D-019

トピックグラフを用いた相対的文書ランキング
The topics graph based relative document ranking

畠中 翔太† 島田 諭‡ 三浦 孝夫†
Shota Hatakenaka Satoshi Shimada Takao Miura

1. はじめに

近年，インターネット上ではBlogやtwitterなど，多様で大量の情報が蓄積されている．これらの情報からユーザが必要とする情報を効率よく入手するためには，これまで主流だった情報検索だけでは必ずしも対応しきれない状況となってきている．

これまでの情報検索は，基本的には入力されたクエリと文書とのマッチングにより，候補文書をユーザに返すものである．クエリと文書のマッチングにおいては，単語の出現頻度などの情報が中心的に用いられてきた．しかし，ユーザの検索意図は多様であり，必ずしも出現頻度に基づく重み付けだけで対応できるとは限らない．また，ユーザや利用環境の多様化により，ユーザが常に適切なクエリを入力できるということも期待できなくなってきている．このため，従来の情報検索を補完，代替できるような情報アクセス技術の必要性が高まっている．

そのような技術の一つとして，ユーザが入力した少数または必ずしも情報検索において適切ではないクエリに対し，検索対象の文書集合において効果的な検索が可能である有用なクエリを提示，追加する「クエリ拡張」がある． 本研究では，ユーザの検索意図に追随したクエリ拡張の実現を目指す．
本稿では，提案手法の基礎となる，クエリに対する関連語および関連文書の相対的なランキング手法について検討した結果を報告する．

2. クエリ拡張

クエリ拡張は，情報検索を実用化するために重要な技術であり，様々な手法が提案されている．例えば，文書クラスタおよび単語クラスタからトピックを 抽出し関連語で拡張する手法，相対的にクリック数が多いもの (人気があるもの) に誘導による拡張手法，ユーザの閲覧履歴をもとに，ユーザの嗜好(検索履 歴)に合わせてパーソナライズな拡張手法などがあるが，予め決められた共通の テーマやトピックは存在しないために正確なクラスタリングは難しく，履歴情報 の不足による精度の低下などの問題がある．

また，ユーザの検索意 図は検索の過程で大幅に変遷 す る こ と が 知 ら れ て い る ．

†法政大学大学院 工学研究科
‡法政大学マイクロ・ナノテクノロジー研究センター

このようなユーザの検索意図に追随したクエリ拡張を行うためには，局所的なト ピックを高精度に特定できる手法と，大域的なトピックを効果的に用いてトピッ ク間をジャンプできる手法を併用することが必要になると考えられる．

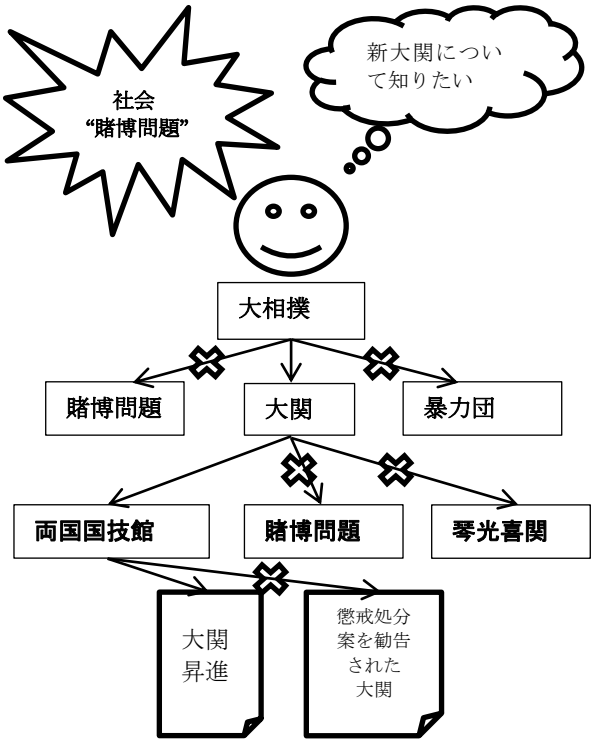

図 1,提案手法

3. 提案手法

本研究では，クエリに対する 1 日の文書頻度をヒストグラムで表し 2 つのヒストグラム間の類似度を計算する Bhattacharyya 係数[6]を用いることで、クエリ拡張を行う．クエリと抽出した関連語の Bhattacharyya 係数を用い，文書を相対的にランキングする．

