DBSJ Journal, Vol.7, No.3
December 2008

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

Topic Clustering Based on Partial Similarity of the Search Tendency

小野田 透♥ 湯本 高行♠ 角谷 和俊♦

Toru Onoda Takayuki Yumoto
Kazutoshi Sumiya

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候補の提示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自分の調べたいものに関して，どのようなキーワードを入力したら良いかわからない場合などに効果的である．しかしながら，従来のシステムではクエリ同士の関連性を考慮した提示までは行われていない．本稿では，クエリの過去の検索頻度データをクエリログから取得し，関連するクエリを抽出する手法を提案する．本手法では，異なるクエリの検索傾向における部分的な類似によって関連するクエリの抽出を行う．これにより，従来の検索時期の一致や検索傾向の全体的な類似を用いる手法では抽出不可能であった関連するクエリの抽出が可能になる．

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. In this paper, we propose a method to extract related queries used at different times and having different tendencies from query-log data.

1. はじめに

近年，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は，Google，Yahoo!といった検索エンジンを用いるのが一般的であり，ユーザは検索エンジンに対して適切なクエリを入力することで，Webページを取得することができる．しかし，ユーザが適切なクエリを入力する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，ユーザが入力したキーワードに応じて，システムがクエリの候補を提示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[1]では，ユーザがシステムにキーワードを入力することでユーザに対しクエリの候補を自動的に提示する．

♥学生会員 兵庫県立大学大学院環境人間学研究科
nd07o007@stshse.u-hyogo.ac.jp
♠正会員 兵庫県立大学大学院工学研究科
yumoto@eng.u-hyogo.ac.jp
♦正会員 兵庫県立大学環境人間学部
sumiya@shse.u-hyogo.ac.jp

1 http://www.google.com/webhp?hl=ja&complete=1

従来のサービスによって提示されるクエリは，ユーザが入力したキーワードとの関連性は高いものの，提示されるクエリ間の関係は考慮されていない．提示されるクエリの中には，非常に近いイベントを検索するクエリや，全く関係しないイベントを検索するクエリなど様々なクエリが提示される．そして，ユーザはそれらのクエリ間の関係を知ることはできない．

本稿では，時間的な検索傾向の部分的な類似によって互いに関連するクエリを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，ある限られた時区間における類似，つまり部分的な類似によってクエリ間の類似性の判定を行う．クエリ間の検索傾向の類似を部分的な時区間を用いて判定することで，全体的な検索傾向の類似を求めるだけでは抽出不可能なクエリ間の関係が抽出可能になる．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，関連するイベント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．

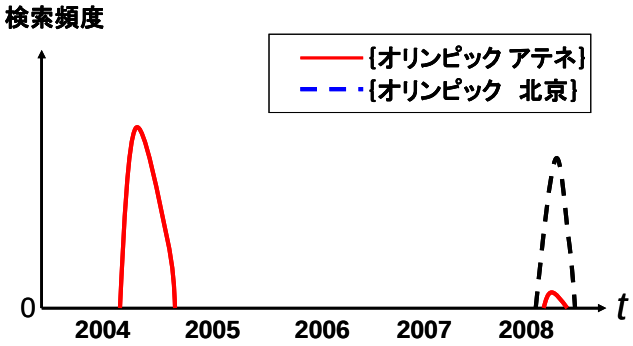

図1 クエリ間の部分的な類似
Fig.1 Comparison between queries in partial time section

図1は横軸を時間軸，縦軸をクエリの検索頻度とし，あるクエリの検索頻度の時間的変化をグラフで表したものである．図中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．ここで，{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時区間に限ってクエリの検索傾向を比較すると，{オリンピック，アテネ}と{オリンピック，北京}は検索頻度が増加し始めた時期，検索頻度が最大に達する時期，検索頻度が減少する時期など，時間的な検索の傾向が類似しているといえる．

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も興味を持ったWeb利用者が検索を行ったために発生したものと考えられる．つまり，北京オリンピックとアテネオリンピックの間には，ユーザにこの2つのイベントを関連して検索させるような関係があると考えられる．本稿では，このようにして検索が発生したクエリ間の関係を「連想関係」と呼ぶ．ここでいう連想とは，ある

イベントについて情報を得たユーザが，そのイベントによって他のイベントを思い出すことである．そして，そのような関係が存在する{オリンピック，アテネ}と{オリンピック，北京}は互いに関連を持つクエリであると考える．

