International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2020, Vol. 4, No. 2, 1 – 18

Analyzing Searching Behavior in Online Shopping Sites based on Product-Specificity of Query Words

Kei Wakabayashi *, Genkou Ou *, Tetsuji Satoh *

Abstract

As the internet continues to spread as a crucial element of social infrastructure, more and more people are shopping online. Online sites that formerly dealt with such specific products as books and clothing have expanded to mall-type shopping sites by incorporating various kinds of stores. This transition has made product searches by users more complicated and prolonged. In this paper, we propose a method that analyzes the transition patterns of the product-specificity of queries in a product-searching behavior. As a key concept in our proposed method, we adopt the notion of information content, which represents the amount of information contained in a query, to quantitatively define product-specificity. We conducted an experiment on an actual shopping log dataset to confirm the effectiveness of our proposed method. The result demonstrates that the proposed method extracts illuminating behavioral patterns such as "narrowing-down behavior" that keeps adding query words and "expanding behavior" that keeps removing query words to increase the search results.

Keywords: Searching Behavior, Online Shopping, Information Content, Log Analysis

1 Introduction

Due to the spread of online shopping sites, opportunities to purchase products from online sources continue to increase. If such online sites were actual stores, desired products could be narrowed down before being purchased based on consultation with a clerk. If a user's desired product is not clear, people can window-shop at several stores to find items that satisfy their shopping desires. However, in online shopping, all purchasing behaviors must be performed through such network terminals as personal computers and smartphones. Users iterate between narrowing down product choices by keyword searches and checking products by browsing product lists on screens. Unlike experiences with informative, intelligent store clerks, unless an adequate keyword is given, the number of products in the search result may be too large or too small for browsing, inconveniencing the user [19]. Therefore, determining efficient, useful search keywords is a critical requirement in purchasing behavior in online shopping.

Search keywords include language of various product-specificity: words that refer to a specific product or brand, such as "iPhone8," words that indicate a product category, such

* University of Tsukuba, Ibaraki, Japan

as “smartphones” and “winter coat,” and words that co-occur with all other words, such as “free shipping” and “outlet products.” Users enter such search keywords alone or in a phrase as a query and modify it while viewing the obtained results until they find their desired product.

In this paper, we propose a method that characterizes searching behavior by examining the transition of the product-specificity levels of search queries issued by a user. Our proposed method's key concept is a query's “information content,” which represents the amount of information contained in a given query. We adopt information content as a measure of a query's product-specificity to characterize the searching behavior. We define a sequence of purchasing queries continuously input by users as a session and propose a method that analyzes the item-searching behavior of users from the pattern of the amount of query information content in it.

A common search word frequently entered by many users only has slight information content value. Contrarily, a search word that is occasionally entered by a specific user, for example, the name of a specific product model, tends to have higher information content value. We experimentally examined the effectiveness of our proposed method using an actual dataset. In this paper, we report that the proposed method identified interesting searching behavior tendencies, which include “narrowing-down behavior” that keeps adding query words and “expanding behavior” that keeps removing them to increase the search results.

The rest of the paper is organized as follows. In Section 2, related work is described to clarify the position of this paper. In Section 3, we propose a method that analyzes item-searching behavior based on the product-specificity of queries. In Section 4, we show our experimental results that applied our proposed method to an actual search log in an online shopping mall. In Section 5, we scrutinize our results and draw insights. We describe our conclusion in Section 6.

2 Related Work

Many studies have analyzed user search behaviors for applications that provide navigational support [12, 20, 25] and arrange personalized search results [18, 27, 28, 29]. Many insights have been particularly reported in web search behavior [10, 13]. Broder [8] classified web search queries into three classes: *navigational query*, which directly goes to a specific website; *informational query*, which seeks information; and *transactional query*, which wants to take an action like a purchase or a download. Jansen and Booth [15] added a second-level classification of four categories to Broder's taxonomy: *directed queries*, which directly ask a question; *undirected queries*, which only specify a broader topic; *list queries*, which list something; and *find queries*, which seek a website that satisfies the user. Levene and Loizou [17] focused on continuous web interaction (i.e., sessions) during web surfing and suggested that users often engage in confusing search behaviors. They developed a method that calculates the extent to which a user's search session has a specific direction for its purpose based on the entropy of a Markov chain of the visited websites. Although these insights are related to the analysis of user behaviors on online shopping sites, mismatches do occur because online shopping is more specific than general web searches and users behave differently.

Analysis that focuses on user behaviors at online shopping sites is also useful for sending appropriate messages or supporting site navigation by stores [20]. Other studies [20, 24]

reported that user behaviors in online shopping can be classified into the following patterns: a *buying pattern* where users directly go to purchase a specific item; a *browsing pattern* where users compare and consider what to buy; a *searching pattern* where users look for something to buy; and a *knowledge-building pattern* where users collect information about a shop's items. They confirmed that their classification works well by analyzing actual data and applying an unsupervised clustering method. In contrast, Sondhi et al. [26] applied a data-driven approach and derived five categories from a clustering analysis: *shallow exploration*, *major-item shopping*, *targeted purchases*, *minor-item shopping*, and *hard-choice shopping*. A session's features, which are used in the clustering method in these analyses, are based on such action types as the number of query searches, the number of page views, and the number of times the item categories are changed. In this paper, we focus on the transition of the contents of queries that are issued in a session and provide new insight from a different perspective compared with the classification in these studies.

Some Studies including ones by Huang and Efthimiadis [11], Juiang and Ni [16], and others [1, 5, 14, 23] focus on query reformulations in search sessions. In these studies, query reformulations (e.g., adding or replacing words) are classified from multiple aspects and associated with contextual information such as user's satisfaction, the number of search results, etc. In contrast to this line of work that classifies the query formulation itself, we attempt to classify whole search sessions by using features based on the series of query reformulations. The work that is the most similar to ours is one by Nozaki and Satoh [21], which proposed a method that classifies search sessions in online shopping sites by using time-series features of search actions such as issuing a query and clicking a search result. The important difference of our work from Nozaki's [21] is that we look into the query contents at word level. By making use of the word-level features, our proposed method draws a new insight from search logs in the online shopping domain.