Bhattacharyya係数とは，二つの正規化したヒストグラム間の類似度の計算などに用いられている．式は以下で表される．

$$\mathrm{L} = \sum_{u=1}^{m} \sqrt{\mathrm{P}_u Q_u} \quad (0 \leqq \mathrm{L} \leqq 1) \qquad (1)$$

ここで，p，qは比較対象となる正規化された（$\sum_{u=1}^{m} P_u = \sum_{u=1}^{m} Q_u = 1$）ビンを掛け合わせて算出する．

ある事象が発生したとき，電子文書，blog，twitter などは，文書ストリームを時系列に沿って観測すると，ある期間において，ある単語を含む文書の時間軸方向の密度が高くなるような状態が発生する．クエリに関連した単語はクエリと類似した時間軸方向の密度が高くなるので，クエリに対して Bhattacharyya 係数が高い語は関連語である．

ユーザが与えたクエリに対して Bhattacharyya 係数を用いて文書ランキングを行う．式は以下で表される．

$$d_q = \sum_{n=1}^{N} \frac{tf_i * Bha_{qi}}{N} \qquad (2)$$

クエリ q が与えられた時の文書 d のランク値を d_qとする．tf_iは文書 d での単語 i の頻度，Bha_{qi}はクエリ q と単語 i の Bhattacharyya 係数，N は文書 d の単語数である．これにより，クエリ q に関連した文書は高いランク値が与えられる．

ユーザがクエリ拡張から関連語 q’を選択すると，選択したクエリ以外の関連語はユーザの検索意図ではないことになる．

ユーザが初期に選んだクエリ q の Bhattacharyya 係数から関連語 q’以外の関連語の Bhattacharyya 係数を 0 に修正した Bhattacharyya 係数を作成する．

次に，関連語 q’の Bhattacharyya 係数と修正したクエリ q の Bhattacharyya 係数を足し合わせ，新しく Bhattacharyya 係数を作成する．

新しく作成した Bhattacharyya 係数を用いて，クエリ拡張を行うことで，検索者の検索意図でない関連語を下げる．

ユーザが初期に選んだクエリ q とクエリ拡張から選択した関連語 q’の Bhattacharyya 係数からクエリ拡張から選択しなかった関連語の Bhattacharyya 係数を 0 に修正し，足し合わせ，新しく Bhattacharyya 係数を作成する．

式２で，新しく作成した Bhattacharyya 係数を用いて，クエリ q と q’が与えられた時の文書 d のランク値を算出することで，検索意図に合わない文書のランク値を下げる．

4. 実験

4.1 実験方法

本稿では，新聞記事集合に対して提案手法を適用し，ユーザの検索意図を反映したクエリ拡張を試みる．

毎日新聞記事データ集 2010 年版から，1 面，2 面，3 面の各面に掲載された記事 1 年分，9,911 記事を用いて実験する．記事のタイトルおよび本文を形態素解析 し，名詞，動詞，未知語を使用する．名詞については，以下に示すルールで複合 語として扱う．

・連続する名詞は結合する
・接尾語に後続する語は結合しない
・数字は“＊”に置き換える

これらのルールにより，人名はフルネームで 1 語とし，数字を含む語は数字の桁 数にのみ着目する．

クエリ拡張において，極端な低頻度語および高頻度語は有用ではないと考えられることから，抽出された単語の文書頻度が 20 以上 1,000 未満となる 6,788 語を使用する．

