1L3-1

コンテクスト検索エンジンを対象としたランキング機能の提案

Proposal of Ranking Function Designed for Context Search Engine

手塚 拓哉 *1　山口 晃一 *1　高間 康史 *1
Takuya Tezuka　Koichi Yamaguchi　Yasufumi Takama

*1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

This paper proposes a ranking function for context search engine, which is a search engine designed for answering trend-related queries. Query-independent features specific to context search engines, such as intensity and periodicity of temporal change, and Increasing / decreasing trend are proposed. A method for generating listwise training data for rank learning is also proposed. This paper discusses the effectiveness of the proposed ranking function through preliminary experiment.

1. はじめに

本稿では，コンテクスト検索エンジン [高間 15] のランキング機能を提案する．コンテクスト検索エンジンは，動向に対する問いに答えることを目的として開発されており，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でコンテクスト検索エンジンの有効性が示されている．しかし，現在のシステムでは検索結果が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．そこで本稿では，より効率的な検索を実現するために，ランキング学習を用いたランキング機能の導入を提案する．

提案手法では，既存の Web 検索エンジンと同様のランキング学習 [Li 11] を用いた枠組みによりランキング機能を導入する．しかし，既存 Web 検索エンジンとは検索対象などが異なるため，コンテクスト検索エンジンの検索タスク・対象データに適した素性を新たに作成する必要がある．本稿では，クエリ独立の素性として変動の激しさ，周期性，増加減少傾向を提案する．また，学習に必要なリスト型の訓練データを人手により作成することは困難であるため，検索行動から自然に作成することができるクリックログとブックマークから，リスト型の訓練データを作成する手法も提案する．

提案する素性およびランキング機能をコンテクスト検索エンジンに実装し，予備実験を行った．検索意図を規定して検索を行ったログから訓練データを作成し，ランキング学習を行った結果から，提案手法の有効性について考察する．

2. コンテクスト検索エンジン

既存 Web 検索エンジンの基本検索機能と，ユーザの情報要求の乖離が指摘されており，その解決策の一つとしてコンテクスト検索エンジンが提案されている [高間 15]．コンテクスト検索エンジンは，タスクを「動向に関する問い」に限定することにより，幅広いドメインで利用可能であり，かつ高度な検索機能を提供する次世代検索エンジンである．検索の際には，アイテム名，期間，変動タイプを組み合わせてクエリを入力する．動向情報は月を基本単位としており，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 [高間 15]．

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，ytakama@tmu.ac.jp

3. 提案するランキング機能

3.1 ランキング機能で用いる素性

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 つの素性を提案する．これらの素性は動向情報ごとに求められる．

3.1.1 変動の激しさ

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と最小値の差の 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 [高間 15]．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．

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

ここで，V_{start}，V_{end} は急上昇あるいは急降下として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．

3.1.2 周期性

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．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 [MathWorks] 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．

3.1.3 増加/減少傾向

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．

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

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，$x(n) = n - 1$ 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．

3.2 暗黙的フードバックを利用した訓練データの作成

クリックログから 2 対のデータに対する適合評価を持った訓練データを作成する既存手法 [Joachims 02] を応用し，クリックログとブックマークからリスト型の訓練データを作成する手法を提案する．クリックログとブックマークにより評価されたデータは，評価されてないデータよりもユーザの興味が高いとの考えに基づき，元のランキングから評価されたデータとされていないデータの順位を入れ替える．入れ替え後の順位をリスト型の訓練データとして利用する．クリックログとブックマークでは，ブックマークの方がユーザの興味が強いと考えられるため，ブックマークが優位に働くように，クリックログ，ブックマークの順で入れ替え処理を行う．

4. 予備実験

一過性の大きな変動があるデータの発見を目的に，コンテクスト検索エンジンを利用して検索を行った．この検索による，123 クエリ 1190 データのクリックログと，73 クエリ 295 データのブックマークを用いて，123 クエリに対するデータセットを作成した．そのうち，101 クエリを訓練データ，22 クエリをテストデータとして利用した．素性には，3.1 節で提案した素性に加え，データ数，分散，標準偏差，平均値，コンテクスト検索エンジンで利用可能な各特徴的変動（6 種類）の数の合計 13 個を利用した．3.1 節で述べた以外の素性に関しては，全データの最大値と最小値の差との比を用いて正規化した．ランキング学習には，確率的損失関数と多層ニューラルネットワークを利用した Coordinate Ascent[Burges 05] を用いた．評価指標には NDCG を用いて，上位 10 件の評価を行った．

単一の素性を用いた場合と比較し，ランキング学習の効果について考察する．2011 年 1 月から 12 月までに急上昇したアイテムの検索を行った際の検索結果ページを図 1,2 に示す．図 1 は提案手法によるランキング，図 2 は変動の激しさに関するスコアのみに基づくランキングの上位 10 件である．

提案手法によるランキング結果には，東京電力やミネラルウォーター，原発，JR 東日本などの，一過性の大きな変動が見られる動向情報（図中赤枠）が上位に表示されていることがわかる．これらの多くは，2011 年 3 月に発生した東日本大震災に関連する動向情報であり，その影響を大きく受けているためと考えられる．変動の激しさによるランキング結果にも，震災関連の動向情報はいくつかみられるが，一過性の変化が見られない動向情報（図中青枠）も多く含まれている．これは，単一素性のスコアのみを用いた場合は，ノイズデータやデータ点の少ない動向情報など，データの性質の影響を大きく受けるためと考える．

提案手法による NDCG は，訓練データが 0.4657，テストデータが 0.4195 であり，単一素性による NDCG は，0.3971 であり，提案手法が上回る結果が得られた．このことからも，提案手法によるランキング結果は妥当であると考える．

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，ランキング学習を用いたランキング機能を提案した．変動の激しさ，周期性，増加/減少傾向の 3 種類の素性と，クリックログとブックマークを用いた訓練データの作成手法を提案した．今後，期間や変動タイプなどのクエリ依存の素性を追加することで，よりユーザの検索意図を反映したランキング機能を実現する予定である．

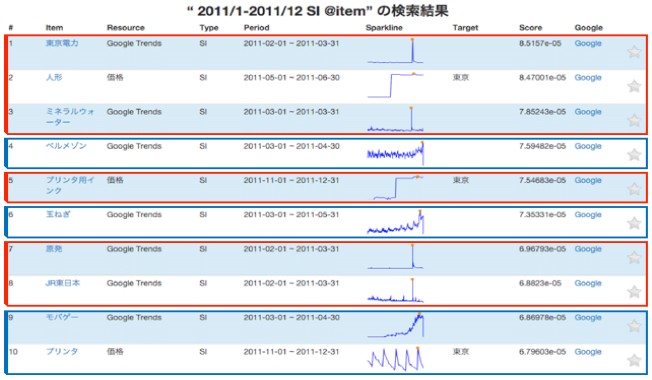

図 1: 提案手法によるランキング結果

図 2: 変動の激しさによるランキング結果

参考文献

[高間 15] 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

[Li 11] H. Li: A Short Introduction to Learning to Rank, IEICE Transactions on Information and Systems, Vol. E94-D, No. 10, pp. 1854-1862, 2011

[MathWorks] 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` (2016/1/30 現在)

[Joachims 02] T. Joachims: Optimizing Search Engines using Clickthrough Data, KDD'02, pp. 133-142, 2002

[Burges 05] C.J.C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender: Learning to rank using gradient descent, ICML2005, pp. 89-96, 2005