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

2006－MPS－59 (15)
2006／5／26

ベクトル表現可能な機械抽出トピックの定量的評価法

福井 健一[†], 斉藤 和巳[††], 木村 昌弘[‡], 沼尾 正行[†]

[†] 大阪大学 産業科学研究所
[††] NTTコミュニケーション科学基礎研究所
[‡] 龍谷大学 理工学部 電子情報学科

大規模文書群からのトピック自動抽出は文書群全体像の把握や文書分類などに有用であるとして，研究が行われている．しかし，ここで機械抽出トピックの評価は重要な課題であるが，人手によってトピック分類された文書群が得られたとしても，トピック表現の多様性によりトピック間の対応付けが困難となるため，単純に機械抽出トピックと比較評価できない．そこで本稿では，潜在的意味解析などによるベクトル表現可能なトピック抽出法を対象として，人手によってトピック分類された文書群を用いて抽出トピックの解釈可能度を定量的に評価することを試みた．日本語および英語の新聞記事などから機械抽出したトピック群にて，提案評価法の妥当性を検証した．

Evaluation of Vector Representable Topics that were Extracted Automatically

Ken-ichi Fukui[†], Kazumi Saito[††], Masahiro Kimura [‡], and Masayuki Numao[†]

[†] The Institute of Scientific and Industrial Research, Osaka University
[††] NTT Communication Science Laboratories
[‡] Department of Electronics and Infomatics, Ryukoku University

Automatic topic extraction from a large number of documents is useful to figure out an entire picture of the documents or to classify the documents. Here, it is an important issue to evaluate the automatically extracted topics, however, even if manually-labeled documents are obtained, it is impossible to compare automatically and manually derived topics due to complexity and uncertainty of the topics' structure. As the objective is vector representable topic extractions such as Latent Semantic Analysis, in this paper we tried to evaluate the interpretability of automatically extracted topics using the manulally-labeled documents. We validated the proposed evaluation method using topics extracted from Japanese and English news articles.

1 はじめに

近年，インターネットを通じてニュース記事，ブログ，電子メールなどから大規模な文書群を容易に得ることができるようになった．これら文書群の内容は，世界の最新の事件や出来事など，あるいは文書提供者間での議論などと様々で刻々と文書内容が変化する．このような文書群から，適切な主要トピックを自動抽出し，各トピックに関連する文書群を同定することにより，文書群を体系的に整理することができれば，文書群の全貌を容易に把握することができるとして研究が行われている．

現在，文書群からの機械的トピック抽出法は，文書内に出現する有意な単語群を基底とした特徴空間を考え，各文書を単語頻度ベクトルとして取り扱うベクトル空間モデルに基づく方法が一般的である．ここで，トピックとは同じ事柄について述べられている文書群が属するクラスを指し，同じトピックに属する文書は出現する単語頻度が似ていると考えられる．ベクトル空間モデルに基づいたトピック抽出には様々な方法により研究がなされている．例えば，古典的クラスタリング手法による方法[1]や，同じトピックに属する文書はある特徴的なひとつの軸の周りに分布していると考えると，特徴空間から特徴軸を発見する問題として定式化する方法がある．特徴軸を発見する方法としては，情報検索の分野で広く用いられている潜在的意味解析（*LSA*）[2]や，主成分分析（*PCA*）に基づく方法[3]，さらには近年信号処理の分野で発展した独立成分分析（*ICA*）を用いる方法[4]が

ある．

ここで，どのトピック抽出法を用いたとしても機械抽出されたトピックの評価は重要であるが難しい問題である．それは，トピックは一般的に多様な捉え方が可能であるからと考えられる．例えば，「米国大統領」に関するトピックと大きく捉えるか，「米朝実務会談」とより細かく捉えるかといった階層の問題もあれば，普通，人はこのような捉え方はしないと考えられるが「各国の議員選挙」という視点の違いの問題もある．このようなトピック表現の多様性により，機械的に抽出されたトピックと認知的に抽出されたトピック間の対応付けが困難になり，直接両者を比較することは難しいと考える．

