2F4-2

Synvie: ビデオブログコミュニティから獲得されたアノテーションに基づく応用

Synvie: Applications Based on Annotation Acquired From Videoblog Community

山本 大介[*1] 増田 智樹[*1] 大平 茂輝[*2] 長尾 確[*3]
Daisuke Yamamoto Tomoki Masuda Shigeki Ohira Katashi Nagao

[*1]名古屋大学 情報科学研究科
Graduate School of Information Science, Nagoya University

[*2]名古屋大学 エコトピア科学研究所
EcoTopia Science Institute, Nagoya University

[*3]名古屋大学 情報メディア教育センター
Center for Information Media Studies, Nagoya University

In this paper, we propose a mechanism which acquires semantics of video contents as annotations from related Weblog communities. In particular, we have implemented a Web-based tool which user can generate a Weblog entry quoting video scenes easily. This tool can acquire relationships which associate multiple video scenes with a document structure of a Weblog entry from editing histories. In the result, we can acquire higher quality annotation than previous works which acquire information from communication with online message board systems. We have developed an online video quotation system "Synvie." Moreover, we have analyzed real annotation data which were accumulated using the public beta service which we are providing, and confirmed the utility of our system. As examples of applications based on these annotation, we proposed a video scene retrieval system.

1. はじめに

近年，YouTube[*1]や Google Video[*2]などといったユーザ投稿共有サービスが盛んに提供されている．これらのサービスでは，ビデオコンテンツの投稿や閲覧のための機能のみを提供するだけではなく，ビデオコンテンツに対するコメントの付与や共有・投票・評価・タグ付け等を行うための機能や，ブログエントリー上でのビデオコンテンツの埋め込みを支援する機能を提供しているものも多い．その一方で，これら膨大なコンテンツを効率よく配信・検索・蓄積したいという要求が高まっている．

従来，映像コンテンツの検索や要約などの応用を行う場合，映像認識や音声認識等の自動解析技術を利用し映像に関するメタ情報を取得する自動アノテーション方式 [Wactlar 96] や，専任の作業者が映像のシーンに対するアノテーション情報を専用のツールで付与する半自動アノテーション方式 [Davis 93, Smith 00, Nagao 02] によって付与されたアノテーション情報を利用する必要があった．しかしながら，とりわけ個人が作成したコンテンツの場合，手ぶれ・ピンぼけ・雑音・不明瞭な声などといった撮影者の技能の問題や，カメラ付き携帯電話やデジカメといった撮影機器の性能問題等から映像や音声の品質のばらつきが大きく自動認識・解析は極めて限定的にしか利用できない．また，専任の作業者による半自動アノテーションを行うためには，視聴者が限定され費用対効果が見合わない等の理由から，全ての映像コンテンツに対するアノテーションを施すことは困難である．

そこで，筆者らは以前の研究で，マルチメディアコンテンツの配信とそれらを取り巻く Web コミュニティの活動とを効果的に連携させる仕組みを提案した．その仕組みは，それらのコミュニティにおけるユーザの自然な知的活動からコンテンツに関する知識をアノテーション [Nagao 01] として獲得・蓄積・解析することを目的として設計されている．具体的には，二つのコミュニケーション手段を提供する．一つ目は，映像コンテンツの任意のシーンに対して，コンテンツの内容に対する感想や評価などの情報の関連付けを支援する掲示板型コミュニケーションの仕組み [山本 05] であり，二つ目は，任意の映像シーンを引用した Weblog エントリーを生成し，映像シーンとそれらの記事の文書構造との関連付けを支援する Weblog 型コミュニケーションの仕組み [山本 07] である．これらの仕組みを作成することによって，ユーザらによる映像を題材としたコミュニケーションを支援する．さらには，コンテンツの内容とこれらのコミュニケーションとを詳細に結びつけることによって，コンテンツに付随する様々な情報をアノテーションとして獲得する．このような方式ならば，映像の画質や音質に左右されず，また，アノテーションにかかるコストも発生しない利点があり，上述した自動・半自動アノテーションの問題を回避できる．これらのコミュニケーションは単体のコンテンツのみに閉じているのではなく，コンテンツを部分引用する仕組みによって，Web 全体を対象とするより広がりをもつコミュニティの構築を支援する．また，人気のある映像コンテンツほどより多くのアノテーションの取得が可能になる利点がある．

本論文では，システムの公開実験に基づく分析・評価を行い，コミュニケーションから得られるアノテーションを用いたアプリケーション作成について議論する．

連絡先: 〒 464-8603 名古屋市千種区不老町 IB 電子情報館南棟 395 号室
名古屋大学 情報科学研究科 長尾研究室
TEL : (052)-789-5878 / FAX : (052)-789-5875
email: yamamoto@nagao.nuie.nagoya-u.ac.jp

