# Searching for Old News: User Interests and Behavior within a National Collection

Tessel Bogaard
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Tessel.Bogaard@cwi.nl

Laura Hollink
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
L.Hollink@cwi.nl

Jan Wielemaker
Centrum Wiskunde & Informatica
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
J.Wielemaker@cwi.nl

Lynda Hardman
Centrum Wiskunde & Informatica
Universiteit Utrecht
Amsterdam/Utrecht, The Netherlands
Lynda.Hardman@cwi.nl

Jacco van Ossenbruggen
Centrum Wiskunde & Informatica
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
Jacco.van.Ossenbruggen@cwi.nl

## ABSTRACT

Modeling user interests helps to improve system support or refine recommendations in Interactive Information Retrieval. The aim of this study is to identify user interests in different parts of an online collection and investigate the related search behavior. To do this, we propose to use the metadata of selected facets and clicked documents as features for clustering sessions identified in user logs. We evaluate the session clusters by measuring their stability over a six-month period.

We apply our approach to data from the National Library of the Netherlands, a typical digital library with a richly annotated historical newspaper collection and a faceted search interface. Our results show that users interested in specific parts of the collection use different search techniques. We demonstrate that a metadata-based clustering helps to reveal and understand user interests in terms of the collection, and how search behavior is related to specific parts within the collection.

## CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; *Clustering*; *Query log analysis*; *Search interfaces*;

## KEYWORDS

User interest; Search behavior; Digital libraries; Metadata; Log analysis; Clustering

**ACM Reference Format:**
Tessel Bogaard, Laura Hollink, Jan Wielemaker, Lynda Hardman, and Jacco van Ossenbruggen. 2019. Searching for Old News: User Interests and Behavior within a National Collection. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3295750.3298925


*CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom*

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298925

## 1 INTRODUCTION

Understanding user interests and related search behavior can reveal the different types of system support users need. Collections are not always homogeneous and users may have different information needs depending on which parts of a collection they are interested in. This begs the question of how we can identify different "parts" of a collection – for example through different usage patterns and/or by professionally curated categorizations of the documents in the collection. If we are able to identify different usage patterns corresponding to identifiable parts of the collection then we can better help collection owners in providing support for these.

The research questions addressed in this paper are thus:

- (RQ1) What are the user interests in terms of the different parts of a collection? How can we detect these?
- (RQ2) What is the related search behavior within these parts?

Understanding the answers to these questions may lead to more targeted search interfaces, better search algorithms, and a fine-tuning of strategies for collection management.

In a digital library or archive, metadata categorizations of the documents are often reflected in facets, allowing the user to filter the search results. We use the metadata of facets selected and of documents clicked to detect user interests. As we do not know in advance in which (combinations of) metadata categories users are interested, we apply a data-driven partitioning of sessions: a clustering of sessions based on the metadata features of both search (selected facet values) and clicks (document metadata) in each session, and we analyze the behavior in the resulting clusters.

Evaluating the resulting clustering is nontrivial, for different reasons: first, an interpretation of the results is subjective, and second, we have no ground truth available in the data as a way to measure the "correctness" of the clustering. Nevertheless, as we are interested in stable clusters that reappear over different periods, we test the stability of the clusters in each month over a six-month period and interpret stability as an indicator of the quality of the clustering, similar to the cluster stability measured between two periods in [3].

We apply our approach to data from the National Library of the Netherlands, a digital library with a richly annotated historical newspaper collection spanning 400 years, and a faceted search interface. The library has granted us access to both user logs[1] and the metadata descriptions of the documents in the collection.

Our results show that the detected user interests are stable, and that the related search behavior varies within the different parts of the collection. Examples of user interests are: specific types of news items, such as family announcements (relating to births, marriages, deaths), specific periods, such as 1930-49 (including the Great Depression and World War II), or specific regions, such as Suriname (one of the former Dutch colonies). We observe users focusing exclusively on specific parts of the collection, in some parts spending less time and few search techniques, in other parts spending a lot of time and a variety of search techniques. As a result this approach can help to find and investigate these highly-focused users. This can inform the design of more targeted user interfaces, or help to improve search systems or collection management. We contribute to the research field by demonstrating that a partitioning of sessions into clusters based on the metadata of a collection and an investigation of related search behavior reveals specific user needs in specific parts of a collection, where in an overall analysis these patterns would disappear.

