

検索エンジン・サジェストの統計分布を用いた市場シェア推定

今田 貴和[†] 守谷 一朗[†] 井上 祐輔[†] 宇津呂武仁^{††} 河田 容英^{†††}
神門 典子^{††††}

[†] 筑波大学大学院システム情報工学研究科 〒305-8573 茨城県つくば市天王台 1-1-1

^{††} 筑波大学 システム情報系 知能機能工学域 〒305-8573 茨城県つくば市天王台 1-1-1

^{†††} (株) ログワークス 〒151-0051 東京都渋谷区千駄ヶ谷 5-13-18

^{††††} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

あらまし 本論文では、検索エンジン・サジェストによって測定される関心事項の情報を最大限に有効活用するタスクとして、特定商品ジャンルにおける製品・サービス等の供給者である複数の企業の間で、検索における関心の割合を比較するというタスクを設定する。そして、検索における関心の割合が、実社会における市場シェア統計との間でどの程度の相関を持つのかについて分析を行う。また、両者の間の中間的な統計情報として、価格.com 等の商品比較レビューサイトにおけるページビュー統計との間の相関の有無についての分析も併せて行う。

キーワード 検索エンジン・サジェスト, トピックモデル, 収集・集約, 商品ジャンル, 市場シェア

Market Share Estimation based on Statistics of Search Engine Suggests

Takakazu IMADA[†], Ichiro MORIYA[†], Yusuke INOUE[†], Takehito UTSURO^{††}, Yasuhide
KAWADA^{†††}, and Noriko KANDO^{††††}

[†] Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba 305-8573 Japan

^{††} Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573 Japan

^{†††} Logworks Co., Ltd. Tokyo 151-0051, Japan

^{††††} National Institute of Informatics, Tokyo 101-8430, Japan

1. はじめに

インターネットの普及により、日頃からウェブサイトを閲覧する機会が増えている。ウェブ閲覧者の多くは、自らの関心事項について Google や Yahoo!, Bing といった検索エンジンを用いてウェブ検索を行っている。各種検索エンジンの会社では、検索ログを分析することにより、景気の動向^(注1)、選挙結果の動向^(注2)、インフルエンザの動向^(注3) ^(注4)、等を予測する試みがなされている。例えば、Yahoo! JAPAN ビッグデータレポート^(注5)において検索語と景気との相関を分析した結果の報告においては、好景気との相関が高い検索語、不景気との相関が高

い検索語等が報告されている。同レポートの、選挙結果の動向を見る取り組みにおいては、特定の検索語の急上昇度などを用いて、各党の得票率を予測した結果が報告されている。近年このように、検索者の関心の割合と実社会の動向を比較する取り組みが盛んである。

本研究では、このような検索者の関心の割合と実社会の動向を比較する取り組みに関連して、特定商品ジャンルにおける製品・サービス等の供給者である複数の企業の間で、検索における関心の割合を比較するというタスクを設定する。そして、検索における関心の割合が、実社会における売れ行きシェア統計との間でどの程度の相関を持つのかについて分析を行う。本研究の方式によって、検索における関心の割合を事前に測定することができれば、市場シェアの事前予測情報として有効であると考えられる。話題性の大きい商品ジャンルや各企業の命運を握る商品ジャンルが対象の場合には、各企業の株価等にも影響しうだけの情報を持つ可能性も考えられる。

本研究の方式では、検索における関心の割合を示す指標とし

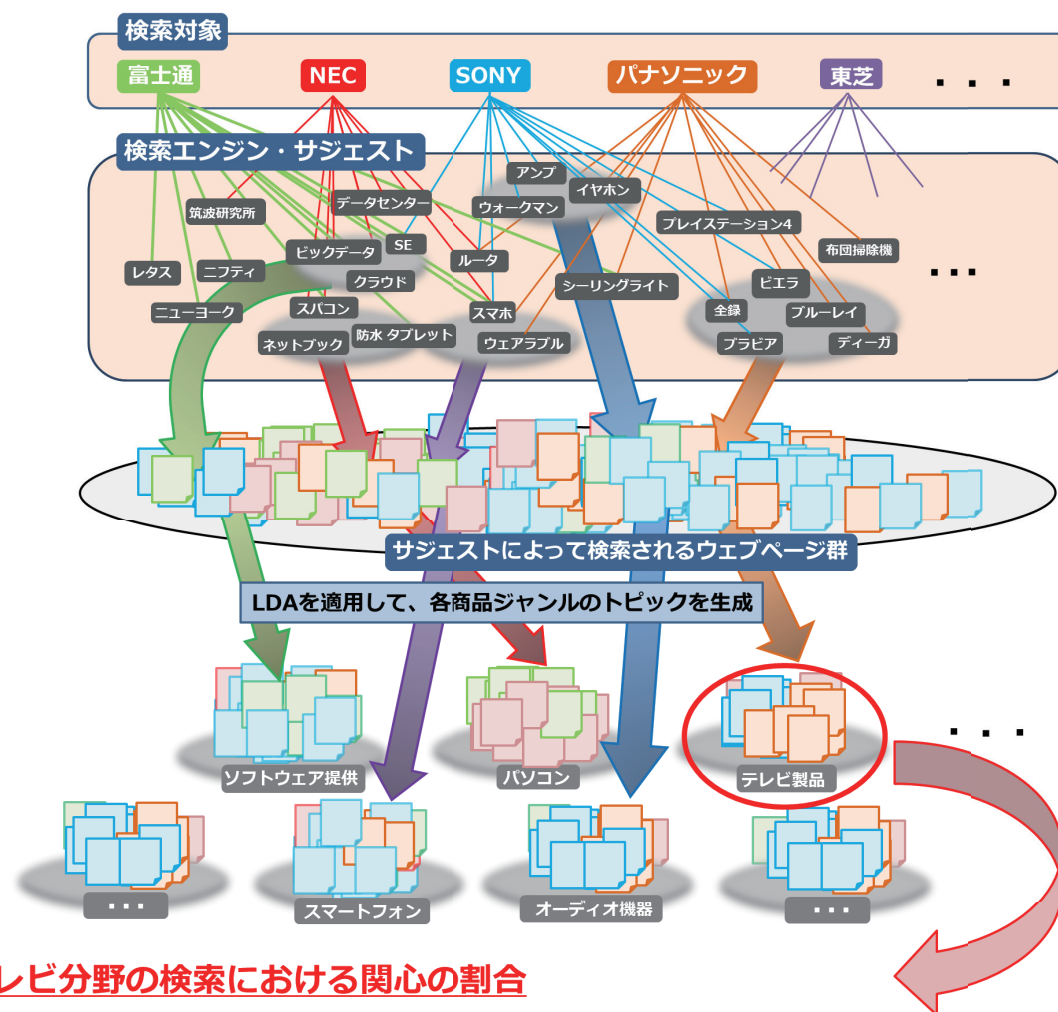
(注1) : <http://docs.yahoo.co.jp/info/bigdata/economiccondition/2014/01/>

