ユーザの検索意図を反映させた文書ランキングに関する研究

畠中, 翔太 / HATAKENAKA, Shota

(発行年 / Year)
2013-03-24
(学位授与年月日 / Date of Granted)
2013-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2012年度　修 士 論 文

ユーザの検索意図を
反映させた文書ランキングに関する研究

STUDIES ON RANKING DOCUMENTS WITH QUERY-INTENT SENSITIVITY

指導教官　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

11R3129　畠中 翔太
Shota HATAKENAKA

目 次

第 1 章　序論 **3**
1.1　問題の背景 3
1.2　扱う問題 4
1.2.1　文書特性に基づく重要な文書 4
1.2.2　トピックを考慮した重要度 4
1.2.3　クエリ拡張を用いた検索意図の抽出 5
1.2.4　適合性フィードバックに基ずく検索意図を反映 5
1.3　論文の構成 5
1.4　発表論文 6

第 2 章　文書間の類似度を用いたランキング法 **8**
2.1　前書き 8
2.2　ベクトル空間モデルと重要な文書 9
2.2.1　ベクトル空間モデル 9
2.2.2　重要な文書 10
2.3　文書ランキング 10
2.3.1　PageRank アルゴリズム 10
2.3.2　新ランキングアルゴリズム 11
2.4　実験 13
2.4.1　実験方法 13
2.4.2　実験結果 13
2.4.3　考察 14
2.5　結論 14

第 3 章　トピックに適合する PageRank **17**
3.1　前書き 17
3.2　トピックドリフト問題 18
3.3　ベクトル空間モデル 19
3.4　PageRank アルゴリズム 19
3.5　トピックを考慮した重要度ランキングアルゴリズム 20
3.5.1　Topic Sensitive PageRank 20
3.5.2　トピックを考慮した重要度 21

3.5.3 検索時のトピックを考慮した文書の重要度 23
3.6 実験 23
3.6.1 実験方法 23
3.6.2 実験結果 24
3.6.3 考察 25
3.7 結論 26

第 4 章 トピックグラフを用いた相対的文書ランキング 30
4.1 はじめに 30
4.2 クエリ拡張 31
4.3 提案手法 31
4.4 実験 32
4.4.1 実験方法 32
4.4.2 実験結果 32
4.5 考察 33
4.6 結論 34

第 5 章 適合性フィードバックに基づく検索意図を反映させた文書ランキング 36
5.1 前書き 36
5.2 Rocchio アルゴリズム 37
5.3 提案手法 38
5.3.1 単語の関連度 38
5.3.2 フィードバック 39
5.3.3 手順 40
5.4 実験 40
5.4.1 実験方法 40
5.4.2 評価方法 40
5.4.3 実験結果 41
5.4.4 考察 41
5.5 結論 41

第 6 章 結論 48

謝辞 49

参考文献 50

第1章　序論

1.1　問題の背景

近年のインターネットやイントラネットの発展により, 電子化された文書が容易に, 大量に, また高速に入手できるようになった．電子文書・書籍は, 保存の容易さや手軽で高速に検索できることが利点であり，例えば情報爆発には，この検索技術なしにはあり得ない．実際, 典型的な例として Web, Wiki, blog, twitters などでは, 大量の情報を効率よく管理し, 同時に高性能な情報検索を実現する機構を前提としている．

代表的には，検索キーワードの形で与えられた問い合わせ (query) に対して, その検索キーワードにマッチする文書を選び, その存在有無 (term-matching) や出現頻度 (term frequency) の多いものを検索する．また, 少ない文書に含まれる単語はそれらを特徴付けることを逆文書頻度 (inverse document frequency) と呼び, これに基づいた TF*IDF 法 が知られる．

しかし, ユーザは検索対象の文書集合について知らないため, どのようなクエリを与えれば適切な検索結果が返されるかわからないため，ほとんどのユーザは短いクエリや曖昧なクエリを与えてしまうために, 検索意図にマッチする文書が検索できず, 検索結果の上位に検索意図とは関係ない文書が混じってしまうことがある．これは, 検索キーワードが同音異義語にある場合に起こる．例えば, 検索者が”マウス”で検索したときにコンピュータのポインティングデバイスのマウスについての文書を検索したいが, ネズミや口の”マウス”の出現頻度などが高いために, 検索結果の上位に検索意図とは関係ない文書が上位にランキングされてしまう．

このことから, 情報検索では, ユーザの検索意図を反映させた文書ランキングを行うアルゴリズムの実現が求められている．現在, この問題は様々な分野で研究が行われており, 確率モデルやクラスタリングを用いた研究もあり，文書が有する文脈をトピックモデルを用いて推定し, 同時にクエリの意図を推定して適合させる研究, 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」, ユーザが与えたクエリで得られた検索結果からどの文書が適合・不適合をシステムに学習させることで，ユーザにとって有用なクエリに改善する「適合性フィードバック」などの応用が期待されている．

1.2 扱う問題

本研究では, 新聞記事集合に対して提案を行う.

1.2.1 文書特性に基づく重要な文書

文書特性に関する取扱いは，例えば検索エンジンを用いた情報検索ですでに利用されている．問合わせ質問を手掛かりに，確率や特殊なアルゴリズム, 経験的な知識により重み付けを行い，何らかの特別な評価により文書を一定の順序に並べるが, ユーザは上位数個の文書しか閲覧しないため，ランキング技術は大切である．

従来, Webページの重要性を考えるときは，テーマ性 (theme)，正統性 (authoritative)，分配性 (distributive) の側面が考えられる．

本研究では, 正統的で分配的な文書を重要な文書と定義することで, 文書の重要度の度合いに文書間の類似度を用いる．しかし Web ページのように，文書間にリンクがないので文書の類似度によって関連を生成し，PageRank アルゴリズムを拡張して適用する手法を提案する．

この結果，2 つの重要な記事ランキングと話題性な記事ランキングの上位 10 記事で重複度 90 %を算出でき，本手法が重要な記事をランキングできたといえる．[1].

1.2.2 トピックを考慮した重要度

多くのランキングアルゴリズムの主要なアイデアは, 重要度とキーワードとの関連性から決定され, 重要度の算出のためには Web 空間上では PageRank アルゴリズムや HITS アルゴリズムがあるが，誤った重要度の算出により検索結果の上位に検索意図とは関係ない文書が上位にランキングされてしまう．問題の一つとしてトピックドリフト問題があり，ランキング手法においてトピックドリフト問題により正確な情報の入手が困難になっている．

本研究では, 図**??**のように検索キーワードからトピックを用いて検索意図を抽出し, 文書の重要度の度合いにトピックと文書間の類似度を用いる．これにより, トピックを考慮した重要度による文書ランキングすることで, トピックドリフトの問題の解決する．

この結果，提案手法のようにトピックを考慮した重要度ランキングアルゴリズムである Topic Sensitive PageRank との比較を行い，10 個の検索キーワード中 4 個の検索キーワードで Topic Sensitive PageRank が提案手法よりトピックドリフトしている記事を下位にランキングしていることから，本手法が検索意図にあった文書を上位にランキングするのに有効に働くことがわかる [2].

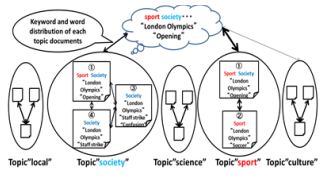

図 1.1: トピックを考慮した重要度

1.2.3　クエリ拡張を用いた検索意図の抽出

ユーザはどのようなクエリを与えれば，適切な検索結果が返されるのかわからないため，ほとんどのユーザは短いクエリや曖昧なクエリを与えてしまう．そのため，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．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，図??のように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．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を併用することが必要になると考えられるため，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を提案する．

この結果，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．本手法がユーザの検索意図に合致する文書を上位にランキングするのに有効に働くことがわかる.[3]

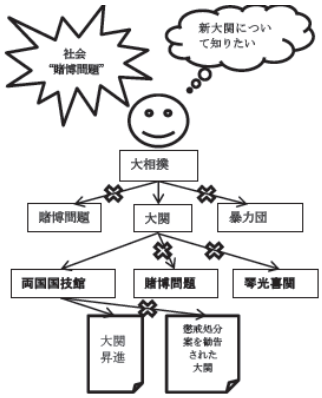

図 1.2: クエリ拡張を用いた検索意図の抽出

1.2.4　適合性フィードバックに基ずく検索意図を反映

ユーザは検索対象の文書集合について知らないので，どのようなクエリを与えれば，適切な検索結果が返されるかわからないため，ほとんどのユーザは短いクエリや曖昧

なクエリを与えてしまい，文書集合の特徴を活かしたランキング手法などでは, 文書集合の特徴（重要度，トピック）が必ずしもユーザの重要度やトピックの定義に一致しているとは限らない.

本研究では，適合性フィードバックを用いることで，ユーザにトピックを定義させ，またユーザが高速にほしい情報を手に入れるためにクエリに関連する複数のトピックを含む文書を上位にランキングさせることで，検索結果のリランキングを行い，検索結果の精度を向上させる手法を提案する.

この結果，一般的な適合性フィードバックである Rocchio アルゴリズムと比較した場合，5 クエリ中 3 クエリで提案手法が Rocchio アルゴリズムより高い適合率を算出され，本手法がユーザの検索意図に合致する文書を上位にランキングするのに有効に働くことがわかる．[4]

1.3 論文の構成

本研究では，以上の問題について以下の構成で論じる．第 2 章では，文書間の類似度を用いたランキング法について論じる．第 3 章では，トピックに適合する PageRank について論じる．第 4 章では，トピックグラフを用いた相対的文書ランキングについて論じる．第 5 章では，適合性フィードバックに基づく検索意図を反映させた文書ランキングを提案する．第 6 章で結論とする.

1.4 発表論文

1. 畠中翔太, 三浦孝夫: “文書間の類似度を用いたランキング手法”，第 3 回データ工学と情報マネジメントに関するフォーラム (*DEIM*), 2011.
   ベクトル空間モデルなどを用いた文書同士の類似度をリンクの重みとした仮想のリンクを生成することで，PageRank アルゴリズムにより文書間のリンク構造のみで文書の相対的な重要度を算出し，文書をランキングする．そこで本稿では，文書間の類似度を用いたランキング手法について提案する.

2. 畠中翔太, 三浦孝夫: “Ranking Documents using Similarity-based PageRanks”, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (*PacRim*), 2011.
   ベクトル空間モデルなどを用いた文書同士の類似度をリンクの重みとした仮想のリンクを生成することで，PageRank アルゴリズムにより文書間のリンク構造のみで文書の相対的な重要度を算出し，文書をランキングする．そこで本稿では，文書間の類似度を用いたランキング手法について提案する.

3. 畠中翔太, 三浦孝夫: “トピックに適合する PageRank”, 第 4 回データ工学と情報マネジメントに関するフォーラム (*DEIM*), 2012.
本稿では, トピックを考慮した重要度による文書ランキングすることで検索エンジンで発生するトピックドリフト問題の解決をする. トピックドリフトの原因と思われる 2 つの原因を解決することにより, ユーザが与えた問い合わせに意図しないページのランキングを下げる.

4. 畠中翔太, 三浦孝夫: “Ranking Documents with Query and Topic Sensitivity”, 7th International Conference on Digital Information Management (*ICDIM*), 2012.
本稿では, トピックを考慮した重要度による文書ランキングすることで検索エンジンで発生するトピックドリフト問題の解決をする. トピックドリフトの原因と思われる 2 つの原因を解決することにより, ユーザが与えた問い合わせに意図しないページのランキングを下げる.

5. 畠中翔太, 三浦孝夫: “Query and Topic Sensitive PageRank for General Documents”, 14th IEEE International Symposium on Web Systems Evolution (*WSE*), 2012.
本稿では, トピックを考慮した重要度による文書ランキングすることで検索エンジンで発生するトピックドリフト問題の解決をする. トピックドリフトの原因と思われる 2 つの原因を解決することにより, ユーザが与えた問い合わせに意図しないページのランキングを下げる.

6. 畠中 翔太, 島田 諭, 三浦 孝夫: “トピックグラフを用いた相対的文書ランキング”, 第 11 回 FIT（情報科学技術フォーラム）, 2012.
本稿では, ユーザの検索意図を質問に反映させるため,「クエリ拡張」を導入し, 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について提案する. 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示す.

7. 畠中 翔太, 島田 諭, 三浦 孝夫: “Ranking Documents with Query-Topic Sensitivity”, International Workshop on Web Information Retrieval Support Systems in IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology(*WIRSS*), 2012.
本稿では検索エンジンで発生するトピックドリフト問題の解決する. トピックドリフトの原因と思われる原因を文書学習し, バチャタリア係数を用いてこれを評

価することでクエリ解の品質が向上できることを示す.

