Vol.2015-DPS-165 No.11
2015/12/11

トピックモデルと WEB 閲覧履歴による ユーザの意図を考慮した検索システムの開発

唐澤貴大[†1]　中沢実[†1]

概要：現在のインターネットは情報の流通量が増加し続けているにもかかわらず，情報収集の手段として検索エンジンでのキーワード検索が主流のままである．これは知りたい情報に関してキーワードを入力することで情報検索を行えるが，キーワードの組み合わせや端的に言語化しにくい情報を見つける上では困難な手段となってしまう．そこで本研究では，検索キーワードのみに依存しない情報検索システムを提案する．また，試作システムによる現時点での評価実験の結果と今後の展望について述べる．

キーワード：トピックモデル，閲覧履歴，検索システム，キーワード検索

Development for Search System considering User's Intention using Topic Model and WEB Browsing history

TAKAHIRO KARASAWA[†1]　MINORU NAKAZAWA[†1]

Abstract: Despite the growing amount of information is the current Internet, as a means of collecting the information, remains the mainstream is the keyword search in the search engine. It becomes a difficult way on finding keyword combinations and difficult language of informal information, but it can make a search by entering a keyword for the information you want to know. We shouldn't rely only on search keywords in the present study, information retrieval systems. Also, as a result of the experiments in the present system described prospect.

Keywords: Topic Model, WEB Browsing history, Search System, Keyword Search

1．はじめに

近年，インターネットの普及により誰もが様々な情報の発信を簡単に行え，情報収集ができるようになった．しかし，情報収集の手段としては依然として検索エンジンを利用したキーワード検索が主流となっている．これは自らが知りたい情報に対して適切な単語を選定し組み合わせることで成立する．また，キーワードを含む web ページのみが検索の対象とされるため，欲しい情報に対してキーワードを変えた再検索を強いられることがしばしばある．このことから検索エンジンを使いこなせなかったり，端的に言語化しにくい情報を見つけ出したりすることは難しい．そこで，本研究では検索キーワードのみに依存しない情報検索システムを提案する．

本システムでは,ユーザから知りたい情報に関するキーワードを受け取り，それに関連する単語から新たな情報を収集・選別し，ユーザが求めていると思われる情報を提示する事を可能としている．

2．関連研究

仲川らは WWW 検索支援として動的にカテゴリ構造を変化させるディレクトリ検索サービスを提案している[1]．このシステムではユーザから受け取ったキーワードをもとに情報のカテゴリ構造を構成することで容易に必要な情報を得られるとする検索システムである．このシステムではキーワードに関連のない文書が検索対象にならないことや,目的の情報にたどり着くまでに何度もカテゴリの選択をする必要がある．本システムではキーワードに幅をもたせることと一般的に利用されているキーワード検索を拡張している点が異なる．

また,堀らのシステムではユーザのページ閲覧行動と検索意図に関係があるとして閲覧履歴に出現した単語を独自の方法でクラスタリングし検索語拡張を行っている[2]．このシステムではクラスタリングの基準が一ユーザのコンテキストによって定められているため，基準が偏ってしまったり，新規の単語に対しては良い効果が得られなかったりすると考える．さらに閲覧履歴によってキーワードに類似する単語を選択するため,意味的な関連が強いとは限らない．閲覧履歴を利用する点では同じであるが，それ自体でキーワードを拡張するのではなく結果の絞り込みに利用している点や，複数の文書から単語の意味的な関連を探る点が異なる．

森らのシステムでは,ソーシャルブックマークのタグ情報

†1 金沢工業大学
Kanazawa Institute of Technology

をもとに web 上の情報の絞り込みを行いユーザが情報の選別をする手間を省くというものである[3]．この手法ではソーシャルブックマークのタグとキーワードを一致させることでブックマークされた情報から選別を行いユーザに提示している．ソーシャルブックマークを利用するユーザやそこに登録された情報には限りがあるため検索対象をそのような基準で限定してしまうと新しい情報や誰も注目していない情報は得られない．しかし，本システムでは WWW を検索対象としているためそのような問題は生じない．

