社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－146 (31)
2008／9／22

検索傾向の部分的な類似に基づくトピッククラスタリング

小野田 透† 湯本 高行†† 角谷 和俊†††

† 兵庫県立大学大学院環境人間学研究科 〒 670-0092 兵庫県姫路市新在家本町 1-1-12
†† 兵庫県立大学大学院工学研究科 〒 671-2280 兵庫県姫路市書写 2167
††† 兵庫県立大学環境人間学部 〒 670-0092 兵庫県姫路市新在家本町 1-1-12
E-mail: †nd07o007@stshse.u-hyogo.ac.jp, ††yumoto@eng.u-hyogo.ac.jp, †††sumiya@shse.u-hyogo.ac.jp

あらまし ユーザが検索を行う際，入力されたクエリに応じてシステムがクエリの推薦を行うサービスが普及している．このようなサービスはユーザが適切なクエリを入力できない場合などに効果的である．しかしながら，推薦されるクエリ同士の関連性などはあまり考慮されていない．そのような問題に対し，関連するクエリの抽出を行う研究が行われているが，それらは語の共起情報や検索傾向の類似性から関連するクエリを抽出するものが主である．本稿では，クエリの過去の検索データをクエリログをから取得し，異なる時期に検索が行われたクエリや検索の傾向が異なるクエリから，関連するクエリを抽出する手法を提案する．関連するクエリの判定が可能になることで，あるクエリを用いて検索を行ったユーザに対して別のクエリを推薦し，トピックに関して異なる視点を提供したり，ユーザに必要のないクエリを排除するなど，検索を効率的に行う支援を行うことが可能になると考えられる．
キーワード クエリログ，クラスタリング

Topic Clustering Based on Partial Similarity of the Search Tendency

Toru ONODA†, Takayuki YUMOTO††, and Kazutoshi SUMIYA†††

† Graduate School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
†† Graduate School of Engineering, University of Hyogo
2167 Syosya,Himeji,Hyogo,671-2280,Japan
††† School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
E-mail: †nd07o007@stshse.u-hyogo.ac.jp, ††yumoto@eng.u-hyogo.ac.jp, †††sumiya@shse.u-hyogo.ac.jp

Abstract Query recommendation systems based on inputted queries became widespread. These services are effective if users cannot input relevant queries. However, the conventional systems don't consider the relevance between recommended queries. There are some earlier studies to extract related keywords, but most of them base on co-occurring keywords or similarity between the search tendencies. In this paper, we propose a method to extract related queries used at different times and having different tendencies from query-log data.
Key words Query-log, Clustering

1. はじめに

近年では，Web を用いた情報の収集が一般的なものとなり，日々，多くのユーザが Web から情報を得ている．ユーザが Web から情報を得る手段として，Google，Yahoo!といった検索エンジンを用いるのが一般的である．ユーザは検索エンジンに対して適切なクエリを与えることで，ユーザが入力したクエリに関する Web ページを取得することが出来る．しかし，ユーザが適切なクエリを思いつくことができない場合，Web からの情報収集は困難なものとなる．このような場合にユーザを支援するサービスとして，ユーザが入力したクエリに応じて，システムが関連するクエリを推薦するサービスが提供されている．代表的なサービスの一つである Google サジェスト(注1) では，ユーザが検索窓にクエリを入力することで関連する複数のクエリを自動的に提示する．

しかし，推薦されるクエリは，ユーザが入力したクエリとの関連性は高いものの，推薦されるクエリ間の関係はあまり考慮されていない．推薦されるクエリの中には，非常に近いトピッ

(注1)：http://www.google.co.jp/webhp?complete=1& hl=ja

クを検索するものが複数含まれている場合もあれば，全く関係しないクエリが推薦されている場合もある．そして，ユーザはそれらの推薦されるクエリ間の関係を知ることは出来ない．クエリ間の関連を判定する手法として，クエリの過去の検索傾向に基づき類似性を判定する手法が提案されている [1]．しかし，この手法では検索傾向が全体的に類似するクエリのみ関連があるとしており，一部分が類似するクエリなどは考慮されていない．本稿では，関連するクエリを，クエリ間の時間的な検索傾向の部分的な類似によって抽出する手法を提案する．

