

Googleニュースにおけるニュース・メディアの特性

News Media Features of Google News

2007年 9月

新井 豪

Go Arai

塩谷 勇

Isamu Shioya

坂本 祐司

Yuji Sakamoto

Googleニュースにおけるニュース・メディアの特性

News Media Features of Google News

新井 豪^{*}
塩谷 勇
坂本 祐司

この研究ノートは、どのようなインターネット上のニュース記事がGoogleニュース記事となっているかについての調査の過程で、日本版Googleニュースを一定期間収集し、分析を行った結果について報告する。

In this research note, we describe statistical results obtained by collecting and analyzing Google news items for a certain period.

1. はじめに

Googleニュース[1]はインターネットから複数のウェブサイトのニュース記事を収集して、記事の内容をクラスタ分類して利用者にわかりやすいKWIC形式でニュース記事を再構成して提供している。このGoogleの試みは、現在の世の中の状況をウェブサイトのニュース記事から要約していると言うことができる。Googleの検索エンジンの利用者が多いことから、Googleニュースの利用も多いと推測することができ、Googleのニュース記事の編集方針が大きな影響を与える可能性があると考えられる。Googleニュースの記事はインターネット上のニュース記事を人間が内容を確認して分類するのではなく、自然言語処理などの技法を通じて自動的に分類している。しかし、詳細は明らかにされていない。

Googleニュースは、トップニュース、トップニュース（続き）、社会、経済、スポーツ、科学・技術、国際、政治、文化・芸能、の9つのカテゴリに分類されている。

この研究では、検索キーが与えられていない場合の日本版Googleニュースを一定期間収集（注釈:パーソナライズド検索は使用していない）して、どのようなニュース記事が採用されてGoogleニュース記事となっているか分析を行った。その結果、実メディアとの特性の違いが明らかになった。

2. 実験

2006年11月23日から12月10日までの18日分、延べ総数884件のGoogleニュース記事を、日付、閲覧時刻、カテゴリ、タイトル、記事の日付、ソース源（新聞社のウェブサイトのアドレス）、リンク先のカテゴリの7つの属性データを収集した。また、記事の文書についてもリンク先から収取した。Googleニュースは頻繁に更新されるが、手作業による収集のため、一日2回行った（この内、手作業のミスから11月30日は一日3回収集）。

収集時刻は、1回目が午前1時頃から4時前後、2回目が昼から夕方である。意図的に時間が選ぶのではなく、収集者の都合で時間が選ばれた。

3. ニュース・メディアの分析結果

3.1. Googleニュースのメディア別の件数

延べ884件（3件ソース源不明。手作業のミスと推測される）の記事の中に、サイト数69（ソース源不明1サイトを含む。Googleによると610以上のウェブサイトから最新のニュース記事を収集）であった（図1）。「日本版Googleニュース」となっているが、海外から発信し

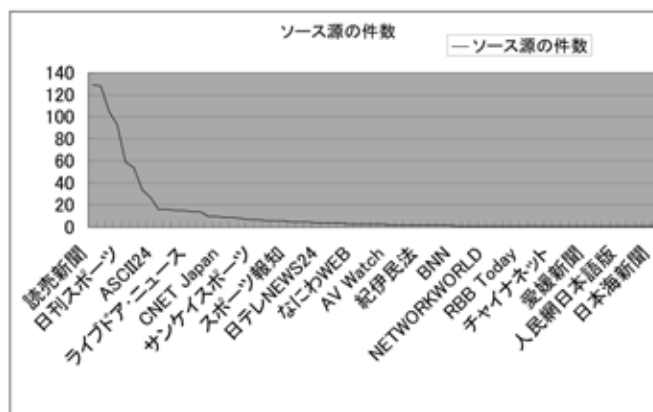


図1 Googleウェブサイト毎の分布

ている日本語のニュースサイトも含まれており、「日本語版Googleニュース」であり、多量の日本語のニュース記事を海外から発信してGoogleニュースを経由してプロパガンダ

(Propaganda) として利用する事が考えられ（例えば人民網日本語版（図1）、政治問題にも関係すると想像できる。記事の多い順に挙げると、読売新聞（129件）、朝日新聞（128件）、東京新聞（105件）、日本経済新聞（93件）、日刊スポーツ（59件）、産経新聞（54件）となり、6つのウェブサイトのニュースが全体の64%を占める（図2）。