3．提案手法

3.1 クローリング

検索キーワードに対する関連語の算出のため，事前に代表となる幾つかの文書を収集する．収集した文書はテキストデータに変換しておく．

3.2 モデリング

収集した文書に対してトピックモデル[4]を用いる．これによりそれらの内容から 1 文書を複数項目(トピック)に分類することができる．トピックモデルではトピックごとに，登録された全単語の生起分布が存在する．1 つのトピックはある基準によって単語の生起確率が求められる．そのため，1 トピック内で生起確率が近い単語同士はそのトピックが暗に意味する基準で近しいと言える．あるトピック内では 1 つの文書で使われた異なる幾つかの単語の生起確率が共に大きくなる性質がある．これらの性質から，異なる文書に現れる単語であっても同じトピックで生起確率が近い単語同士は意味的な関連が強いとみなすことができる．この性質より任意の単語に対して関連を持つ別の単語を求める．（図 1）

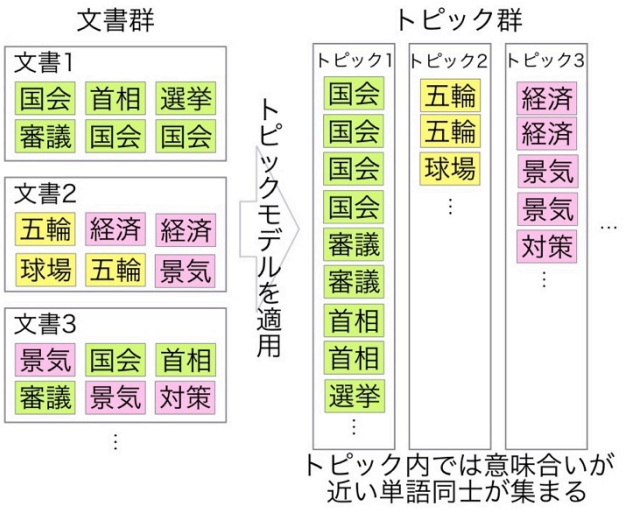

図 1　トピックモデルの概要

3.3 形態素解析

トピックモデルを利用した文書の分類をする上で文書を単語に分割する必要がある．そのため今回は MeCab[5]を用いた形態素解析を行い単語に分割する．その際，単語の種類数（単語の意味による種類数ではなく，文字列としての種類数）を減らすためにすべての単語に対して標準形に直す処理を施す．これは本システムでは単語の意味を重視するため，同じ単語の活用形を別の単語とすると，ある意味に対する表現の仕方にゆれが生じ，良い基準とは言えなくなってしまうからである．

また，後述する文書同士の比較の前段階としても閲覧履歴に対して形態素解析を行う．

3.4 文書の比較

閲覧履歴を用いた評価のために文書同士を TF-IDF コサイン類似度で比較をする．この手法は 2 つの文書の TF-IDF をパラメータとしてコサイン類似度を算出するものである．TF-IDF により文書の特徴を算出しコサイン類似度によって文書間の距離を算出する．これによりユーザの閲覧履歴を基準とした任意の web ページ群との 1 対多の絶対的な比較を行い，ユーザが求めていると思われる情報を提示する．閲覧履歴を基準とすることで，ユーザの趣味，思考に沿った比較が可能になる．

3.5 提案手法

本システムでは事前に準備した文書から，ある単語のトピック毎の生起確率，あるトピック内の単語の生起分布，また，ある文書と閲覧履歴の類似度の 3 つの要素に注目した．

