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

2008－NL－187 (12)
2008／9／25

# トピック関連語の推定と文書ランキングへの適用

石川 浩一郎[†1] 近藤 真史[†1] 櫻井 彰人[†1]

文書集合中から特定のトピックに関する文書を検索する際，ユーザが入力する検索キーワード（トピック語）を用いて，トピックに関連がある語を推定し，この語を用いて，トピック語を必ずしも含まないがトピックに関連する文書をランキングする方法を提案する．トピックと関連ある語の推定にはアソシエーション分析で用いるリフト値を応用する．被験者による実験により，提案手法の有効性が確認された．

# Ranking Documents by Topic-Related Words

Koichiro Ishikawa,[†1] Masafumi Kondo[†1] and Akito Sakurai[†1]

We propose a method to rank documents based on a search keyword (topic word), which a user inputs, and the (topic-related) words expected to be related to the topic word and extracted from the documents where we employ a measure similar to the lift used in the association analysis. Experimental results show effectiveness of the proposed method.

## 1. 緒　　論

情報処理技術の進歩により，膨大な量の文書が電子的に蓄積されている．この大量の文書群から参照したい情報が記述されている文書を探すための，文書自動検索技術の開発・普及が進んでいる（例えば，Webページ検索[1)]）．

文書検索の代表的な方法である全文検索においては，例えば，検索したい文書の条件を，参照したい情報を表現する単語（キーワード）で表現する．しかし参照したい情報を表現する単語が思いつかない場合やその単語自体を検索したい場合がある．例えば，フランス旅行中に訪れるべき名所に関する情報を知りたい場合，名所の名称を知らないのが普通である．

全文検索に慣れていれば，フランスの名所名を検索し，それから，名所を個別に調べることになるであろう．しかし，コンピュータに慣れていないユーザに，それを期待するのは難しい．

本論文では，検索要求は，発見したい対象を順次制約していく，助詞「の」を用いた言葉の連鎖で表現される場合が多いと考え，またこの表現方法であれば，コンピュータに詳しくないユーザでも利用できると考え，この表現方法による検索要求の表現を用いることを考えた．その正しさを検証するために，まず，検索要求が，2 単語で表現される場合，例えば，「フランスの名所」，「モネの作品」などのように表される場合を考える．本論文では，この 2 単語の列のうち，第一の単語をテーマ，第二の単語をトピックと呼ぶことにする．これは，ユーザが用いるある検索キーワード（上述のトピック）（検索対象を最も明確に表すとしてユーザが選択した単語）の意味の多様性を減じるには，または意味・意図を明確にするためには，より上位の概念を表す単語で当該単語を修飾し，それは，テーマと呼んで差し支えない場合が多いであろうと考えたためである．

ところで，このような検索要求に対して，テーマ語（テーマを表している単語）とトピック語（トピックを表している単語）を 2 個の検索キーワードとして，全文検索を行うことができる．しかし，上述のように，例えば「フランスの名所」を検索したい場合，目的の対象はその名称（例えばルーブル美術館）を用いて記述され，“名所” と書かれているとは限らないが，いずれかの文書では，“名所” と名所の個別名がともに書かれている可能性がある．

本論文では，特定のトピックについて言及された文書を文書集合から検索する際，ユーザが入力するトピック語に加え，トピックと関連が深い名詞（すなわちトピック関連語）を推定し，両者の出現頻度情報に基づいて各文書をランキングする方法を提案する．

本論文の構成は以下の通りである．2. で提案手法について解説した後，3. で実験とその結果を報告する．4. で考察した後，5. で全体をまとめる．

†1 慶應義塾大学理工学部
Keio University
{koichiro, m_kondo, sakurai}@ae.keio.ac.jp

## 2. 提案手法

### 2.1 方針

既存の文書検索システムでは，ユーザの検索要求に対し，ユーザが入力した検索クエリのみを用いて適合文書の検索を行っている．しかし，ユーザが入力した検索クエリはユーザが求める情報を象徴する文字列であって，ユーザが求める情報そのものではない．例えば，ユーザがある映画の感想を読みたいと思ったとき，検索クエリには “(映画のタイトル) ”と “感想”という文字列を入力する場合が多いと予想されるが，感想を書き込む人が必ずしも “感想”という文字列を用いて感想を書くわけではない．すなわち，「感想」に “感想”という文字列を書くわけではない．