提案手法では，クエリの時間的な検索傾向において，全体的な傾向の類似ではなく，部分的な類似によってクエリの類似性の判定を行う．我々は，検索傾向の部分的な類似性を判定することにより，クエリの関連を抽出できると考えた．例えば，北京オリンピックについて検索を行い情報を得たユーザが，関連するトピックとして過去のアテネオリンピックについての情報を検索するような場合は少なくない．このとき，北京オリンピックについての検索に用いられたクエリと，アテネオリンピックの情報の検索に用いられたクエリの時間的な検索傾向は部分的に類似する (図 1)．図 1 の実線で描かれたグラフは，クエリ { オリンピック，アテネ } の時間的な検索傾向，破線で描かれたグラフはクエリ { オリンピック，北京 } の時間的な検索傾向を示している．{ オリンピック，アテネ } と { オリンピック，北京 } の全体的な検索傾向を比較した場合，クエリ間の類似性は低くなると考えられる．しかし，{ オリンピック，北京 } が検索されている部分的な区間においては { オリンピック，アテネ } と { オリンピック，北京 } の検索傾向は類似している．このような部分的な類似は，北京オリンピックの開催によってアテネオリンピックに興味持った Web 利用者が，アテネオリンピックに関する検索を行ったために発生したものと考えられる．クエリ間の部分的な類似性から { オリンピック，アテネ } と { オリンピック，北京 } は関連するクエリであると判定する．

アテネオリンピックの開催時には { オリンピック，アテネ } 以外にも様々なクエリでアテネオリンピックについての検索が行われていたと考えられる．そのようなクエリは，検索に用いられたキーワードは異なるが，ユーザが検索の対象として想定したものは同一である．このような { オリンピック，アテネ } と検索傾向が類似するクエリを同一の対象を検索するものとして分類することで，アテネオリンピックに関連する異なるクエリを間接的に { オリンピック，北京 } と関連しているクエリとして捉えることができる．しかし，単純に検索傾向の類似だけで関連を抽出すれば，偶然検索傾向が類似したクエリも取得してしまう可能性がある．そこで，本手法では関連を抽出するクエリ間に共通のキーワードが含まれていることを分類の条件とした．つまり，{ オリンピック，アテネ } と検索傾向が類似するクエリとして，{ オリンピック，スケート } と { 海外，旅行 } というクエリが存在したとしても，共通のキーワードを含まない { アテネ，旅行 } は同一の対象を検索するクエリとして分類しない．このような共通のキーワードをクエリの検索するメインのトピックとして捉え，同一の対象を検索するクエリの集合を，メイントピックに対するサブトピックと呼ぶ．

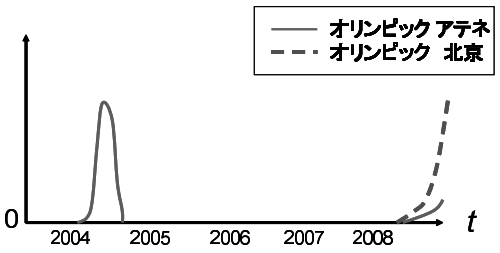

図 1　クエリ間の部分的な類似例

本手法では，最初にユーザが入力したクエリをユーザが検索しようとしているメインのトピックであるとし，入力クエリを共通して含むクエリをクエリログより収集する．次に，収集したクエリの過去の検索頻度の時系列データを用いて検索傾向が類似するクエリを判定し，それらを同一のクラスタに分類することで，入力クエリに関するサブトピックを生成する．そして，生成したサブトピックに属するクエリの部分的な類似性によって，関連するトピックのクラスタリングを行う．

以降，2 節では関連研究について，3 節ではクエリの部分的な類似性によるトピッククラスタリング手法について述べる．4 節で実験と実験の結果について，5 節で考察を述べる．6 節では本研究のまとめと今後の課題について述べる．