The existing studies also suggest that a query's keywords play a significant role in a session. Zhai et al. [31] applied an unsupervised parsing algorithm to queries made at a shopping site and suggested that keywords can be naturally classified into several roles. Some studies such as ones by Cummins [9], Barathi and Shanmugam [3] apply topic models to search queries for extracting topically coherent word groups. We analyze the transition of keywords and focus on the product-specificity of queries that deliver different insights from the previous analysis.

3 Session Analysis based on Product-Specificity

3.1 Overview

We assume that a large amount of query logs is available for analyzing user purchase behavior. Each query log is a tuple of timestamp t, user ID u, and query q. q is regarded as a set of words (e.g., $q = \{\text{usb, 64GB}\}$). We denote the set of all possible words in a dataset by $\mathscr{V}$ so that q is expressed as a member of $2^{\mathscr{V}}$.

Our proposed method (Figure 1) is composed of the following steps:

- calculate the information content of the query word;
- extract search sessions;
- calculate the query's information content;
- classify the session behavior.

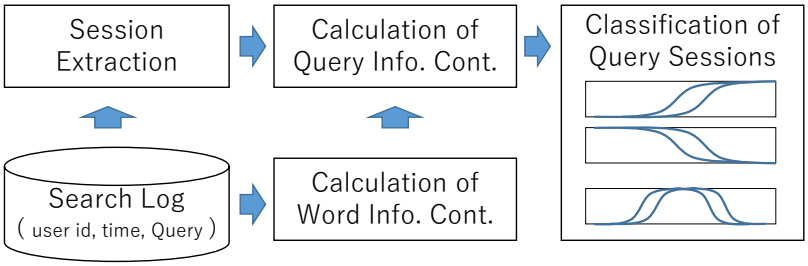

Figure 1: Method overview

3.2 Information content of query words

An event's information content is a measuring index that represents the amount of information gained from its observation. The information content of event e is defined as $-\log p(e)$, where $p(e)$ is its probability. When $p(e)$ is small, which means event e is rare, event e will get information content with a higher value. Conversely, the information content of a frequent event will have lower value. In the proposed method, we apply information content to query words by regarding their occurrence as a probabilistic event. This is because the probability of query words is closely related to product-specificity that we are analyzing in this paper. Query words that are frequently input by many users are general words for item searches in shopping malls. On the other hand, words that are seldom entered are specific terms that identify an item. Another important reason for applying information content is that it can be added when multiple words are used in a query.

Information content $I(w)$ of word $w \in \mathcal{V}$ is calculated by the following equation:

$$I(w) \quad = \quad -\log p(w). \tag{1}$$

The probability of word w is estimated by maximum likelihood estimation using an entire log dataset, as follows;

$$p(w) \quad = \quad \frac{n(w)}{\sum_{w' \in \mathcal{V}} n(w')},$$

where $n(w)$ is the number of occurrences of word w.

The distribution of the query words in the logs is shown in Figure 2, where the entire tendency follows a power law. Almost $1,000$ words were input more than $5,000$ times, and about $100,000$ words were input over 30 times.

The definition of information content is essentially the same as inverse document frequency (IDF), which is usually employed in information retrieval and text mining. We prefer to call this quantity information content because we need to deal with queries that contain multiple words. As we explain in section 3.4, we can treat a query as an event where multiple words jointly occur whose information content can be defined in a more natural way than by combining multiple IDF values of different words. However, since these two notions are mathematically identical, mentioning them gives a different perspective for the proposed method.

3.3 Extraction of search sessions

A search session is a series of search actions for reaching a particular item desired by a user. Search sessions are extracted from search logs by the following process [6, 7]:

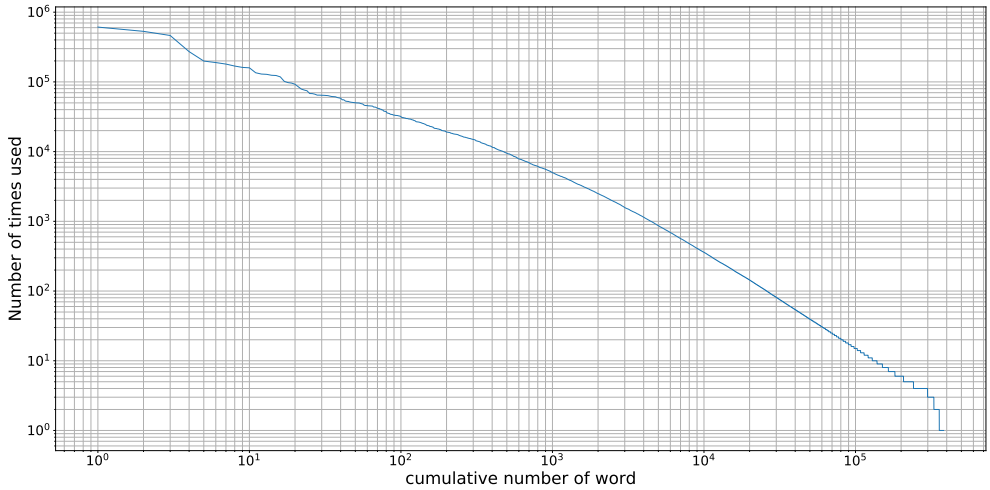

Figure 2: Number of words used

- classify the search logs by user IDs;
- sort the classified search logs by time;
- scan the sorted search logs.
  - If their time difference is less than 30 minutes, they are categorized as identical sessions.
  - If it exceeds 30 minutes, they are treated as separate sessions.

We denote the set of extracted sessions by S and each session by $s = (t_1, u_1, q_1), \ldots, (t_{L_s}, u_{L_s}, q_{L_s})$ where $t_i < t_{i+1}$ and $u_i = u_j$ for all $1 \leq i < L_s$ and $1 \leq j \leq L_s$. L_s indicates the length of session s. Table 1 shows an example of the extraction of search sessions.

