単語の共起グラフを用いた潜在的意味に基づく効果的な文書分類への取り組み

A Study on Efficient Text Classification Based on Latent Semantic Used a Graph of Co-occurring Terms

小倉由佳里 [1*]　　小林一郎 [2]
Yukari Ogura[1]　　Ichiro Kobayashi[2]

[1] お茶の水女子大学理学部情報科学科
[1] Dept. of Information Sciences, Faculty of Science, Ochanomizu University
[2] お茶の水女子大学大学院人間文化創成科学研究科理学専攻
[2] Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Abstract: In this paper, we propose a method to raise the accuracy of text classification based on latent topics, reconsidering the techniques necessary for good classification – for example, to decide important sentences in a document, the sentences with important words are usually regarded as important sentences. In this case, *tf.idf* is often used to decide important words. On the other hand, we apply the PageRank algorithm to rank important words in each document. Furthermore, before clustering documents, we refine the target documents by representing them as a collection of important sentences in each document. We then classify the documents based on latent information in the documents. As a clustering method, we employ the k-means algorithm and investigate how our proposed method works for good clustering.

1　はじめに

近年，インターネットの発達に伴い，爆発的に増大した莫大な量のテキストデータを扱う問題がある．そのため大量のテキストを，自動でカテゴリごとに分類できるような文書分類手法が必要とされている．本研究では，文書の潜在的意味を考慮した分類手法を提案する．文書分類の方針として，まず語彙の重要度に基づき重要文抽出を行い，元の文書を重要文のみで構成し，分類対象となる文書の精錬化を図る．語彙の重要度を決める指標としては，一般に $tf \cdot idf$ や語彙の頻度などが用いられるが，本研究では，語の共起関係からグラフを構成し，PageRank アルゴリズムを用いて重要語の決定を行う．次に，潜在的意味解析手法を用いて，文書の潜在トピックごとの確率分布をもとに，k-means 法でクラスタリングを行う．実験を行い，語の重要度の決定に PageRank を用いた場合と，$tf \cdot idf$ を用いた場合の文書分類の精度を比較することにより，提案手法の有効性を検討する．

*連絡先：お茶の水女子大学理学部情報科学科小林研究室
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920509@is.ocha.ac.jp

2　関連研究

文書分類の研究において，分類精度を上げるため数多くの研究がなされており，特に，文書中の語の重要度を決めるアルゴリズムを改良することにより，分類精度の向上が出来ることが報告されている．Hassan ら [1] は，n-グラムを用いて，単語の共起関係に基づき構成したグラフに，PageRank アルゴリズムを用いることにより，文書分類の精度が向上することを示した．Zaiane ら [2] や，Wang ら [3] は，文書分類における，語の重要度の決定手法を提案した．Wang ら [3] は，語の重要度の決定に PageRank アルゴリズムを用いることが，文書分類に有効であることを示した．PageRank アルゴリズムは，センチメント分析や，トピック推定にも用いられており，Kubek ら [4] は，語の共起関係に基づき構成したグラフに，PageRank アルゴリズムを用いることにより，トピック推定を行っている．語の重みづけは，文書要約やにおいても重要な課題である．Erkan ら [5] は，LexRank や TextRank と呼ばれる，PageRank アルゴリズムを用いた文書要約の手法を提案している．文をノードとしてグラフを構成し，高い PageRank スコアを持つ，中心性の高い文を抽出することにより，文書要約を行っている．

本研究では，文書を潜在情報に基づいて分類することを目的とし，Newman ら [8] による潜在的情報の首尾一貫性は単語の共起関係により形成されるという報告を参考に，共起語からなるグラフを構築し，それに PageRank アルゴリズムを適用することにより，抽出された重要語から重要文を決定する．その重要文を用いて，潜在情報に敏感な文書群を再構成し，文書分類を行う手法を提案する．