(注2) : <http://docs.yahoo.co.jp/info/bigdata/election/2014/02/>

(注3) : <http://docs.yahoo.co.jp/info/bigdata/influenza/2014/01/>

(注4) : <http://google.org/flutrends/about/how.html>

(注5) : <http://docs.yahoo.co.jp/info/bigdata/>



テレビ分野の検索における関心の割合

パナソニック30%	SONY21%	東芝19%	その他30%
ビエラディーガチャンネル 無線lan diga Diga リモコン 設定 テレビ チャンネル録画 ブルーレイディーガ 録画チャンネル設定 ミュージックプレイヤー dimora ブルーレイ Diga リモコン 全録 ir diga 価格 ブルーレイディスク 外付けhdd レコーダー ミュージックプレイヤー 録画モード 全録機 ブルーレイ価格 dmr-bwt660 diga リモコン 故障 モニター	おでかけ放送 全録 録画モード レコーダー ホームネットワーク プラビア 録画 ブルーレイ 全番組録画 著作権 学習リモコン ルームリンク iPhone 外からどこでも視聴 録画 全録レコーダー 無線リモコン ルームリンク 対応機種 ダビング リモート視聴	wooo 録画 wowow 逆質問 ハードディスクレコーダー パズーカ タイムシフト レクザ nas ブルーレイリモコン ブルーレイ 評判 ミュージックタイマー ぜんろく ir ハルディア tiemon 外付けhdd	. . .

図1 検索における関心の割合を企業間で比較する処理の流れ

て、検索エンジン・サジェストを使用する。検索エンジン・サジェストとは、検索エンジン会社に蓄積された検索語のログを用いて提示される語である。検索エンジン・サジェストには、検索における関心の割合が反映されている。そこで、本論文では、図1に示す方式によって、特定商品ジャンルにおける製品・サービス等の供給者である複数の企業の間で、検索における関心の割合を比較する。提案手法においては、まず、分析対象の企業を検索対象として検索エンジン・サジェストを収集し、収集された検索エンジン・サジェストを索引としてウェブページ集合を収集する。ここで、「パナソニック」、「SONY」等の電気メーカ各社のように、特定の商品ジャンルにおいて競合する企業の集合を検索対象とする。これによって、競合企業数社全体で最大数万ページのウェブページを収集する。このウェブページ集合は、複数の検索対象に対して収集されたウェブページから構成される混合文書集合となっている。この混合文書集合に対してトピックモデルを適用し、トピック集合を生成する。以上

の手順によって生成されたトピックにおいては、「テレビ関連製品」、「パソコン関連製品」等、各企業が競合している商品ジャンルの話題が含まれる。そこで、それぞれの商品ジャンルに対応するトピックの一つ一つについて、検索エンジン・サジェストの企業別割合の統計分布、及び、収集されたウェブページの企業別割合の統計分布を作成し、これを用いて各商品ジャンルにおいて高い関心が持たれている企業の内訳の比較分析を行う^(注6)。

さらに、検索エンジン・サジェストを用いて作成した関心の企業別割合が、実際の商品売り上げ統計との間で相関を持つか否かを調べるために、価格.com^(注7)における統計情報との間の相関について分析する。具体的には、価格.comにおいて掲載されている実社会の市場シェア統計と、検索エンジン・サジェス

