動画共有サイトのタグを利用した検索情報システムの開発

Development of Search Information System that Using the Tag in the Video-Sharing Site

宮城　健治 *1,　谷口　祐治 *2
Kenji MIYAGI*1, Yuji TANIGUCHI*2
*1 国立大学法人琉球大学 工学部 情報工学科
*1 Department of Information Engineering, Faculty of Engineering, University of the Ryukyus
*2 国立大学法人琉球大学 総合情報処理センター
*2 Computing and Networking Center, University of the Ryukyus
Email: tomo@osn.u-ryukyu.ac.jp, taniguchi@cc.u-ryukyu.ac.jp

あらまし：動画検索をする際にキーワードやカテゴリによって動画を検索する．しかし，キーワードがわからないと検索が行えず，また大量の動画の中から動画を探すのは労力が要る．本稿では，動画検索にあたり動画のタグ間の共起頻度から新たな検索情報となるタグの推薦システムを提案する．タグから関連タグを得ることで動画を探す際の検索情報として使用できる．本研究ではニコニコ動画のデータを使用し，本稿のシステムに実装した．動画の類似度を示す係数として Jaccard，Simpson，NWD（Normalized Web Distance），ϕ^2（X^2統計量）を使用した．それぞれの係数からタグ間の類似度を求め，関連するタグを取得し，タグの特徴を考察した．
キーワード：動画投稿サイト，動画検索，タグ，共起頻度

1.　はじめに

動画サイトの人気が高まっている．動画サイトを利用するにあたり，キーワードやカテゴリなどを用いて検索を行う．しかし，自分が知っているキーワードでしか動画検索を行うことができず，新たに興味のある動画の発見ができないという問題がある．またカテゴリの検索では大枠での検索はできるが範囲が大きすぎてカテゴリ内から動画を探すのに労力が掛かってしまう．

新たな動画の検索方法としてタグを利用したタグの推薦による動画発見手法を提案する．タグとは動画の種類や特徴を表したものであり，タグによって動画の検索に利用している．タグは動画の情報を表す上で非常に有益だと考える．本稿ではユーザーがタグを自由に登録できるニコニコ動画 (1) でタグを使用し，動画データのタグ間の類似性から新たな検索情報となるタグの推薦システムを提案する．

2.　関連研究

ニコニコ動画のタグの共起頻度を利用した研究として，榊らの出現頻度 X^2 値を用いて出現頻度の正規化を利用した単語間の関連度の指標を使って，ニコニコ動画のタグ間の関連度を求めている (3)．また，ISR 手法を用いたタグ間の親子関係を示し，動画タグの階層化が行われている (4)．

3.　タグ検索システムの構築

3.1　システムの実装環境と構成設計

本システムの実装環境を表１に本システムの構成設計図を図１に示す．

表１　本システムの実装環境

OS	Mac OS X
RDB	SQLite3
Web アプリケーション フレームワーク	Ruby on Rails

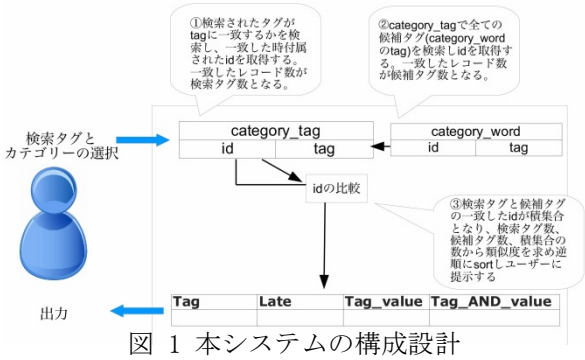

図 1 本システムの構成設計

4.　システムを利用した関連タグの取得

検索タグとの関連度を調べるため，類似度の指標係数を使用する．カテゴリータグとして使用されている「VOCALOID」「スポーツ」「アニメ」のタグが付いている動画のメタデータを集計する．カテゴリータグの中で動画に５０個以付けられたタグを候補タグ群とし，検索タグと候補タグの類似度を計上する．本稿の検索タグシステムを使用し，上位３０個のタグを取得する．予稿では「VOCALOID」に関する検索結果と考察を述べる．検索するタグとして，「VOCALOID」に関連性の強いタグである「初音ミク」で検索を行った．それぞれの類似指標