そこで，ユーザが求める文章を特徴づける名詞を収集し，それをユーザが入力した検索クエリと併せて用いて文書を検索することで，ユーザの情報要求を満たす適切な文書が特定可能であると考えられる．

本論文では，ユーザが

「“文字列 X” の “文字列 Y” について調べたい」

という情報要求をもつ局面を想定する．情報要求の具体例としては X = 京都，Y = 観光名所，あるいは X = 秋葉原，Y = 電気屋 等が考えられる．ここで，“文字列 Y” が本論文におけるトピック語に相当する．

### 2.2 手法

本論文では，文書集合中で出現する名詞について，トピック語との共起頻度を基にトピック関連度を算出し，文書中に出現した名詞の出現頻度とトピック関連度の積の総和を，ユーザの情報要求に対する文書の重要度とする手法を提案する．ここでは，以下に示すトピック関連度が高い名詞がトピック関連語となる．

文書集合内でのトピック語の出現を事象 $A$，ある名詞 $w_i$ の出現を事象 $B$ としたとき，$w_i$ のトピック関連度 $rel(w_i)$ を次式で定める．

$$rel(w_i) = P(B|A) \times \frac{P(B|A)}{P(B)} \qquad (1)$$

これはアソシエーション分析等で用いられる信頼度 (confidence) とリフト値 (lift)[6] の積である．信頼度は事象の発生確率が大きいほど有利であり，リフト値は事象の発生確率が小さいときは，事象と事象に関係性が無くても偶然に大きくなる場合がある．このため，双方が大きくなるような名詞がトピック関連語として相応しいと考えた．

トピック関連度はトピック語自身についても上式に従って同様に計算する．その際，$P(B|A) = 1$ とする．なお，次節の実験では，式 (1) の $P(B|A)$ が 0.1 以上の名詞のみを用いた．これは計算機の処理速度上の配慮であり，かつ $P(B|A)$ が小さい名詞はトピック関連度が小さくなると考えられ，そのような名詞は文書の重要度を判定する上で不要だと考えたからである．

次に，ある文書 $D_k$ 中での名詞 $w_i$ の出現頻度を $freq_k(w_i)$，文書 $D_i$ 中に存在する名詞の集合を $W_k$ としたとき，トピックに対する文書の重要度 $imp(D_k)$ を以下の式で定義する．

$$imp(D_k) = \sum_{w_i \in W_k} rel(w_i) \times freq_k(w_i) \qquad (2)$$

上式の通り，トピック関連度の高い名詞の出現頻度が大きい文書ほど，そのトピックに対する重要度が大きくなる．

**表 1** 検索ワードの詳細

| | 文書集合収集の際の検索クエリ | | トピック語 |
|---|---|---|---|
| | クエリ | 追加クエリ | |
| 試行 1 | 朝鮮戦争 | ⟶ | ソ連 |
| 試行 2 | モネ | ⟶ | 作品 |
| 試行 3 | OB 訪問 | ⟶ | お礼状 |
| 試行 4 | フランス | 旅行 | 名所 |
| 試行 5 | (映画名) | ⟶ | レビュー |
| 試行 6 | ゴッホ | ⟶ | ひまわり |
| 試行 7 | (小説作家名) | ⟶ | 感想 |
| 試行 8 | 大阪 | 名物 | 美味いもの |
| 試行 9 | (ラーメン店名) | ラーメン | 評価 |
| 試行 10 | 中古車 | ⟶ | 選び方 |
| 試行 11 | オスマン帝国 | ⟶ | 民族 |
| 試行 12 | (漫画名) | ⟶ | 感想 |

