PageRankアルゴリズムを応用した重要文抽出システム

坪井 康徳　鳥澤 健太郎
北陸先端科学技術大学院大学 情報科学研究科
{y-tsuboi, torisawa}@jaist.ac.jp

1 はじめに

本稿では, 情報検索手法である PageRank[1] のアルゴリズムを応用する事でテキスト中の各文の重要度を決定する方法を提案する. PageRank は Web ページ間のリンク構造に注目して情報検索結果をスコアリングする手法であるが, 本研究では Web ページをテキスト中の文に, リンクを文間の単語の共有で置き換え PageRank の適用を行う. 単語の共有は行列で表され, その行列の固有ベクトルを計算することで各文の重要度を算出する.

2 PageRank アルゴリズムを応用した重要文抽出

2.1 PageRank 概要

PageRank[1][1] は, Web サーチエンジン Google の中核技術の一つである. PageRank は「多くの良質な Web ページからリンクされている Web ページは, やはり良質な Web ページである」という再帰的な関係を基に, 全ての Web ページの重要度を判定する. PageRank は Web ページ間のリンク関係を表した接続行列[2] その固有値問題を解くことで, 各 Web ページの評価を行う. PageRank の計算手順は以下の通りである. ある Web ページ i が, 別の Web ページ j からリンクが張られていたとき, その接続行列の成分 a_{ij} を 1 とし, そうでない場合その成分を 0 とする. すなわち,

$$a_{ij} = \begin{cases} 1 & \text{i が j からリンクを受けている場合} \\ 0 & \text{i が j からリンクを受けていない場合} \end{cases}$$

と定義する. 得られた行列の最大固有値に属する固有ベクトルを求め, 固有ベクトルの各成分を対応する Web ページの重要度とする.

次に例を示して考える. 図 1 中の四角 A, B, C は

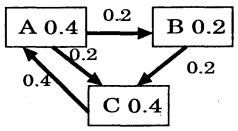

図 1 簡単な PageRank の例

Web ページを表しており, この 3 つの Web ページ以外にリンク関係はないものとする. 被リンク数を見るとページ A 及び B が 1 本, ページ C は 2 本のリンクを受けている. ここで多くのページからリンクを受けているページ C は高いスコアとなる. 同じ被リンク数でありながらページ A とページ B ではスコアに違いがある. これはこれらのページに至るリンクのリンク元の発するリンク数に関係がある. ページ A のリンク元であるページ C は, 1 本のリンクをページ A に発しており全てのスコアをページ A に伝えている. ページ B のリンク元であるページ A は, 2 本のリンクを発しており, そのうちの 1 本をページ B に発しておりページ A の持つスコアの半分を伝えている. この差がページ A, B の差である. 一般に PageRank のアルゴリズムにおいて高いスコアを得る Web ページは, 以下の性質を満たしている.

1. 多くの Web ページからリンクを受けている Web ページである
2. その Web ページに至るリンクのリンク元の Web ページが多くのリンクを受けている Web ページである
3. その Web ページに至るリンクのリンク元の Web ページが発しているリンク数が少ない Web ページである

すなわち, PageRank のアルゴリズムでは被リンク数

[1] PageRank (TM) は米 Google 社の登録商標である.
[2] より厳密には Web ページ間は強連結であるとし, その接続行列も既約であるとする

