

特定商品ジャンルにおける検索エンジン・サジェストの統計分布と市場シェアの相関分析

Analyzing Correlation of
Statistics of Search Engine Suggests in a Specific Products Domain and Market Share

今田 貴和*1
Takakazu Imada

井上 祐輔*1
Yusuke Inoue

李 雪山*1
Xueshan Li

宇津呂 武仁*2
Takehito Utsuro

河田容英*3
Yasuhide Kawada

神門 典子*4
Noriko Kando

*1筑波大学大学院システム情報工学研究科
Grad. Sch. Sys. & Inf. Eng. Univ. of Tsukuba

*2筑波大学システム情報系
Fclty. Eng. Inf. & Sys. Univ. of Tsukuba

*3(株) ログワークス
Logworks Co., Ltd.

*4国立情報学研究所
National Institute of Informatics

This paper studies a task of comparing rates of concerns of those who search for Web pages among several companies which supply products or services, given a specific products domain. In this paper, we measure concerns of those who search for Web pages through search engine suggests. Then, we analyze whether rates of concerns of those who search for Web pages have certain correlation with actual market share. Furthermore, as an intermediate statistics between the rates of concerns of those who search for Web pages and market share, we also examine correlations between the page view statistics at the kakaku.com site and the other two statistics. The results of the analysis show that those three statistics have certain correlations among each other.

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開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*1, 選挙結果の動向*2, インフルエンザの動向*3 *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*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。

本論文では、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[今田 15]。本論文の方式によって、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るだけの情報を持つ可能性も考えられる。

本論文の方式では、検索における関心の割合を示す指標として、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*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*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

