Analyzing Time Series Changes of Correlation between Market Share and Concerns on Companies measured through Search Engine Suggests

Takakazu Imada[1], Yusuke Inoue[1], Lei Chen[1], Syunya Doi[1], Tian Nie[1], Chen Zhao[1], Takehito Utsuro[2], Yasuhide Kawada[3]

[1]Graduate School of Systems and Information Engineering, University of Tsukuba
[2]Faculty of Engineering, Information and Systems, University of Tsukuba
[3]Logworks Co., Ltd.

Abstract

This paper proposes how to utilize a search engine in order to predict market shares. We propose to compare rates of concerns of those who search for Web pages among several companies which supply products, given a specific products domain. 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. We show that those statistics have certain correlations. We finally propose how to predict the market share of a specific product genre based on the rates of concerns of those who search for Web pages.

Keywords: search engine suggest, topic model, market share, products genre, aggregation

1. Introduction

Search engine companies recently examine how to predict specific information in real world such as economic trend [1], results of elections[2], spread of influenza[3] [4] by analyzing concerns of those who search for Web pages within the search engine log data. For example, in the analysis of correlation between economic trend and search queries reported by Yahoo! JAPAN Big Data Report[5], search queries that have high correlation with economic boom as well as economic depression are identified. Thus, it has become popular recently to analyze concerns of those who search for Web pages and to compare them with people's behavior in real world.

Considering those recent studies on correlations between concerns of those who search for Web pages and real world statistics, this paper proposes to utilize a search engine as a social sensor which is to be used for predicting market shares. In this paper, we measure concerns of those who search for Web pages through search engine suggests. Let us consider a case where we are given a query keyword "パナソニック" (Panasonic). Here, the search engine collects user search logs including the query keyword "パナソニック" (Panasonic) and then presents suggests keywords such as "洗濯機" (washing machine), "ブルーレイ" (Blu-ray), " 冷蔵庫" (refrigerator), and "ビエラ" (VIERA) which have strong relation to the query term "パナソニック" (Panasonic).

In our framework, given the query company names for analyzing concerns of those who search for Web pages, we first collect search engine suggests of those query company names. Next, by specifying those query company names as well as search engine suggests, we retrieve Web pages for all of the search engine suggests and collect them into a mixture of Web pages for several companies that are competitive with each other in certain product genres. Then, we apply a topic model to those mixture of Web pages and generate a set of topics. In our framework, it is expected that, out of the whole set of generated topics, certain percentage can be regarded as certain product genres. Actually, in the case of 10 company names we examine in this paper as search queries, out of the total 80 topics generated by the topic modeling procedure, we observed 23 topics that can be clearly regarded as certain product genres.

Next, for each of those 23 topics regarding certain product genres, we compare the statistics of search engine suggests among the 10 company names. In the case of the TV and related product genres shown in Figure 1, for example, companies which have higher rates in the statistics of search engine suggests include "パナソニック" (Panasonic) (30%), "SONY" (21%), and "東芝" (TOSHIBA) (19%). Next, we examine whether rates of concerns of those who search for Web pages represented in terms of statistics of search engine suggests have certain correlation with actual market share at `kakaku.com`. We also examine an intermediate statistics between the rates of concerns of those who search for Web pages and market share. More specifically, we examine the page view statistics[6] at the `kakaku.com` site and its correlation with the other two statistics. Figure 1 is an example of analyzing those correlations in TV and related product genres. Those three statistics have quite high correlations. In this paper, we propose how to analyze correlation between concerns of those who search for Web pages and actual market share in certain product genres. We also study time series changes of those correlations and show that their correlations continue to be fairly high for about five months periods. We finally propose how to predict the market share of a specific product genre based

[1] `http://docs.yahoo.co.jp/info/bigdata/economiccondition/2014/01/` (in Japanese)

[2] `http://docs.yahoo.co.jp/info/bigdata/election/2014/02/` (in Japanese)

[3] `http://docs.yahoo.co.jp/info/bigdata/influenza/2014/01/` (in Japanese)