だけでなく，リンク元のページのスコア，リンク元の発するリンク数を含めて総合的にスコアを決定している為，スコアの高いページは広い視点で評価されているといえる．

2.2 PageRankアルゴリズムの重要文抽出システムへの対応

これまでに説明したPageRankのアルゴリズムを重要文抽出に応用する為に，本研究ではWWWを対象のテキスト，Webページ間のリンクを文と文の単語の共有，WWW上にあるWebページをテキスト中の文に対応させる．これを図示したものを図2に示す．本研究では，「多くの重要な文からリンク，すなわち単語の共有を受けている文は重要である」と仮定しPageRankのアルゴリズムをテキスト中の文の重要度計算に適用する．アルゴリズムをテキストに対応させる為には，Webページ間のリンクに相当する文間のリンクの決定，Webページ間のリンクの方向に相当する文間のリンクの方向性，の2点を解決する必要がある．Webページ間のリンクに相当する文間のリンクとして，本研究では文間における語彙的結束性に注目し，その中でも特につながりの強い同一単語の繰り返しを文間のリンクとして利用した．テキスト中のある文と別の文で各々の文中に同一の単語が現れたとき，その文間にはリンクがあると仮定する．今後，この文間のリンクの事を単語共有リンクと呼ぶことにする．また，使用する単語は形態素解析プログラムJUMANを用い，形態素解析結果の品詞が名詞，動詞，形容詞，未定義語である単語とする．

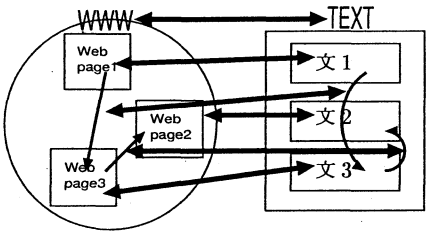

図2 WWW上のWebページとテキスト中の文の対応

単語共有リンクの方向に関して，本研究では吉見ら[2]の手法を基にした手法によりテキスト中の各文にスコアを与え，そのスコアによってつながりの方向を決める．吉見らは，標題を最も重要な文であると仮定し，重要な文とつながりの強い文を重要とする事でテキスト中の各文のスコアを決定した．タイトルへの重み付けを以下のように行う．

$$\text{タイトル}\ S_1\ \text{の重要度} = \frac{S_1\text{の重要語の重みの和}}{S_1\text{の重要語の数}}$$

ここで重要語とは品詞が名詞，人称代名詞，動詞，形容詞，副詞であるとする．また，吉見らはタイトル中に含まれる重要語を特別に標題語として扱った．

$$\text{重要語}\ w\ \text{の重み} = \begin{cases} 5 & w\ \text{が標題語，かつ}\ w\ \text{を含む} \\ & \text{文数が総文数の}\ 1/4\ \text{以下} \\ 1 & \text{その他} \end{cases}$$

タイトルであるS_1以外の文S_jの重要度はS_jから先行文S_iへのつながりの強さ(関連度)とS_iの重要度によって決まる．ここで，S_iとS_jの関連度R_{ij}は，

$$R_{ij} = \frac{S_j\text{の重要語から}\ S_i\text{の題述中の重要語につながる語の重みの和}}{S_i\text{題述中のの重要語の数}}$$

ここで文S_jの重要度をW_j，文S_iの重要度をW_iとする．S_jの重要度W_jは

$$W_j = \max_{j-5 \le i < j} \{W_i \times R_{ij}\}$$

のように定義した．以上により各文の重要度を求める．吉見らの手法は結果としてテキスト中の先頭に近い文であるほどスコアが高い傾向にある．また，吉見らの手法は英文を対象にしており，人称代名詞，先行（代）名詞の前方照応も文のつながりとして利用している．しかしながら，日本語では代名詞が省略される場合があり，テキスト中に存在するとは限らない（ゼロ代名詞問題）．そこで，本研究では，人称代名詞，先行（代）名詞の前方照応を行わず，他は吉見らの手法を基に，文間のつながりを決定する方法を実装する．今後この手法を，手法Aと呼ぶ．

本手法では，テキスト中の文iと文jについて，文iと文jが同一の単語を持っている時，単語共有リンクがあるとする．この時，手法Aによって決定される各文のスコアに基づき単語共有リンクの方向を決定する．手法Aによる文iのスコアをW_i，文jのW_jとすると単語共有リンクを表す行列の要素を次のように定義した．

$$a_{ij} = \begin{cases} 1\ \text{を加える} & (W_i \ge W_j) \\ 0 & (W_i < W_j) \end{cases}$$

この式は単語共有リンクがある場合，手法Aによるスコアの大きい文をS_iスコアの小さい文S_jとするとリンクの方向は，$S_j \rightarrow S_i$となり，その要素a_{ij}に1を加える．以上の定義から得られた行列に対する固有ベク