3 提案方法

3.1 PageRank アルゴリズムによる重要語の決定

PageRank とは，Brin ら [6] によって提案された，Web ページ間に存在するハイパーリンク関係を利用することでページの順位付けを行うアルゴリズムである．PageRank の基本的な考え方は，推薦である．例として，図 1 の場合，V_a から V_b へリンクが張られているため，これは V_a から V_b への推薦と考えることができる．他の重要な Web ページから推薦されている Web ページは重要である，という考え方が PageRank において中心となっている概念である．Web ページをノード，ページ間のリンク関係をエッジとした有向グラフとして構成され，このグラフに基づいて順位のスコアが計算される．グラフ $G = (V, E)$ が与えられたときに，$In(V_a)$ は，点 V_a を指している点の集合，$Out(V_a)$ は，点 V_a が指している点の集合である．点 V_a の PageRank スコアは，式 (1) を反復的に処理することにより，全てのノードの PageRank スコアを求める．d は，制動係数 (dumping factor) であり，ある一定の割合でリンクのないノードからの影響を考慮するパラメータであり，$[0, 1]$ の値をとる．

$$S(V_a) = \frac{(1-d)}{N} + d * \sum_{V_b \in In(V_a)} \frac{S(V_b)}{|Out(V_b)|} \qquad (1)$$

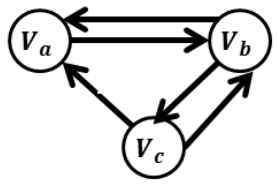

図 1: リンク関係の例

反復計算には，べき乗法を用いる．べき乗法とは，行列の主固有値と主固有ベクトルを見つけるための反復法であり，マルコフ連鎖の定常ベクトルがマルコフ行列の左側主固有ベクトルであること，および，求めたい PageRank ベクトルが Web ページ間のリンク関係を表した推移行列をもつマルコフ連鎖の定常ベクトルであることにより，PageRank の計算に用いられる．

語の重要度を決定するのには，$tf \cdot idf$ などが頻繁に用いられるが，語同士の様々な関係をグラフ構造で表現し，語の重要度を決定する手法が提案されている [3][1][10]．特に，Hassan ら [1] は，PageRank を用いてランクづけされた語の重要度は $tf \cdot idf$ よりも重要度を明確に差別化できることを示している．本研究でも彼らの手法を参考にして，語の重要度を PageRank アルゴリズムを用いて決定する．

3.2 潜在情報による分類

文書内の潜在的トピックの確率分布を表わすモデルとして Latent Dirichlet Allocation(LDA)[7] がある．このモデルでは，文書内にはいくつものトピックが潜在しており，トピックごとに出現しやすい単語があると考える．各トピックはそのトピックに対する出現確率を持った単語群で表され，複数文書内に存在している総単語に対して，各トピックごとに総和が 1 になる出現確率が割り当てられる．トピック自身にも文書セット内において出現確率の総和が 1 となるトピック比率として確率が付与される．本研究においては，文書に対する潜在トピックの確率分布を用いて，各文書をトピックで構成されるベクトルで表現し，文書間の類似度を測る．

3.3 提案手法における処理の流れ

本手法における，文書分類の流れを説明する．

step1 単語の共起関係の抽出
文書を文で区切り，文脈を考慮して，文中の単語の共起度を自己相互情報量 (PMI:Point-wise Mutual Information) に基づき算出する．

step2 重要単語の決定
step1 で得られた共起関係に基づき，ノードを単語，エッジの重みには PMI を用いたグラフを構成する．図 2 は，共起関係を基に構成したグラフの一例である．ここで，グラフを単語間の PMI で構成する理由は，文書分類を潜在的意味に基づき行うとしており，潜在トピックの一貫性は語の共起関係が影響を与えているとする Newman ら [8] の研究に基づき，潜在トピックを考慮した単語の重要度を算出するためである．このグラフに対し，多くの単語と高い共起度を持つ単語は重要