## 2 RELATED WORK

To answer our research questions, definitions of user interests and sessions are needed. Additionally, we need a method to group the sessions. In this section we discuss relevant literature with respect to how to detect user interests, define sessions, and what methods can be used to group the identified sessions.

*Detecting user interests.* User interests are frequently derived from queries, for example by categorizing user queries in [15], or finding search topics by semantic linking of user queries [10]. Alternatively, interests can be detected in logs using the context of search [23], or search histories [24]; and in [9] mouse hovering is used to help understand user interests within a digital library, in combination with query analysis and the (analyzed) metadata of document clicks in a statistical analysis.

Similar to this research, we use a form of categorization to identify user interests, and similar to [9], we make use of the metadata categories of the collection. However, we use the metadata directly as found in facets selected and documents clicked, rather than the query input, to identify user interests, as we aim for a definition of user interests in terms of parts of the collection.

*Defining sessions.* Search behavior is often interpreted using a bounded sequence of search actions by a user [12]. Sessions have been studied to understand search in context and to evaluate it in terms of success or failure [12]. Sessions help to provide information about repeated visits [13], to examine query modification [10], to obtain information about learning in search [5], or to find patterns in search behavior [3, 19].

We use sessions to put user interactions in a context and so to enable the detection of user interests and behavior. This requires a computational method for specifying the beginning and end of a session. Sessions can be specified based on query boundaries using the IP address as identifier. For example, in [7] a session is defined as a search query and the following clicks until the next query, and in [11] a session is bounded by the presence of overlapping terms in successive queries until there is no more common term. Sessions are frequently bounded by a timeout, a period of inactivity by a user, e.g. [2, 5, 10, 12]. In the context of studying web navigation, the concept of a *clickstream* is more often used, as in [22]. A clickstream is the navigational path a user follows, consisting of consecutive HTTP requests from a single IP address. We adopt this definition of a session, as it enables the identification of multiple users behind a single IP and we want to avoid breaking up longer sessions by using a timeout.

*Grouping user logs.* Several approaches exist to group user logs in order to find patterns, for example logs can be classified or clustered. To classify different types of behavior, queries have been grouped into *why* versus *what* questions [5], into DBpedia concepts [16], or into categorizations based on a thesaurus related to the collection [11]. Alternatively, Niu and Hemminger have provided an analysis of faceted versus non-faceted search, grouping the logs based on user actions, showing in their work that facets play an important role in search [18]. In our study, we not only include the facets, we also enrich the clicked documents with their metadata descriptions, and use this metadata explicitly to group the user logs for the detection of user interests.

Clustering techniques can also be used to detect patterns in logs. For example, Wang et al. use unsupervised hierarchical clustering to detect user behavior patterns in social networks [22]. In our work, we also use unsupervised clustering and not supervised classification, for similar reasons: we do not have a ground truth available in the data, nor do we know in advance which patterns we want to detect. However, since our data is skewed we use a different algorithm that is more robust to outliers.

Clustering techniques have been used before in the context of a digital library. Chen and Cooper applied a hybrid clustering technique to detect different types of users in the logs, combining an initial clustering using k-means with hierarchical clustering to get to the final clusters [3]. In this research, sessions are represented using a set of features based on user interactions with the search system. More recently, Niu and Hemminger have reproduced this research with an added focus on the facets present in more recent search interfaces [19]. In our study the goal is different, as we aim to find the user interests in terms of the collection and relate these user interests to search behavior. Nevertheless, we use a similar clustering technique and a similar representation of the sessions to be clustered as in [3] and [19], even though we focus exclusively on the bibliographic metadata features of search and clicks.

To evaluate the clustering we look at stability [21], similar to the approach in [3]. This approach was more recently investigated as a validation method for a clustering in a log analysis of a digital library in [6].

## 3 METHOD

In this study we use a clustering algorithm to detect the user interests and investigate the relation between these user interests and

[1]Logs collected from the search platform http://www.delpher.nl, access granted under a strict confidentiality agreement.

search behavior in the collection. For the clustering of the sessions, we base the features on the metadata of facets and clicked documents (the metadata of the facets are the selected values used in search). To do this we need both user logs and metadata records of the collection being searched.

### 3.1 Session Identification and Representation

We identify sessions in the logs based on a *clickstream* model, using the IP address as identifier and connecting sequential HTTP requests to follow the user navigating the search platform.

