リンク解析に基づく複数文書の自動要約

酒井明奈　　福本文代
山梨大学工学部コンピュータ・メディア工学科
{t05kg014, fukumoto}@yamanashi.ac.jp

1　はじめに

現在 Web 上では, あらゆる情報が氾濫している. さらにこれらは時間の経過と共に次々に更新されるため, ある話題に関する古い情報から新しい情報まで全てを読み, 理解することは容易ではない. そこで同じ話題に関する複数の文書から, 各文書の内容を端的に示す文のみを抽出し, それをまとめて要約とする研究が多く行われている. 文抽出に基づく文書要約手法として, 近年 Markov Random Walk モデルを利用した要約手法が提案されている [1, 2]. これは, グラフのノードとリンクを要約対象となる文書中の各文と文間の類似度で表現し, グラフに対して PageRank や HITS モデルなどのランキングアルゴリズムを適用することで重要な文を抽出する手法である.

Wan らは, 冗長な文の抽出を避け, 各文書の内容を端的に示す文のみを高精度で抽出するため, 従来の Markov Random Walk モデルに対し, 話題という観点で要約対象となる文を分類した結果を取り入れた Cluster-based Conditional Markov Random Walk(ClusterCMRW) モデルを提案した [3]. 彼らは, k-means を用いて同じ話題を持つ文を同一のクラスタに分類した. しかし, k-means は, 少数の文から成る文書を要約対象とした場合には高い精度が得られるものの, 大量の文数から成る複数文書の場合にはクラスタの割り当てに偏りが生じるため精度が下がってしまう. 本研究では, Weiss らが提案した Spectral Clustering[4] を用いて次元削減を行うことで, 要約の精度を向上させることを目的とする.

2　ClusterCMRW モデル

Wan らの ClusterCMRW モデルは, 複数文書に含まれる類似した文の集合であるクラスタの重要度と, 文とクラスタの相互関係を考慮した複数文書要約を実現するため, 文とクラスタから成る 2 層のリンク構造に対して Markov Random Walk モデルを適用した手法である. ClusterCMRW モデルを図 1 に示す.

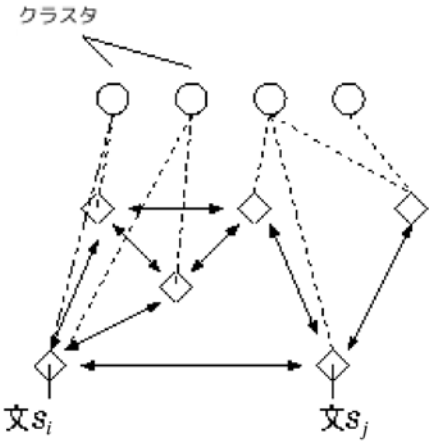

図 1: ClusterCMRW モデル

図 1 において, 四角は要約対象となる文を示し, 丸印はクラスタを表す. また, 矢印は文間の関係を示し, 破線は文とクラスタとの関係を表す. 下層は Markov Random Walk モデルと同様であるが, ClusterCMRW モデルは上層においてクラスタとの関係を加えている. クラスタに属する文 s_i から s_j への関係の度合を示す確率は, 式 (1) で示される.

$$p(i{\rightarrow}j|clus(s_i), clus(s_j)) = \begin{cases} \dfrac{f(i{\rightarrow}j|clus(s_i), clus(s_j))}{\sum_{k=1}^{|V|} f(i{\rightarrow}k|clus(s_i), clus(s_k))}, & \text{if } \Sigma f \neq 0 \\ 0\ , \text{ otherwise} \end{cases} \quad (1)$$

式 (1) において, $|V|$ は要約対象となる文の個数を示す. $f(i{\rightarrow}j|clus(s_i), clus(s_j))$ はクラスタに属する文 s_i から s_j へのリンクの重みを示し, 式 (2) を用いて求める.

$$\begin{aligned} &f(i{\rightarrow}j|clus(s_i), clus(s_j)) \\ &\quad = f(i{\rightarrow}j){\cdot}(\lambda{\cdot}\pi(clus(s_i)){\cdot}\omega(clus(s_i)) \\ &\qquad +((1-\lambda){\cdot}\pi(clus(s_j)){\cdot}\omega(clus(s_j)) \end{aligned} \quad (2)$$

