

ブログマイニングによるマーケットシェア推定 —ブログと既存統計指標の関係の分析—

長谷川 真吾[†] 藤村 考^{†‡}

[†]電気通信大学情報システム学研究科 〒182-8585 東京都調布市調布ヶ丘 1-5-1

[‡]NTT サイバーソリューション研究所 〒239-0847 神奈川県横須賀市光の丘 1-1

E-mail: [†]hasegawa@ohta.is.uec.ac.jp, [‡]fujimura.ko@lab.ntt.co.jp

あらまし ブログ空間での言及数は、現実世界における人間の活動を反映するものであり、分野によってはマーケットシェアを推定する有益な指標となるものと考えられる。本稿では、テレビ視聴率や商品の売り上げデータなどの既存の統計指標とブログから抽出される言及数との相関関係を分析することにより、ブログマイニングによるマーケットシェア推定がどのような分野で有効かを調査した。

キーワード ブログマイニング、視聴率、マーケットシェア、感情表現

Estimating Market Share by Blog Mining —Analysis of the relationship between blog and existing statistical indicator—

Shingo HASEGAWA[†] Ko FUJIMURA^{†‡}

[†]The University of Electro-Communications 1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585 Japan

[‡]NTT Cyber Solutions Laboratories 1-1 hikarinooka, Yokosuka-shi, Kanagawa, 239-0847 Japan

E-mail: [†]hasegawa@ohta.is.uec.ac.jp, [‡]fujimura.ko@lab.ntt.co.jp

Abstract The number of mentions in blogosphere reflects activities in the real world. It is thus considered as a useful indicator of market share in some industry segments. To clarify the area in which blog mining is effective, this paper reports an investigation of the correlation between the number of mentions in the blogs and existing statistical indicators such as TV audience rating or CD sales ranking.

Keyword blog mining, audience rate, market share, sentiment analysis

1. はじめに

生活水準の向上や余暇時間の増大などにより、人々の価値観の多様化が進んでおり、以前にも増して多様な分野で消費者動向を分析するニーズが高まっている。一方、マーケットシェアを示す既存の統計指標は視聴率等の一部の分野に限られている。そこで、幅広い分野のマーケットに対応するため消費者一人ひとりのマーケティング情報抽出が可能であるブログマイニングに期待が集まっており、それをマーケティングに生かそうというサービスが多数出現している。例えば、kizasi.jp[1]、BuzzPulse[2]、Dentsu Buzz Research[3]などである。また、評判分析などの新たなブログ検索機能、例えば、Yahoo!ブログ検索[4]、goo-ブログの詳細検索[5]なども提供されている。しかしながら、これらの有用性の前提となる言及数と既存の統計指標との関係性に関する報告は少ない。特定のマーケットにおいて既存の統計指標と比較した先行研究としては、プログラムの感情と映画の興行収益との関係の研究[6]やブ

ログの言及数と Amazon における本の売り上げとの関係の研究[7]など、限られている。

本稿では、このような背景からブログマイニングにより抽出される様々な統計データを分析し、マーケットシェア推定に有用な指標を明らかにすることを目的とする。今回は特にブログマイニングにより抽出されたテレビ番組情報の統計データとテレビ視聴率との関連性を分析する。また、テレビ視聴率以外のマーケットシェアデータ（東京都知事選・音楽 CD 売り上げ）に関しても相関関係を分析し、ブログマイニングによるマーケットシェア推定がどの分野において有用かを調査する。

尚、本稿では、同じ分野における商品・サービスの利用者の相対比をマーケットシェアと定義するものとし、必ずしも商品の売り上げには限定しない。

以下、2 章ではテレビ視聴率について、3 章では東京都知事選について、4 章では音楽 CD 売り上げについて、5 章でまとめ、6 章で今後の課題について述べる。

2. 調査 1 テレビ視聴率における調査

2.1. 調査概要

(調査期間)

2007/2/4～2007/4/20

(調査対象)

2007/1～2007/4の21～24時に放映されたドラマを対象とする。多くのテレビ番組の中からドラマを選択した理由は、放映されるドラマのほとんどが同じ時期に始まり放映される時間帯もほぼ同じといった同一の条件で調査ができるからである。

(調査方法)

