# 時間変化を考慮した検索クエリのクラスタリングにおけるクエリ間類似度の検討

木田 巧† 豊田 正史‡ 喜連川 優‡

東京大学大学院情報理工学系研究科† 東京大学生産技術研究所‡

## 1. はじめに

多くのインターネットユーザーが検索エンジンでキーワード検索を行う今日，検索エンジンに蓄積された検索クエリのログが持つ価値は年々大きくなっている．クエリログから有用な情報を得るための手段の１つとしてクエリのクラスタリングが挙げられる．クエリのクラスタリングを通して，頻出クエリや類似クエリのグループを抽出することができ，その結果をユーザーへのクエリ推薦やトレンドの分析に活用することができる．また，クエリをクラスタリングする際のクエリ間の類似度の決め方は，クリックスルーログや類語関係，検索頻度の時間的傾向など様々な視点の類似度が提案されている[1][2]．しかし，検索クエリのクラスタが時期によって動的に変化しうることを考慮していない．そこで時期による検索ニーズの変動も考慮して連続的に検索クエリをクラスタリングすることはトレンドを考慮したクエリ推薦に役立つと考えられる．本研究では時間的な一貫性を保ったクラスタリングを実現する手法に注目し，その手法によって検索クエリを分類するためのクエリ間類似度やパラメータの設定について考察を行う．これに先駆けた我々の研究[3]の中で検索結果のURLのJaccard係数に基づいたクラスタリングにおいて時系列を考慮したクラスタリング手法の有効性を確認したが，本稿ではそれに加えBlog記事の出現頻度の時系列変化もｓ考慮に入れたクラスタリングについて考察を行う．

実験には，Yahoo! Japan[4]から提供された2007 年の検索語データを用いた．このデータには各月の検索頻度上位の10,000 語が含まれており，この中からクラスタリングの対象とする検索語を選択する．また，クエリの類似度の時系列変化を得るには，過去の時点での検索結果が必要となるが，現在の検索エンジンから得ることは難しいため，我々が2006年より収集しているBlog アーカイブを利用し，各時間帯においてクエリを含む記事の集合を取得して類似度を算出する．

## 2. クエリの分類手法

本研究ではクエリログのクラスタリングを行うにあたって，各時期における対象のクエリに対し，その時期におけるクエリ間の類似度を定義する．そして検索クエリをノード，クエリ間の類似度を重みに持つエッジとみなした，構造が時期に応じて変化するようなグラフをクラスタリングすることで時系列を考慮したクエリのクラスタリングを実現する．クエリの類似度は，特定の幅の時間区間において存在する記事を用いて算出し，区間の幅を保ってスライドさせることで時間変化を表す．以降では時間を，$T=[t_1,t_2,\cdots,t_n]$で表し，対応する時間区間を$I=\{i_1,i_2,\cdots,i_n\}=\{[t_1,t_1+k],[t_2,t_2+k],\cdots[t_n,t_n+k]\}$と表す．ただしk は区間の幅である．また，時間の単位は1ヵ月程度を想定している．

A Study on Similarity Measures for Temporal Clustering of Search Queries
†Takumi KIDA
‡ Masashi TOYODA, Masaru KITSUREGAWA
†Graduate School of Information Science and Technology, The University of Tokyo
‡Institute of Industrial Science, The University of Tokyo

### 2. 1 類似度

本研究では，以下の２種類の類似度を考慮した．

- 検索クエリ間の類似度に検索クエリに対応するBlog 記事集合から取得したURLの共起に基づく類似度
- 検索クエリに対応するBlog記事の出現頻度の時系列変化に基づく類似度

まずURLの共起に基づく類似度は，URLある時系列t におけるクエリ間の類似度は，$t$ に対応する区間$i$で取得されたURL 集合同士のJaccard 係数で定義した．すなわち検索クエリq における検索結果に含まれるURL の集合を$u(q,i)$とすると，以下のような式で表される．

$$\mathrm{Similarity}_{url}(q1,q2,t)=\frac{u(q1,i)\cap u(q2,i)}{u(q1,i)\cup u(q2,i)}$$

次に，Blog記事の出現頻度に基づく類似度は，区間$i$に含まれる出現頻度ベクトル同士のコサイン類似度を用いた．検索クエリq に対する記事の出現頻度のベクトルを$\mathbf{q_i}$とすると以下の式で表される．

$$\mathrm{Similarity}_{freq}(q1,q2,t)=\frac{\mathbf{q1_i}\cdot\mathbf{q2_i}}{|\mathbf{q1_i}||\mathbf{q2_i}|}$$

