D-022

動画に対するコメントを利用した自動 Web 検索システム

Automatic Web Search System Using Comments for Video Contents

成田 龍太* 全 眞嬉 * 徳山 豪 *

Ryouta Narita Jinhee Chun Takeshi Tokuyama

1 はじめに

動画映像に対し関連した Web 情報などをリアルタイムに表示することは，宣伝やユーザの利便性を考えると将来的に非常に有用なアプリケーションであるが，一般には画像解析等のコストの高い処理が必要とされる．しかし，映像に何らかの文字情報が含まれていた場合，映像自体を解析して情報を抽出するよりも文字情報を積極的に利用するほうがコストが低く，リアルタイムでのシステム実現のために有功なアプローチである。テレビのニュース放送の字幕テロップから関連記事を検索する先行研究 [1] があるが，本論文ではこれを拡張し，動画映像への不特定多数からの (制御されていない) コメントを利用して関連記事の自動検索システムを実装し，実験によりリアルタイムで動作する事を示した．

2 関連研究と本研究の比較

文書情報に対する関連ページ検索の研究は多く行われており，Web advertising という大きな分野として注目されている [2, 3, 4, 5]．しかしながら，動画に対しては計算コストの高さと情報抽出の精度に問題があり，いまだに研究が開始された段階である [6]．

一つの有望な取り組みは，画像を解説する文字情報を含む動画を対象とすることであり，Henzinger ら [1] は CNN のニュース放送の字幕から WEB 上の関連記事を自動検索する研究を行った．現実には CNN ニュースでは前処理を行い，人の手で関連記事を選び，オンエア時に提示するというシステムが用いられている．それに対し，Henzinger らは前処理をせずに，ニュースに流れる字幕からオンラインに特徴語を取り出し，その特徴語を利用して WEB 上の関連記事を検索する全自動システムを提案し，リアルタイムで高い精度で関連する記事を選び出すことに成功している．

CNN ニュースの字幕テロップはそれ自体がニュースの内容を効果的に示した，推敲された文字情報となっている．それに対して，本論文ではより難しい設定として，動画にその内容に関連した文字情報が与えられているが，推敲されず目的も多様であるもの，たとえば視聴者コメントのようなものである場合を考え，それを利用して関連記事の自動検索を行えることを検証した．

技法的には，Henzinger らの研究では字幕の名詞を特徴語として取り出し，字幕中での重要性を表す重みを与えて特徴語の分類を行う．そして重みの大きい特徴語 2 つを検索語として関連記事を検索・提示を行っている．

それに対し，本論文でも基本的に特徴語を利用したアイデアを用いるが，複合語の利用や出力ページの関係から生じるグラフを利用し，上記のようなより難しい設定でも精度の高い情報検索を実現している．形態素解析などの技法を用いた日本語独自の処理を含めても，アルゴリズムは非常に高速であり，動画の映写に遅れないリアルタイムでのシステム実現を与えている．

3 本研究で利用する技法の説明

3.1 ベクトル空間モデル

関連 Web ページを検索する手段としてベクトル空間モデルを使用する．ベクトル空間モデルは情報検索の分野で多く利用されている技法であり，出現する単語に基づいて文書を 1 つのベクトルで表現し，ベクトルによって文書の内容を近似するという方法である [7]．

*東北大学大学院情報科学研究科システム情報科学専攻

重み付け

文書をベクトルで表現するために，単語それぞれの重み付けを行う．単語の重み付けに tf-idf(term frequency・inverse document frequency) 法を使用する．この方法は一般的に広く使われている手法であり，これは文書中に多く現れ，かつ，少ない数の文書しか表れない単語は特徴的であるという考え方である．

文書中の単語出現頻度である tf 値と，単語の希少性を示す idf 値を用いて，重み $w(T,t)$ を以下のように定める．

$$w(T,t) = tf(T,t) \times idf(t)^2$$

特徴ベクトル作成

文書 T 中に特徴語 $t_1, t_2, ..., t_n$ があるとき，T の特徴ベクトル $\vec{d}$ は tf-idf 重み w を用いて以下のように表される．

$$d(\vec{T}) = (w(T,t_1), w(T,t_2), ..., w(T,t_n))^T$$

類似度の計算

ベクトル空間モデルでは，特徴ベクトルの向きが近いほど文書の類似度が高い．したがって，2 つの文書の類似度を以下のように定める．

$$\cos\theta(\vec{d_1}, \vec{d_2}) = (\vec{d_1} \cdot \vec{d_2})/(|\vec{d_1}||\vec{d_2}|)$$