[4] `http://google.org/flutrends/about/how.html`

[5] `http://docs.yahoo.co.jp/info/bigdata/` (in Japanese)

[6] `https://ssl.kakaku.com/trendsearch/index.asp`

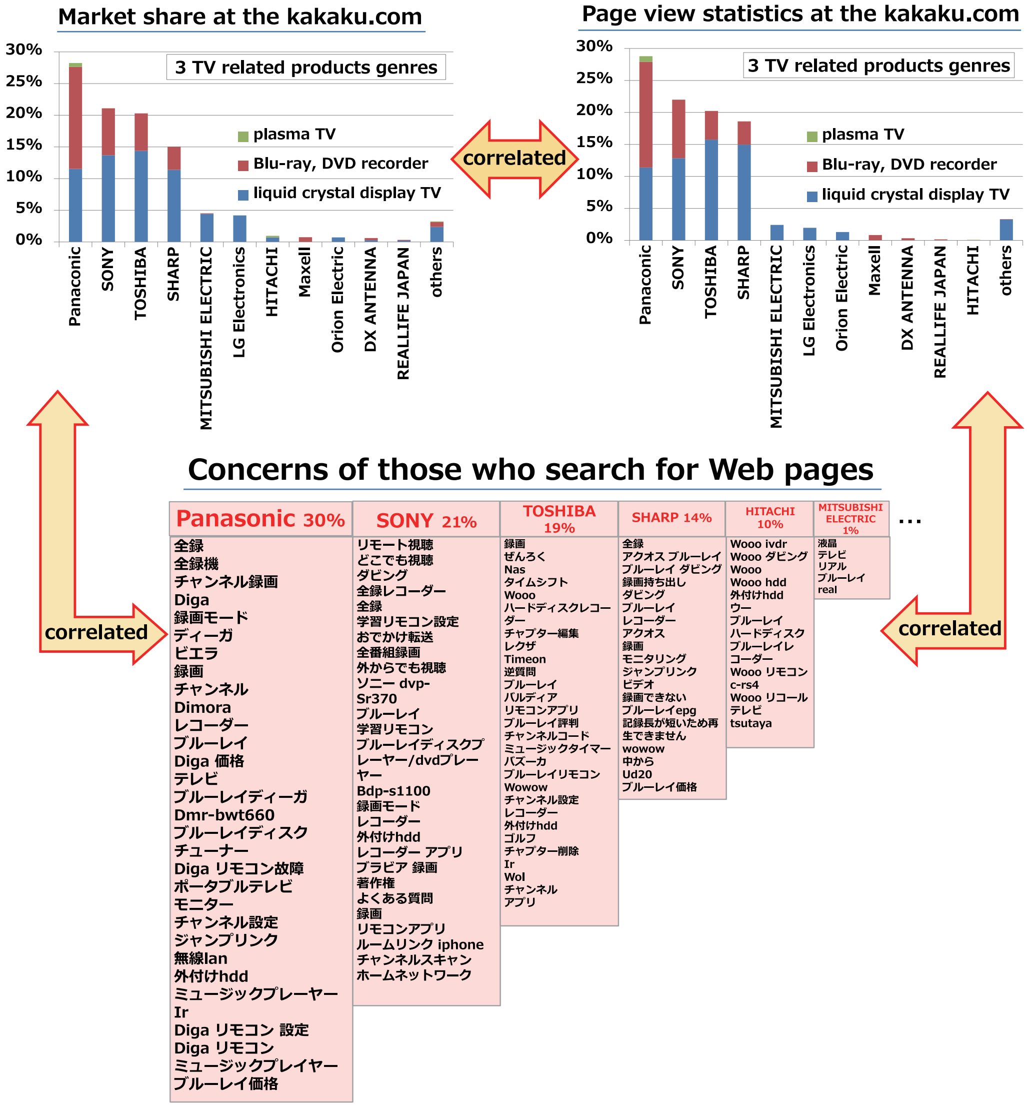


Figure 1: Analyzing Correlation in TV and related Products Genres among Rates of Concerns of those who Search for Web Pages, Page View Statistics at kakaku.com, and Market Share at kakaku.com

on the rates of concerns of those who search for Web pages.

