単語の共起グラフを用いた重要文抽出による文書分類

小倉由佳里　(指導教員：小林一郎)

1　はじめに

近年，潜在意味解析手法の発展に伴い，様々な文書処理への応用がなされている．本研究では，文書分類において重要な情報となる文書中の重要語を，潜在的意味で文書を分類するのに適するように決め，分類対象とする文書をその重要語を用いて抽出された重要文集合とする，潜在意味に基づく文書分類手法を提案する．

2　提案手法

2.1　PageRank アルゴリズムによる重要語の決定

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 [2]．Hassan ら [2] 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．

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．点 V_a の PageRank スコアは，式 (1) を反復的にべき乗法を適用することにより，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．

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

2.2　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)[3] がある．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．

2.3　提案手法における処理の流れ

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 [4] 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．

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．

3　実験

3.1　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．重要文抽出の有効性を検討するため，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．重要文抽出を行わない場合の元の文書群の分類精度の結果を基準とするため，元の文書の最適トピック数 11 を，本実験のトピック数と設定した．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．

3.2　評価手法

評価には，文献 [5] 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．

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

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

$\delta(x, y)$ は，$x = y$ ならば 1 になり，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．

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

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

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

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．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．

3.3 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 10 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，重要文抽出を行った後の文書群の語彙数の変化を表 3 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	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	3 語	4 語	5 語
PageRank	13,589	13738	13895
$tf \cdot idf$	14,446	14675	14688

3.4 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，文書が分類に適するように精錬されていることが確認された．また表 3 から，重要文抽出した後の語彙数の比較では，$tf \cdot idf$ と比較して，PageRank を用いた場合に，より語彙数が減っていることが分かる．このことから，重要文抽出の段階において，PageRank を用いた場合の方が，抽出される文章数が少ないことが推測できる．少ない文章数で，よい精度が出ているため，PageRank を用いて精錬化を行った文書群は，クラスタリングに対して適切な文章が抽出されていると考えられる．$tf \cdot idf$ 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章を抽出すると，自然と多くの文章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，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．

4 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．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，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，より文書分類に適う内容を捉え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．

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

[3] 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.

[4] 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.

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