トルの各要素が各文の重要度になる．本手法がテキスト中の文で高いスコアを与える文は以下の性質を満たしている．

1. 多くの文から単語共有リンクを受けている文である
2. 多くの文から単語共有リンクを受けている文がリンク元の文である
3. 単語共有リンクを受けている文から太いリンク(複数のリンク)を持っている

これらの条件はPageRankの高いスコアを与える条件にほぼ対応している．1番目の条件は，本研究で仮定した，「多くの重要な文から単語共有リンクを受けている文は重要である」の中の「多くの文から単語共有リンクを受ける」を満たすための条件である．また多くの文からリンクを受けている文は，全くリンクを受けていない文に比べて重要であろうとの直感を表したものでもある．2番目の条件は1番目の条件を受けて「1番目の条件を満たす重要な文から単語共有リンクを受ける文はやはり重要である」．これは，本研究の仮定の「重要な文からのリンク」を満たす条件である．すなわちリンク元の文の重要度を考慮している事になる．3番目の条件はWebページにおけるリンクとの違いがでている．文から発せられる単語共有リンクは，Webページからのリンクと異なり，同じ文に対して複数リンクを発する場合が多くある．複数のリンクは，リンク元が持つ多くのスコアをリンク先に伝えることになり，リンク先は多くのスコアを受け取ることになる．本手法では以上の条件を満たす文に対して高いスコアを与える．TFなどの単語の頻度のみを使った手法では，多くの単語を持つ文が高いスコアになる傾向にある．本手法では複数の文で現れる単語共有リンクを作る単語のみが文のスコアに影響を与えるので，多くの単語を持つ文に高いスコアが与えられるとは限らない．

3 実験

本実験の正解コーパスとして，第2回NTCIRワークショップサブタスクTSCで作成された正解コーパスを用いる．記事は毎日新聞94年，95年，98年から抜粋されており，1面，2面記事，社会記事，社説記事，解説記事から構成されている．その内容を表1に示す．ベースラインシステムとして手法A，lead手法，TF based手法を用いた．ここでlead手法はテキストの先頭から要約率ごとに指定された文数だけ出力するものとし，TF based手法は，本文の各文ごとに内容語のTFの和を計算し，このスコアの高い文を要約率ごとに指定された文数だけ選択する．選択した文を元の文の出現順に戻して出力する．内容語として，その品詞が名詞，動詞，形容詞，未定義語の単語を使用する．以下の表2，表3，表4に文選択率10%，文選択率30%，文選択率50%，の結果を示す．まず多くの文選択率で本手法が，前処理として利用している手法Aを上まわる性能を出していることに注意してほしい．これはPageRank適用の有用性を示している．また，本手法は各文選択率で他のベースラインシステムと比べ高い精度を示した．詳しく見ると，10%の文選択率ではlead，手法Aなどのテキストの先頭付近に高いスコアを与える手法が比較的よい精度となっている．本手法はシステムAの結果に基づいてリンクの方向付けを行っているので，10%の文選択率においてもよい精度になったと考えられる．一方50%の文選択率になると，テキストの先頭部分以外からも正解文になる事があるためテキストの先頭付近に高いスコアを与える手法は精度が低くなった．一方，TFなどの文の位置に

表1

	1, 2面	社会	社説	解説
総記事数	16	76	41	47
総文数	342	1721	1362	1096
重要文数 (10%)	34	172	143	112
重要文数 (30%)	103	523	414	330
重要文数 (50%)	174	899	693	555

表2 文選択率10%

Genre	本手法	手法A	TF based	lead
1, 2面	46.9	50.0	42.2	47.9
社会	54.0	53.3	31.3	51.8
社説	31.0	30.7	18.7	31.6
解説	31.6	25.9	20.2	15.9
平均	42.3	40.7	26.5	37.4

表3 文選択率30%

Genre	本手法	手法A	TF based	lead
1, 2面	54.5	54.2	48.5	50.5
社会	53.5	53.0	47.9	54.3
社説	41.4	38.2	42.2	36.7
解説	44.1	35.9	41.9	32.4
平均	48.4	45.2	45.1	44.2