本稿で対象とするような文書群は，一般的にはトピックを特徴付ける単語群を予め選択することは困難である．*LSA* を代表とするトピックを特徴軸として抽出する手法では，予め特徴語の選択を必要とせず，文書群の大まかな内容や概念に基づいた微妙なトピックを抽出できるとされ未知のトピックの発見に適していると考えられる．そこで本稿では，これらベクトル表現可能なトピック抽出法を対象として，任意の認知的方法[1]によってトピックラベルの付与された文書群をベンチマークとして，抽出トピックの解釈可能度の定量化を試みた．

2 トピックの解釈可能度

提案する機械抽出トピックの解釈可能度についての定式化および解法について述べる．各文書は出現単語頻度に基づく Bag-of-Words(BOW) モデルによりベクトル表現され，トピックは *LSA* などによって特徴軸として抽出されているとする．基本的な考え方は，機械抽出トピックが認知トピック群の線形結合で近似できれば解釈可能で，そうでなければ解釈困難であると考える．ここで，本稿では認知トピックは各トピックに属する文書群の重心ベクトルにより定義した．

具体的には，ある v 次元-機械抽出トピックベクトル Ψ に近くなるように，K 個の認知トピック群の重み付き和ベクトル $\sum_{k=1}^{K} w_k \tilde{\Psi}_k$ の K 次元-重みベクトル $\mathbf{w}$ を求める．ただし，$||\Psi|| = ||\tilde{\Psi}_1|| \cdots ||\tilde{\Psi}_k|| = 1$ に正規化されているとする．各ベクトルを列ベクトルとすると，$v \times K$ 行列 $\mathbf{A} = [\tilde{\Psi}_1, \cdots, \tilde{\Psi}_K]$ を定義できる．すると、重み付き和ベクトルは $\sum_{k=1}^{K} w_k \tilde{\Psi}_k = \mathbf{Aw}$ と表される．

いま，ベクトル Ψ と $\mathbf{Aw}$ の *cosine* 類似度の自乗を最大化するように，ベクトル $\mathbf{w}$ を求めるとする．すると、目的関数は以下で定義できる．

$$P = \frac{(\Psi^t \mathbf{Aw})^2}{\mathbf{w}^t \mathbf{A}^t \mathbf{Aw}} \tag{1}$$

上記目的関数は数理物理学における一般化レイリー商と同形である[5]．その解 $\hat{\mathbf{w}}$ は次式により与えられる:

$$\hat{\mathbf{w}} \propto (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \Psi \tag{2}$$

ここで，目的関数は *cosine* 自乗で評価しているため，$\mathbf{w}$ の定数倍には不変なことに注意する．また，認知トピック群である行列 $\mathbf{A}$ の列ベクトルは，人が分類したトピックであるため一次独立であると考えられる．故に，行列 $\mathbf{A}^t\mathbf{A}$ は非特異となるため解は存在する．従って，機械抽出トピックの解釈可能度 p は $\hat{\mathbf{w}}$ を式 (1) に代入して得られる．この時，評価値 p は [0,1] の値を取り，$p = 1$ の時，その機械抽出トピックは，認知トピック群の線形結合で *cosine* 類似度の意味で一致させることが可能であることを意味する．一方，機械抽出トピックと認知トピック群は互いに直交する時，$p = 0$ となる．

また，上記目的関数では *cosine* 類似度の自乗で評価しているため，*cosine* 類似度が負になる場合でも両者は近いと評価される．しかし，その解は $\mathbf{w}$ の定数倍に対して不変であるので，$-\mathbf{w}$ と取ることで *cosine* 類似度として近くなる $\mathbf{w}$ を得ることができる．

機械抽出トピックが認知トピック群の線形結合で近似できれば，それは解釈可能で，そうでなければ解釈困難である，という前提について検討する必要がある．最も単純な場合，すなわち単一の認知トピックと一致する場合，解釈可能な事は明らかである．しかし，たとえ評価値 p が 1 に近かったとしても，必ずしも単一のトピックを表しているとは限らない．そのため，各認知トピックとの類似度と併用して解釈する必要がある．しかしながら，逆に線形結合で近似できない場合，客観性のある認知トピック群が得られているとしたら，それらには含まれないようなトピックで構成されている機械抽出トピックは解釈困難と考えられる．

