2006－DBS－139 (5)
2006－ＦＩ－ 83 (5)
2006／5／30

ブログマイニング

奥村 学
(東京工業大学 精密工学研究所)

blogWatcher

Agenda

- blogとは?
- blogマイニング: blogから社会の何が見えてくるのか?
- いくつかのblogマイニング技術
- blogWatcher: インターネットから社会の関心，意見を収集・分析する

blogWatcher

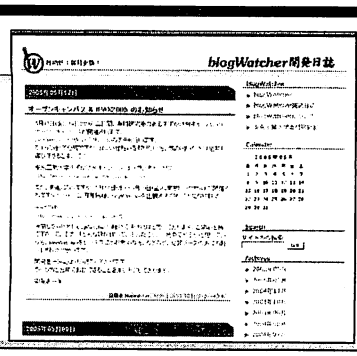
blogとは?

blogWatcher

ブログの歴史

- アメリカでは，他サイトをリンクし，それに簡潔なコメントをつけて紹介するフィルタサイトが最初
  - その後，その作者達が同じようなサイトを意識することで一気にコミュニティが形成された
  - 1999年のblogger登場を契機として，ブログ作成用の高機能なツール，サービスが普及し，それに伴う形でブログコミュニティは拡大，多様化している
- 「ニュージャーナリズムの新しい波」!?
  - アメリカでは9.11を契機に飛躍的発展

blogWatcher



blogWatcher

ブログのサーチエンジン

- ブログ検索エンジン≒RSS検索エンジン
- アメリカでは，2000年から2001年にかけて，日本では，2003年に登場
- 現在では，国内外問わず，ブログの検索エンジンは，数え切れないくらい増えてきている
  - 基本的には，pingサーバベースで収集しているので，持っているコンテンツにはそれほど違いはない
  - ただ検索サービスを提供するだけではなく，独自の分析，サービスをそれぞれが展開

blogWatcher

従来のサーチエンジンとの違い (1)

- Webを大規模にクローリングする必要がない
  - 更新を通知してくるブログに関しては，そのタイミングでクローリングするだけでよい
  - 個人，研究室などでも十分運用可能
- World Wide WebからWorld Live Webへ
  - Technorati (http://www.technorati.com/)
  - よりリアルタイム性のある検索へ
  - 検索対象になるまで，更新通知を送ってから数分以内

blogWatcher

従来のサーチエンジンとの違い (2)

- 時系列データとしてのブログ
  - コンテンツの書かれた日付がメタデータとして利用可能 → マイニングの対象に
  - ランキングも「最新のものから順に」が多い
- 検索エンジンからから分析エンジンへ
  - BBSなどと同様，ブログは個人と密着したメディア
  - Trackbackによるインタラクティブなつながり

blogWatcher

blogマイニング：blogから社会の何が見えてくるのか?

blogWatcher

(ビジネス的?)背景

- 一般の人々からの情報発信が盛んに
- 速報性のある新鮮な大量の情報を有効に活用したい
- 現在注目されている情報源 (CGM; Consumer Generated Media)
  - ✓ 掲示板 (BBS)
  - ✓ blog (Weblog)
  - ✓ chat
  - ✓ …

blogWatcher

blogマイニングの要素技術(1)

- Authority分析(被リンク数によるランキング)
- トレンド分析
  「いつどんな話題が盛り上がっているのか?」
- 評判分析
  「日吉でおいしいラーメン屋さんは?」
- コミュニティ抽出

blogWatcher

blogマイニングの要素技術(2)

- blogの書き手の属性推定
  「若い女性に人気のお店は?」
- 実世界の動向との相関分析
- トピック分類
- Spam filtering
- 自動要約
- 情報の重要性，信頼性評価

blogWatcher

AAAI SS on CAAW 2006

- 48件中34件が関連研究
- 内訳:
  - 評判分析(13件)
  - 書き手の属性推定(6件)
  - トピック分類(3件)
  - 自動要約(3件)
  - ...

