企業名に関する検索エンジン・サジェストおよびトピックモデリングを用いた市場シェアの分析*

今田 貴和† 井上 祐輔† 陳 磊† 徐 凌寒† 宇津呂 武仁‡ 河田 容英§
筑波大学大学院 システム情報工学研究科† 筑波大学システム情報系‡ (株) ログワークス§

1 はじめに

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度合いを比較するというタスクを設定する (図 1)．そして，検索における関心の度合いが，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 [6]．そして，これらの分析の結果をふまえて，検索エンジン・サジェストによって得られる統計分布に基づいて市場シェアを予測する手法を提案する．

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 クエリ・フォーカス

本論文では，クエリ・フォーカスとして「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 $q_j (j = 1, \ldots, 10)$ とする．

2.2 サジェスト及びウェブページの収集

選定した評価用クエリ・フォーカスに対して，Google[1] 検索エンジンを用いて，一クエリ・フォーカス当り約 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，あるクエリ・フォーカスに対して収集されたサジェストの集合を S として，$s \in S$ となるサジェスト s に対して，クエリ・フォーカス 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 d の集合 $D(q_j, s, N)$ (ただし，本論文においては，$N = 10$ とする) を作成する．ここで，ウェブページの収集には Yahoo! Search BOSS API[2] を用いる．また，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．各クエリ・フォーカスごとのサジェスト数およびウェブページ数の一例を表 1 に示す．

表 1: クエリ・フォーカスごとのサジェスト数および収集されたウェブページ数 (2015 年 8 月 6 日収集)

クエリ・フォーカス	サジェスト数	ウェブページ数
ASUS	840	5,012
Lenovo	839	5,163
NEC	909	6,329
SONY	812	5,695
シャープ	900	5,885
パナソニック	938	6,541
三菱電機	847	5,301
富士通	896	6,071
日立	912	6,568
東芝	896	6,367
混合文書集合	—	57,582

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，クエリ・フォーカス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 d に対して，$d \in D(q_j, s, N)$ となるサジェスト s を集めた集合を $S(q_j, d)$ とする．

3 トピックモデルを用いた文書集合中の話題の集約

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

*Analyzing Market Share based on Topic Modeling of Results of Web Search with Search Engine Suggests of Company Names

†Takakazu Imada, Yusuke Inoue, Lei Chen, Linghan Xu, Graduate School of Systems and Information Engineering, University of Tsukuba

‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba

§Yasuhide Kawata, Logworks Co., Ltd.