8. 畠中　翔太, 島田 諭, 三浦 孝夫: “適合性フィードバックに基ずく検索意図を反映させた文書ランキング”, 第5回データ工学と情報マネジメントに関するフォーラム (*DEIM*), 2013.
   多くの検索エンジンは，あらかじめ決められたランキングアルゴリズムによって，Webページや文書の集合から抽出した特徴（重要度，トピック）やクエリなどから抽出したユーザの検索意図などを用いてユーザの検索意図に近い文書をランキングする．しかし，ユーザは検索システムに対して数語程度のキーワード入力しか行なわないため，検索意図を抽出するのは難しい．また，文書集合の特徴が必ずしもユーザの重要度やトピックの定義に一致しているとは限らない．本研究では，ユーザに与えた検索結果に対するユーザの意図をユーザには少ない負担でシステムにフィードバックさせることで，検索結果のリランキングを行い，検索結果の精度を向上させる．

第2章　文書間の類似度を用いたランキング法

ベクトル空間モデルなどを用いた文書同士の類似度をリンクの重みとした仮想のリンクを生成することで，PageRank アルゴリズムにより文書間のリンク構造のみで文書の相対的な重要度を算出し，文書をランキングする．そこで本稿では，文書間の類似度を用いたランキング手法について提案する．また，本提案をベースにした応用について議論を行う．

2.1　前書き

近年のインターネットの発展により, 電子化された文書が容易に, 大量に, しかも高速に入手できるようになった．電子文書・書籍は，手軽で高速に検索できることが利点があり，例えば情報爆発はこの検索技術なしにはあり得ない．実際，典型的な例として Web, Wiki, blog, twitters などでは，山のような情報を整理するために情報検索が欠かせない．

しかし, これらの情報から所望するものをどのように特定するのできるのか．従来, 情報検索に基づく技術と文書集合の特性を用いた手法が提案されている．

従来の情報検索技術では，キーワードの形で与えられた問い合わせ (query) に対して, そのキーワードにマッチする文書を選び, その存在有無 (term-matching) や出現頻度 (term frequency) の多いものを検索する．また，少ない文書に含まれる単語はそれらを特徴付けることを逆文書頻度 (inverse document frequency) と呼び，これに基づいた TF*IDF 法 が提案されている．

検索結果の適合度を改善するため, 文書の重みを算出する適合性フィードバックアルゴリズムが提案されている．しかし，従来の情報検索方法では，検索者が必要とする重要な文書を検索できているとはいえない．

文書特性に関する取扱いは，例えば検索エンジンを用いた情報検索ですでに利用されている．問合わせ質問を手掛かりに，確率や特殊なアルゴリズム, 経験的な知識により重み付けを行い，何らかの特別な評価により文書を一定の順序に並べる (ランキングする)．しかし，上位数個の文書しか閲覧しないことが多く，ランキング技術は大切である．

従来, Webページの重要性を考えるときは，テーマ性(theme)，正統性 (authoritative)，分配性 (distributive) の側面が考えられる．

通常, 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, 多くの人にとって興味ある共通のトピックやクラスタを抽出できればよい．テーマ検出・整理のための方法や特性化する方法は知られていない．テーマ抽出のためには, 頻出・相関するキーワードを用いて共通性を有する文書を求める方法や, 参照関係などからコミュニティの検出を行う方法が知られてるが, いずれも発見的である．

正統的な文書とは, 多くの影響力のあるトッピクを含み，他の重要な文書からその内容に言及されているものを言う．しかし，他の重要な文書が同じトッピクを有する文書とはいえず, 双方向的ではない．一方, 分配的な文書とは，必ずしも主要な内容を共有するわけではない文書であるが, 大量で多様な文書を言及する．本稿では, 正統的で分配的な文書を重要な文書と定義する．

検索結果の順位を決めるアルゴリズムをランキングアルゴリズムと呼ぶ．順位は多数のアルゴリズムの組合わせやチューニングパラメタにより決定される．多くのランキングアルゴリズムの主要なアイデアは, 重要度とキーワードとの関連性から決定される．重要度の算出のためには，Web 空間上では PageRank アルゴリズムや HITS アルゴリズムがある．本研究では, 文書の重要度の度合いに文書間の類似度を用いる．しかし Web ページのように，文書間にリンクがないので文書の類似度によって関連を生成し，PageRank アルゴリズムを拡張して適用する手法を論じる．

2.2 ベクトル空間モデルと重要な文書

2.2.1 ベクトル空間モデル

ベクトル空間モデルとは，大量の情報の中から利用者が与えた質問に対して利用者が必要と思われる文書の集合を提示するデータ表現方法である．

データ記述では文書を単語の多重集合で表現する．単語の並びを無視し，文書を多次元ベクトルとして表す．このモデルで重要な語を扱うには十分である．位置情報を失うため精密な表現ではないが，意味内容を表現しない語や記号などを無視できる．文書 d をベクトル $d =< v_1, ..., v_n >$ で表すとき $i = 1, ..., n$ は予め定まった語 w_i を表し，v_i はその語に与える重みを意味する．重みとしては，2 値や出現頻度，また TF*IDF 値が用いられることが多い．

またベクトル空間モデルでは，情報検索操作を自然に表現することができる．実際, ベクトルの各要素は語の重みを表すため，当該部の要素を対応させればよい．データベース検索と異なり，情報検索では，完全一致解よりも，質問に多く関連する文書を探索することが要求される．質問に類似する度合いを数値で表現し，この度合い順に文書をランキングして提示する．

例えば，与えられた質問をベクトル化した質問ベクトル q と文書の単語をベクトル化した多次元の文書ベクトル d_i のなす角度により類似度を計算する余弦類似度がある．

余弦類似度を $cos(d_i, q)$ とすると，$cos(d_i, q)$ は次式で表せる．

$$cos(d_i, q) = \frac{d_i * q}{|d_i||q|}$$

質問ベクトル q と全ての文書ベクトル $d_1, d_2 \cdots d_n$ の余弦類似度を計算し，余弦類似度の大きいほど質問に似た文書であることがわかる．

2.2.2 重要な文書

質問者が必要とする文書情報を見つけるには，何らかのキーワードで与えられた検索指示に従って結果を質問者に与えるか，キーワードから想定されるトピックにおいて重要と思える文書を抽出し，文書を与えるかのいずれかである．

検索による結果には，キーワードを多く含むか，人工的に様々な評価方法，キーワードの確率モデルやスコアリングなどによる重み付けされた文書をランキング結果が解として与えられる．

確率モデルやスコアリングでは， 語の重み付けにより重要な 文書と質問との類似度や適合度の度合いによりランキングするが，形態素解析の精度により順位が変動するので， あまり正確な順 位付けができない．

人工的に様々な評価方法では， アンケートや閲覧数などの人 工的に文書の重要度を算出する場合には，評価者の主観性な評 価による文書のランキングになってしまう．

これに対して，重要性を扱うためには，参照重要度を有する文書を算出する，つまり支持の高い文書の判定が必要となる．本研究では，文書集合から相対的に重要度を算出し，文書のランキングを行う．

2.3 文書ランキング

2.3.1 PageRank アルゴリズム

PageRank アルゴリズムとは，Web 空間のハイパーリンク構造を利用して Web ページの相対的な重要度を算出し，Web ページをランキングするアルゴリズムである．また，PageRank アルゴリズムで算出される Web ページの重要度を示す数値を PageRank 値と呼ぶ．

例えば，ページＡからページＢへの参照リンクをページＡによるページＢへの支持投票とみなし，この投票数によりそのページの重要度を判断する．また，投票数（リンク数）を見るだけではなく，票を投じたページについても考慮される．重要度の高いページからの支持票（被参照リンク）は高く評価され，重要度の高いページに支持せれたページは「重要なページ」になる．こうした繰り返しによって高評価を得た重要なページには高い PageRank（ページ順位）が与えられ，検索結果内の順位も高くな

る．しかし，リンクの構造のみでページの重要度を算出するので，ページの内容は考慮されない．

ページ P_i の PageRank 値を PR_i とし，入力リンクされているページ P_j の PageRank 値を PR_j とした時，次式で記述できる．

$$PR_i = \sum_{P_j \in B_{P_i}} \frac{PR_j}{|PR_j|}$$

B_{P_i} は P_i を指すページの集合であり，$|PR_j|$ はページ P_j からの出力リンクの総数である．この繰り返しにより PageRank の値は，最終的には安定した値に収束すると期待され繰り返される．また，行列記法を用いることで繰り返しごとに 1 × n ベクトルを使い，すべての PageRank を計算することができる．そのために行列 H と 1 × n ベクトル π^t を用い，次式で記述できる．

$$\pi^t = \pi^{t+1} H$$

行列 H は，ハイパーリンク構造に対する 2 値隣接行列を行で正規化した行列である．つまり，PageRank の計算は行列 H の固有ベクトル求めることである．しかし，現実の Web 空間はランクシンクや閉路などがあるために，固有ベクトルが収束しないので 2 回の調整をする．1 回目の調整は，出力リンクを 1 つも持たない Web ページで PageRank 値が集中しないように，出力リンクを 1 つも持たない Web ページはすべての Web ページに移動できる様に調整する．2 回目の調整は，固有ベクトルを急速に収束させるために，すべての Web ページがすべての Web ページに移動できる様に調整する．この 2 回の調整により行列 H は，次式の行列 H' のように書き換える．

$$H' = (1 - d)H + \frac{d}{N}$$

N は Web ページの総数であり，d は利用者がハイパーリンク構造に従わず移動する割合のパラメータである．

2.3.2 新ランキングアルゴリズム

本研究では，相対的に文書の重要度を算出させることで，文書ランキングをする．Web 空間で，Web ページに相対的に重要度を算出する PageRank アルゴリズムを用い，文書集合自体から文書の相対的な重要度を算出する．しかし，文書集合には，Web 空間でのリンクがないのでベクトル空間モデルを用い，余弦類似度による文書間の類似度により仮想リンクを生成する．文書の類似度によって関連を設定し，あたかもリンクが存在するかのようにし PageRank アルゴリズムを適用する．この仮想的な関連を仮想リンクと呼ぶ．Web 空間のような仮想リンク構造ができ，仮想リンクの重み付け

に文書間の類似度を用いることで，文書の内容を考慮した文書の重要度を算出することができる．

記事 d_i と記事 d_j の類似度を求めるとき，次式から求める．

$$cos(d_i, d_j) = \frac{d_i * d_j}{|d_i||d_j|}$$

このとき，類似度が 0 以外は記事間に仮想リンクができる．また，仮想リンクの重み付けに文書間の類似度を用いることで，リンクが対称性になる．

従来の PageRank アルゴリズムでは，ハイパーリンク構造に対する 2 値隣接行列を行で正規化した行列を扱っているので，ページの内容関係なくリンクの重みを等価になる．記事 d_0 が記事 $d_1 \sim d_n$ の N 個にリンクされている (支持されている) とき，記事 d_1 から記事 d_0 への入力リンクの重みを r_{01} としたとき

$$r_{01} = \frac{1}{N}$$

となり，r_{01} は記事 $d_1 \sim d_n$ から記事 d_0 への入力リンクの重みになる．提案手法では，記事 d_0 から記事 $d_1 \sim d_n$ へのリンクがあるとき，記事 d_1 から記事 d_0 への入力リンクの重みを rw_{01} としたとき

$$rw_{01} = \frac{w_{01}}{\sum_{i=1}^{n} w_{0i}}$$

w_{01} は記事 d_0 と記事 d_1 の類似度とする．仮想リンクに文書間の類似度をリンクの重みに使用することで，類似した内容の記事間のリンクの重みは高くなる．

これにより，記事の内容を考慮したリンクの重み付けができる．ページの類似度からしきい値を設定し，しきい値以下の仮想リンクは削除することで，記事の内容に関連性がないリンクを削除することができる．

作成した行列は余弦類似度により，リンクの重み付けをすると記事 d_0 から記事 d_0 のリンクの重みは 1 になり，自分自身にリンクがあることになる．自分自身にリンクは存在しないので，作成した行列から類似度の 1 を取り除く．

次に，本と書籍などの同じ物だが名称の違いにより，類似度が 0 になってしまうのをカバーするために，すべてのリンクの重みにパラメータ α を加える．このときのパラメータ α は従来の PageRank の 2 回目の調整と同じ $1/N$ とする．N は文書数である．余弦類似度により求めた行列を M' としたとき，次式が記述できる．

$$M' = (1 - d)M + d\ \alpha$$

d には，同じ物だが名称の違う単語が全文書にあるとして 0.15 の割合与える．上の式の行列 M の固有ベクトルを求めることで，PageRank を求める．これにより，例え

