トレンドを考慮した検索クエリの分類手法の一検討

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
〒 113-0033 東京都文京区本郷 7-3-1
†† 東京大学生産技術研究所
〒 153–8504 東京都目黒区駒場 4–6–1
E-mail: †{kida,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし 検索エンジンのログはユーザーの興味を示唆する情報として注目されており，クエリ推薦やスペル修正など様々な応用が研究されている．本稿では，時期に応じたクエリ間類似度の変化を考慮した検索クエリのクラスタリング手法を提案する．提案手法では時間と共に構造の変化するグラフの時系列を，時間的な一貫性を考慮してクラスタリングする手法を用いる．各時間における検索語間の類似度は，約 100 万の記事を毎日継続して収集して構築された Blog アーカイブから取得した検索語間に共通する Blog 記事 URL に基づいて計算する．また，Yahoo! Japan から提供された 2007 年のクエリログを用いた実験を行い，人手で正解付けしたデータによる評価を行うことを通して，時間的一貫性を考慮したクラスタリング手法を用いることで，時期に応じたクラスタの時系列が抽出できることを示す．
キーワード クエリログ，クラスタリング，時系列，ブログアーカイブ

A Method for Clustering Search Queries by Changes in Trends

Takumi KIDA†, Masashi TOYODA††, and Masaru KITSUREGAWA††

† Graduate School of Information Science and Technology, University of Tokyo
Hongo, 7–3–1, Bunkyo-ku, Tokyo, 113–0133 Japan
†† Institute of Industrial Science, University of Tokyo
Komaba 4–6–1, Meguro-ku, Tokyo, 153–8504 Japan
E-mail: †{kida,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract Web search query logs are important data for observing users' interest, and used for various applications such as query suggestion and spell correction. In this paper, we propose a method for clustering search queries reflecting changes in similarities over time. Our method is based on an evolutionary approach for clustering a time series of graphs avoiding drastic changes in clustering results. To calculate similarities between queries in each time, we use occurrences of queries in our daily blog archive including about 1 million blogs. We perform experiments on monthly lists of top queries in 2007 provided by Yahoo! Japan, and show that our method can extract relevant clusters over time.
Key words Query Log, Evolutionary Clustering, Time Series, Blog Archive

1. はじめに

多くのユーザーがインターネットを利用するようになった今日，キーワードによる Web 検索が日常化している．コムスコア [1] が発表した統計によると，2009 年 1 月の日本におけるインターネットユーザー一人当たりの検索回数は平均 100 回である．このことから，多くの検索クエリが日々検索エンジンに入力されていることが予想され，検索エンジンに蓄積された検索クエリのログはユーザーの興味を反映するデータとして重要性を増していると考えられる．たとえば頻繁に検索されるテレビドラマのタイトルや人気の芸能人の名前などはその時の流行を知る判断材料になる．また，季節によっても頻繁に検索されるクエリは異なり，春であれば「桜」,「花見」,「花粉症」などのクエリが多く検索され，年末年始には「クリスマス」,「年賀状」といったクエリが頻繁に検索されることが予想される．

検索エンジンに蓄積したクエリログを利用した研究分野の 1 つとしてクエリのクラスタリングがある．クエリをクラスタリングしておくことで，頻出クエリや，関連クエリの発見が可能

である [2]．例えば検索エンジンのキーワード補完機能のように，関連するクエリを推薦することでユーザーの検索支援に役立てることができる．また，それを目的としたクエリ分類や関連クエリ発見の研究が数多く行われている [3], [4]．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など様々な視点の類似度が提案されているが，類似度の時間変化に着目した例は少ない．時期による検索ニーズの変動も考慮して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．例えば,「宿泊予約」などの旅行・レジャーに関するクエリはシーズンによって関連するクエリは異なることが予想される．本研究では単にその時々のデータを使用してクラスタリングを行うよりも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適切なパラメータ設定について考察を行う．