4 関連 Web ページ検索システムの設計

上記のようなベクトル空間モデルでの類似度を用いて，動画へのコメントを利用した関連 Web ページ検索のシステムの設計を行った。

特に，対象とする日本語対応映像コンテンツとしてニコニコ動画 [8] を考え，それに適したシステムを試作した．ニコニコ動画では，視聴者が動画にコメントを貼り付けることができ，これらのコメントを動画に対する追加の文字情報として得ることができる．本研究ではこれらの貼り付けられた日本語コメントを利用して関連ページを自動出力する．

いわば，画像解析の部分を一般視聴者に任せ，文字情報への翻訳として利用するのである．ニューステロップとは異なり，視聴者のコメントは時として無責任であり，画像の本質を抽出した推敲された文章ではない．しかしながら，複数の視聴者からのコメントを統計的に利用すれば，ニューステロップのような推敲された文字情報に近い情報が得られると期待した。

以下にシステム実装の具体的な手順を示す．

特徴語ベクトルの生成と候補 Web ページの検索

特徴語としてコメント中の名詞を取り出し，特徴語ベクトルを作成する．名詞は形態素解析で抽出するが，複合語も特徴語として利用するので，結びつきの強い複合語の同定を行わなければならない．具体的には idf 値を利用して，複合語を分割した場合と結合した場合の特徴語としての有効性を比較し，適切な複合語の同定を行う．特徴語を抽出した後重みを計算し，重みが大きい上位 10 語を要素とする（すなわち 10 次元空間の）ベクトル空間を考え，この中でニコニコ動画でのコメント文書の特徴語ベクトルを作成する．

これらの特徴語のうち Henzinger らは w がもっとも重い 2 語を用いて AND 検索して関連ページを探した．本研究ではこれでは不足であったので，少し拡張し，特徴語ベクトルの重みが大きい上位 3 語のうち 2 つの組合せを取り出し，各々に対

し Google で AND 検索を行い上位 10 件を関連ページ候補として取り出す．それぞれの候補ページに対して動画の特徴語ベクトルと同じ語句の *tf-idf* 重みを求め，特徴語ベクトルを作成する．

候補ページからの出力ページの選択

候補ページから，動画と関係の深いページの抽出を行う作業が必要である．Henzinger らは類似のページをできるだけ排除し，多様のページを出力したが，本稿ではそれに当てはまらない．

本研究では，動画のメインテーマに関連した複数のページを取り出すことを目的とした手法を考えた．そのため候補ページ同士の関係を表すグラフを作成し，クリークを用いることで関連ページの抽出を行う．クリークを用いることで検索結果のクラスタリングを行うことができ，メイントピックに絞った関連ページを抽出できると期待される．

グラフは候補ページをノードとした構造を考える．動画の特徴語ベクトルと候補ページの特徴語ベクトルとの類似度をノードの重み $\sigma(v)$ とし，候補ページ u，v 同士の特徴語ベクトルの類似度 $\mathrm{sim}(u,v)$ を用いてエッジ $e=(u,v)$ の重みを $\tau(e)=\frac{1}{1-\log \mathrm{sim}(u,v)}$ とする．ただし，エッジの重みがある閾値以下の場合は候補ページ同士に関連性が無いとみなし，エッジを結ばないこととする．図 1 に作成例を示す．関連ページ抽

辺	類似度	辺	類似度
e_{12}	0.76	e_{24}	0.76
e_{13}	0.66	e_{25}	0.2
e_{14}	0.87	e_{34}	0.4
e_{15}	0.3	e_{35}	0.1
e_{23}	0.66	e_{45}	0.58

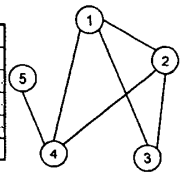

図 1: グラフ作成例

出の方法としては，検索結果のグラフから次数 5 のクリークを取り出し，クリーク重み W を求める．クリーク重みはエッジ重みの総和とノード重みの総和の積 W で求めている．クリーク K=(V(K),E(K)) としたとき，クリーク重み W は以下のように表される．

$W=\sum_{e\in E}\tau(e)\sum_{v\in V}\sigma(v)$

すべてのクリークに対してクリーク重みを求め，もっとも大きいクリークの web ページを出力する．本システムの出力は web ページのタイトル，要約，URL を表示する．図 2 にシステム実行例を示す．

