科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１７　日現在

機関番号：　１７１０２
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００２９９

研究課題名（和文）集合知に基づく高品質コンテンツ検索

研究課題名（英文）High quality content mining using collective intelligence

研究代表者
伊東　栄典（Ito, Eisuke）

九州大学・学内共同利用施設等・准教授

研究者番号：９０２９４９９１

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：急増する利用者投稿型コンテンツを，視聴者が与えるタグやコメントを用いて検索する手法を研究した。コンテンツのランキング手法では，コメント内の感情語数に基づくものと，お気に入り登録を２部グラフとして捉えてグラフ構造解析により重み付けする手法を提案した。後者の手法は，将来人気になるコンテンツを早く見つけることが出来る手法である。コンテンツの自動カテゴリ分けでは，単語（タグ）をカテゴゴリと捉え，タグ群を出現頻度および共起頻度に基づき意味階層を与える手法を開発した。階層は非循環有向グラフの形で出力される。しかし意味的な階層化は充分とはいえない。今後も詳細な研究を行う必要がある。

研究成果の概要（英文）：A large number of contents are being uploaded as online novels. I proposed two ranking methods. One is based on viewer's sentimental comments, and the other is based on the users' favorite lists (bookmarks). Relation between users and favorites can be represented as a bipartite graph. In several genres, the later one can predict future popular contents.
I studied semantic hierarchization of tags, which is given by many users. The hierarchy is based on frequency and co-occurrence of tags, and a hierarchy is represented as a direct acyclic graph. However, the automatic generation of tag hierarchy is not enough for automatic categorization. More careful study is need for automatic categorization.

研究分野：　情報学

科研費の分科・細目：　情報学・図書館情報学・人文社会情報学

キーワード：　情報検索　コンテンツ検索　consumer generated media　集合知　sentiment analysis　semantic analysis

１．研究開始当初の背景

近年人気のある YouTube やニコニコ動画などの利用者投稿型動画提供サービスでは，毎日多数の動画が投稿されており，サイト全体では膨大な数の動画が蓄積されている。中には TV などの従来メディアで提供されていた動画と同程度の品質をもつコンテンツも含まれている。動画投稿サイトには楽曲も投稿されている。動画以外に小説・漫画・静止画・3D オブジェクトモデルの投稿サービスが立案されつつある。これらが有機的に連携することで，電子コンテンツが爆発的に生産されると思われる。電子書籍など，過去のコンテンツの電子化も進んでおり，これらが進むとネット上のコンテンツ数は膨大になる。

膨大なコンテンツから，利用者が求めるコンテンツを探すためには検索システムが必要である。コンテンツを高い品質で検索するためには以下の 2 つの機能が必要である。

・コンテンツの品質（面白さ）評価尺度
・コンテンツのカテゴリ分類

面白さの評価については，現在の YouTube やニコニコ動画でも，検索語入力および検索語に対する適合動画の並び替え手法の選択による動画検索が提供されている。並び替えには，投稿日時，閲覧回数，評価値（平均や累積値）によるものが提供されている。他に，カテゴリ指定による検索対象動画の制限機能もある。

品質評価に閲覧回数・閲覧者数を用いるのは十分ではない。これらの尺度は累積値であるため，古いコンテンツほど評価が高いことになりがちである。各閲覧者が付与する評価値の平均値を尺度とする方法は，閲覧回数の累積値よりも精度が高い。しかし，ほとんどの閲覧者は評価値として中間値（5 点満点の場合は 3 点）を与えることが多く，特定分野を好む人にとっては，その分野を好まない人の評価値がノイズとなる。評価値を用いる方法としては，Amazon で使われている協調フィルタリングがある。協調フィルタリングは，自分の評価値を入力する必要があり，その手間が問題である。

既存サービスのカテゴリ分類も不十分である。YouTube やニコニコ動画では現在，静的なカテゴリ分類が用意されている。ただし，コンテンツ作成者が付ける静的なカテゴリ分類・タグ付けに基づくもので，カテゴリの階層化や，類似カテゴリの提示などの機能はない。類似動画については，一つの動画への関連動画の提示機能が提供されている。関連動画の提示方法は明示されていないものの，視聴者の共起アクセスに基づく手法を用いているものと思われる。しかし，視聴した動画の関連動画から自分の好みの動画を見つけることは困難である。関連動画へのリンクを辿りつつ好みの動画を見つける確率は，動画空間の膨大さから考えて低い。

カテゴリ分類に近い手法としてタグクラウドがある。タグクラウドは，タグ数が少ない場合か，人気の高いメジャーなカテゴリを見つけるには有用である。しかし，コンテンツ数が膨大な場合，タグの数も膨大になる。