*1 http://www.youtube.com/
*2 http://video.google.com/

2. Synvie によって収集されるアノテーション

我々は，Synvie というビデオ共有サービスを試験的に提供[*3]している．任意の映像シーンに対するユーザコメントの投稿やボタン評価などのアノテーション機能や，任意の映像シー

*3 http://video.nagao.nuie.nagoya-u.ac.jp/

ンを引用したブログの執筆を支援するための機能などを備えている．

Synvie によって収集されたアノテーションを，対象単位，行為，データ型で分類したものを表 1 にまとめる．映像の任意のシーンに対するアノテーションをシーンアノテーションと呼び，特に，コメント投稿をシーンコメントアノテーション，画面矩形領域に対するコメント投稿をシーン領域コメントアノテーション，ボタンによる評価をシーンボタンアノテーションと呼ぶ．映像の任意のシーンを引用してブログエントリー上でコメントを記述し映像シーンとコメントとの関連付けすることをシーン引用アノテーションと呼び，特に，連続するシーンを引用した引用アノテーションを連続シーン引用アノテーション，連続しない複数シーンを引用した引用アノテーションを非連続シーン引用アノテーションと呼ぶ．複数の映像シーンを引用したブログエントリーをビデオブログ (図 1) と呼ぶ．また，映像コンテンツ全体に対するコメント付与をコンテンツコメントアノテーション，タグ付与をコンテンツタグアノテーションと呼ぶ．それぞれのアノテーションに対応した専用のインタフェースを開発し，ユーザがそれぞれのアノテーションに最適化されたツールを用途や目的によって使い分けることができる．また，タイトルやコンテンツの紹介文などあらかじめ入力されているメタデータ情報もアノテーションとして利用する．

Synvie によって取得されるアノテーションには以下のような特徴がある．アノテーションの量はコンテンツの面白さや話題性などに依存しやすい．一般ユーザにとって関心のあるシーンにより多くのアノテーションが集まりやすいため，時間軸に対するアノテーション密度の偏りが大きい．また，アノテーションの質や信頼性は，アノテーションを付与する人の知識や性格に依存し，また，シーンアノテーションよりもシーン引用アノテーションの方がよりテキスト品質が高いなどアノテーションの編集法に依存することが分かっている．そのため，アノテーションの量や偏りにロバストであること，アノテーションの質は人や編集スタイルに依存しやすいという特徴を考慮する必要がある．

3. アノテーションの解析

本システムでは，コメントアノテーションやシーン引用アノテーションを，なるべく情報の欠落がないがない形式で蓄積する．そのため，本研究で意味するところのアノテーションはユーザコメントの列挙にすぎず，それ自身が機械によって理解可能な情報とは限らない．つまり，本研究によって取得されたアノテーションを用いたアプリケーションを構築するためには，アノテーションを解析し，機械が理解可能な情報に変換する必要がある．そこで，本章では 3 つの視点からアノテーションを解析する手法を提案する．一つは，アノテーションのテキスト情報からコンテンツの意味内容を表す情報の抽出を行う仕組

表 1: 公開実験によって取得されたアノテーション

対象単位	行為	データ型	アノテーションタイプ
コンテンツ	投稿	単語	コンテンツタグ
		文	コンテンツコメント
	—		コンテンツメタデータ
シーン	投稿	ボタン情報	シーンボタン
		文	シーン領域コメント
			シーンコメント
	引用		連続シーン引用
			非連続シーン引用

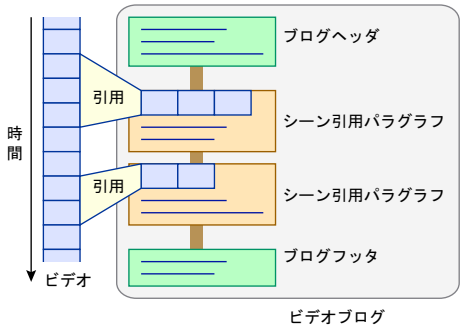

図 1: ビデオブログの構造

みであり，具体的には，ビデオコンテンツ全体およびシーンの内容を表現するキーワード（一般にタグと呼ばれる）の抽出を目指す．二つ目は，アノテーションや映像シーンの各々の重要性の計算手法の提案であり，三つ目は，各々のアノテーション間やシーン間の関連性についての考察である．これらは，アノテーションに基づく応用を実現するために重要な情報である．

