしたコンテクスト検索を提案する．既存の検索エンジンは汎用的に利用可能な反面低機能なため，情報要求をクエリに分解するのに要するユーザの負担は大きい．本稿で提案するコンテクスト検索は，タスクを限定することで高度な検索機能を提供する．動向に関する問いは広く一般に見られるものであり，提案手法は幅広いドメインに貢献することが期待できる．

1　はじめに

本稿では，「去年流行したアイテムは？」や「東日本大震災の影響を受けたアイテムは？」といった「動向に関する問い」に対して行う検索をコンテクスト検索と定義し，これに適した基本検索機能を提供する次世代検索エンジンについて提案する．

現在，検索エンジンを利用した情報収集・分析作業はドメイン・タスクを問わず広く一般に行われているが，既存検索エンジンが提供する機能と，ユーザの情報収集目的との乖離が大きいという問題がある．すなわち，既存検索エンジンが提供するのは，キーワードベースの検索要求指定，ページ単位での結果出力といった低機能にとどまったままであり，情報要求を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と考える．

次世代検索エンジンの実現に向けて，自然言語文での問い合わせを受付け，ユーザの問いに直接回答するような検索エンジンの知的化のアプローチも考えられるが，本稿では検索エンジンが提供する基本検索機能を見直すことにより，ユーザの情報要求とのギャップを小さくするアプローチを採用する．基本検索機能として，「動向に関する問い」というタスクに着目する．近年，人気や流行といったアイテムの動向に関する問いは一般的なものと考える．

検索エンジンの知的化において，十分な性能を得るためには対象ドメインを限定する必要があると考えられるのに対し，本稿ではドメインに依存しないタスクを対象とすることにより，広く一般的に利用可能な検索エンジンの実現を目指す．現在の検索エンジンがユーザを限定せず，日常的に用いられる存在である以上，対象ドメインを限定しない本稿のアプローチは，次世代検索エンジン実現において重要な視点と考える．

本稿では，コンテクスト検索のコンセプトについて提案すると共に，現在構築中のプロトタイプシステムについて述べる．web で入手可能な動向情報は，検索エンジンでの検索数やヒット数などに表れる主観的動向情報と，官公庁を含めた様々な組織・機関が公開する価格や生産量のデータ，統計データなどの客観的動向情報に大別できる．本稿では，それらの動向情報を Web 上から抽出し，データベースを構築する．システムが提供する基本検索機能として，「指定アイテムに関する動向情報のピーク時期検索」，「指定期間に動向情報のピークを持つアイテム検索」を提案する．構築したプロトタイプシステムを用いて検索を行った事例を示す．

*連絡先：　高間 康史
首都大学東京大学院システムデザイン研究科
〒191-0065 東京都日野市旭が丘６－６
E-mail: ytakama@sd.tmu.ac.jp

2 関連研究

2.1 次世代検索システムへの試み

Web が普及してから 20 年弱が経過し，Web 上には膨大な量の情報が蓄積されている．現在，最も用いられている情報検索手法は，検索エンジンを利用する方法である．しかし，既存の検索エンジンによるキーワードベースの検索は，ユーザが入力したキーワードを含むページを探すという低機能なものにとどまったままであり，情報要求をキーワードに分解する際のユーザの負担が大きいという問題がある．このようなユーザへの負担を軽減するために次世代検索システムの開発・研究がなされている[2][5].

亀井ら[2]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検索するための検索エンジン構築を提案している．多くのソフトウェアが開発されているが，それらの知識は必ずしも有効に蓄積・利用されていないために，似たようなソフトウェアが開発されていたり，同じようなミスでソフトウェア開発が滞ることがある．それらの問題を解決するため，巡回ロボットにより，Web 上に存在するソフトウェア資源を収集し，ソフトウェアメトリクスやパッケージ名，クラス名などの指定によりユーザに適切な情報を提供する検索エンジンを構築している．

小久保ら[5]は，新たな専門検索エンジンの構築手法として，「検索隠し味」を用いる方法を提案している．検索隠し味とは，機械学習の一種である決定木学習アルゴリズムを元に，Web ページ集合から抽出したブール式であり，ユーザの入力クエリに加えることで，汎用検索エンジンの検索結果をある特定ドメインに特化させることが可能となる．

これらを含めた多くの次世代検索システムの研究では，ドメインを狭い領域に限定することで検索性能の向上を図っている．自然言語によるクエリを受け付ける検索エンジンも次世代検索エンジンの一つとみなせるが[1]，この場合も性能向上のためにはドメインの限定が必要になると考える．これに対し，本稿で提案する検索システムでは，ドメインに依存しないタスクを対象とすることにより，広く一般的に利用可能な検索エンジンの実現を目指す．