2. 関 連 研 究

クエリログを用いた先行研究として，Chien らはクエリの検索傾向の類似度を相関係数によって表し，類似するクエリを発見する手法を提案している [1]．Wang らは検索に用いられたキーワードのアスペクトを用いて，検索結果の分類とラベル付けを行う手法を提案している [2]．Zao，Baeza-Yates らは，あるクエリを入力したユーザがどのような Web ページを閲覧したかによって，クエリの類似性を計算，Web ページの改善などに用いる手法を提案している [3] [4]．しかし，彼らは検索キーワードの時間的な関係は考慮していない．我々は，クエリの時間的な検索傾向によって関連するクエリの判定を行い，分類を行う．また，我々は，クエリの検索頻度の時系列データのみを用いて分類を行っており，クエリを入力したユーザが閲覧したページなどの情報は利用していない．

トピック間の関係を抽出する手法として，森，三浦らは，時制的な側面を持つ Web ページ集合から関連するトピックの追跡を行う手法を提案している [5]．また，森，山田らは Web ページに記述されている事件などの出来事を抽出し，時間順序とトピックの関係間の表現を主とした情報の提示手法を提案している [6]．これらの研究は，トピック及びトピック間の関係判定に用いる情報を Web 上に存在するコンテンツから得ており，コンテンツの供給側が発信したトピックに関する関係抽出であると考えることができる．我々は，ユーザが入力した検索クエリのみからトピック間の関係を判定することを試みており，コンテンツ側の扱いに関わらず，ユーザが関心を持ったトピックに関して関係を判定している．

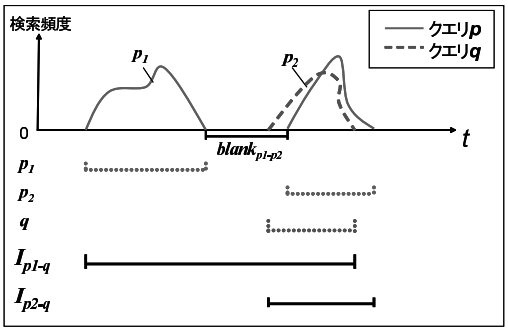

図 2　検索区間の分割と抽出

3. トピッククラスタリング

本節では，クエリ間の部分的な類似性に基づきトピックのクラスタリングを行う方法について述べる．まず，ユーザが入力したクエリをもとにしてクエリログからクエリを収集する．本稿では，ユーザが最初に入力したクエリを入力クエリ，入力クエリをもとにクエリログのデータを用いて拡張を行ったクエリを，拡張クエリと呼ぶ．収集した拡張クエリの時間的な検索傾向の類似度を算出し，サブトピックの生成を行う．

サブトピックとは，時間的な検索傾向が類似する拡張クエリの集合であり，同一のトピックの検索に用いられると考えられる拡張クエリの集合である．類似度が一定以上の値を持つ拡張クエリは，検索頻度の時間的な変化の傾向が近いクエリである．そのようなクエリは，検索に用いられたキーワードは異なるが，ユーザが検索の対象として想定したものは同一であると考えられる．例えば，{ 秋葉原，事件 } というクエリと { 秋葉原，通り魔 } というクエリは，用いられているキーワードは異なるが，共に秋葉原で発生した通り魔事件を検索するために入力されたクエリであると考えられる．このようなクエリを検索傾向の類似性によって同一のサブトピックを検索するものとして分類し，さらに生成されたサブトピックに属する拡張クエリ間の部分的な類似性によって，サブトピックのクラスタリングを行う．

3.1 拡張クエリの取得

クエリの拡張は，クエリログから過去に検索されたクエリのデータを取得し，入力クエリが含まれるクエリを抽出することで行う．

入力クエリを q とする．q は 1 語以上のキーワードで構成される．最初に，クエリログから q を構成するキーワードが全て含まれているクエリを拡張クエリとして取得する．本手法では，ユーザがキーワードを入力した順序は考慮していない．よって，仮に q が 2 つのキーワード a, b で構成されていた場合，キーワード a, b が含まれていることが取得の条件となり，q においてもクエリログ中のクエリにおいても a，b の入力順は関係しない．取得したクエリの集合から検索頻度の高いものから順にさらに n 件を抽出し，拡張クエリとする．