2. Search Query and Search Engine Suggests

We examine 10 query company names "ASUS", "Lenovo", "NEC", "SONY", "シャープ" (SHARP), "パナソニック" (Panasonic), "三菱電機" (MITSUBISHI ELECTRIC), "富士通" (FUJITSU), "日立" (HITACHI), and "東芝" (TOSHIBA). They include well known Japanese electronics makers. We then compare rates of concerns on those companies with respect to the products genre of electronics. We denote those 10 query company names as q_j $(j = 1, \ldots, 10)$ in this paper.

For a given query keyword, we specify about 100 types of Japanese hiragana characters[7] to Google search engine[8] and then collect at most about 1,000 suggests. For example, when we type in "パナソニック そ" ("Panasonic", "so") into the Web search window, we can collect suggests which start with the reading character "そ" ("so") such as "掃除機 (soojiki)" ("cleaner") and "ソーラー (sooraa)" ("solar") are collected. For each query company name q_j $(j = 1, \ldots, 10)$, we denote the set of collected search engine suggests as $\mathbb{S}(q_j)$. In Table 1, we show an example the numbers of collected search engine suggests as well as those of collected Web pages for each of the 10 query company names.

3. Collecting Web Pages

Let s $(\in \mathbb{S}(q_j))$ be a search engine suggest for the search query q_j. Then, we collect top N ranked Web pages retrieved by AND search of the given query q_j and the search engine suggest s into the set $D_N(q_j, s)$[9]. Next, we collect those retrieved Web pages for all of the search engine sug-

[7] About 100 types of Japanese hiragana characters include Japanese alphabet consisting of about 50 characters, voiced and semi-voiced variants of voiceless characters, and Youon.

[8] https://www.google.com/

[9] We use the number N as 10 in this paper.

Table 1: Number of Collected Suggests and Web Pages (collected on July 3rd, 2015)

query q_j	# of suggests $\lvert\mathbb{S}(q_j)\rvert$	# of Web pages $\lvert D_N(q_j)\rvert$
ASUS	879	5,483
Lenovo	875	5,203
NEC	943	6,465
SONY	871	5,881
“シャープ” (SHARP)	942	6,210
“パナソニック” (Panasonic)	971	6,653
“三菱電機” (MITSUBISHI ELECTRIC)	879	5,520
“富士通” (FUJITSU)	950	6,620
“日立” (HITACHI)	947	6,766
“東芝” (TOSHIBA)	933	6,534
total # of Web pages $\lvert D_N\rvert$	—	61,529

gests s in $\mathbb{S}(q_j)$ into $D_N(q_j)$ as below:

$$D_N(q_j) \quad = \bigcup_{s\in\mathbb{S}(q_j)} D_N(q_j, s)$$

We use Yahoo! Search BOSS API[10] to collect N $(= 10)$ URLs. Then, we collect those Web pages within $D_N(q_j)$ that are collected for each query q_j of the 10 company names into the the set D_N of the mixture of Web pages as below:

$$D_N \quad = \bigcup_j D_N(q_j)$$

Table 1 shows an example of the numbers of collected Web pages for each of the 10 query company names as well as that of the set of their mixture.

Since each Web page is retrieved by AND search of the given query q_j and the search engine suggest s, one or more suggests are assigned to each Web page. For each Web page d, we collect search engine suggests s which satisfy $d \in D_N(q_j, s)$ into the set $\mathbb{S}(q_j, d, N)$ as below:

$$\mathbb{S}(q_j, d, N) \quad = \quad \left\{ s \in \mathbb{S}(q_j) \;\middle|\; d \in D_N(q_j, s) \right\}$$