実験には，Yahoo! Japan から提供された 2007 年のクエリデータ [5] 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クエリを選択する．また，クエリの類似度の時系列変化を得るために，過去の時点での検索結果が必要となるが，これは現在の検索エンジンでは得ることができない．このため，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．

以降の章では，2 章で我々が行ったクエリの分類手法，3 章で評価実験，4 章で評価実験の結果，5 章で関連研究，6 章でまとめと今後の課題について述べる．

2. クエリの分類手法

本研究では検索クエリ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グラフをクラスタリングすることで時間変化を考慮したクエリのクラスタリングを実現する．グラフの時系列の作成では，各時期における検索クエリ間の類似度を計算し，類似度が閾値を超えた場合にエッジが存在すると考える．クエリ間の類似度は，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，$T = \{t_1, t_2, t_3, ..., t_n\}$ で表し，対応する時間区間を $I = \{i_1, i_2, i_3, ..., i_n\} = \{[t_1, t_1 + W], [t_2, t_2 + W], ...\}$ と表す．ただし W はあらかじめ決められた区間の幅である．また，後述の実験では時間の単位は 1ヵ月を想定している．

2.1 類 似 度

類似度には，検索クエリ間の類似度に検索クエリに対応する Blog アーカイブから取得した URL の集合間の Jaccard 係数を使用した．ある時間 t におけるクエリ間の類似度は時間 t に対応する区間 i で取得された URL 集合間の Jaccard 係数で定義した．すなわち検索クエリ q の区間 i 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，以下のように表される．

$$similarity(q1, q2, t) = \frac{|u(q1, i) \cap u(q2, i)|}{|u(q1, i) \cup u(q2, i)|}$$

また，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の対数を取り再度 0 から 1 へ正規化することでエッジの重みを決定した．

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 らが提案した FacetNet [6] を用いた．FacetNet は時間とともに構造が変化するグラフの時系列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．このアプローチを取ることで，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，本研究では FacetNet をクラスタリングアルゴリズムに用いたクエリの分類を行った．

この手法では Snapshot Cost と Temporal Cost という二種類のコスト関数を定義する．Snapshot Cost は得られたクラスタ構造がどの程度その時点のスナップショットを反映しているかを表し，クラスタ構造がスナップショットに忠実であるほどコストが小さくなるような関数を定義する．Temporal Cost はクラスタ構造の時間的な一貫性を表し，クラスタ構造とその 1 時点前のクラスタ構造と近いほどコストが小さくなるような関数を定義する．そして二種類のコスト関数をパラメータ α を用いて線形結合した全体のコスト関数を最小にするようなクラスタ構造を求める．

FacetNet では各ノードが確率的な重みづけをもってすべてのクラスタに所属するというモデルによってソフトクラスタリングを行う．まず，それぞれの時点 t における n ノード，m クラスタからなるグラフが与えられており，グラフ中の各ノード間には類似度を重みに持つエッジが行列 $W \in \mathcal{R}_{+}^{n \times n}$ として与えられているものと仮定する．

次に各ノードがすべてのクラスタに確率的な重みづけをもって参加しているというモデルをノード・クラスタ間の 2 部グラフとして表現する．ここで i 番目のノードが k 番目のクラスタに属する確率を $p_{k \to i}$ と表す．ただし $\sum_k p_{k \to i} = 1$ である．また，k 番目のクラスタが各ノードに対して影響している事前確率を p_k とおく．この時，同じクラスタに所属しているノード間の類似度は高くなるという考えに基づき，ノード i, j の間の類似度 w_{ij} をクラスタへの所属確率 $p_{k \to i}, p_{k \to j}$，およびクラスタに関する確率変数 p_k を用いて以下の式で推定する．

$$w_{ij} \approx \sum_k p_k \cdot p_{k \to i} \cdot p_{k \to j}$$

このモデルを $x_{ik} = p_{k \to i}$ であるような行列 $X \in \mathcal{R}^{n \times m}$ と $\lambda_k = \lambda_{kk} = p_k$ であるような $m \times m$ の対角行列 Λ で表すと