3.1 タグの抽出

アノテーションによって取得されたテキストからコンテンツやシーンの内容を表現するキーワードの抽出を行う．コンテンツと対応付けられたキーワードをタグと呼ぶ．特に，コンテンツ全体の内容を表現するタグをコンテンツタグといい，シーンの内容を表現するタグをシーンタグと呼ぶ．コンテンツタグ・シーンタグともに以下の手法によって抽出する．まず，それぞれの自由コメントを形態素解析器茶筌 [奈良先端科学技術大学院大学 03] を用いて形態素に分割する．それぞれの形態素から，名詞・動詞・形容詞・形容動詞・未知語を抽出する．ただし，代名詞や非自立名詞・非自立動詞は除外し，未知語は固有名詞として扱った．さらに一般的に不要語と判断可能な形態素（たとえば，する，ある，なる，できる，いる，等）も除外した．それぞれの形態素の基本形をタグとする．

3.2 アノテーションとシーンの重み

アノテーションやシーンの重みの計算手法を議論する．ここでいうアノテーションの重みとは，そのアノテーションが対象となる映像シーンの内容をどれだけ的確に，かつ，信頼性が高く表現しているかを示す指標であり，シーンの重みとは，そのシーンがその映像の中でどれだけ重要なシーンであるかを示す指標である．

本論文では，アノテーション A の重み $w(A)$ は，アノテーションの対象粒度 $g(A)$，アノテータの信頼性 $r(A)$，アノテーションタイプの信頼性 $t(A)$ から推定する．つまり，信頼できる人がより正確にアノテーションを作成できるツールを用いて，より粒度の細かい対象（コンテンツよりもシーン，長いシーンよりも短いシーン）に対するアノテーションを付与した場合に，より高い重みを与える．本来ならばアノテーションの意味内容を加味したアノテーションの重み付けをすることが望ましいが，本論文では意味内容を考慮したテキスト解析は一般に困難であるため見送った．具体的な計算式は以下のとおりである．

$$w(A) = g(A) \times r(A) \times t(A) \quad (1)$$

なお，$g(A), r(A), t(A)$ はアノテーションの種類や内容に応じて，あらかじめ定義した実数を返す関数である．具体的には，$g(A)$ は，シーンに対するアノテーションの場合は 2，コンテ

ンツ全体に対するアノテーションの場合は 1 に，$r(A)$ はアノテーターが対象となる映像コンテンツの投稿者である場合は 3，登録ユーザの場合は 2，匿名ユーザの場合は 1 に，$t(A)$ はアノテーションタイプがシーンコメントアノテーションの場合は 1，シーン引用アノテーションの場合は 2 とした．

また，映像シーン S の重み $w(S)$ は，より多くの，よりアノテーションの重みが大きいアノテーションから参照されているシーンほど重要であると仮定し，それぞれのシーンを参照するアノテーションの重みの合計がその映像シーンの重みであるとし，以下の式で表す．

$$w(S) = \sum_{A \in refer(S)} w(A) \qquad (2)$$

これらの式や定数値は暫定的なものである．十分なデータが不足している，コンテンツの種類やコミュニティに依存しやすいため検証が困難などの理由から，妥当性の検証は今後の課題としたい．

3.3 アノテーション構造の活用

映像シーンに対するコメントアノテーションは，図 2 のように，対応する映像シーンとコメントとを「シーンコメントアノテーション」というラベルのついたグラフで表現される．コメントは映像シーンに関する情報を含んでいる場合が多く，映像シーンに対するアノテーションとして利用可能である．

その一方，ビデオブログエントリーは，図 3 のように，引用した映像シーンと Weblog エントリーのパラグラフとを「シーン引用」というラベルのついたグラフで表現され，他のシーンやコンテンツ，Weblog エントリーとの何らかの関連性の抽出が期待できる．

具体的には，連続シーン引用アノテーションによって選択された連続するショットからなる引用シーンでは，それに対応するコメント内容という観点に基づきシーンの連続性があるとみなすことができる．また，非連続シーン引用アノテーションを用いて選択されたショットの集合は，対応するコメントの意味内容という観点に基づいて，シーンの関連性があると考えられる．さらに，一つのビデオブログエントリーで複数のコンテンツを同時に引用した場合，そのビデオブログエントリーの内容に基づいて，これらのコンテンツの意味的な関連性があると捉えることが可能になる．複数のコンテンツを引用したビデオブログエントリーの例としては，CG アニメーション「ノラネコピッピ 1 話」とその元になった実写映像である「ノラネコピッピのモデルになった猫♪」を同時に引用し比較する記事などである．