4. Topic Modeling

We apply a topic model to the set D_N of the mixture of Web pages $D_N(q_j)$ collected for each query q_j of the 10 company names. As a topic model, this paper employs LDA (Latent Dirichlet Allocation) (Blei et al., 2003). In this paper, we estimate the distributions $p(w \mid z_n)$ $(w \in V)$ and $p(z_n \mid d)$ $(n = 1, \ldots, K)$ by GibbsLDA++[11], where the parameters are tuned through a preliminary evaluation by examining the number of topics as $K = 60 \sim 100$, and are then determined as $K = 80$ as well as $\alpha = 50/K, \beta = 0.1$. Let d be a Web page in the set D_N of the mixture of Web pages $D_N(q_j)$ collected for each query q_j $(j = 1, \ldots, 10)$ out of the 10 company names, and K be the number of topics. Then, to each Web page d, we assign the topic z_n with the highest probability $P(z_n|d)$. Next, for each topic z_n, we collect Web pages d to which the topic z_n is assigned into the set $D_N(z_n)$:

$$D_N(z_n) \quad = \quad \left\{ d \in D_N \;\middle|\; z_n = \underset{z_u\ (u=1,\ldots,K)}{\arg\max}\ P(z_u|d) \right\}$$

Furthermore, for each query q_j $(j = 1, \ldots, 10)$ out of the 10 company names, we extract Web pages collected by specifying q_j as a query from the set $D_N(z_n)$ into $D_N(z_n, q_j)$ as below:

$$D_N(z_n, q_j) \quad = \quad D_N(z_n) \bigcap D_N(q_j)$$

5. Search Engine Suggests of a Topic

Next, for each topic z_n, this section describes how to collect search engine suggests from the the set $D_N(z_n)$ of Web pages for z_n. Then, given a topic z_n and a query q_j out of the 10 company names, for each Web page d included in the set $D_N(z_n, q_j)$, we collect those search engine suggests within the set $\mathbb{S}(q_j, d, N)$ into the set $\mathbb{S}(z_n, q_j, N)$ as below:

$$\mathbb{S}(z_n, q_j, N) \quad = \bigcup_{d\in\ D_N(z_n, q_j)} \mathbb{S}(q_j, d, N)$$

[10] `http://developer.yahoo.com/search/boss`

[11] `http://gibbslda.sourceforge.net/`

Table 2: Results of Measuring Time Series Changes of Correlation of Market Share and Page View Statistics with Rates of Search Engine Suggests (Pearson product-moment correlation coefficient)

(a) TV and related products genre

		Mar. 2015	May 2015	Jul. 2015
rates of search engine suggests	kakaku.com market share	0.84	0.80	0.86
	kakaku.com page view statistics	0.81	0.93	0.90
kakaku.com page view statistics	kakaku.com market share	0.97	0.88	0.99

(b) PC and related products genre

		Mar. 2015	May 2015	Jul. 2015
rates of search engine suggests	kakaku.com market share	0.76	0.53	0.64
	kakaku.com page view statistics	0.91	0.66	0.68
kakaku.com page view statistics	kakaku.com market share	0.76	0.87	0.73

When analyzing and comparing statistics of search engine suggests among the 10 companies for evaluation in the next section, we consider the lower bound θ_{lbd} of the probability $P(z_n|d)$. Then, from the set $D_N(z_k, q_j)$, we collect Web pages which satisfy the lower bound θ_{lbd} into the set $D_N(z_k, q_j, \theta_{lbd})$ as below:

$$D_N(z_k, q_j, \theta_{lbd}) \;=\; \Big\{ d \in D_N(z_n, q_j) \;\Big|\; P(z_n|d) \geq \theta_{lbd} \Big\}$$

Furthermore, we collect search engine suggests assigned to each Web page d in the set $D_N(z_k, q_j, \theta_{lbd})$ into the set $\mathbb{S}(z_n, q_j, N, \theta_{lbd})$ as below:

$$\mathbb{S}(z_n, q_j, N, \theta_{lbd}) \;=\; \bigcup_{d \in\, D_N(z_n, q_j, \theta_{lbd})} \mathbb{S}(q_j, d, N)$$