[1] http://www.google.com/

[2] http://developer.yahoo.com/search/boss

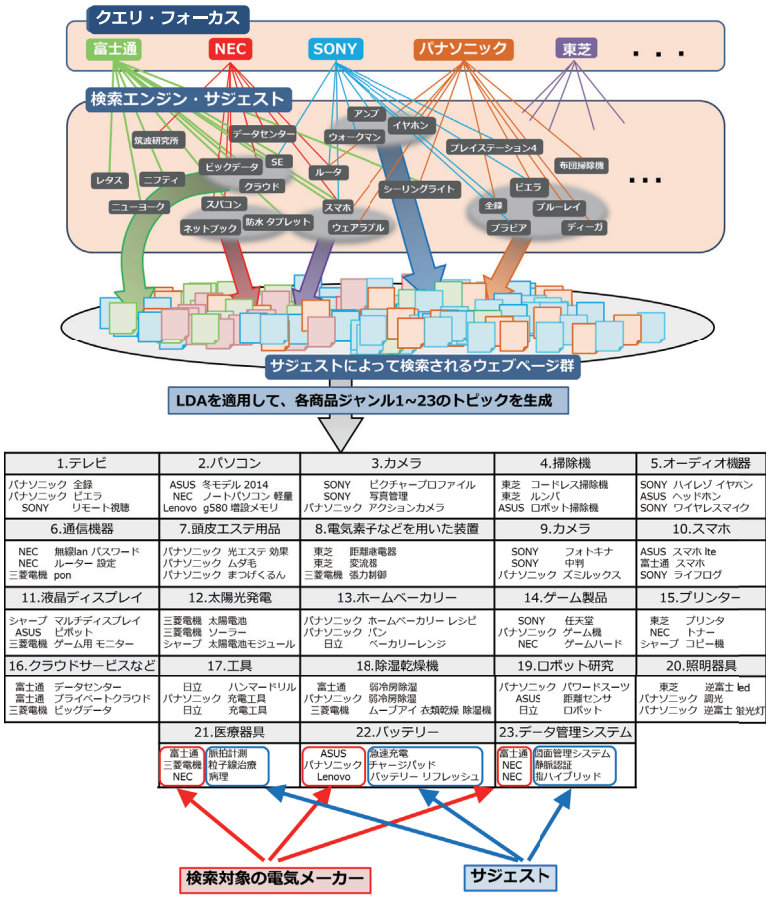


図 1: 検索における関心の割合を企業間で比較する処理の流れ

て，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)$ $(w \in V)$，及び，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ を推定する．これらを推定するためのツールとしては，GibbsLDA++[3]を用いた．

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．本論文においては，$W = 20$ 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．

本論文では，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．文書 d におけるトピックの分布において，確率が最大のトピックに，文書 d を割り当てている．

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．クエリ・フォーカス 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 $S(z_n, q_j)$ は次式となる．

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

[3]http://gibbslda.sourceforge.net/

トピック z_n の話題分析を行う際には，全クエリ・フォーカス $q_j (j = 1, \ldots, 10)$ に対する集合 $S(z_n, q_j)$ 中のサジェストのうち，全クエリ・フォーカス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．

2.2 節において作成された混合文書集合に対して，確率値 $P(z_n|d)$ 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 [6]．確率値 $P(z_n|d)$ 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．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 $S(z_n, q_j, \theta_{lbd})$ とする．

4 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の分析

4.1 サジェストの統計分布の分析

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 $S(z_n, q_j, \theta_{lbd})$ 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．