それぞれ類似度の分布を考慮し，類似度が閾値以上だったクエリ間の類似度を再度0 から1 へマッピングを行った．

また，最終的な類似度はパラメータβによる線形結合を考え，

$$\mathrm{Similarity}=\beta\cdot\mathrm{similarity}_{url}+(1-\beta)\cdot\mathrm{similarity}_{freq}$$

によって与え，Similarityを重みとしたエッジをノード間に張る．

### 2. 2 クラスタリング

時系列を考慮したクラスタリングにはLinら[5]が提案したFacetNet を用いた．FacetNet は時間とともに構造が変化するグラフをソフトクラスタリングするアルゴリズムで，1時点前のクラスタ構造を維持しながらソフトクラスタリングを行うことでノイズデータの混ざりくい，時間的に一貫性をもったクラスタ構造の抽出が期待される．また，検索クエリのような多義語が想定されるものを確率的に複数のクラスタに分類することが有効であると考え，本研究でFacetNetをクラスタリングアルゴリズムに用いてクエリの分類を行った．

## 3. 評価実験

クエリログから得られた検索単語に対して，Blog 記事集合から取得した検索語を含む共通のURL を類似度としたクラスタリングを行い，人手で作成した正解データに基づく評価を行った．

### 3. 1 データセット

クラスタリングの対象とする検索語データは特定領域研究「情報爆発時代に向けた新しいIT 基盤技術の研究」（情報爆発プロジェクト）で，ヤフー株式会社から国立情報学研究所に提供された2007 年1 月から12 月までの各月の検索回数上位1 万語のリストからユニークな検索語18730 語を対象に，以下の３つの手順によって1057 語を選出した．ユニ

ークな検索語18730 語から検索頻度が大きい上位5000 語を選出した．検索頻度の大きさは検索回数の全時間における合計である．その中で変化が一定以上ある検索語1515 語を選んだ．変化の大きさはクエリの検索頻度の前月比の変化率を基準にし，ある時期$t$ における検索頻度を$q(t)$とした時に

$$\left|\frac{q(t)-q(t-1)}{\min(q(t),q(t-1))}\right| > \theta$$

を満たすクエリを選出した．$\theta$は変化の閾値で，$\theta = 1.0$としてクエリの選出を行った．さらに，抽出された1515 語の中で，1 月から12 月までのリストの中に少なくとも2 回以上出現している1057 語を選出した．

次に選んだ1057語のクエリに対してBlog 記事データベースを検索し，検索語をタイトルまたは本文に含む記事を取得した．Blog 記事集合は2006 年2 月から約100 万のRSS, ATOMフィードを毎日収集したもので，その中から検索クエリに対応する記事のURL を月ごとに集計したデータを用いた．また，各クエリに対応する記事の出現頻度は日単位で集計した．

### 3. 2 評価

クラスタリング結果の評価は意味的な類似性，時間的な類似性の両方を考慮して人手で正解付けした正解データをもとに行った．1つの正解データは（正解ラベル，単語リスト，出現時期）から構成され，出現時期は各月にそのラベルに対応する単語が出現しているかどうかを表す{0,1}からなる12次元のベクトルで表現した．まずクラスタリング対象にした1057 語のうち503 語を検索語間の類似性を考慮して132 のラベルに人手で分類した．ラベル付けは同じ意味を指すと考えられる単語（totobig，サッカーくじ）や，関連の深いと考えられる単語（スキー，スノーボード，積雪情報）などを1つのラベルとして人手の分類を行い，検索頻度に基づいて出現時期を決定した．評価は，各クラスタの純度にクラスタの大きさをかけた重みつき平均を用いて評価した．純度の定義は以下の通りである．

$$\mathrm{Purity}(L,C) = \frac{1}{N}\sum_{c\in C} max_j |C_c \cap L_j|$$

ただしLはラベルの集合，Cはクラスタの集合，N はラベル付けされた単語の総数である．FacetNetに与えるパラメータとして，Temporal Costの割合を決めるパラメータは，$\alpha = 0.9$とし，クラスタ数は150とした．各類似度をそれぞれ単体で用いた場合，と類似度の重みを決めるパラメータβを変えた場合で評価を行った．