全体の流れは次のようになる．

(1) 代表となる文書を収集しトピックモデルを構築する．

(2) (1)でのトピックモデルをもとにユーザからの検索キーワードに類似する単語を新しいキーワードとして選択する．

(3) 新しいキーワードを用いてインターネットから情報検索を行う．

(4) 検索結果とユーザの閲覧履歴を比較し，情報の選択とランク付けを行いユーザに提示する．

図 2 に全体の動作図を示す．

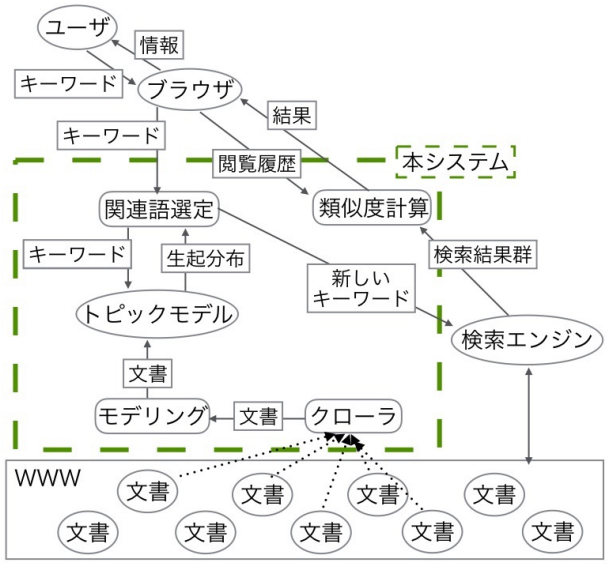

図 2　システムの動作

手順(1)は単語の意味的な関連を構築するために行う．web 上の情報など不特定多数の著者による多数の文書が好ましいと考える．収集した文書を先述の方法で形態素解析を行いシステムに登録する．トピックモデルの生成には中華料理フランチャイズを適用した階層ディリクレ過程[]を用いる．これにより，必要なトピック数を推定しトピックごとの単語の正規分布を求める．

手順(2)ではユーザから検索のためのキーワードを受け取り，構成したトピックモデルから類似する単語を抽出し，新しいキーワードとする．単語抽出には 3 つのステップを要する．

1. close topic と near topic の選定
2. 各トピックで near words の選定
3. close topic の near words から near topic の near words を除外し，新しいキーワードを決定する．

手順 1 ではキーワードをもとに close topic と near topic を決める．

$$T_c = \max_t \left(\sum_{T=1}^{t} \frac{\sum_{j=1}^{n} K_{tj}}{\sum_{i=1}^{N_t} W_{ti}} \right) \qquad (1)$$

T_cを選出するトピック，Kをトピック tでのキーワードの生成確率，nをキーワードの数，w_{ti}をトピック tの単語の生成確率，N_tをトピック tの単語数とする．選出された T_c を close topic，その他のトピックを near topic とする．

手順 2 では全トピックに対してキーワードに近い単語群を抽出する．各トピックである 1 キーワードの生起確率に最も近い 1 単語を near word, キーワードごとの near word の集合を near words とする．

手順 3 では close topic のみで near words と判定された単語を抽出する．close topic の near words から near topic の near words を除外し，close topic のみに存在する near words を選出する．この手順を行うことで単語を厳選しキーワードを絞り込むことができる．

例として図 3 のようなトピック 1 からトピック 3 までの 3 つのトピックが生成されたとし，ユーザからキーワード「アルバム」と「ジャケット」を受け取ったとする． 手順 1 に従って close topic と near topic を選出した場合，トピック 1 での単語「アルバム」の生起確率が 0.0200, 単語「ジャケット」の生起確率が 0.0190, 合計値が 0.0390 となる．トピック 2 とトピック 3 においても同様に求めることで，合計値 0.0281 と 0.0309 が求まる．よって，それが最も大きいトピック 1 が close topic となり，その他のトピック 2, トピック 3 は near topic となる．（図 4）次に，手順 2 に従い，near words を選出する．例としてトピック 2 を用いた場合，キーワード「アルバム」の生起確率 0.0280 に最も近い生起確率 0.0300 を持つ単語は「子供」，キーワード「ジャケット」に最も近い生起確率を持つ単語は「ジーンズ」と「コンサート」となる．よって，「子供」,「ジーンズ」,「コンサート」が near word であり，トピック 2 の near words となる．同様にトピック 1 とトピック 3 に関して near words を求めた結果が表 1 上部である．（図 5）そして，手順 3 に従い単語の厳選を行う．close topic であるトピック 1 の near words から near topic のトピック 2，トピック 3 の near word と重複する単語を除外する．よって新しいキーワードとして選出される単語は「歌」となる．（表 1）

トピック1

単語	生起確率
アルバム	0.0200
歌	0.0200
コンサート	0.0190
ジャケット	0.0190
歌手	0.0170
:	:
コート	0.0013
:	:
子供	0.0001
赤ちゃん	0.0001
七五三	0.0000
:	:
:	:

トピック2

単語	生起確率
子供	0.0300
アルバム	0.0280
赤ちゃん	0.0220
七五三	0.0220
:	:
歌	0.0090
アウター	0.0080
コート	0.0030
:	:
ジーンズ	0.0001
ジャケット	0.0001
コンサート	0.0001
:	:

トピック3

単語	生起確率
アウター	0.0320
コート	0.0311
ジャケット	0.0300
スカジャン	0.0200
ジーンズ	0.0189
:	:
:	:
コンサート	0.0100
子供	0.0100
七五三	0.0080
アルバム	0.0009
:	:
:	:

