E-060

文書クラスタリングにおけるクラスタタイトルの自動生成
Automatic generation of cluster title for document clustering

黒田 知宏[1]　内海 彰[2]
Tomohiro Kuroda　Akira Utsumi

1 はじめに

今日，Web を利用することで様々な文書を入手できるようになった．そのため膨大な文書の中から必要とする文書を効率的に得るために情報検索技術が必要となっている．情報検索技術の分野において，文書の集合を内容的に類似している文書どうしでまとめ，いくつかの文書群（以下，クラスタとする）に分けるために，文書クラスタリングという手法が用いられている．似ている内容の文書がまとまるので文書群から求めている文書を探すことが容易になり，検索処理における作業効率が改善される．

文書クラスタがどのような内容かを表すためにクラスタ中の文書から抽出した重要語をラベルとして付与することが多い．しかし，重要語のみではクラスタの内容を把握しづらいことも少なくない．このとき，クラスタにタイトルが付けられていればクラスタの内容がより把握しやすくなると考えられる．

クラスタに対してタイトルを生成する研究はないが，単一文書に対してタイトルを生成する研究には以下のようなものがある．長安ら [1] は社説記事に対して，重要語抽出と単語の並びによるタイトルパタンを用いてタイトルを生成している．タイトルパタンは社説記事の実際のタイトルに出現する語の組み合わせをもとに決定している．そのため，社説記事と同じ内容の文書でなければ適切なタイトルは生成できないので，汎用性に欠ける．南野ら [2] は 3-gram に基づく確率を用いたタイトルの生成を行っている．しかし，日本語は英語などに比べて語順は割と自由な言語であるため，係り受け構造の方が重要であると考えられる．

そこで本研究では，クラスタ間の差異を考慮して重要語を抽出し，タイトルパタンを適用することで，クラスタの内容に沿ったタイトルを自動生成する手法を提案する．どの分野の文書に対しても対応可能にするために，文節間の係り受け関係を利用したタイトルパタンを用いる．山本ら [3] は体言止めや助詞止めによる文末の整形によりタイトルを生成できるとしているので，タイトルパタンは体言止めの形とする．

2 タイトルの自動生成手法

本手法では Web ページを文書クラスタリングし，生成されたクラスタ集合を入力して各クラスタに対するタイトルを生成する．以下に本手法の概要を示す．

1. **文分割**
   クラスタ内のページからタグを除去し，区切り文字によって各ページのテキストを分割し文とする．
2. **形態素解析**
   各文に対して形態素解析を行い名詞を抽出する．ここで名詞とは形態素解析によって名詞と判断された単名詞の他に，単名詞やアルファベットによって成り立つ複合名詞も含む．形態素解析には「茶筅」[3]を用いる．
3. **重要語の決定**
   形態素解析で抽出された名詞に対して重要度を計算する．重要度上位の名詞をクラスタの重要語とする．また，Web の検索結果のクラスタに対して本手法を適用する場合は，検索語も重要語とする．
4. **タイトルパタンの適用**
   重要語を含む文に対して係り受け解析を行い，タイトルパタンを適用することでタイトル候補を生成する．係り受け解析には「南瓜」[4]を用いる．
5. **タイトルの決定**
   名詞の重要度などを用いてタイトル候補の重要度を算出し，上位 1 件をタイトルとして出力する．

以下に各手順の詳細を述べる．

2.1 文分割

クラスタ内のページごとに title タグで囲まれているページタイトルと，body タグで囲まれている部分を抽出する．このとき，以下の 2 つのルールを適用する．

- コメント，script，style，strike，select，option タグ間の文字は考慮しない
- img タグ内の alt 属性部分は考慮する

抽出した部分からタグを除去し，改行，「。」，「！」，「？」で区切り，文に分割する．このとき，以下の 2 つのルールを適用する．

- h，div，tr，td，dd，dl，dt，li，blockquote，ul，p タグを改行コードに置き換える
- 「。」の後に br タグがあるか，br タグが 2 回以上連続で用いられる場合，br タグを改行コードに置き換える

ただし，分割したもので助詞を含まない場合は文としない．これは，Web ページの場合のメニュー項目などの，ページに書かれているがページの内容を表してはいない語を考慮しないためである．

2.2 形態素解析

複合名詞は複数の語で 1 つの意味を持っているため，単名詞とは別の意味を表す語である．例えば，「情報」と「情報検索」では同じ「情報」を含むが意味は異なる．そこで，形態素解析によって名詞と判断された語以外に以下の 2 種類の複合名詞も名詞として考慮する．