6. Statistics of Search Engine Suggests

For each topic z_n, we collect search engine suggests for a query q_j ($j = 1, \ldots, 10$) out of the 10 company names into the set $\mathbb{S}(z_n, q_j, N, \theta_{lbd})$. We analyze those sets $\mathbb{S}(z_n, q_j, N, \theta_{lbd})$ ($j = 1, \ldots, 10$) of collected search engine suggests. Then, we compare their statistics among the 10 company names $q_1, \ldots, q_{10}$. For each query company name q_j, we measure the following rate of search engine suggests within the topic z_n and consider them as concerns on those companies in the certain product genre represented by the topic z_n:

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

In the case of 10 company names we examine in this paper, out of the total 80 topics generated by the topic modeling procedure, we observed 23 topics that can be clearly regarded as certain product genres. Among those 23 topics on certain product genres, we pickup that on TV and related product genres as well as that on PC and related product genres. For those two topics, we further examine the lower bound θ_{lbd} of the probability $P(z_n|d)$ as 0 $\sim$ 0.5, and select $\theta_{lbd} = 0.4 \sim 0.5$ for TV and related product genres and $\theta_{lbd} = 0.1 \sim 0.4$ for PC and related product genres, respectively. These lower bounds are selected so that they result in the highest correlation between statistics of search engine suggests and actual market share at kakaku.com.

7. Correlation of Statistics of Search Engine Suggests for Company Names and Market Share

We next analyze whether rates of concerns of those who search for Web pages presented in the previous section have certain correlation with actual market share at kakaku.com. Furthermore, as an intermediate statistics between the rates of concerns of those who search for Web pages and market share, we also examine the page view statistics[12] at the kakaku.com site and its correlation with the other two statistics. More specifically, for TV product genre, Figure 1 shows market share as well as page view statistics at the kakaku.com site. Roughly speaking, those two statistics have certain correlation with the rates of concerns of those who search for Web pages, which are represented as statistics of search engine suggests shown in Figure 1. Actually, as shown in Table 2, we measure time series changes of Pearson product-moment correlation coefficient for about five months periods. For TV domain, they continue to correlate very well, while for PC domain, their correlation is relatively lower, although they correlate fairly well. These results support the claim that rates of concerns of those who search for Web pages contribute to estimating actual market share in product genres such as electronics domain.

8. Predicting Market Share based on Concerns of Those Who Search for Web Pages

This section describes how to predict market share statistics as well as page view statistics at the kakaku.com site based on rates of concerns of those who search for Web pages. The overall procedure is illustrated in Figure 2. Suppose that, given a certain products genre, we predict market share statistics or page view statistics of N-th month based

