人工知能学会 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(第11回)
SIG-AM-11-02

コンテクスト検索エンジンへの ランキング機能の導入に関する検討

Consideration of Introducting Ranking Function to Context Search Engine

手塚 拓哉[1]　山口 晃一[2]　諸 琰俊[2] 桑折 章吾[2] 高間 康史[2]

Takuya Tezuka[1], Koichi Yamaguchi[2], Yanjun Zhu[2], Shogo Kori[2], and Yasufumi Takama[2]

[1] 首都大学東京システムデザイン学部
[1] Faculty of System Design, Tokyo Metropolitan University

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

Abstract: This paper aims to introduce a ranking function to context search engine. Context search engine has been developed for answering trend-related queries. In order to achieve a more efficient search, ranking retrived results, which is one of important function of exsisting Web search engines, is expected to be effective also for the context search engine. This paper discusses several features that could be used for ranking function of context search engines, such as intensity and periodicity of temporal change. The result of ranking retrieved results with the intensity of temporal change is also shown.

1. はじめに

本稿では, 動向に関する問いにタスクを限定したコンテクスト検索エンジンにおいて, より効率的な検索を実現することを目標として, ランキング機能の導入について検討する.

Web 上には多種多様で膨大な量の情報が日々蓄積され続けられている. Web を利用することにより発見することのできる情報も増え, ユーザの求める情報も多様化している. その結果, ユーザの情報要求と既存検索エンジンの提供する基本検索機能の乖離が大きくなっている. この問題に対する解決策の一つとして, タスクを「動向に関する問い」に限定することにより, ドメインを限定せずに高度な検索機能を提供するコンテクスト検索エンジンが提案されている[1]. コンテクスト検索エンジンを利用することにより, 時間的変動の観点から関係のあるアイテムを発見するタスクなどにおいて, 有効性を確認している. 検索対象となる動向データとして Web 上のオープンデータを収集しており, 2015 年 7 月の時点で 27,848 アイテムが検索可能となっている[2]. 今後もデータベースを拡張していくことが予定されているが, 検索対象アイテム数の増加に伴い, 検索結果として返されるアイテム数も増加している. 現在のコンテクスト検索エンジンでは検索結果は順位づけられていないため, 結果の確認にかかるユーザの負担が課題となっている. そこで, より効率的な検索を実現するために, 本稿ではランキング機能の導入について検討する.

現在の Web 検索エンジンでは, PageRank スコアや文書適合度など多様な素性を用いて Web ページのスコアを計算する[3]. また, 各素性のスコアにおける重みはランキング学習[4]を用いて決定することが一般的になっている. 本稿でも, 同様の枠組みによりランキング機能を実装することを検討する.

上述の枠組みによるランキング機能の導入においては, スコア計算に用いる素性, およびランキング学習に用いる訓練データについて検討する必要があるが, 本稿では素性について検討する. 既存検索エンジンは, 基本検索機能としてキーワードベースのクエリを入力とし Web ページを検索結果として出力する. しかし, コンテクスト検索エンジンの検索対象は時系列データであり, クエリはアイテム名と期間, 変動タイプといった違いがある. このため, 既存検索エンジンと同様の素性を用いることができないため, 検索タスク・対象データに適した素性を新たに検討する必要がある.

本稿では, クエリ独立の素性として, 変動の激し

さや周期性などについて検討する．また，変動の激しさを利用したランキング機能を実装した結果を示し，その効果について考察する．

2. 関連研究

2.1. コンテクスト検索エンジン

現在，Web における情報アクセス手段として，キーワードを用いて Web ページを検索する検索エンジンが普及している．しかし，これら既存の検索エンジンには，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や，個々の情報要求に合わせ，適切なクエリに分解するのに要するユーザの負担が大きいことが問題として指摘されている．

この問題に対して，動向に関する問いに答える問いう幅広いドメインで見られるタスクに限定することにより，ドメインによらず利用可能という既存検索エンジンの利点を維持しつつ，より高度な検索機能をもつコンテクスト検索エンジンが提案されている[1].