それぞれのドラマについて各ドラマのタイトルをクエリーとしたブログ検索を行い、そこで得られた言及数とテレビ視聴率（ビデオリサーチ調べ）との間の相関関係を調べる。

ブログ検索のために、NTTサイバーソリューション研究所が goo lab にて公開しているブログサーチエンジン BLOGRANGER API (<http://ranger.labs.goo.ne.jp/>) を利用し、ブログ上から特定のキーワードを抽出するシステムを作成した。(図 1)

TV視聴率調査ツール
(出力:日時,Blog名,記事URL,本文)
キーワード 時間間隔
期間:2007年 月 日 時 分 ～ 2007年 月 日 時



～ログ～
～クローリング結果～

図 1.キーワード抽出システム

2.2. 調査仮説

ある特定のドラマのタイトルがブログに多く書き込まれるということは、そのドラマがブロガーたちの間で多く関心を持たれているということを表している。したがって、その関心がドラマを見るという行為につながれば、テレビ視聴率が高くなる可能性がある。

また、ブロガーは関心のあるドラマに関する記事を必ずしも当日に書くとは限らない。更には、ドラマ放映日から長い日数がたっている場合、そのドラマのタイトルが含まれている記事であっても、他のドラマへの感想の引き合いに出される可能性があるなど、必ずしも関心の度合いを表すとは言えないかもしれない。

したがって、最適なデータ収集時期・期間を持つ可能性がある。

また、ドラマのタイトルの他に「面白い」や「楽しい」などのポジティブな感情表現が記事に含まれている場合、ただ単純にドラマのタイトルだけを書いている場合に比べ、強い関心を示している可能性がある。

以上を踏まえ、調査に際して以下の仮説について検証を行った。

(仮説 1) 多くの人がブログで言及する番組は視聴率が高い

(仮説 2) 番組放映当日だけのデータよりも、放映後数日間のデータを集めた方がより視聴率を反映する

(仮説 3) 「面白い」等のポジティブな感情表現を含むものだけを抽出した方がより視聴率を反映する

2.3. 調査結果

仮説 1

多くの人がブログで言及する番組は視聴率が高い、という仮説を検証するため、テレビ視聴率と言及数(テレビ放映後 2 日間の累積数)についてまとめたグラフを図 2 に示す。

このときテレビ視聴率と言及数との間の相関係数は 0.895 であり、99%の有意水準で相関があった。このことにより、ブログの言及数の増加と視聴率の増加には関連性があると言える。

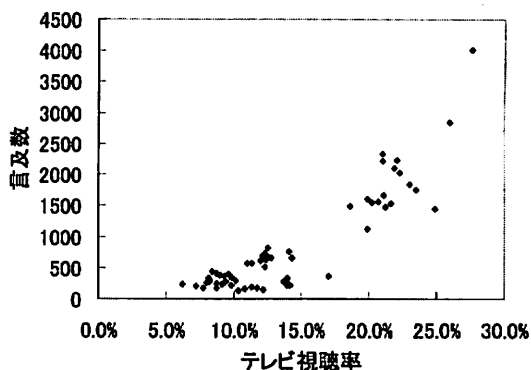


図 2.テレビ視聴率と言及数との関係

仮説 2

番組放映当日だけのデータよりも、放映後数日間のデータを集めた方がより視聴率を反映する、という仮説を検証するためドラマ放映後の経過日数と、テレビ視聴率と言及数との相関係数についてまとめたグラフを図 3 に示す。

このとき最も相関係数が高くなったのは放映 6 日後

まで抽出した場合である。相関係数の差の検定を行ったところ有意にはならなかったが、日数が増えるにつれて相関係数が増加していく傾向にある。

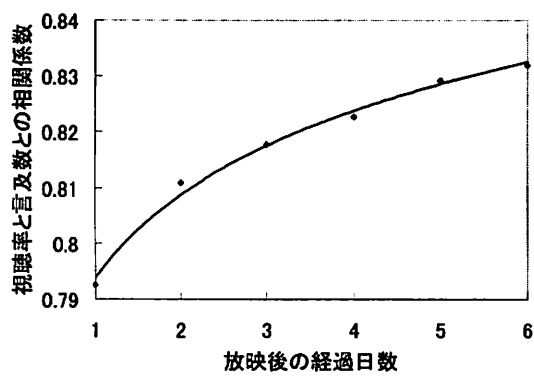


図 3.放映後の経過日数と相関係数との間の関係

仮説 3

「面白い」等のポジティブな感情表現を含むものだけを抽出した方がより視聴率を反映する、という仮説を検証するため、既存のクエリー（ドラマのタイトル）と「面白い」というキーワードを AND でつないだ新たなクエリーを作りデータ抽出をした。感情表現（「面白い」）を含まない場合と含む場合のテレビ視聴率と言及数（テレビ放映後 6 日間）についてまとめたグラフを図 4、図 5 に示す。また、感情表現を含まない場合と含む場合についてドラマ放映後の経過日数と、視聴率と言及数との相関係数についてまとめたグラフを図 6 に示す。