ソース源	件数	累積%
読売新聞	129	14.9
朝日新聞	128	29.1
東京新聞	105	41.0
日本経済新聞	93	51.5
日刊スポーツ	59	58.1
産経新聞	54	64.3
デイリースポーツ	34	68.1
西日本新聞	27	71.2
ASCII24	16	73.0
ITmedia	16	74.8
中央日報	15	76.5
中日新聞	15	78.2
ライブドア・ニュース	14	79.8
中日スポーツ	14	81.3
ITpro	10	82.5
スポーツニッポン	10	83.6
CNET Japan	9	84.6
中国新聞	9	85.6
MYCOM ジャーナル	8	86.5
TBS	7	87.3
その他	112	100.0

図2. ソース源に対する件数

3.2. ニュース・カテゴリ別の件数

9つのGoogleのカテゴリに対して、ソース源のカテゴリ数は168（カテゴリ不明を1つ含む）であった。

科学・技術のカテゴリはリンク先のカテゴリとの違いが多々見受けられた。例えば、図3のようなカテゴリ「科学・技術」でありながら「飲酒運転の摘発」の内容であり、明らかに分類の誤りと思われる記事がある。スポーツ、文化・芸能のカテゴリはほぼ一致し、経済、国際、社会、政治のカテゴリは約半分が一致していたが、Googleでは経済に分類されていたカテゴリがリンク先では国際に分類されているといった相違などがあった。もしリンク先のカテゴリを参照しているとする経済、社会、政治、科学・技術のカテゴリの分類誤りはないと考えられる。これらの結果からGoogleはリンク先のニュース記事のカテゴリを参照してい

Googleニュースにおけるニュース・メディアの特性

るのではなく、少なくとも記事の内容からGoogleが独自で判断し、カテゴリに分類していることが確認できる。

閲覧日付	閲覧時刻	カテゴリ	タイトル	記事の日付	ソース源	リンク先カテゴリ
2006/12/10	18:37:00	科学・技術	飲酒運転摘発、昨冬の半分 全国一斉の取り締まり	2006年12月08日 12時02分	東京新聞	社会

図3. Googleの分類誤りの例

3.3. ニュース・カテゴリと記事件数のクラスタリング

図4に、Googleの分類の基づいた各サイトの記事の件数についての階層型のクラスタリングを行ってデンドログラムを作成した。クラスタリングのパラメータの詳細は述べないが、大手の主要新聞、スポーツ新聞、IT技術系のウェブサイト、それ以外に分けるられる。

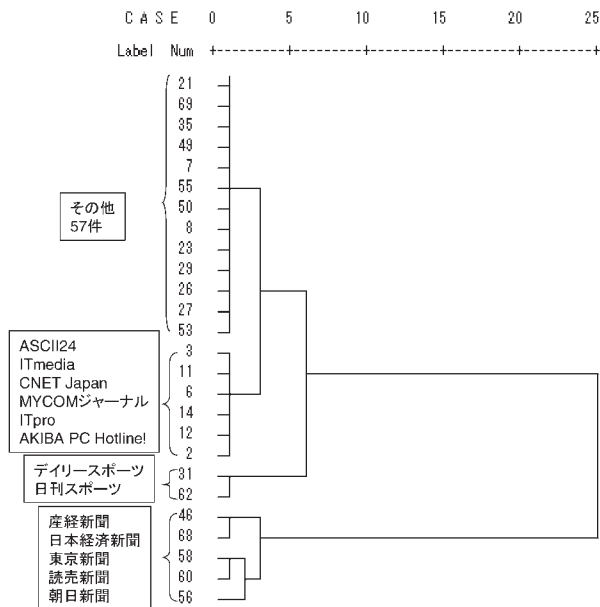


図4 Googleのカテゴリと件数に基づいたデンドログラム

3.4. 大手新聞社の発行部数と記事件数

大手新聞社に限定して、2005年の朝刊の発行部数（図5）との関係を図示すると、図6の関係が見えてくる。これより、東京新聞や日本経済新聞は朝刊の発行部数が少ないが、高い頻度でGoogleニュースの記事になっていることが解かる。新聞社のウェブサイトによっては、