Table 1: Example of search sessions

User ID	Time stamp	Query	Session ID
00001	2017-04-06 16:36:12	apple	1
00001	2017-04-06 16:36:22	apple juice	1
00001	2017-04-06 16:39:22	apple juice 500 ml	1
00001	2017-04-06 16:42:04	apple juice	1
00001	2017-04-06 16:43:31	fruit juice	1
00001	2017-04-06 17:24:27	usb	2
00001	2017-04-06 17:26:19	usb 64 GB	2
00001	2017-04-06 17:37:58	usb 128 GB	2
00002	2017-04-07 08:38:13	cola	3
00002	2017-04-07 08:38:43	cola 2L	3
⋮	⋮	⋮	⋮

3.4 Information content of query

The information content of query $q \in 2^{\mathcal{V}}$ is naturally defined as the sum of the information content of each word $I(w)$ where $w \in q$. The information content of query $I(q)$ is defined:

$$I(q) = \sum_{w \in q} I(w). \tag{2}$$

An example of the information content of queries is shown in Table 2. In the proposed method, we consider the sequence of the information content of queries as a feature of a session and characterize the sessions based on their sequential patterns. The feature of session $s = (t_1, u_1, q_1), \ldots, (t_{L_s}, u_{L_s}, q_{L_s})$ is denoted by $f(s)$:

$$f(s) = I(q_1), \ldots, I(q_{L_s}) \tag{3}$$

We apply time-series clustering to the set of features $\{f(s)|s \in S\}$ to determine the patterns of the session behaviors as we explain in the next section.

Table 2: Example of information content of query

User ID	Session ID	Query	$I(q)$
00001	1	apple	3.1110
00001	1	apple juice	5.1120
00001	1	apple juice 500 ml	10.1125
00001	1	apple juice	5.1120
00001	1	fruit juice	4.5450
00001	2	usb	1.2569
00001	2	usb 64 GB	5.4458
00001	2	usb 128 GB	5.7569
00002	3	cola	0.5456
00002	3	cola 2L	1.7126
⋮	⋮	⋮	⋮

3.5 Clustering of session behavior

We apply a time-series clustering to sessions to form groups that have a similar sequential pattern of the information content of queries. For the clustering of sessions, we adopt the k-shape algorithm [22], which is a robust and efficient time-series clustering method based on a shape-based feature. Before the clustering, we apply the following preprocessing steps that include sampling, padding, and normalizing:

- **Sampling**: Our preliminary observation of the sequence of the information content revealed primary patterns that prevent effective clustering of sessions. The most dominant pattern is sessions that have few or no changes in the query words during the search session. To avoid domination, we selectively sample sessions from groups derived by a simple analysis to examine the trend of increasing or decreasing the information content values. We apply linear regression to each sequence of the information content values to characterize the session by the slope of the regression line. The following is the sampling process:

  - For calculating slope feature $g(s)$ for session s, linear regression is applied to a set of points $\{(i, I(q_i)) | 1 \leq i \leq L_s\}$ (i.e., each point is $(x^{(i)}, y^{(i)}) = (i, I(q_i))$ and the number of data points is L_s). The slope of the obtained regression line is used as the value of $g(s)$.
  - We classify the sessions into groups by the slope feature $g(s)$. The groups are determined based on the distribution of the slope values for all the sessions in the dataset, $\{g(s) | s \in S\}$.
  - We randomly choose the same number of sessions from each group.

  $g(s) = 0$ indicates that the user browsed the site without changing the queries during session s. By sampling sessions from groups that have different ranges of $g(s)$, we expect to find salient sequential patterns more easily using time-series clustering.

- **Padding**: To apply the k-shape algorithm, the sequences of all the sessions must be the same length. We align the length of the sessions by padding the information content of the last query in the session at the end of the sequence until its length equals the longest session $\max_{s \in S} L_s$. For example, since the longest session in Table 2 has length 5, the session of Session ID 2, which originally has a sequence of information content $(1.2569, 5.4458, 5.7569)$, is extended with padding: $(1.2569, 5.4458, 5.7569, 5.7569, 5.7569)$.

- **Standardization**: Since the scale of the magnitude of $I(q)$ is varied over the sessions, we standardize the information content so that the mean is zero and the standard deviation is one. Given a sequence of the information content of queries $I(q_1), \ldots, I(q_{L_s})$, the sequence of standardized values $z_1, \ldots, z_{L_s}$ is obtained:

$$z_i = \frac{I(q_i) - \mu_s}{\sigma_s}, \tag{4}$$

  where $\mu_s = \frac{1}{L_s} \sum_{i=1}^{L_s} I(q_i)$ and $\sigma_s^2 = \frac{1}{L_s} \sum_{i=1}^{L_s} (I(q_i) - \mu_s)^2$.

After preprocessing the dataset, we apply the k-shape algorithm to the selected set of sessions for clustering. The k-shape algorithm is a time-series clustering method based on a shape-based distance, which is robust against various kinds of distortion [22]. By obtaining the clusters of the sequences of information content, we expect to analyze the patterns of searching behavior from the perspective of product-specificity transition. In the next section, we empirically examine the effectiveness of the proposed method.

4 Experimental Results

4.1 Dataset

We applied our proposed method to actual data and analyzed the user searching behaviors with the log data of Ponpare Mall, a Japanese online shopping site provided by Recruit Technologies Co., Ltd. Identical items are sold in different stores. The period of the logs ran from January 1, 2016 to December 30, 2017. The logs have $24,582,912$ queries in its data. During the data period, $2,460,782$ users accessed the Ponpare Mall. The number of queries input by users is shown in Figure 3. The horizontal axis represents the cumulative number of users. About $10,000$ users input more than 200 queries, and about $100,000$ users input more than 3 queries.