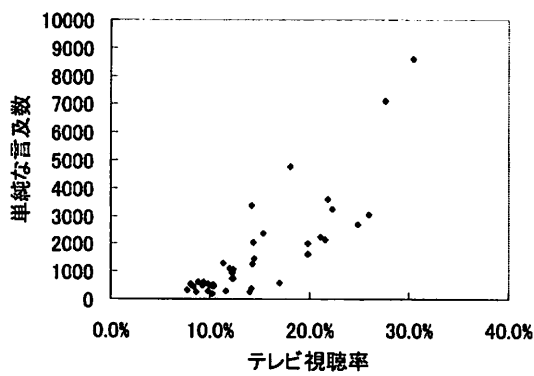


図 4.テレビ視聴率と単純な言及数との関係

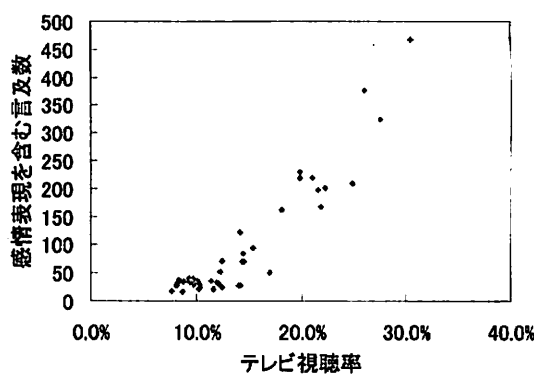


図 5.テレビ視聴率と感情表現を含む言及数との関係

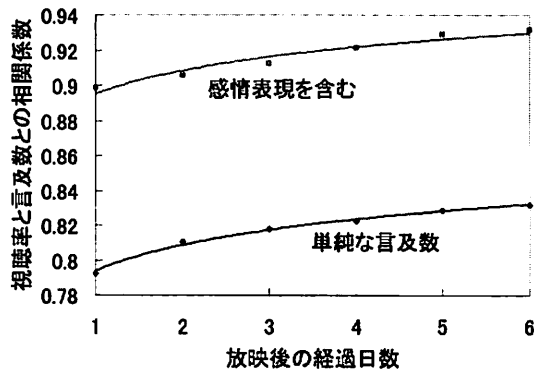


図 6.感情表現の有無による相関係数の差

1 ～ 6 日間までの全ての経過日数において、感情表現を含むクエリーを用いて抽出した場合の方が高い相関係数を示した。5、6 日目においては、感情表現を含む場合と含まない場合の相関係数の差の検定において、95%の有意水準で採択された。今回は 1～4 日目までは 95%の有意水準で採択されなかったが、これは抽出データ数を増やすことによって改善できると考えられる。経過日数と信頼区間・差の検定の結果を表 1 に示す。

経過日	単純な言及数		感情表現を含む		有意さ
	下限	上限	下限	上限	
1 日	0.638	0.885	0.814	0.945	n.s.
2 日	0.668	0.896	0.827	0.949	n.s.
3 日	0.679	0.900	0.840	0.953	n.s.
4 日	0.687	0.902	0.854	0.958	n.s.
5 日	0.698	0.906	0.869	0.962	*
6 日	0.702	0.908	0.873	0.964	*

* 95%有意

表 1.感情表現の有無による相関係数の差の検定

3. 調査 2 東京都知事選

3.1. 調査概要

(調査期間)

2007/3/4～2007/4/8

(調査対象)

2007/4/8 投票の東京都知事選に出馬した 5 名（石原・浅野・吉田・黒川・中松）

(調査方法)

それぞれの候補者について「都知事選 AND 候補者名」をクエリーとしたブログ検索を行い、そこで得られた言及数と 2007/4/8 に行われた選挙での得票数との間の相関関係を調べる。