blogWatcher

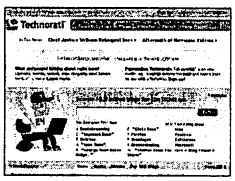
blogにおけるランキングあれこれ

- Entry
  いわゆるblog検索におけるランキングアルゴリズム(relevance, 日付順, ...)
- 人(blog)
  Authority分析
- 参照対象(本, ニュース, ...)
  Authority分析, 評判分析

blogWatcher

Technorati

- URL
  http://www.technorati.com/
- 運営者
  Technorati, Inc.
- サービス開始
  2002/11
- index
  1700万のブログ
  15億のlinkを収集



blogWatcher

Technorati

- リンク解析
  - あるURLにリンクしているブログを検索
- リンクベースのランキング(Authority分析)
  - News, Books, Movies, Blogs
    - ニュースサイトやamazonなどにリンクしているブログをカウント, 関連づけて表示
  - Top 100 Technorati
- microformats (http://microformats.org/)
  - Tags
    - 著者がエントリにTagを埋め込むことによって, カテゴライズ
  - hReview
    - 機械可読なレビューを書くためのフォーマット

blogWatcher

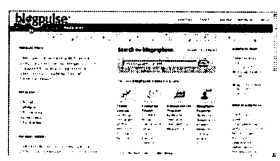
Technorati

- Indexingが非常に高速
  - pingを送ってから数分以内にindexing
- APIを広く公開
  - 様々なアプリケーションがユーザによって開発
- 前回の大統領選挙特集
  - Liberal, Uncategorized, Conservativeなどのカテゴリ分け
- Technorati.jpスタート
  - 2005年7月正式リリース
  - 先日の衆議院総選挙特集
  - Technorati mobile 2005/9/16 リリース

blogWatcher

BlogPulse

- URL
  http://www.blogpulse.com/
- 運営者
  Intelliseek, Inc.
- サービス開始
  2004/05/11
- index
  1600万のブログ



blogWatcher

BlogPulse

- Trend Search(トレンド分析)
  - 頻度に基づくグラフを生成
  - Featured Trends
    - News, Sports, Businessなどのカテゴリ毎のトレンドを紹介
- Conversation Tracker
  - 話題の伝播をトラッキング，視覚化
  - Blog Epidemic Analyzer – HP lab.
- BlogPulse Profiles
  - あるブログの言及数，更新頻度，情報源，類似ブログなどを検索

blogWatcher

いくつかのblogマイニング技術

blogWatcher

blogの書き手の属性推定

- 性別，年齢，居住地域，...
- 性別，年齢
  - CAAWに多数
  - [池田，他，2006]
- 居住地域
  - [安田，他，2006]

blogWatcher

blogにおけるコミュニティ抽出

- blog間のリンク構造を元にコミュニティを抽出
- Webにおけるコミュニティ抽出手法の適用
  - HubとAuthority
  - PageRank, HITS, ...
  - Clusteringアルゴリズム

blogWatcher

blogのトピック分類[平野，他，2005]

- Naïve Bayes法による多重トピック分類
- 91クラス
- 名詞，動詞，形容詞のみ利用
- 7-8割が妥当な分類

blogWatcher

Spam Filtering

- Spam対策は必須
  - Spam blog, Ping Spam, Trackback Spam
    - アフィリエイト関連のblogを大量に自動生成
    - 検索結果にも大量に含まれる
    - 分析時に影響
- [Kolari et al., 2006]
  - SVM
  - 素性として，bag-of-words, bag-of, anchors, bag-of-urls
  - F値0.881

blogWatcher

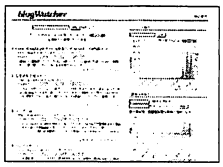
相関分析

- [Gruhl et al., 2005]
- 本の売り上げランクとblogの言及数の推移の相関
- WebFountainのデータを利用
- 立ち上がりのある売り上げランクの場合，十分な数の言及数があれば，言及の立ち上がりも精度良く相関する
- 言及数の急な増加は，売り上げランクの立ち上がりを予測できる!?

blogWatcher: インターネットから社会の関心，意見を収集・分析する

blogWatcher

- IPA 平成15年度 未踏ソフトウェア創造事業
- ブログやWeb日記を収集して，マイニング
  - HTML文書を構造解析，RSSを使用しない収集
  - バースト検出
  - 評判情報検索
  - メタブログ
  - ニュースとブログの対応づけ
  - なんでもRSS!