## 3. 実験

### 3.1 実験設定

実験では，ユーザの

「“文字列 X” の “文字列 Y” について調べたい」

という情報要求（詳細は，2.1 参照）に対して，異なる手法によって選択された文書群を，被験者に提示・評価させることで有効性を検証した．

文書群を選択する母集合として，上記の “文字列 X” を既存の Web 検索エンジンである Google に検索クエリとして入力し，検索結果として提示された Web ページを用いた．Google の検索結果として提示される Web ページ数の上限は 1,000 件であるため，実験データとなる文書集合の総数は高々1,000 件となる．得られた文書集合から，“文字列 Y” というトピックに関連する適切な文書を検索することが，情報検索手法の目的となる．

文書集合収集時及び実験時の検索ワードは，表 1 の通りとした．提示された情報要求の中には，“文字列 X” が意味する概念が大きすぎて，Google から取得した高々1,000 件の中に正例となる Web ページが存在しないケースも存在した．そのような場合は，Google に入力する検索クエリに単語を適宜追加し，文書集合内に検索対象の候補となる文書が含まれるようにした．

例えば，「“フランス” の “名所” について知りたい」という情報要求の場合，検索クエリ “フランス” で取得した 1,000 件の中には “名所” についての話題が殆ど無かったため，検索クエリ “フランス” AND “旅行” を Google に入力して文書集合を得た上で，その中から “名所” というトピックに関する文書を検索した．以上のデータ前処理における文書からの名詞の抽出には，形態素解析ツール mecab[7] と Perl モジュール Term Extract[8] を利用した．

評価対象とした手法は，以下の 2 つである．

(1) 提案手法

(2) Google での “文字列 X” と “文字列 Y” の AND 検索結果

ここで，(1) については “文字列 X” によって Google から取得した文書集合中で検索を行なうが，(2) は Web 全体から Web ページを検索するため，検索精度の面では不利である．しかし，(2) を被験者に同じ判断基準で評価させることで，普段から使用している Web 検索エンジンと比較した各手法の満足度が調査出来る．

上記 2 通りの文書ランキング方法のそれぞれについて，検索結果上位 10 件の各 Web ページを被験者に閲覧・評価させた．評価は，検索結果として提示された各 Web ページについて，

$h$ : 情報要求を満たすページであり，満足

$a$ : 情報要求を満たすページだが，物足りない

$b$ : 情報要求とは直接，あるいは全く関係のないページ

の 3 段階とした．$h$ と $a$ の違いは，情報要求に対する回答としての情報量や新規性，信頼性などである．

適合性が未知な文書集合を用いた場合の情報検索システム評価手法として使われる Cumulative Gain (CG)[4] を用いて評価を行った．CG は多段階適合性に適した評価手法であり，$d_i$ を $i$ 番目にランクされた文書，$G_i$ を $d_i$ の得点，$CG_i$ を上位 $i$ 番目までにランクされているページの累積得点とすると，以下のように定義される．

$$CG_i = \begin{cases} G_i, & \text{if } i = 1 \\ CG_{i-1}, & \text{otherwise} \end{cases}$$

$$G_i = \begin{cases} h, & \text{if } d_i \in H \\ a, & \text{if } d_i \in A \\ b, & \text{if } d_i \in B \end{cases}$$

ここで，$H, A, B$ は 3 段階に評価された文書の集合であり，一般に，それぞれ高適合文書，適合文書，部分適合文書と呼ばれる．$h, a, b$ は各々に割り当てられた重みであり，任意に定められる．検索エンジンを国際的に比較評価するプロジェクト NTCIR[5] においては $(h, a, b) = (3, 2, 0)$ とされた．

2 通りの検索手法それぞれの評価の CG を集計し，検索手法を比較した．上記の $h, a, b$ には

$$(h, a, b) = (3, 1, 0) \tag{3}$$

というスコアを与えた．

CG を用いることから分かる通り，本実験では検索結果上位 10 件中の順位の差異は，手法の評価で考慮しない．また，順位を考慮した発展型の CG（Discounted Cumulative Gain など）よりも，単なる CG のほうが検索手法に対するユーザの満足度を強く反映するという調査結果[9] も参考にした．