3.2 部分的な類似性の判定

3.2.1 検索区間の抽出

拡張クエリ間の部分的な類似性を判定するためには，類似度判定を行う区間を分割する必要がある．図 1 の { オリンピック，アテネ } と { オリンピック，北京 } の部分的な類似性を判定するためには，アテネオリンピックの開催時に検索されていた { オリンピック，アテネ } の検索区間と，北京オリンピックの開催時に検索された { オリンピック，アテネ } の検索区間はそれぞれ別に { オリンピック，北京 } との類似性を判定しなければならない．よって，拡張クエリの検索頻度の時間的変化に基づいて検索区間の分割を行い，部分的な検索区間を抽出する．

クエリ p の検索頻度の時間的変化を以下のようなベクトルで表す．

$$\boldsymbol{v}_p = (v_p^{(1)}, v_p^{(2)}, \cdots, v_p^{(n)}) \tag{1}$$

ただし，$v_p^{(i)}$ は時刻 t_i における検索回数である．本手法では，検索区間の部分的な類似性を判定するため検索区間の分割を行う．区間の分割は，$\boldsymbol{v}_p$ の検索頻度が閾値 α 以下になる期間が β 以上連続した場合に行う．これはクエリの検索頻度が閾値以下に減少し，それが一定区間以上続いた後，再び検索頻度が閾値以上に増加したとしても，それは異なるトピックに対するクエリである可能性が高いと考えられるためである．$\boldsymbol{v}_p$ は以下の式で表される．

$$\boldsymbol{v}_p = \boldsymbol{v}_{p,1} + \boldsymbol{v}_{p,2} + \cdots + \boldsymbol{v}_{p,m} + \boldsymbol{\epsilon}_p \tag{2}$$

また，$\boldsymbol{v}_{p,i}$ は以下の形式で表すことが出来る．

$$\boldsymbol{v}_{p,i} = (0, \cdots, 0, v_{p,i}^{(j)}, \cdots, v_{p,i}^{(j')}, 0, \cdots, 0) \tag{3}$$

ただし，この場合 $j \leqq l \leqq j'$ において $v_{p,i}^{(l)} > \alpha$ であり，$v_{p,i}^{l} > 0$ ならば任意の $j \neq i$ において $v_{p,j}^{(l)} = 0$ である．また，$\boldsymbol{\epsilon}_p$ は閾値 α 以下の成分からなり，$\epsilon_p^{(k)} > 0$ ならば任意の i において $\boldsymbol{v}_{p,i}^{(k)} = 0$ となる．つまり，検索頻度が α 以下ならばその時刻における検索頻度は 0 として扱う．

図 2 に検索区間の抽出例を示す．図 2 上部のグラフは縦軸がクエリの検索頻度，横軸が時間経過を表している．実線で描かれた曲線はクエリ p の検索頻度の時間的推移を，同様に破線で描かれた曲線はクエリ q の検索頻度の時間的推移を表している．図 2 下部の点線は，クエリ p，q の検索頻度が閾値 α 以上になる時区間を表している．分割された検索区間は以下の式で表される．

$$\boldsymbol{v}_{p,i} + \boldsymbol{v}_{p,i+1} = \tag{4}$$
$$(0, \cdots, 0, v_{p,i}^{(j)}, \cdots, v_{p,i}^{(j')}, \underbrace{0, \cdots, 0}_{\geq \beta}, v_{p,i+1}^{(h)}, \cdots, v_{p,i+1}^{(h')}, 0, \cdots, 0)$$