コンテクスト検索エンジンでは，動向情報として「コンテンツとしての動向情報」と「ユーザ活動による動向情報」の 2 種類を Web から収集し，検索対象としている．前者の例として商品の価格や生産量，人口，後者の例として GoogleTrends やきざしランキングから得られるデータなどが収録されている．

それらの時系列データに対する基本検索機能は，Google を利用した検索作業において観測された検索意図[5]を元に決定されている．具体的には，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，指定した期間に特徴的変動を示したアイテムの検索，指定したアイテムに関する動向が特徴的変動を示したアイテムの検索の 3 つの基本検索機能が利用可能となっている．また，最大値や急上昇などの 6 種類の特徴的変動タイプを時系列データから抽出し，検索条件として指定可能である．

2.2. ランキング機能に用いる素性

既存のキーワードを用いて Web ページの検索を行う検索エンジンでは，Web ページの検索結果としての適合度を計算するために，多種多様な素性を用いている[3]．それらは，クエリ依存の素性，クエリ独立の素性に大別することができる．クエリ依存の素性とは，入力されたクエリと Web ページの関係からスコアを求めるものであり，BM25[7]や TF-IDF などクエリと Web ページの関連度に関する素性がよく用いられる[8].

これに対して，クエリ独立の素性とは，入力されたクエリに関係なく Web ページのスコアを決定するものであり，Web のリンク構造を利用した Web ページの重要度である PageRank[9]やアンカーテキストなどがある．

他にも Web 検索エンジンで利用されていると思われるランキングの素性として，Twitter や facebook などの SNS のアカウントに信頼度を付与し，投稿された短文のリンクに重みをつけるソーシャルシグナルや，検索履歴やクリックログなどのユーザシグナルを利用した素性がある[10].

ランキングの素性に利用されたものではないが，時系列データの特徴としては，その時間的変動に基づくものが考えられる．蓮井らは，言語表現を用いた時系列データ検索システムを提案している[6]．グラフの変動，変化の度合い，グラフの概形に着目し，グラフの始点と終点の範囲から「上昇」「下降」「安定」，傾きから「大きく」「小さく」「なだらかに」などの特徴を抽出している．

周期性の検出には周波数分析がよく用いられる．動向情報の周期性を判定する方法として，綱元らは web ページがブックマークされるタイミングの周期性を離散フーリエ変換とパワースペクトルを用いて判定し，ブックマークが周期的に利用されるページに関する調査を行っている[11].

3. 提案するランキング機能に用いる素性

本節ではランキングに用いるクエリ独立の素性として，変動の激しさ，周期性，増加/減少傾向の 3 つの素性について検討する．

3.1. 変動の激しさ

動向情報は外的要因の影響で激しい変動をすることがある．顕著な例として，2011 年 3 月の東日本大震災の前後で激しい変動をした動向情報は多く，震災に関連する動向情報として多くのユーザに有益である事が期待できる．動向情報の特徴的変動から関連アイテムや期間を検索するコンテクスト検索エンジンにとって，激しい変動を持つ動向情報の重要性は高いと考える．

本稿では，激しい変動とはデータ値が短期間に大きく変動することと定義する．激しい変動を行う期間と変動の大きさは重要な要素であると考えられる．激しい変動を行う期間を検出するために，コンテクスト検索エンジンで指定可能な変動タイプである急上昇と急降下を利用する．急上昇/急降下は，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として定義されている[1]. これを利用して, 急上昇と急降下が発生した期間を動向情報が短期間に大きな変動を行う期間と判断する.

変動の大きさに関して, 動向情報ごとに単位や平均値が異なるため, 固定的な閾値で判断することは現実的ではない. そこで, データ内において変動の占める割合を以下の式で定義する.

$$Intensity = \frac{|V_{start} - V_{end}|}{V_{max} - V_{min}} \qquad (1)$$

ここで, V_{start}, V_{end} は抽出された期間の開始時, 終了時のデータ値をそれぞれ示す. $V_{\max}$, $V_{\min}$ はその動向情報における最大値, 最小値である. この値を動向情報の変動の激しさの素性として扱い, 動向情報ごとに付与する. 複数の激しい変動がある場合は, その動向情報における最大の値を用いる.