ば文書 A と文書 B が類似し，文書 B と文書 C が類似していても，文書 A と文書 C が類似しているとはいえないことから，文書集合から重要な文書 (多くの話題性を含んだ文書) ランキングができる．

2.4 実験

2.4.1 実験方法

実験では，新聞記事集合に対して提案手法を行う．新聞記事は，本文の第一段落のみで記事の大筋が要約することができる．毎日新聞の 2009 年の 1 月から 6 月の記事の内からランダムに抽出した 1 万記事の第 1 段落のみを使用する．記事の第 1 段落は形態素解析し，名詞のみを抽出したコーパスを使用する．

実験結果の評価方法として，PageRank のランキング上位 10 とすべての記事に対する余弦類似度の和のランキング上位 10 の重複度を用いる．PageRank のランキング上位 10 には，新聞記事集合の中で重要な記事が上位 10 にランキングされる．

提案手法では，記事間のリンクの重み付けに余弦類似度を用いたことにより，多くの話題性を含んだ記事が重要性な記事である．

重要な記事ランキングと話題性な記事ランキングの重複度が高いほど，重要な記事をランキングできたといえる．

二つのランキング r_1，r_2 の上位 10 の重複の割合は，次のように定義する．

$$Sim(r_1, r_2) = \frac{(A \cap B)}{k}$$

$A \cap B$ は二つのランキング r_1，r_2 の共通記事，k はランキングされている記事数である．PageRank のランキング上位 10 とすべての記事に対する余弦類似度の和のランキング上位 10 の重複度が高いほど，多くの話題性を含んだ記事は，重要性な記事であるといえる．

2.4.2 実験結果

表 1 は，PageRank ランキングは PageRank，記事のタイトル，記事の第一段落を示す．表 2 は，余弦類似度ランキングはすべての記事に対する余弦類似度の和，記事のタイトル，記事の第一段落を示す．表 3 は，PageRank ランキングと余弦類似度ランキングの上位 10 の重複の割合を示す．表 1 のランキング上位 10 の記事と表 2 のランキング上位 10 の記事が 10 記事の内，9 記事が同記事になっている．表 3 よりランキング上位の重複の割合は 90 %であることがわかる．また，表 1 のランキングの上位 4 記事と表 2 のランキングの上位 4 記事の順位が同じである．

また，表 1 の上位 9996 から 10000 の記事はタイトルや第一段落から，コラムであることがわかる．

2.4.3 考察

実験結果から，提案手法について考察を行う．

表 1 と表 2 のランキング上位 10 の記事は，重複の割合は 90 %であるので，重要な記事ができた．また，表 1 と表 2 の上位 4 から上位 9 まで順位は違うが，上位 4 から上位 9 まで同じ記事が重複している．このことから話題性のある記事ではなく，多くの話題性を含んだ記事がランキングすることができている．

表 1 のランキング下位には，コラム類がランキングされているのは，コラムは新聞記事の集合では話題性な記事ではないので，多くの話題性を含んだ記事は重要性な記事なのでコラム類が重要な記事ではないのといえるので，コラム類がランキング下位付けされている．また，上位 9997 から上位 10000 に関しては第 1 段落が＜ｖａｎｃｏｕｖｅｒ２０１０＞などからランキング下位の原因である．

提案手法での余弦類似度による仮想リンクの生成により，文書間のリンクは対称性になる．リンク関係を行列記法すると対称行列になるので，固有値は実数になり，べき乗より固有値と固有ベクトル (PageRank) を求めることができる．

また，本研究の提案手法は情報検索への適応することができる．本研究で使用したコーパスの記事の集合の代わりに Web ページの集合を使用することで，検索結果の上位に質問に対して重要なページに，類似したページの集合が検索結果の上位に表示される．これにより，従来の PageRank アルゴリズムとよりも良い検索結果が上位に表示されると考えられる．

2.5 結論

本研究では，文書間の類似度を用いたランキング手法を提案した．本手法により，文書集合から重要な記事を抽出することができた．そして，本稿では新聞記事集合に対してランキング手法を示したが，Web 空間に対して適応することができると考えられる．

表 2.1: PageRank ランキング

Top10	PR	タイトル	記事の第１段落
1	0.032	野球：ＷＢＣ　日本がカブスなどと練習試合	ＷＢＣ日本代表が２次ラウンドに出場 …
2	0.031	桜開花予想：早ければ3月２０日ごろ	民間気象会社「ウェザーニューズ」…
3	0.030	利益供与要求：容疑で元総会屋の男逮捕－－警視庁	東証１部上場の八千代銀行（本店・東京都新宿区）に利益供与を求めたとして …
4	0.0307	新型インフルエンザ：新たに静岡などで感染確認	静岡，千葉，徳島県で３日，男性１人，女性２人の新型インフルエンザ感染が …
5	0.0294	衆院：本会議開会，２１日午後で決定　「麻生降ろし」収束	政府は１７日，麻生太郎首相が２１日に衆院解散踏み切ることを踏まえ …
6	0.0290	ＧＷ：全国で晴れ多く	気象庁は２８日，ゴールデンウイークに …
7	0.0288	高速道路料金:トラック，バス計８日間半額に－－お盆	金子一義国土交通相は３０日の閣議後会見で休日（土日祝日）の …
8	0.0288	インフルエンザ：新人警官ら５２人感染－－栃木県警察学校	栃木県警察学校（宇都宮市若草）に７日入校した新人警察官（初任科生）ら …
9	0.0286	梅雨明け：九州南部で	鹿児島地方気象台は１２日，九州南部 …
10	0.0284	麻生首相：中国，欧州を訪問へ	河村建夫官房長官は２４日午前の …
～	～	～	～
9996	0.001	ガンマ線銀河：早稲田大や広島大などが発見 ブレーザーとは別の種類	非常に波長の短い高エネルギーのガンマ線を出す新しい種類の銀河を …
9997	0.001	１０年バンクーバー冬季五輪:…	＜ｖａｎｃｏｕｖｅｒ２０１０＞
9998	0.001	１０年バンクーバー冬季五輪:…	＜ｖａｎｃｏｕｖｅｒ２０１０＞
9999	0.001	１０年バンクーバー冬季五輪:…	＜ｖａｎｃｏｕｖｅｒ２０１０＞
10000	0.001	水と緑の地球環境 :…	＜マイＥＣＯ＞

表 2.2: 余弦類似度ランキング

TOP10	余弦類似度の和	タイトル	記事の第１段落
1	1436.3933	野球：ＷＢＣ　日本がカブスなどと練習試合	ＷＢＣ日本代表が２次ラウンドに出場 …
2	1379.4339	桜開花予想：早ければ３月２０日ごろ	民間気象会社「ウェザーニューズ」…
3	1348.871501	利益供与要求：容疑で元総会屋の男逮捕－－警視庁	東証１部上場の八千代銀行（本店・東京都新宿区）に利益供与を求めたとして …
4	1344.996201	新型インフルエンザ：新たに静岡などで感染確認	静岡，千葉，徳島県で３日，男性１人，女性２人の新型インフルエンザ感染が …
5	1270.10094	インフルエンザ：新人警官ら５２人感染－－栃木県警察学校	栃木県警察学校（宇都宮市若草）に７日入校した新人警察官（初任科生）ら …
6	1258.451663	衆院：本会議開会，２１日午後で決定　「麻生降ろし」収束	政府は１７日，麻生太郎首相が２１日に衆院解散に踏み切ることを踏まえ …
7	1248.997552	ＧＷ：全国で晴れ多く	気象庁は２８日，ゴールデンウイークに …
8	1244.844031	高速道路料金：トラック，バス計８日間半額に－－お盆	金子一義国土交通相は３０日の閣議後会見で休日（土日祝日）の …
9	1242.542849	梅雨明け：九州南部で	鹿児島地方気象台は１２日，九州南部 …
10	1238.969157	大相撲：夏巡業を追加	日本相撲協会は８日，夏巡業の日程を …

表 2.3: ランキングの重複度

比較するランキング	k	Sim
(PageRank ランキング, 余弦類似度ランキング)	10	0.9

第3章　トピックに適合するPageRank

本稿では, トピックを考慮した重要度による文書ランキングすることで検索エンジンで発生するトピックドリフト問題の解決をする．トピックドリフトの原因と思われる2つの原因を解決することにより, ユーザが与えた問い合わせに意図しないページのランキングを下げる．

3.1　前書き

近年のインターネットやイントラネットの発展により, 電子化された文書が容易に, 大量に, また高速に入手できるようになった．電子文書・書籍は, 保存の容易さや手軽で高速に検索できることが利点であり，例えば情報爆発には，この検索技術なしにはあり得ない．実際, 典型的な例として Web, Wiki, blog, twitters などでは, 大量の情報を効率よく管理し, 同時に高性能な情報検索を実現する機構を前提としている．

しかし, 検索者が必要とするものをどのように特定することができるのであろうか？情報検索に基づく技術と文書集合の特性を用いた様々な手法が提案されている．

代表的には，検索キーワードの形で与えられた問い合わせ (query) に対して, その検索キーワードにマッチする文書を選び, その存在有無 (term-matching) や出現頻度 (term frequency) の多いものを検索する．また, 少ない文書に含まれる単語はそれらを特徴付けることを逆文書頻度 (inverse document frequency) と呼び, これに基づいた TF*IDF 法 が知られる．

しかし, 残念なことに, しばしば検索意図にマッチする文書が検索できず, 検索結果の上位に検索意図とは関係ない文書が混じってしまうことがある．これは, 検索キーワードが同音異義語にある場合に起こる．例えば, 検索者が”マウス”で検索したときにコンピュータのポインティングデバイスのマウスについての文書を検索したいが, ネズミや口の”マウス”の出現頻度などが高いために, 検索結果の上位に検索意図とは関係ない文書が上位にランキングされてしまう．

文書特性に関する取扱いは, 例えば検索エンジンを用いた情報検索ですでに利用されている．検索キーワードを手掛かりに，確率や特殊なアルゴリズム, 経験的な知識により重み付けを行い, 何らかの特別な評価により文書を一定の順序に並べる (ランキングする)．

通常, 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, 多くの人にとって興味ある共通のトピックやクラスタを抽出できればよい．テーマ検出・整理のための

方法や特性化する方法は知られていない．テーマ抽出のためには, 頻出・相関するキーワードを用いて共通性を有する文書を求める方法や, 参照関係などからコミュニティの検出を行う方法が知られているが, いずれも発見的である．

このような何らかの特別な評価によるランキングでは，誤った評価により検索結果の上位に検索意図とは関係ない文書が上位にランキングされてしまう．特に，検索意図を満たすとした上位結果が同一の特徴を有する場合，検索意図と意味的なつながりを有するかのように錯覚してしまうことがある．この状況をトピックドリフティング (Topic Driftting) という．もとの検索意図（トピック）が移ってしまう（ドリフト）様を表す表現であって，検索意図と関係ない文書が上位を占める「検索エラー」とは異なることに注意したい．

検索結果の順位を決めるアルゴリズムをランキングアルゴリズムと呼ぶ．順位は多数のアルゴリズムの組み合わせやチューニングパラメタにより決定される．多くのランキングアルゴリズムの主要なアイデアは, 重要度と検索キーワードとの関連性から決定される．重要度の算出のためには，Web 空間上では PageRank アルゴリズムや HITS アルゴリズムがある．

本研究では, 検索キーワードからトピックを用いて検索意図を抽出し, 文書の重要度の度合いにトピックと文書間の類似度を用いる．これにより, トピックを考慮した重要度による文書ランキングすることで, トピックドリフトの問題の解決する．しかし Web ページのように，文書間にリンクがないので文書の類似度によって関連を生成し，PageRank アルゴリズムを拡張して適用する手法を論じる．

3.2 トピックドリフト問題

トピックドリフト問題とは，検索結果にユーザの検索意図を満たす上位結果が同一の特徴を有する場合，検索意図と意味的なつながりを有するかのように錯覚してしまうことである．

ランキング手法は，検索者が大量な検索結果から素早く必要な情報を入手するための手法だが，トピックドリフト問題により正確な情報の入手が困難になる．

トピックドリフトが発生する原因として，2 つの問題が考えられる．

1 つ目は，検索キーワードが同音異義語にある場合に起こる．例えば，マウスで検索したときにコンピュータのポインティングデバイスのマウスやネズミや口のマウスが混合してしまう．これでは，検索者の検索意図が曖昧になってしまい，検索結果の上位に検索意図とは関係ない文書が混じってしまう．