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ほぼ同一であると考えられる．よって，{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，{オリンピック，アテネ}と同一の対象を検索するクエリとして分類することができると考えた．本稿では，このようなクエリの集合を，同一のトピックを検索するクエリの集合という意味で「トピック」と呼ぶ．そして，{オリンピック，アテネ}と{オリンピック，北京}が連想関係を持つとき，{オリンピック，アテネ}と同一の対象を検索するクエリもまた，{オリンピック，北京}と関連を持つと考えられる．よって，それぞれ異なるトピックに属するクエリが連想関係を持つとき， クエリが属している2つのトピック間には関連が存在すると考え，クエリ間の連想関係を用いたトピックのクラスタリングを行う．

本手法では，まずユーザが入力したクエリを構成するキーワードをすべて含むクエリをクエリログより取得する．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，同一の対象を検索するクエリの集合（トピック）を生成する．そして，生成したトピックに属するクエリの部分的な類似性によってクエリ間の連想関係を判定し，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する．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でき事を抽出し，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．

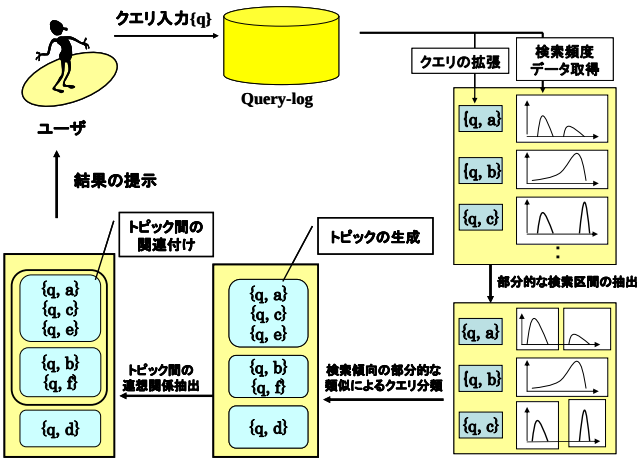

図2 本手法の手順
Fig2. Procedure of our method.

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．図2に本手法の手順を示す．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システムが拡張を行ったクエリを拡張クエリと呼ぶ．拡張クエリの取得と同時にクエリログから拡張クエリの検索頻度の時系列データを取得する．取得したデータから拡張クエリ間の時間的な検索傾向の類似度を算出し，トピックの生成を行う．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．このようなクエリは，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対象を検索するものとして分類し，さらに生成された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，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とすると，q は1語以上のキーワードで構成される．

まず，クエリログからqを構成するキーワードがすべて含まれているクエリを拡張クエリ候補として取得する．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．そのようにして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によるトピック生成

3.2.1 部分的な検索区間の抽出

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．例えば，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は以下のようなベクトルで表される．

$$v_p = (v_p^{(1)}, v_p^{(2)}, ..., v_p^{(n)}) \quad (1)$$

ただし，$v_p^{(i)}$ は時刻t_i における検索頻度である．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，v_pの検索頻度が閾値α以下になる期間がβ日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．v_pは以下の式で表される．

$$v_p = v_{p,1} + v_{p,2} + ... + v_{p,m} + \varepsilon_p \quad (2)$$

また，$v_{p,i}$は以下の式で表すことができる．

$$v_{p,i} = (0,...,0,v_{p,i}^{(j)},...,v_{p,i}^{(j')},0...,0) \quad (3)$$

ε_pは閾値α以下の成分からなり，$\varepsilon_p^{(k)}>0$ならば任意のiにおいて$v_{p,i}^{(k)}=0$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．

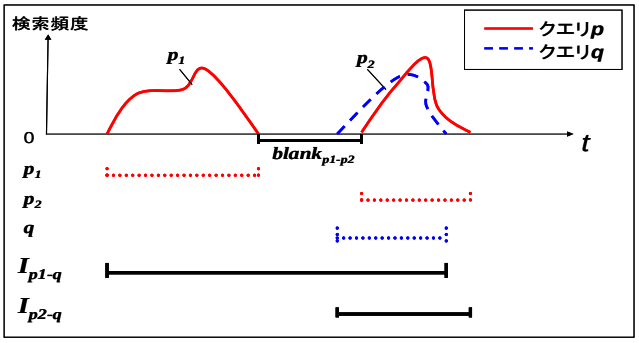

図3 検索区間の分割と抽出
Fig.3 Division and extraction of search section

図3に検索区間の抽出例を示す．図3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3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．

$$v_{p,i} + v_{p,i+1} = (0,...0,v_{p,i}^{(j)},...,v_{p,i}^{(j')},0,...,0,v_{p,i+1}^{(h)},...,v_{p,i+1}^{(h')},0,...,0) \quad (4)$$