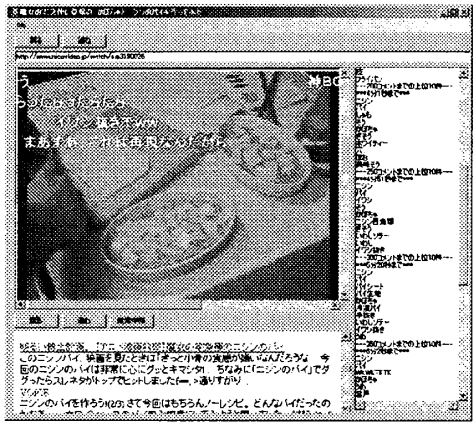

図 2: システム実行例

5 実験・考察

システムの性能を調べるために出力の正確性を計測する．正確性とは，出力されたページのうち関連ページ数の割合で表される．

本稿では後処理のアルゴリズムを 2 種類用意することで，後処理の性能を比較する．動画の特徴語ベクトルとの類似度が最も高い上位 5 つの Web ページを出力するシンプルなアルゴリズム *Alg*1 と前述で述べたクリークを用いたアルゴリズム *Alg*2 で比較実験を行った．尚，出力された結果が関連があるかどうかは人間が判定する．表 1 に実験結果を示す．

表 1: 後処理による正確性の違い

動画	動画 1	動画 2	動画 3	動画 4
*Alg*1(%)	62	72	24	66
*Alg*2(%)	64	78	46	76

表 1 の実験結果からクリーク (*Alg*2) を採用することで正確性が向上していることがわかる．特徴ベクトルに動画と関連性が低いあいまいな単語が含まれると，検索結果において動画と関連性が低いあいまいなページの特徴語ベクトルに対する類似度が高くなってしまう．しかし，あいまいなページというのは他のページとの類似度が低いと考えられる．また，特徴語に少数のあいまいな単語が含まれていても，動画と関係あるページが複数含まれている場合は特徴ごとの類似度が低くてもページ間の類似度が高くなると考えられる．このためクリークを採用することで，あいまいな単語によって動画との類似度が低かった関連ページは他ページとの類似度の高さで補うことにより，結果として出力することができるようになる．

また，*Alg*2 のシステムの実行時間を表 2 に示す．

表 2: システム実行時間

動画	動画 1	動画 2	動画 3	動画 4
download(msec)	7110	6288	7537	5910
計算時間 (msec)	14.1	3.4	3.75	9.5

表 2 より検索結果の html ダウンロードとクリーク計算を合わせて 10 秒を切っている．コメントは平均で約 50 秒ごとに分けて入力されるので実時間で関連ページが出力できることとなる．

6 まとめ

本稿では，日本語の映像コンテンツに対する視聴者からのコメント情報を利用し，関連ページを自動に検索するシステムを実装した．また，入力を形態素解析する際に，複合語の判定をする際に確率を用いて複合語の判定を行った．検索後の関連ページ抽出をベクトル空間モデルによりグラフを作成し，クリークを用いて関連ページ抽出手法を提案した．既存手法の直接の適用では，ニコニコ動画では動画に対して有効なコメントがあまりなく関連ページが検索できないと思われたが，本論文にあるような工夫を行うことにより，高い正確性で関連ページを表示することができた．

参考文献

[1] Monika Henzinger, Bay-Wei Chang, Brian Milch, Sergey Brin, "Query-Free News Search", *World Wide Web* 2005, 101-126

[2] 河重貴洋, 小山聡, 大島裕明, 田中克己, "質問修正と再ランキングを用いた文脈依存 Web 検索", DEWS,2006

[3] 成田宏和, 太田学, 片山薫, 石川博 "Web 文書検索のための非排他的クラスタリング手法の提案" DEWS,2003

[4] 阿倍倫子, 細野公男, 中川裕志 "コメント文を利用する映画ナビゲーション" 言語処理学会大会,2001

[5] Andrei Broder,Marcus Fontoura,Vanja Josifovski,Lance Riedel "A Semantic Approach to Contextual Advertising" *Proceedings of the 30th annual international ACM SIGIR conference on Research and development information retrieval*, 2007

[6] Qiang Ma, Akiyo Nadamoto, Katsumi Tanaka "Complementary Information retrieval for Cross-media News Content" Information Systems, Vol.31, 2006, 659〜678

[7] 大谷紀子, "情報検索におけるベクトル空間モデルの応用" 武蔵野工業大学環境情報学部研究論文 3-6

[8] ニコニコ動画: http://www.nicovideo.jp/