$$W \approx X \Lambda X^T$$

となる．行列 $X\Lambda$ の行を正規化したものが求めるクラスタ構造に相当しており，以下で述べるコスト関数を最小にするような

X, Λ を求めることでクラスタリングを行う．

このモデルにおける二つのコスト関数について述べる．Snapshot Cost(CS) は，実際の類似度行列 W とクラスタ構造から類推される W の構造である $X\Lambda X^T$ の間の KL-Divergence を Snapshot Cost とする．

$$CS = D(W||X\Lambda X^T)$$

Temporal Cost(CT) は抽出したいクラスタ構造に対応する $X\Lambda$ とその一時点前の構造との間の KL-Divergence を Temporal Cost とする．

$$CT = D(Y||X\Lambda)$$

ただし Y は 1 時点前におけるクラスタ構造で，$Y = X_{t-1}\Lambda_{t-1}$ である．

また，KL-Divergence はカルバック・ライブラー距離で，

$$D(A||B) = \sum_{i,j}(a_{ij}\log\frac{a_{ij}}{b_{ij}})$$

という式で定義される．
全体のコスト関数は CS, CT をパラメータ α で線形結合したもので，以下のように決定される．

$$Cost = \alpha \cdot CS + (1-\alpha)\cdot CT$$

α はユーザが決めるパラメータ $(0 \leqq \alpha \leqq 1)$ である．α が 1 に近い程 Snapshot Cost が重視され，その時点のクエリ間の類似関係を反映したクラスタが得られることが期待される．逆に α が 0 に近づくと Temporal Cost が重視され，時間的な一貫性を反映したクラスタが得られることが期待される．FacetNet ではラグランジュの未定乗数法に基づく繰り返し計算による最適化で，コスト関数を局所的に最小にするクラスタ構造を求める．

3. 評価実験

提案手法によるクラスタリングを評価するために，クエリログから得られた検索語に対して，Blog アーカイブ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う実験を行った．

3.1 データセット

- クエリログ

クラスタリングの対象とするクエリ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 [5] からユニークなクエリ 18730 語を対象に，以下の 3 つの手順によって 1048 語を選出した．

（1） ユニークなクエリ 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．

（2） その中で変化が一定以上あるクエリ 1515 語を選んだ．ここで，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

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

という式で定義した．θ は変化の閾値である．ここでは $\theta = 1.0$ としてクエリの選出を行った．

（3） (2) のステップで選出した 1515 語の中から，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クエリ 1048 語を選出した

- Blog アーカイブ

Blog アーカイブは 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をタイトルまたは本文に含む記事のうち，2007 年に書かれたものを対象に URL を月ごとに集計したデータを用いた．

3.2 評　　価

クラスタリング結果の評価は意味的な類似性，時間的な傾向の両方を考慮して人手で正解付けを行った正解データを用いて行った．1 つの正解データは（ラベル ID，検索語リスト，出現時期）から構成される．出現時期は各月にそのラベルに対応するクエリが出現しているかどうかを表す 12 次元のベクトルで表現し，出現している時に 1, 出現していない時に 0 を取るものとする．各時間における正解データは，ラベルのうち対応する時間区間中にクエリが出現しているものをその時間における正解とみなした．すなわち，出現時期ベクトルの要素のうち時間区間に少なくとも 1 箇所の 1 を含むものを正解とした．

3.2.1 正解データの作成

まずクラスタリング対象にした 1048 語のうち 453 語を検索語間の類似性を考慮して 158 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クエリ（例えば totobig とサッカーくじ) や，関連の深いと考えられるクエリ（例えばスキー，スノーボード，積雪情報）などを 1 つのラベルとして人手による分類をおこなった．ただし，ここでは 1 つのクエリが複数のラベルに重複して分類されることを許しており，n 種類のラベルが重複して付けられたクエリは $\frac{1}{n}$ の重み付けをして評価した．表 1 は用いた正解データの一例である．