ただし，$j \leqq k \leqq j'$ において $v_{p,i}^{(k)} > \alpha$ かつ $h \leqq l \leqq h'$ において $v_{p,i+1}^{(l)} > \alpha$ である．例えば，図 2 では，$\boldsymbol{v}_p = \boldsymbol{v}_{p,1} + \boldsymbol{v}_{p,2}$ となり，$\boldsymbol{v}_{p,1}$ が図 2 中の p_1 に，$\boldsymbol{v}_{p,2}$ が p_2 に対応する．p_1，p_2 間の検索頻度が閾値 α 未満となる区間を $blank_{p1-p2}$ とし，その時間的な長さが閾値 β 以上であるとき，p_1，p_2 がそれぞれ独立の検索区間として抽出される．$blank_{p1-p2}$ が閾値 β 未満

である場合，p_1，p_2 は一つの検索区間として抽出する．

抽出された検索区間から類似度判定区間の生成を行う．類似度判定区間は，対象となる検索区間において最初に検索が発生した時点を始点とし，最後に検索が行われた時点を終点とする区間である．図 2 下部の 2 本の実線は，検索区間 p_1 と q の間で生成される類似度判定区間，p_2 と q の間で生成される類似度判定区間をそれぞれ表している．

3.2.2 類似性判定

拡張クエリ間の類似度を計算する手法について述べる．類似度の計算には相関係数を用い，2 つの拡張クエリの時間的な検索傾向がどの程度類似しているかを調べる．検索区間 $\boldsymbol{v}_{p,i}$，$\boldsymbol{v}_{q,j}$ の類似度を以下のように定義する．

$$sim(\boldsymbol{v}_{p,i}, \boldsymbol{v}_{q,j}) = cor(\boldsymbol{v}_{p,i}, \boldsymbol{v}_{q,j}) \tag{5}$$

ただし，$cor(\boldsymbol{v}_{p,i}, \boldsymbol{v}_{q,j})$ は $v_{p,i}^{(k)} \neq 0$ または $v_{q,j}^{(k)} \neq 0$ を満たすデータ列 $\{(v_{p,i}^{(k)}, v_{q,j}^{(k)})\}$ の相関係数である．相関係数とは二つの連続変数の関係を示す統計量である．相関係数は -1 から $+1$ までの値をとり，絶対値が大きくなるほど強い相関があるとされる．本稿では負の相関については考慮しておらず，拡張クエリ間の時間的な変化がどれだけ類似しているかを求めるために相関係数を用いている．2 組の数値からなるデータ列 $(x, y) = \{(x_i, y_i)\}(i = 1, 2, \ldots, n)$ があたえられたとき，x, y の相関係数は以下の式で求められる．

$$cor(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - m_x}{s_x} \right) \left(\frac{y_i - m_y}{s_y} \right) \tag{6}$$

m_x，m_y は x, y の平均値を s_x, s_y は x, y の標準偏差を表す．

計算された拡張クエリ間の類似度をもとに，拡張クエリを分類しサブトピックの生成を行う．類似度の値が閾値 γ 以上になる拡張クエリを同じ傾向で検索されていると見なし，同一のサブトピックに分類する．仮に p における検索区間 $\boldsymbol{v}_{p,1}$，$\boldsymbol{v}_{p,2}$ が異なるクラスタに存在する場合，それらは独立のクエリとしてそれぞれのサブトピックに分類する．また，サブトピック間の類似度の計算には最短距離法を用い，類似度の値が閾値 γ 以上になる場合はサブトピック同士の結合を行う．

3.3 部分的な類似性に基づくクラスタリング

サブトピックに属する拡張クエリの関係を用いて，サブトピックのクラスタリングを行う．判定を行うサブトピックを，それぞれ s_1, s_2 とする．s_1, s_2 の間に関連が存在する場合，サブトピック s_1 の情報を取得した Web 利用者によって，サブトピック s_2 に対しても検索が発生すると考えられる．よって，サブトピック間の関係の判定は，s_1, s_2 間に同一の拡張クエリが含まれているか否かで判定を行う．そして，s_1, s_2 に同一の拡張クエリが含まれており，検索区間が異なるものであるとき，s_1, s_2 には関連性があると判定する．