(注6)：本論文では、検索エンジン・サジェストの企業別割合の統計分布のみを示す。

(注7)：<http://kakaku.com/>

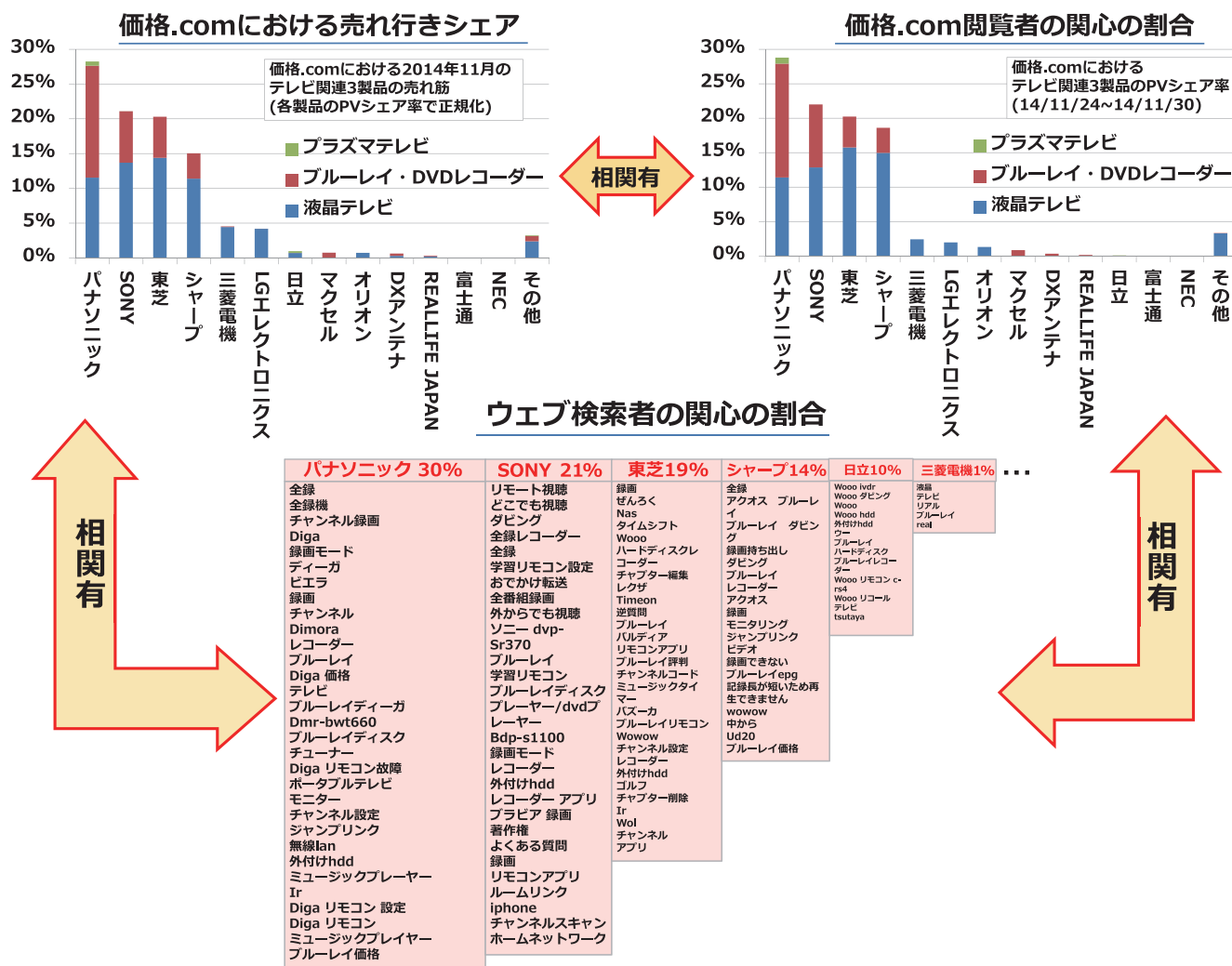


図2 「テレビ関連製品」分野におけるウェブ検索者の関心の割合、価格.com 閲覧者の関心の割合、価格.com における市場シェアの相関分析

トを用いて作成した関心の企業別割合との間の相関を分析する。ここで、一例として、「テレビ関連製品」分野において、ウェブ検索者の関心の割合、価格.com 閲覧者の関心の割合、および、価格.com における市場シェアの間の相関を分析した結果を図2に示す。この結果においては、三種類の統計はいずれも相互に相関が高くなった。以上の考え方にに基づき、本論文では、検索における関心の割合を手がかりとすることによって、実社会における市場シェア統計を推定する方式の有効性について分析を行った結果について述べる。

2. 検索エンジン・サジェストを用いたウェブページの収集

2.1 検索エンジン・サジェスト

各検索エンジン会社においては、ウェブ検索者の検索ログが蓄積されており、多数のウェブ検索者が検索したキーワードに対して、検索者が強い関心を持つ語を抽出し、検索エンジン・サジェストとして提示するサービスを提供している。ここで、本論文では、詳細な情報を検索したい対象を「検索対象」と呼

ぶ。また、検索対象に対して、検索者が AND 検索の形で二つ目以降のキーワードとして指定し、検索対象に対して詳細な情報を得るために用いる観点を「情報要求観点」と呼ぶ。すると、検索エンジン・サジェストとして提示される言葉は、「検索対象」に対して、多数のウェブ検索者が「情報要求観点」として指定した語に相当しており、ウェブ検索者の関心事項そのものを反映していることが分かる^(注8)。そこで、本論文では、検索エンジン・サジェストに着目することによって、ウェブ検索者に焦点を当て、ウェブ検索者の関心事項の収集を行う。

2.2 検索対象

本研究では、検索対象として「ASUS」、「Lenovo」、「NEC」、「SONY」、「シャープ」、「パナソニック」、「三菱電機」、「富士通」、「日立」、「東芝」の電気メーカー10社を指定し、各種電気

(注8)：図3の例では、検索窓に「パナソニック」を入力すると、「洗濯機」、「ブルーレイ」、「冷蔵庫」などが検索エンジン・サジェストとして提示される。この例では、「パナソニック」が検索対象であり、「洗濯機」、「ブルーレイ」、「冷蔵庫」等がサジェストである。また、実際の検索ログにおいては、「パナソニック AND 洗濯機」のように、検索対象とサジェストの AND 検索の形式で表現された検索要求が蓄積されている。