2 つ目は誤った評価によるランキングの場合に起こる．例えば，類似度，出現頻度，逆文書頻度，重要度などにより文書をランキングされる．検索者が”マウス”で検索したときにコンピュータのポインティングデバイスのマウスについての文書を検索したいが，ネズミや口の”マウス”の出現頻度や文書集合から抽出した逆文書頻度などが高いために，検索結果の上位に検索意図とは関係ない文書が上位にランキングされてしまう．

2つの原因により，検索結果にユーザの検索意図とは異なる文書が上位にランキングされる．

3.3 ベクトル空間モデル

ベクトル空間モデルとは，大量の情報の中から利用者が与えた質問に対して利用者が必要と思われる文書の集合を提示するデータ表現方法である．

データ記述では文書を単語の多重集合で表現する．単語の並びを無視し，文書を多次元ベクトルとして表す．このモデルで重要な語を扱うには十分である．位置情報を失うため精密な表現ではないが，意味内容を表現しない語や記号などを無視できる．文書 d をベクトル $d =< v_1, ..., v_n >$ で表すとき $i = 1, ..., n$ は予め定まった語 w_i を表し，v_i はその語に与える重みを意味する．重みとしては，2 値や出現頻度，また TF*IDF 値が用いられることが多い．

またベクトル空間モデルでは，情報検索操作を自然に表現することができる．実際，ベクトルの各要素は語の重みを表すため，当該部の要素を対応させればよい．データベース検索と異なり，情報検索では，完全一致解よりも，質問に多く関連する文書を探索することが要求される．質問に類似する度合いを数値で表現し，この度合い順に文書をランキングして提示する．

例えば，与えられた質問をベクトル化した質問ベクトル q と文書の単語をベクトル化した多次元の文書ベクトル d_i のなす角度により類似度を計算する余弦類似度がある．余弦類似度を $cos(d_i, q)$ とすると，$cos(d_i, q)$ は次式で表せる．

$$cos(d_i, q) = \frac{d_i * q}{|d_i||q|}$$

質問ベクトル q と全ての文書ベクトル $d_1, d_2 \cdots d_n$ の余弦類似度を計算し，余弦類似度の大きいほど質問に似た文書であることがわかる．

3.4 PageRank アルゴリズム

PageRank アルゴリズムとは，リンク構造を利用して文書の相対的な重要度を算出し，文書をランキングするアルゴリズムである．また，PageRank アルゴリズムで算出される文書の重要度を示す数値を PageRank 値と呼ぶ．

例えば，文書Aから文書Bへの参照リンクを文書 A による文書 B への支持投票とみなし，この投票数によりその文書の重要度を判断する．また，投票数（リンク数）を見るだけではなく，票を投じた文書についても考慮される．重要度の高い文書からの支持票（被参照リンク）は高く評価され，重要度の高い文書支持せれたページは「重要な文書」になる．こうした繰り返しによって高評価を得た重要な文書には高い PageRank（文書順位）が与えられ，検索結果内の順位も高くなる．しかし，リンク構造のみで文書の重要度を算出するので，文書の内容は考慮されない．

文書 d_i の PageRank 値を PR_i とし，入力リンクされている文書 d_j の PageRank 値を PR_j とした時，次式で記述できる．

$$PR_i = \sum_{d_j \in B_{d_i}} \frac{PR_j}{|PR_j|}$$

B_{d_i} は d_i を指す文書集合であり，$|PR_j|$ は文書 d_j からの出リンクの総数である．この繰り返しにより PageRank の値は，最終的には安定した値に収束すると期待され繰り返される．また，行列記法を用いることで繰り返しごとに 1 × n ベクトルを使い，すべての PageRank を計算することができる．そのために行列 H と 1 × n ベクトル π^t を用い，次式で記述できる．

$$\pi^t = \pi^{t+1} H$$

行列 H は，リンク構造に対する 2 値隣接行列を行で正規化した行列である．つまり，PageRank の計算は行列 H の固有ベクトル求めることである．しかし，現実の Web 空間はランクシンクや閉路などがあるために，固有ベクトルが収束しないので 2 回の調整をする．1 回目の調整は，出力リンクを 1 つも持たない文書で PageRank 値が集中しないように，出力リンクを 1 つも持たない文書はすべての文書に移動できる様に調整する．2 回目の調整は，固有ベクトルを急速に収束させるために，すべての文書がすべての文書に移動できる様に調整する．この 2 回の調整により行列 H は，次式の行列 H' のように書き換える．

$$H' = (1-d)H + \frac{d}{N}$$

N は Web 文書の総数であり，d は利用者がリンク構造に従わず移動する割合のパラメータである．

3.5 トピックを考慮した重要度ランキングアルゴリズム

3.5.1 Topic Sensitive PageRank

Topic Sensitive PageRank [7] は，1 つのみトピックを人手によりあてられた Web ページをカテゴリ分けされた ODP を用いて，ODP のトピックそれぞれに関して，PageRank アルゴリズムの 2 回目の調整に偏りを持たせた PageRank ベクトルを算出し，ODP のトピック c_j の PageRank ベクトルでの文書 d の PageRank 値 $rank_{jd}$ を算出する．ユーザの検索キーワード q がどの PageRank ベクトルに重要なのかの重みに単純ベイズを使用し，トピック c_j に検索キーワード q が重要な重み $P(c_j|q')$ を算出する．

$$P(c_j|q') = \frac{P(c_j) * P(q'|c_j)}{P(q')} \propto P(c_j) * \prod P(q'|c_j)$$

最後に，以下の式にしたがってすべてのページの重要度 s_{qd} を算出するアルゴリズム．

$$s_{qd} = \sum_{j}^{n} P(c_j|q') * rank_{jd}$$

これにより，ユーザの検索キーワードから ODP のトピックを用いて検索意図をトピックにより判定することで，トピックを考慮した重要度を算出しているアルゴリズムである．

3.5.2 トピックを考慮した重要度

本研究では，トピックを考慮した重要度を用いて，文書ランキングをする．

リンクのない文書集合から重要度を抽出するアルゴリズム [?] があるが，リンク付けに言語モデルやベクトル空間モデルを用いているが，文書は複数の話題を所有しているので、例えばある企業が球団を買収した主要な内容の文書があるとき、企業に社長が就任や球団に監督が就任などの主要な内容の文書に類似度により仮想リンクが張られ、文書の主要な話題が移り、誤った仮想リンク付けになり，正確な重要度が算出できていない．

文書の複数の話題に複数のトピックと文書間の類似度で対応する．文書に複数のトピックを指定し，同トピックの文書間に類似度で仮想リンクを生成し，仮想リンク構造に PageRank アルゴリズムを適応することで重要度を抽出する．学習データを用いて，文書と学習データの余弦類似度により，文書へ複数のトピックを与える．この仮想リンク付けにより、同じ主要な話題の内容の文書には類似度が高くなり，また 2 つの文書間に同単語を含んでいても、同トピックではない文書間にはリンクが張られないので，主要な内容が移るのを防ぐ．仮想リンクには余弦類似度を用いる．これにより，トピックを考慮した文書の重要度を算出することができる．

文書 d_i と学習データのトピック t_j との類似度 T_{ji} を求めるとき，次式から求める．

$$T_{ji} = \frac{d_i * t_j}{|d_i||t_j|}$$

文書と学習データとの類似度から閾値を設定し，閾値以下はトピック t_j を割り当てない．これにより，文書に複数のトピックを割り当てることで，トピック t_j を含む文書集合 s_j ができる．トピック t_j を含む文書集合 s_j 内で仮想リンクを生成することで，文書間の内容の類似性を考慮した仮想リンクを生成する．

文書 d_i と文書 d_j の類似度 w_{ij} を求めるとき，次式から求める．

$$w_{ij} = \frac{d_i * d_j}{|d_i||d_j|}$$

このとき，類似度が 0 以外は記事間に仮想リンクができる．また，仮想リンクの重み付けに文書間の類似度を用いることで，リンクが対称性になる．

従来の PageRank アルゴリズムでは，ハイパーリンク構造に対する 2 値隣接行列を行で正規化した行列を扱っているので，文書の内容関係なくリンクの重みを等価になる．文書 d_0 が文書 $d_1 \sim d_n$ の N 個にリンクされている (支持されている) とき，文書 d_1 から文書 d_0 への入力リンクの重みを r_{01} としたとき

$$r_{01} = \frac{1}{N}$$

となり，r_{01} は文書 $d_1 \sim d_n$ から文書 d_0 への入力リンクの重みになる．提案手法では，文書 d_0 から文書 $d_1 \sim d_n$ へのリンクがあるとき，文書 d_1 から文書 d_0 への入力リンクの重みを r'_{01} としたとき

$$r'_{01} = \frac{w_{01}}{\sum_{i=1}^{n} w_{0i}}$$

仮想リンクに文書間の類似度をリンクの重みに使用することで，類似した内容の文書間のリンクの重みは高くなる．

これにより，文書の内容を考慮したリンクの重み付けができる．文書の類似度から閾値を設定し，しきい値以下の仮想リンクは削除することで，文書の内容に関連性がないリンクを削除することができる．

作成した行列は余弦類似度により，リンクの重み付けをすると文書 d_0 から文書 d_0 のリンクの重みは 1 になり，自分自身にリンクがあることになる．自分自身にリンクは存在しないので，作成した行列から類似度の 1 を取り除く．

PageRank アルゴリズムと同じように，固有ベクトルを急速に収束させるために，収束パラメータ α を従来の PageRank の 2 回目の調整と同じ $1/N$ とする．N はトピック t_j を含む文書集合の文書数である．余弦類似度により求めた行列を M' としたとき，次式が記述できる．

$$M' = (1 - d)M + d\ \alpha$$

d には，PageRank アルゴリズムと同じように 0.15 の割合与える．上の式の行列 M の固有ベクトルを求めることで，PageRank を求める．

次に，文書集合 s_j の文書の PageRank 値と他のトピックの文書集合の PageRank 値を文書数に差異があるのに，同等に扱うのは誤っているので正規化を行う．文書集合 s_j の文書 d_i の正規化した PageRank 値を PR'_{ji} としたとき

$$PR'_{ji} = \frac{|s_j|}{\sum_{k=1}^{n} |s_k|} * PR_{ji}$$

文書集合 s_j の文書 d_i の PageRank 値を PR_{ji} とする．

3.5.3 検索時のトピックを考慮した文書の重要度

ユーザが与えた検索キーワードに対して，トピック t_j にどのトピックに属するのかので重みを与えることで，検索キーワードが同音異義語などによる検索者の検索意図の曖昧性を解消する．検索キーワードに対して，トピック t_j に属する重みを与えるのに，検索キーワード q_m と文書集合 s_j の単語分布との類似度 Q_{mj} を重みとして与える．文書集合 s_j の単語分布を w_j としたとき

$$Q_{jm} = \frac{q_m * w_j}{|q_m||w_j|}$$

類似度 Q_{jm} から閾値を設定し，閾値以下は重みを与えない．次に，検索キーワード q_m が与えられた時の文書 d_i の重要度を次式より求める．

$$v_{mi} = \sum_{j=1}^{n} T_{ji} * PR'_{ji} * Q_{jm}$$

検索キーワード q_m が与えられた時の文書 d_i の重要度 v_{mi} とする．これにより，トピックを考慮した重要度による文書ランキングする．

例えば，検索者が“ 最下位の○×球団 ”について知りたく，また世の中では“ △□企業が（最下位の）○×球団を買収 ”が話題の場合，リンク構造のない文書集合に PageRank アルゴリズムを適応させる手法 [?] では，全文集集合から文書の重要度を抽出すると，検索キーワード“ 最下位　○×球団 ”では“ △□企業が（最下位の）○×球団を買収 ”の話題の文書が上位にランキングされる．提案手法では，検索キーワード“ 最下位 ○×球団 ”から検索意図をトピックを用いてスポーツや経済を抽出する．文書の複数の話題をトピックを用いて表し,“ 最下位の○×球団 ”の内容の文書では、学習データからスポーツなどのトピックが指定され、“ △□企業が（最下位の）○×球団を買収 ”の内容の文書では，経済，スポーツなどが指定される．検索キーワードがスポーツなどに属する重みと　の文書のスポーツなどの割合をかけることで，検索者の検索意図と文書の主要な話題が重要度に反映されるので,“ 最下位の○×球団 ”の内容の文書のトピックを考慮した重要度は高くなる．検索者によって文書のトピックを考慮した重要度が変わり，検索者の検索意図にあった文書ランキングが行える．しかし，学習データで正確なトピック指定ができないと，トピック“ スポーツ ”の文書集合でも“ △□企業が（最下位の）○×球団を買収 ”の内容の文書が多くなってしまい，重要度がトピック“ スポーツ ”,“ 経済 ”の文書集合で高くなってしまう．