- 名詞，アルファベット，未知語によって構成される語
  例：情報検索技術，Google 検索
- 形容詞＋「さ」
  例：広さ，高さ

ただし，「もの」や「こと」などの非自立の名詞，「彼」や「私」などの代名詞，「2007」などの数字のみの名詞はページの内容とあまり関係がない場合が多いため処理の対象外とする．

2.3 重要語の決定

あるクラスタにおいて出現回数が多く，クラスタ内の多くの文書にも出現し，他クラスタには出現しない名詞が重要であるという考えに基づき重要語を決定する．

クラスタ C_i の名詞 w_j について，式 (1) で重要度 NI_{ij} を計算する．

$$NI_{ij} = tf_{ij} \cdot df_{ij} \cdot log(N/cf_j) \quad (1)$$

[1]電気通信大学大学院電気通信学研究科システム工学専攻
[2]電気通信大学電気通信学部システム工学科
[3]URL;http://chasen-legacy.sourceforge.jp/
[4]URL;http://chasen.org/ taku/software/cabocha/

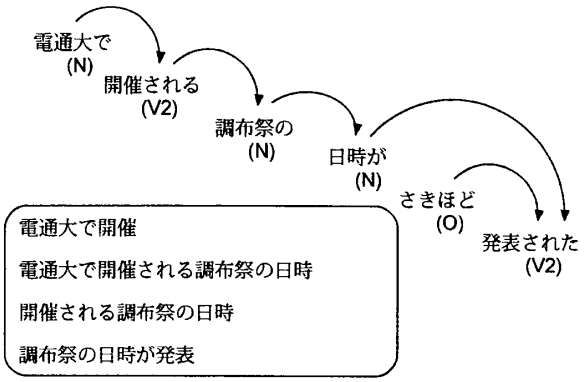

図 1: タイトル候補生成例

ここで，tf_{ij} はクラスタ C_i における名詞 w_j の出現数，df_{ij} はクラスタ C_i における名詞 w_j の出現する文書数，N は全クラスタ数，cf_j は全クラスタ集合での名詞 w_j の出現するクラスタ数である．

そして重要度が上位の名詞をクラスタ C_i の重要語とする．ただし，NI_i の最大値の 1/4 未満の重要度の名詞は重要語としない．これは話題が限定されている場合に，重要ではない語を選択しないためである．

2.4 タイトルパタンの適用

文を構成する要素のうち，タイトルとなり得るものを主節，目的節，体言止めと考える．これらは全て名詞節を中心としたものである．そこで，重要語を含む文に対して係り受け解析を行い，以下の 4 パタンを適用してタイトル候補を生成する．なお各文節において，名詞があれば名詞節 (N)，動詞もしくは助動詞「だ」,「です」があれば動詞節 (V)，サ変名詞とサ変動詞の両方があればサ変動詞節 (V2)，それ以外はその他の節 (O) と節タグを付ける．

2.4.1 タイトルパタン

各文の文節ごとに係り先の節タグを用いてどのタイトルパタンに当てはまるかを調べる．なお，名詞節（N）どうしの係り受けの場合，まとめて 1 つの名詞節（N）とする．ただし，生成された表現が名詞もしくは複合名詞のみの場合はラベルと同じなのでタイトル候補としない．また，重要語を含まないタイトル候補は生成しない．適用するタイトルパタンは非排他的で，同一の名詞節に対して複数のタイトルパタンが適用されることもある．

1. **名詞節（N）**
   名詞節がサ変動詞節以外に係るか係り先がない場合，名詞節（N）をタイトル候補とする．
2. **名詞節（N）＋サ変名詞**
   名詞節がサ変動詞節に係る場合，名詞節（N）とサ変名詞をつないだものをタイトル候補とする．
3. **名詞節（N）＋他の文節＋名詞節（N）**
   名詞節が他の文節に係り，それがさらに名詞節に係る場合，全てつないだものをタイトル候補とする．
4. **名詞節以外の文節＋名詞節（N）**
   名詞節以外の節が名詞節に係る場合，全てつないだものをタイトル候補とする．