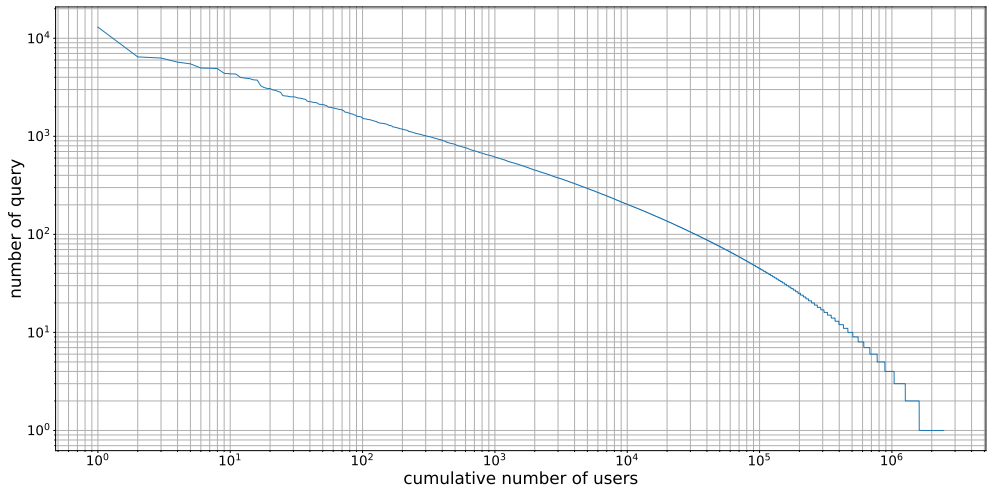

Figure 3: Number of queries input by users

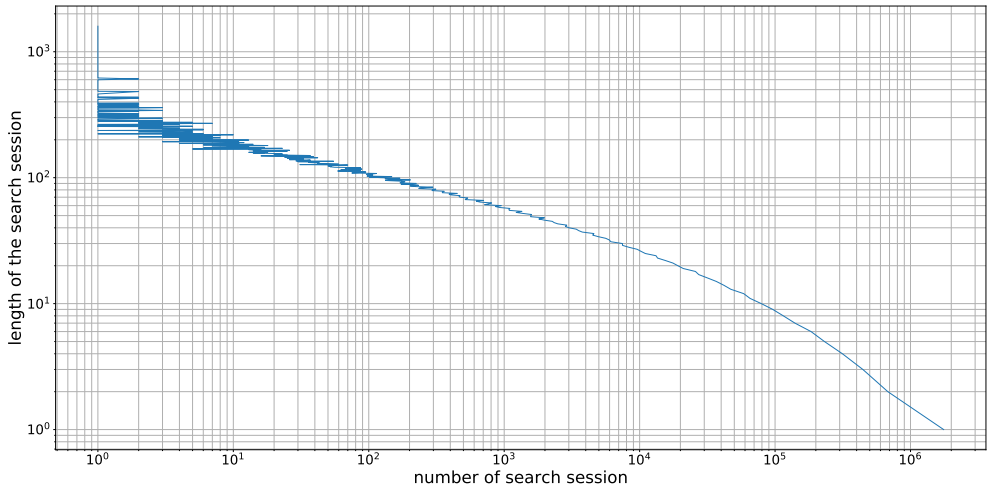

Figure 4: Length of search sessions

We extracted 4,628,521 search sessions by the method described in Section 3.3. The length of the sessions is widely diverse, as Figure 4 shows. The number of short sessions is larger than the number of long sessions.

We excluded sessions that were too long or too short from subsequent analysis to avoid deleterious effects caused by padding. If the dataset to be analyzed includes sessions whose length is too diverse, much shorter sessions than the long sessions will suffer from a loss of sequential characteristics because of too many padded values. For this reason, we used search sessions whose length ranges from 10 to 15. We found 316,325 such search sessions.

We classified these sessions into six groups based on the slope of the regression line (i.e., session's slope feature), as explained in Section 3.5. The slope feature's distribution is shown in Fig. 5. We determined the boundaries of the groups as -2.5, -1, 0, 1, and 2.5, which means we made the following six groups of sessions: $g(s) < -2.5$, $-2.5 \leq g(s) < -1$, $-1 \leq g(s) < 0$, $0 \leq g(s) < 1$, $1 \leq g(s) < 2.5$, and $2.5 \leq g(s)$. We excluded the sessions of $g(s) = 0$ in advance and randomly chose 250 sessions from each group.

To implement the linear regression, we used the implementation in the scikit-learn li-

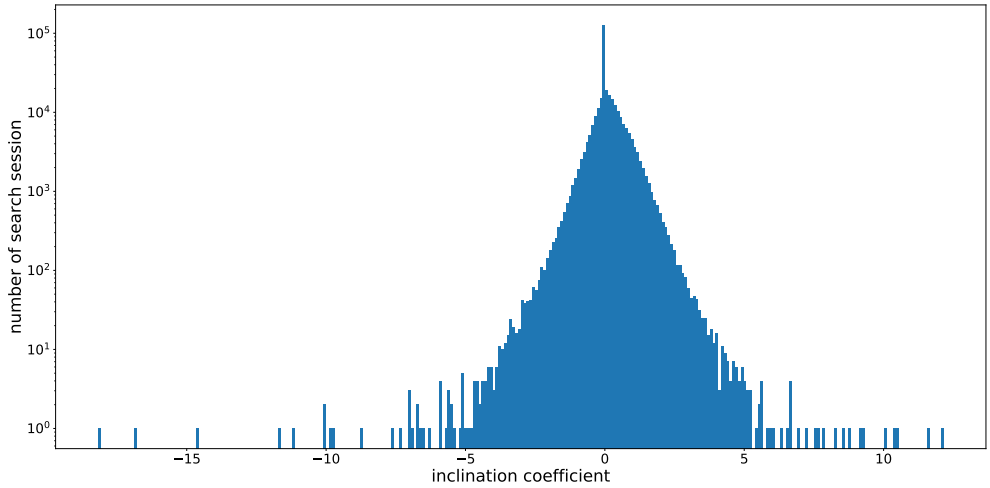

Figure 5: Distribution of slope of regression line

brary[1]. For the k-shape algorithm, we used the implementation in the tslearn library[2]. For executing the k-shape algorithm, we set parameter k, which indicates the number of clusters. We applied the k-shape algorithm with different k values to the dataset in a preliminary experiment. As a result, we observed that k does not significantly affect the explainability of the session clusters. In this paper, we report the result details for $k = 6$ to demonstrate the effectiveness of the proposed method.