であると考え，PageRank アルゴリズムを用い，単語の重要度のランク付けを行う．

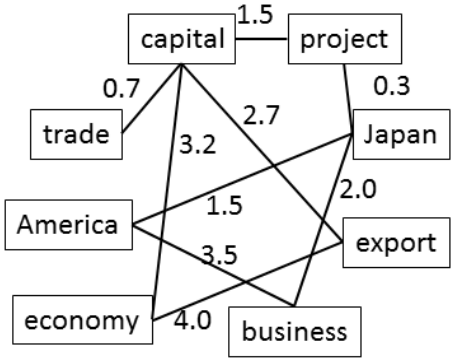

図 2: 類似度グラフ

step3 重要文の抽出
step2 で得られた単語のランキングに基づき，ランキング上位の単語を含む文を重要文とみなし，これを文書から抽出し，元の文書を重要文のみで構成する．

step4 分類
新たに構成された文書群に対し，LDA を用いてそれぞれの文書の潜在トピックごとの確率分布を得る．各文書のトピックに基づくベクトルを Jensen-Shannon 距離を用いて類似度を測り，k-means 法により分類する．

4 実験

4.1 実験仕様

実験対象データには，Reuters-21578[1] のテストセットからタグを除去したものを使用した．提案する手法は，対象文書から重要文を抽出し，文書を精錬してから文書分類を行うため，文数の少ない文書では提案手法の効果が判別できないため，1 文書中の文章数が 5 文以上である文書を利用した．カテゴリは，文書分類の他研究 [9]，[11] においても用いられている上位 10 件のカテゴリ，acq，corn，crude，earn，grain，interest，money-fx，ship，trade，wheat を利用した．その結果，文書数 792 件，語彙数 15,835 語，カテゴリ数 10 の文書群を対象に，ステミング処理とストップワード除去を施し実験を行った．

また，LDA で用いるパラメータは，$\alpha = 0.5$，$\beta = 0.5$ とし，サンプリングにはギブスサンプリングを用い，イテレーションは 200 回とした．トピック数は，パープレキシティにより決定することにした．トピック数を 1 から 30 まで変化させたときのパープレキシティの値の 10 回の平均をとり，パープレキシティが最小になるときのトピック数を最適トピック数とした．計算の結果，元の文書群のトピック数が 11 となった．重要文の抽出を行わない元の文書群の分類精度をベースラインとするため，実験に使用するトピック数は 11 とした．分類手法には，k-means 法を用い，トピックで構成された文書ベクトルを用いて分類を行う．

[1] http://www.daviddlewis.com/resources/testcollections/reuter21578

4.2 評価手法

評価には，文献 [9] を参考にして，正解率と F 値の 2 つの評価指標を用いる．文書 d_i に関して，l_i はクラスタリングアルゴリズムにより d_i に与えられたラベル，α_i は d_i の正解のラベルである．そのとき，正解率は式 (2) で表される．

$$\text{正解率} = \frac{\sum_{i=1}^{n} \delta\left(map\left(l_i\right), \alpha_i\right)}{n} \tag{2}$$

$\delta(x, y)$ は，$x = y$ ならば 1 となり，そうでなければ 0 となる関数である．$map(l_i)$ は，k-means 法により d_i に与えられるラベルである．

評価には，各カテゴリの F 値を求め，全カテゴリの平均を算出した．カテゴリ c_i の F 値は，精度を $P(c_i)$，再現率を $R(c_i)$ とすると，式 (3) のように表される．

$$F\left(c_i\right) = \frac{2 \cdot P\left(c_i\right) \cdot R\left(c_i\right)}{P\left(c_i\right) + R\left(c_i\right)} \tag{3}$$

カテゴリごとの F 値 (式 (3)) を測り，全カテゴリの平均を評価指標として用いた．(式 (4))

$$F = \frac{1}{|C|} \sum_{c_i \in C} F\left(c_i\right) \tag{4}$$

また，k-means 法において初期値には，それぞれのカテゴリの正解データの文書ベクトルをランダムに選び，1 つ与えることにする．分類する際，文書群におけるカテゴリ数 k を事前に知っていること，それぞれのカテゴリから 1 つだけ正解例を見つけることは，計算コストがかからないことから，妥当な方法であると判断できる．この方法により，分類結果のクラスが，どのカテゴリであるか判断できるようになる．