## 4．結果

表1 は各類似度を単体で使用してクラスタリングを行った時の評価値である．URL の共起による分類のほうが純度の高いクラスタを抽出できることがわかる．

評価実験において，$\beta = 0.90, 0.91, \cdots, 1.0$と0.01刻みで変えながら評価を行った結果を図1に示す．結果からは$\beta = 0.97$の時にもっとも純度が高く，URL の共起のみに基づく場合($\beta = 1.0$)に比べて約 1%程度純度の高いクラスタが抽出できることがわかる．

記事の出現頻度の時系列情報を加えることで，レジャーに関するクラスタがシーズン毎に細かいクラスタに分割された．また，ドラマのタイトルと出演者の組，結婚した芸能人の組などもクラスタとして抽出することができた．これらの新たな組み合わせをクラスタとして分類できたことが精度向上に寄与していると考えられる．

表1 使用した類似度と純度

| 類似度 | 純度 | $\alpha$ |
|---|---|---|
| URL の共起 | 0.562 | 0.9 |
| 記事頻度の時系列 | 0.352 | 1.0 |

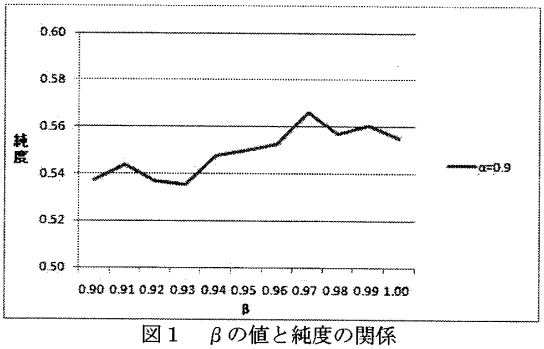


図1　βの値と純度の関係

表2　抽出されるクラスタの例

| 類似度 | 単語リスト |
|---|---|
| URL の共起 | 年末ジャンボ, 年末ジャンボ宝くじ, オータムジャンボ, toto, サマージャンボドリームジャンボ, サマージャンボ宝くじ, … |
| 記事頻度の時系列 | 紅白歌合戦, 年末ジャンボ宝くじ, ガキの使い，年末ジャンボ, 年賀状 画像, 福袋, 忘年会, 年賀状 無料, コミケ, 明治神宮, 年賀, … |

表2は各類似度を使用した際に抽出できたクラスタの典型的な例である．URL の共起を用いた場合では意味的な繋がりの強いクエリが同じクラスタになるのに対して，記事頻度を用いた場合は時期的な繋がりの強いクエリ同士が同じクラスタになる．例えば「年末ジャンボ宝くじ」に注目すると URL の共起に基づく場合は宝くじに関する話題がクラスタになり，記事頻度の時系列の場合は年末年始の話題がクラスタになっている．

## 5．まとめと今後の課題

本研究では時系列を考慮したクラスタリング手法によって検索クエリを分類する際に使用する類似度によるクラスタリング結果の違いについて考察し，Blog 記事 URL の Jaccard 係数と，記事の出現頻度の時系列変化を組み合わせることで URL の共起情報のみによる分類に比べてより精度の高い分類を実現できることを評価実験によって確認した．また，各クラスタ内部において URL の共起情報のみに頼った場合分類できなかったクエリのセットをクラスタとして抽出することが可能であることを確認した．

今後の課題として本研究で用いなかった他の類似度やクラスタリング手法を用いることについて，扱っていきたいと考えている．

**謝辞**　本研究の実験に用いる検索語データを提供していただいたヤフー株式会社に感謝いたします

**参考文献**

[1]J. Wen, J. Nie and H. Zhang: “Clustering user queries of a search engine”, Proceedings of the 10th international conference on World Wide Web, pp. 162-168 (2001).
[2] S. Chien and N. Immorlica: “Semantic similarity between search engine queries using temporal correlation”, Proceedings of the 14th international conference on World Wide Web, pp. 2-11 (2005).
[3] 木田，豊田，喜連川: “トレンドを考慮した検索クエリの分類手法の一検討”, DEIM (2010). (発表予定)
[4] “Yahoo! Japan”, http://www.yahoo.co.jp/.
[5] Y. Lin, Y. Chi, S. Zhu, H. Sundaram and B. Tseng: “Facetnet”: a framework for analyzing communities and their evolutions in dynamic networks”, Proceedings of the 17th international conference on World Wide Web, pp. 685-694 (2008)