3 検証実験

3.1 データセット

本実験では，1994 年 1 月〜6 月までの毎日新聞国際面記事（日本語），および TDT3[2] で公開されているデータセット（英語）の 2 種類を用いた．TDT3 のデータセットには，1998 年 10 月〜12 月までの

[1] 本稿では，“機械的” と対比させて単純に人手によることを指す．“機械的” は “自動的” と読み替えられる．

[2] http://projects.ldc.upenn.edu/TDT3/

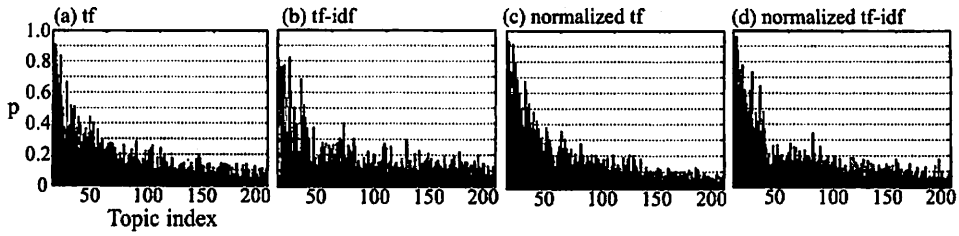

図 1: *LSA* による抽出トピックの解釈可能度. (毎日新聞データ)

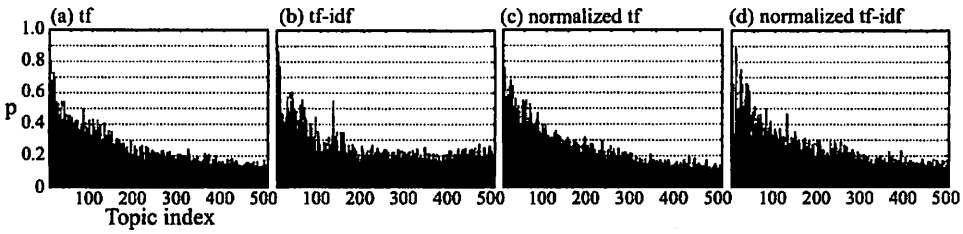

図 2: *LSA* による抽出トピックの解釈可能度. (TDT3 データ)

New York Times などの新聞記事や CNN, ABC ニュース番組などの音声から抽出された文字情報が含まれており, 各文書はガイドラインに従って人手によってトピックが付与されている. 各データセットの基本統計量を表 1 に示す. 毎日新聞記事については, 斉藤らの認知科学的実験 [6] によってトピックが付与されたデータセットを用いた. また, BOW に用いる単語集合を選択する前処理として, 不要な単語を設定し stop word の除去を行い, TDT3 のデータに関しては英語の語幹を取り出す Porter stemming も施した.

表 1: データセットの基本統計量

	総文書数	総単語種類数	トピック数
毎日新聞	2,694	18,070	39
TDT3	34,413	78,452	115

3.2 検証の考え方

本検証実験では, これらの仮定や情報検索分野での事実に基づいて, 本提案評価法によってトピック抽出においても同様の事が示せることを確かめる. すなわち, 有意であると仮定もしくは考えられるトピックに関する評価が高く, そうでなければ低い結果が得られたならば, 提案評価方法が正しく機能している事実のひとつになると考える. 本実験では, 代表的なトピック抽出法として *LSA* を適用し次元縮約を行った. また, *tf*, *tf-idf*, *normalized tf*, *normalized tf-idf* の 4 種類の重み付けを比較した.

3.3 トピックの解釈可能度の妥当性

機械抽出トピックの解釈可能度の妥当性の検証を行った. 有意な抽出トピックの近傍には人が同じトピックであると判断する文書群が集まっているので, 認知トピック群の線形結合で表される可能性が高いと考えられる. すなわち, *LSA* の性質から有意であると考えられる上位の抽出トピックの評価値は高く, 有意ではない中下位の機械抽出トピックおよびランダムなトピックは低くなるはずである.