係数で，上位３０個のタグを取得し，候補タグ名，候補タグのタグ数，検索タグと候補タグとの積集合，類似度から，取得されたタグの特徴と関連性を調べる．

カテゴリタグ「VOCALOID」が付属した全ての動画データは 1,716,520 件で候補タグ群の数は 2,142 個である．

5. 類似度の指標係数

検索タグと候補タグとの類似度を計る係数の説明を行う．

Jaccard 係数は分母に２つの集合の和集合，分子に２つの集合の積集合を用いた，単語間の類似度を示す係数である．次式で与えられる．

$$\mathrm{Jaccard(x,y)} = \frac{x \cap y}{x \cup y} \qquad (1)$$

simpson(overlap)係数は分母に２つの集合の和集合，分子に２つの集合の内小さい集合を用いた，単語間の類似度を示す係数である．片方の単語のヒット数との差があまりにも大きいと関連性が低い単語の値が大きくなる場合がある．本稿では閾値を１，０００以上に設定し候補タグを集計した．次式で与えられる．

$$\mathrm{Simpson(x,y)} = \frac{x \cap y}{\min(\mathrm{x,y})} \qquad (2)$$

NWD(Normalized Web Distance)は NID(Normalized InformationDistance) を近似したものである．正規化情報距離の数式は次式（３）で求められる．NID は計算が原理的に不可能であるコルモゴロフ複雑性を含んいるが NWD ではコルモゴロフ複雑性の代わりに，検索エンジンで検索し得られたページ数（ヒット数）で近似することで計算できるようにした[5]．数式は次式で求められる．

$$\mathrm{NID(x,y)} = \frac{K(\mathrm{x,y}) - \min\big(K(x),\mathrm{K}(y)\big)}{\max\big(K(x),\mathrm{K}(y)\big)} \qquad (3)$$

$$\mathrm{NWD(x,y)} = \frac{\max(logf(x),logf(y))}{logN - \min(logf(x),logf(x))} \qquad (4)$$

X^2統計量に使われいてる φ^2を使用する[5]．X^2統計量の式から φ^2は次式のように変換ができ，$\mathrm{R}_{x\to y}$と $\mathrm{R}_{y\to x}$を掛け合わせた値となる．次式（5）で与えられる．

$$\varphi^2 = \mathrm{R}_{x\to y}\mathrm{R}_{y\to x} \qquad (5)$$

6. 検索結果

それぞれの係数で得た関連タグのタグ名，タグ数，検索タグとの積集合から取得されたタグの特徴を考察する．図２に Jaccard，Simpson，NWD，φ^2で取得した関連タグの棒線グラフを示す．

7. 考察

jaccard 係数では，共起頻度順の高い順に並び，タグ数の多い関連タグ，simpson 係数では，検索するタグとの類似が非常に高いタグ，NWD と φ^2では Simpson 係数と Jaccard 係数の２つの要素をもったタグが取得できた．類似係数によって異なる特徴のタグを取得できた．

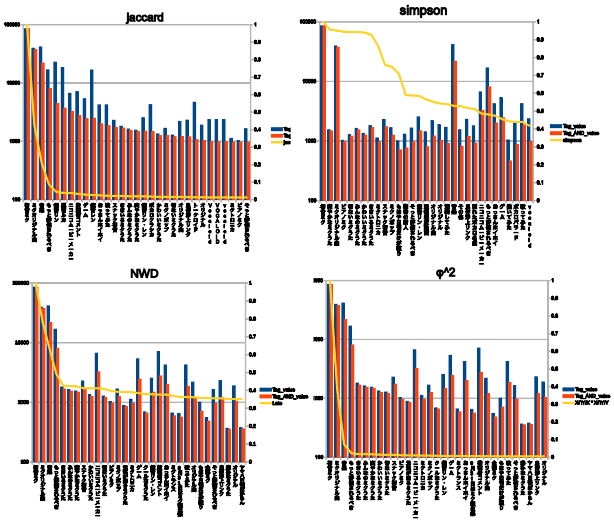

図２ Jaccard，Simpson，NWD，ϕ^2で取得した関連タグの棒線グラフ

8. 考察

jaccard 係数では，共起頻度順の高い順に並び，タグ数の多い関連タグ，simpson 係数では，検索するタグとの類似が非常に高いタグ，NWD と φ^2では Simpson 係数と Jaccard 係数の２つの要素をもったタグが取得できた．類似係数によって異なる特徴のタグを取得できた．

9. まとめ及び今後の課題

今後の課題として動画データ数の多さから検索が遅滞してしまう為，検索結果のレスポンスを高める必要がある．また，それぞれの類似度の指標係数から得たタグがユーザーにとってどれだけ有効なタグであるかテストする必要がある．

参考文献

(1) NicoNico ： http://www.nicovideo.jp/ （参照 2015.2.19）
(2) CodeZine ：”X² 乗値を関連度としたニコニコ動画タグ関連ネットワークの解析”，http://codezine.jp/article/detail/3516?p=3 （参照 2015.2.19）
(3) 村上 直至，伊東 栄典：動画投稿サイトで付与された動画タグの階層化，情報処理学会研究報告．MPS，数理モデル化と問題解決研究報告，Vol.2010-MPS-81 No.17，Vol.2010-BIO-23 No.17 pp 1-6 2010
(4) Cilibrasi, R.L., P.M.B. Vitanyi,“ Normalized Web Distance and Word Similarity, ”Handbook of Natural Language Processing, 2nd ed, pp. 293-314, 2010.
(5) 佐々木 靖弘，佐藤 理史，宇津呂 武仁：用語間の関連度を測る指標の提案，言語処理学会第１０回年次大会，pp.25-28