2004/08/16に一般公開
2005/05/09にVer.2.0を公開
http://blogwatcher.pi.titech.ac.jp/

モチベーション (1)

- 現在，「ブログ」というと，
  - Movable Typeなどの「ブログツール」やまた，これらのホスティングサービスを利用しているサイト
  - Trackbackなど，ブログツール特有の機能を有しているサイト

  を指すことが多い．
- しかしながら，日本では...
  - ブログツール登場以前から，「Web日記」が盛んであった
  - ブログとは，内容的にも形式的にも，ほとんど区別の必要はない．

モチベーション (2)

- このようなブログも含めて網羅的に収集したい
  - ブログツールによって書かれたブログだけなら...
    - RSS Feed
    - pingサーバ

    を利用することにより，容易に収集ができる
  - 「Web日記」のような特定のツールやhostingサービスを利用しないものも網羅的に収集したい
    - Webページの一部として記述されるものも多い

モチベーション (3)

- さらに，ブログをマイニングしたい
  - マイニングの対象として，非常に魅力的
    - 人々の生の声が大量に含まれる
    - 記事の書かれた日付が付いているため，時系列マイニングができる

blogWatcherの特徴 (1)

- 対象は，広義のブログ
  - Web日記，テキスト系サイトを含む
  - メタデータに依存しない，HTML文書の解析に基づいた手法でブログ判定
- バースト分析(トレンド分析)
  - キーワードがいつ，どの程度盛り上がっていたかを自動分析
- 評判情報検索(評判分析)
  - キーワードに関連する評判情報を抽出
  - ポジティブとネガティブに自動分類

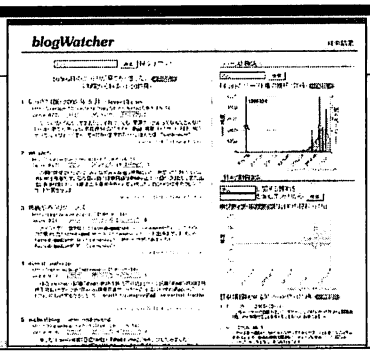
blogWatcherの特徴 (2)

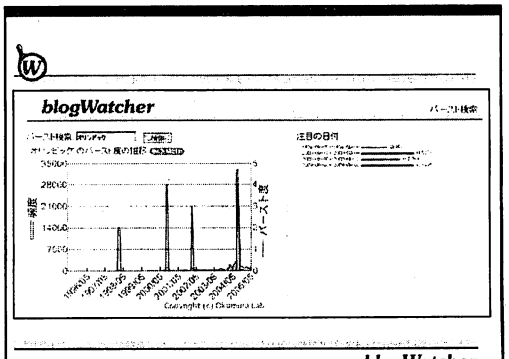
- ニュースとブログのマッピング
  - あるニュースに関連するブログ，あるブログに関連するニュースを検索可能に
- メタブログ
  - 今話題のトピックを紹介
  - 完全自動生成されるブログとして紹介
- API公開
  - 通常検索，バースト，評判検索など，各種APIを公開
- なんでも RSS
  - blogWatcherの収集エンジンを利用したRSS Feed自動生成ツール

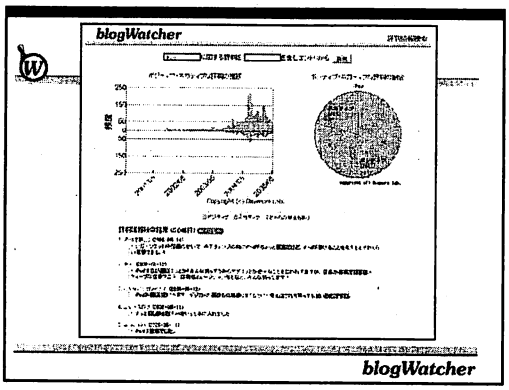
blogの定義

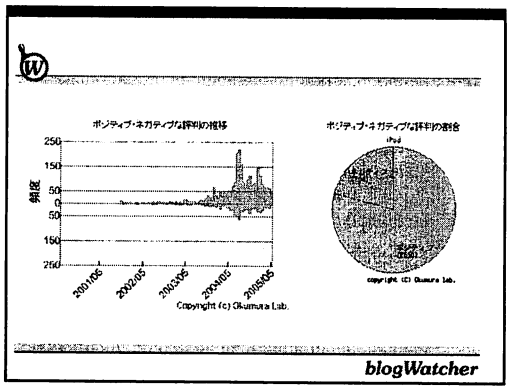
- 我々は，
  - 不特定多数でない個人(あるいは団体)が
  - 関心を持ったニュースやできごとについて書いたもので，何らかのコメントを含み，
  - 日付表現を伴い，時系列に沿って掲載されており，
  - ある程度の頻度で更新されるWebページ