3.6 実験

3.6.1 実験方法

実験では，新聞記事集合に対して提案手法を行う．新聞記事は，本文の第一段落のみで記事の大筋が要約することができる．

毎日新聞の種別コードより 8 種類のトピック (国際，経済，家庭，文化，科学，芸能，スポーツ，社会) の 2010 年の 1 月から 6 月の記事の内からランダムに抽出した 1000 記事の第 1 段落のみを使用する．記事の第 1 段落は形態素解析し，名詞のみを抽出したコーパスを使用する．

学習データは，毎日新聞の種別コードより 2010 年の 1 月から 6 月の 8 種類のトピックの記事の全タイトルを使用する．記事のタイトルは形態素解析し，名詞のみを抽出した単語集合を学習データとして使用する．

実験結果の評価方法として，人手によって作られた正解データをトピックを PageRank スコアの算出に反映させた Topic Sensitive PageRank と比較する．

しかし，Topic Sensitive PageRank はハイパーリンクを用いて PageRank 値を算出するが，新聞記事集合にはハイパーリンクがないので，提案手法と同様に余弦類似度により仮想リンクを生成し，2 値化することで仮想ハイパーリンク構造を生成する．また，ODP の代わりに毎日新聞の種別コードの 8 トピックを用いて，PageRank ベクトルに偏りを与える．

表 8 は人手によって作られた正解データの検索キーワード”冬季 五輪”の一例．1000 記事に検索キーワード”冬季 五輪”をキーワードマッチングさせ，マッチングした 8 個の文書である．検索キーワード”冬季 五輪”から検索者は，”冬季五輪に関連した記事”という検索意図のあることがわかる．表 8 からわかるように記事 988 は明らかに”冬季五輪に関連した記事”ではないことがわかる．この正解データを用いて，どちらの手法がトピックドリフトしている記事を下位にランキングしているのかを比較する．

3.6.2 実験結果

表 3.1: TSPR と MYPR の比較

検索キーワード	TSPR	MYPR	記事数
サッカー 日本 代表	3	6	8
トヨタ自動車	9	9	10
バンクーバー 五輪	15	15	15
研究 受賞	2,3,5	3,4,5	5
高校 野球 大会	10	11	11
初 優勝	5	8	10
大会 開幕	9	9	9
冬季 五輪	8	4	8
米 大統領	5	5	6
毎日新聞社 主催	11	10	11

表 1 は検索キーワード，Topic Sensitive PageRank と提案手法のトピックドリフトしている文書のランキング，1000 記事中検索キーワードにマッチングした記事数を表し

ている．
10 個の検索キーワード中 4 個の検索キーワードで提案手法が Topic Sensitive PageRank よりトピックドリフトしている記事を下位にランキング，その他の 4 個の検索キーワードで同順位のランキング，残りの 2 個の検索キーワードでは Topic Sensitive PageRank が提案手法よりトピックドリフトしている記事を下位にランキングしている．

3.6.3 考察

実験結果から，提案手法について考察を行う．

表 1 から提案手法が Topic Sensitive PageRank よりもトピックドリフトしている記事を下位にランキングしている．この中でもランキング差に 3 位以上した検索キーワードの内”サッカー 日本 代表”，”冬季 五輪”について考察する．

検索キーワード”サッカー 日本 代表”では，トピックドリフトしている記事は記事 694．提案手法が Topic Sensitive PageRank より記事 694 を下位にランキングしている．

表 2，3，7 から Topic Sensitive PageRank 手法では，すべての記事の重要度と検索キーワードのトピックに一番”スポーツ”を反映させていることがわかる．記事 694 は直観的にスポーツの記事だが，サッカー日本代表に関する記事ではない．しかし，記事 694 の重要度には一番”スポーツ”を反映され，その他のトピックは反映されていない．Topic Sensitive PageRank 手法では，記事 694 を上位 3 位にランキングされた．

表 2，4，7 から提案手法ですべての記事の重要度に一番”スポーツ”を反映され，全記事は直観的にスポーツの記事である．また，記事によって記事の重要度へのトピックの反映度に違いがあり，記事 694 ではトピック”国際”，”経済”が”スポーツ”の次に高く，トピック”科学”，”家庭”は反映されていない．検索キーワードのトピックには”家庭”以外のトピックを反映され，その中でもトピック”経済”，”科学”，”スポーツ”が高い．提案手法により，記事 694 を上位 6 位にランキングしている．

検索キーワード”冬季 五輪”では，トピックドリフトしている記事は記事 988．Topic Sensitive PageRank が提案手法より記事 988 を下位にランキングしている．

表 2，5，8 から Topic Sensitive PageRank 手法で記事 988 以外の重要度には一番”スポーツ”を反映され，記事 988 の重要度には一番”社会”を反映され，その他のトピックは反映されていない．記事 988 は直観的にトピック”社会”の記事であり，その他の記事はトピック”スポーツ”とわかる．検索キーワードのトピックには，一番”スポーツ”を反映させていることがわかる．Topic Sensitive PageRank 手法では，記事 988 を一番下位にランキングしている．

表 2，6，8 から提案手法ですべての記事の重要度に一番”スポーツ”を反映され，記事によって記事の重要度へのトピックの反映度に違いがあり，記事 988 はトピック”スポーツ”，”芸能”，”社会”が高く反映されている．検索キーワードのトピックには”家庭”以外のトピックを反映され，その中でもトピック”スポーツ”，”社会”，”芸能”が高い．提案手法により，記事 988 を上位 4 位にランキングしている．

表 3.2: 単純ベイズと余弦類似度による検索キーワードが各トピックに属する重み

検索キーワード	国際	経済	家庭	文化	科学	芸能	スポーツ	社会
(TSPR) サッカー 日本 代表	4.56E-11	0	0	3.55E-10	0	0	2.44E-08	0
(MYPR) サッカー 日本 代表	0.1326	0.1509	0	0.1324	0.1492	0.1358	0.1399	0.1294
(TSPR) 冬季 五輪	0	0	0	0	0	0	1.07E-06	3.41E-08
(MYPR) 冬季 五輪	0.1388	0.1331	0	0.1222	0.14063	0.15065	0.1591	0.1552

このことから，提案手法は文書の重要度と検索キーワードに複数のトピックを反映させることで，トピックを考慮した重要度による文書ランキングができている．

Topic Sensitive PageRank も検索キーワードに複数のトピックを反映させているが，単純ベイズによりほぼ 1 つのトピックしか記事の重要度に反映されていない．文書の重要度に 1 つのトピックのみ指定させるので，検索キーワード”冬季 五輪”のように正確なトピックの指定やトピックドリフトしている記事の主要なトピックが他の記事の主要なトピックと異なっていれば，トピックドリフトしている記事を下位にランキングできる．しかし，検索キーワード”サッカー 日本 代表”のようにトピックドリフトしている記事が他の記事と同じ主要なトピックのときには，トピックドリフトしている記事を上位にランキングしてしまう．

また，Topic Sensitive PageRank は ODP や種別コードの人手による正確なトピックを指定された Web ページや記事集合以外の文書集合などには適合できない．提案手法では学習データによりどのような文書集合にも適合できる．

3.7 結論

本研究では，トピックを考慮した重要度による文書ランキング手法を提案した．本手法により，検索者の検索意図と文書の主要な話題を文書の重要度に反映させ，トピックドリフトを起こしている文書を下位にランキングすることができた．本稿では新聞記事集合に対してランキング手法を示したが，学習データによりどのうような文書集合にも適合できる．

表 3.3: Topic Sensitive PageRank の各トピックでの記事の PageRank 値 (サッカー 日本 代表)

記事	国際	経済	家庭	文化	科学	芸能	スポーツ	社会
677	0.01628549	0.01543551	0.01642703	0.01434725	0.01414995	0.0158754	0.05786429	0.01602105
694	0.02621034	0.02478334	0.02624309	0.02300665	0.02275057	0.02519013	0.06747065	0.02585367
707	0.0279393	0.02634675	0.02806745	0.02461977	0.02420318	0.02692705	0.06919561	0.02758511
740	0.02331051	0.02204983	0.02318837	0.02055867	0.02012315	0.02235665	0.06482445	0.02300719
748	0.02692198	0.02539224	0.02695318	0.02380562	0.02334913	0.02608048	0.06836908	0.02664311
768	0.02320812	0.02191476	0.02323929	0.02053375	0.0201492	0.02245379	0.06472778	0.02300338
770	0.02416074	0.02280096	0.02406911	0.02120447	0.0209023	0.02313225	0.06550857	0.02380637
819	0.02508562	0.02368776	0.02501543	0.02208009	0.02170155	0.02411804	0.06649766	0.02476353

表 3.4: 提案手法の各トピックでの記事の PageRank 値*記事の各トピックの割合 (サッカー 日本 代表)

記事	国際	経済	家庭	文化	科学	芸能	スポーツ	社会
677	0.000441584	0.000433684	0	0	0.00034025	0.000388661	0.00144729	0.000433972
694	0.000977919	0.000798366	0	0.000398389	0	0.000732579	0.002108727	0.000688847
707	0.001440559	0.001033526	0	0	0.000820688	0.001343557	0.00249171	0.001186773
740	0.000951179	0.000591478	0	0.000516558	0	0.000810154	0.002673234	0.000597504
748	0.00110303	0	0	0.001230758	0.0009621	0.001061925	0.002487807	0.000899565
768	0.001117554	0.000452364	0	0.000600762	0.000757061	0.000582189	0.001677901	0.0005313
770	0.000867845	0.000759111	0	0	0.000448544	0.000682351	0.002246525	0.000590367
819	0.001420582	0.000596109	0	0.000720661	0	0.000958302	0.002508245	0.000889678

表 3.5: Topic Sensitive PageRank の各トピックでの記事の PageRank 値 (冬季 五輪)

記事	国際	経済	家庭	文化	科学	芸能	スポーツ	社会
692	0.02374389	0.02237004	0.02367169	0.02091047	0.02048126	0.02284057	0.0650916	0.02348498
712	0.01978951	0.01869293	0.01978766	0.01746852	0.01709531	0.01899625	0.06135691	0.019642
721	0.02067844	0.01954869	0.02058044	0.0182153	0.01788259	0.01993703	0.06221481	0.02041511
730	0.01943435	0.01839565	0.01966527	0.01715896	0.01693975	0.0189899	0.06083285	0.01911198
742	0.02041298	0.01924279	0.02033717	0.0180109	0.01761391	0.01968302	0.06192326	0.02019657
775	0.00912514	0.00861247	0.00914079	0.00808068	0.00788348	0.00883151	0.050626	0.00898688
782	0.02532507	0.02394848	0.02524539	0.02227661	0.02184676	0.02437671	0.06665876	0.0249562
988	0.02325761	0.02185824	0.02330936	0.02057891	0.02016374	0.02250221	0.02282538	0.06228756

表 3.6: 提案手法の各トピックでの記事の PageRank 値*記事の各トピックの割合 (冬季五輪)

記事	国際	経済	家庭	文化	科学	芸能	スポーツ	社会
692	0.001156051	0.000665472	0	0	0	0.000723613	0.001425498	0.00098955
712	0.001026558	0	0	0	0	0.000909314	0.002742536	0.000951096
721	0.000461732	0	0	0.000635344	0.000684857	0.000524384	0.001514467	0.000426173
730	0.000671832	0.000554855	0	0	0.000478327	0.000645225	0.001602642	0.000567606
742	0	0	0	0.000795937	0.000786941	0.000552131	0.001344852	0.000469283
775	0.000226757	0	0	0	0.000194107	0.000449187	0.00087587	0.000324726
782	0.001061882	0.000595193	0	0.000446785	0.000660955	0.001168546	0.00152643	0.000711374
988	0.000565488	0	0	0.000725272	0	0.001175026	0.001189103	0.000974241