本システムにより，Web と映像コンテンツの垣根を越えた引用に基づく詳細なネットワークを形成する．これにより Weblog ネットワークと映像コンテンツを統合することが可能になる．Weblog と映像コンテンツの統合されたネットワークでは，コンテンツを扱う粒度がコンテンツ/エントリー単位から映像シーン/パラグラフ単位へとより詳細になり，コンテンツに関連するコミュニティが共有サイト内から Web 全体に拡大されている．さらに，コンテンツ間のリンクをナビゲーションのための 1 方向的な Hyperlink から引用に基づく意味的な双方向リンクへと拡張させることができる．これにより，我々の提案する仕組みはコンテンツに付随する様々な知識を抽出するためのフレームワークとして機能し，それによって収集されるデータは検索やコンテンツ推薦などの様々な応用のための基礎的データとして利用されることが期待できる．

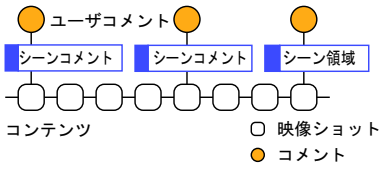

図 2: 映像シーンへのアノテーションのモデル

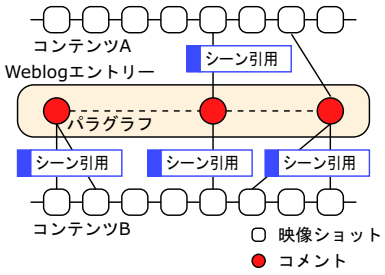

図 3: 映像シーン引用に基づくアノテーションのモデル

4. アノテーションに基づく応用

本実験によって取得されたアノテーションに基づく応用の例として，ビデオシーン検索システムなどを提案する．具体的な応用を試作することによって，本実験によって収集されたアノテーションの有用性を示す．なお，対象コンテンツの数が１００個前後と少ないこと，アノテーションの量は時間とともに増えていき，それが応用の精度や質に直結すること，母体となるコミュニティやコンテンツに強く依存することなどから定量的な評価が困難であるため，詳細な評価は今後の課題とする．本論文ではアノテーションに基づく応用の可能性について言及することに留めておく．

4.1 映像シーン検索

映像シーン検索とは，映像をコンテンツ単位ではなくシーン単位で検索しようとする仕組みである．我々の手法の特徴は，アノテーションから抽出されたタグを検索することによって，ビデオシーンを検索しようとしている点である．

具体的な検索プロセスは以下のとおりである．ユーザは，目的のシーンを検索するために，一つないし複数の検索クエリをタグ形式で入力する．それらのタグと一致するタグを含むアノテーションの検索を行い，一致したアノテーションをコンテンツ毎に列挙する．一致したアノテーションに対応するシーンを検索結果候補とする．

一つのコンテンツ内に多くのアノテーションが一致した場合は，検索結果候補が膨大かつ時間軸上で細切れになる危険がある．そこで，対象となるシーンが連続する，あるいは時間的に近い場合は類似するシーンである可能性が高いと考え検索結果候補を統合する．逆に，一致したアノテーションの数が少なく，また，分散しており，検索結果のシーンを特定できない場合はコンテンツ全体を検索結果候補とする．検索結果候補内に属するアノテーションの重みの合計が，その検索結果候補の重みとする．このような仕組みにより，アノテーションが多数存在する場合にも，アノテーションが少量しか存在しない場合にも，ある程度対応可能になる．検索結果候補の重みに基づき，検索結果候補のランク付けを行う．

検索結果候補の内容を理解するために，シーンの内容を表現するサムネイル画像を提示することは有効である．サムネイ

図 4: ビデオシーン検索システム

ル画像は，検索結果候補内のアノテーションに関連付けられているシーンに属するサムネイルを候補とする．ただし，サムネイル画像が一定個数以上存在する場合には，そのサムネイル画像が属する映像シーンの重みに基づいて絞る．

ビデオシーン検索システムのインタフェースを図 4 に示す．

検索が成功する例としては，検索したいシーンに的確なキーワードを含むアノテーションが存在する場合である．逆に，検索が失敗する例としては，検索したいシーンに的確なキーワードが含まれないなど，アノテーションの量が不足している場合が考えられる．しかしながら，人気のあるシーンやコンテンツには，より多くのアノテーションが集まりやすく，また，人気のあるシーンほど検索ニーズが高い，このようなシーンやコンテンツには自然にアノテーションが増えていくことが考えられる．すなわち，ある程度の時間が経過すれば，この問題は解決される可能性が高い．また，同じ内容を異なるタグで表現している場合にも検索に失敗する．その場合は，シソーラスを用いて類義語や語彙の上位概念・下位概念の関係を考慮する必要がある．