**表 2** 提案手法のユーザ満足度

| | CG | |
|---|---|---|
| | 提案手法 | 比較手法 |
| 試行 1 | 7 | 14 |
| 試行 2 | 22 | 16 |
| 試行 3 | 8 | 21 |
| 試行 4 | 13 | 9 |
| 試行 5 | 20 | 12 |
| 試行 6 | 14 | 20 |
| 試行 7 | 24 | 14 |
| 試行 8 | 16 | 14 |
| 試行 9 | 15 | 28 |
| 試行 10 | 2 | 19 |
| 試行 11 | 4 | 3 |
| 試行 12 | 22 | 20 |
| 平均 | 13.9 | 15.8 |

### 3.2 結　　果

各検索手法における，12 人の被験者から得た評価結果から算出した CG の平均値を表 2 に示す．なお，式 (3) の通り，各 Web ページに与えられるスコアは最大で 3 であり，検索結果の上位 10 件を評価対象としたため，CG の最大値は 30 となる．

さらに，提案手法において，トピック関連語として検索トピックに対し高い重要度を得た名詞の例を表 3 に示す．表 3 の左右は，“フランス” AND “旅行” というクエリで収集した文書集合から “名所” というトピックに関する文書を検索する際のデータである．表 3 左は文書集合全体での頻出名詞上位 20 語，表 3 右はトピック関連度上位 20 語を表示した．

## 4. 考　　察

### 4.1 トピック関連語

提案手法は，表 3 右に示されたような名詞が出現している文書は “名所” というトピックについて記述された文書である可能性が高い，という方針で検索する．トピック語自身もトピック関連語に含まれるので，トピック関連度が最も高い名詞はトピック語である “名所” となる．

この手法では，文書内で名詞が 1 回出現するたびに，その名詞のトピック関連度が重要度として文書に加算される．表 3 右では，“世界遺産” や “凱旋門”，“美術館”，“庭園” など，トピックである “名所” に直

表 4　満足度による手法の比較

| 提案手法の成績が優れた課題 | | | | 提案手法の成績が劣った課題 | | | |
|---|---|---|---|---|---|---|---|
| | 名詞 | トピック語 | 試行番号 | | 名詞 | トピック語 | 試行番号 |
| ① | フランス | 名所 | 試行 4 | ④ | OB 訪問 | お礼状 | 試行 3 |
| ② | モネ | 作品 | 試行 2 | ⑤ | ゴッホ | ひまわり | 試行 6 |
| ③ | (映画名) | レビュー | 試行 5 | ⑥ | 中古車 | 選び方 | 試行 10 |

接関係がありそうな名詞や，“唯一”，“歴史”，“発見”など，“名所”について話題にしている文書内で使われそうな名詞が並んでいる．

その一方で，“ホテル”や“日本”など，トピックとは関係が無さそうな名詞も見られる．これらの名詞がトピック関連語の上位に現れたのは，表 3 左の通り，これらの名詞の文書集合全体での出現頻度が高く，式 (1) の $P(B|A)$ が大きくなったことが原因である．しかしながら，このような文書集合全体での出現頻度が高い名詞に高いトピック関連度を与えても，多くの文書に一様に重要度が付与されるだけで，検索上の誤りは発生しにくいと考えられる．

実験結果が良好であったことから，トピック語である“名所”という文字列のみに注目するよりも，表 3 左に示されたような名詞の出現頻度全体で文書を評価したほうが“名所”というトピックに関する文書をより高精度に検索出来る，という提案手法の方針の正当性が検証された．

### 4.2　提案手法の有効性

表 2 のユーザ満足度は，12 試行の平均では，提案手法が比較手法に劣っている．しかし，個別の試行を見てみると，トピック関連語で文書を評価した方が有効である事例が見受けられる．各々の文書ランキング方法で高い満足度を得た情報要求の例を表 4 に示す．