[12] https://ssl.kakaku.com/trendsearch/index.asp

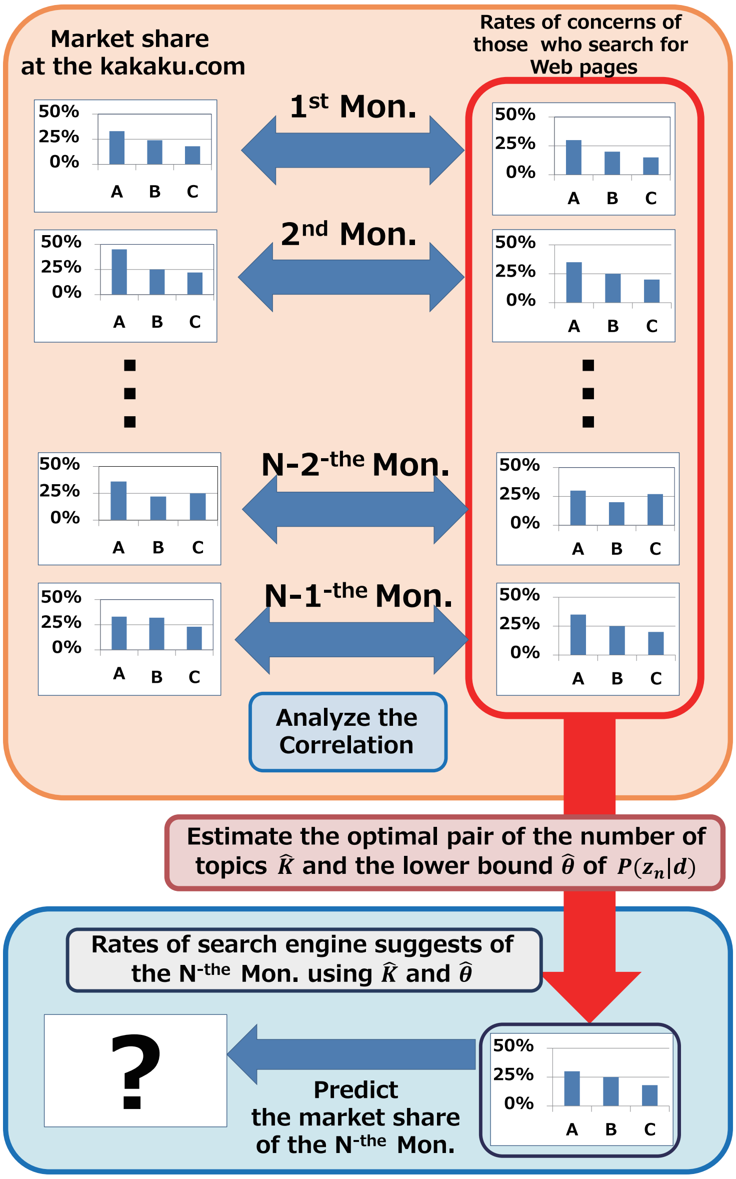
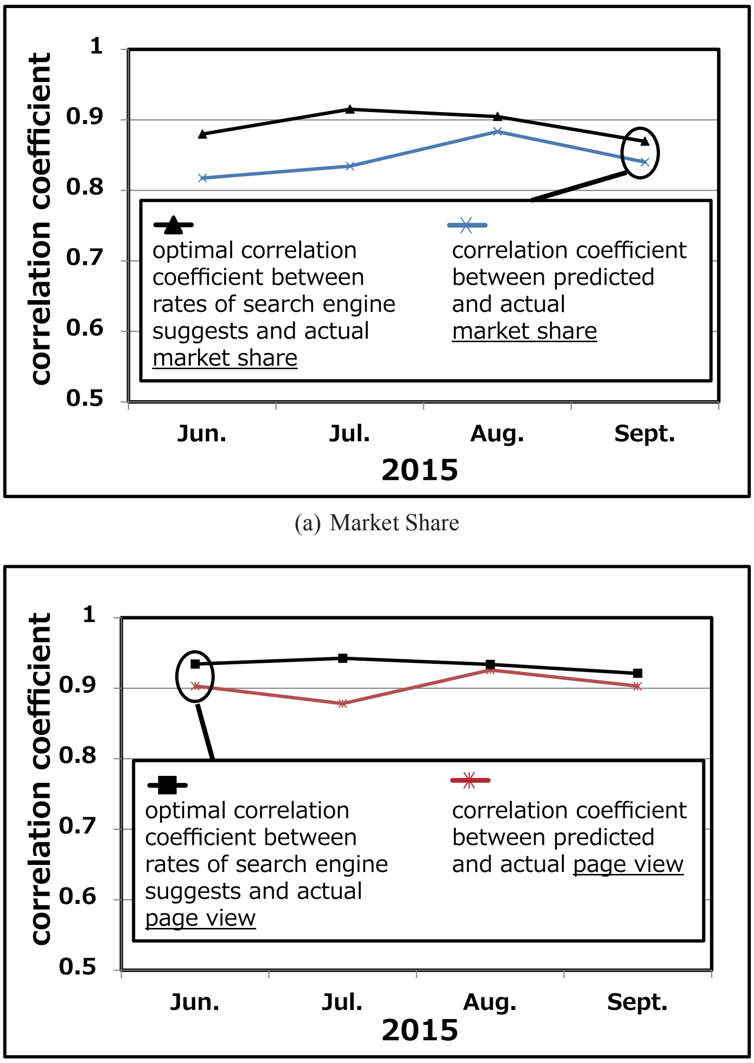