表 3.7: 検索キーワード”日本 サッカー”の正解データ

TD	記事	第 1 段落
	677	サッカーのアジア・カップ最終予選イエメン戦を控える日本代表で、大学生ＦＷがフル代表デビューの機会をうかがっている。「裏に抜け出すスピードをアピールしたい」と話す福岡大３年の永井謙佑には、何としても試合に出たい理由がある。
×	694	日本サッカー協会の犬飼基昭会長は１１日、アフリカ選手権への視察スタッフ派遣を見送ったことを明かした。トーゴ代表を乗せたバスが武装集団の襲撃を受けたことに対応した。２０１８、２２年W杯の招致活動のため、出張予定だった田嶋幸三専務理事のアンゴラ入りも見合わせた。
	707	日本サッカー協会は１４日の理事会で、日本女子代表の強化指定選手制度を４月から導入することを決めた。１１年W杯ドイツ大会、１２年ロンドン五輪で活躍が期待できる選手の海外リーグ挑戦を後押しすることが目的。対象は日本女子代表選手で１年ごとに更新し、第１回は５人程度、来年以降は１０人程度を選定する。
	740	サッカー日本代表は２日、国際親善試合のキリンチャレンジカップでベネズエラと大分・九州石油ドームで対戦する。１月６日のアジアカップ最終予選イエメン戦（サヌア）に若手主体で臨んだ日本にとって、国内組の主力を集めた今回は、ワールドカップ（W杯）イヤーの実質的な初戦となる。
	748	サッカーの男子日本代表は４日、東アジア選手権（６～１４日、東京）に向けて２３人全員が参加して約１時間半、千葉県内で練習した。このうち約５０分間を非公開とし、メンバーを入れ替えながら紅白戦を行った。
	768	サッカーの男子日本代表は１４日、最終戦の韓国戦（１９時１５分、東京・国立競技場）に臨む。日本は中国と並んで勝ち点４、韓国は同３、香港は同０。先に行われる中国―香港戦で中国が勝つと、日本の大会初優勝は勝利が絶対条件となる。
	770	日本サッカー協会の犬飼基昭会長と原博実技術委員長が１５日、東アジア選手権で３位に終わった日本代表の岡田武史監督を協会に呼んで緊急会談を行い、岡田監督を全面支援することを確認した。
	819	サッカー日本代表の岡田武史監督は１６日、６月のワールドカップ（W杯）南アフリカ大会に、登録メンバー以外の若手４～５人を「育成枠」として帯同させる方針を明らかにした。５月中旬、日本代表２３人と同時に発表を行い、５月２４日のキリンチャレンジカップ・韓国戦（埼玉スタジアム）以降、行動を共にするという。

表 3.8: 検索キーワード”冬季 五輪”の正解データ

TD	記事	第1段落
	692	ジャンプのW杯創設に尽力し、長野冬季五輪で同種目の技術代表を務めたノルウェー人のトールビョルン・イグセト氏が１０日、死去した。７５歳だった。国際スポーツ記者協会（ＡＩＰＳ）によると死因は前立腺がん。
	712	スケルトンのW杯は１５日、スイスのサンモリッツで第７戦を行い、男子でバンクーバー冬季五輪出場が確実な４５歳のベテラン、越和宏（システックス）は1分9秒８８で１４位だった。田山真輔（システックス）は1分１０秒３６で２３位。コース状態に不備があったために1回目の結果が取り消され、2回目だけで争われた。エリック・バーノタス（米国）が1分9秒１５で優勝した。
	721	１７日、フランスのショヌーブで個人第１３戦を行い、バンクーバー冬季五輪代表で臨んだ日本は小林範仁（東京美装）が8位に入り、加藤大平（サッポロノルディッククク）が9位で続いた。小林は前半飛躍（ＨＳ１００メートル、K点９０メートル）で２０位につけ、後半距離（１０キロ）で巻き返した。飛躍4位の加藤は距離も粘った。
	730	バンクーバー冬季五輪の開幕まで2週間を切った。前回トリノ大会で日本選手団が獲得したメダルは、フィギュアスケート女子・荒川静香の「金」1個だけ。金5個を含むメダル１０個を得た９８年長野大会を頂点に、日本勢は低落傾向が続く。バンクーバーで歯止めはかかるのか。世界の強豪に挑戦する、日本のエースたちを紹介する。
	742	3日、ドイツのクリンゲンタールで個人第１８戦（ＨＳ１４０メートル、K点１２５メートル）を行い、バンクーバー冬季五輪代表を逃した湯本史寿（東京美装）が１１９・5メートル、１２４メートルの２１３・3点で１２位になった。
	775	男子モーグルの現場で、記念すべき瞬間に立ち会った。地元のアレクサンドル・ビロドーが、冬季夏季含め3度目となる自国開催の五輪で、カナダに初めての金メダルをもたらした熱戦だ。
	782	２０１８年冬季五輪に招致を表明しているアヌシー（フランス）、ミュンヘン（ドイツ）、平昌（ピョンチャン、韓国）の3都市が１５日、それぞれバンクーバー市内で記者会見を開き、6月の国際オリンピック委員会（ＩＯＣ）理事会での1次選考に向けてアピールをした。
×	988	勤務先のスケート教室の雇用を打ち切られたのは無効として、インストラクターの坂田清治さん（６２）＝横浜市＝が5日、教室を運営する財団法人「神奈川体育館」を相手に、雇用の確認を求める仮処分を横浜地裁に申し立てた。坂田さんは日本スケート連盟強化コーチや長野五輪の強化コーチなどを歴任。浅田真央選手らの靴の刃を調整したことなどで知られ、バンクーバー冬季五輪では金妍児（キムヨナ）選手の靴を調整した。

第4章　トピックグラフを用いた相対的文書ランキング

近年，インターネットの利用環境は多様化している．ユーザが必ずしも効果的なクエリを入力できない場合にも的確な検索結果を返すためには，クエリ拡張が有効であるが，履歴情報の不足による精度の低下が問題となる．また，ユーザによるクエリの修正を支援するには，複数のトピックにまたがる検索結果の相対的なランキングも重要である．本研究では，履歴情報を用いずに多面的なクエリ拡張を行い，得られたトピックに基づいて文書を相対的にランキングする手法を提案する．具体的には，入力クエリに関連する複数のトピックを抽出して拡張クエリを生成し，その検索結果から構築するトピックグラフを用いて文書に相対的な重みを付与する．

4.1　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．

4.2 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．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．

4.3 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．以下の式 1 で表される．

$$L = \sum_{u=1}^{m} \sqrt{P_u Q_u}$$

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．以下の式 2 でで表される．

$$d_q = \sum_{n=1}^{N} \frac{tf_i * Bha_i}{N}$$

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，$Bha_q i$ 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．また，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．

4.4 実験

4.4.1 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．

4.4.2 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 4.1 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 4.2 に，さらに“ 大相撲 大関 両国国技館 ”したときの Bhattacharyya 係数を表 4.3 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 4.4 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 4.1: クエリ ”大相撲 ”での Bhattacharyya 係数

Bhattacharyya 係数	記事 ID	拡張語
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 4.2: クエリ ”大相撲 大関 ”での Bhattacharyya 係数 (クエリ拡張)

Bhattacharyya 係数	記事 ID	拡張語
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
1.109182	3057	師匠

4.5 考察

実験結果から，提案手法について考察を行う．

表 4.1，表 4.2，表 4.3 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 4.3 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

表 4.3: クエリ ”大相撲 大関 両国国技館 ”での Bhattacharyya 係数 (クエリ拡張)

Bhattacharyya 係数	記事 ID	拡張語
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

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 4.4 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，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．

4.6 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．

表 4.4: クエリ“ 大相撲　大関　両国国技館 ”にマッチした上位 5 文書

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。

第5章　適合性フィードバックに基づく検索意図を反映させた文書ランキング

5.1　前書き

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．しかし，ユーザは検索対象の文書集合について知らないので，どのようなクエリを与えれば，適切な検索結果が返されるかわからないため，ほとんどのユーザは短いクエリや曖昧なクエリを与えてしまう．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．これは, 検索キーワードが同音異義語にある場合に起こる．例えば, 検索者が”マウス”で検索したときにコンピュータのポインティングデバイスのマウスについての文書を検索したいが, ネズミや口の”マウス”の出現頻度などが高いために, 検索結果の上位に検索意図とは関係ない文書が上位にランキングされてしまう．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．

このような問題を解決するための手法は多く提案されている．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．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，ユーザが与えたクエリで得られた検索結果からどの文書が適合・不適合をシステムに学習させることで，ユーザにとって有用なクエリに改善する「適合性フィー

ドバック」がある．適合性フィードバックは，3つのタイプに分けることができる．明示的フィードバックは，あるクエリによって検索された初期結果から適合文書と不適合文書をユーザに選択してもらうことで，クエリを修正する．暗示的フィードバックは，検索結果の文書をユーザがどの文書を閲覧したか，閲覧時間，ブラウジング，スクローリングなどの行動から適合情報を抽出する．疑似フィードバックは，あるクエリによって検索された文書ランキングの上位数件を適合文書であると仮定することで，クエリを修正する．

また，確率モデルやクラスタリングを用いた研究もあり，文書が有する文脈をトピックモデルを用いて推定し、同時にクエリの意図を推定して適合させる研究やクラスタを作成し，そのクラスタを用いて適合性フィードバックを行う研究などもある．

しかし，適合文書や不適合文書の選択による負担，システムの重要度とユーザの重要度が不一致などで検索結果を改善するのにはユーザに大きく負担がかかってしまう．また，初期検索の性能に大きく依存してしまうという問題や予め決められた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い．

本研究では，検索結果からユーザの意図を抽出しやすいように，クエリに関連する複数のトピックを含む文書を上位にランキングさせる．その文書を上位にランキングさせることで．ユーザの検索意図にあった情報を早い段階で入手させ，またユーザがフィードバックする負担を減らし，検索結果のリランキングを行うことで検索結果の精度を向上させる．また，提案手法の有効性を示すため，Rocchio アルゴリズムとの比較を行う．

5.2 Rocchio アルゴリズム

ベクトル空間モデルとは，大量の情報の中から利用者が与えた質問に対して利用者が必要と思われる文書の集合を提示するデータ表現方法である．

データ記述では文書を単語の多重集合で表現する．単語の並びを無視し，文書を多次元ベクトルとして表す．このモデルで重要な語を扱うには十分である．位置情報を失うため精密な表現ではないが，意味内容を表現しない語や記号などを無視できる．文書 d をベクトル $d =< v_1, ..., v_n >$ で表すとき $i = 1, ..., n$ は予め定まった語 w_i を表し，v_i はその語に与える重みを意味する．重みとしては，2値や出現頻度，また TF*IDF 値が用いられることが多い．

またベクトル空間モデルでは，情報検索操作を自然に表現することができる．実際，ベクトルの各要素は語の重みを表すため，当該部の要素を対応させればよい．データベース検索と異なり，情報検索では，完全一致解よりも，質問に多く関連する文書を探索することが要求される．質問に類似する度合いを数値で表現し，この度合い順に文書をランキングして提示する．

例えば，与えられた質問をベクトル化した質問ベクトル q と文書の単語をベクトル化した多次元の文書ベクトル d_i のなす角度により類似度を計算する余弦類似度がある．余弦類似度を $cos(d_i, q)$ とすると，$cos(d_i, q)$ は次式で表せる．

$$cos(d_i, q) = \frac{d_i * q}{|d_i||q|}$$

質問ベクトル q と全ての文書ベクトル $d_1, d_2 \cdots d_n$ の余弦類似度を計算し，余弦類似度の大きいほど質問に似た文書であることがわかる．

Rocchio アルゴリズム [15] は，ベクトル空間モデルを用いて適合性フィードバックを行い，クエリベクトルを更新することで検索結果の適合率を上げる手法である．例えば，ユーザが与えたクエリベクトル q に対する文書ランキングの中に含まれる適合文書集合を D_r，不適合文書集合 D_n とする．ユーザは下の式を用いて，ユーザが適合と判断した文書に含まれる語の重みを出現回数や TF*IDF 法などにより大きくし，逆に不適合文書に含まれる語の重みは小さくすることで，クエリベクトル q を q' に修正する．

$$q' = \alpha\ q + \frac{\beta}{|D_R|} \sum_{d_i \in D_R} d_i - \frac{\gamma}{|D_N|} \sum_{d_j \in D_N} d_j$$

上式で，α，β，γ は重み付け係数であり，$|D_R|$ および $|D_N|$ は，それぞれの文書集合に含まれる文書数である．

5.3 提案手法

5.3.1 単語の関連度

本研究では，検索結果からユーザの意図を抽出しやすいように，クエリに関連する複数のトピックを含む文書を上位にランキングさせるため，クエリと各単語との関連度を抽出する．この関連度が高い単語を多く含む文書は，クエリに関連する複数のトピックを含む文書となる．