図 3　トピックモデルの例

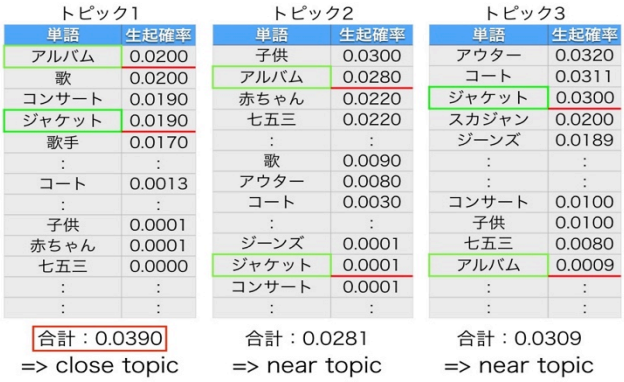

図 4　トピックの分類

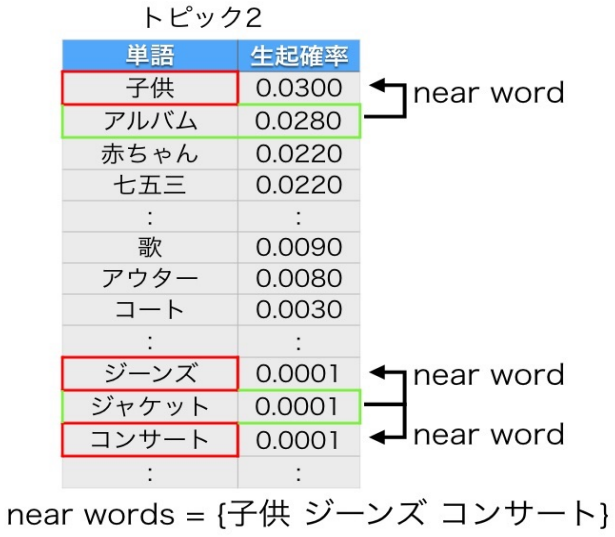

図 5　near words の選出

表 1　near words と選出される新しいキーワード

	トピックの種類	near words
トピック 1	close	歌 コンサート
トピック 2	near	子供 ジーンズ コンサート
トピック 3	near	七五三 コート
新しいキーワード	-	歌

手順(3)では新しいキーワードで Google[6]を用いて情報検索を行い，結果の上位 20 件を取得する．これは日本全国平均で 80%以上が 20 件目までしか閲覧しない[7]ことによる．

手順(4)ではユーザの嗜好をもとに検索結果の順位付けを行う．ユーザの閲覧履歴に含まれる web ページを取得，形態素解析を施し，TF-IDF 値を算出する．同じく検索結果ページに関しても同様の処理を行い，値の算出を行う．閲覧履歴の TF-IDF 値と各検索結果ページの TF-IDF 値を用いてコサイン類似度を算出する．このコサイン類似度が 0.7 に近いものからスコアが高いとする．全ての検索結果に対して比較が終了した時点でのランキングでユーザに結果ページを提示する．コサイン類似度の基準を 0.7 と設定するのは，ユーザの嗜好から離れすぎず，過去に閲覧したページではないものを優先的に提示するためである．

4. 実験と評価

4.1 評価方法

本システムに対してユーザからの検索クエリを与え，新しいキーワードとして選出された単語について観察する．今回の評価実験のデータセットとして，情報学広場：情報処理学会電子図書館[8]，人工知能学会論文誌[9]，映像メディア学会誌[10]にて公開されている文書 12,795 件を使用した．(表 2) 単語数は文書に対して形態素解析処理を行った後の値である．複数のデータセットとそれぞれについてトピックモデルの学習回数を 50，100，150 回と変えたものを用いる．

表 2　データセット

	文書数	単語数
人工知能学会論文誌 +映像メディア学会誌	5,725	33,977
情報処理学会電子図書館	7,070	41,399
合計	12,795	75,376

4.2 評価結果

人工知能学会論文誌の文書と映像メディア学会誌の文書を合わせたもの(jstage)と情報処理学会電子図書館(ipsj)の学習結果を表 3，表 4 に示す．ただし学習時間は概算である．ユーザからの検索キーワードとして与えたものを表 5 に示す．