クエリ拡張には上位 10 単語を使用する．

評価方法としては，検索意図にあった文書が上位にランキングしているのかを評価する．

例えば，“大相撲の大関”について知りたいが，“大相撲野球賭博問題”が原因で“大相撲・・・”などのクエリを与えると“大相撲野球賭博問題”に関する記事が上位に上がってしまうのを防ぐ．

4.2 実験結果

初期クエリ“大相撲”に対する関連語の Bhattacharyya 係数を表 1 に示す．また，クエリ拡張を行い“大相撲 大関”としたときの Bhattacharyya 係数を表 2 に，さらに“大相撲 大関 両国国技館”したときの Bhattacharyya 係数を表 3 に示す．

次に，クエリ“大相撲 大関 両国国技館”にマッチする 22 件の記事を提案手法に よりランキングして得られる上位 5 記事を表 4 に示す．22 件の記事中には，賭博問 題に関する記事が 15 件，理事選挙に関する記事が 3 件あり，ID2675 の記事のみが “大相撲 大関”についての記事である．

表 1. クエリ”大相撲”での Bhattacharyya 係数

大相撲		
Bha 係数	記事 ID	term
1	2644	大相撲
0．828794	2653	大関
0．800367	3899	日本相撲協会
0．756451	1929	力士
0．742017	6489	関脇
0．722026	4278	横綱
0．720092	6384	野球賭博
0．669799	2056	協会
0．658522	5731	親方
0．648716	4338	武蔵川理事長
0．639959	2281	名古屋場所

表 2. クエリ”大相撲 大関”での
Bhattacharyya 係数(クエリ拡張)

大相撲 大関		
Bha 係数	記事 ID	term
1．828794138	2644	大相撲
1．828794138	2653	大関
1．30290872	3974	時津風
1．272601265	4953	相撲協会
1．262531914	4789	琴光喜関
1．236375728	1930	力士ら
1．225277334	6021	賭博問題
1．194493536	1101	両国国技館
1．131219219	5742	角界
1．117314461	4718	特別調査委員会
1．114044308	2057	協会員

表 3. クエリ”大相撲 大関 両国国技館”での
Bhattacharyya 係数(クエリ拡張)

大相撲 大関 両国国技館		
Bha 係数	記事 ID	term
2．435927321	2644	大相撲
2．416154491	2653	大関
2．194493536	1101	両国国技館
1．590458825	3068	幕内
1．58998331	5460	臨時理事会
1．559216392	4782	理事会
1．53386505	2032	十両
1．481068325	2601	外部理事
1．477399884	4783	理事長
1．476763755	6489	関脇
1．461201778	4278	横綱
1．436617236	2460	土俵
1．431435544	3899	日本相撲協会

5. 考察

表 1，表 2，表 3 に示したように，クエリと同一のトピックに属する関連語とみなせる語は，Bhattacharyya 係数が高くなっている．特に，表 3 においては，新大 関に関する記事の検索を意図するクエリである“大相撲 大関 両国国技館”に対応 する関連語の Bhattacharyya 係数が，関連しない語の Bhattacharyya 係数よりも高 くなっている．このことから，Bhattacharyya 係数はピンポイントでのトピック の特定能力が高く，ユーザの検索意図を反映するクエリ拡張において有用である ことが確認できた．

また，表 4 に示したように，クエリに対する Bhattacharyya 係数の高い語を用い て，文書のランキングを行った．クエリ“大相撲 大関 両国国技館”に対して，“大相撲”という語は出現するもののユーザの検索意図とは合致しない別のトピッ クである“賭博問題”などの記事は下位にランキングされ，検索意図に合致する記 事が上位にランキングされた．

このことから，提案手法によりユーザの検索意図を反映したクエリ拡張を行い，関連する文書を適切にランキングできることを確 認した．

一方，Bhattacharyya 係数では，複数の検索 意図や曖昧な検索意図に対応するような大域的なトピックに属する語を得ることは困難である．ユーザの検索意図 は，検索の過程で大幅に変遷することが知られている．このような検索意図に追随したクエリ拡張を行うためには，Bhattacharyya 係数だけでなく，大域的なト ピックを参照したり，トピック間をジャンプするようなクエリを生成する必要が ある．このためには，文書のクラスタリング結果や出現語の共起関係などに基づくトピックグラフを構築して併用することが必要になると考えられる．