各単語の 1 日の文書頻度をヒストグラムで表し，各ヒストグラム間の類似度を計算する Bhattacharyya 係数 [?] を用いることで、各単語間の関連度を抽出する．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．

$$Bha = \sum_{u=1}^{m} \sqrt{P_u Q_u} (0 \leq Bha \leq 1)$$

ここで，p，q は比較対象となる正規化された $(\sum_{u=1}^{m} P_u = \sum_{u=1}^{m} Q_u = 1)$ ヒストグラムを用いて算出する．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にある．

しかし，クエリと共起性の高い単語は，ヒストグラムが類似しているため，Bhattacharyya 係数では関連語になってしまうので，共起度の逆数をかけることで，クエリ q と関連のあるトピックに含まれる単語との関連度は高くなる．クエリ q と単語 i の関連度 BC_{iq} は，次式で表される．

$$BC_{iq} = Bha_{iq} * log(\frac{1}{CO_{iq}})$$

ここで，Bha_{iq} はクエリ q と単語 i の Bhattacharyya 係数，CO_{iq} はクエリ q と単語 i の共起度である．クエリが 2 個以上のときは，Bha と CO はそれぞれ掛け合わせて用いる．また，縦軸を Bha，横軸を単語にして Bha が左側から高い順に並べたとき，曲線の傾きが急になる単語の Bha 以上の単語のみをクエリとの関連語として用いる．CO_{iq} は次式で表される．

$$CO_{iq} = \frac{c}{a+b-c}$$

ここで，a は単語 i を含む文書数，b はクエリ q を含む文書数，c は単語 i とクエリ q を含む文書数である．

5.3.2 フィードバック

本研究ではユーザがクエリ q を与えたとき，クエリ q と各単語の関連度 BC を用いて全文書にランク値 dr を与え，検索結果の上位 N 件を用いて適合性フィードバックを行う．

ユーザが閲覧した文書を適合文書，閲覧しなかった文書を不適合文書と見なし，単語の関連度を修正していく．そして．新しい単語の関連度を用いてリランキングを行う．単語の関連度の更新式は，次式で表される．

$$BC_{iq,n+1} = BC_{iq,n}(1 + \frac{w_{i,D^+}}{|N^+|} - \frac{w_{i,D^-}}{|N^-|})$$

$BC_{i,n}$ は，n 回目のクエリ q と単語 i の関連度である．N^-，N^+ は n 回目の検索での適合文書数，不適合文書数であり，w_{i,D^+}，w_{i,D^-} は適合文書集合での単語 i の出現頻度，不適合文書集合での単語 i の出現頻度を示している．n+1 回目の文書 j のランク値 $dr_{j,n+1}$ は，次式で表される．

$$dr_{j,n+1} = \frac{\sum_{n=1}^{N} tf_{i,j} * BC_{iq,n+1}}{N}$$

$tf_{i,j}$ は文書 j での単語 i の出現頻度，N は文書 j の単語数である．また，文書 j の共起度 CO_j が閾値以下の文書はランキングから除外する．式は以下で表される．

$$CO_j = \frac{\sum_{n=1}^{N} tf_{i,j} * CO_{iq}}{N}$$

5.3.3 手順

本研究の適合性フィードバックの一連の流れを説明する．

1. ユーザがクエリをシステムに与えるシステムはクエリ q と単語 i の関連度 BC_{iq} を用いて，クエリに関連する複数のトピックを含む文書を上位にランキングさせる．
2. ユーザは検索結果の上位 N 件から適合性フィードバックを行うクエリに関連する複数のトピックを含む文書含む検索結果からフィードバック行うため，検索意図の曖昧なクエリからも検索意図を抽出できる．システムはユーザのフィードバックから $BC_{i,n}$ を算出し，文書集合の特徴をユーザの好みに訂正する．
3. ユーザにリランキングした検索結果を与える

この一連の流れからユーザの検索意図にあった情報を早い段階で入手させ，ユーザの負担を減らし，検索結果のリランキングを行うことで検索結果の精度を向上させる．

5.4 実験

5.4.1 実験方法

実験では，新聞記事集合に対して提案手法を行う．

毎日新聞記事データ集 2010 年版から，1 面から 3 面に掲載された記事 1 年分，7,999 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．

5.4.2 評価方法

評価には，平均適合率を用いて Rocchio アルゴリズムとの比較を行う．3 つのパラメータ α，β，γ は，一般的な α =1.0，β =0.8，γ =0.1 を使用する．初期検索の検索結果の上位 7 件に対して、適合不適合の判定を行う．適合不適合の判定は，ユーザの検索意図に合っている単語が、タイトルや第 1 段落に含まれていれば適合文書とする．例えば．クエリ”中国漁船衝突事件”のみ与えたユーザが”ビデオ流出事件について知りたい” という検索意図だったときは．”ビデオ流出”，”ビデオ映像 流出”，”ビデオ映像 流出”などが含まれていれば適合文書になる．

5.4.3 実験結果

表 1 は Rocchio アルゴリズムと提案手法で，3 回フィードバックを行った時の top10 での適合文書数を表している．
また，クエリの横のカッコ内はユーザの検索意図，R) は Rocchio アルゴリズムの結果を表している，3 回目ののフィードバック時では，5 クエリ中 3 クエリで提案手法が Rocchio アルゴリズムより高い適合率を算出している．

5.4.4 考察

実験結果から，提案手法について考察を行う．
本研究では，検索結果からユーザの意図を抽出しやすいように，クエリに関連する複数のトピックを含む文書が上位にランキングさせたが，表 3 からわかるように．記事 7815 はクエリ ”中国漁船衝突事件 ”に関連する複数のトピックを含む文書が上位にランキングされている．表 2 の Rocchio アルゴリズムの初期検索結果では．7 文書中 5 文書が ”ビデオ流出 ”についての文書のため，ユーザが ”ビデオ流出 ”以外のトピックについて検索する場合は，時間がかかってしまう．
実際に，表 1 でのクエリ ”中国漁船衝突事件 (中国人船長) ”では, 適合文書が 1 つしかないのと不適合文書が ”ビデオ流出 ”についての文書が多いために，”ビデオ流出 ”に不適合情報はシステム側に大きく反映されたが, その他の不適合情報が反映されないために，ユーザの検索意図がシステムに反映されるのに時間がかかってしまっている．
表 4 と表 5 からわかるように Rocchio アルゴリズムは余弦ベクトルで文書を検索するために，文書の単語数が少ない場合やクエリの出現頻度が文書内で高いなどが原因でクエリ”民主 大敗”を与えても，”民主 ”のみや ”大敗 ”のみを含む文書でもランキングの上位に順位付けしてしまう．そのため，適合文書が上位にないためにフィードバックが出来ない．
提案手法は．クエリと関連度の高い単語を用いて文書ランキングを行うので，トピック ”参院選挙で民主党が大敗 ”に含まれる単語はクエリと高い関連があるので，このトピックに関する文書が上位にランキング出来ている，また．仮にこのトピックに関連する文書がクエリの単語を含んでいなくても，時系列に単語間の関連性を算出しているので，表 6 のようにフィードバックを繰り返すことでトピックに関連する文書はランキング上位に上位にランクされる．

5.5 結論

本研究では，検索結果からユーザの意図を抽出しやすいように，クエリに関連する複数のトピックを含む文書を上位にランキングさせることで，ユーザの検索意図にあった情報を早い段階で入手させ，またユーザがフィードバックする負担を減らし，検索

結果のリランキングを行うことで検索結果の精度を向上させた．また，時系列に単語間の関連性も求めているので，クエリを含んでいなくても検索意図に合った文書を上位にランキングできる．

表 5.1: 提案手法と Rocchio の適合文書数

クエリ（検索意図）	FB[0]	FB[1]	FB[2]	FB[3]
R) 中国漁船衝突事件（ビデオ流出）	8	9	10	10
R) 中国漁船衝突事件（仙谷由人官房長官）	2	7	7	7
R) 中国漁船衝突事件（中国人船長）	1	3	1	1
R) 民主　大敗 (参院選挙で民主党が大敗)	0	0	0	0
R) 郵便不正事件 (虚偽有印公文書作成・同行使罪に問われた事件)	3	9	9	10
中国漁船衝突事件（ビデオ流出）	5	8	8	10
中国漁船衝突事件（仙谷由人官房長官）	1	5	6	10
中国漁船衝突事件（中国人船長）	1	1	1	2
民主　大敗 (参院選挙で民主党が大敗)	2	3	4	4
郵便不正事件 (虚偽有印公文書作成・同行使罪に問われた事件)	5	10	10	10

表 5.2: Rocchio アルゴリズムのクエリ”中国漁船衝突事件”の初期検索結果

記事	タイトル	本文
6776	中国漁船・尖閣領海内接触：ビデオ流出　菅首相、原因究明指示	菅直人首相は 5 日午前の閣僚懇談会で、中国漁船衝突事件の映像の流出問題について、「しっかり調査して、原因究明しなければならない」と情報管理態勢の点検を指示した。
6812	中国漁船・尖閣領海内接触：ビデオ流出　鳩山前首相「クーデターだ」	民主党の鳩山由紀夫前首相は 6 日、中国漁船衝突事件の映像流出問題について「情報流出によるクーデターのようなことを政府内の人間が行うのは政権にとって大変厳しい話だ」と述べた。
6816	中国漁船・尖閣領海内接触：ビデオ流出　鳩山前首相「クーデターだ」　【大阪】	民主党の鳩山由紀夫前首相は 6 日、佐賀市で講演し、中国漁船衝突事件の映像流出問題について「情報流出によるクーデターのようなことを政府内の人間が行うのは政権にとって大変厳しい話だ」と述べた。…
5728	中国漁船・尖閣領海内接触：SMAP 上海公演のチケット販売停止	【上海・鈴木玲子】中国漁船衝突事件で日中関係が悪化する中、上海で 10 月 9、10 の両日に開催される予定の人気グループ「SMAP」のコンサートのチケット販売が停止されていたことが 20 日分かった。…
6393	中国漁船・尖閣領海内接触：船長処分、国際関係も考慮　政府が答弁書	政府は 19 日、沖縄県・尖閣諸島周辺で起きた中国漁船衝突事件に関し、容疑者を起訴するかどうかの刑事処分を検察官が判断するに当たっては「国際関係への影響などについても、犯罪後の状況として考慮できる」との答弁書を閣議決定した。…
6832	中国漁船・尖閣領海内接触：ビデオ流出　馬淵国交相、調査状況を仙谷長官に報告　【大阪】	仙谷由人官房長官は 7 日、首相官邸で馬淵澄夫国土交通相と会い、沖縄県・尖閣諸島沖での中国漁船衝突事件に関する映像流出問題の調査状況について報告を受けた。, 会談後、馬淵氏は記者団に「調査の報告をした」と話した。
6765	中国漁船・尖閣領海内接触：衝突映像、ネット流出か	沖縄県・尖閣諸島沖の中国漁船衝突事件のビデオ映像とみられる映像が、インターネット上に流出していることが分かった。映像は数種類あり、海上保安庁の巡視艇に、中国漁船が衝突している場面などが映っていた。海保は「コメントできない」としている。

表 5.3: 提案手法のクエリ”中国漁船衝突事件”の初期検索結果

記事	タイトル	本文
6765	中国漁船・尖閣領海内接触：衝突映像、ネット流出か	沖縄県・尖閣諸島沖の中国漁船衝突事件のビデオ映像とみられる映像が、インターネット上に流出していることが分かった。映像は数種類あり、海上保安庁の巡視艇に、中国漁船が衝突している場面などが映っていた。海保は「コメントできない」としている。
7815	ことば：中国漁船衝突事件	◇中国漁船衝突事件, 沖縄・尖閣諸島付近で 9 月 7 日、中国漁船が海上保安庁の巡視船に衝突。石垣海上保安部が公務執行妨害の疑いで中国人船長を逮捕した。中国側は反発し、船長の釈放を要求。…
6776	中国漁船・尖閣領海内接触：ビデオ流出　菅首相、原因究明指示	菅直人首相は 5 日午前の閣僚懇談会で、中国漁船衝突事件の映像の流出問題について、「しっかり調査して、原因究明しなければならない」と情報管理態勢の点検を指示した。
6832	中国漁船・尖閣領海内接触：ビデオ流出　馬淵国交相、調査状況を仙谷長官に報告　【大阪】	仙谷由人官房長官は 7 日、首相官邸で馬淵澄夫国土交通相と会い、沖縄県・尖閣諸島沖での中国漁船衝突事件に関する映像流出問題の調査状況について報告を受けた。, 会談後、馬淵氏は記者団に「調査の報告をした」と話した。
5773	中国漁船・尖閣領海内接触:「機会があれば中国側に説明」??NY で前原外相	【ニューヨーク山科武司】中国漁船と海上保安庁巡視船の衝突事件について、前原誠司外相は 21 日、「機会があれば中国側に日本の立場を説明したい」と語った。国連総会のため訪問したニューヨークで記者団に語った。…
6055	菅首相：尖閣は固有領土、説明の意向 ASEM に出発	菅直人首相は 3 日、ブリュッセルで開かれるアジア欧州会議（ASEM）首脳会議に出席するため、政府専用機で羽田空港を出発した。沖縄県・尖閣諸島周辺での中国漁船衝突事件をめぐって首相は 2 国間会談の場で、尖閣諸島は日本固有の領土とする立場を説明する意向だ。…
6923	中国漁船・尖閣領海内接触：ビデオ流出　中国も関係修復を模索 首脳会談へ努力要求	【北京・成沢健一】中国外務省の洪磊・副報道局長は 11 日の定例会見で、尖閣諸島沖衝突事件の映像を流出させたと神戸の海上保安官が認めていることについて、「報道を注視している。中国側は、いわゆるビデオ問題が中日関係を妨げ続けることを望まない」と述べた。…