ニュース記事のサイトに広告を表示させている場合もあり、広告費との関連も興味有る点である。

発行組織	発行部数	Google件数
読売新聞	10,033,215	129
朝日新聞	8,146,130	128
日刊スポーツ	3,955,525	59
毎日新聞	3,845,646	0
日本経済新聞	3,034,481	93
産経新聞	2,172,039	54
デイリースポーツ	998,921	34
西日本新聞	849,361	27
東京新聞	596,626	105

図5 2005年新聞の発行部数、朝刊

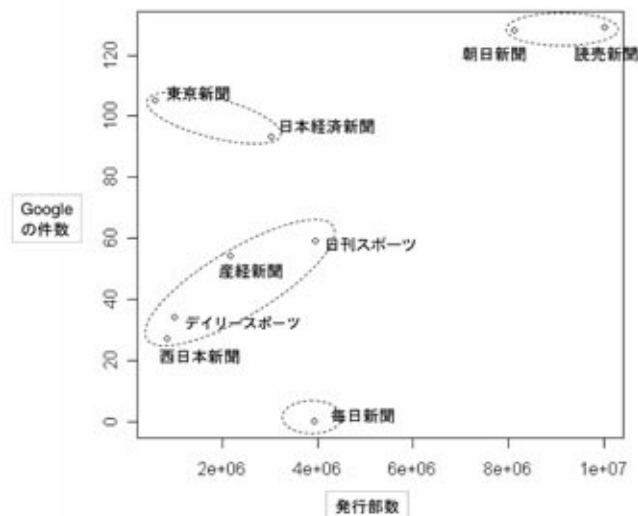


図6 発行部数とGoogleの件数

Googleカテゴリとソース源のクロス集計から各ウェブサイトの得意カテゴリが見えてきた。各ウェブサイトの記事がどのカテゴリに分類されているかを調べ、各カテゴリの頻度を調べた。その数が多かったもの上位を得意分野としている。

朝日新聞、東京新聞、読売新聞、日本経済新聞などの大手新聞社はトップニュースに選ばれることが多い（図7）。ASCII24、ITmedia、日刊スポーツなどのいわゆる専門のウェブサイトは、やはりその分野に秀でている（図8）。ニュース・カテゴリの「社会」に限定すると、主要新聞社の記事が多くを占めるが、それ以外の社会に無関係と思われるサイトも選ばれて

Googleニュースにおけるニュース・メディアの特性

いる（図9）。これは、分類の誤りでなく、実際の社会に該当する記事が選ばれている。

ソース源	件数
読売新聞	12
東京新聞	11
朝日新聞	8
西日本新聞	7
日本経済新聞	6
日刊スポーツ	3
産経新聞	3
TBS	3
ライブドア・ニュース	2
中日新聞	2
その他	11

図7 Googleの「トップニュース」カテゴリに対する件数

ソース源	件数
ASCII24	15
ITmedia	13
東京新聞	10
ITpro	9
CNET Japan	9
朝日新聞	8
MYCOMジャーナル	8
読売新聞	7
AKIBA PC Hotline!	6
日本経済新聞	5
その他	12