2.2 動向情報に着目をした研究

動向情報とは，ある商品の価格や売上げの状況，ある会社の業績状況，内閣や政党の支持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，それらを総合的にまとめ上げることで得られるものである[3]．これら動向情報は，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おり，世の中の社会活動に深く関わっている．近年，官公庁を含めた様々な組織・機関による情報公開が進み，Web 上には，多種多様で大規模な動向情報が蓄積されている．この流れは，今後も益々進んでいくことが予想される．このような背景から動向情報を利用した研究が多くなされている[4][6][7].

松下ら[6]は，動向情報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．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が可能となる．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，動向の理解を支援している．

山本ら[7]は，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，その動向情報を効率的に獲得する手法を提案している．山本らが提案するシステムを用いることにより，「ある会社の株価の変動と同期している株価をもつ会社を探したい」や，「ある製品の売上げの変動とともに使用されるようになった単語を知りたい」といった問いに答えることができる．

3 動向情報を対象とした コンテクスト検索システム

3.1 システム構成

提案するコンテクスト検索システムの構成を図 1 に示す．提案システムでは，Web 上から抽出した動向情報を事前に抽出し，データベースに格納しておく．データベース管理システム（DBMS）には MySQL を利用し，Web サーバの実装には Webrick を用いている．Web アプリケーションフレームワークには，Ruby on Rails を使用した．

動向情報は，検索エンジンの検索数やヒット数などの主観的動向情報と，アイテムの価格や生産量データ，統計データなどの客観的動向情報に分けられる．3.2 節，3.3 節に主観的動向情報および客観的動向情報の抽出手法をそれぞれ示す．

あるアイテムに関する動向を調査する際には，アイテムの人気や流行に応じて変動する動向情報において，その変動の最大値の検索が重要であると考える．そのため，本稿で紹介するプロトタイプシステムでは「指定アイテムに関する動向情報のピーク（最

大値）時期の検索」,「指定期間に動向情報の最大値を持つアイテムの検索」の 2 つを基本検索機能として実装している.

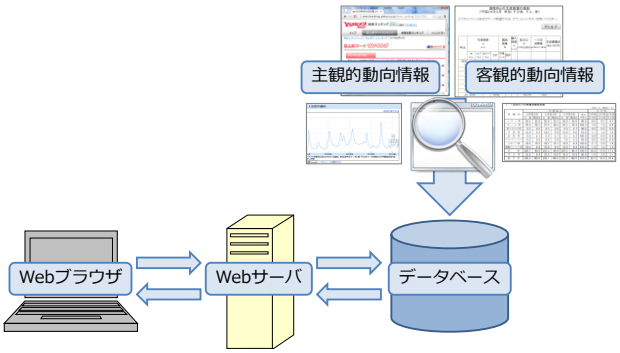

図 1　コンテクスト検索システムの構成

3.2 主観的動向情報抽出

主観的動向情報とは，検索エンジンの検索数やヒット数，ブログの記事数などのユーザの興味や関心に基づいて値が増減する情報であり，それらを Web から抽出する．現在，抽出対象としている主観的動向情報とその情報源を表 1 に示す.

表 1 抽出対象の主観的動向情報

分類	情報源	URL
検索数（指数）	Google Trends	http://www.google.com/trends/
ヒット数	Yahoo!検索（ウェブ検索）	http://www.yahoo.co.jp/
ブログ記事数	Yahoo!検索（ブログ検索）	http://search.yahoo.co.jp/blog
急上昇ワードランキング	Yahoo!検索ランキング	http://searchranking.yahoo.co.jp/
きざしランキング	kizasi.jp	http://kizasi.jp/
HOT ワード	ついっぷるトレンド	http://tr.twipple.jp/

検索数は，検索数の推移を調査することができるサービスである Google Trends[1]から取得している．Google Trends で取得できる値は，各単語が Google で検索された回数を 1 週間単位で集計し，検索された総回数に対する相対値を 0～100 の指数で表したものである．ヒット数・ブログ記事数は，Yahoo!検索サービスにおいてウェブ検索・ブログ検索を利用して検索した際の検索結果件数を取得している．急上昇ワードランキングは，Yahoo!JAPAN が運営する Yahoo!検索ランキング[2]，きざしランキングは kizasi.jp[3]，HOT ワードは，Twitter 話題ランキングサイトのついっぷるトレンド[4]がそれぞれ提供しているランキング結果を Web スクレイピングによって Web ページから抽出している.