表 3　文書集合中の頻出名詞（左）及び トピック関連語（右）

| 名詞 | 出現頻度 | トピック関連語 | トピック関連度 |
|---|---|---|---|
| パリ | 572 | 名所 | 3.413 |
| フランス旅行 | 390 | パリ | 1.083 |
| 写真 | 381 | ホテル | 0.847 |
| ホテル | 373 | 唯一 | 0.805 |
| 日本 | 357 | 世界遺産 | 0.739 |
| 記事 | 337 | 歴史 | 0.708 |
| コメント | 335 | アルザス | 0.669 |
| ページ | 297 | 凱旋門 | 0.647 |
| ブログ | 284 | 写真 | 0.642 |
| 情報 | 280 | ツアー | 0.608 |
| 名前 | 240 | 庭園 | 0.605 |
| ヨーロッパ | 226 | 日本 | 0.595 |
| 検索 | 221 | 発見 | 0.590 |
| 場所 | 207 | 情報 | 0.585 |
| レストラン | 205 | 意味 | 0.578 |
| 世界 | 200 | 美術館 | 0.574 |
| フランス語 | 197 | パリ市内 | 0.568 |
| 最新 | 190 | 作品 | 0.539 |
| トラックバック | 189 | セーヌ | 0.538 |

表 4 の結果から，トピック語がユーザの欲する情報の対象そのものを指し示す固有名詞であったり，ユーザの情報要求を具体的に表現している名詞である場合は，トピック語の出現頻度のみを重視したほうが良く，逆にトピック語が抽象的であったり，トピック語が表す概念が広い場合には，トピック関連語による重要度算出が有効であると推測される．

例えば，① の場合，ユーザはフランスの名所である凱旋門やエッフェル塔などについての情報を求めており，文書の中に“名所”という名詞が出現している必要は無い．しかし，④ では OB 訪問のお礼状の話題の中で“お礼状”という名詞が出現しないことはあり得ないと考えられ，その出現頻度が高いほどトピックに関する記述が豊富でユーザにとって有意義な文書である可能性が高い．

画家の作品に関する情報要求という同じ条件で対比させた ② と ⑤ についても同様である．② の場合は“作品”という名詞自体が重要なのではなく，モネの作品の具体的な情報が知りたいのに対し，⑤ では“ひまわり”というゴッホの作品について知りたいのだから“ひまわり”という名詞の出現頻度が重要となる．

このように，トピック語の出現頻度のみを重視するか，トピック語はトピック関連語の 1 つに過ぎないとするか，いずれの方針が有効かはユーザの情報要求の内容によって異なる．提案手法は，少なくともその一方において，既存の Web 検索を上回ることが示された．提案手法は，Web ページ内のテキスト情報のみに着目していることから文書集合全般に適用可能であることを考えると，有意義な結果が示されたと言える．

なお，文書の内部構造を数学的により厳密にモデル化し，有効性をさらに高めた，提案手法の拡張[10] に関しても評価を進めており，稿を改めて結果報告の予定である．

## 5.　結　　論

本論文では，文書集合中から特定のトピックに関する文書を検索する際に，ユーザによって入力されるトピック語だけでなく，そのトピックと関連があるトピック関連語を用いて，そのトピックに対する文書の重要度を算出する文書ランキング方法を提案した．その実現のために，名詞毎に算出されたトピック関連度を，その名詞の出現頻度とリフト値に従って文書に重要度として付与する実験を行った．

実験の結果，

- 検索トピックに対する文書の重要度を，トピック語のみの出現頻度を重視して重要度をスコアリングする方法は，従来の手法と同等のユーザ満足度を得ることが出来る
- この文書ランキング方法は，互いに異なる性質の検索トピックの一方に対してより有効である場合がある

ことが示された．本研究の今後の展望としては，

- ユーザ満足度のさらなる向上
- とくに，提案手法が有効でないクエリへの対応

の通りである．

## 参 考 文 献


1) Google: Google. http://google.com.
2) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).
3) 北研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版 (2002).
4) Järvelin, K. and Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents, *Proc. of the 23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp.44–48 (2000).
5) NTCIR-WEB: http://research.nii.ac.jp/ntcweb/
6) 人工知能学会（編）：人工知能学事典，共立出版 (2005).
7) 形態素解析ツール：mecab http://mecab.sourceforge.net/
8) Term Extract: http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html
9) Al-Maskari, A., Sanderson, M. and Clough, P.: The Relationship between IR Effectiveness Measures and User Satisfaction, *Proc. of the 30th Annual Intl. ACM SIGIR Conf.*, pp.773–774 (2007).
10) 近藤真史：トピック関連語に基づく文書の重要度ランキング，慶應義塾大学理工学研究科 修士論文 (2008).