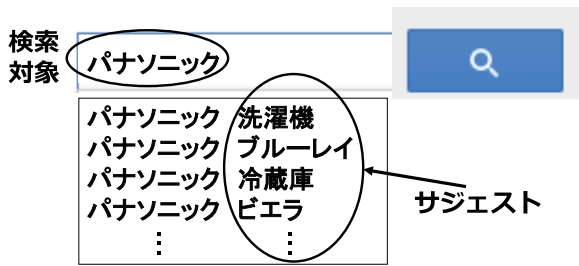


図 3 検索エンジン・サジェストの例

表 1 検索対象ごとのサジェスト数および収集されたウェブページ数

検索対象	サジェスト数	ウェブページ数
ASUS	838	5,253
Lenovo	840	5,296
NEC	902	6,611
SONY	812	5,632
シャープ	902	6,272
パナソニック	937	6,628
三菱電機	843	5,652
富士通	898	6,643
日立	925	7,031
東芝	901	6,540
混合文書集合	—	60,919

製品ジャンルにおける関心の割合を比較する。以降では、これらの検索対象を $q_j (j = 1, \dots, 10)$ とする。

2.3 検索エンジン・サジェストの収集

選定した評価用検索対象に対して、Google^(注9) 検索エンジンを用いて、一検索対象当たり約 100 通りの文字列を指定し、最大約 1,000 語のサジェストを収集する。100 通りの文字列とは具体的には、五十音、濁音、半濁音及び「きゃ」や「びゃ」などの開拗音である。例えば検索窓に「パナソニック そ」と入力すると、「掃除機」や「ソーラー」などがサジェストとして掲示されるので、それらの収集を行う。ある検索対象に対して収集されたサジェストの集合を $\mathbb{S}$ とする。検索対象毎に得られたサジェストの数を表 1 に示す。

2.4 検索エンジン・サジェストを用いたウェブページの収集

$s \in \mathbb{S}$ となるサジェスト s に対して、検索対象 q_j との AND 検索により上位 N 件以内に検索されるウェブページ p の集合を $\mathbb{P}(q_j, s, N)$ (ただし、本論文においては、 $N = 10$ とする) とし、各検索対象あたりのウェブページの文書集合 $D(q_j)$ を以下のように定義する。

$$D(q_j) = \bigcup_{s \in \mathbb{S}} \mathbb{P}(q_j, s, N)$$

なお、ウェブページの収集には Yahoo! Search BOSS API^(注10) を用いた。

2.5 ウェブページへの検索エンジン・サジェストの割り当て

各ウェブページは、検索対象および各サジェストの AND 検索によって検索されたものである。したがって、あるウェブペー

ジには、一つ以上のサジェストが対応することになる。

各ウェブページ p に対して、 $p \in \mathbb{P}(q_j, s, N)$ となるサジェスト s を集めた集合を $\mathbb{S}(q_j, p)$ とし、以下のように定義する。

$$\mathbb{S}(q_j, p) = \{ s \in \mathbb{S} \mid p \in \mathbb{P}(q_j, s, N) \}$$

2.6 複数の検索対象に対する混合文書集合の作成

2.4 節において、各企業 q_j ごとに収集したウェブページ集合 $D(q_j)$ を混合し、混合文書集合 D を作成する。

$$D = \bigcup_j D(q_j)$$

各検索対象ごとおよび混合文書集合のウェブページ数を表 1 に示す。

3. トピックモデルを用いた文書集合中の話題の集約

3.1 トピックモデル

本論文では、トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [3] を用いる。LDA を用いたトピックモデルの推定においては、語 w の集合を V として、語 $w (w \in V)$ の列によって表現された文書の集合と、トピック数 K を入力として、各トピック $z_n (n = 1, \dots, K)$ における語 w の確率分布 $P(w|z_n) (w \in V)$ 、及び、各文書 b におけるトピック z_n の確率分布 $P(z_n|b) (n = 1, \dots, K)$ を推定する。これらを推定するためのツールとしては、GibbsLDA++^(注11) を用いた。LDA のハイパーパラメータである α, β には、GibbsLDA++ の基本設定値である $\alpha = 50/K, \beta = 0.1$ を用いた。LDA ではトピック数 K を人手で与える必要があるが、

今回は、トピック数を 60 から 100 程度まで変化させてトピック推定を行い、得られたトピックを人手で見比べ、トピックの推定結果の性能がより高くなったトピック数を採用するという手順を採った。なお、このツールは推定の際に Gibbs サンプリングを用いているが、その反復回数は 2,000 とした。

また、本論文において、語 w の集合 V としては、日本語 Wikipedia 中のタイトルの集合^(注12) を用いる

また、GibbsLDA++ では、各トピック z_n において確率 $P(w|z_n)$ の高い順に語 w を W 件出力することができる。本研究においては、 $W = 20$ として、トピックの話題分析の際に参考情報として用いている。

3.2 文書に対するトピックの割り当て

本研究では、各文書に対してトピックを一意に割り当てることで、各文書を分類することとした。記事集合を D 、トピック数を K 、1 つの文書を $d (d \in D)$ とすると、トピック $z_n (n = 1, \dots, K)$ の記事集合 $D(z_n)$ は以下の式で表される。

$$D(z_n) = \left\{ d \in D \mid z_n = \underset{z_u (u=1, \dots, K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