4.2 Result

Figure 6 shows the sessions grouped into six clusters. The horizontal axis indicates the time points that correspond to the number of queries in a session. The vertical axis indicates the query's standardized information content issued at the corresponding time point. Each line in the figure expresses a trajectory of the information content values in a single session. The value in parentheses is the number of sessions in the cluster.

By looking at Figure 6, the information content values in clusters 3, 5, and 6 tend to increase, and the values in clusters 1, 2, and 4 tend to decrease. The characteristics of each cluster are described below:

- **Cluster 1**: 265 sessions belong to this cluster. The cluster has many sessions showing zig-zag shape, which involves both the increase and decrease of information content values, in comparison with other clusters. Since the information content represents the product-specificity of the queries, this kind of session may indicate that the user changed the mind during it.
- **Cluster 2**: 53 sessions belong to this cluster. Most sessions in this cluster contain a very high value of information content that exceeds 3. This means that the users of these sessions searched with a very specific query but changed to a more common query, perhaps because they failed to find what they wanted.

[1] `https://scikit-learn.org`
[2] `https://github.com/rtavenar/tslearn`

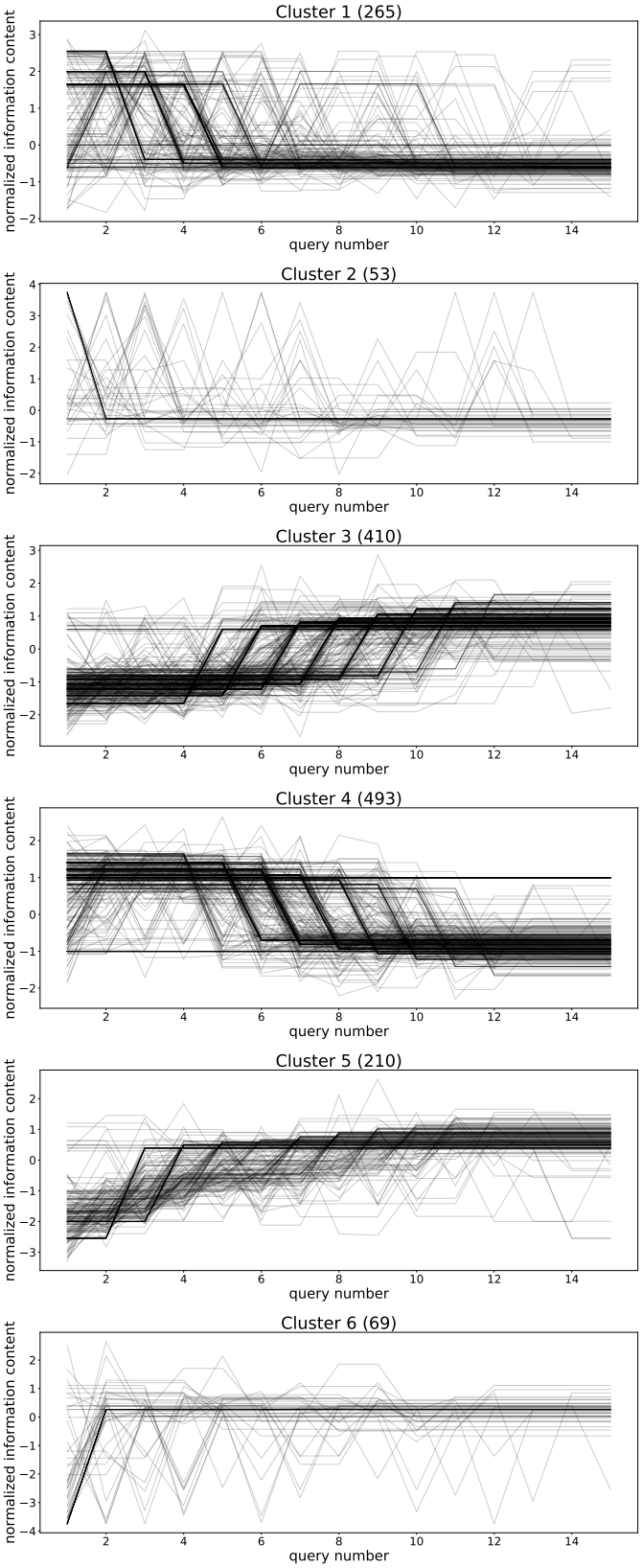

Figure 6: Clustering result

- **Cluster 3**: 410 sessions belong to this cluster. In most sessions in it, the information content values monotonically increase, which probably indicates that the user kept narrowing down the search result to find a specific product by changing the query.
- **Cluster 4**: 493 sessions belong to this cluster. It contains sessions that have the opposite trend to sessions in cluster 3. The decrease of the information content is typically an action that broadens the search result with a more general query.
- **Cluster 5**: 210 sessions belong to this cluster. Although it has an increasing trend in common with cluster 3, its information content grows in an earlier phase. Perhaps the user suddenly found better keywords in these sessions.
- **Cluster 6**: 69 sessions belong to this cluster. Its salient feature is the lowest value of information content, which is less than -3, denoting that a very general query was issued.

5 Discussion

5.1 Average trajectories

We plotted the average trajectories of the information content values in each cluster, as Figures 7 and 8 show. Figure 7 shows the average trajectories of clusters 3, 5, and 6, which are the "up group" that has an increasing trend of information content values. Figure 8 shows the average trajectories of the "down group" that has a decreasing trend.