表 4 文選択率 50 %

Genre	本手法	手法 A	TF based	lead
1, 2 面	64.8	63.2	58.4	60.4
社会	66.7	62.6	66.2	62.5
社説	56.7	53.0	57.0	51.0
解説	61.4	56.0	61.8	50.4
平均	62.9	58.7	62.3	56.1

よらない手法では文の位置を使う手法に比べて高い精度となった. 本手法は, 手法 A の結果に基づいてリンクの方向を決定しているが, TF と比べて精度の落ち込みは見られない. これは, 手法 A で得られたスコアはリンクの方向を決定する為に用いているが, 手法 A で得られたスコアが低くても, 同じスコアのもの, あるいは更にスコアの低いものからはつながりを受ける事が出来る. これによりテキストの先頭部分以外の文も選択可能になったと考えられる.

次に 難波ら [3] の報告に基づき, 第 2 回 NTCIR ワークショップサブタスク TSC(Text Summarization Challenge) で行われたテキスト自動要約システムのコンテストの結果と提案手法を比較し検討を行った. その結果, 本手法は平均の値より低い値となった. コンテストは毎日新聞 95, 98 年の中から社会記事 15 記事, 社説記事 15 記事の合計 30 記事で行われた. 最も精度の良いシステムとコンテスト参加 10 システムの平均値を表 5 に示す. 提案手法では, 特に文選択率 10 %の

表 5 TSC 結果

Genre	本手法	Best	Average
社会 10 %	27.0	32.4	29.7
社会 30 %	50.0	39.5	47.2
社会 50 %	59.7	56.6	61.2
社説 10 %	15.4	35.4	20.5
社説 20 %	37.4	44.4	37.5
社説 50 %	56.1	60.9	55.1
TOTAL	41.5	46.3	43.7

社説記事において著しく低い値となった. これは社説記事の特徴として, 正解文が先頭付近にあるとは限らないことが挙げられる. 本手法は結果として先頭部分にバイアスをかける手法である為, 選択文の制限の厳しい文選択率 10 %においては精度が下がると考えられる. その結果, 全体として精度が平均の値よりも低下した.

4 考察

提案手法によって正解文に与えられた重要度が小さく, 正解文が選択されなかった原因を分析した. 一つの原因として, テキストが複数のサブトピックから構成されている場合が上げられる. 文と文のつながりは同一トピック内で起りやすい. このため文の少ないトピックからは, そのトピック中で重要な文であっても低い重要度しかあたえられないため, テキストが複数のサブトピックから構成されている場合には, 正解文として選択すべき文に対して高い重要度を与えることが難しいと考えられる. このような複数のサブトピックから構成されているテキストへの対応は今後の課題である.

5 まとめ

本稿では, PageRank のアルゴリズムを用いることにより, 再帰的に重要度を考慮した, 重要文抽出を提案した. 毎日新聞の記事から 180 記事を対象とした実験では, 文選択率を 10 %としたとき平均で 42.3 %, 文選択率を 30 %としたとき平均で 48.4 %, 文選択率を 50 %としたとき平均で 62.9 %, の精度を得た. これらの精度は, ベースラインシステムと比べて良い精度を示しており, PageRank のアルゴリズムを用いることで精度向上が図れることを示唆している. 今後の課題として, 複数のサブトピックから構成されるテキストへの対応や, 今回精度の低かった社説記事や解説記事への対応は今後の課題である.

参考文献

[1] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winogard, *"The PageRank Citation Ranking: Bringing Order to The Web"*.In Technical Report, 1998.

[2] 吉見 毅彦, 奥西 稔幸, 山路 孝浩, 福持 陽士, "表題へのつながりに基づく文の重要度評価", 自然言語処理, Vol. 6, No. 1, pp.43–56, 1999.

[3] 難波 英嗣, 奥村 学, "第 2 回 NTCIR ワークショップ 自動要約タスク (TSC) の結果および評価法の分析", 情報処理学会自然言語処理研究会報告, 144–20, pp.143–150, 2001.