2.4.2 適用手順

重要文の先頭の文節からタイトルパタンにマッチすればタイトル候補を生成し，次の文節に移る．

例えば重要語が「電通大」と「調布祭」の時,「電通大で開催される調布祭の日時がさきほど発表された」という文（図 1）に対して以下のようにタイトルパタンを適用する．図 1 において，例文の各文節の節タグを括弧内に示す.「電通大で」（N）の係り先はサ変動詞節（V2）なので，パタン 2 より「電通大で開催」が生成される．係り先の「開催される」が「調布祭の」（N）に係り，それがさらに「日時が」（N）に係っているので，パタン 3 より「電通大で開催される調布祭の日時」が生成される．同様に,「開催される」は「調布祭の」に係り，それがさらに「日時が」に係っているので，パタン 4 より「開催される調布祭の日時」が生成される.「調布祭の」は「日時が」に係り，さらに「発表された」に係っているので，パタン 2 より「調布祭の日時が発表」が生成される.「日時が」以降の文節には重要語が存在しないので，タイトルパタンは適用しない．最終的に図 1 の枠内に示すタイトル候補が生成される．

表 1: 助詞の置き換え規則

助詞	置き換え前	置き換え後
で	電通大で開催	電通大での開催
と	日時と場所	日時と場所
は	調布にある大学は電通大 (サ変名詞以外に係る)	調布にある大学
	調布祭は延期 (サ変名詞に係る)	調布祭の延期
その他	調布祭の時間が発表	調布祭の時間の発表

2.4.3 助詞の置き換え

タイトル候補の最後の節の前にある助詞を表 1 に示す規則に従い置き換える.「は」の場合，係り先がサ変動詞節以外であれば,「は」の前後は同格であると判断し,「は」の前までをタイトル候補とする．また,「と」,「や」の場合は「日時と場所」のように並列関係にあることが多いと考え，そのままとする．その他の助詞の場合は「の」に置き換える．

2.5 タイトルの決定

式 (2) によりクラスタ C_i のタイトル候補 T_{ij} の重要度 TI_{ij} を計算する．

$$TI_{ij} = \frac{n_{ij}}{m_i} \cdot \frac{1}{l_{ij}} \sum_{w_{ik} \in T_{ij}} NI(w_{ik}) \qquad (2)$$

ここで，l_{ij} はタイトル候補 T_{ij} の文節の数，w_{ik} は T_{ij} に含まれる名詞，n_{ij} は T_{ij} に含まれる名詞を全て含むクラスタ C_i 内の文の数，m_i はクラスタ C_i 内の総文数とする．第 1 項はクラスタ内に多く出現する表現かどうかを計算し，第 2 項は重要な語を多く含んでいるかを計算する．同一タイトル候補が複数回生成される場合，TI_{ij} 値を生成回数倍する．

3 評価実験

タイトルと重要語のどちらがよりクラスタの内容を表しているかを比較するために，本システムが出力したタイトルと重要語それぞれに対して 5 段階評価による評価実験を行った．

3.1 クラスタ作成

学生 5 人にそれぞれ Google に 1 つクエリを入力してもらい，検索結果をクラスタリングしてもらった．合計 5 個のクラスタ集合ができ，それらの平均クラスタ数は 3.40，クラスタに含まれる平均文書数は 6.06 であった．

3.2 評価方法

人手により作成された各クラスタに対して TI_{ij} 上位 3 個のタイトルと NI_{ij} 上位 3 個の重要語を出力した．学生 11 名に出力したタイトルと重要語についてクラスタの内容を表しているかを評価してもらった．1（全く表していない）から 5（完全に表している）の 5 段階評価を用いた．また，タ

表 2: 出力例（オーバーラン）

クラスタ	重要語	タイトル
電車	電車 日記 運転士	「オーバーランの検索」 「オーバーランの発生」 「オーバーランの事故」
プログラム	バッファオーバーラン 実行 リターンアドレス	「コードの実行」 「プログラムの実行」 「任意のコードの実行」
スイスの地方	スイス アルプス 風景	「アルプスの名峰」 「スイスの四季」 「スイス マッターホルンとアルプス紀行 DVD 紀行」

表 3: 出力例（スカイライン）

クラスタ	重要語	タイトル
車	日産 搭載 モデル	「スカイライン・ニューマンスカイライン・史上最強のスカイライン」 「スカイラインの検索」 「エンジンの搭載」
乗鞍	乗鞍岳 ほお タクシー	「乗鞍岳の映像」 「乗鞍岳の自然」 「乗鞍岳の天気」
富士	五合目 富士山 富士山スカイライン	「富士山の南斜面」 「富士山新五号目へと登る道のT字路」 「ここから五合目」

イトルについて日本語として正しい表現かどうかを 1（おかしい）から 5（正しい）の 5 段階で評価してもらった．なお，回答時には 1 つのクラスタ集合の結果全てに答えてもらい，各クラスタにつき 5 人分のデータを集めた．