3.2. 周期性

野菜の価格や自転車の販売量, Amazon の Google 検索数など周期性を持つ動向情報がある一方で, 乾電池の価格や内閣支持率など周期性の見られない動向情報がある. 周期性を持つ情報の特徴は, 気候,あるいは入学やクリスマスなどといった定期的な行事など周期性を持って発生する要因に影響を受けている点である. これらの要因について関心がある場合, 周期性を持つ動向情報は価値ある情報と考える.

本稿では, 自己相関を用いて動向情報の周期性を判定する. コンテクスト検索エンジンでは動向情報の粒度が月単位であるものがほとんどである[1]ため, 1 年周期の動向情報を主な対象とする. そのため, 動向情報のデータ点が 12 点以下であるか, データに欠損がある動向情報は除外した. 自己相関の計算とピーク値の推定には Matlab の xcorr 関数と findpeaks 関数を用いた. 推定したピークの中にはノイズによるものが含まれるため, 文献[12]を参考に,自己相関が 0.3 以上のピークに限定し, 時差なし以外に 1 つ以上主要な周期を含むものを周期性があると判断する. 周期性がある場合は 1, ない場合は 0 と設定しランキングの素性とする.

例として, 日本梨の価格の時系列データと自己相関のグラフを図 1, 2 に示す. また, ノートブックの価格の時系列データと自己相関のグラフを図 3, 4 に示す. 図 1 から日本梨の価格は毎年 6 月に周期的にピークを迎えていることがわかる. この時, 図 2 においても閾値を超えるピークが複数存在することがわかる. しかし, 図 3 のグラフではその様な傾向は読み取れず, 図 4 においても, 閾値を超えるピークが存在しないことがわかる.