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

本論文では，トピック数 K を 60 から 100 程度まで変化させてトピック推定を行った．また，各トピック

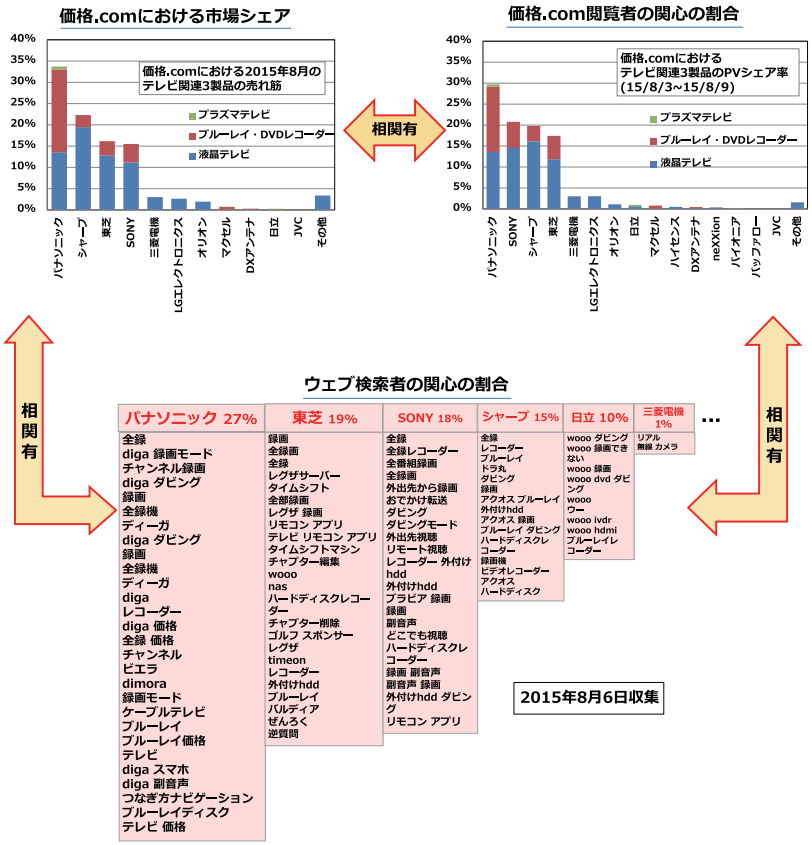


図 2: 「テレビ関連製品」分野におけるウェブ検索者の関心の割合，価格.com 閲覧者の関心の割合，価格.com における市場シェアの相関分析

数 K において，確率値 $P(z_n|d)$ の下限値 θ_{lbd} の値を 0~0.9 の範囲で変化させて，トピック数 K および確率値 $P(z_n|d)$ の下限値 θ_{lbd} の組について，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，表 1 において示したウェブページ集合を対象として,「テレビ関連製品」分野に相当するトピックにおいて，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係数の平均値が最大となるトピック数 K および確率値 $P(z_n|d)$ の下限値 θ_{lbd} を求めた結果，$K = 90$，$\theta_{lbd} = 0.4$ となった．この場合について，検索エンジン・サジェスト数の企業別割合を，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との間で比較した結果を図 2 に示す．

4.2 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予測

本節では，図 3 に示す手順によって，検索エンジン・サジェストの統計分布を用いて，価格.com における市場シェア統計，および，ページビュー統計を予測する手法について述べる．本論文の手法においては，まず，第 1 月目から第 $N-1$ 月目までの各月ごとに，価格.com において，各商品ジャンルごとの市場シェア統計およびページビュー統計を収集する．同様に，第 1 月目から第 $N-1$ 月目までの各月ごとに，表 1 においてクエリ・フォーカスとして示す企業についての検索エンジン・サジェストおよびウェブページ集合を収集し，混合文書集合に対してトピックモデルを適用する．そして，各商品ジャンルに対応するトピックを人手で選定し，検索エンジン・サジェストの企業別割合を求め，価格.com における市場シェア統計，および，ページビュー統計との間の相関係数を求める．そして，各商品ジャンルごとに，第 1 月目から第 $N-1$ 月目における相関係数の平均値を最適化するトピック数 K および確率値 $P(z_n|d)$ の下限値 θ_{lbd} を求める．最後に，このパラメータ値を用いて，第 N 月目における検索エンジン・サジェストの企業別割合を求め，これを第 N 月目における市場シェア統計，あるいは，ページビュー統計の予測値とする．

以上の手順によって,「テレビ関連商品」分野および「パソコン関連商品」分野を対象として，2015 年 3 月を第 1 月目，同年 5 月を第 $N-1$ 月目として，第 N 月目である同年 6 月の市場シェア統計およびページビュー統計を予測した結果と，実際の市場シェア統計およびページビュー統計との間の相関係数をプロットした結果を図 4 に示す．図 4 中には，同様の手順により，同年 7 月，8 月，9 月をそれぞれ第 N 月目と