ただし，j≦k≦j’において$v_{p,i}^{(k)} > \alpha$かつh≦1≦h’において$v_{p,i}^{(1)} > \alpha$である．例えば，図3では，$v_p = v_{p,1} + v_{p,2}$となり$v_{p,1}$が図3中のp_1に，$v_{p,2}$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1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3下部の2本の実線I_{p1-q}，I_{p2-q}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 部分的な類似性の判定

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$v_{p,i}$，$v_{q,j}$の類似度を以下のように定義する．

$$sim(v_{p,i}, v_{q,j}) = cor(v_{p,i}, v_{q,j}) \quad (5)$$

ただし，$cor(v_{p,i}, v_{q,j})$は$v^{(k)}_{p,i} \neq 0$，または$v^{(k)}_{q,j} \neq 0$ を満たすデータ列$\{(v^{(k)}_{p,i}, v^{(k)}_{q,j})\}$の相関係数である．相関係数とは2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，．．．，n) が与えられたとき，x，yの相関係数は以下の式で求められる．

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

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みなし，同一のトピックに分類する．仮にp における検索区間$v_{p,1}$，$v_{p,2}$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トピックに分類する．また，トピック間の結合のため，トピック間の類似度の計算を行う．計算はそれぞれのトピックに属するクエリの類似度を最短距離法によって計算し，トピック間の類似度とする．類似度の値が閾値γ以上になる場合は，2つのトピックを1つに結合する．

3.3 連想関係に基づくクラスタリング

生成したトピックに対し，トピックに属するクエリ間の連想関係に基づいたクラスタリングを行う．連想関係を持つクエリは必ず複数の部分的な検索区間を持ち，検索区間が他のクエリと類似している．例えば図1の｛オリンピック，アテネ｝はアテネオリンピックの開催時，北京オリンピックの開催時において2つの部分的な検索区間が発生しており，それぞれアテネオリンピックに関して検索するクエリ，北京オリンピックについて検索するクエリと検索傾向が類似する．このとき，｛オリンピック，アテネ｝は2つの異なるトピックに属するクエリとなる．

連想関係を持つクエリを抽出するため，生成したトピックから複数のトピックに属する拡張クエリの抽出を行う．トピックA，トピックBの間に同一の拡張クエリが含まれており，

かつ，Aに属しているクエリとBに属しているクエリの部分的な検索区間が異なるものであるとき，トピックA，トピックBには関連が存在するとみなし，クラスタリングを行う．

表1 実験に用いた入力クエリ
Table.1 Queries used for our experiment.

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}

4. 実験

拡張キーワードが適切に分類されているか，トピック間の関連が適切に抽出されているかを検証するため，実験を行った．評価実験は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，トピックを生成する
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クエリが適切に分類されているか評価を行う
Step4 連想関係によって関連付けられたトピック間の関係が妥当であるか評価を行う

4.1 実験データ

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 Search[2] を用いた．Google Insights for Search 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．また，｛オリンピック，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，出現する時区間が異なるものである．それらを区別するために，表中では同一の拡張クエリであるが，出現する時区間が異なるものに対し異なる番号を割り振り表記している．

表2 {ライブドア}の予想分類結果
Table.2 Expected classification result in the case of {livedoor}.

No.	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}

表3 {ライブドア}の分類結果
Table.3 Classification result in the case of {livedoor}.

No.	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$_3$}

4.3 考察

4.3.1 分類に関する考察

表3に｛ライブドア｝の拡張クエリの分類結果を示す．この結果では，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イベントを想定し，拡張クエリを選定した．予想では｛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｝は同じトピックに分類されているが，実際には｛ライブドア，近鉄｝，｛ライブドア，買収｝と｛ライブドア，球団｝，｛ライブドア，新球団｝は検索された時期にずれがあり，異なるトピックとして分類された．図4に｛ライブドア｝の拡張クエリの時間的な検索傾向を示す．

表5に｛オリンピック，北京｝の拡張クエリの分類結果を示す．この結果では，予想よりも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つのイベントを想定している．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｛オリンピック，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