3.2.2 評 価 基 準

評価はラベル付けをしたクエリについて，クラスタへの所属確率，ラベルの重複による重みを考慮した適合率，再現率，F 値の三種類について評価を行った．

ラベル付きの各クエリに対する要素が，そのクエリがクラスタ C_i に含まれる確率であるようなベクトルを $\boldsymbol{c_i}$ とおく．これは FacetNet で計算された $X\Lambda$ を行を正規化した行列の第 i 列からラベルの付いていないクエリに相当する要素を除いたものである．また，ラベル L_j に含まれるクエリに対応する要素がそのクエリの重み，それ以外が 0 であるようなベクトルを $\boldsymbol{l_j}$ とする．この時の各評価値の定義は以下の通りである．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クエリの総数である．

- 適合率

表 1　正解データの例

ラベル ID	検索語リスト	出現時期
1	積雪情報, 苗場, スキー, スキー場, 野沢温泉, スノーボード	[1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1]
2	レイトン教授, ナゾトキ, レイトン	[0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1]
3	toto big,totobig, サッカーくじ,toto, トト	[0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1]
4	脳内メーカー, 脳内メーカ, 脳内メイカー, 脳内	[0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0]

クラスタ C_i, ラベル L_j に対する適合率 $precision(C_i, L_j)$ は以下の式で表される．

$$precision(C_i, L_j) = \frac{\boldsymbol{c_i} \cdot \boldsymbol{l_j}}{|\boldsymbol{c_i}|}$$

また，全体の適合率 $Precision(C, L)$ は各クラスタ C_i について最も適合率が高いラベル L_j と重み付き平均を取り，以下のように算出した．

$$Precision(C, L) = \frac{1}{N} \sum_{C_i \in C} |\boldsymbol{c_i}| \cdot \max_j \{precision(C_i, L_j)\}$$

- 再現率

クラスタ C_i, ラベル L_j に対する再現率 $recall(C_i, L_j)$ は以下の式で表される．

$$recall(C_i, L_j) = \frac{\boldsymbol{c_i} \cdot \boldsymbol{l_j}}{|\boldsymbol{l_j}|}$$

全体の再現率 $Recall(C, L)$ は各ラベル L_j に対して最も再現率の高いクラスタ C_i を選び，ラベルサイズによる重み付き平均を取り，以下のように算出した．

$$Recall(C, L) = \frac{1}{N} \sum_{L_j \in L} |\boldsymbol{l_j}| \cdot \max_i \{recall(C_i, L_j)\}$$

- F 値

F 値は適合率・再現率の調和平均で，クラスタ C_i, ラベル L_j に対する F 値 $f(C_i, L_j)$ は以下の式で表される．

$$f(C_i, L_j) = \frac{2 \cdot precision(C_i, L_j) \cdot recall(C_i, L_j)}{precision(C_i, L_j) + recall(C_i, L_j)}$$

また，各ラベル L_j に対して最も F 値の高いクラスタ C_i を選び，ラベルサイズによる重みづけ平均をとり，以下の式で全体の F 値を算出した．

$$F(C, L) = \frac{1}{N} \sum_{L_j \in L} |\boldsymbol{l_j}| \cdot \max_i \{f(C_i, L_j)\}$$

また FacetNet のアルゴリズムの性質上クラスタリングの結果が局所解に収束することを考慮して，評価値は 10 回試行した平均値を採用した．

4. 結　　果

4.1 パラメータによる違い

まずクラスタリングにあたって設定するパラメータ（区間幅, クラスタ数，α）による評価結果の違いについて述べる．図 1 はクラスタ数 k を横軸，評価値（F 値）を縦軸にとったグラフ