連絡先: 今田 貴和, 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, 029-853-5427

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

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

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

*4 <http://google.org/flutrends/about/how.html>

*5 <http://docs.yahoo.co.jp/info/bigdata/>

*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。

*7 <http://kakaku.com/>

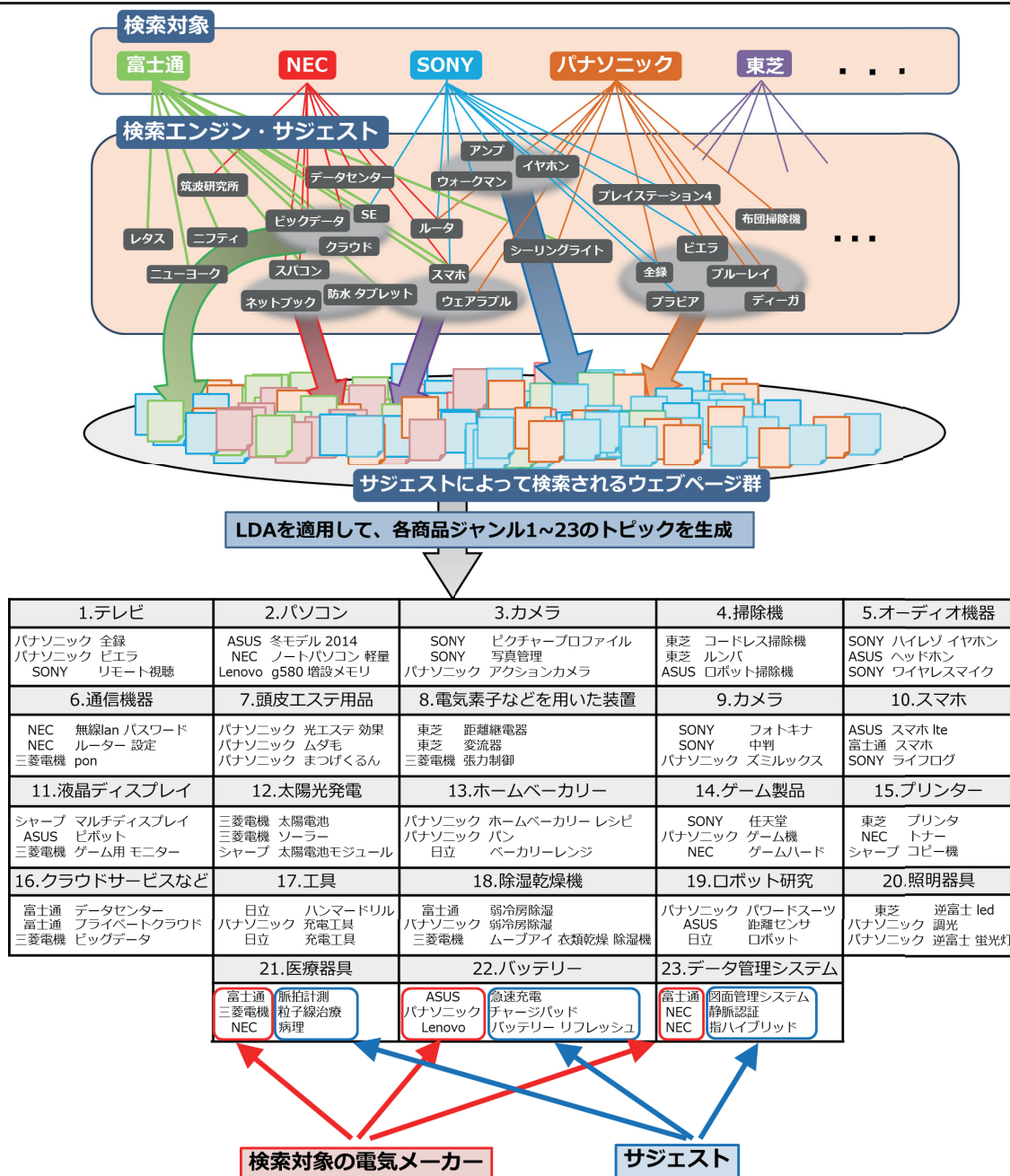


図 1: 検索における関心の割合を企業間で比較する処理の流れ

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 検索対象

本論文では、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。

2.2 検索エンジン・サジェスト及びウェブページの収集

選定した評価用検索対象に対して、Google^{*8} 検索エンジンを用いて、一検索対象当り約 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。さらに、ある検索対象に対して収集されたサジェストの集合を S とし、 $s \in S$ 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 $P(q_j, s, N)$ (ただし、本論文においては、 $N = 10$ とする) を作成する。

2.3 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、 $p \in P(q_j, s, N)$ となるサジェ

*8 <http://www.google.com/>

り当てている。

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。

2.4 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 [今田 15]。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。

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

4. 検索エンジンサジェストの統計分布

3. 節で抽出したサジェストの集合に対して、サジェスト数の企業別割合を算出する。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。

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

本論文では、下限値 θ_{lbd} の値を 0~0.9 の範囲で変化させて、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 $\theta_{lbd} = 0.4$ とした場合について、検索エンジン・サジェスト数の企業別割合と、価格.com 閲覧者の関心の割合、価格.com における市場シェアとの間で相関分析を行った結果を図 2 に示す。

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 [那須野 14]、ツイートに対してインフルエンザ流行予測を行う手法 [荒牧 11]、Twitter の感情分析に基づいて株式市場の予測をする手法 [Bollen 11]、ツイートの内容から映画の興行収入の予測をする手法 [Asur 10]、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 [保住 14]、ブログから選挙結果や株価の予測を行う手法 [松尾 13]、Wikipedia の閲覧数から株価の動きの予想をする手法 [Moat 13] 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 の閲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、[上野山 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。

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論文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。

参考文献

- [荒牧 11] 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)
- [Asur 10] Asur, S. and Huberman, B. A.: Predicting the Future with Social Media, in *Proc. WI-IAT*, pp. 492–499 (2010)
- [Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)
- [Bollen 11] Bollen, J., Mao, H., and Zeng, X.: Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011)
- [保住 14] 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)
- [今田 15] 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: 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, 第 7 回 DEIM フォーラム論文集 (2015)
- [松尾 13] 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)
- [Moat 13] Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T.: Quantifying Wikipedia Usage Patterns before Stock Market Moves, *Scientific Reports*, Vol. 3, No. 1801 (2013)
- [那須野 14] 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)
- [上野山 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)