ブログ検索のためにセクション 2.1 で述べた図 1 のシステムを使用した。

3.2. 調査仮説

ある特定の候補者の名前がブログに多く書き込まれるということは、その候補者がブロガーたちの間で多く関心を持たれているということを表している。関心の度合いが大きければそれが投票の意思へとつながる可能性がある。

また、投票日の一ヶ月以上前から候補者の開示はされており、ブロガーは関心のある候補者に関する記事を必ずしも当日あるいは前日に書くとは限らない。更には、投票日からかなり過去に遡っている場合、その候補者の名前が書かれた記事であっても、他の候補者に対する引き合いに出された可能性があるなど、必ずしも関心の度合いを表すとはいえないかもしれない。また、出馬の意思を示す時期はばらばらであるので、強い誤差要因が生じる日があるかもしれない。したがって、最適なデータ収集時期・期間を持つ可能性がある。

以上を踏まえ、調査に際して以下の仮説について検証を行った。

(仮説 1) 多くの人がブログで言及する候補者は得票数が高い

(仮説 2) 選挙前日だけのデータよりも、ある程度の日数を遡ってデータを蓄積した場合の方が、言及数と得票数の相関があがる

3.3. 調査結果

仮説 1

多くの人がブログで言及する候補者は得票数が高い、という仮説を検証するため、言及数（選挙の日より一週間前までの累積言及数）と得票数についてまと

めたグラフを図 7 に示す。

このとき 5 人の候補者の総言及数は 473 件、言及数と得票数との間の相関係数は 0.975 であり、99% の有意水準で相関があった。このことにより、ブログの言及数の増加と得票数の増加には関連性があると言える。

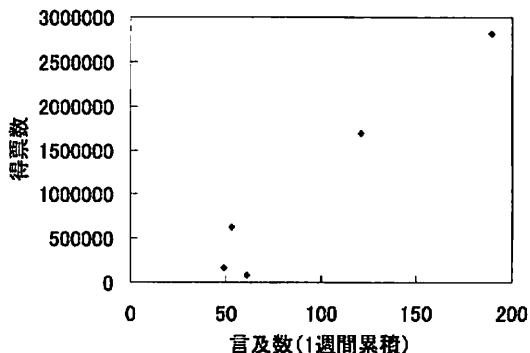


図 7.言及数と得票数との間の関係

仮説 2

選挙前日だけのデータよりもある程度の日数を遡ってデータを蓄積した場合の方が言及数と得票数の相関があがる、という仮説を検証するため、投票日から遡った日数と相関係数、投票日から遡った日数と累積言及数についてまとめたグラフを図 8、図 9 に示す。

このとき、累積言及数がほぼ線形に増えていることから、投票日直前に投稿が集中するという現象は起こっていないことがわかる。また、投票日から 21 日遡った付近より相関係数が大きく跳ね上がっているが、これはデータ数を 5 個しかとっていないために、一人の候補者への言及数の増減が大きく影響してしまったため、長い期間データ収集したほうが言及数と得票数の相関が上がるとは言い切れない。

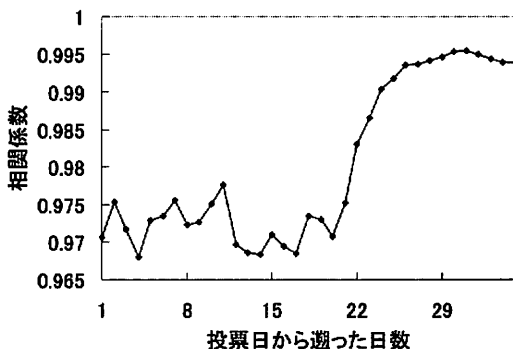


図 8.投票日から遡った日数と相関係数の関係

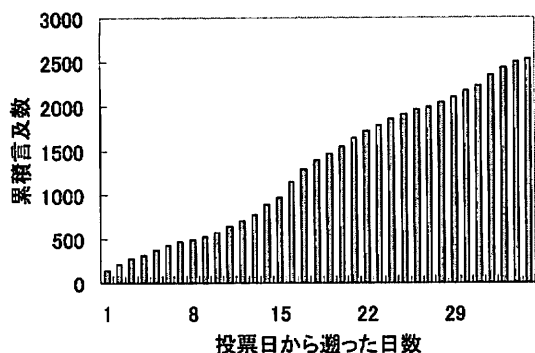


図 9.投票日から遡った日数と累積言及数

4. 調査 3 音楽 CD 売り上げ

4.1. 調査概要

(調査期間)