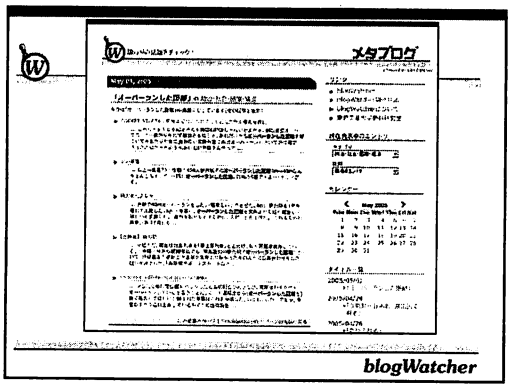
をblogと考える











キーワードのburst

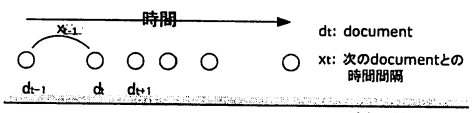
- ある単語の出現頻度が急激に上がることがある
  - ワールドカップ中は「サッカー」「日本」「代表」など
- そのような語を収集するために、単純に頻度を計算するのでは問題がある
  - 助詞などのいわゆる「ストップワード」が大量にとれてしまう
- そこで出現間隔に基づいて盛り上がり(burst)度を計算する手法である[Kleinberg, 2002]を拡張して用いた

blogWatcher

document streamとは

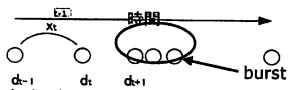
- Kleinbergの手法ではdocument streamを扱う
  - 掲示板への連続する書き込み
  - 連続する一連の記事

  などのように，ある間隔で到着するdocument列(情報ストリームなどとも呼ばれる)
- ここでは，以下のようなモデルとして考える



blogWatcher

document streamにおけるburst



- burstとは，document間の間隔が他と比べて短くなっている部分
- ある単語を含むdocumentのみを集めて作られたdocument streamでburstが存在する場合，burst部分ではその単語に関連するイベントが発生していると考えることができる
- Kleinbergの手法とは，このようなburst部分を検出し，その度合い(burstの大きさ)を得るための手法

blogWatcher

Kleinbergの手法について

- 以下の点で他の手法に比べ優れているとされる
  - 一時的に間隔が長くても，一続きのburstと判定する
  - 通常の出現間隔と比べてどう変化したかに基づくため，ストップワード処理をする必要がない
  - 比較的計算量が少ない
- しかし，この手法は平常状態におけるdocument出現間隔が一定であることを仮定しているため，出現間隔に偏りのあるdocument streamを対象にする場合は問題がある

対象とするデータの性質

- 電子掲示板では朝は書き込み数が少なく，夜に書き込みが集中するという性質がある
- 収集されたblogにはインターネット普及率などに起因するblog総数の増加が存在する
- そこで，このような対象にも適用できるように手法を拡張する

電子掲示板への書き込み数

収集されたblog数

ホットキーワード(メタブログ)

- 収集したblogの中から，注目すべき話題をユーザに提示したい
- あらかじめblog中に出現する単語に対し，日ごとのburst度を計算し，ランキングを作成する
- システム側ではこの計算結果を利用することで，「今日(今月)のホットキーワードリスト」として表示することができるようになる

評価情報の要素組の抽出

- Chasenによる形態素解析，
- Cabochaによる係り受け解析
- 対象-属性-評価表現候補の3つ組を抽出
- 3つ組分類(positive/negative/neutral)

評価情報の要素組の分類

- Semi-Supervisedな学習手法を用いる
- 少量の種となる3つ組を人手で用意
- ブートストラップ的に学習
- 3つ組の周辺の文脈情報も考慮
- Naïve Bayes + EMアルゴリズム + SVM
- 実験的には
  - 3つ組数 約57万
  - 約80%の精度

Positive or Negative or Neutral?