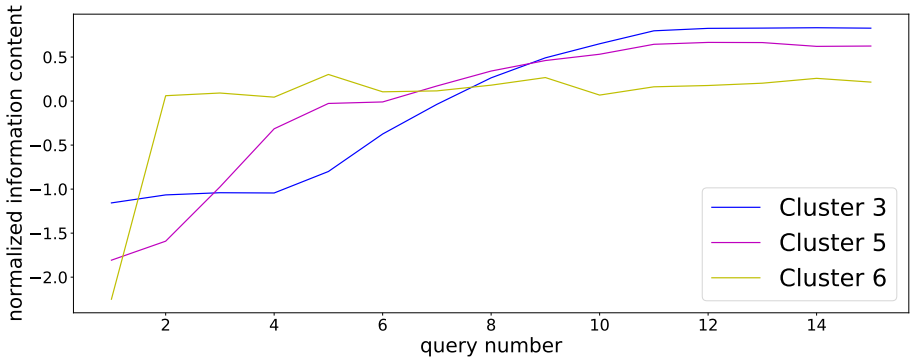

Figure 7: Average trajectories of the information content values in each cluster (up group)

As shown in Figure 7, the first query's average value of the information content of cluster 6 is the lowest in the up group, which means its sessions tend to begin with a general query. Additionally in cluster 6, the average time point at which the information content goes up is the earliest in the up group, which implies that its users quickly changed their queries into highly product-specific ones. On the other hand, the information content value in the average trajectory of cluster 3 slowly increases, which indicates that its users gradually changed their queries toward more product-specific ones. The trajectory of cluster 5 goes up earlier than the trajectory of cluster 3, which is the same trend seen in Figure 6.

According to Figure 8, the first query's average value of the information content of cluster 2 is the highest in the down group. This indicates that its sessions tended to begin

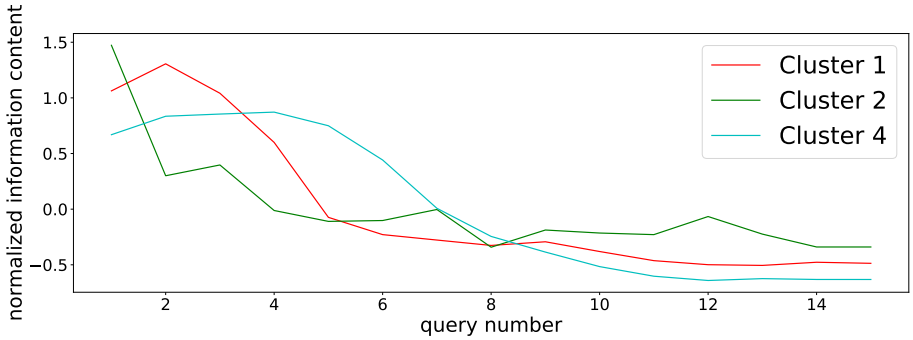

Figure 8: Average trajectories of the information content values in each cluster (down group)

with a highly specific query. Its average trajectory also goes down quickly, which means its users made the query more general in an early phase of the sessions. The information content value in the trajectory of cluster 4 decreased slowly, which implies that the users gradually changed the queries into more general ones.

We focused on the relationship between the information content values of the first and last queries. According to Figure 7, the average information content values of both the first and last queries in cluster 6 are the lowest in the up group, and both values in cluster 3 are the highest in the group. Figure 7 shows that the order of the value heights of the information content between the first and last queries are identical. The order of the values in the down group are also identical (Figure 8). The average information content of all the clusters is shown in Figure 9. The clusters in the up group are shown as red lines, and blue lines indicate those in the down group. The last query's information content of the sessions in the down group is lower than those in the up group.

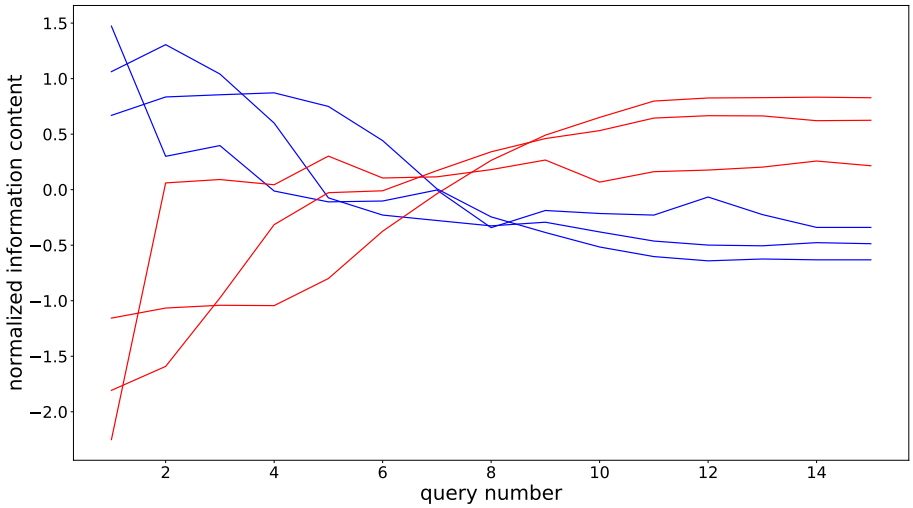

Figure 9: Average trajectories of the information content values in each cluster

Table 3: Example of session in cluster 3

Query	$I(q)$
xperia z5	10.74
xperia z5 case	16.35
xperia z5 case	16.35
xperia z5 case	16.35
xperia z5 so-01 case	28.15
xperia z5 so-01 case leather	35.75
xperia z5 so-01 case leather	35.75
xperia z5 so-01 case leather	35.75
xperia z5 so-01 case leather	35.75
xperia z5 so-01 case leather	35.75

Table 4: Example of session in cluster 4

Query	$I(q)$
flannel rug high-resistance	24.65
flannel rug high-resistance	24.65
flannel rug high-resistance	24.65
flannel rug high-resistance	24.65
high-resistance rug	17.77
high-resistance rug	17.77
high-resistance rug	17.77
high-resistance rug	17.77
high-resistance rug	17.77
rug	8.10

Table 5: Example of session in cluster 5

Query	$I(q)$
bargain	5.20
toothbrush	8.97
toothbrush spare	20.09
toothbrush spare	20.09
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74
toothbrush spare brush	28.74