4.2 その他の応用

映像と同期して関連性のある他のコンテンツのサムネイル画像とキーワード，及びその根拠となるビデオブログエントリーを表示することによって，ビデオ推薦を行うシステムを開発している．従来からあるソーシャルフィルタリング [Shardanand 95] を用いたコンテンツ推薦システムでは，映像のシーン単位で推薦を行うのは困難であること，サンプル数が少ない場合には必ずしも精度が良くないこと，さらに，推薦となる根拠を統計的にしか示すことができないなどの欠点がある．本システムでは，ユーザが複数コンテンツを引用したビデオビデオブログエントリーを執筆した情報を用いて，統計情報に頼らない詳細なコンテンツ推薦を実現している．

また，ビデオスキミングシステムを開発している．ビデオスキミング [是津 00] とは，映像の重要なシーンのみを通常の速さで再生し，それ以外のシーンを早送りで再生する仕組みである．これにより，映像の内容を短時間で把握するのに適している．具体的には，ある一定時間内に収まるように，映像シーンの重みが高い順に選別を行う．

5. おわりに

本論文では，映像シーンへのアノテーション，映像シーン単位でのコンテンツの引用に基づく Weblog エントリーからのアノテーション取得方法の提案，コミュニケーションに特化した具体的なインタフェースの提案と公開実験に基づく評価を行った．これにより，それぞれのアノテーションタイプによって得られるアノテーションの傾向をアノテーションの量と質の観点から分析を行い，それぞれのアノテーションに特有の傾向が見られることが分かった．特に，関連する Weblog エントリーから情報を抽出することが質の高いアノテーションを抽出する手助けになることが示せたことが有用であると考えている．これは，シーンコメントアノテーションが掲示板文化を引き継いでいるのに対して，シーン引用アノテーションは Weblog 文化を引き継いでいることを反映していると考えられる．さらに，これらのアノテーションに基づく応用システムをいくつか開発した．また，これらのアノテーションは，二つの観点により映像の構造的・意味的情報も抽出可能である．一つは，コンテンツを引用することによってそれぞれのショット間の意味的な関係の抽出が期待できる．もう一つは，引用によって複数のコンテンツ間の意味的な関係の抽出が期待できる．

今後の課題として，応用システムの評価に基づく Synvie のアノテーションシステムとしての有用性の検証と改良がある．

謝辞

本研究は独立行政法人情報処理推進機構 (IPA) による 2005 年度上期未踏ソフトウェア創造事業の支援を受けた．

参考文献

[Davis 93] Davis, M.: An Iconic Visual Language for Video Annotation., in *Proceedings of the IEEE Symposium on Visual Language*, pp. 196–202 (1993)

[Nagao 01] Nagao, K., Shirai, Y., and Squire, K.: Semantic Annotation and Transcoding: Making Web Content More Accessible, *IEEE MultiMedia*, Vol. 8, No. 2, pp. 69–81 (2001)

[Nagao 02] Nagao, K., Ohira, S., and Yoneoka, M.: Annotation-Based Multimedia Summarization and Translation, in *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING-02)*, pp. 702–708 (2002)

[Shardanand 95] Shardanand, U. and Maes, P.: Social Information Filtering: Algorithms for Automating “Word of Mouth”, in *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, Vol. 1, pp. 210–217 (1995)

[Smith 00] Smith, J. R. and Lugeon, B.: A Visual Annotation Tool for Multimedia Content Description, in *Proceedings of the SPIE Photonics East, Internet Multimedia Management Systems*, pp. 49–59 (2000)

[Wactlar 96] Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M.: Intelligent Access to Digital Video: Informedia Project, *IEEE Computer*, Vol. 29, No. 5, pp. 140–151 (1996)

[山本 05] 山本 大介, 大平茂輝, 長尾 確：オンラインビデオコンテンツを中心としたコミュニティ支援システム, 情報処理学会第 67 回全国大会 (2005)

[山本 07] 山本 大介, 増田 智樹, 大平 茂輝, 長尾 確：Synvie:映像シーン引用に基づくアノテーションシステムの構築とその評価, インタラクション 2007, pp. 11–18 (2007)

[是津 00] 是津 耕司, 上原 邦明, 田中 克己：映像の意味的構造の発見, 情報処理学会論文誌, Vol. 41, No. 1, pp. 12–23 (2000)

[奈良先端科学技術大学院大学 03] 奈良先端科学技術大学院大学 自然言語処理学講座：形態素解析システム 茶筌, http://chasen.aist-nara.ac.jp/ (2003)