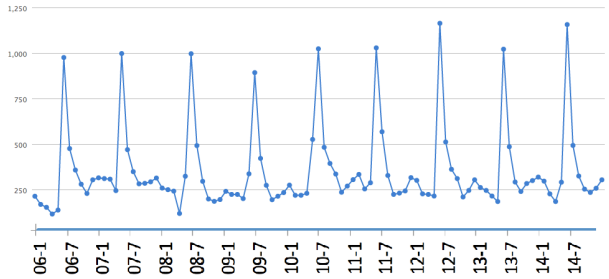

図 1. 日本梨の価格の時系列データ

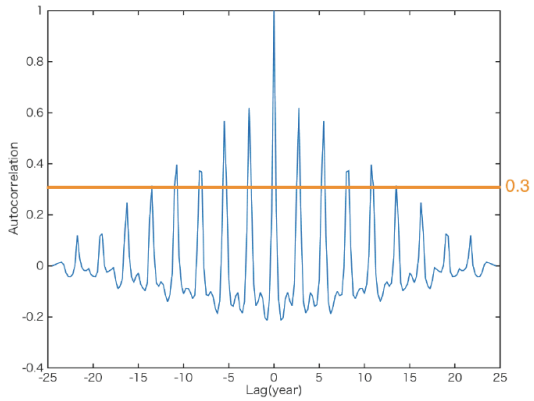

図 2. 日本梨の価格の自己相関

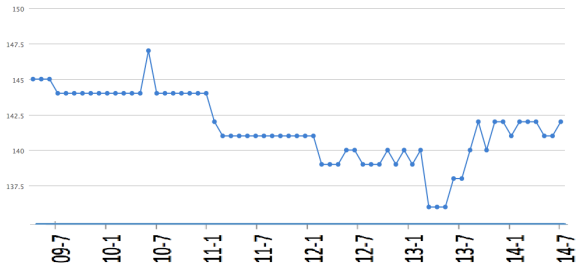

図 3. ノートブックの価格の時系列データ

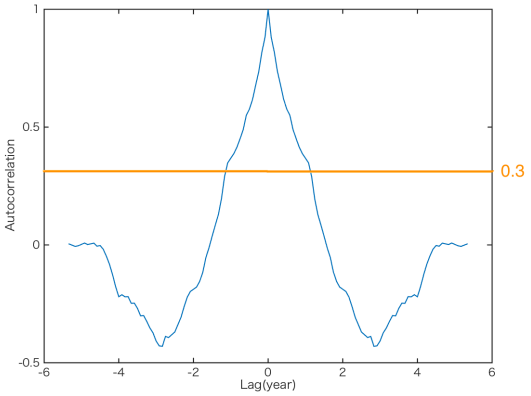

図 4. ノートブックの価格の自己相関

3.3. 増加・減少傾向

動向情報には，一時的な激しい変動や周期的な変動以外にも，右肩上がり/下がりといった傾向を示す変動がある．このような動向情報には，突発的な要因や周期的な要因とは異なる，長期的に普遍な要因の影響があると考える．変動の激しさや周期性とは異なる関連が期待できるため，検索者の目的によっては重要であると考える．

増加・減少傾向を判定するために，式(2)で定義されるピアソンの積率相関係数を用いる．

$$r(y) = \frac{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})(y(n)-\bar{y})}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n)-\bar{x})^2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y(n)-\bar{y})^2}} \quad (2)$$

ここで，$y(n)$ は N 点からなる時系列データの n 番目の点であり，$x(n) = n\text{-}1$ とする．

例えば，何らかの要因によりあるアイテムの生産量が減少し，価格が高騰するなど，同じアイテムに関する動向情報が，同じ要因により反対の変動を示すことがある．このため，ランキングの素性とする場合，増加・減少傾向を区別する必要はないと考え，得られた相関係数の絶対値をランキングの素性とする．

4. ランキング機能の実装と考察

本節では，3 節で述べた変動に関する素性のうち，変動の激しさを用いたランキング機能を実装した結果を示し，その性質について考察する．

2008 年 1 月から 2011 年 12 月に最大値を示した動向情報を検索した結果を図 5 に示す．図 5 から，たちあがれ日本の政党支持率や東京電力の検索数(GoogleTrends) などが上位で検索されていることがわかる．これらの動向情報は 2011 年 3 月に発生した東日本大震災と関連が深いことから，それらがランキングの上位となっていることは，妥当であると考える．

次に，レギュラーガソリンの動向情報が最大値をとる時期に同様に最大値をとる動向情報を検索した結果を図 6 に示す．また，レギュラーガソリンの動向情報の一つである卸価格の動向を図 7 に示す．図 7 からレギュラーガソリンの卸価格は 2008 年 8 月をピークに急激に減少していることがわかる．また，図 6 からレギュラーガソリンと同時期に最大値を示した動向情報のランキング上位には，ストック(生花) や西洋梨など同時期に価格に大きな変動がある動向情報が検索されている．原油価格の高騰が，花しや果物の栽培用の燃料費の増加につながり，販売価格に影響を与えることが指摘されており[13]，これらが上位にランキングされていることは妥当であると考える．