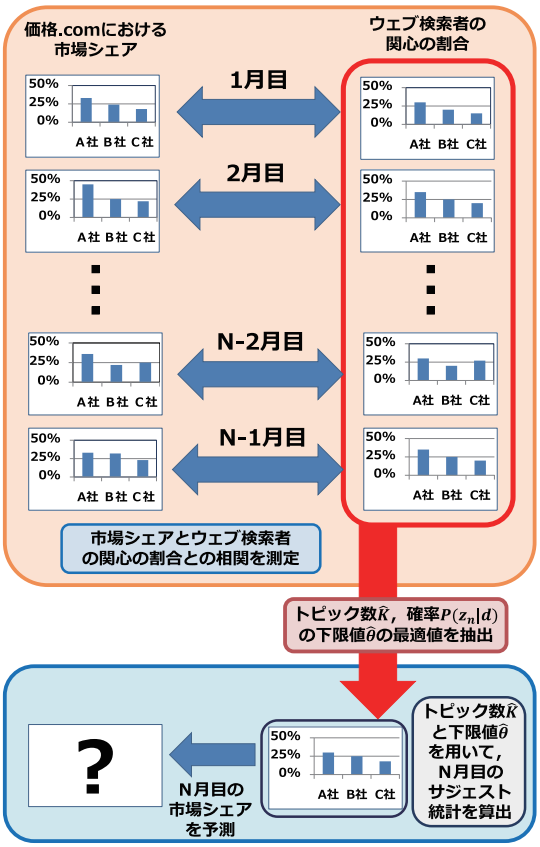


図 3: ウェブ検索者の関心の割合を用いた市場シェアの予測手順

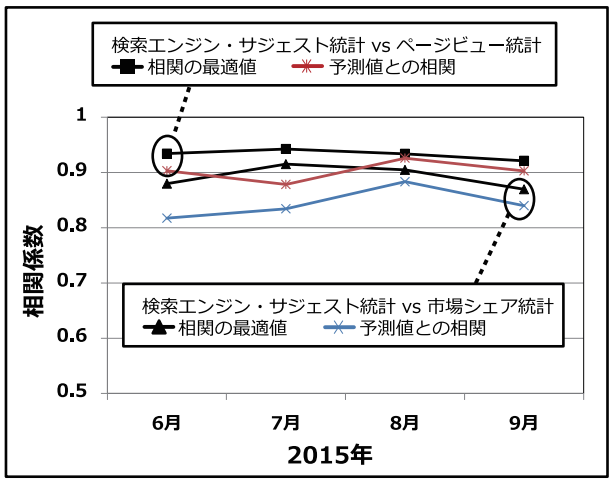
して，予測結果と実際の統計値との間の相関係数をプロットした結果を示す．また，比較対象として，各月において，検索エンジン・サジェストの企業別割合と市場シェア統計，および，ページビュー統計との間の相関係数の最適値を求めた結果のプロットも併せて示す．この結果から分かるように，提案手法によって，検索エンジン・サジェストの企業別割合を用いることによって，実際の市場シェア統計およびページビュー統計との間の相関係数の最適値とほぼ同等の相関係数を示す予測結果が得られている．

5 関連研究

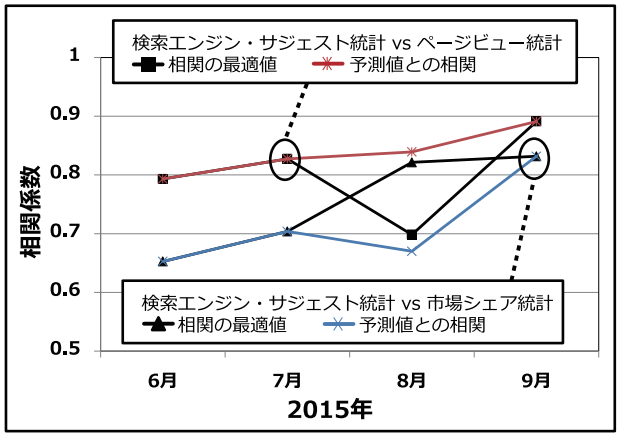
本論文に関する関連研究として，Twitter，検索エンジンの検索数，ブログ，Wikipedia の閲覧数等の情報に基づいて，実社会の動きを予測する手法 [1,2,4,5,7,8] が提案されている． 一方，本論文では，実社会の動きを予測するための情報源として，検索エンジン・サジェストを用いる手法を提案している．

6 おわりに

本論文では，検索エンジン・サジェストの企業別割合が，実社会における市場シェア統計，および，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でどの程度の相関を持つのかについて分析を行った．さ



(a) 「テレビ関連製品」分野



(b) 「パソコン関連製品」分野

図 4: 検索エンジン・サジェストの企業別割合を用いて予測された市場シェア統計・ページビュー統計と実際の市場シェア統計・ページビュー統計との相関の推移

らに，これらの分析の結果をふまえて，検索エンジン・サジェストによって得られる統計分布に基づいて市場シェアを予測する手法を提案した．

参考文献

[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] 保住純, 飯塚修平, 中山浩太郎, 高須正和, 嶋田総理子, 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.

[6] 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.

[7] 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.

[8] 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.