F-021

記事関連度を利用した話題ごとのブログ記事分割手法

Blog Article Division Method for Each Topic using Degree of Association between Articles

奥田 将好† 三品 賢一† 土屋 誠司‡ 渡部 広一‡
Masayoshi Okuda Kenichi Mishina Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，インターネットの普及により，膨大な情報が Web 上に溢れるようになった。これらの情報には誤った情報が多く含まれるため，文書を短時間で読み，正しい情報を選び出す必要がある．しかし，文書には様々な話題の文章が含まれており，自分が知りたい話題以外の文章は，文書を短時間に読むうえでの障害となる．ここで，記事が話題ごとのブロックに分割されていれば，自分が知りたい話題について書かれている文章を選択して読むことができ，効率的に情報収集を行うことができる．そこで，本稿ではブログ記事を対象に，記事間の関連の強さを判断できる記事関連度[1]を利用して，記事を話題ごとのブロックに分割する手法を提案する．

2. 関連技術

2.1 概念ベース

概念ベース[2]とは，電子化した国語辞書などから機械的に構築した大規模なデータベースである．ある概念Aはm個の属性a_iと重み$w_i(w_i > 0, \sum_{i=1}^{m} w_i = 1)$の対により以下のように定義される．

$$A = \{(a_i, w_i) | i = 1 \sim m\} \quad (1)$$

2.2 関連度計算方式

関連度計算方式[3]とは概念と概念の関連の強さを定量的に評価する手法である．関連度は，概念間の関連の強さを 0 と 1 の間の数値で表し，関連が強いほど高い数値となり，以下の式で求める．A, Bは概念，a_i, b_iは属性，u_i,v_iは重みであり，DoAを関連度，DoMを一致度と呼ぶ．

$$DoA(A,B) = \sum_i DoM(a_i, b_i) \times (u_i + v_i)/2 \times \min(u_i, v_i)/\max(u_i, v_i) \quad (2)$$

$$DoM(A,B) = \sum_{a_i = b_j} \min(u_i, v_j) \quad (3)$$

2.3 EMD を用いた記事関連度計算方式

Earth Mover's Distance[4]（以降，EMD とする）とは分布間の距離を表すもので，一方の分布を他方の分布に変換するための最小コストを分布間の距離として定義した距離尺度である．

EMD を用いた記事関連度計算方式[5]は，記事に含まれる単語とその重みからなる分布と，関連度計算方式によるコスト関数を用いて，記事と記事の関連の強さを定量的に表現する手法である．この手法では，記事間の単語を多対多で柔軟に対応をとることができるため，記事の単語数の差異にかかわらず，記事同士の関連の強さを求めることができる．

†同志社大学大学院　理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

3. 提案手法

提案手法の流れを図 1 に示す．

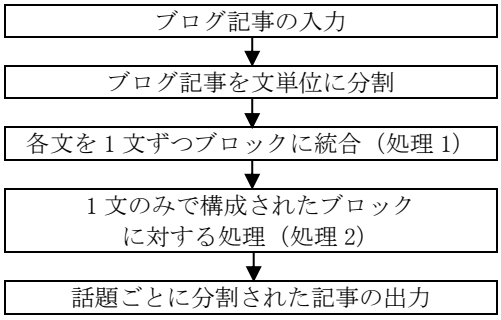

図 1　提案手法の流れ

以下の節で図 2 のブログ記事[6]を話題ごとに分割するときの処理の流れを示す．

長野県では、「大北森林組合の補助金不正問題」が大きな関心事になっています。今回（4 日）、大北森林組合に直接現地調査に行ってきました。
⋮
「『県の指導をしっかり守って事業を行うこと』と確認しながら事業を行ってきました。私たちにも責任はありますが…」
とのお話でした。不正に受給されたお金がどうなったのか？真相究明のために引き続き追及していきたいと思います。昼食に道の駅中条でいただいたほうとう風のおうどん「おぶっこ」の味と、帰りの北アルプスの景色は最高でした。

図 2　ブログ記事の例

3.1 ブログ記事を文単位に分割

入力されたブログ記事を文単位に分割する．ここで，文とは括弧内の句点を除いた，句点もしくは改行で区切られる一文を指す．図 2 のブログ記事を分割すると，図 3 のような 10 個の文に分割される．

文 1	長野県では、「大北林組合の補助金不正問題」が大きな関心事になっています。
文 2	今回（4 日）、大北森林組合に直接現地調査に行ってきました。
⋮	⋮
文 7	「『県の指導をしっかり守って事業を行うこと』と確認しながら事業を行ってきました。私たちにも責任はありますが…」
⋮	⋮
文 10	昼食に道の駅中条でいただいたほうとう風のおうどん「おぶっこ」の味と、帰りの北アルプスの景色は最高でした。

図 3　ブログ記事を文単位に分割した例

3.2 各文を 1 文ずつブロックに統合

まず，各文と直前のブロックの記事関連度を計算し，記事関連度が閾値以上であれば，同じブロックに統合される．ここで，ブロック全体との記事関連度のみを閾値として使用する方法を処理 1-1，それに加えてブロック内のそれぞ

れの文との記事関連度を使用する方法を処理 1-2 とする．

図 2 のブログ記事に対して，閾値 0.01 で処理 1-2 を行った場合，図 4 のようなブロックに統合される．

ブロック 1	長野県では、「大北森林組合の補助金不正問題」が大きな関心事になっています。 ⋮ 「『県の指導をしっかり守って事業を行うこと』と確認しながら事業を行ってきました。私たちにも責任はありますが…」
ブロック 2	とのお話でした。
ブロック 3	不正に受給されたお金がどうなったのか？ 真相究明のために引き続き追及していきたいと思います。
ブロック 4	昼食に道の駅中条でいただいたほうとう風のおうどん「おぶっこ」の味と、帰りの北アルプスの景色は最高でした。