2007/4/8～2007/4/25

(調査対象)

2007/4/25 に発売されたシングル CD の中で、オリコンシングル週間ランキングが 30 位以内であった 14 作品

(調査方法)

それぞれのシングル CD について「CD タイトル」をクエリーとしたブログ検索を行い、そこで得られた言及数とオリコンによって発表された CD 売り上げ枚数との間の相関関係を調べる。

ブログ検索のためにセクション 2.1 で述べた図 1 のシステムを使用した。

4.2. 調査仮説

ある特定の CD の名前がブログに多く書き込まれるということは、その CD がブロガーたちの間で多く関心を持たれているということを表している。関心の度合いが大きければそれが CD 売り上げ増加へとつながる可能性がある。

また、発売日以前から CD のプロモーション活動は行われており、ブロガーは関心のあるシングル CD に関する記事を必ずしも発売日当日あるいは前日に書くとは限らない。更には、発売日からかなり過去に遡っている場合、その CD について書かれた記事であっても、プロモーション活動の行われるタイミングの影響を大きく受けるなど、必ずしも CD 発売日の関心の度合いを表すとはいえないかもしれない。したがって、最適なデータ収集時期・期間を持つ可能性がある。

よって、調査に際して以下の仮説について検証を行った。

(仮説 1) 多くの人がブログで言及するシングル CD

は売り上げ枚数が多い

(仮説 2) 発売日だけのデータよりも、ある程度の日数を遡ってデータを蓄積した場合の方が、言及数と売り上げ枚数の相関があがる

4.3. 調査結果

仮説 1

多くの人がブログで言及するシングル CD は売り上げ枚数が多い、という仮説を検証するため、言及数(発売日より一週間前までの累積言及数)と売り上げ枚数についてまとめたグラフを図 10 に示す。

このとき 14 枚の CD に関する総言及数は 3231 件、言及数と売り上げ枚数との間の相関係数は 0.764 であり、99%の有意水準で相関があった。このことにより、ブログの言及数の増加と売り上げ枚数の増加には関連性があると言える。

今回、CD タイトルをクエリーにしたため、各タイトルの記述のしやすさなどにより誤差要因が発生している可能性が考えられる。また、一般的に使われる単語がタイトルの場合、アーティスト名を付加して検索を行ったため、その点についても公平な抽出が出来ていない可能性が考えられる。

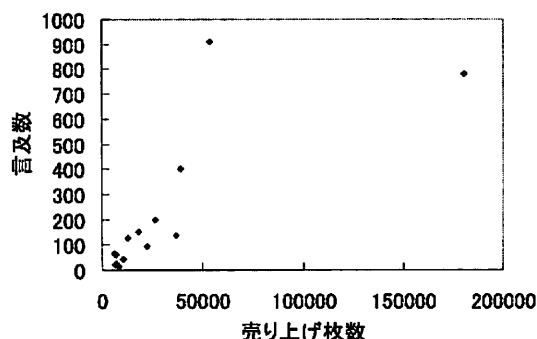


図 10.CD 売り上げ枚数と言及数との関係

仮説 2

発売日だけのデータよりも、ある程度の日数を遡ってデータを蓄積した場合の方が、言及数と売り上げ枚数の相関があがる、という仮説を検証するため、発売日から遡った日数と相関係数、発売日から遡った日数と累積言及数についてまとめたグラフを図 11、図 12 に示す。

このとき、累積言及数が対数的に増加していることから、発売日直前に言及数が増えていることがわかる。さらに、発売日当日の相関係数が高いことから、過去のデータを多く抽出しなくても発売日から数日間のデ