$clus(s_i)$ はクラスタ c_{ij} に文 s_i が含まれていることを示す. π は複数文書全体におけるクラスタ $clus(s_i)$

の重要度である．また ω は s_i が $clus(s_i)$ の中でどの程度重要であるかを示す．π 及び ω は余弦尺度を用いて算出する．λ は原因となるクラスタから目的のクラスタへ相対的な助力を制御する結合重みを示す．ClusterCMRW モデルでは, 話題ごとのクラスタ $clus$ を得るために, k-means を用いる．k-means は非階層的クラスタリング手法の 1 つであり, クラスタの重心値を用いて与えられた k 個のクラスタに割り当てる手法である．クラスタの重心値は割り当てられたデータの平均値によって求める．各文は各クラスタの重心値との距離を測り, 距離の近いクラスタに再度割り当てられる．これをクラスタの重心値が変化しなくなるまで繰り返すことで同じ話題に属する文を同一のクラスタに割り当てる．またクラスタ数は文章数 V としたとき, 式 (3) とする．

$$k = \sqrt{V} \tag{3}$$

要約対象となる n 文から成る文の集合を $\{S_i\}_i = 1^n$ とする．各文間の類似度を式 (1) により求め, $n \times n$ 行列を作成する．要約となる文は, この行列の絶対値が最大となる固有値に対応する固有ベクトルを求めることにより得られる．本研究では ClusterCMRW モデルにおいて, 話題という観点で要約対象となる文を分類する際に Spectral Clustering を用いて分類することで最終的に要約となる文を抽出した．

3 Spectral Clustering による文の分類

Spectral Clustering は高次元空間においてデータ $\{S_i\}_{i=1}^n$ を k 次元空間の特徴空間に次元削減し, 特徴空間内で k-means 法を適用する非階層的クラスタリング手法であり [4], 単語の意味分類など自然言語処理の多くのタスクで利用されている [5]．Spectral Clustering を用いた文の分類手順を以下に示す．

1. n 個からなる文の集合を $S = s_1, s_2, s_3, ..., s_n$ とする．ここで $\{S_i\}_{i=1}^n$ は出現回数を要素としたベクトルで表現されている．

2. ユークリッド距離尺度により, 各文間の類似度を求め, 式 (4) を用いて, $n \times n$ 行列 A を作成する．

$$A_{ij} = \begin{cases} exp\left(-\frac{D_{ij}^2}{2\sigma^2}\right) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{4}$$

式 (4) 中, D_{ij} は, 文 s_i と s_j の類似度であり, 式 (1) を用いて求める．σ はパラメータである．本研究では, 訓練データによるパラメータ学習により, σ=4.5 とした．

3. 対角行列 B を考える．対角要素 (i, i) は行列 A の i 行目の要素の合計 $\sum_j^n A_{ij}$ とする．この対角行列 B を利用し, 式 (5) を用いて行列 L を構築する．

$$L = B^{-1/2} A B^{-1/2} \tag{5}$$

4. 行列 L の固有値, 固有ベクトルを求める．ここで必要なクラスタ数を k 個とする．このとき上位 k 個の固有値に伴う固有ベクトルを取得し, $n \times k$ 行列となる行列 X を構築する．

5. 行列 X を正規化し, 行列 Y として構築する．

6. 行列 Y の各行を k 次元空間上にプロットする．これを用いて k-means でクラスタリングを行う．

7. 行列 Y の i 行目がクラスタ c に割り当てられたとき, データ s_i はクラスタ c に割り当てられる．

Spectral Clustering を用いてクラスタの割り当てをした後, ClusterCMRW モデルに適用して重要度を算出し, 重要度の高い順に一定個数の文を抽出する．