これはつまり、文書 d におけるトピックの分布において、確率

(注11) : <http://gibbslda.sourceforge.net/>

(注12) : 2014 年 3 月にダウンロードした、エントリ数約 140 万 7,000 のものを用いた。

(注9) : <http://www.google.com/>

(注10) : <http://developer.yahoo.com/search/boss>

が最大のトピックに、文書 d を割り当てていることになる。また、検索対象 q_j ごとの記事集合 $D(z_n, q_j)$ を次式で定義する。

$$D(z_n, q_j) = D(z_n) \cap D(q_j)$$

3.3 トピックに対する検索エンジンサジェストの割り当て

また、各ウェブページには、トピックが対応付けられている。一つのトピックに対して割り当てられた一つ以上のウェブページに対応するサジェストを収集することにより、一つのトピックに一つ以上のサジェストが割り当てられていることになる。検索対象 q_j に対してあるトピック z_n に割り当てられたウェブページ集合を $D(z_n, q_j)$ とすると、トピックに割り当てられたサジェスト集合 $\mathbb{S}(z_n, q_j)$ は次式となる。

$$\mathbb{S}(z_n, q_j) = \bigcup_{p \in D(z_n, q_j)} \mathbb{S}(q_j, p)$$

トピック z_n の話題分析を行う際には、全検索対象 $q_j (j = 1, \dots, 10)$ に対する集合 $\mathbb{S}(z_n, q_j)$ 中のサジェストのうち、全検索対象に対する総頻度の上位 20 個を参照することによって話題を分析する。

3.4 トピックモデル適用結果における話題分析の手順

本節では、2.6 節において作成した混合文書集合 D に対して、トピックモデルにより推定されたトピックについて、各トピックに割り当てられたウェブページの話題がどの程度まとまっているのかの評価を行う。評価に際しては、各 z_n の混合記事集合 $D(z_n)$ について、確率の降順に 20 記事選定し、手でウェブページの内容を分析し、20 記事中、10 記事以上について話題が同一となれば、そのトピック z_n のウェブページ集合 $D(z_n)$ は、話題がまとまっていると判定した。

3.5 検索対象ごとのサジェストおよび文書の集合の作成

2.6 節において作成された混合文書集合に対して、確率 $P(z_n|d)$ の下限値、あるいは、ウェブページ数の上限値を設定し、企業別のウェブページ集合および検索エンジン・サジェスト集合を作成する。

(1) 確率 $P(z_n|d)$ の値が下限値 θ_{lbd} 以上のウェブページを収集し、集合 $D(z_k, q_j, \theta_{lbd})$ を作成する。

$$D(z_k, q_j, \theta_{lbd}) = \left\{ d \in D(z_n, q_j) \mid P(z_n|d) \geq \theta_{lbd} \right\}$$

また、それらのウェブページに割り当てられているサジェストを収集した集合を $\mathbb{S}(z_n, q_j, \theta_{lbd})$ とする。

$$\mathbb{S}(z_n, q_j, \theta_{lbd}) = \bigcup_{p \in D(z_n, q_j, \theta_{lbd})} \mathbb{S}(q_j, p)$$

(2) $P(z_n|d)$ の降順に N_{lbd} 個のウェブページを収集し、集合 $D(z_n, q_j, N_{lbd})$ を作成する。また、それらのウェブページに割り当てられているサジェストを収集した集合を $\mathbb{S}(z_n, q_j, N_{lbd})$ とする。

$$\mathbb{S}(z_n, q_j, N_{lbd}) = \bigcup_{p \in D(z_n, q_j, N_{lbd})} \mathbb{S}(q_j, p)$$

4. 検索エンジンサジェストの統計分布および市場シェアの分析

4.1 検索エンジン・サジェストの統計分布の分析

3.5 節で抽出したサジェストの集合に対して、サジェスト数

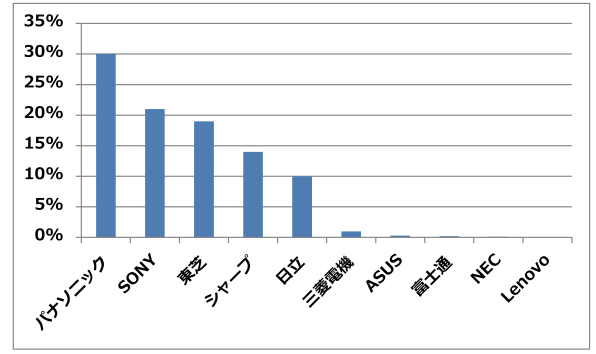


図4 「テレビ関連製品」分野における検索エンジン・サジェスト数の割合 ($\theta_{lbd} = 0.4$ の場合)

の企業別割合を算出する。

(1) 集合 $\mathbb{S}(z_n, q_j, \theta_{lbd})$ における検索エンジン・サジェスト数の企業別割合を次式で表す。

$$rate(z_n, q_j, \theta_{lbd}) = \frac{|\mathbb{S}(z_n, q_j, \theta_{lbd})|}{\sum_i |\mathbb{S}(z_n, q_i, \theta_{lbd})|}$$

(2) 集合 $\mathbb{S}(z_n, q_j, N_{ubd})$ における検索エンジン・サジェスト数の企業別割合を次式で表す。

$$rate(z_n, q_j, N_{ubd}) = \frac{|\mathbb{S}(z_n, q_j, N_{ubd})|}{\sum_i |\mathbb{S}(z_n, q_i, N_{ubd})|}$$

本論文では、下限値 θ_{lbd} 、および、上限値 N_{ubd} の値を、それぞれ、0~0.9 の範囲、および、50~900 の範囲でそれぞれ変化させて、検索エンジン・サジェスト数の企業別割合の分析を行った。このうち、「テレビ関連製品」分野において $\theta_{lbd} = 0.4$ とした場合、および、「パソコン関連製品」分野において $\theta_{lbd} = 0.1$ とした場合について、検索エンジン・サジェスト数の企業別割合を図4、および、図5に示す^(注13)。これらの結果のうち、「テレビ関連製品」分野においては、「パナソニック」が30%、「SONY」が21%、「東芝」が19%と大きな割合を示していた。一方、「パソコン関連製品」分野においては、「Lenovo」が23%、「ASUS」が21%、「NEC」が20%と大きな割合を示していた。