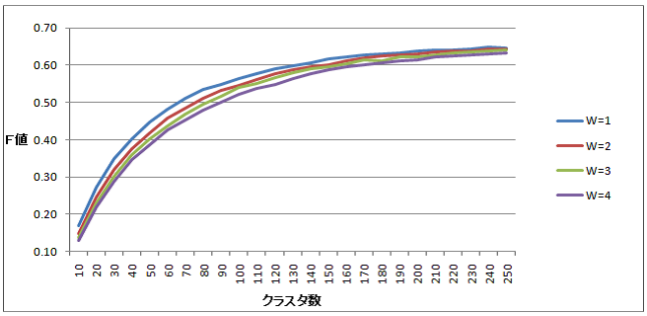

図 1　区間幅, クラスタ数と評価値の関係

である．まず，k と評価値の関係であるが，クラスタ数を増やすに従って評価値は増加するが，増加の度合いは k が小さい時ほど急激で，k が増えるにつれて徐々に緩やかになり，$k = 200$ の周辺では飽和している．この時の各クラスタの平均クエリ数は約 5 クエリであり，評価セットにおける 1 つのラベルあたりの平均クエリ数とほぼ同じ値である．これは設定するクラスタ数によって得られるクラスタの粒度が決まるため，$k = 200$ 程度で評価値が高くなったのは正解データの粒度と得られたクラスタの粒度が近いためと考えられる．したがって k は得たいクラスタの粒度に応じて設定する必要がある．また，時間区間の幅 W との関係は，k が小さい時は W が小さいほうが F 値が高い．しかし，クラスタ数が十分に大きく，クラスタの粒度がこれ以上細かくならない場合は ($k > 150$) では W の値に関わらずほぼ同じとなった．

次に区間の幅を $W = 4$ヵ月，クラスタ数を $k = 200$ クラスタとして FacetNet に与えるパラメーター α を 0.1,0.2,...1.0 と 0.1 刻みで変化させ，α の値が評価値に与える影響について調べた．実験結果を図 2 に示す．適合率，再現率，F 値の関係は全ての評価基準で α の変化に対して似たような傾向を示し，いずれの基準においても最も高い評価値であったのは $\alpha = 0.9$ の時であった．これは Temporal Cost を 10% 考慮してクラスタリングを行った時が最も評価値が高いということを表している．表 2 に $\alpha = 0.9$ の時と， Temporal Cost を考慮しない ($\alpha = 1.0$) 時の評価値の全時間に対する平均と，最も差が開いた時間における差を表した．Temporal Cost を考慮した場合のほうが平均で 1.9% ，最大で 4.3% の評価値の向上を見ることができる．

また，時間経過に伴う評価値の変化を図 3 に表す．図 3 から，

表 2　Temporal Cost の有無による評価値の比較

Temporal Cost なし ($\alpha = 1.0$)(%)	Temporal Cost あり ($\alpha = 0.9$)(%)	平均差 (%)	最大差 (%)
61.8	63.7	1.9	4.3

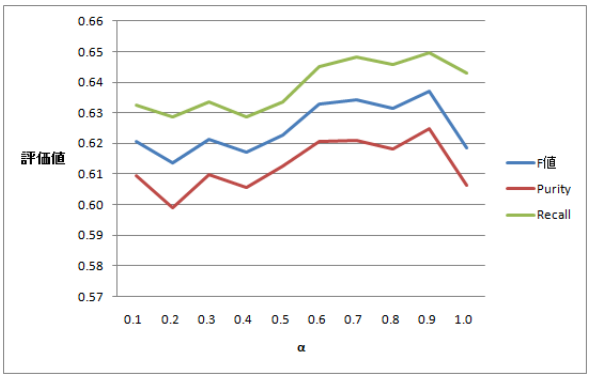

図 2　α と評価値の関係

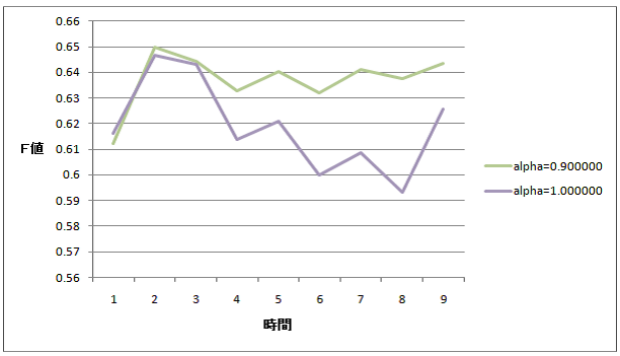

図 3　時間経過と評価値の関係

t_1, t_2 のようなクラスタリング開始時点に近い早期の時点では Temporal Cost の有無による評価値の差はほとんど無いが，時間が経過した後の時点になるほど評価値の差が大きくなり，t_8 の時点で最大 4.3% の差となった．Temporal Cost を考慮した $\alpha = 0.9$ の場合の方が時間経過による対象の変化に対して安定していたと言える．