Table 6: Example of session in cluster 1

Query	$I(q)$
wallet	6.28
wallet men's	11.15
wallet men's leather	20.87
wallet men's leather	20.87
wallet men's bi-fold	18.91
wallet men's bi-fold	18.91
wallet men's	11.15
wallet men's	11.15
wallet bi-fold	14.04
wallet men's	11.15
wallet	6.28
wallet bi-fold	14.04

5.2 Instance-based analysis

To confirm that the interpretations of the clusters' characteristics reflect the actual situations of the searching behaviors, we scrutinized the sessions in each cluster. Table 3 shows an example of a session in cluster 3, which is one of the largest clusters. In this session, the user seems to be looking for a smartphone case. At first, only the name of the smartphone model, "xperia z5," was provided, and then "case" is quickly added to the query. After a while, the user discovered a more specific model number, "so-01," and a keyword that specifies its material, "leather," is added. By looking at Figure 6, we can confirm that the sessions in cluster 3 tend to have increasing information content value (i.e., product-specificity), which is consistent with the query transitions in the example. Many other sessions in the cluster have the same trend that narrows down the search results by adding words to the queries.

Table 4 shows a session in cluster 4, which is another large cluster. The session begins with a specific type of rug with keywords that specify its material ("flannel") and hardness ("high-resistance"). However, the user does not seem satisfied with the products displayed and gradually removes the words to expand the search results with broader conditions. This behavior is well described by the trend in cluster 4 shown in Figure 6. Such expanding

Table 7: Example of session in cluster 6

Query	$I(q)$
carpet	9.80
carpet 3 square meters wood flooring	30.43
carpet 3 square meters wood flooring	30.43
carpet 3 square meters wood flooring	30.43
carpet 3 square meters wood flooring	30.43
carpet 3 square meters wood flooring	30.43
carpet 3 square meters wood flooring	30.43
carpet 5 square meters wood flooring	30.49
carpet 5 square meters wood flooring	30.49
carpet 5 square meters wood flooring	30.49
carpet 5 square meters wood flooring	30.49
carpet 5 square meters wood flooring	30.49
carpet 5 square meters wood flooring	30.49

Table 8: Example of session in cluster 2

Query	$I(q)$
nintendo 2ds yellow ftr-s-yadn	41.58
nintendo 2ds	12.70
nintendo 2ds yellow ftr-s-yadn	41.58
nintendo 2ds	12.70
nintendo 2ds	12.70
nintendo 2ds	12.70
nintendo 2ds	12.70
nintendo 2ds	12.70
nintendo 2ds	12.70
nintendo 2ds	12.70

behavior that broadens the product-specificity of a search can also be observed in many other sessions in the cluster.

Table 5 shows an example of a session in cluster 5. The session starts with a very general query, “bargain,” without specifying any product. The user looked at the first search result and perhaps remembered that a toothbrush is needed, and then added more specific words “spare brush” to the query. The salient feature of cluster 5 in the plot of Figure 6 is a trend that increases the information content in an earlier phase. The example explains why the information content rises so quickly: a lack of a concrete idea about what to purchase. Many similar examples can be found in cluster 5.

Table 6 shows an example of a session in cluster 1. The user keeps adding and removing keywords, probably because he or she does not have a strong preference for wallet type. As a result, the information content value fluctuates. Such searching behavior can often be observed in cluster 1 and shows consistent characteristics in the plot of Figure 6.

Table 7 shows a session in cluster 6. Here the user starts with a general query, “carpet,” but instantly adds very specific keywords, probably based on checking the first search results. In this cluster, we found many similar searching behaviors that issue a general query in the first search and add highly product-specific words. In such sessions, users spend the most time checking products with specific properties. This trend is also consistent with the plot in Figure 6.

Table 8 shows an example of a session in cluster 2. The session begins with a specific model of game hardware (“ftr-s-yadn” is the name of a limited edition of “nintendo 2ds” that is yellow). However, the user removed the model name and started to look for a standard “nintendo 2ds” model perhaps because he or she did not like the first target very much. The same tendency, which issues highly product-specific queries and changes them to general queries, can be observed in many sessions in cluster 2.

Overall, in every cluster, we confirmed interesting search behavior tendencies in terms of product-specificity. Since each tendency is consistent with the interpretation of Figure 6, we conclude that clustering by our proposed method effectively characterizes the searching behaviors based on the product-specificity of queries.

5.3 Related methods

We clarify the essential differences between our proposed method and some related existing methods based on the examples in the experiment.

- The studies on query reformulation attempt to uncover the reason why the user changed the query. For example, the method proposed by Jiang and Ni [16] estimates the correlation between individual query reformulations and observable session-level variables (e.g., the type of specified task that the user is engaging currently) by using hierarchical logistic regression. While this kind of method can attribute the individual (local) query reformulation to some hidden reasons, the session-wide (global) characteristics, such as "narrowing-down behavior" and "expanding behavior", cannot be estimated.

- From the perspective of text analysis, one of the relevant work is the study that applies topic modeling (e.g., latent Dirichlet allocation [4], biterm topic model [2, 30], etc.) to queries by regarding them as bags of words. For instance, the method proposed by Cummins [9] analyzes query words by using topic models to uncover hidden correlated word groups as topics. This method is useful to extract knowledge such as "nintendo" and "2ds" are in the same group of words that are frequently used in the same context, which is yet a different target from ours.

- The method proposed by Nozaki and Satoh [21] applies a clustering algorithm to the set of search sessions in online shopping sites along with sequential features based on a series of actions such as issuing a query and clicking a search result. Their clustering algorithm yields clusters that basically correspond to how quickly users can find a query that satisfies the user's information need. For example, Nozaki's method intends to assign the same cluster to the sessions in Table 5 and Table 8 because both sessions stop to change the query at the early phase. This example clarifies the difference from our proposed method, which pays attention to the increase/decrease of the product-specificity of queries and assigns different clusters to these sessions.