3.3 結果と考察

3.3.1 出力例

クエリとして「オーバーラン」を用いた場合，クラスタ，重要語，タイトルは表 2 のようになった．「スカイライン」を用いた場合は表 3 のようになった．

3.3.2 日本語として正しい表現か

全タイトルの平均評価値は 3.99 であった．このことから係り受け解析により生成されたタイトルは日本語として正しい表現であると言える．

3.3.3 内容を表しているか

タイトルの平均評価値は 3.33，重要語の平均評価値は 3.52 であった．タイトルの方が重要語より内容を表していないという結果となった．原因として，適用したタイトルパタンの一部の評価のみが悪くなったためにタイトルの平均値が下がった可能性がある．

そこで，タイトルパタンごとの評価値を求め，有用なパタンがあるかどうかを調べる．パタンごとの生成されたタイトル候補数と出力されたタイトルに含まれていた数を表 4 に示す．表 4 より，パタン 1 の「名詞節（N）」が出力されたタイトルのほとんどを占めていることがわかる．またパタン 3 の「名詞節（N）＋他の文節＋名詞節（N）」の候補数は多いものの，タイトルに用いられていないことがわかる．

「名詞節（N）」は係り先がない場合とサ変動詞節（V2）以外に係る場合がある．そこで係り先がない場合とサ変動詞節（V2）以外に係る場合に分けて，それぞれの平均評価値を求める．また日本語として正しくないタイトル候補は，タイトルパタンの評価としては不適切と考え，日本語の正しさの平均評価値が中間値 3 より小さいタイトル候補は除外して平均評価値を計算する．各タイトルパタンごとの平均評価値を表 5 に示す．

表 4: タイトルパタン毎の候補数と出力数

タイトルパタン	候補数	出力数
名詞節	640	39
名詞節＋サ変名詞	582	11
名詞節＋他の文節＋名詞節	691	0
名詞節以外＋名詞節	206	4

表 5: タイトルパタンごとの平均評価値

タイトルパタン	内容を表しているか
名詞節（係り先なし）	3.31
名詞節（サ変動詞節以外に係る）	3.64
名詞節＋サ変名詞	3.52
名詞節以外＋名詞節	3.40

表 5 の N（V2 以外に係る）の平均評価値が 3.64 であることから，「名詞節（N）」のサ変動詞節（V2）以外に係る場合が重要語の平均評価値（3.52）よりも良いことがわかる．このタイトルパタンでの名詞節は文における主語や目的語にあたり，これらは話題となることが多いため評価が高かったと考えられる．

逆に「名詞節（N）」の係り先がない場合と「名詞節以外＋名詞節（N）」の平均評価値はよくない．「名詞節（N）」の係り先がない場合は体言止めから生成されたと考えられる．体言止めは強調として使われることが多く，文書の内容を示していると考えるが，複数の文書の場合には共通の話題となることが少なかったため評価が下がったと考えられる．これらの不適切なタイトルパタンを適用していたため，全体的な評価でタイトルが重要語より低かったと言える．

4 おわりに

クラスタに対して，単語の出現情報，係り受け情報を用いてタイトルを生成する手法を提案した．名詞節がサ変動詞節以外に係る場合のタイトルパタンを用いることで，重要語によるラベル以上の内容示唆が可能であることが示された．

しかし，ラベルよりも有用であるタイトルパタンは 1 種類しかなかったので，より良いタイトルパタンを見つける必要がある．また，タイトルを付ければ全てのクラスタの内容が把握しやすくなるというわけではない．そこで，ラベルとタイトルを使い分けられるようにするとよりクラスタの内容を把握しやすくなると考えられる．さらに，人間がタイトルを付ける場合，文書には出現しない語を用いたり，別の文にある語を組み合わせてタイトルを付けることが少なくない．このようなタイトルも生成できる手法を開発すればよりクラスタの内容を表すことができると思われる．

参考文献

[1] 長安義夫, 山本和英. タイトルパタンによる文書の一文概要生成. 言語処理学会第 13 回年次大会,pp.684-687 (2007).
[2] 南野朋之, 奥村学.RSS 自動生成のためのタイトル生成. 言語処理学会第 11 回年次大会,pp.57-60, (2005).
[3] 山本和英, 池田諭史, 大橋一輝.「新幹線要約」のための文末の整形. 自然言語処理,Vol.12,No.6,pp.85-112 (2005).