２．研究の目的

Web 上のコンテンツ投稿サービスにより，動画・楽曲・小説・マンガの電子コンテンツが爆発的に増加している。従来の紙メディアと異なり，Web 上のコンテンツ投稿サイトでは編集者やプロデューサーといったフィルタを通さないため，コンテンツへの品質保証がない。既存の閲覧回数などの評価尺度には問題がある。そこで多数の閲覧者が与える評価を集合知として用いるコンテンツの品質評価尺度とカテゴリ分類手法を提案し，それを用いた高品質な検索を実現する。

３．研究の方法

(1)集合知データ収集

Ruby 言語による収集プログラムを構築して，動画コンテンツとしてニコニコ動画の，動画メタデータおよび動画への視聴者コメントを収集する。他に，小説投稿サイト「小説を読もう」を対象に，小説メタデータと，小説へのコメントや，お気に入り小説（ブックマーク） のデータを集める。

(2) 品質評価尺度の提案と評価

視聴者コメントから感情語を抽出し，良い感情，悪い感情に分ける。また感情語を笑い，怒り，揶揄・嘲笑などの分類を行い，それらの出現回数を基に，コンテンツの品質評価を行う手法を提案する。

利用者とコンテンツの関係をグラフ構造として捉え，グラフ構造からコンテンツの良さを評価する手法も提案する。

(3) カテゴリ分類手法の提案と評価

タグ（単語）の出現頻度と，複数タグの共起出現頻度から，タグ間の上位語・下位語の関係を抽出する手法を提案する。

他にもタグを用いるコンテンツ２次元分類表示手法を提案する。

(4) 動画以外のコンテンツへの，提案手法の適用

動画以外として，オンライン小説を対象にする。

(5) 外部ソーシャルサービスからのタグおよびコンテンツ評価コメントの収集

Twitter や Facebook 等で，利用者がコンテンツへのリンクや評価を行うデータを収集する予定であった。

４．研究成果

(1)集合知データ収集

ニコニコ動画を対象に，動画メタデータと動画への視聴者コメントを収集した。当時の全動画 1700 万件について動画メタデータを収集した。集合知となる視聴者コメントの収集は，大変であるため，研究対象とした「音楽」動画だけを収集した。2011～2012 年前半には，ニコニコ動画サイトからの研究用データの提供が無かったため，研究室で Ruby 言語による収集プログラムを構築し，データ収集を行った（現在は国立情報学研究所経由でデータが提供されている）。

他に，小説投稿サイト「小説を読もう」を対象に，約 200 万件の小説メタデータと，約 150 万人の利用者が提供するお気に入り小説（ブックマーク）の情報を集めた。小説サイトの情報も，Ruby 言語による収集プログラムを構築してデータ収集した。

(2) 品質評価尺度の提案と評価

ニコニコ動画を対象に，視聴者コメントの感情語を統計解析して与える品質評価尺度を提案した[10,11]。この方法で，埋もれた面白い動画を探すことを実現した。しかしながら「内輪受け」をするような動画が高いランキングになり，万人が受け入れるような評価にはならない事がわかった。

うまくいかない原因は，コメントを投稿する視聴者の特性（好み）を考慮していないためである。歴史に関する動画に，歴史に詳しい視聴者と，そうでない視聴者が与えるコメントは，価値が異なる。より詳しい利用者のコメントを重要視することが必要であると分かった。

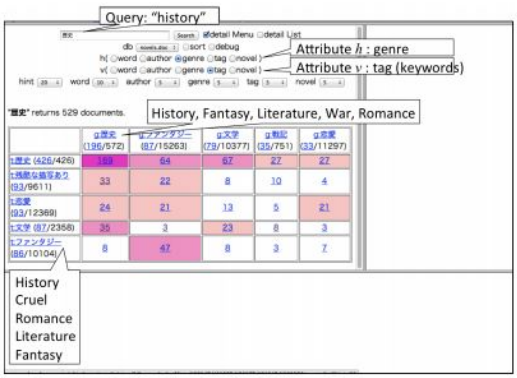
そこで，次に小説投稿サイトを対象に，小説の品質評価を行う事にした。ニコニコ動画は視聴者 ID がランダム化されて個人を特定できないのに対し，小説投稿サイト「小説を読もう」では読者コメントや，読者のお気に入り小説（ブックマーク）の情報が，利用者 ID と一緒に公開されている。