4.2　抽出されたクラスタの例

次に得られたクラスタの内容について述べる．表 3 はある試行結果において得られたあるクラスタの内容を定点観測した一例である．クラスタに一定以上の確率で所属している検索クエリのリストの時系列変化を表している．ここでは「チョコレート」,「クッキー」を中心としたクエリのクラスタの内容に着目して，$\alpha = 1.0$ と $\alpha = 0.9$ の場合を比較してみると，まず t_2, t_3 では「バレンタイン」,「ホワイトデー」などの語と同じクラスタに所属しているが，t_4, t_5 では「父の日」,「母の日」と同じクラスタになり，t_6, t_7 では「ハロウィン」とクラスタを作っている．このように時期に合わせて同じクラスタに所属するクエリが変化するクラスタの抽出に成功していることがわかる．また，Temporal Cost を考慮している $\alpha = 0.9$ の場合に比べて，Temporal Cost を考慮していない $\alpha = 1.0$ の場合は，例えば t_4 において「明和水産」,「早乙女太一」のようにノイズとして混る関係のないクエリが多い．

5.　関 連 研 究

5.1　クエリの分類

クエリのクラスタリングによるクエリ推薦や，頻出クエリの発見の研究には多くの先行研究があるが，分析するログデータの性質や，クエリ間の類似度の定義は様々である．これらの先行研究の多くは，クエリログデータとして検索クエリとクリックされた URL からなるクリックスルーログやユーザーのセッション情報を含むセッションログを利用しており，本研究で利用したクエリログデータのような検索クエリと月ごとの検索頻度からなるログとは性質が異なる．また，それ以外の場合も，クエリ間類似度をログデータ自体から抽出した特徴や検索エンジンの API を用いて取得した情報に基づいて計算する例はあるが，本研究のようにブログアーカイブから抽出した特徴を用いている例は我々の知る限りはない．

Wen ら [7] は検索クエリ同士に共通する単語とクリックした URL の類似性に着目してクラスタリングを行っている．Cao ら [8] は，ユーザーの入力したクエリとクリックされた URL の関係を 2 部グラフでモデル化しクラスタリングした結果と同一セッション内で入力されたクエリの関連性に基づいてクエリ推薦を行う手法を提案している．Baeza-Yates ら [3] の研究ではクエリとそのクエリに対してクリックされた URL, その URL の先のページに含まれる語をもとに類似度を計算して k-means 法を用いてクラスタリングを行っている．Zhang ら [4] は同一セッション内で発生したクエリに対して定義される Consecutive Similarity と，キーワード間の TFIDF 類似度の代わりに SF(Search Frequency) を用いた SFIDF を基づく Content Similarity の線形結合を類似度としたクラスタリング手法を提案している．甲谷ら [9] はクエリとクリックされた URL 情報を元に，Web サイトに到達するために頻繁に使用されるクエリを発見することでクエリ推薦を行う手法を提案している．また大規模なログデータを人手で分類したセットと機械学習によって分類する手法 [10] や，検索結果のドキュメントに含まれる語によるクロスリファレンスグラフを作成することで，分類のためのカテゴリを自動で獲得する手法 [11] なども提案されている．

一方で，検索クエリに対する URL の類似性やテキスト類似性ではなく検索傾向の時間的な類似性に着目した類似クエリの発見，推薦の研究もなされている．Chien らの手法 [12] では特定の期間におけるクエリの検索頻度の時間平均と標準偏差に基づいた類似度を用いた類似クエリの発見を行っているが，検索頻度の時間変化情報のみで分類を行っている ことや時間の変化に伴う類似クエリの変化については考慮していない点が本研究と異なる．上田ら [13] は，検索頻度の時間的な類似性に着目することで検索結果に共起しにくいクエリのセットを抽出する手法を提案しているが，検索時期に合わせた共起関係を持つクエリのセットを抽出する本研究の手法と異なっている．

検索クエリのクラスタリング以外にもクエリログを利用した研究として，検索クエリとそれによる検索結果から関連語を獲得する研究や，関連語による検索支援による研究事例などが挙げられる [14], [15].