4. 実　　験

以下の 2 点を検証するため，実験を行った．

- 拡張キーワードが適切に分類されているか
- サブトピック間の関連が適切に抽出されているか

表 1　実験に用いた入力クエリ

No.	入力クエリ
1	{ チベット }
2	{ 船場吉兆 }
3	{ 年金 }
4	{ ライブドア }
5	{ 硫化水素 }
6	{winny}
7	{iphone，日本 }
8	{ オリンピック，北京 }

表 2　{ ライブドア } の予想分類結果

StNo.	拡張クエリ
1	{ ライブドア，近鉄 } { ライブドア，買収 } { ライブドア，球団 } { ライブドア，新球団 }
2	{ ライブドア，フジテレビ } { ライブドア，ニッポン放送 }
3	{ ライブドア，捜査 } { ライブドア，粉飾 } { ライブドア，強制捜査 } { ライブドア，違法 }

評価実験を以下に示す手順で行った．

Step1　入力クエリと，それに対する拡張クエリを準備する

Step2　拡張クエリ間の類似度を求め，サブトピックに分類する

Step3　入力クエリごとに予想分類結果を想定し，実際の分類結果と照らし合わせて評価を行う

4.1 データ準備

実験に用いるデータとして，入力クエリを 8 件，そして各々の入力クエリに対して拡張クエリを 10 件準備した．入力クエリは任意で選定を行い，拡張クエリはそれぞれ Web 上のニュース記事から選定を行った．実験に用いたクエリを表 1 に示す．次に，拡張クエリの過去の検索データの取得を行った．検索データの取得には Google Insights for Serch を用いた(注2)．Google Insights for Serch では，任意のクエリを入力することで，入力したクエリの過去の検索データを取得することができる．検索データは，7 日間を 1 単位として，その間の検索数を集計したものが返される．また，返されるのは検索頻度の実数データではなく，データ取得の対象区間の中で最も検索頻度が高くなる時点を 100 とした相対的な検索頻度のデータである．今回の実験では，対象区間を 2004 年 1 月 4 日から 2008 年 8 月 10 日までとし，データを取得した．検索区間の抽出に用いる閾値は，α を検索頻度の最大数の 100 分の 1，β を 14 日間とした．また，拡張クエリの分類に用いる閾値 γ を 0.7 とした．

4.2 実 験 結 果

分類結果の例として，{ ライブドア } の拡張クエリの予想分類結果と実際の分類結果をそれぞれ表 2，表 3 に示す．{ オリ

(注2)：http://google.com/insights/search/#

表 3 { ライブドア } の分類結果

StNo.	拡張クエリ
1	{ ライブドア，近鉄 } { ライブドア，買収$_1$}
2	{ ライブドア，球団 } { ライブドア，新球団 }
3	{ ライブドア，フジテレビ$_1$} { ライブドア，ニッポン放送 } { ライブドア，買収$_2$}
4	{ ライブドア，捜査 } { ライブドア，粉飾 } { ライブドア，強制捜査 } { ライブドア，違法 } { ライブドア，フジテレビ$_2$} { ライブドア，買収$_3$}

表 4 { オリンピック，北京 } の予想分類結果

StNo.	拡張クエリ
1	{ オリンピック，北京，予選 } { オリンピック，北京，代表 } { オリンピック，北京，選考 }
2	{ オリンピック，北京，開催 } { オリンピック，北京，開会式 }
3	{ オリンピック，北京，放送 } { オリンピック，北京，テレビ } { オリンピック，北京，中継 } { オリンピック，北京，競技 } { オリンピック，北京，種目 }

表 5 { オリンピック，北京 } の分類結果

StNo.	拡張クエリ
1	{ オリンピック，北京予選 }
2	{ オリンピック，北京，開会式 } { オリンピック，北京，開催 } { オリンピック，北京，放送 } { オリンピック，北京，競技 } { オリンピック，北京，種目 } { オリンピック，北京，テレビ } { オリンピック，北京，代表 } { オリンピック，北京，中継 } { オリンピック，北京，選考 }

ンピック，北京 } の拡張クエリの予想分類結果と実際の分類結果を表 4，表 5 に示す．複数のサブトピックに出現しているクエリは，同一の拡張クエリであるが，検索区間が異なるものである．それらを区別するために，標柱では同一の拡張クエリであるが，検索区間が異なるものに対し番号を振り表記している．

