6T-7

ストーリー性を考慮した映画あらすじからの類似度計算

村手 宏輔　　黒岩 眞吾　　堀内 靖雄　　篠崎 隆宏

千葉大学　大学院融合科学研究科

1. はじめに

近年，インターネットの普及により書籍や音楽，映画など様々なコンテンツの情報を手軽に収集することが可能になった．一方で，手に入る情報が膨大なため，必要な情報を見つけられないという情報過多の問題が発生している．この問題を解決するため，情報検索の分野では様々な類似文書検索の技術が考案されている．代表的なものとして，文書に含まれる単語の頻度情報等から特徴ベクトルをつくり，コサイン尺度等で類似度を計算する手法がある[1]．しかし，単語の頻度情報だけで文書の内容を表すことは難しい．例えば人が映画や小説等のコンテンツを探す際には，単語やキーワードだけでなく，物語の進み方や結末など，ストーリーも重要な情報となる．

そこで本稿では，あらすじが書かれた文書を対象として，文書が持つストーリーの類似度を計算する方法を提案する．ストーリーという指標で文書間類似度を計算できれば，文書分類や情報推薦等に役立つと考えられる[2]．

2. ストーリーを比較する上での問題

ストーリーとは映画や小説などの物語がもつ話の筋のことであり，出来事の時系列によって表される．しかし，厳密な記述方式は定められていないため，文書が持つストーリーの類似度を計算する際，以下の点を考慮する必要がある．

① 記述量の差: あらすじは物語を説明するための文書であり，人物の行動や状況を説明する記述で出来事が表され，その出来事の流れによりストーリーが表されている．しかし，文の長さや説明の詳細さはあらすじ毎に異なるため，これを考慮した類似度計算が必要である．

② 単語の出現順序によるストーリーの差: 文を構成する単語が同じでも，文内の出現順序が異なるとストーリーは変化する場合がある．例えば「主人公が一度は勝つが，最終的に負ける」と「主人公が一度は負けるが，最終的に勝つ」では「勝つ」と「負ける」の 2 単語の出現順序が異なるだけだが，出来事の時系列が変わるため，ストーリーが異なっていると言える．本稿は頻度情報では考慮することができない上記の差の定量化を目指す．

Similarity Measures for Movie Outlines Based on Differences in Story
Kosuke Murate, Shingo Kuroiwa, Yasuo Horiuchi, Takahiro Shinozaki, Graduate School of Advanced Integration Science, Chiba University

3. 提案手法

3.1 概要

前章で挙げた点を考慮し，本稿では出来事に関係する単語の並び方を比較することでストーリーの差を定量化する方法を提案する．

まず，単語を単語がどのような出来事に関係しているかの特徴ベクトルとして表現する(以下，単語特徴ベクトルと呼ぶ)．そして，文書を構成する単語の並びを単語特徴ベクトルの系列データとして表し，DP マッチング[3]を用いて系列間の類似度を計算する．また本稿では，名詞，動詞，形容詞が出来事に関係すると考え，他は不要語として削除した．

3.2 単語特徴ベクトル

単語特徴ベクトルの要素として，本稿では映画を分類するジャンルを用いる．ジャンルにより，出現しやすい出来事と出現しにくい出来事がある．例えば，ロマンス映画に分類される映画のあらすじでは恋愛に関係する出来事が多い．しかし，出来事の系列全てが恋愛に関係するわけではなく，別のジャンルに出現しやすい出来事も含まれる．このことから，単語がどのジャンルに出現しやすいかをベクトルとして表すことを考える．単語 i の単語特徴ベクトルは $w_i = (w_{i1}, \cdots, w_{ij}, \cdots, w_{iJ})$ で表される．ここで，j は映画が分類されるジャンルに対応する．w_{ij} はジャンル j に分類されるあらすじ全体の中の単語 i の出現頻度である．ただし，1 つの文書が複数のジャンルに分類される場合が多く，その場合は複数のジャンルに属するものとした．ジャンル分類としては映画紹介サイト The Internet Movie Database(IMDb)[4]上で公開されている以下の映画のジャンル 21 種を用いた．

Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, History Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War, Western

3.3 系列データの作成

3.2 で作成した単語特徴ベクトルを文書の出現順に並べることで，あらすじを系列情報で表現する．例として表 1 に映画ターミネーターのあらすじを表現した単語特徴ベクトル系列の一部を示す．単