図 4　処理 1 で分割されたブログ記事

3.3 1 文のみで構成されたブロックに対する処理

図 4 のブロック 2 のように，処理 1 により，1 文のみで構成されたブロックによって同じ話題のブロックが分割されてしまう場合があるため，1 文のみで構成されたブロックの前後のブロックで記事関連度を計算し，閾値以上であればすべて同じブロックに統合する．ここで，処理 1 と同様に，記事関連度を計算する対象によって処理 2-1，処理 2-2 とする．

図 4 のように分割された記事に対して，閾値 0.01 で処理 2-2 を行った場合，図 5 のようなブロックが出力される．

ブロック 1	長野県では、「大北森林組合の補助金不正問題」が大きな関心事になっています。 ⋮ とのお話でした。 不正に受給されたお金がどうなったのか？ 真相究明のために引き続き追及していきたいと思います。
ブロック 2	昼食に道の駅中条でいただいたほうとう風のおうどん「おぶっこ」の味と、帰りの北アルプスの景色は最高でした。

図 5　出力されるブロック

4. 評価と考察

提案手法によるブログ記事の分割精度を，次の方法で評価した．まず Yahoo!ブログ[7]から無作為に抽出したブログ記事 50 件を用意する．ブロックへの分割手法として，処理 1 と処理 2 で使用する手法の組み合わせにより，以下の 6 種類の手法を用いる．ここで括弧内の値は，実験により求めた，それぞれの処理で使用する閾値である．

手法 A:処理 1-1（0.01）のみ
手法 B:処理 1-2（0.01）のみ
手法 C:処理 1-1（0.04）+処理 2-1（0.02）
手法 D:処理 1-2（0.04）+処理 2-1（0.02）
手法 E:処理 1-1（0.01）+処理 2-2（0.01）
手法 F:処理 1-2（0.01）+処理 2-2（0.02）

分割精度の評価方法として，まず人手で評価セットのブログ記事をブロックに分割し，それを正解ブロックとする．そして，提案手法で分割されたブロック（出力ブロック）と正解ブロックを用いて以下の式で表される適合率，再現率，F 値で評価する．本稿では F 値を分割精度として扱う．出力ブロックを正解とする条件は，正解ブロックに出力ブロックがすべて含まれていること，正解ブロックに含まれる文を出力ブロックが半分より多く含んでいることである．

$$\text{適合率} = \frac{\text{正しく検出されたブロックの総数}}{\text{手法で検出したブロックの総数}} \quad (4)$$

$$\text{再現率} = \frac{\text{正しく検出されたブロックの総数}}{\text{正解ブロックの総数}} \quad (5)$$

$$\text{F 値} = \frac{2\times\text{適合率}\times\text{再現率}}{\text{適合率}+\text{再現率}} \quad (6)$$

4.1 評価結果

評価結果を表 1 に示す．

表 1　評価結果

手法	適合率	再現率	F 値
A	9.2%	31.5%	14.3%
B	9.2%	29.9%	14.0%
C	15.3%	28.8%	20.0%
D	15.5%	28.8%	20.2%
E	23.3%	30.4%	26.4%
F	19.2%	27.2%	22.5%

4.2 考察

表 1 の結果より，手法 E の精度が最も高く，26.4%の精度となった．

手法 A と手法 B を比較した場合，手法 B の精度が低いことから，処理 1 では，ブロックに含まれるそれぞれの文を使用した場合，別の話題の文が同じ話題のブロックに統合されたため，精度が低下したと考えられる．

処理 2 を追加した手法は，処理 1 のみの手法と比較して再現率は大きく低下せず，適合率が向上していることから，1 文のみで構成されたブロックで途切れてしまっていた同じ話題のブロックを，処理 2 によって適切に統合することができたと考えられる．

5. おわりに

本稿では，記事関連度を用いてブログ記事を話題ごとのブロックに分割する手法を提案した，記事分割の精度は 26.4%となった．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けて行ったものです．

参考文献

[1] 倉田篤史, 渡部広一, 河岡司, “概念ベースと関連度計算を用いた記事関連度計算方式”, 情報処理学会研究報告, 2006-NL-171, pp.19-24(2006).
[2] 奥村紀之, 土屋誠司, 渡部広一, 河岡司, “概念間の関連度計算のための大規模概念ベースの構築”, 自然言語処理, Vol.14, No.5, pp.41-64(2007).
[3] 荒木孝允, 奥村紀之, 渡部広一, 河岡司, “比較対象概念の共通属性を重視する動的関連度計算方式”, 同志社大学理工学研究報告, Vol. 48, No. 3, pp.14-24(2007).
[4] X.Wan, Y.Peng, “The Earth Mover’s Distance as a Semantic Measure for Document Similarity”, Proceeding of the 14th ACM international conference on Information and knowledge management， pp.301-302(2006).
[5] 藤江悠五, 渡部広一, 河岡司, “概念ベースと Earth Mover’s Distance を用いた文書検索”, 自然言語処理, Vol.16, No.3, pp.25-49(2009).
[6] “ふじおか義英は長野県佐久から政治をかえます！”, https://blogs.yahoo.co.jp/fujioka_nagano_saku/47205693.html(2017-6-19)
[7] “Yahoo!ブログ - 無料で 10GB の大容量のブログ(Blog)をはじめよう！”, http://blogs.yahoo.co.jp(2017-6-19)