Market share
at the kakaku.com
Rates of concerns of those who search for Web pages
1st Mon.
2nd Mon.
N-2-the Mon.
N-1-the Mon.
50%
25%
0%
A B C
Analyze the Correlation
Estimate the optimal pair of the number of topics $\hat{K}$ and the lower bound $\hat{\theta}$ of $P(z_n|d)$
Rates of search engine suggests of the N-the Mon. using $\hat{K}$ and $\hat{\theta}$
?
Predict
the market share
of the N-the Mon.

Figure 2: Procedure of Predicting Market Share based on Concerns of Those Who Search for Web Pages

on market share statistics or page view statistics of from 1st to $N-1$-th months as well as rates of concerns of those who search for Web pages of from 1st to N-th months. In order to realize this, for each month from 1st to $N-1$-th months, we collect market share statistics as well as page view statistics. For each month, we also collect search engine suggests and Web pages for the 10 company names listed in Table tab:suggest, and then, for the given products genre, calculate the rates of concerns of those who search for Web pages for various number K of topics as well as the lower bound θ_{lbd} of the probability $P(z_n|d)$. Next, we optimize the number K of topics as $\hat{K}$ and the lower bound θ_{lbd} of the probability $P(z_n|d)$ as $\hat{\theta}_{lbd}$ by maximizing the average of the sum of the correlation coefficient between the market share statistics and the rates of concerns those who search for Web pages and that between page view statistics and the rates of concerns of those who search for Web pages. Finally, we use the rates of those who search for Web pages at the N-th month with the optimal $\hat{K}$ and $\hat{\theta}_{lbd}$ as the prediction of the market share statistics and the page view statistics of the N-th month.

For the products genres of TV and related products genres as well as PC and related product genres, Figure 3 and Figure 4 plot the results of prediction in terms of the correlation coefficient between predicted statistics and the actual market share / page view statistics. Prediction results for the following months are plotted in Figure 3 and Figure 4.

- predicting the N-th month as June, 2015 by optimizing K and θ_{lbd} from March to May, 2015.
- predicting the N-th month as July, 2015 by optimizing K and θ_{lbd} from March to June, 2015.
- predicting the N-th month as August, 2015 by optimizing K and θ_{lbd} from March to July, 2015.
- predicting the N-th month as September, 2015 by optimizing K and θ_{lbd} from March to August, 2015.



(a) Market Share

(b) Page View

Figure 3: Time Series Changes of Correlation of Actual Market Share / Page View and their Prediction based on Search Engine Suggests (TV and related Products Genres)

For comparison, in Figure 3 and Figure 4, we also show optimal correlation coefficient between the actual market share statistics and the rates of concerns of those who search for Web pages, as well as between the actual page view statistics and the rates of concerns of those who search for Web pages, by simply selecting K and θ_{lbd} which maximize those correlation coefficients. Those results show that the predicted statistics have comparatively high correlation against the actual statistics of market share and page view.