4.3 考　　察

4.3.1 分類に関する考察

{ ライブドア } の拡張クエリの分類では，ほぼ予想通りの結果が得られた．このクエリでは，ライブドアによる球団買収，フジテレビ，ニッポン放送株の大量取得，粉飾決済による強制捜査という 3 つのライブドアに関連するサブトピックを想定し，クエリを選定している．本文中では拡張クエリ ライブドア，近鉄を省略して 近鉄 と表記する．その他の拡張クエリも同様に入力クエリとの重複部分を省略して表記する．予想では { 近鉄 }，{ 買収 }，{ 球団 }，{ 新球団 } は同じサブトピックに分類されているが，実際には { 近鉄 }，{ 買収 } と { 球団 }，{ 新球団 } は検索された時期にずれがあり，異なるサブトピックとして分類された．図 3 に { ライブドア } の拡張クエリの時間的な検索傾向を示す．

{ オリンピック，北京 } の拡張クエリの分類では，予想よりもサブトピックが細かく分類されなかった．このクエリでは，北京オリンピックの代表選考，開会式，開催期間中の 3 つのサブトピックを想定している．予想通りの分類が行われなかった原因として，拡張クエリ { 選考 } などは，実際に選考などが行われていた時期でなく，北京オリンピックの開催前後に検索頻度が高くなったため，開催期間中の検索を想定した拡張クエリと類似度が高くなった．

このように，拡張クエリに関してユーザの検索履歴だけを用いていることで，実際にイベントが発生した時期ではなく，ユーザが興味を持った時期によって分類結果が左右される．今回の実験では，Web 上の情報を参考に時期に合わせたクエリを予想したため，分類結果と異なる結果となったと考えられる．

4.3.2 サブトピック間の関係抽出に関する考察

サブトピック間の関係抽出に関して，考察を行う．8 件の入力クエリのうち，サブトピック間の関係が抽出されたのは { ライブドア }，{ 年金 }，{winny} の 3 件であった．{ 年金 }，{winny} の拡張クエリの分類結果を表 6，表 7 に示す．

{ ライブドア } では，StNo.1，StNo.3，StNo.4 が関係するサブトピックとして判定された．これらのサブトピックは，{ ライブドア，買収 } というクエリによって結合されている．これらは，ライブドアに関する一連のサブトピックとして，関係があると判定されたのは妥当であると考える．しかし，StNo.1 に分類される { 買収 } は「近鉄球団の買収」を想定したクエリと考えられるが，StNo.2，StNo.3 に分類される {買収$_2$}，{買収$_3$} は「近鉄球団の買収」を想定し検索されたものか，「フジテレビの買収」を想定し検索されたものなのかを判別することは出来ず，別の判別手法が必要であると考えられる．

{ 年金 } では StNo.1，StNo.2 が関連するトピックとして判定された．StNo.1 は社会保険庁の年金記録消失に関係する拡張クエリ，StNo.2 には年金制度の改革や，同時期に発生した国会議員の年金未納問題に関する拡張クエリが分類されている．{winny} では StNo.1，StNo.2 が関連するトピックとして判定された．StNo.1 では winny の開発者逮捕に関する拡張クエリ，StNo.2 では開発者に対する裁判と判決に関する拡張クエリが分類されている．いずれも，入力クエリから想定される事件の一連の流れとして認識できるサブトピックが関係があると判定されている．

5. おわりに

本稿では，ユーザが入力したクエリに関する推薦クエリを，関連のあるクエリごとに分類して提示する手法の提案を行った．まず，ユーザの入力したクエリをもとにクエリの拡張を行い，