- 形容詞・形容動詞に注目
  - おいしい：Positive ☺
  - まずい：Negative ☹
  - 大きい：？
- 対象・属性も考慮
  - 「あのHDDは容量が大きい。」→ ☺
  - 「あのHDDは動作音が大きい。」→ ☹
    - [対象]＝HDD、[属性]＝動作音/容量、[評価]＝大きい

この三つ組があれば大抵は判定できる

3つ組辞書の効率的な作成

- 3つ組辞書の例
  - HDD / 容量 / 大きい →Positive ☺
  - HDD / 動作音 / 大きい → Negative ⊗
  - 部屋 / - / 暗い → Neutral ☺
  - 彼 / 性格 / 暗い → Negative ⊗

・人手で作成するのは大変なので、自動的に！
・Semi-supervisedな機械学習手法を利用

blogWatcher

評判分析モジュール

- 評価語発見
  - Blogエントリ(テキスト)中から評価を表す可能性のある語を発見
- 対象・属性同定
  - 表現が評価しているモノとその属性をテキストから探索
- 評価極性判定
  - <対象, 属性, 評価語>の3つ組をpositive/negative/両方ありうる/評価でないに分類

blogWatcher

評判分析モジュール: 例

- 評価語発見
  - テキスト中から一定の基準で評価語候補を発見
- 対象・属性同定
  - "鈍い"が指す特徴(属性)何の評判なのか(対象)言語処理の技術で同定
- 評価極性判定
  - 辞書・周辺文脈・統計情報を利用し判定
  - 対象"パソコン"にnegativeな評価があることがわかる

パソコンの 反応が 鈍い (評価語)

パソコン(対象)の 反応(属性)が 鈍い (評価語)

パソコン(対象)の 反応(属性)が [■] (評価語)

blogWatcher

評価語発見

- 現在扱っている評価語
  - 形容詞(終止形/連体形など、特定の活用形)
    - 「おいしい」,「まずい」,「うれしい」,「厚い」, etc...
  - 形容動詞語幹
    - 「健康的だ」,「難解だ」 etc...

blogWatcher

対象・属性同定(1/3)

- 3つ組
  - 評価極性を判定する基本単位
  - <対象, 属性, 評価> [鈴木 04]
  - スープの味が濃い→<スープ, 味, 濃い>
- 属性抽出: 属性名詞辞書
  - "属性"に入る単語を限定

blogWatcher

対象・属性同定(2/3)

- 対象抽出
  - 評価語との係り受け関係
    - りんごの色が赤い → <りんご, 色, 赤い>
    - 色が赤いりんご → <りんご, 色, 赤い>
    - 甘くておいしいりんご
      → <りんご, , 甘い>
      → <りんご, , おいしい>

blogWatcher

対象・属性同定(3/3)

- 対象抽出(つづき)
  - 以前の文から持ってくる
    - "センタリング理論"に基づくルールを独自に作成
    - 例:
      「先日、りんごが実家から届きました。
      甘くておいしいです。」
      → <りんご, , 甘い>, <りんご, , おいしい>

blogWatcher

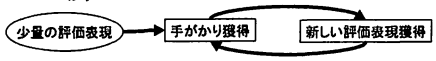
評価情報の3つ組の分類

- 機械学習を用いて分類
  - 通常は人手により正解を与え、学習を行う
  - 本研究では、正解付けの手間を軽減するために、正解が与えられていないデータも学習に取り入れて、性能を向上させる

blogWatcher

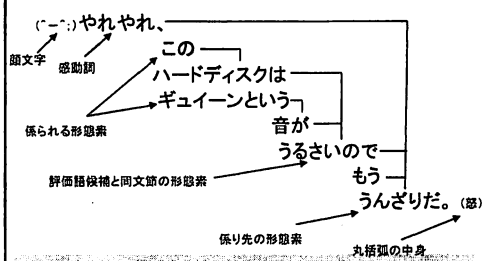
3つ組分類の手がかり

- 仮定:評価情報の3つ組は特徴的な周辺情報を伴なうことが多い
  - 例:顔文字「(^ー^)」が文末にある
    →文中の形容詞は大抵 Positive☺
- 周辺情報と3つ組を繰り返し獲得
  - 「Aなので良い」→評価表現A:☺
  - 評価表現A:☺ +「AなのでB」→周辺情報「B」:Positive☺の手がかり