4 実験

4.1 実験データと評価方法

Spectral Clustering を用いた ClusterCMRW モデルの有効性を検証するため, 情報検索システム評価用テストコレクション構築プロジェクト (NTCIR) の NTCIR-3 SUMM を用いて実験を行った．NTCIR による要約文は毎日新聞の 1998 年と 1999 年の 2 年間分の新聞記事の話題を用いて作成されているため, 同年度の記事を使用した．また, テストデータとして用いた話題数は Formalrun の FBFREE の 30 とし, パラメータ推定のため, 訓練データとして Dryrun の FBFREE から 3 つの話題を無作為に抽出したものを用いた．文の抽出数は NTCIR による正解データの要約文数と同じ個数とした．NTCIR-3 SUMM ではそれぞれの話題に長さの異なる 2 種類の要約を作成しているので, 長い要約を long, 短い要約を short とし, 合計 60 の結果を用いて手法の有効性を検証した．実験では, k-means のみによる手法と提案手法によって抽出した文とのそれぞれの要約結果を比較した．

表 1: 訓練データによる余弦尺度の結果

		余弦尺度	
ID	文数	k-means	Spectral
0020	22	0.4130	0.4136
0040	51	0.1227	0.4003
0100	24	0.4035	0.4749

評価尺度は余弦尺度と DUC[1] で用いられている ROUGE-score を用いて評価した. 余弦尺度による評価は, 正解要約文と各手法で得られた要訳文との余弦尺度を求めた結果であり, 1 に近いほど正解要約文と類似していることを示す. また, ROUGE-score を式 (6) に示す.

$$ROUGE = \frac{\sum_{S \in \{RefSum\}} \sum_{word \in S} Count_{match}(word)}{\sum_{S \in \{RefSum\}} \sum_{word \in S} Count(word)} \quad (6)$$

式 (6) の $word$ は形態素解析「茶筅」[6] を用いて得られた単語とした.

5 実験結果

パラメータ推定のため訓練データを用いた結果を表 1 に示し, テストデータを用いた実験結果を表 2, 及び表 3 に示す.

表 2 は短い語彙数からなる要約での結果をを示し, 表 3 は長い要約での結果を示す. 余弦尺度の結果を比較すると, k-means の方が値が高いものが short, 及び long でそれぞれ 6 と 6 話題, 本手法による結果が高い値が得られているものが 10, 13 話題であることから, 全体的に Spectral Clustering の方がよい結果が得らえれていることがわかる. 次に再現率では表 2 の short での結果を見ると, 精度が向上したものとそうでないものとの半々の結果が得られた. また表 3 の long での結果を見ると, short よりも比較的 Spectral Clustering の方が精度が高くなっていることがわかる. 話題ごとにみると, Spectral Clustering を用いることで最も精度 (余弦尺度) が向上した話題は ID0210 であった. 要約対象となる文数を求めると 191 文であり他の話題よりも文数が多かったことから, Spectral Clustering は比較的長い文書を対象とした場合にも有効であると言える.

[1] http://www-nlpir.nist.gov/projects/duc/guidelines/2001.html

6 おわりに

本研究では, ClusterCMRW にモデルに対して意味的に類似した文同士を一つの話題として高精度で分類するため Spectral Clustering を用いて次元圧縮を行うことで重要文を抽出する手法を提案した. NTCIR テストコレクションを用いた実験の結果, k-means 法を用いた従来手法よりも高い精度で重要文を抽出できることを確認した. 一般にクラスタリング手法を含む学習アルゴリズムはパラメータ値により精度が変わるため, その空間を探索することが重要である. 従ってパラメータ値の細かい設定による精度の検証を行う必要がある. またさらなる精度向上のため, EM や IB(Information Bottleneck) 法 [7, 8] など他のクラスタリング手法を用いた検証も今度の課題である.

参考文献

[1] G.Erkan and D.Radev, LexPageRank: Prestige in Multi-Document Text Summarization. In Proc. of EMNLP'04 pp. 365-371, 2004.

[2] R.Mihalcea and P.Tarau, Language Independent Extractive Summarization, In Proc. of ACL, pp. 49-52, 2005.

[3] X.Wan and J.Yang, Multi-Document Summarization using Cluster-based Link Analysis, In Proc. of the 31st ACM SIGIR, pp. 299-306, 2008.

[4] Y.Weiss, Segmentation using Eigenvectors: A unifying view. In ICCV (2), pp. 975-982, 1999.

[5] C.Brew and S.S.Walde, Spectral Clustering for German Verbs, In Proc. of EMNLP2002, pp. 117-123, 2002.

[6] 松本裕治他形態素解析システム「茶筅」version 2.2.1 使用説明書, 奈良先端科学技術大学院大学松本研究室, 2000.