4.2 市場シェアの分析

本節では、前節で示した検索エンジン・サジェストの企業別割合が、価格.comにおける市場シェア統計との間でどの程度の相関を持つのかについて分析を行う。また、両者の間の中間的な統計情報として、価格.comにおけるページビュー統計^(注14)との間の相関の有無についての分析も合わせて行う。具体的には、2014年11月において、価格.comの「テレビ関連製品」分野における市場シェア、および、「パソコン関連製品」分野における市場シェアを、それぞれ、図7、および、図9に示す。また、価格.com閲覧者の「テレビ関連製品」分野における関心の割合(ページビュー)、および、「パソコン関連製品」分野にお

(注13)：これらの下限値 $\theta_{lbd} = 0.4$ および $\theta_{lbd} = 0.1$ は、いずれも、次節で述べる 価格.com における市場シェア統計との間の相関が高いパラメータである。

(注14)：<https://ssl.kakaku.com/trendsearch/index.asp>

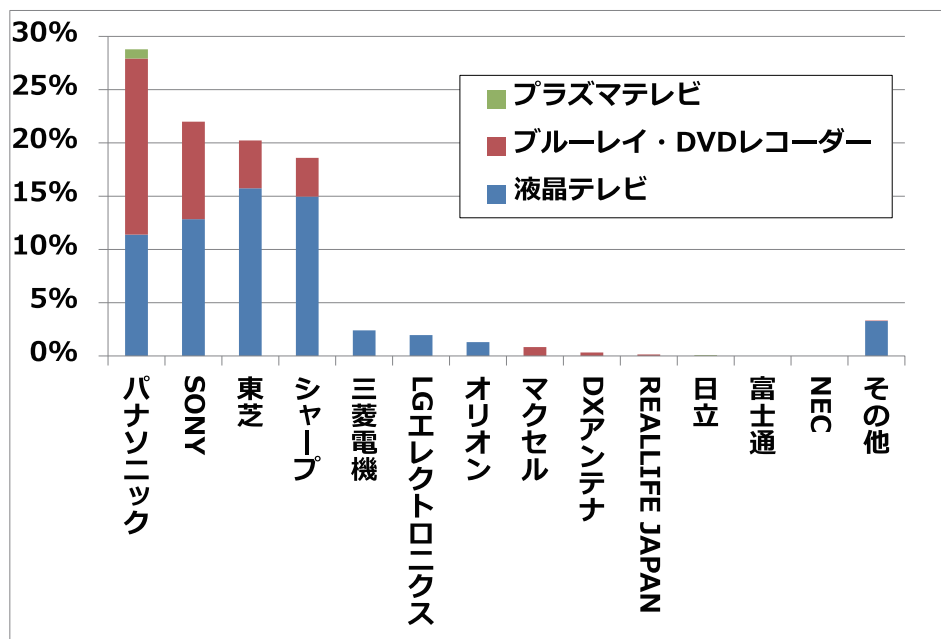


図 6 価格.com 閲覧者の関心の割合 (「テレビ関連製品」分野)

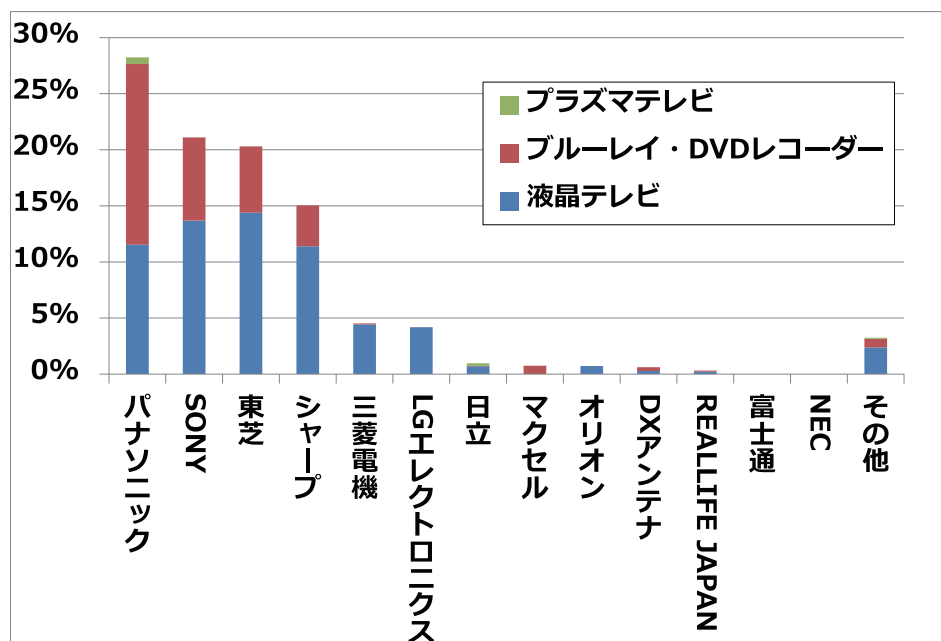


図 7 価格.com における市場シェア (「テレビ関連製品」分野)

ける関心の割合 (ページビュー) を、それぞれ、図 6、および、図 8 に示す^(注15)。以上の統計分布は、概ね、図 4、および、図 5 に示す検索エンジン・サジェスト数の企業別割合との間で相関がとれている。これらの結果から、検索における関心の割合を手がかりとすることによって、実社会における市場シェア統計を推定する方式が有効であると期待できる。

