ネットワーク構造を利用したWikipediaからの意外性のある情報の抽出

野田陽平† 清田陽司‡ 中川裕志‡

東京大学大学院学際情報学府† 東京大学情報基盤センター‡

はじめに

近年，ウェブログやマイクロブログ，BBS，Wiki などの投稿者は膨大となり，CGM(consumer generated media)の記事数は急激に増加している．例えば日本においては，ウェブログの合計サイト数は 1,600 万サイトであり，その記事数は 130 億記事である．これらの集合知がウェブ上に蓄積するにつれ，それらの CGM の中から有用な情報を抽出する技術への要求が高まってきた．

これまでに，ウェブ上の文書からの情報抽出に関する多くの研究がなされてきた．それらの研究の多くはクエリ型やディレクトリ型の検索システムである．Google や Yahoo!に代表されるクエリ型検索エンジンは，ユーザから与えられたクエリに関連した記事を取得することを目的として作られている．一方，ディレクトリ型検索エンジンはディレクトリを辿ることで，ウェブページに辿りつくことができる．

本研究では，Wikipedia のカテゴリネットワークの構造を利用して，Wikipedia から意外性のある情報を抽出することを目的とする．

関連研究

レコメンドエンジンに関する研究においては，その評価指標として，ユーザの嗜好に対する一致の度合い以外に，意外性や新規性を導入するべきであるという主張がなされてきた[1]．これに対して Murakami ら[2]は，意外性のあるアイテムを推薦する手法を提案し，満足度が向上したことを報告している．

また，Nadamoto[3]らは，BBS や SNS などのコミュニティ型コンテンツの議論の中で，抜け落ちている視点を，その議論のトピックとなっている項目に関する Wikipedia の記事を軸に発見し，提示するシステムを提案している．

Discovering Serendipitous Information from Wikipedia by using its Category Networks

†Yohei Noda (noda@r.dl.itc.u-tokyo.ac.jp): Graduate School of Interdisciplinary Information Studies, University of Tokyo

‡Yoji Kiyota, Hiroshi Nakagawa: Information Technology Center, University of Tokyo

提案手法

本研究では，１つの記事が２つのカテゴリに同時に属しているような関係性を Wikipedia のダンプデータ[1]から，約 160 万件抽出し，その中から意外性のある関係性をもった情報を発見する．このデータの例を，表１に示す．

表１．カテゴリ A，B，記事の例

カテゴリ A	カテゴリ B	記事
Category：日本の内閣総理大臣	Category：オリンピック射撃競技日本代表選手	麻生太郎
Category：薬学	Category：伝統医学	漢方医学

本研究では，Wikipedia のカテゴリネットワークから特徴量を抽出し，SVM と回帰分析により，意外性のある情報の抽出を行う．

本研究で使用した特徴量は，下記の通りである．

- カテゴリ A，B の子項目数
- カテゴリ A，B の階層
- カテゴリ A，B の共通子項目数
- カテゴリ A，B の距離

教師データは，正例 200 件，負例 200 件を収集した．教師データの分布を，表２，３に示す．

表２．正例の分布

Feature	Mean	Standard Diviation
カテゴリ A の子項目数	280.465	660.16
カテゴリ A の階層	3.895	1.20
カテゴリ B の子項目数	119.18	262.21
カテゴリ A の階層	4.31	1.24
カテゴリ A，B の共通子項目数	1.14	0.82
カテゴリ A，B の距離	4.79	1.34

表３．負例の分布

Feature	Mean	Standard Diviation
カテゴリ A の子項目数	1126.30	1952.08
カテゴリ A の階層	5.60	1.45
カテゴリ B の子項目数	649.29	1682.50
カテゴリ A の階層	5.67	1.36
カテゴリ A，B の共通子項目数	150.42	373.93
カテゴリ A，B の距離	3.83	1.65