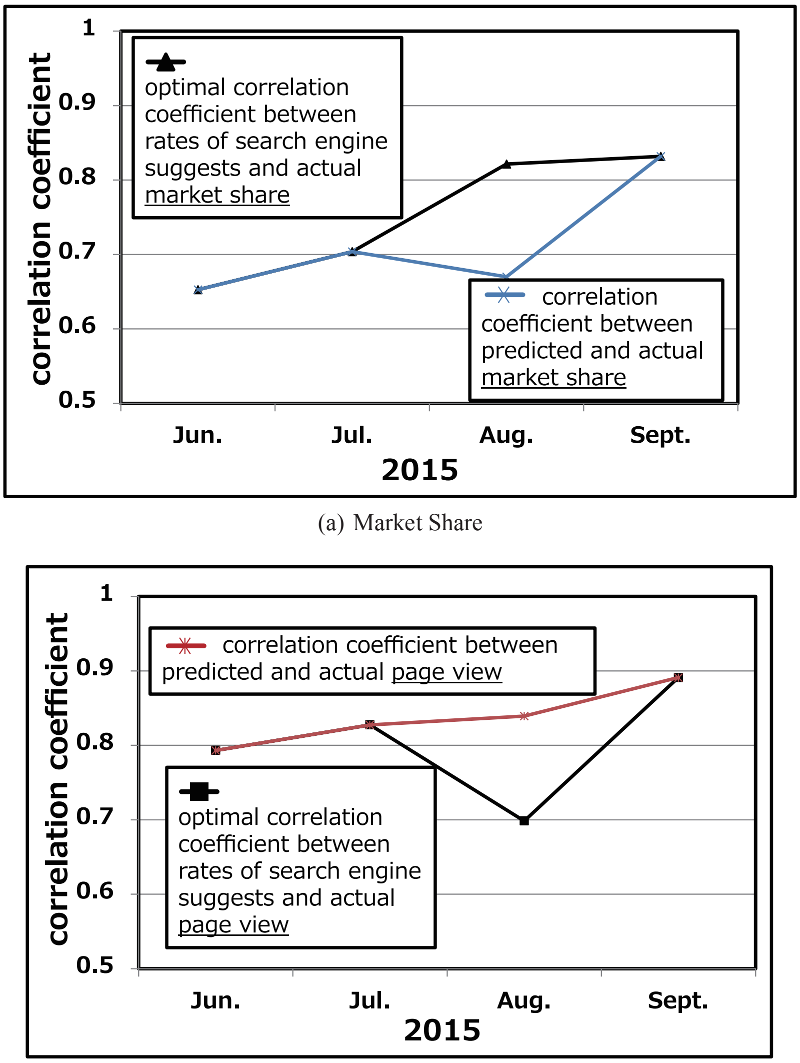
9. Related Work

Related work include a technique of detecting influenza epidemics from Twitter (Aramaki et al., 2011), that of predicting stock market from sentiment analysis of Twitter (Bollen et al., 2011), that of predicting movie ranking based on Twitter analysis before the release of the movie (Asur and Huberman, 2010), and that of predicting stock market based on Wikipedia page view statistics (Moat et al., 2013). Those previous related methods predict changes in real world based on statistics available through Internet such as that of Twitter and Wikipedia page view. In the method proposed in this paper, on the other hand, it is shown that real world statistics such as market share within certain product genres can be predicted based on rates of concerns of those who search for Web pages, which are measured by collecting search engine suggests of company names.

The work presented in this paper is based on the our previous framework (Moriya et al., 2015), which studied how to overview the knowledge of a given query keyword through search engine suggests. Based on this previous work, this paper studied how to compare rates of concerns of those who search for Web pages among several companies which supply products, given a specific products domain.

10. Conclusion

This paper proposed how to utilize a search engine in order to predict market shares. We proposed to compare rates of concerns of those who search for Web pages among several companies which supply products, given a specific prod-



(a) Market Share

(b) Page View

Figure 4: Time Series Changes of Correlation of Actual Market Share / Page View and their Prediction based on Search Engine Suggests (PC and related Products Genres)

ucts domain. We measure concerns of those who search for Web pages through search engine suggests. Then, we analyzed whether rates of concerns of those who search for Web pages have certain correlation with actual market share. We showed that those statistics have certain correlations. We finally proposed how to predict the market share of a specific product genre based on the rates of concerns of those who search for Web pages. Future work includes scaling up the proposed method by applying it to all of the 23 topics mentioned in section 1., that can be regarded as product genres.

11. Bibliographical References

Aramaki, E., Masukawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proc. 2011 EMNLP*, pages 1568–1576.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proc. WI-IAT*, pages 492–499.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3(1801).

Moriya, I., Inoue, Y., Imada, T., Utsuro, T., Kawada, Y., and Kando, N. (2015). Overviewing the knowledge of a query keyword by clustering viewpoints of Web search information needs. In *Proc. 29th WAINA*, pages 535–540.