表 5.4: Rocchio アルゴリズムのクエリ”民主　大敗”の初期検索結果

記事	タイトル	本文
1663	日米密約：首相経験者らに国会出席呼びかけへ	民主、社民、国民新の与党 3 党の国対委員長は 10 日、国会内で会談し、日米両政府による外交密約の存在が明らかになったことを受け、歴代首相経験者や外相経験者に対し国会への出席を呼びかける方針で一致した。…
3686	小沢・民主前幹事長：辞任タイミング、参院選前「ぎりぎりセーフ」	民主党の小沢一郎前幹事長は 12 日午前、連合和歌山の村上正次会長らと会談し、幹事長辞任について「もう少し早ければとも思ったが、鳩山由紀夫前首相と話し合ってああいうタイミングになった」と説明。…
7551	11 年度予算：普天間予算、見送り要求??社民	社民党は 10 日午前、首相官邸で開かれた 11 年度予算編成を巡る民主、国民新両党との協議で、米軍普天間飛行場（沖縄県宜野湾市）の名護市辺野古崎地区への移転に関し、環境影響評価など関連経費の計上をすべて見送るよう求めた。…
1020	亀井金融・郵政担当相：「夫婦別姓」にも反対表明	国民新党代表の亀井静香金融・郵政担当相は 11 日、福井市内で講演し、民主、社民両党が導入を目指す永住外国人の地方選挙権と選択的夫婦別姓に反対する考えを示した。， 亀井氏は夫婦別姓について、「一家の表札が（複数になって）アパートみたいになってしまう」と批判。「（この二つは）国民新党が反対するので、絶対に成立しない。法案提出も私がノーと言えばできない」と話した。【大久保陽一】
1197	日本航空：過去の問題点を検証、民主が PT 発足　24 日にも初会合	民主党は 18 日、経営破綻（はたん）した日本航空の過去の問題点などを検証するプロジェクトチーム（PT）を発足させることを決めた。24 日にも初会合を開く。自民党政権の実質的な保護下で粉飾決算がなかったかという問題や、空港建設に与党政治家がどう関与したかなどについても調査する。…
836	基幹労連:政治と金、民主の対応批判??内藤委員長	民主党を支持している「基幹労連」の内藤純朗委員長は 3 日、東京都内で開いた労連の中央委員会で「民主党の対応に一抹の不安と不信を抱くのは私だけではない」と、同党の政治と金にまつわる対応を批判した。…
1188	10 年度予算案：年度内成立へ	衆院は 18 日の本会議で、自民党が提出した鹿野道彦予算委員長（民主）の解任決議案を与党 3 党の反対多数で否決した。公明党とみんなの党は賛成し共産党は棄権した。予算委は 19 日に地方公聴会、24 日に中央公聴会が開かれることが決まり、…

表 5.5: 提案手法のクエリ”民主　大敗”の初期検索結果

記事	タイトル	本文
3686	小沢・民主前幹事長：辞任タイミング、参院選前「ぎりぎりセーフ」	民主党の小沢一郎前幹事長は 12 日午前、連合和歌山の村上正次会長らと会談し、幹事長辞任について「もう少し早ければとも思ったが、鳩山由紀夫前首相と話し合ってああいうタイミングになった」と説明。参院選を念頭に「ぎりぎりセーフかなと思う」と述べた。…
4500	菅首相：小林・宮台両教授招き、参院選大敗の原因を聞く	◇励まされたり、反省したり，　菅直人首相は 25 日、首相公邸に小林良彰慶応大教授（政治学）と宮台真司首都大学東京教授（社会学）を招き、参院選で民主党が大敗した原因について意見を聞いた。…
4731	菅首相：消費税率「10％」事実上撤回　「党政調で」	菅直人首相は 4 日夜、自らの「消費税率 10 ％」発言について記者団に「超党派の協議を呼びかけた中で自民党の案に触れたことがいろいろと誤解を生んだ。今は（民主）党の議論を待っている」と語った。…
4268	菅首相：消費税争点化で党に迷惑かけた??稲盛氏に	菅直人首相は 14 日午前、首相官邸で内閣特別顧問の稲盛和夫日航会長と会談し、参院選敗北について「唐突に消費税問題を挙げたからこういう結果になり、（民主）党全体に迷惑をかけた。大変反省している」と語った。…
194	山口・公明代表：「民主と連携、検討も」　参院選後へ秋波、選挙協力にも含み	公明党の山口那津男代表＝似顔絵＝は 10 日、NHK の報道番組で、次期参院選後の民主党との連携について「選挙結果に応じて対応を検討するのが国民への責任だ。国民のニーズを実現する視点で選択したい」と述べ、…
4616	臨時国会：ヤワラちゃん、笑顔の初登院【大阪】	参院選で 35 万票余りを獲得して当選した柔道家、谷亮子氏（34）＝民主、比例＝は 30 日午前 8 時 40 分ごろ、ベージュのスーツ姿で初登院。多くの報道陣が集まり、警備にあたる衛視らともみ合いになる場面も。柔道との両立が可能か議論を呼んでいるが、意気込みを問われた谷氏は「公務をおろそかにすることなく務めたい」と話した。
3591	荒井国家戦略相：事務所費問題　自民・石破氏「辞任に値する」	自民党の石破茂政調会長は 9 日の記者会見で、荒井聡国家戦略担当相の事務所費問題に関し、「当然、辞任に値する。菅直人首相の任命責任も問われる」と辞任を求めた。

表 5.6: 提案手法のクエリ”民主　大敗”のフィードバック 2 回目の検索結果

記事	タイトル	本文
4500	菅首相：小林・宮台両教授招き、参院選大敗の原因を聞く	◇励まされたり、反省したり，　菅直人首相は25日、首相公邸に小林良彰慶応大教授（政治学）と宮台真司首都大学東京教授（社会学）を招き、参院選で民主党が大敗した原因について意見を聞いた。…
4268	菅首相：消費税争点化で党に迷惑かけた??稲盛氏に	菅直人首相は14日午前、首相官邸で内閣特別顧問の稲盛和夫日航会長と会談し、参院選敗北について「唐突に消費税問題を挙げたからこういう結果になり、（民主）党全体に迷惑をかけた。大変反省している」と語った。稲盛氏が記者団に明らかにした。首相は「政権運営は大変難しくなる。今後、慎重にする」とも語ったという。…
4320	菅首相：「任期3年、続投」　意欲示す	菅直人首相は16日夜、東京都内のホテルで法政大学の五十嵐敬喜、江橋崇両教授と会食し、首相続投に強い意欲を示した。五十嵐氏によると、同氏らが「（衆院の残る任期の）3年間、絶対に辞めないでください」と求め、首相は「任期まで全部やる」と述べ、今後の政権運営に強い意欲を示したという。…
4731	菅首相：消費税率「10％」事実上撤回　「党政調で」	菅直人首相は4日夜、自らの「消費税率10％」発言について記者団に「超党派の協議を呼びかけた中で自民党の案に触れたことがいろいろと誤解を生んだ。…
3686	小沢・民主前幹事長：辞任タイミング、参院選前「ぎりぎりセーフ」	民主党の小沢一郎前幹事長は12日午前、連合和歌山の村上正次会長らと会談し、幹事長辞任について「もう少し早ければとも思ったが、鳩山由紀夫前首相と話し合ってああいうタイミングになった」と説明。参院選を念頭に「ぎりぎりセーフかなと思う」と述べた。…
4351	ファイル：民主・細野氏も予算委に前向き	民主党の細野豪志幹事長代理は18日のフジテレビ番組で、臨時国会について「野党側が求める予算委員会は少なくともやらないといけない」と述べ、参院で多数を握る野党側の主張を取り入れるべきだとの考えを示した。…
4235	枝野・民主幹事長：消費増税、年度内取りまとめ断念示唆	民主党の枝野幸男幹事長は12日の会見で、消費税増税に関する具体案の取りまとめについて「当初想定していた期限にこだわるのではなく、幅広い国民の理解と合意を得られるペースで進めることを基本に考えないといけない」と述べた。…

第6章　結論

本研究では, ユーザの検索意図を反映させた文書ランキングを行うアルゴリズムの提案を行った.

多くのランキングアルゴリズムでは，Web ページや文書の集合から抽出した特徴（重要度，トピック）やクエリなどから抽出したユーザの検索意図などを用いてユーザの検索意図に近い文書をランキングするが, ユーザは検索システムに対して数語程度のキーワード入力しか行なわないため，検索意図を抽出するのは難しい.

この問題に対し, 文書特性に基づく重要な文書, トピックを考慮した重要度, クエリ拡張や適合性フィードバックを用いた検索意図の抽出する手法を行い, 実験によりその有効性を示した.

これにより, ユーザが与えたクエリから本手法のクエリ拡張や適合性フィードバックにより, 高速に検索意図を抽出を出来ることができ, さらに抽出した検索意図を用いて文書集合の特徴をユーザの好みに変形させた文書集合から重要度を本手法で算出することで, ユーザの検索意図に合致する文書を上位にランキングすることが可能になったと言える.

謝辞

本研究を遂行するにあたり，日頃より適切な御指導，御鞭撻をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます．

並びに，法政大学マイクロ・ナノテクノロジー研究センター 島田 諭博士研究員にも多くの有益なご助言をいただきました．深く感謝いたします．

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適で能率的な研究室環境を整えていただきました．御礼申し上げます．

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます．

最後に，今までの学生生活を支えてくださった私の両親に深く感謝したいと思います．

参考文献

[1] Hatakenaka, S. and Miura, T.: Ranking Documents using Similarity-based PageRanks, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing(PacRim), 2011.

[2] Hatakenaka, S. and Miura, T.: Query and Topic Sensitive PageRank for General Documents, 14th IEEE International Symposium on Web Systems Evolution(WSE), 2012.

[3] 畠中翔太，三浦孝夫，三浦孝夫: トピックグラフを用いた相対的文書ランキング, 第 11 回 FIT（情報科学技術フォーラム） , 2012.

[4] 畠中翔太，三浦孝夫，三浦孝夫: 適合性フィードバックに基ずく検索意図を反映させた文書ランキング, 第5回データ工学と情報マネジメントに関するフォーラム (DEIM), 2013.

[5] S. Brin and L. Page The anatomy of a large scale hypertextual Web search engine. *ComputerNetworks and ISDN Systems* , 30 , 107-117 , 1998.

[6] Oren. Kurland and Lillian Lee PageRank without hyperlinks: Structural reranking using links induced by language models. *Proceedings of the 28th annual international ACM SIGIR* , 2005.

[7] T. H. Haveliwala Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web* , 2002.

[8] Amy N. Langville, Carl D. Meyer, 岩野 和生, 黒川 利明, 黒川 洋:Google PageRank の数理，共立出版，2009.

[9] J. Kleinberg Bursty and hierarchical structure in streams. *Proc. 8th SIGKDD*,2002, 91-101.

[10] 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* , 2007.

[11] Hang Cui, Ji-RongWen, Jian-Yun Nie andWei-YingMa, Probabilistic Query Expansion Using Quer Logs *Proceedingsof the 11th international conference on World Wide Web* 2002, 325-332

[12] Georges E. Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations *ACM SIGIR*2008.331-338,

[13] KMamoru Komachi, Shimpei Makimoto, Kei Uchiumi, and Manabu Sassano. Learning semanticcategories from clickthrough logs *ACLIJCNLP* 2009.189-192,

[14] 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,

[15] Rocchio, J.J Relevance fssdback in information retrieval. *The SMART Retrieval Systems* , pp313-323 , Prentice-Hall , 1971.