[1] http://download.wikimedia.org/

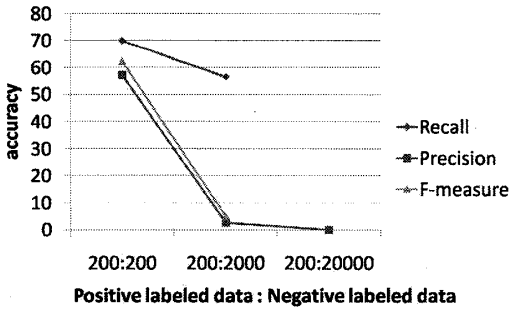
SVMによる意外性情報の発見

表３に，SVMによる判別結果を示す．２次と３次の多項式カーネルを用いた際の精度が最も高い．しかしながら，負例を10倍，100倍に増やして実験を行ったところ，図１のように，精度が著しく低下した．意外性のある情報は，全体のほんの一部分であるため，正例の教師データを収集することが困難であり，教師データを大量に得ることができないこのような問題に対しては，SVMのような判別手法は適切ではない．

表３．Results by SVM

Kernel	Recall	Precision	F-measure
Linear kernel	72.88%	54.43%	62.31%
Polynomial kernel (2)	69.84%	57.14%	**62.86%**
Polynomial kernel (3)	73.33%	55.00%	**62.86%**
Gaussian kernel	76.36%	51.85%	61.76%

図１．負例を増加させた際のSVMの精度推移



そこで我々は，回帰により，意外性の順位付けを行うことにした．図２に示すように，意外性のある情報には１を，意外性のない情報には０のラベルをつけ，これを目的変数として使用し，線形回帰分析及び，２次と３次の非線形回帰分析を行い，各データにあてはめた際の予測値を意外性のスコアとした．実験の結果は，表４の通りである．

図２．Regressionによる意外性情報の発見

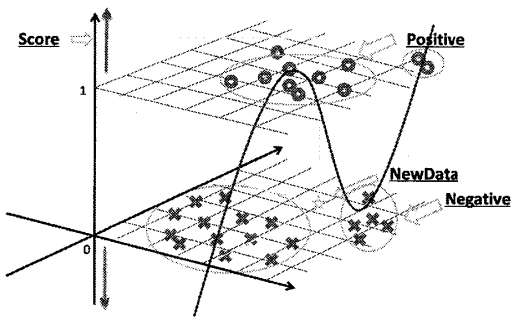


表４．Average precision

positive : negative	200:200	200:2000	200:20000
Category:マイケル・ジャクソン	0.190	0.239	0.236
Category: 地球温暖化	0.420	0.373	0.380

回帰分析に対しても，負例を増やした実験を行ったところ，SVMと比較して回帰による手法は精度が安定しており，この問題に対して妥当な手法であるといえる．すべてのデータを評価することは困難なので，ある特定のカテゴリが含まれる関係性の中で，予測値による順位付けを行った．表５，６は上位の正解例である．

表５．「Category:マイケル・ジャクソン」が含まれる関係性の中で，順位が高い正解例

CategoryA	CategoryB	Article	Rank
LGBTの人物	マイケル・ジャクソン	リサ・マリー・プレスリー	3
マイケル・ジャクソン	著名な動物	バブルス	10

表６．「Category:地球温暖化」が含まれる関係性の中で，順位が高い正解例

CategoryA	CategoryB	Article	Rank
都市伝説	地球温暖化	ワールド・ジャンプ・デー	1
地球温暖化	ドキュメンタリー番組	地球温暖化詐欺（映画）	9

まとめ

本研究では，Wikipediaのカテゴリ構造から特徴量を抽出し，それを教師データとして意外性のある情報を発見する手法を提案し，SVMと回帰分析による手法を提案した．しかしながら判別手法はこの問題には適切ではなく，回帰による手法が比較的安定性の高い手法であった．

参考文献

[1] Swearingen, K. and Sinha, R.: Beyond algorithms: An HCI perspective on recommender systems, ACM SIGIR 2001Workshop on Recommender Systems, 2001

[2] Murakami, T. and Mori, K. and Orihara, R.: A Method to Enhance Serendipity in Recommendation and its Evaluation, Transactions of the Japanese Society for Artificial Intelligence, vol.24, pp.428-436, 2009

[3] Nadamoto, A. and Aramaki, E. and Abekawa, T. and Murakami, Y.:Content hole search in community-type content, Proceedings of the 18th international conference on World wide web, pp.1223- 1224, 2009