読者のお気に入り小説（ブックマーク）の情報を使い，読者と小説の関係を２部グラフとして構造化し，読者の好みを反映して小説へ評点を与える手法を提案した[7,8,9]。提案手法は，将来人気が高くなるコンテンツを早期に予想することが可能で，大変良い成果を得られた。

(3) カテゴリ分類手法の提案と評価

タグ（単語）の出現頻度と，複数タグの共起出現頻度から，タグ間の上位語・下位語の関係を抽出する手法を提案した[4,9]。

タグを用いて，コンテンツ２次元分類表示ツールを提案した[2,10]。このツールでは下図のように表形式で，コンテンツを分類表示できる。



(4) 動画以外のコンテンツへの，提案手法の適用

動画以外として，オンライン小説を対象に分析を行った。ニコニコ動画には視聴者の感情語の分析が適していたが，オンライン小説は感情的な単語が少ないため，グラフ構造を用いた品質評価の方が適していた。

(5) 外部ソーシャルサービスからのタグおよびコンテンツ評価コメントの収集

Twitter や Facebook 等で，利用者がコンテンツへのリンクや評価を行うデータを収集する予定であった。しかしながら，時間的な都合から，Twitter データの収集は行わなかった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　２件）

[1] Jun ZENG, Brendan FLANAGAN, Sachio HIROKAWA and Eisuke ITO: A Web Page Segmentation Approach Using Visual Semantics,IEICE TRANSACTIONS on Information and Systems, Vol.E97-D, No.2, pp.223-230, Feb.01 2014.

[2] Eisuke Ito, Kazunori Shimizu, Sachio Hirokawa, Development of Facet Analysis System for Diverse Online Novels, Journal of Data Processing, Volume 2, Issue 3, September, 2012, pp.113-119, 2012.

〔学会発表〕（計１０件）

[3] 呉 沢臣，伊東 栄典: Bilibili 動画サービスにおける感情コメント分析,人工知能学会 SIG-AM-06-03, pp.16-19, Mar.15, 2014.

[4] Eisuke Ito and Sachio Hirokawa: Keyword Relation Analysis Using Concept Graph Toward Automatic Categorization of Online Novels, Proc. of IIAI AIT2013, Nov.2013.

[5] Eisuke Ito, Brendan Flanagan,

Chengjiu Yin, Tetsuya Nakatoh and Sachio Hirokawa: A Private Cloud Environment for Teaching Search Engine Construction, Proc. of ICCE2013 (21st International Conference on Computers in Education), pp.391-397, Nov.2013.
[6] Eisuke Ito, Takahiro Urakawa, Brendan Flanagan, Sachio Hirokawa: Keywords frequency trend analysis of online novels, Proc. of IIAI AAI2013, pp.68-73, Sep.1, 2013.
[7] Kazunori Shimizu, Eisuke Ito, Sachio Hirokawa: Predicting Future Ranking of Online Novels based on Collective Intelligence, Proc. of ICDIPC2013 (The Third Int'l Conf on Digital Info. Processing and Communications), SDIWC, pp.261-272, Jan.30-Feb.1, 2013.
[8] 清水一憲, 伊東栄典, 廣川佐千男: 集合知に基づくオンライン小説のランキング手法,信学技法, Vol.112, No.346, DE2012-33, 電子情報通信学会, pp.107-112, Dec.13, 2012.
[9] Eisuke Ito, Kazunori Shimizu: Frequency and link analysis of online novels toward social contents ranking, Proc. of SCA2012 (The 2nd International Conference on Social Computing and its Applications), pp.531-536, Nov. 2012.
[10] Eisuke Ito, Sachio Hirokawa, Kazunori Shimizu: Introducing faceted views in diversity of online novels, Proc. of ICDIM2012 (Seventh International Conference on Digital Information Management), IEEE, pp.145-148, Aug. 2012.
[11] 村上直至, 伊東栄典: 動画コンテンツの視聴者コメントに基づくランキングとその評価, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 日本データベース学会, F8-3, 2012.
[12] Naomichi Murakami, Eisuke Ito: Emotional video ranking based on user comments, Proc. of iiWAS2011, pp.499-502, ACM, Dec. 2011.
[13] Kensuke Baba, Maso Mori, Eisuke Ito: Identification of Scholarly Papers and Authors, Proc. of NDT2011 (Networked Digital Technologies), Springer LNCS CCIS136, pp.195-202, Jul., 2011.

〔図書〕（計　0件）

〔産業財産権〕
○出願状況（計　0件）

○取得状況（計　0件）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
伊東　栄典（ITO Eisuke）
九州大学・情報基盤研究開発センター・准教授
研究者番号：90294991