6Q-5

PageRank アルゴリズムを用いた重要文抽出による潜在的意味に基づく文書分類

小倉由佳里 †　　小林一郎 ††

† お茶の水女子大学理学部情報科学科　†† 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

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．そこで本研究では，文書の潜在的意味を考慮した分類手法を提案する．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 tfidf 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．

2　PageRank アルゴリズム

PageRank とは，Brin ら [1] 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．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．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，$[0, 1]$ のパラメータである．反復計算にはべき乗法を用いる．

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

3　Latent Dirichlet Allocation

複数文書内の潜在トピックを確率的に求めるトピックモデルとして Latent Dirichlet Allocation(LDA)[2] がある．このモデルでは，文書はトピックの集まりによって構成されており，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，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．

4　提案手法

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．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 [3] 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 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．

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．

5　実験

5.1　実験仕様

対象データには，Reuters-21578† のテストセットからタグを除去したものを使用した．重要文抽出の有効

† http://www. daviddlewis. com/resources/testcollections/reuter21578

Text Classification based on Latent Topics of Important Sentences Extraction by PageRank Algorithm

† Yukari OGURA(g0920509@is. ocha. ac. jp)
†† Ichiro KOBAYASHI(koba@is. ocha. ac. jp)
Department of Informatoin Science, Faculty of Sciences, Ochanomizu University (†)
Advanced Sciences, Graduated School of Humanities and Sciences, Ochanomizu University (††)
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

性を検討するため，1 文書中の文章数が 5 文以上である文書を利用した．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．しかし，重要文抽出した文書群のトピック数を測った結果，クラスタリング目標のカテゴリ数より少なくなったため，本実験ではトピック数を 11 に設定した．

5.2 評価

評価には，文献 [4] を参考にして，相互情報量 (MI:Mutual Information) を用いる．$L = \{l_1, l_2, \cdots, l_k\}$ は，k-means 法により分類された文書のラベルの集合，$A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ は，分類された文書の正解のラベルである．これら 2 つのラベルの相互情報量は，式 (2) で定義される．

$$MI(L, A) = \sum_{l_i \in L, \alpha_j \in A} P(l_i, \alpha_j) \cdot \log_2 \frac{P(l_i, \alpha_j)}{P(l_i) \cdot P(\alpha_j)} \quad (2)$$

$P(l_i)$ は，アルゴリズムによって l_i にラベル付けされる確率，$P(\alpha_j)$ は，正解データにおいて α_j である確率，$P(l_i, \alpha_j)$ は，これら 2 つが同時に起こる確率である．相互情報量を [0，1] の値で得られるようにするため，正規化を行う．正規化された相互情報量は式 (3) で表される．

$$\widehat{MI} = \frac{MI(L, A)}{MI(A, A)} \quad (3)$$

また，k-means 法において初期値にそれぞれのカテゴリの正解データを 1 つ与えることにより，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する．

5.3 実験結果

k-means 法を 10 回行い，その平均値を測った．LDA を用いて，文書のトピックごとの確率分布で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．

表 1: 重要文抽出した場合

単語の重要度	類似度指標	$\widehat{MI}$
PageRank	Jenshen-Shannon 距離	0．6883
	コサイン類似度	0．1549
$tf \cdot idf$	Jenshen-Shannon 距離	0．6273
	コサイン類似度	0．1314

表 2: 重要文抽出しない場合

類似度指標	$\widehat{MI}$
Jenshen-Shannon 距離	0．6190
コサイン類似度	0．1616

5.4 考察

実験結果より，重要文抽出を行った場合において，行わない場合よりも精度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．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．

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，Jenshen-Shannon 距離と比較すると，文書間の類似度の値の差が小さく，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．

6 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った．実験を行った結果，$tf \cdot idf$ を用いた場合と比較して，PageRank を用いた場合に，分類精度が向上することが分かった．

今後の課題としては，重要文をどの程度選択したかにより，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したいと考えている．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での比較を行うつもりである．

参考文献

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

[2] David M.Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. : Latent dirichlet allocation, Journal of Machine Learning Research, Vol.3，pp.993–1022, 2003.

[3] Newman, David and Lau, Jey Han and Grieser, karl and 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, Los Angeles, 2010.

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