5.2　時間変化を考慮したクラスタリング

時間変化を考慮したクラスタリングを行う際に Snapshot Cost と Temporal Cost を用いてコストの最小化をするという

表 3　抽出されたクラスタの時系列変化の例

時系列	α	検索語リスト
...	...	...
t_2	0.9	ホワイトデー, チョコレート, クッキー, チョコ, バレンタイン, バレンタイン レシピ
	1.0	ホワイトデー, チョコレート, クッキー, チョコ, バレンタイン, バレンタイン レシピ
t_3	0.9	ホワイトデー, チョコレート, クッキー, チョコ, バレンタイン, バレンタイン レシピ
	1.0	父の日, ホワイトデー, 赤福, チョコレート, クッキー, ルタオ, 母の日, チョコ, バレンタイン, バレンタイン レシピ
t_4	0.9	チョコレート, クッキー, 母の日, チョコ, バレンタイン
	1.0	父の日, 福袋, クリスマスプレゼント, ホワイトデー, 明和水産, チョコレート, お歳暮, 早乙女太一, 母の日, バレンタイン,...
t_5	0.9	父の日, 赤福, チョコレート, クッキー, 母の日, チョコ
	1.0	バーゲン, ホンコンマダム, 父の日, 福袋,t シャツ, チョコレート, クッキー, サンダル, 母の日, チョコ
t_6	0.9	赤福, チョコレート, クッキー, チョコ
	1.0	新丸ビル, ミッドランドスクエア, ミッドタウン, チョコレート, クッキー, 東京ミッドタウン, チョコ
t_7	0.9	ハロウィン, チョコレート, クッキー, チョコ
	1.0	ハロウィン, サマーソニック, フジロック, チョコレート, サマソニ,summer sonic, チョコ
t_8	0.9	ハロウィン, チョコレート, クッキー, チョコ
	1.0	箱根駅伝, ハロウィン, なんばパークス, クリスマスケーキ, チョコレート, 三井グリーンランド, 京都 紅葉, チョコ...
...	...	...

アプローチを最初に導入したのは 2006 年に Chakrabati らによって提案された Evolutionary Clustering [16] である．Evolutionary Clustering では，入力データは n 個の時系列のグラフデータ $\{G_1, G_2, ...G_n\}$ で，それに対応したクラスタの時系列 $\{C_1, C_2, ..., C_n\}$ が出力される．各時間 t におけるグラフデータ G_t にはそれぞれ対応するノード間の類似度が行列 M_t の形で与えられている．Chacrabarti らは全体のコスト関数を以下のように定義している．

$$cost = CS(M_t, C_t) - cp \cdot CT(C_{t-1}, C_t)$$

ただし cp は Temporal Cost の重みを決定するためにユーザーが定義するパラメータである．Snapshot Cost と Temporal Cost の詳細な定義は利用するアルゴリズムによって異なり，アルゴリズムに合わせて適切に定義してやる必要がある．Chakrabati らはこの手法を階層クラスタリングと k-means クラスタリングについて適用しているが，コスト関数を最小化するために Greedy アルゴリズムによる近似でクラスタリングの結果を得ている．このアプローチを利用した研究として，本研究で使用した FacetNet [6] や Chi らが提案した Evolutionary Spectral Clustering [17] が挙げられる．