3.3 客観的動向情報抽出

客観的動向情報とは，販売量や売上高のデータ，統計データなどの定量的な測定が可能な情報であり，これらも Web から抽出する．主観的動向情報と異なる点として，これらのデータは集約されておらず，各企業・団体などでそれぞれ公開されている点，その公開形式も様々である点が挙げられる．一般的な公開形式として，Web ページに HTML で直接記載されている他，CSV・PDF・Excel などが用いられる.

HTML から情報を抽出する場合には，Ruby のライブラリである nokogiri を用いて HTML 解析を行う．多くの Web ページ内では，数値情報は表形式となって表されているため，HTML の<table></table>タグで囲まれた箇所から，各セルを意味する<td></td>タグ内の情報を抽出する.

Excel・CSV 形式の場合は，ファイルをダウンロードし，数値などの重要な情報が記載されているセルから情報を抽出する.

PDF の場合には，PDF ファイルの全文をテキストファイルに変換可能なツールである xdoc2txt[5]を用いてテキストファイルに変換し，不要な情報を除去して情報を抽出する.

前述の通り，客観的動向情報は多くの Web サイトに分散して存在するため，網羅的な収集は困難である．現状では，野菜や即席めんなどの価格や生産量などに関する情報を中心に 31 種類の客観的動向情報を収集しているが，今後も拡充していく予定である.

4　提案システムを用いた検索事例

プロトタイプシステムを用いて，想定する検索タスクについて，検索を行った事例を紹介する．プロトタイプシステムでは，主観的動向情報として Yahoo!検索や Google Trends など 3.2 節に示した 6 つの情報源から取得した動向情報を，客観的動向情報

[1] http://www.google.com/trends/

[2] http://searchranking.yahoo.co.jp/
[3] http://kizasi.jp/
[4] http://tr.twipple.jp/
[5] http://www31.ocn.ne.jp/~h_ishida/xdoc2txt.html

として 3.3 節に示した抽出方法によって，統計局や産業振興協会など 7 つの Web サイトから取得した 31 種類の動向情報をそれぞれデータベースに格納している．3.1 節で述べたように，プロトタイプシステムでは基本検索機能として「アイテムから探す」と「期間から探す」の 2 つを提供しており，ユーザは自身の「動向に関する問い」を，これらの基本検索，および既存検索エンジンへのクエリに分解して調べることを想定している．

提案システムの入力画面を図 2 に示す．上部のラジオボタンによって「アイテムから探す」，「期間から探す」を選択可能である．「アイテムから探す」を選択した場合は，検索ボックス内に検索したいアイテム名を入力することで，指定したアイテムに関する主観的・客観的動向情報の最大値およびその時期，情報を公開している Web サイトの URL，動向情報の変化を表したグラフが出力される（図 3）．「期間から探す」を選択した場合は，セレクトボックスに検索したい期間を月単位で指定することで，指定した期間内に動向情報の最大値を持つアイテム名，動向情報の最大値，URL，グラフを出力する（図 4）．

4.1 節に基本検索機能を用いた検索事例を，4.2 節にプロトタイプシステムと既存検索エンジンを併用した検索事例を示す．

図 2 提案システムの入力画面

4.1 基本検索機能を用いた検索事例

以下に，基本検索機能である「アイテムから探す」を選択した場合の検索事例と「期間から探す」を選択した場合の検索事例を示す．

● 「アイテムから探す」を利用した検索

ユーザが「野菜」に関する動向情報について調査したいと考えた場合を想定する．この場合，ユーザは入力フォーム上部の「アイテムから探す」を選択した上で，検索ボックスに「野菜名」（例えば，にんじん）を入力し，検索を実行する．プロトタイプシステムによる検索結果を図 3 に示す．システムによる出力から，ユーザは「にんじんの価格」の最大値が 2006 年 8 月の 517 円であることを知ることができる．

図 3 「アイテムから探す」を選択した場合の出力画面

● 「期間から探す」を利用した検索

ユーザが「過去に流行したアイテム」について関心を持ち，該当するアイテムを調査したいと考えた場合を想定する．この場合，ユーザは入力フォーム上部の「期間から探す」を選択した際に表示されるセレクトボックスに検索対象の期間（例えば，2011 年 3 月～2011 年 9 月）を指定し，検索を実行する．プロトタイプシステムによる検索結果から，ユーザは，対象期間に「自転車の販売量」などが最大値を迎えたことを知ることができる．