6. 結論

本稿では，ユーザの検索意図に追随したクエリ拡張を行うための基礎となる．

Bhattacharyya 係数を用いた関連語および関連文書の相対的なランキング手 法について報告した．

実験により，Bhattacharyya 係数はピンポイントでのト ピックの特定能力が高く，ユーザの検索意図を反映するクエリ拡張において有用 であることを確認した．また，提案手法により文書をランキングし，ユーザの検 索意図に合致する文書を上位にランキングできることを確認した．

今後の課題としては，検索の過程で大幅に変遷するユーザの検索意図に追随したクエリ拡張を実現するため，大域的なトピックを扱うトピックグラフの構築が挙げられる．

参考文献

[1] J. Kleinberg. Bursty and hierarchical structure in streams.In Proc. 8th SIGKDD, pp. 91–101, 2002.

[2] Masaya Murata, Hiroyuki Toda, Yumiko Matsuura and Ryoji Kataoka, “A Query Expansion Method Using Access Concentration Sites in Search Result”. Proceedings of the DataBase and Web symposium (DBWeb 2007)

[3] Hang Cui, Ji-RongWen, Jian-Yun Nie andWei-YingMa, “Probabilistic Query Expansion Using Quer Logs”. Proceedings of the 11th international conference on World Wide Web 2002, 325-332

[4] Georges E. Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In ACM SI-GIR, pp. 331–338, 2008.

[5]Mamoru Komachi, Shimpei Makimoto, Kei Uchiumi, and Manabu Sassano. Learning semantic categories from clickthrough logs. In ACL-IJCNLP, pp. 189–192, 2009.

[6] Qingshan LIU and Dimitris N. METAXAS. Unifying Subspace and Distance Metric Learning with Bhattacharyya Coefficient for Image Classification. Lecture Notes in Computer Science, 2009, Volume 5416/2009, 254-267

表 4:クエリ“大相撲　大関　両国国技館”

順位	記事 ID:第 1 段落
1	2675:日本相撲協会は 31 日午前、大阪府立体育会館で大相撲夏場所（5 月 9 日初日、両国国技館）の番付編成会議と理事会を開き、把瑠都（25）＝本名カイド・ホーベルソン、エストニア出身、尾上部屋＝の大関昇進を決めた。
2	5072:大相撲の賭博問題で、日本相撲協会の外部有識者で作る特別調査委員会（座長＝伊藤滋・早稲田大特命教授）は 2 日、東京・両国国技館で会合後に会見を行い、野球賭博への関与を認めている力士や親方ら協会員に関して、懲戒処分や謹慎を勧告した者以外についても氏名公表を勧告すると決めた。
3	4797:大相撲の賭博問題を受け、日本相撲協会は 21 日、東京・両国国技館で臨時理事会を開き、名古屋場所（7 月 11 日初日・愛知県体育館）の開催可否について、来月 4 日の理事会で最終決定することを決めた。
4	4955:大相撲の賭博問題で、日本相撲協会の外部有識者による特別調査委員会（座長＝伊藤滋・早大特命教授）が 28 日、謹慎勧告された武蔵川理事長（元横綱・三重ノ海）の代行職に外部理事の村山弘義・元東京高検検事長を推薦する意向であることが分かった。
5	4960:◇処分勧告案を協議，◇大嶽親方引退届、協会は不受理，大相撲の賭博問題で、日本相撲協会の外部有識者による特別調査委員会（座長＝伊藤滋・早大特命教授）が 28 日、謹慎勧告された武蔵川理事長（元横綱・三重ノ海）の代行職に外部理事の村山弘義・元東京高検検事長を推薦する意向であることが分かった。