6. おわりに

6.1 まとめ

本研究ではトレンドを考慮した検索クエリのクラスタリングを行う上で，クエリと対応する時期に書かれた Blog 記事の URL 集合に関する Jaccard 係数を類似度とした構造が時間とともに変化するグラフのクラスタリングを FacetNet を用いて行った．そして，その結果に対して人手で正解付けを行ったデータを用いて行った評価実験を通して，Temporal Cost によってクラスタ構造の時間的一貫性を考慮することがクラスタリングの精度を向上する上で有効であり，本研究の評価実験においては $\alpha = 0.9$ が最適であることを確認した．

また，実際に抽出されたクラスタの内容を調査することによって，時間的一貫性を考慮した場合のほうがそうでない場合に比べて，時期によって同じクラスタに所属する組み合わせが変わるケースが確認されやすく，ノイズに相当するデータが混ざりにくいことを確認した．

6.2 今後の課題

本研究においても，クエリに対応する URL 集合の Jaccard 係数の他に，Chien ら [12] や上田ら [13] の手法にあるようなクエリログから得られる検索頻度の時系列や，Blog 記事数の時系列を類似度として考慮することによって URL の共起以外の類似性を持つクエリのクラスタリングが可能になり，今回の実験結果よりもトレンドを反映したクラスタを得られることを期待している．そのような複数の類似度を組み合わせた場合などについて今後の課題として取り扱っていきたい．また，今回の評価実験は 1048 語という限られたデータに対する実験のみであったため，より大規模なデータセットに対する評価実験や大規模化に向けたアルゴリズムの効率化についても取り組んでいきたい．

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様に感謝いたします

文　献

[1] “comscore releases rankings for top japanese web properties”, http://www.comscore.com/Press_Events/Press_Releases/2007/08/Top_Sites_in_Japan.

[2] 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).

[3] R. Baeza-Yates, C. Hurtado and M. Mendoza: “Query recommendation using query logs in search engines”, EDBT Workshops, **3268**, pp. 588–596 (2004).

[4] Z. Zhang and O. Nasraoui: “Mining search engine query logs for query recommendation”, Proceedings of the 15th international conference on World Wide Web, p. 1040 (2006).

[5] “「情報爆発時代のサーチ技術研究を加速する産学連携の開始　～yahoo!検索の検索語データの開放による研究の推進～」プレスリリース資料”, http://www.nii.ac.jp/news_jp/2008/03/yahoo.shtml.

[6] 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).

[7] J. Wen, J. Nie and H. Zhang: "Query clustering using user logs", ACM Transactions on Information Systems, **20**, 1, pp. 59–81 (2002).

[8] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen and H. Li: "Context-aware query suggestion by mining click-through and session data", Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 875–883 (2008).

[9] 甲谷, 湯本, 小山, 田中："Web ページに対する典型的なクエリの発見 (マイニング, 夏のデータベースワークショップ 2007 (データ工学, 一般))", 電子情報通信学会技術研究報告. DE, データ工学, **107**, 131, pp. 49–54 (2007).

[10] S. BEITZEL, E. JENSEN, D. LEWIS, A. CHOWDHURY and O. FRIEDER: "Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs", ACM Transactions on Information Systems, **10**, pp. 1229179–1229183 (2007).

[11] E. Diemert and G. Vandelle: "Unsupervised query categorization using automatically-built concept graphs", 18th International World Wide Web Conference (WWW2009) (2009).

[12] 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).

[13] 上田, 小野田, 角谷："クエリログの時系列的関係性を用いた非共起的関連語句の抽出とその応用", DEWS (2008).

[14] 安川, 横尾："クエリログから獲得した関連語のクラスタリングに基づく Web 検索", 電子情報通信学会論文誌 D, **90**, 2, pp. 269–280 (2007).

[15] 山口, 大島, 小山, 田中："サーチエンジンのクエリログを利用した同位語の発見", DBSJ Letters, **5**, 2, pp. 17–20 (2006).

[16] D. Chakrabarti, R. Kumar and A. Tomkins: "Evolutionary clustering", ACM SIGKDD (2006).

[17] Y. Chi, X. Song, D. Zhou, K. Hino and B. Tseng: "Evolutionary spectral clustering by incorporating temporal smoothness", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 162 (2007).