LSA による抽出トピックの解釈可能度を評価した結果を 1 および 2 に示す. 図の縦軸は解釈可能性評価値 p, 横軸は抽出トピックインデックスを表している. まず, *LSA* による抽出トピックでは (a) ～ (d) いずれの重み付けにおいても, 評価値は上位数トピックは高く, その後急速に減少していっている. しかし, ランダムに生成した正規直交基底からなるランダムトピックの場合 (3) にその傾向は見られず, トピック毎の際だった差はない.

次に, 重み付けによる違いを図 2 と図 3 によっ

表 2: 解釈可能度の各階級のトピック数および最大値. (毎日新聞データ)

p	(a)	(b)	(c)	(d)
> 0.9	2	0	4	3
> 0.8	2	2	4	4
> 0.7	4	5	7	7
> 0.6	8	6	10	13
> 0.5	9	11	17	15
> 0.4	16	13	22	23
> 0.3	31	23	35	27
MAX	0.9094	0.8229	0.9134	**0.9597**

表 3: 解釈可能度の各階級のトピック数および最大値. (TDT3 データ)

p	(a)	(b)	(c)	(d)
> 0.9	1	0	1	0
> 0.8	1	3	1	2
> 0.7	4	4	2	6
> 0.6	9	5	9	16
> 0.5	14	17	23	26
> 0.4	49	40	54	45
> 0.3	116	70	107	97
MAX	0.9140	0.8632	**0.9364**	0.8883

て比較する. それぞれ (a) *tf*, (b) *tf-idf*, (c) *normalized tf*, (d) *normalized tf-idf* を表している. 表中の数値は, *LSA* によって抽出されたトピックの内, 一定以上の解釈可能度が得られたトピック数を表しており, 最下段は抽出トピック中の最大値を表している. 毎日新聞データ (図 2) では, (a) (b) よりも (c) (d) の方が明らかに多くのトピックで高い評価値を得ている. しかし, (c) と (d) の優劣は付け難い. また, TDT3 データ (図 3) では, 特徴空間の適合度同様, 毎日新聞データほど明らかな差は出ていないが, (a) (b) よりも (c) (d) の方が若干良い評価値が得られている. (c) と (d) では, 最大値は (c) の方が高いが, (d) では $p > 0.6$ のトピック数が多いため, トピック全体としては (d) の方が若干良い結果であると考えられる.

4 まとめ

本稿では, 人手によってトピック分類された文書群を用いて, *LSA* を代表とするベクトル表現可能な機械抽出トピックの解釈可能度の定量化を試みた. 本稿では, 従来研究の知見に基づく仮定を元に, 異なる手法間を比較することで提案評価尺度に関する検討を行った. 実際のニュース記事など 2 種類のデータセットに適用した結果, その仮定の範囲においては概ね妥当な結果が得られた.

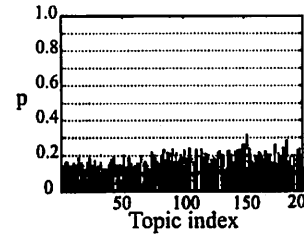

図 3: ランダムトピックの解釈可能度. (毎日新聞データ)

参考文献

1) Schultz, J. M. and Liberman, M.: Topic Detection and Tracking using idf-Weighted Cosine Coefficient, *Proc. DARPA Broadcast News Workshop*, pp. 189–192 (1999).

2) Landauer, T. K. and Dumais, S. T.: A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review*, Vol. 104, pp. 211–240 (1997).

3) Kimura, M., Saito, K. and Ueda, N.: Multinominal PCA for extracting major latent tooics from document streams, *Proceedings of 2005 International Joint Conference on Neural Networks*, pp. 238–243 (2005).

4) 濱本雅史, 北川博之, Pan, J. Y., Faloutsos, C.: 独立成分分析を用いたテキストデータからのトピック検出, 電子情報通信学会第 15 回データ工学ワークショップ (DEWS2004) (2004).

5) Duda, R. O., Hart, P. E. and Stork, D. G.: *Pattern Classification Second Edition*, Wiley-Interscience (2000).

6) 斉藤和巳, 木村昌弘, 上田修功: 文書トピックに関する認知科学的実験, *SIG-KBS-A405-10*, pp. 57–62 (2005).