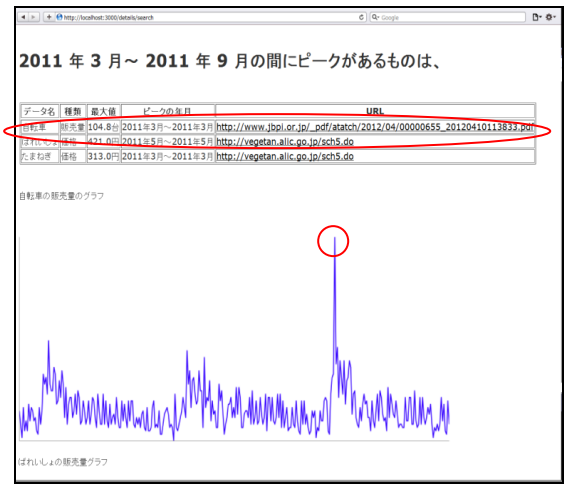

図 4 「期間から探す」を選択した場合の出力画面

4.2 提案システムと既存検索エンジンを併用した検索事例

提案システムを利用した実際の動向情報調査では，

4 節冒頭で述べたプロトタイプシステムの基本検索機能の他，既存検索エンジンの併用を想定している．本節では，基本検索機能と既存検索エンジンの両方を用いて「東日本大震災の影響を受けたアイテムの調査」と「過去において同時期流行したアイテムの調査」という動向に関する問いに答える検索事例を示す．

● 東日本大震災の影響を受けたアイテムの調査

ユーザが「東日本大震災がアイテムに与えた影響」に関心を持ち，様々なアイテムに関する動向情報の震災後における変化について調査したいと考えた場合を想定する．この場合，ユーザは「期間から探す」を選択し，クエリとして「2011 年 1 月～2011 年 12 月」を指定し，検索を実行する（図 5）．

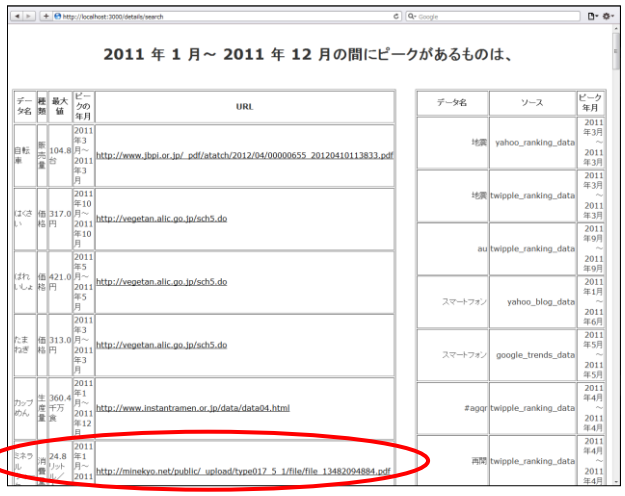

図 5 プロトタイプシステムの検索結果
（クエリ：2011/01～2011/12）

このとき，ユーザは検索結果から「ミネラルウォーターの消費量」が対象期間に最大値を迎えていることに興味を持ったとする．この場合には，続いてプロトタイプシステムの「アイテムから探す」から「ミネラルウォーター」をクエリに検索を実行し，さらに詳しい情報を得ることができる（図 6）．図より，「消費量」だけでなく，「検索数」や「ブログ記事数」などの主観的動向情報においても同期間に最大値を迎えていることが読み取れる．そこで既存検索エンジンを用いて，「ミネラルウォーター 2011」で検索した結果（図 7）から，ユーザはミネラルウォーターの消費量や検索数，ブログ記事数などの動向情報が大きく値を伸ばし，最大値を迎えたのは，東日本大震災の影響を受けたためではないかと推測することができる．この様に，提案する基本検索機能を用いることで，関心のあるアイテムを絞り込み，既存検索エンジンで効率良い情報収集が可能となる．