(注15)：「テレビ関連製品」分野におけるページビュー統計としては 2014 年 12 月 08 日～14 日の期間のものを、また、「パソコン関連製品」分野におけるページビュー統計としては 2014 年 11 月 24 日～30 日の期間のものを、それぞれ用いた。

5. 関連研究

本論文に関連する研究事例として、Twitter の情報拡散に着目して、国政選挙の当選者予測を行う手法 [12]、ツイートに対してインフルエンザ流行予測を行う手法 [1]、Twitter の感情分析に基づいて株式市場の予測をする手法 [4]、ツイートの内容から映画の興行収入の予測をする手法 [2]、検索エンジン、Twitter、Wikipedia の三種類の情報源から得られる情報を素性として消費トレンドを予測する手法 [6]、ブログから選挙結果や株価の予測を行う手法 [9]、Wikipedia の閲覧数から株価の動きの予想をする手法 [10] 等がある。これらの手法においては、Twitter、検索エンジンの検索数、ブログ、Wikipedia の関

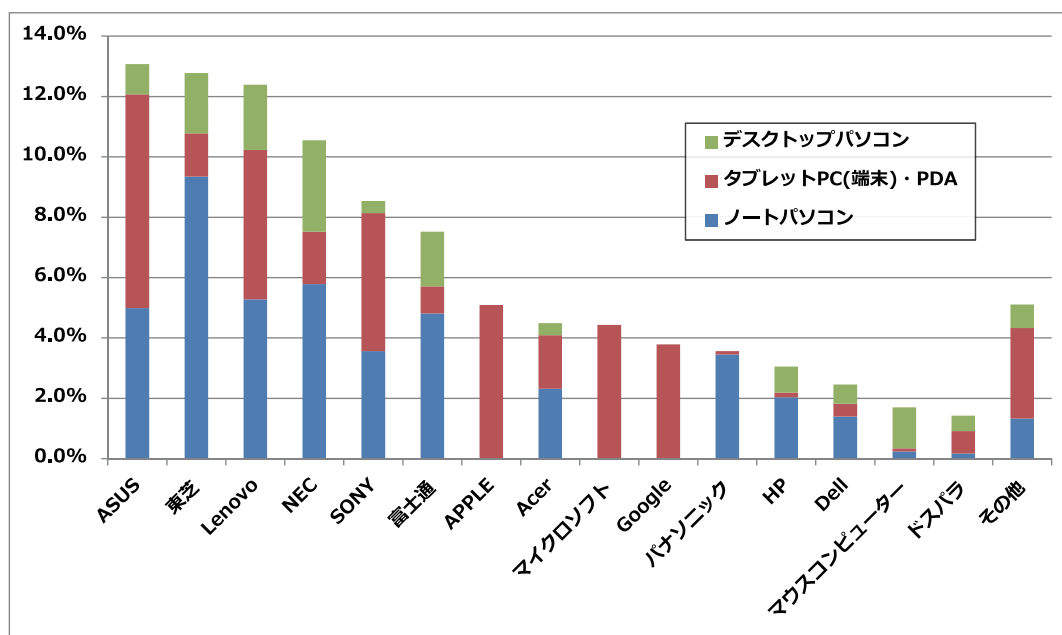


図 8 価格.com 閲覧者の関心の割合 (「パソコン関連製品」分野)

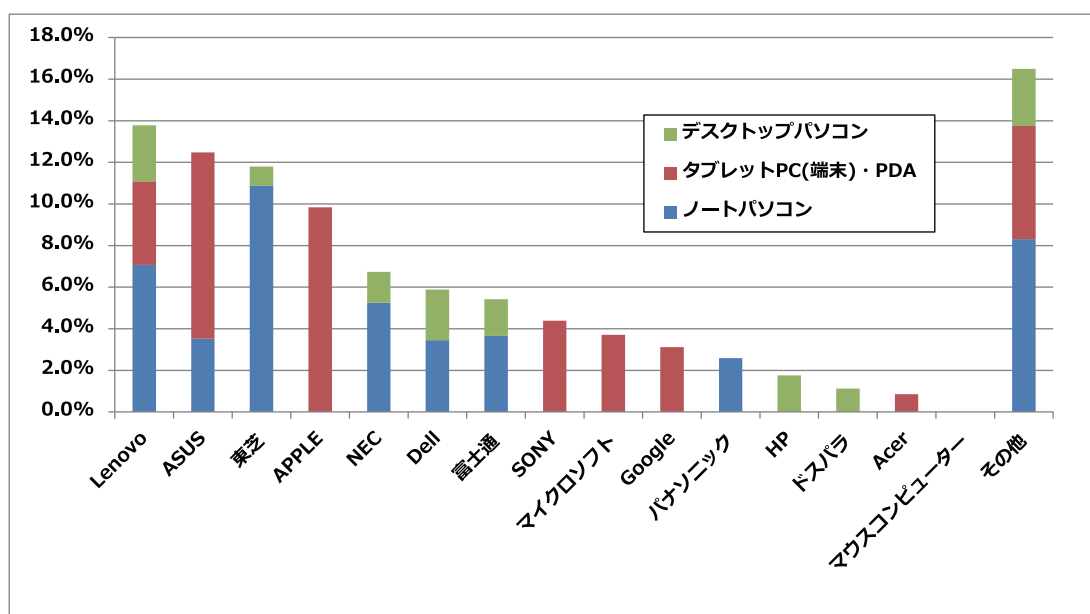


図 9 価格.com における市場シェア (「パソコン関連製品」分野)