ータを抽出することにより売り上げ枚数を予測できる可能性がある。

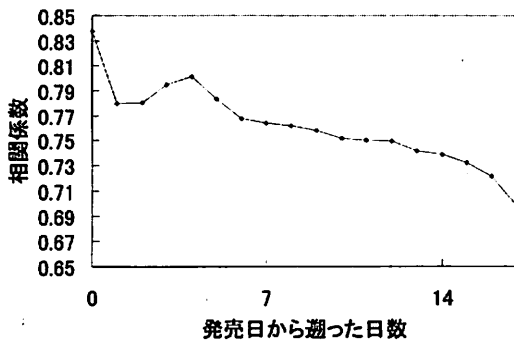


図 11. 発売日から遡った日数と相関係数の関係

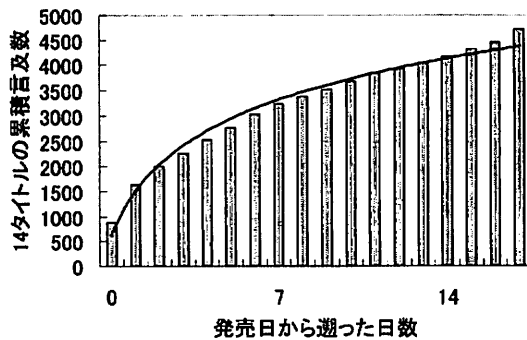


図 12. 発売日から遡った日数と累積言及数

5. まとめ

今回の調査では、テレビ視聴率・東京都知事選・音楽 CD 売り上げの全ての場合において、既存の統計指標とブログの言及数に相関関係があることがわかった。しかし、それぞれの場合において相関係数にはばらつきがあり、特に音楽 CD 売り上げの相関係数が一番低かった。これは、音楽 CD のタイトルがドラマのタイトルや都知事選の候補者に比べて複雑なものや一般的過ぎるものが多いため、また、アーティスト名に略称が多いため最適で公平なクエリーを作成できなかった可能性がある。また、アイドルグループの音楽 CD は言及数が多い傾向にあり、これが外れ値として作用し、データが外れ値に頑健でない可能性がある。同じように、東京都知事選挙においても黒川氏や中松氏、外山氏など政策の話だけではなく派手なパフォーマンスを行った候補者にブログの言及数が集まる傾向にあり、これが相関係数を下げる原因になっていると考えられ

る。

感情表現を用いたことによりテレビ視聴率とブログの言及数の相関関係は、感情表現を用いない場合と比べ強くなるということが 95% の有意水準で採択された。他のマーケット分野においても適用することができるかは、今後調査が必要である。

データ収集するための最適な時期・期間に関しては、調査 1～3 のそれぞれで異なっていた。

映画の興行収益とブログの言及との相関は感情解析を用いた方が高くなると [4] で述べられているが、今回テレビ番組においても同様であることがわかった。

6. 今後の課題

本稿においては単純な相関係数の比較にとどまったが、回帰分析を導入するなど更なる統計分析が必要であろう。また、今回比較的相関係数が低かった音楽 CD の売り上げデータに関しては、クエリーの精練やそれぞれのコミュニティによる差を補正するモデルを作るなどして対応していく。これは今回述べた東京都知事選などにも応用していけるかもしれない。

[5] ではブログでの言及により本の売り上げを予測することは出来ないが、言及数の急激な上昇が起こった場合に、本の売り上げの急激な上昇を予測する可能性があることを述べており、今後他の分野にも適用できるか調査したい。

今回はサンプルとなる分野が 3 種類だったが、分野を拡大することによりブログマイニングによるマーケットシェア予測の可能性を検討する。さらに、テレビ番組においても、テレビドラマだけでなくニュースやバラエティーなど異なるカテゴリーで調査を行い適用できる範囲を検討したい。

文 献

- [1] kizasi.jp, <http://kizasi.jp/> (2007 年 5 月アクセス)
- [2] @nifty BuzzPulse, <http://www.nifty.com/buzz/> (2007 年 5 月アクセス)
- [3] Dentsu Buzz Research, <https://www.dbuzz.jp/> (2007 年 5 月アクセス)
- [4] Yahoo! ブログ検索, <http://blog-search.yahoo.co.jp/> (2007 年 5 月アクセス)
- [5] goo-ブログの詳細検索, <http://blog.search.goo.ne.jp/>, (2007 年 5 月アクセス)
- [6] G.Mishne and N.Clance. "Predicting Movie Sales from Blogger Sentiment" AAAI-CAAW 2006.
- [7] D.Gruhl, R.Guha, R.Kumar, J.Novak, and A.Tomkins. "The Predictive Power of Online Chatter" In KDD, pages 78-87, 2005.