図8 Googleの「科学・技術」カテゴリに対する件数

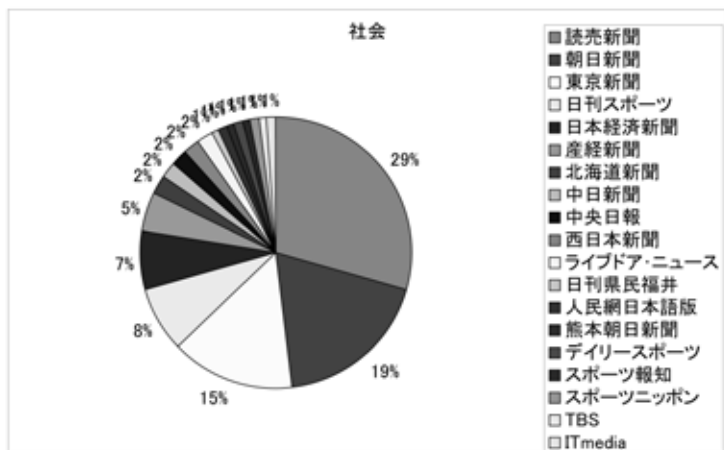


図9 社会

3.5. 総合新聞とカテゴリ別の件数

ウェブサイト限定して、Googleのカテゴリ別には選ばれる件数を調べた。読売新聞、朝日新聞、東京新聞の3つを合わせて総合新聞と呼ぶことにすると総合新聞の中で、朝日新聞は文化・芸能、スポーツ、経済に選ばれることが多い。読売新聞は社会に選ばれることが多く、東京新聞は政治では読売と同じ、国際でも朝日新聞と同じ、しかし、スポーツと文化・芸能は全く弱いと言える。日本経済新聞と産経新聞を合わせて、経済紙と呼ぶことにすると、圧倒的に日本経済新聞が勝っている（図11）。デイリースポーツと日刊スポーツを合わせて、2大スポーツ紙と呼ぶと、日刊スポーツはスポーツに選ばれることが多いが、文化・芸能で負けている。

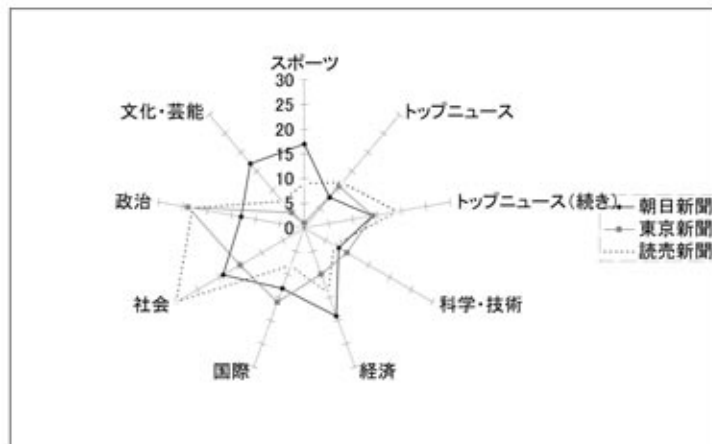


図10 総合新聞紙

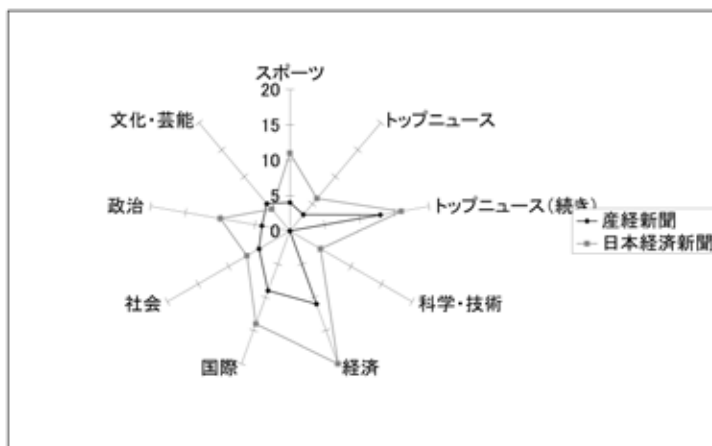


図11 経済紙

Googleニュースにおけるニュース・メディアの特性

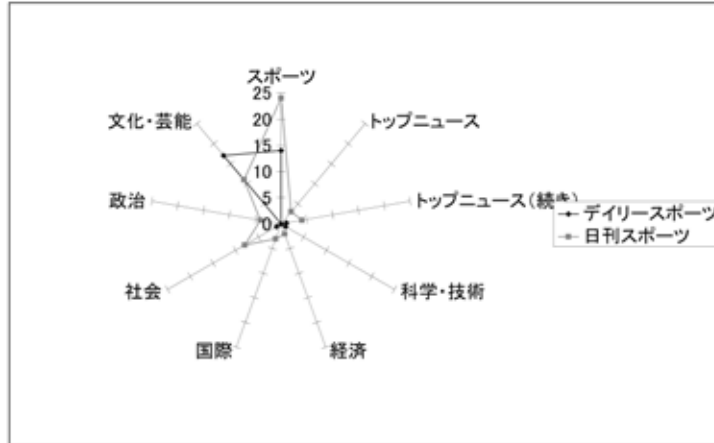


図12 2大スポーツ紙

3.6. ソース源の種類別の件数

図4から、大手新聞、スポーツ新聞、IT系、その他にソース源を分類して、Googleニュースの記事に現れる割合をグラフで示すと図13となる。大手新聞が多いということは想像できるが、IT系の新聞を発刊していないウェブサイトも7%を占めている。