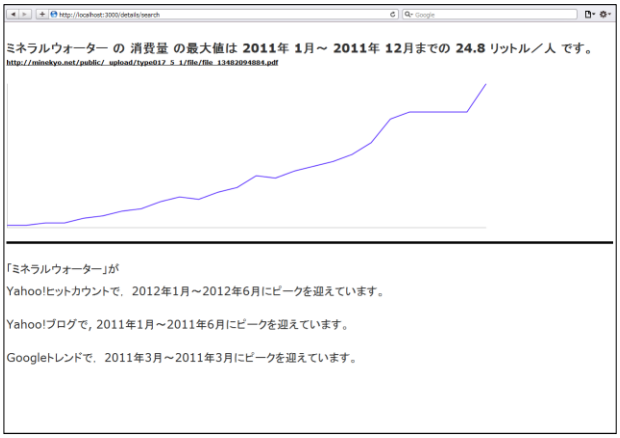

図 6 プロトタイプシステムの検索結果
（クエリ：ミネラルウォーター）

図 7 既存検索エンジンの検索結果
（クエリ：ミネラルウォーター 2011）

● 過去において同時期流行したアイテムの調査

ユーザが「過去において同時期に流行したアイテム」について調査したいと考えた場合の検索の流れを図 8 に示す．この検索には，状況に応じて，いくつかの異なる方法が考えられる．

一つは，ユーザが調査したい対象アイテムを想定している場合である．この場合には，プロトタイプシステムの「アイテムから探す」を用いて，対象アイテムの動向ピーク期間を調べたあとで，「期間から探す」によって，基準となる対象アイテムが動向の最大値を迎えた期間に，同じく動向の最大値を迎えているアイテム群を検索可能である．さらにその際

に，既存検索エンジンによる検索を併用し，実際にどのように話題となったのかを確認することで，流行の根拠を知ることができると考えられる．

ユーザが調査したい期間を予め想定している場合には別の方法が考えられる．その場合には，プロトタイプシステムの「期間から探す」を実行し，得られた結果から，興味を抱いたアイテムについて，「アイテムから探す」の実行や，既存検索エンジンでの検索により，調査を進めていくことが可能である．

また，どちらの方法であっても，既存検索エンジンを用いた検索中に，新しく関心の湧いたアイテムを発見した場合には，そのアイテムをプロトタイプシステムの「アイテムから探す」を用いて検索し，そのアイテムの動向情報を得ることも想定している．

図 8 同時期流行アイテム検索の流れ

5 おわりに

本稿では，「動向に関する問い」を対象タスクとして行う検索をコンテクスト検索と定義し，これに適した基本検索機能を提供する検索エンジンのプロトタイプシステムを構築した．また，動向に関する問いの例として，「東日本大震災の影響を受けたアイテムの調査」や「過去において同時期流行したアイテムの調査」というタスクに対して調査を行う事例を想定し，ユーザの情報要求が基本検索機能および既存検索エンジンへのクエリの組み合わせに分解される様子を示した．本稿で提案するコンテクスト検索は，タスクを限定することで高度な検索機能を提供しつつ，幅広いドメインへの適用が期待できるものであり，次世代検索エンジン実現に適した性質を備えていると考える．開発中のプロトタイプシステムでは，指定したアイテムに関する動向情報のピーク時期の検索，指定した期間に動向情報の最大値を持つアイテムの検索という2つの基本検索機能を提供するが，今後より充実させていく予定である．また，検索対象となる情報も，現状では主観的動向情報が6つの情報源から6種類，客観的動向情報が7つのwebサイトから31種類と小規模であるが，今後，収集する動向情報の量を増やすことで，さらに多くの問いに対して答えることが可能となる．構築したシステムを公開し，運用を通じて必要な基本検索機能の検討やユーザインタフェースの改良を行うことも重要であると考える．

参考文献

[1] A. Ferreira, J. Atkinson: Intelligent Search Agents Using Web-Driven Natural-Language Explanatory Dialogs, IEEE Computer, Vol. 38, No. 10, pp. 44-52 (2005)

[2] 亀井 俊之, 門田 暁人, 松本 健一: 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 102(617), pp.59-64 (2003)

[3] 加藤 恒昭, 松下 光範, 平尾 努: 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.89-94 (2004)

[4] 加藤 優, 高間 康史: Web コンテクスト情報に基づく同時期流行アイテム検索手法の提案, ファジィシステムシンポジウム講演論文集 28, pp.115-118 (2012)

[5] 小久保 卓, 小山 聡, 山田 晃弘, 北村 泰彦, 石田 亨: 情報処理学会論文誌 43(6), pp.1804-1813 (2002)

[6] 松下 光範, 加藤 恒昭: 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, 日本知能情報ファジィ学会誌 知能と情報 18(5), pp.721-734 (2006)

[7] 山本 健一, 谷岡 広樹, 殿井 加代子: 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 (JSAI2007), 3H9-3 (2007)