文書分類と複数文書要約を対象とした文書内の重要情報抽出

理学専攻 情報科学コース
小倉 由佳里

1 はじめに

ウェブの発達により，大量のテキストデータを効率的に扱う技術への需要が高まっている．その課題に対する技術として，文書分類と自動文書要約は，盛んに研究がなされている技術である．本研究では，分類精度向上を目的とした，単語の共起グラフを用いた重要文抽出に基づく文書の潜在意味による分類手法の提案および多目的最適化による複数文書要約手法の提案を行う．文書分類では，単語の重要度算出に PageRank[1] アルゴリズムを用い，潜在的意味解析に LDA[2] を用いた手法を開発した．文書要約は，文の組合せ最適化問題として解かれることが多い．本研究では，組合せ最適化に遺伝的アルゴリズム (GA:Genetic Algorithm) による多目的最適化を用いることで，実用的な時間で質の良い要約を生成する手法を開発した．

2 PageRank を用いた重要文抽出による潜在意味に基づく文書分類

2.1 提案手法のアルゴリズム

本手法における，文書分類の流れを説明する．

step i．単語の共起関係の抽出
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 [6] の先行研究に基づき単語の共起関係を抽出する．文書を文を区切り，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．共起を抽出するウィンドウサイズを 3 文とする．

step ii．重要単語の決定
step1 で得た共起関係に基づき，グラフを構成する．グラフのノードは単語，エッジの重みには PMI を用いる．潜在トピックを考慮した単語の重要度を算出するために構成したグラフに，PageRank アルゴリズム [1] を適用し，単語の重要度のランク付けを行う．

step iii．重要文の抽出および文書の再構成
step2 で得られた単語の重要度ランキングに基づき，重要文を決定し抽出する．重要文は，重要度が高い単語を含む文とする．さらに分類対象文書を，抽出した重要文のみで再構成する．

step iv．文書分類
再構成された文書群に対し，LDA[2] 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．分類に用いる素性は，各文書のトピックに基づく文書ベクトルとする．この文書ベクトル間の類似度は，Jensen-Shannon 距離 (JS 距離) を用いて測り，分類手法には，k-means 法を用いる．

2.2 実験

2.2.1 実験設定

実験対象データには，Reuters-21578[1] と 20Newsgroups[2] を使用した．使用する文書は，1 文書中の文章数が 5 文以上である文書とした．各データセットに対し，タグの除去，ステミング処理，ストップワード除去を施した．Reuters-21578 では，カテゴリ数 10 の 792 文書を用いた．20Newsgroups では，カテゴリ数 4 の 800 文書を用いた．LDA で用いるパラメータは，$\alpha = 0.5$，$\beta = 0.5$ とし，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した．抽出する文は，重要度の高い上位 3 単語を含む文とし，文書の再構成を行った．また，k-means 法での初期値には，各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 つ与えることとした．

評価指標には，2 つの指標の正解率と F 値を用いる．

2.2.2 実験結果および考察

分類の正解率と F 値の結果を，表 1，表 2 に示す．実験より，データセットによって異なる結果となった．Reuters-21578 では，重要文抽出により文書が精錬されたことから，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た．しかし，20Newsgroups を用いた実験では，重要文抽出を行わない場合の方が，行う場合よりも精度が高い結果となった．これは，データセットの性質の違いであると考えられる．20Newsgroups は，単語数が Reuters-21578 より少ない．そのため，重要文抽出を行ったことにより，単語数がさらに減り，本来大量の文書の下で行う学習の効果が下がり，LDA の学習精度が下がったのではないかと考えられる．重要文抽出に関しては，Reuters-21578，20Newsgroups 共に，$tf \cdot idf$ を用いた場合に比べ，PageRank を用いた場合に分類の精度の向上が見られた．このことから，単語の共起関係のグラフから，単語の重要度を PageRank により算出することで，分類に適した単語の重要度が得られることが検証された．

表 1: 正解率

単語の重要度	Reuters-21578	20Newsgroups
PageRank	**0.5671**	0.6415
$tf \cdot idf$	0.5500	0.5915
重要文抽出なし	0.5177	**0.8563**

表 2: F 値

単語の重要度	Reuters-21578	4-News
PageRank	**0.4852**	0.6321
$tf \cdot idf$	0.4347	0.5091
重要文抽出なし	0.4262	**0.8494**

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2] http://qwone.com/ jason/20Newsgroups/

3 多目的 GA を用いた要約生成

3.1 要約生成のための多目的遺伝的アルゴリズム

多目的 GA には，Deb らにより開発された NSGA-II[3] を用いる．

3.1.1 個体と表現方法

各個体は，解の候補を示し，出力する要約の文の組合せを表現している．個体における 1 つのマスが遺伝子座であり，各遺伝子座は，要約対象の文書群が含む各文に対応する．i 番目の遺伝子座の持つ値が “1” である時，文 s_i は要約に含まれることを示し，遺伝子座の持つ値が “0” である時，要約に含まれないことを示す．