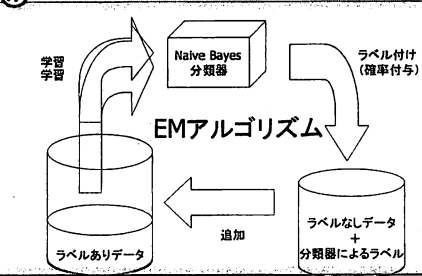
blogWatcher

素性として用いる周辺情報の例



blogWatcher

3つ組分類の繰り返し学習の流れ



blogWatcher

SVMとFisher Kernel

- SVM(Support Vector Machine)
  - Supervisedな機械学習手法
  - 高い分類性能
- Fisher KernelでNaïve Bayes+EMにより生成されたモデルを用いることにより、SVMにラベルなしデータの影響を取り入れることができる

blogWatcher

分類結果を用いた3つ組辞書の作成

- 同じ3つ組でも文脈によって評価極性が異なるものもある
  - →全ての分類結果を利用したのでは、矛盾が起こる可能性がある
- 方針
  - 確信度が高い事例の分類結果を利用
  - ある評価極性で偏って使われている3つ組を採用

blogWatcher

評価極性判定(1/3)

- 3つの分類器・辞書を以下の優先順で利用
  1. 評価語のみの基本辞書による判定
  2. 柔軟な自動分類器による判定
  3. 対象・属性も考慮した辞書による判定
- 否定「遅くない」, 伝聞「おいしいらしい」などへ対応
- 判定結果
  - Positive, negative, neutral, 両方ありうる

blogWatcher

評価極性判定(2/3)

1. 評価語のみの辞書 (1250組)
   - <*, *, おいしい> → positive
   - <*, *, 悲惨> → negative
3. 対象・属性も考慮した辞書 (414335組)
   - <スープ, 深み, ある> → positive
   - <パソコン, 反応,遅い> → negative
   - <ラーメン, *,ぬるい> → negative

blogWatcher

評価極性判定(3/3)

2. 自動分類器
- 周辺情報を利用した機械学習
  - 例:「先日、実家からりんごが届きました。甘くておいしい！(^_^)」
    - 3つ組に係り受け関係のある評価語
    - 文末語・文末記号
    - 同一文中の感動詞 ex.「おおっ」etc...
    - 同一文中の顔文字
    - 3つ組に続く"(笑)"などのような丸括弧の中身
    - 評価語と同じ文節の形態素

blogWatcher

blogとニュースの自動対応付け(1/2)

- (リンクされていなくても)blog中で言及されている記事内容へ
- ニュースに対するblogでの反応は?

blogWatcher

blogとニュースの自動対応付け(2/2)

- 基本は, VSMによる類似度判定
- blog, ニュースでは, 書かれている内容も, 使われている語彙も異なる
- blog, ニュース記事のベクトルの作成方法に工夫
- ニュース記事中の重要な単語は, blog中で言及されることにより, 出現頻度が大きく変動する!?
- 実験的には, 再現率約71%, 精度約89%

blogWatcher

次期ヴァージョンの新しい機能

- 行動分析[野呂, 他, 2006]
  - blogの中の「行動」の抽出
  - 「行動」の時間帯推定(朝, 昼, 夕方, 夜)
  - 「ヨーグルトを食べるのは朝が多い?」
- blogの性別推定[池田, 他, 2006]

blogWatcher

- http://www.lr.pi.titech.ac.jp/blogwatcherをご参照下さい.
- 次期ヴァージョン5月中旬公開予定. お楽しみに.

blogWatcherの使用例

- ブログウォッチャーエンタープライズ(ホットリンク)
  http://www.hottolink.co.jp/service/package/blogwatcher_enterprise/
- 評判.info(blog Watcher Consortium)
  http://www.hyoban.info/
- Yahoo!ブログ検索
  http://blog-search.yahoo.co.jp/

参考文献

[平野, 他, 2005]
平野, 古林, 高橋, 日本語圏ブログの自動分類. 情報処理学会第170回自然言語処理研究会. 2005.
[Kolari et al. 2006]
Kolari, Finin, Joshi, SVMs for the Blogosphere: Blog Identification and Splog Detection, AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs 2006.
[Gruhl, et al., 2005]
Gruhl, Guha, Kumar, Novak, Tomkins, The Predictive Power of Online Chatter, KDD, 2005.
[野呂, 他, 2006]
野呂, 乾, 高村, 奥村, イベントの生起時間帯判定, 言語処理学会第12回年次大会. 2006.
[池田, 他, 2006]
池田, 南野, 奥村, blogの若者の性別推定, 言語処理学会第12回年次大会. 2006.

参考文献(blog)

- 1st-3rd WWW Workshop on Weblogging Ecosystem(2004-2006)
- 人工知能学会第6回SWO研究会 2004
- AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs 2006