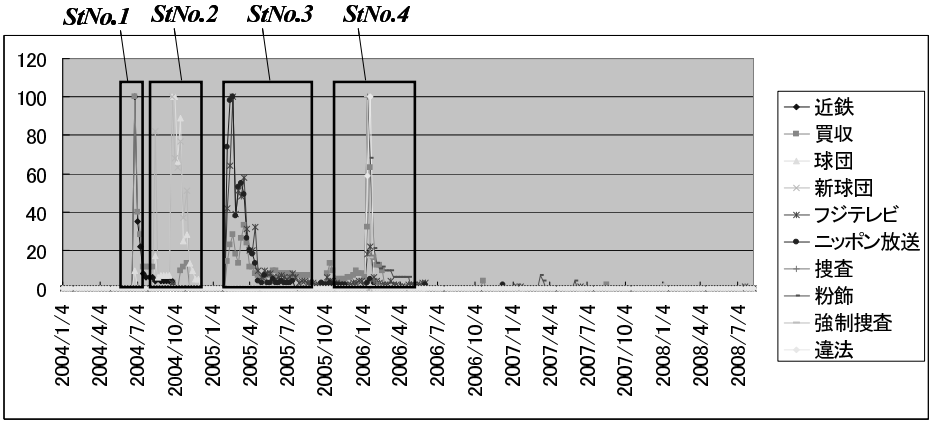

図 3 { ライブドア } の拡張クエリの時間的な検索傾向

表 6 { 年金 } の分類結果

StNo.	拡張クエリ
1	{ 社会保険庁 }
	{ 年金，記録 }
	{ 年金，問題 }
	{ 年金，確認$_1$}
2	{ 年金，未納 }
	{ 年金，改革 }
	{ 年金，仕組み }
	{ 年金，制度 }
	{ 年金，確認$_2$}
3	{ 年金，不正 }
4	{ 年金，保障 }

表 7 {winny} の分類結果

StNo.	拡張クエリ
1	{ 年金，開発$_1$}
	{ 年金，逮捕$_1$}
	{ 年金，開発者$_1$}
	{ 年金，違法$_1$}
	{ 年金，東大 }
2	{ 年金，開発$_2$}
	{ 年金，逮捕$_2$}
	{ 年金，開発者$_2$}
	{ 年金，違法$_2$}
	{ 裁判 }
	{ 判決 }
	{ 有罪 }
	{ 漏洩 }
	{ 個人情報 }

拡張したクエリの過去の検索傾向によって類似度を求め．分類を行う．さらに，分類したクエリ集合をユーザが入力したクエリが表すトピックのサブトピックであると見なし，関連のあるサブトピックをまとめてユーザへの提示を行う．

今後の課題として，従来手法との比較実験を行い必要がある．また，本手法の応用を考えていく必要がある．本手法で行っている，同一の検索対象に対するクエリの集約と関連の抽出はあるトピックに対する Web 全体での注目度の算出などに応用できると考えられる．あるトピックに対し複数の異なるクエリで検索が行われている場合，トピックの真の注目度を求めるのは難しい．本手法を用いることで，関連を持つクエリを抽出できトピックに対する注目度の集約を行うことができる．

謝　　辞

本研究の一部は，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです．ここに記して謝意を表します．

文　　献

[1] Steve Chien, N. I.: Semantic Similarity Between Search Engine Queries Using Temporal Correlation, *Proceedings of the 14th international conference on World Wide Web (WWW2005)*, pp. 2–11 (2005).

[2] Xuanhui Wang, C. Z.: Learn from Web Search Logs to Organize Search Results, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007)*, pp. 87–94 (2007).

[3] Kenneth Ward Church, P. H.: Word Association Norms, Mutual Information, and Lexicography, *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pp. 76–83 (1989).

[4] Qiankun Zhao, S. C. H. H.: Time-Dependent Semantic Similarity Measure of Queries Using Historical Click-Through Data, *Proceedings of the 15th international conference on World Wide Web (WWW2006)*, pp. 543–552 (2006).

[5] 森正輝，三浦孝夫，　塩谷勇：時制クラスタのトピック追跡，第 17 回データ工学ワークショップ (DEWS2006) 6A-i5 (2006).

[6] 森幹彦，山田誠二：Web における話題の時間変化の提示，*The 20th Annual Conference of the Japanese Society for Artificial Intelligence,3G1-2* (2006).