[2] http://google.com/insights/search/#

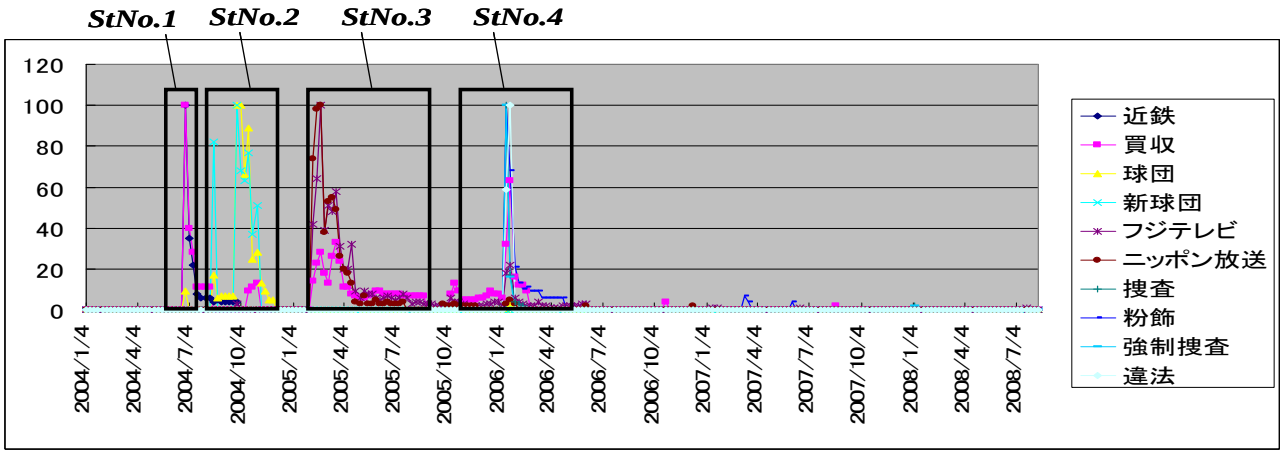

図4 {ライブドア} の拡張クエリの時間的な検索傾向
Fig.4 History of Query Frequency in the case of {livedoor}

表4 {オリンピック, 北京}の予想分類結果
Table.4 Expected classification result in the case of {olympic, beijing}.

No.	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
Table.5 Classification result in the case of {olympic, beijing}.

No.	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}

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 トピック間の関連付けに関する考察

連想関係の抽出によるトピッククラスタリングに関して，考察を行う．8 件の入力クエリのうち，トピック間の関連が抽出されたのは｛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．表3に示す｛ライブドア｝の分類結果では，No.1, No.3, No.4が関連するトピックとして判定された．これらのトピックは，｛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，No.1に分類される｛ライブドア，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，No.2, No.3に分類される｛ライブドア，買収$_2$｝，｛ライブドア，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できず，別の判別手法が必要であると考えられる．

｛年金｝の分類結果ではNo.1，No.2が関連するトピックとして判定された．結果を表6に示す．No.1 は社会保険庁の年金記録消失に関係する拡張クエリ，No.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｝を入力クエリとした場合の分類結果ではNo.1, No.2が関連するトピックとして判定された．結果を表 7 に示す．No.1 ではwinnyの開発者逮捕に関する拡張クエリ，No.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トピックが関連付けされている．

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，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トピックであるとみなし，関

表6 {年金}の分類結果
Table.6 Classification result in the case of {annual pension}.

No.	拡張クエリ
1	{年金, 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, 制度}
4	{年金, 保証}

連のある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う必要がある．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).

表7 { winny}の分類結果
Table.7 Classification result in the case of { winny}.

No.	拡張クエリ
1	{winny, 開発$_1$} {winny, 逮捕$_1$} {winny, 開発者$_1$} {winny, 違法$_1$} {winny, 東大}
2	{winny, 開発$_2$} {winny, 逮捕$_2$} {winny, 開発者$_2$} {winny, 違法$_2$} {winny, 裁判} {winny, 判決} {winny, 有罪} {winny, 漏洩} {winny, 個人情報}

小野田 透 Toru ONODA
兵庫県立大学大学院環境人間学研究科博士前期課程在学中．2006 年兵庫県立大学人間環境人間学科卒業．主に Web アーカイブ・クエリログに関する研究に従事．日本データベース学会学生会員．

湯本 高行 Takayuki YUMOTO
兵庫県立大学大学院工学研究科助教．2007年京都大学大学院情報学研究科博士後期課程修了．博士(情報学)．主に情報検索・情報統合に関する研究に従事．情報処理学会，日本データベース学会，ACM，IEEE 各会員．

角谷 和俊 Kazutoshi SUMIYA
兵庫県立大学環境人間学部環境人間学科教授．1998年神戸大学大学院自然科学研究科博士後期課程修了．工学博士．マルチメディアデータベース，データ放送の研究開発に従事．IEEE Computer Society，ACM，映像情報メディア学会，情報処理学会，日本データベース学会等各会員．