表 6，表 7 に各クエリに対する新しいキーワードの内訳を示す．列がトピックモデルの学習回数，選出された新しいキーワード数，新しいキーワードの一部を示す．このように，新しいキーワードの選出数が多く検索エンジンに渡せないという結果になった．また，選出された単語に関してもキーワードと関係があるように感じられるものは少なかった．

表 3　jstage の学習結果

学習回数（回）	0	50	100	150
学習時間（時間）	0	13.5	28	42.5
トピック数（種）	0	6	9	11
パープレキシティ	1982242468.89	29.916	29.869	29.833

表 4　ipsj の学習結果

学習回数（回）	0	50	100	150
学習時間（時間）	0	16	32	48
トピック数（種）	0	36	35	34
パープレキシティ	139196942.643	82.995	79.265	77.860

表 5　検索キーワード

Q1	語句　抽出　手法
Q2	キーワード抽出　手法
Q3	文章　まとめ　自動

表 6　選出された新しいキーワード（jstage）

	50		**100**		**150**	
	数	**例**	**数**	**例**	**数**	**例**
Q1	12	スライド図	12	エラー無視	8	解 生命ハブ
Q2	1	図	1	図	1	研究
Q3	7	遺　伝優先	580	図書館統括	896	述語 性能

表 7　選出された新しいキーワード（ipsj）

	50		**100**		**150**	
	数	**例**	**数**	**例**	**数**	**例**
Q1	156	修正　わかりやすさ	122	読み取りニッチ	57	茶筌 接続詞
Q2	1	走る	0		1	全国
Q3	896	他動詞解凍	786	わかち書きキータッチ	1239	茶筌 イディオム

4.3 考察と展望

新しいキーワードが大量に選出されるという問題に関して，モデル内の単語の生起確率にばらつきがなさすぎることが原因として挙げられる．これはデータセット不足や学習回数不足によって個々の単語の意味が不明瞭な状態になっていると推測する．

新しいキーワードがキーワードに対して関連を感じにくいという問題がある．大量に選出された新しいキーワード内にはキーワードに対して関連がありそうな単語も含まれている．これらの単語のみを抽出することで精度を上げることができる．そのために，新しいキーワード選出段階でユーザのコンテキストを利用する策や，データセットを増やすという策が挙げられる．

また，クエリ 2 の単語「キーワード抽出」について，この単語はデータセットに含まれておらず，新しいキーワード選出時にはシステムに無視されている．このことから，あらゆる単語をデータセットに含める，何らかの手法によって意味を推測し検索に活用するなどの解決策を講じる必要がある．

今回の実験では，トピックモデルの学習回数を固定の 3 パターンで行った．そのため，言語モデルの評価指標となるパープレキシティが収束する前に学習を終了している可能性がある．パープレキシティが収束したら学習を終了するなどの工夫によって幾つかの問題が改善される可能性がある．

5．おわりに

本論文では，トピックモデルを用いてキーワードのみに依存しない情報検索システムを提案した．この手法では，元となる文書からトピックモデルを生成し，同じトピック内での生成確率が近い単語同士は意味的な関連性が高いと考え，それらに関してシステムが情報を収集しユーザに提示した．また，今後精度向上を目指すためには，ユーザの閲覧履歴だけでなく他のコンテキストについても収集し，システムの情報収集や情報の選別部分に適用することが挙げられる．

謝辞

日本語形態素解析システム MeCab の作者の方々に深く敬意を表します．

参考文献

1) 仲川，高田，関，検索目的を反映したカテゴリ構造に基づく WWW 検索支援：情報処理学会研究報告ヒューマンコンピュータインタラクション(HCI)，1999，9(1998-HI-082)，1999
2) 堀，今井，中山，ユーザの Web 閲覧履歴を用いた検索支援システム：情報知識学会誌 Vol.17，No.2，2007
3) 森　一聡, ブックマーク情報を用いた Web 検索支援システムの開発：高知工科大学フロンティアプロジェクト学士学位論文（未刊行）
4) 岩田 具治（2015）．トピックモデル 講談社
5) http://taku910.github.io/mecab/
6) https://www.google.co.jp/
7) クロスニフティ，クロスフィニティ 全国の男女・年代別の検索エンジン利用動向を調査，http://www.crossfinity.co.jp/pdf/20140131_01.pdf（参照 2015-11-14）
8) https://ipsj.ixsq.nii.ac.jp/ej/
9) https://www.jstage.jst.go.jp/browse/tjsai/-char/ja/
10) https://www.jstage.jst.go.jp/browse/itej/-char/ja/