6 Conclusion

We proposed a method that analyzes searching behavior based on the product-specificity of issued queries and its transition during a session. The proposed method consists of four steps: (1) calculation of the information content of query words; (2) extraction of search sessions; (3) calculation of the information content of queries; (4) clustering of session behavior. From analysis results with actual data, we confirmed that the proposed method effectively characterizes product-searching behavior based on the product-specificity of search queries.

We need to develop a method that introduces another perspective to the product-specificity analysis in future work. One of the promising directions is the application for user profiling. Each user possibly tends to take a particular type of searching behavior characterized by transitions of queries' product-specificity. If we could develop a method that identifies users' tendencies in terms of product-specificity, it will help provide better product recommendations or user navigations.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP16H02904. We participated in the DBSJ Data Challenge and used the Ponpare Mall data provided by Recruit Technologies, Inc.

References

[1] P. Anick, "Using terminological feedback for web search refinement: A log-based study," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 88–95, 2003.

[2] N. Awaya, J. Kitazono, T. Omori, and S. Ozawa, "Stochastic collapsed variational Bayesian inference for biterm topic model," Proceedings of the 2016 International Joint Conference on Neural Networks, pp. 3364–3370, 2016.

[3] M. Barathi and V. Shanmugam, "Topic based query suggestion using hidden topic model for effective web search," Journal of Theoretical and Applied Information Technology, vol. 59, no. 3, pp. 632–642, 2014.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[5] P. Bruza and S. Dennis, "Query reformulation on the internet: Empirical data and the hyperindex search engine," Computer-Assisted Information Searching on Internet, pp. 488–499, 1997.

[6] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The query-flow graph: Model and applications," Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 609–618, 2008.

[7] I. Bordino, C. Castillo, D. Donato, and A. Gionis, "Query similarity by projecting the query-flow graph," Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 515–522, 2010.

[8] A. Broder, "A taxonomy of web search," ACM SIGIR Forum, vol. 36, iss. 2, pp. 3–10, 2002.

[9] Ronan Cummins, "Improved Query-Topic Models Using Pseudo-Relevant Pólya Document Models," Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 101–108, 2017.

[10] L.A. Granka, T. Joachims, and G.K. Gay, "Eye-Tracking Analysis of User Behavior in WWW-Search," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 478–479, 2004.

[11] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," Proceedings of the 18th ACM conference on Information and knowledge management, pp. 77–86, 2009.

[12] P. Huang, N. H. Lurie, and S. Mitra, "Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods," Journal of Marketing, vol. 73, no. 2, pp. 55–69, 2009.

[13] B. J. Jansen, A. Spink, and D. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," Information Processing & Management, vol. 36, iss. 2, pp. 207–227, 2000.

[14] B.J. Jansen, D. L. Booth, and A. Spink, "Patterns of query reformulation during web searching," Journal of the American Society for Information Science and Technology, vol. 60 no. 7, pp. 1358–1371, 2009.

[15] B. J. Jansen and D. Booth, "Classifying Web Queries by Topic and User Intent," CHI '10 Extended Abstracts on Human Factors in Computing Systems, pp. 4285–4290, 2010.

[16] J. Jiang and C. Ni, "What Affects Word Changes in Query Reformulation During a Task-based Search Session?," Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, pp. 111–120, 2016.

[17] M. Levene and G. Lozou, "Computing the entropy of user navigation in the web," International Journal of Information Technology and Decision Making, vol. 2, no. 3, pp. 459–476, 2003.

[18] N. Matthijs and F. Radlinski. "Personalizing web search using long term browsing history," Proceedings of the fourth ACM international conference on Web search and data mining, pp. 25–34, 2011.

[19] L. Milong. "The Analysis of Strengths and Weaknesses of Online-Shopping," Proceedings of International Conference on Information and Management Engineering, pp. 457–464, 2011.

[20] W. W. Moe, "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream", Journal of Consumer Psychology, vol. 13, iss. 1-2, pp. 29–39, 2003.

[21] Y. Nozaki and T. Satoh, "Search Log Analysis Method of Online Shopping Sites for Navigating Item Categories," Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services, pp. 85–93, 2018.

[22] J. Paparrizos and L. Gravano, "k-Shape: Efficient and Accurate Clustering of Time Series," Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1855–1870, 2015.

[23] S. Y. Rieh and H. Xie, "Analysis of multiple query reformulations on the web: The interactive information retrieval context," Information Processing & Management, vol. 42, no. 3, pp. 751–768, 2006.

[24] D. Schellong, J. Kemper and M. Brettel, "Clickstream Data as a Source to Uncover Consumer Shopping Types in a Large-Scale Online Setting," Proceedings of the 24th European Conference on Information Systems, 2016.

[25] A. E. Schlosser, T. B. White, and S. M. Lloyd, "Converting Web Site Visitors into Buyers: How Web Site Investment Increases Consumer Trusting Beliefs and Online Purchase Intentions," Journal of Marketing, vol. 70, no. 2, pp. 133–148, 2006.

[26] P. Sondhi, M. Sharma, P. Kolari and C. Zhai, "A Taxonomy of Queries for E-commerce Search," Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1245–1248, 2018.

[27] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck. "Probabilistic models for personalizing web search," Proceedings of the fifth ACM international conference on Web search and data mining, pp. 433–442, 2012.

[28] J. Teevan, S. T. Dumais, and D. J. Liebling. "To personalize or not to personalize: modeling queries with variation in user intent," Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 163–170, 2008.

[29] J. Teevan, S. T. Dumais and E. Horvitz, "Potential for Personalization," ACM Trans. Comput.-Hum. Interact., vol. 17, no. 1, pp. 4:1–4:31, 2010.

[30] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," Proceedings of the 22nd international conference on World Wide Web, pp. 1445–1456, 2013.

[31] K. Zhai, Z. Kozareva, Y. Hu, Q. Li and W. Guo, "Query to Knowledge: Unsupervised Entity Extraction from Shopping Queries Using Adaptor Grammars," Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 255–264, 2016.