覧数等の情報に基づいて、実社会の動きを予測している。一方、本論文では、実社会の動きを予測するための情報源として、検索エンジン・サジェストを用いることにより、検索者の関心事項を収集している。また、文献 [13] においては、検索エンジンを活用して、同一の業界における複数の企業のウェブ上における露出状況を比較することにより、実社会における企業の業績との相関を分析している。これに対して、本論文では、検索エンジン・サジェストおよびトピックモデルを利用することにより、特定の商品ジャンルにおける競合企業間の市場シェアとの相関の分析を行っている点が異なる。

本論文に関連して、筆者らは、文献 [5,7,8] において、一検索対象当たり最大約 1,000 語あるサジェストに対して、冗長なものを集約するとともに、関連するサジェストを集約的に俯瞰する

枠組みを提案している。さらに、サジェストに対して収集されるウェブページ集合に対して、話題が重複する冗長なウェブページを集約して俯瞰する枠組みを提案している。また、文献 [11] においては、検索エンジン・サジェストの中でも、Wikipedia には掲載されていない観点に焦点を当てて、Wikipedia とは異なる観点についての情報を収集して集約し、提示する方式を提案している。特に、Wikipedia においては、物事を解決するための実用的な知識や経験談、些細な雑談の類いや最新の話題等が掲載されることはあまり多くない。一方、文献 [11] においては、検索エンジン・サジェストを分析することによって、ウェブ検索者がそれらの話題についても高い関心を持っていることを示している。

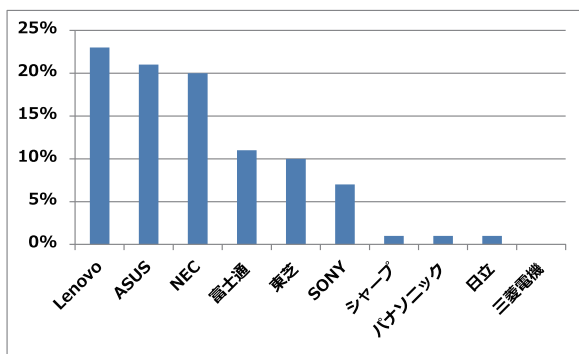


図5 「パソコン関連製品」分野における検索エンジン・サジェスト数の割合 ($\theta_{tbd} = 0.1$ の場合)

6. おわりに

本論文では、検索エンジン・サジェストによって測定される関心事項の情報を最大限に有効活用するタスクとして、特定商品ジャンルにおける製品・サービス等の供給者である複数の企業の間で、検索における関心の割合を比較するというタスクを設定した。そして、検索における関心の割合が、実社会における売れ行きシェア統計との間でどの程度の相関を持つのかについて分析を行った。また、両者の間の中間的な統計情報として、価格.com等の商品比較レビューサイトにおけるページビュー統計との間の相関の有無についての分析も併せて行った。

商品ジャンル「テレビ製品」の例のように、検索における関心の割合と実社会における売れ行きシェア統計との間で相関が認められれば、今後、多様な局面において、検索エンジン・サジェストを索引とする本研究の方式を有効に活用できる可能性が高まると考えられる。一方、商品ジャンルによっては、検索における関心の割合と実社会における売れ行きシェア統計との間で、一部相関が認められない場合も想定される。そのような場合に対しては、その現象の分析を行い、検索における関心の割合と実社会における売れ行きシェア統計との乖離の原因を解明する。

文 献

- [1] 荒牧英治, 増川佐知子, 森田瑞樹. Twitter catches the flu: 事実性判定を用いたインフルエンザ流行予測. 情報処理学会研究報告, Vol. 2011-NL-201, , 2011.
- [2] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proc. WI-IAT*, pp. 492–499, 2010.
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.
- [4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.
- [5] 土井俊弥, 井上祐輔, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. トピックモデルを用いた検索エンジン・サジェストの集約. 第29回人工知能学会全国大会論文集, 2015.
- [6] 保住純, 飯塚修平, 中山浩太郎, 高須正和, 嶋田絵理子, 須賀千鶴, 西山圭太, 松尾豊. Web マイニングを用いたコンテンツ消費トレンド予測システム. 人工知能学会論文誌, Vol. 29, No. 5, pp. 449–459, 2014.
- [7] 井上祐輔, 今田貴和, 守谷一朗, 陳磊, 宇津呂武仁, 河田容英, 神門典子. 冗長な情報要求観点の集約によるウェブ検索結果の集

- 約. 第28回人工知能学会全国大会論文集, 2014.
- [8] 小池大地, 鄭立儀, 今田貴和, 守谷一朗, 井上祐輔, 宇津呂武仁, 河田容英, 神門典子. ウェブ検索者の情報要求観点の集約. 言語処理学会第20回年次大会論文集, pp. 328–331, 2014.
- [9] 松尾豊. ウェブからの実世界の観測と予測. 電子情報通信学会論文誌, Vol. J96-B, No. 12, pp. 1309–1315, 2013.
- [10] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis. Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, Vol. 3, No. 1801, 2013.
- [11] 守谷一朗, 小池大地, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. Wikipedia 掲載事項との間の差分に着目したウェブ検索者の情報要求観点の分析. 第6回 DEIM フォーラム論文集, 2014.
- [12] 那須野薫, 松尾豊. Twitter における候補者の情報拡散に着目した国政選挙当選者予測. 第28回人工知能学会全国大会論文集, 2014.
- [13] 上野山勝也, 松尾豊. Web を用いた企業認知状況の把握と企業 PR への活用. 情報処理学会論文誌, Vol. 54, No. 11, pp. 2392–2401, 2013.