表1：単語特徴ベクトル系列の例

順序	単語	Action	Crime	Horror	Romance	Sci.-Fi	Thriller
	⋮						
13	やってくる	0.07	0.06	0.04	0.07	0.06	0.09
	⋮						
51	核	0.23	0.05	0.05	0.00	0.14	0.05
52	戦争	0.22	0.02	0.02	0.02	0.07	0.02
53	勃発	0.19	0.10	0.00	0.05	0.05	0.05
	⋮						
87	写真	0.09	0.06	0.03	0.09	0.06	0.03
88	もらう	0.09	0.04	0.02	0.06	0.02	0.04
	⋮						
101	恋	0.11	0.01	0.03	0.15	0.04	0.00
	⋮						

語特徴ベクトルの値に着目すると，異なる出来事を表した単語である「戦争」と「恋」では，異なる要素の値が高くなっている．また，連続して出現している「核」「戦争」「勃発」は共通した要素の値が高い．このことから，連続して出現する単語が同じ出来事を表しているとき，その単語特徴ベクトルは類似したものになると期待できる．3.4 ではこれを考慮した系列の比較を行う．

3.4 DP マッチングによる類似度

3.3 で表した特徴ベクトル系列を比較するために DP マッチングを用いる．DP マッチングは，要素数の異なる系列間の類似度を測る方法である[3]．

単語特徴ベクトル系列データ A，B に対して本稿では式(1)により系列間類似度の部分和 $g(i,j)$ を計算した．

$$A=a_1,a_2,\cdots,a_i,\cdots,a_I \quad B=b_1,b_2,\cdots,b_j,\cdots,b_J$$

$$g(i,j)=\max\begin{bmatrix} g(i-1,j)+sim(a_{i-1},a_i) \\ g(i-1,j-1)+sim(a_i,b_j) \\ g(i,j-1)+sim(b_{j-1},b_j) \end{bmatrix} \quad \cdots(1)$$

特徴ベクトル間の類似度 $sim(x,y)$ にはコサイン尺度を用いた．ここで,式(1)内の $sim(a_{i-1},a_i)$ と $sim(b_{j-1},b_j)$ は系列内で連続する単語特徴ベクトル間の類似度であるが，単語特徴ベクトルが類似する単語は削除，挿入が生じても問題が少ないと考え，このように定義した．系列間の最終的な類似度は $g(I,J)$ であり，これに要素数による重み付けを行ったものをストーリーの類似度とした．

4. 主観評価実験

4.1 概要

提案手法を評価するため，類似度が高いあらすじ同士を被験者に読み比べてもらい，アンケート方式でストーリーの似ている度合を回答してもらった．

4.2 実験条件

比較手法：従来手法として文書全体に対してベクトルを求めて類似度を計算した．ベクトルの要素として単語の tf-idf を次元圧縮したものを用い，コサイン尺度で類似度を計算した．

対象文書：利用した文書はキネマ旬報映画データベース[5]で公開されている 1,000 文字前後の映画のあらすじ 255 本である．利用した単語は 11,027 種類であり，それぞれを 21 次元の単語特徴ベクトルで表した．

評価用データ：まず，255 本のあらすじから，ジャンルが異なる 10 本を選び，これを検索用あらすじとした．次に，各々の手法で検索用あらすじと類似度が高い上位 10 件を選び，手法間で重複していない上位 3 件を評価用あらすじとした．

評価値の設定：評価値は事前に行った予備実験とアンケートの結果から，下記の 4 段階で定義した．

1: 似ていない　2:少し似ている
3: 似ている　　4:とても似ている

被験者：学生 20 人を被験者とした．検索用あらすじと評価用あらすじの組み合わせを 1 人 12 件ずつ読み比べてもらい，合計 240 件の評価結果を得た．

4.3 結果

各手法の評価の平均値を表 2 に示す．

表2：評価の平均値

	評価の平均値
提案手法	2.3
従来手法	1.8

表から提案手法によりストーリーが似ているあらすじが選択できていることが確認できる．

5. おわりに

本稿では，文書を単語特徴ベクトルの系列で表し，比較することでストーリーを考慮した類似度計算ができると考えた．具体的には DP マッチングによる類似度計算法を提案した．実際の映画のあらすじを用いた主観評価実験の結果，ストーリーが似ているあらすじを検索できていることを確認した．今後は，対象文書を増やした実験を行う予定である．

参考文献

[1] 土方嘉徳：“情報推薦・情報フィルタリングのためのユーザプロファイリング技術”，人工知能学会論文誌，Vol.19，No.3 (2004)
[2] 赤石美奈：“文書群に対する物語構造の動的分解・再構成フレームワーク”，人工知能学会論文誌，Vol.21，No.5,pp.428-438 (2006)
[3] 迫江博昭，千葉成美：“動的計画法を利用した音声の時間正規化に基づく連続単語認識”,音響学会誌，Vol.27，No.9，pp.483-500 (1971)
[4] “The Internet Movie Database”，http://www.imdb.com/
[5] “キネマ旬報映画データベース”，http://www.kinejun.jp/