3.1.2 交叉

交叉は，1 点交叉，または 2 点交叉を行う．交叉後に，個体の示す要約候補が持つ文字数に応じて文字数に関する操作を行う．この目的は，制約文字数を大幅に超えた個体が生成されることを防ぐことと，交叉により全ての遺伝子座の持つ値が “0” である個体が生成されることを防ぐことである．

3.1.3 目的関数

3 つの目的関数を用いる．式 (1) は，単語の出現頻度による文のスコアを算出する．これは，人が作った要約に多く含まれる単語は，要約対象の文書でも多く出現するという Nenkova ら [5] の先行研究に基づく．

$$f(x) = \sum_{i=1}^{N} \frac{1}{w_i} x_i \tag{1}$$

$$where,\ \ w_i = \sum_{t_i \in s_i} \frac{p(t_i)}{|\{t_i | t_i \in s_i\}|},\ \ p(t_i) = \frac{c}{n}$$

$x_i = \{0, 1\}$ であり，文 s_i が要約に含まれる時 $x_i = 1$，そうでない時 $x_i = 0$ となる．c は，単語 t_i の出現回数，n は総単語数である．式 (2) は，文の出現位置による文のスコアを算出するもので，文の最初と最後に出てくる文は重要な文であるという仮定に基づいている．

$$g(y) = \sum_{i=1}^{N} \frac{1}{p_i} y_i \tag{2}$$

$$where,\ \ p_i = max(\frac{1}{i}, \frac{1}{|S| - i + 1})$$

$y_i = \{0, 1\}$ であり，文 s_i が要約に含まれる時 $y_i = 1$，そうでない時 $y_i = 0$ となる．i は文 s_i の文書 d_j での出現位置，$|S|$ は文書 d_j の総文数である．式 (3) は，生成する要約と，要約対象の文書群の単語出現分布の類似度を，JS 距離で測るものである．

$$h(z) = JS(z||D) \tag{3}$$

z は，要約に含まれる文集合の単語出現頻度分布であり，D は要約対象の文書群の単語出現頻度分布である．

3.2 実験

3.2.1 実験設定

対象データには DUC2004[3]で使われた 10 件の新聞記事群 50 セットを用い，各文書セットに対して，長さ 665 バイト以内の要約を生成する．評価指標には ROUGE-1，ROUGE-2 を用いる．各文書セットあたり 10 回試行し，その平均値を測る．NSGA-II の設定は，個体数 50，世代数 50，交叉率 1.0，突然変異率 0.1 とした．

[3]http://duc.nist.gov/duc2004/tasks.html

3.2.2 実験結果および考察

結果を表 3 に示す．他の手法との比較より，提案手法は，1 つの要因の目的関数に関して，貪欲法で最適化を行う手法よりも高い ROUGE 値を示していることから，要約生成に関して考慮すべき要因は複数必要であることが考えられる．出力される要約に関しては，本手法で出力される解はパレート最適解であるが，他の最適化手法に劣らない要約の出力が可能であることを示している．またパレート最適解には，複数の要因を適当な度合で考慮した要約が出力されていると考えられる．

表 3: ROUGE 値

システム	ROUGE-1	ROUGE-2
提案手法	37.98	9.48
FreqSum[5]	35.30	8.11
Greedy-KL[4]	37.98	8.53
LR[7]	39.00	9.60

4 おわりに

本研究では，重要文抽出による潜在意味に基づく文書分類手法と，多目的 GA を用いた複数文書要約手法の提案をした．前者では，データセットにおいて結果に違いが見られ，その理由はデータセットの性質の違いであると考察した．後者では，提案手法では，他の単目的最適化手法よりも高い ROUGE 値が得られたことから，要約生成においては複数の目的関数が必要であること，パレート最適解であっても質の良い要約の生成が可能であることが示唆された．今後の課題としては，異なるデータセットを用いた実験やパレート最適解における ROUGE の比較を行いたいと考えている．

参考文献

[1] S. Brin, and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems 30.1*, pp. 107–117.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research, vol.3*, pp. 993–1022.

[3] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan. 2002. A fast and elitist multi-objective genetic algorithm: NSGA2. *IEEE Transactionon Evolutionary Computation 6.2*, pp. 149–172.

[4] A. Haghighi, and L. Vanderwende. 2009. Exploring content models for multi-document summarization. *NAACL*, pp. 362–370.

[5] A. Nenkova, L. Vanderwende, and K. McKeown 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. *ACM SIGIR*, pp. 573–580.

[6] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. 2010. Automatic evaluation of topic coherence. *NAACL*, pp. 100–108.

[7] M. Nishino, N. Yasuda, T. Hirao, J. Suzuki, and M. Nagata. 2013. Lagrangian Relaxation for Scalable Text Summarization while Maximizing Multiple Objectives. *Information and Media Technologies 8.4*, pp. 1017–1025.