4.3 実験結果

k-means 法を 10 回行い，その平均値を測った．ただし，LDA を用いて，文書のトピックごとの確率分布から分類を行う場合には，出力される確率分布 θ が毎回

変化する．そのため 1 つの θ に対して k-means 法を 10 回行い，これを 8 セット行ったときの平均値を測った．確率分布を用いて分類を行う場合には Jensen-Shannon 距離を，文書ベクトルを用いて分類を行う場合にはコサイン類似度を用いた．重要文抽出を行った場合の結果を表 1，行っていない場合の結果を表 2 に示す．また，重要文を行った後の文書群の語彙数，文数の変化をそれぞれ表 3，表 4 に示す．また，表 5，表 6 では，PageRank，$tf \cdot idf$ のそれぞれの指標を用いて重要単語のランクを決定し，それに基づき，同じ数だけ文を抽出し，トピック数を変化させて分類を行った場合の実験結果を示す．

表 1: 重要文抽出した場合

単語の重要度	類似度指標	正解率	F 値
PageRank	Jenshen-Shannon 距離	**0.5671**	**0.4852**
	コサイン類似度	0.2870	0.2906
$tf \cdot idf$	Jenshen-Shannon 距離	0.5500	0.4347
	コサイン類似度	0.2753	0.2701

表 2: 重要文抽出しない場合

類似度指標	正解率	F 値
Jenshen-Shannon 距離	0.5177	0.4262
コサイン類似度	0.2875	0.3048

表 3: 語彙数の変化

手法	1 語	2 語	3 語	4 語	5 語
PageRank	12,268	13,141	13,589	13,738	13,895
$tf \cdot idf$	13,999	14,573	14,446	14,675	14,688

表 4: 文数の変化

手法	1 語	2 語	3 語	4 語	5 語
PageRank	1,244	1,392	1,470	1,512	1,535
$tf \cdot idf$	1,462	1,586	1,621	1,643	1,647

表 5: 正解率

トピック数	8	9	10	11	12
PageRank	0.525	0.535	**0.566**	0.553	0.524
tf.idf	0.556	0.525	0.557	0.550	0.541

考察

実験結果より，Jenshen-Shannon 距離を用いて分類を行った場合においては，重要文抽出を行った場合の方が，行わない場合よりも正解率，F 値ともに値が良くなるということが分かった．このことから，重要文抽出することにより，文書が精錬されていることが確認された．文書が精錬されたことにより，文書の特徴を表現するのに必要な文のみが残り，文書のトピックごとの確率分布の差が測りやすくなったのではないかと考えられる．また重要文抽出に関して，$tf \cdot idf$ を用いた場合に比べ，PageRank を用いて重要文の抽出を行った場合に文書分類の精度の向上が見られた．このことから，文書の 3 文中での単語の共起関係からグラフを構成し，単語の重要度を PageRank アルゴリズムを用いて決定することにより，文脈を考慮した単語の重要度が得られていると考えられる．

また表 3，表 4 から，重要文抽出したあとの語彙数，文数の比較では，$tf \cdot idf$ と比較して，PageRank を用いた場合に，より語彙数，文数が減っていることが分かる．$tf \cdot idf$ の場合，特定の文書に多く出現している単語の値が高くなるため，$tf \cdot idf$ が高い単語は，その文書中の多くの文に出現している可能性が高い．そのため，$tf \cdot idf$ の高い単語を含む文を抽出すると，自然と多くの文を抽出することになるのではないかと考えられる．

コサイン類似度で分類を行った場合において，精度が良くなかった原因としては，実験結果から，Jenshen-Shannon 距離と比較すると，文書間の類似度の値の差が小さいことが観測されており，そのため異なるカテゴリの文書の判別がうまくいかなかったのではないかと考えられる．