[7] A.P.Dempster and N.M.Laird and D.B.Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the RST, 39(B), pp. 1-28, 1977

[8] N.Tishby and F.C.Pereira and W.Bialek, The Information Bottleneck Method, In Proc. of the 37th Annual Alleraton Conference on Communication, Control and Computing, pp. 368-377, 1999.

表 2: NTCIR データを用いた要約 (short) 結果

		余弦尺度		ROUGE				余弦尺度		ROUGE	
ID	文数	k-means	Spectral	k-means	Spectral	ID	文数	k-means	Spectral	k-means	Spectral
0010	12	0.786	0.786	0.127	0.127	0160	23	0.452	0.452	0.235	0.235
0020	5	0.191	0.191	0.211	0.211	0170	2	0.771	0.206	0.188	0.149
0030	12	0.643	0.651	0.275	0.269	0180	9	0.495	0.495	0.259	0.259
0040	25	0.595	0.595	0.106	0.106	0190	10	0.836	0.836	0.325	0.325
0050	8	0.221	0.475	0.097	0.124	0200	12	0.855	0.855	0.145	0.145
0060	24	0.543	0.543	0.151	0.151	0210	16	**0.043**	**0.635**	0.092	0.117
0070	12	0.832	0.832	0.212	0.212	0220	15	0.686	0.699	0.239	0.201
0080	15	0.634	0.675	0.211	0.206	0230	9	0.730	0.745	0.094	0.090
0090	10	0.221	0.475	0.097	0.124	0240	12	0.777	0.610	0.156	0.223
0100	13	0.543	0.543	0.151	0.151	0250	4	0.621	0.664	0.172	0.157
0110	19	0.832	0.832	0.212	0.212	0260	9	0.535	0.523	0.301	0.357
0120	14	0.634	0.675	0.211	0.206	0270	6	0.517	0.588	0.226	0.195
0130	5	0.266	0.234	0.213	0.213	0280	8	0.735	0.557	0.184	0.129
0140	9	0.741	0.743	0.203	0.206	0290	19	0.363	0.393	0.101	0.115
0150	11	0.868	0.867	0.234	0.239	0300	9	0.276	0.552	0.154	0.214
						Ave.	11.9	0.575	0.598	0.186	0.189

表 3: NTCIR データを用いた要約 (long) 結果

		余弦尺度		ROUGE				余弦尺度		ROUGE	
ID	文数	k-means	Spectral	k-means	Spectral	ID	文数	k-means	Spectral	k-means	Spectral
0010	23	0.713	0.713	0.251	0.251	0160	47	0.297	0.232	0.387	0.387
0020	7	0.053	0.053	0.368	0.368	0170	4	0.612	0.142	0.417	0.257
0030	18	0.555	0.576	0.386	0.38	0180	19	0.063	0.063	0.386	0.386
0040	47	0.477	0.477	0.214	0.214	0190	10	0.727	0.72	0.539	0.539
0050	11	0.252	0.299	0.151	0.302	0200	27	0.783	0.783	0.290	0.290
0060	43	0.446	0.457	0.255	0.255	0210	29	**0.011**	**0.592**	0.170	0.213
0070	17	0.600	0.600	0.468	0.468	0220	20	0.406	0.503	0.400	0.345
0080	22	0.395	0.395	0.348	0.348	0230	20	0.592	0.610	0.179	0.168
0090	11	0.252	0.299	0.151	0.302	0240	21	0.672	0.376	0.296	0.423
0100	26	0.446	0.457	0.255	0.255	0250	5	0.382	0.386	0.402	0.425
0110	29	0.600	0.600	0.468	0.468	0260	17	0.437	0.437	0.448	0.502
0120	19	0.395	0.395	0.348	0.348	0270	14	0.154	0.530	0.355	0.277
0130	8	0.025	0.025	0.275	0.267	0280	12	0.422	0.410	0.475	0.285
0140	15	0.604	0.604	0.378	0.383	0290	36	0.178	0.136	0.217	0.225
0150	16	0.865	0.865	0.396	0.402	0300	19	0.129	0.475	0.349	0.339
						Ave.	20.4	0.418	0.440	0.334	0.336