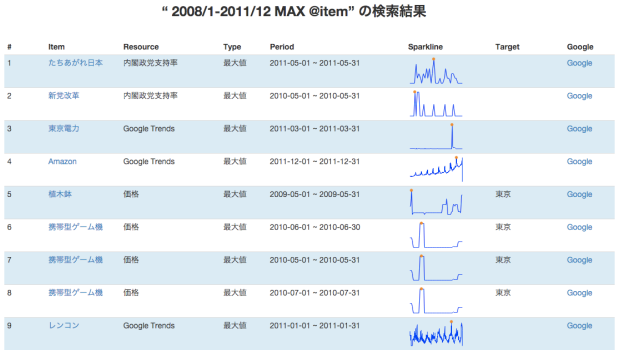

図 5. 2008 年 1 月から 2011 年 11 月に最大値を示した動向情報の検索結果

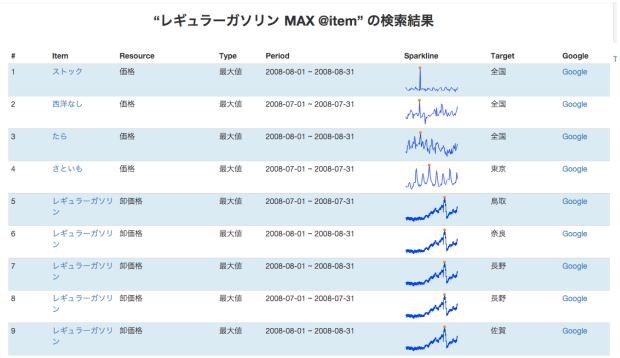

図 6. レギュラーガソリンの動向情報が最大値を示した期間に同様に最大値を示した動向情報の検索結果

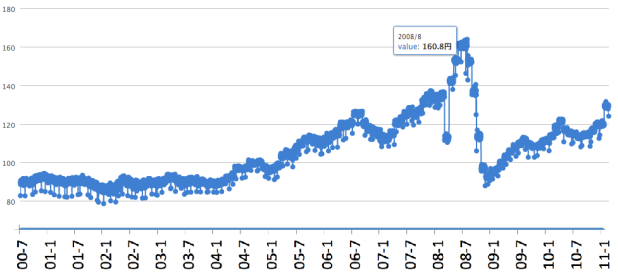

図 7. レギュラーガソリンの卸価格の時系列データ

5. おわりに

本稿では，コンテクスト検索エンジンにおいてより効率的な検索を実現するために，ランキング機能の導入を検討した．変動の激しさ，周期性，増加・減少傾向の 3 つの素性について，期待される役割や計

算方法を示した他, 変動の激しさを素性に用いたランキング機能を実装し, 検索結果の例を示した. 今後は, 本稿で検討したランキング素性を用いてランキング学習を行うために, クリックログやブックマークなどのユーザフィードバックを利用した訓練データの作成を予定している.

参考文献

[1] 高間, 加藤, 桑折, 石川: 動向に関する問いを対象とした検索エンジンの提案, 人工知能学会論文誌, Vol. 30, No. 1, pp. 138-147 (2015)

[2] 高間, Zhu, 桑折, 山口, 瀧口: 動向に関する問いに答える検索エンジンの開発, 人工知能学会第 10 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. 9-15 (2015)

[3] 菱沼, 山口: 検索エンジン最適化の有効性に関する考察, 東京工科大学研究報告, pp. 3-13 (2008)

[4] H. LI: A Short Introduction to Learning to Rank, IEICE Transactions on Information and Systems, Vol. E94-D, No. 10, pp. 1854-1862 (2011)

[5] 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)

[6] 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)

[7] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu: Okapi at TREC-3, 3rd Text REtrieval Conference, pp. 109-126 (1994)

[8] P. Matthew: Determining Relevance: How Similarity Is Scored, https://moz.com/blog/determining-relevance-how-similarity-is-scored

[9] S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, 7th International Conference World-Wide Web 7, pp. 107-117 (1998)

[10] M. Tober, L. Hennig, D. Furch: SEO Ranking Factors and Rank Correlation 2014 –Google U.S.-, searchmetrics Whitepaper (2015)

[11] 綱元, 亀井, 藤田: 周波数分析を利用した周期的にブックマークされる web ページの特定, 第 74 回情報処理学会全国大会講演論文集, Vol. 2012, No. 1, pp. 719-720 (2012)

[12] MathWorks: 自己相関を使用した周期性の検出, http://jp.mathworks.com/help/signal/ug/find-periodicity-using-autocorrelation.html

[13] 大山, 古在: 園芸用施設の暖房費および CO2 排出量削減(1), 農業および園芸, Vol. 83, No. 11, pp. 1157-1163 (2008)