また、発行部数とGoogleの件数の図6の分類に基づいて、全国紙系、Google系、スポーツ系、その他に分類してGoogleのニュース記事に現れる割合を計算すると、Google系と呼んでいる東京新聞と日本経済新聞、スポーツ系の2社がGoogleニュースに強いことが図14から解かる。

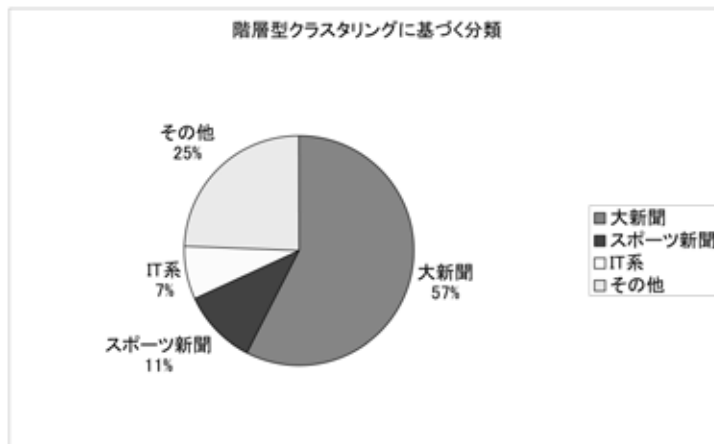


図13 階層型クラスタリングに基づく分類

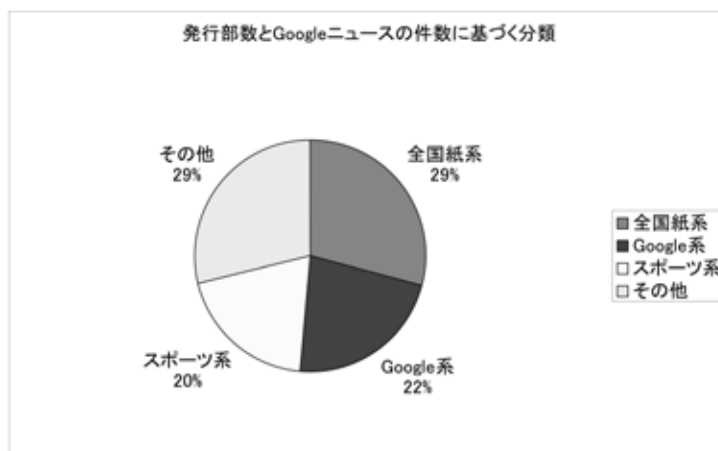


図14 発行部数とGoogle件数の図に基づく分類

3.7. まとめ

以上の結果は、発行部数の図5と対比すると、ネットワーク社会のニュース・メディアの力関係は別な点にあると言える。Googleニュースは公正と言われているが、この論文では新聞の発行部数とGoogleニュースで取り上げた件数を対比させて分析することでGoogleニュースの特性を明らかにした。

4. おわりに

Googleニュースの分類とソース源の分類を対比させることで、Googleニュースの手法が少し見えてきた。

毎日新聞、Yahooニュース、MSN Japan、Excite、共同通信などは、Google ニュースに現れていない。一方、ceek.jpニュース[2]はRSSを表示しており、ニュースのコンテンツに触れていないという差であろうか。各ウェブサイトの戦略の違いがわかる。

今後、Googleニュースで東京新聞、日本経済新聞が健闘している理由を一時的なものであるか明らかにしたい。今回は一日2回のみの調査であるため、Googleニュースの更新時刻に合わせた調査、ニュースの重要度などを含めた調査を行いたい。また、Googleニュースは独自にニュース記事を分類してニュースを提供しており、Googleの分類の精度の検討は今後の研究としたい。

参考文献

[1] <http://news.google.co.jp/>.

[2] <http://news.ceek.jp/>.

[3] メディア・リサーチ・センター、雑誌新聞総かたろぐ2006年版、2006.