文書群から抽出する文数を同じにし，トピック数を変化させて分類を行った場合，トピック数が 9,10,11 の時に $tf \cdot idf$ を用いた場合よりも，PageRank を用いた場合に分類精度が上回った．また，PageRank，$tf \cdot idf$，どちらを用いた場合でもトピック数を 10 に設定した時に，高い精度となった．これは，文書群の正解カテゴリが 10 であることから，正解のカテゴリ数と設定したトピック数が一致しているため，分類精度が高くなったのではないかと考えられる．トピック数 10 の時に PageRank と $tf \cdot idf$ での結果を比較すると，PageRank を用いた場合に良い精度となっている．このことからも，PageRank アルゴリズムにより決定した単語の重要度が，分類精度の向上に寄与することが分かる．

表 6: F 値

トピック数	8	9	10	11	12
PageRank	0.431	0.431	**0.467**	0.460	0.434
tf.idf	0.466	0.430	0.461	0.435	0.445

5 おわりに

本研究では，PageRank を用いた重要語の抽出を行い, それに基づいて重要文を抽出し, 潜在的意味によるクラスタリングを行う手法を提案した．分類対象データとして，Reuters-21578 を用いて実験を行った．提案手法の有効性を検証するため，重要文の抽出に語の重要度を PageRank，または $tf \cdot idf$ を用いて行い，重要文によって再構成された文書集合に対して，潜在情報または表層情報に基づき，k-means 法を用いてクラスタリングを行った．その結果，全体として表層情報よりも潜在情報を用いた分類の方が精度が良く，重要文を抽出する際に，語の重要度を PageRank を用いて決める方が分類精度が向上することがわかった．重要文抽出した後の語彙数の比較から，PageRank を用いた場合のほうが抽出される文章数が少ないということが推測できるため，より文脈を考慮した重要文の抽出がなされているのではないかと考えられる．

今後の課題としては, どの程度の重要文をどのように選択したかにより, 分類の精度が変化すると考えられるため, 適切な重要文選別方法を考察するつもりである．また, 現在は k-means 法での分類しか行っていないため, 他の多クラス分類手法との比較を行うつもりである．

参考文献

[1] Samer Hassan, Rada Mihalcea, Carmen Banea.: Random-Walk Term Weighting for Improved Text Classification, (2007)

[2] Osmar R.Zaiane, Maria-luiza Antonie.: Classifying Text Documents by Associating Terms with Text Categories, *In Proc. of the Thirteenth Australasian Database Conference(ADC'02)*, pp. 215–222

[3] Wei Wang, Diep Bich Do, and Xuemin Lin.: Term Graph Model for Text Classification, *Springer-Verlag Berlin Heidelberg 2005*, pp. 19–30 (2005)

[4] Mario Kubek, Herwig Unger.: Topic Detection Based on the PageRank's Clustering Property, *IICS'11*, pp. 139–148 (2011)

[5] Gunes Erkan.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *Journal of Artificial Intelligence Research 22*, pp. 457–486 (2004)

[6] Sergey Brin, Lawrence Page.: The Anatony of a Large-scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, pp. 107–117 (1998)

[7] David M. Blei, Andrew Y. Ng, Michael I. Jordan, John Lafferty.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, p. 993–1022 (2003)

[8] Newman David, Lau Jey Han, Grieser karl, Baldwin Timothy.: Automatic evaluation of topic coherence, *Human Language Technologies :The 2010 Annual Conference of the North Ametican Chapter of the Association for Computational Linguistics*, pp. 100–108 (2010)

[9] Gunes Erkan.: Language Model-Based Document Clustering Using Random Walks, *Association for Computational Linguistics*, pp. 479–486 (2006)

[10] Christian Scheible, Hinrich Shutze.: Bootstrapping Sentiment Labels For Unannotated Documents With Polarity PageRank, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, (2012)

[11] Amarnag Subramanya, Jeff Bilmes.: Soft-Supervised Learning for Text Classification, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1090–1099, Honolulu (2008)