We represent the sessions based on the metadata values of the search interactions, where available in the facets selected, and clicked documents, linked to the metadata records of the collection. We include all values of the (main) categories in the metadata (such as publication date, origin or type of document). These values are proportional to the number of search interactions or the number of clicked documents per session, and are used as features for the clustering.

To detect the user interests, we apply a clustering algorithm representing the sessions using a metadata feature set. As the features are likely to be correlated, principal component analysis is applied for dimensionality reduction before clustering with a standardized feature set. We retain the principal components with a standard deviation equal to or higher than 1 for the clustering.

In addition, we collect interaction variables based on user interactions within the search interface to analyze the search behavior. These variables include typical variables, such as the total duration of a session, the number of HTTP requests, the proportions of actions that are search or clicks, and specific variables dependent on the search interface, such as facets or reordering of results.

### 3.2 Clustering

We use an unsupervised clustering algorithm, as we have no ground truth available and do not know in advance what kind of patterns are present in the data. Since we cannot assume the data adheres to a normal distribution, we have chosen a k-medoids method [14], partitioning the data into k clusters, as k-medoids is more robust against outliers than k-means is, it is to k-means what the median is to the mean. As we have a high number of sessions and many dimensions in the clustering, we apply the CLARANS algorithm[17], a k-medoids variant optimized for large datasets. We use the Manhattan distance as distance metric for the clustering, because it is suitable for data represented in a high dimensional space [1]. To choose the number of clusters $k$, we apply the silhouette method [20], which measures the separation between the clusters with values ranging from -1 to 1, the higher values indicating a better clustering. We cluster the sessions repeatedly with different values for $k$ and select the $k$ with highest average silhouette width. We use a statistical summary of user behavior in each resulting cluster to analyze differences in behavior between the clusters based on the user interests.

### 3.3 Evaluation of Clustering

Our goal is to find stable patterns that reoccur in different period, so we evaluate the stability of the clustering over time, using this as an indication for clustering quality [21]. For this purpose, we cluster logs collected in separate periods, similar to the approach in [3]. We use a six-month period as it is the maximum period user logs can be retained according to Dutch law and as is common practice to protect the privacy of users. The size of each period is a month, as the sample size used in the collection of the logs was a month and some sessions have a duration longer than two weeks (12% of the sessions).

The stability of the clusters between two periods, the previous period and the target period, is measured as follows:

(1) We cluster the sessions in the previous period using the same value for $k$ as was used for the target period.

(2) For each cluster in the previous period we determine a "center" by taking the original metadata features of the sessions and computing the median for each feature, resulting in a set of medians.

(3) For each session in the target period, we compute the Manhattan distance to each of the centers in the previous period based on the original metadata features.

(4) We assign each session in the target period to the cluster from the previous period with the shortest Manhattan distance, the nearest "center".

(5) For each of the $k$ clusters in the target period, we compute the percentage of sessions in each of the $k$ clusters of the previous period, resulting in $k$ x $k$ percentages .

(6) We define the stability of a cluster in the target period as the highest of the $k$ percentages, the best match.

(7) The stability of a clustering as a whole is the average stability of all its clusters, weighted by cluster size.

We inspect in detail the overlap between the clusters between two periods. We do this with a "stability matrix", that shows the amount of matching (i.e. the percentages per cluster as assigned in step 5) between each of the clusters of the two periods. In the stability matrix, the clusters of the target period are the columns (percentages in the columns sum to 100%), and the previous period the rows.

We remark that cluster stability and silhouette widths measure different things: the first consistency between clusterings over time and the second consistency within a clustering.

## 4 THE NATIONAL LIBRARY OF THE NETHERLANDS

We apply our method to a library that is representative for digital libraries in general, with a richly annotated collection of digitized historical documents and a faceted search interface. The National Library of the Netherlands has granted us access to user logs from their search platform[2], our focus is on the historical newspaper collection, amounting to more than 90% of all HTTP page requests to the library's search platform.

From this collection, users can retrieve full newspaper issues, pages, or individual items on a newspaper page. The documents in the collection are annotated with bibliographic metadata records, including a publication date, distribution zone and type of newspaper item. The distribution zone of a document is the geographical region where the newspaper was distributed, and can be one of

[2]http://www.delpher.nl provides access to collections from the National Library of the Netherlands and other heritage institutions, comprising newspapers, magazines, radio bulletins, and books.

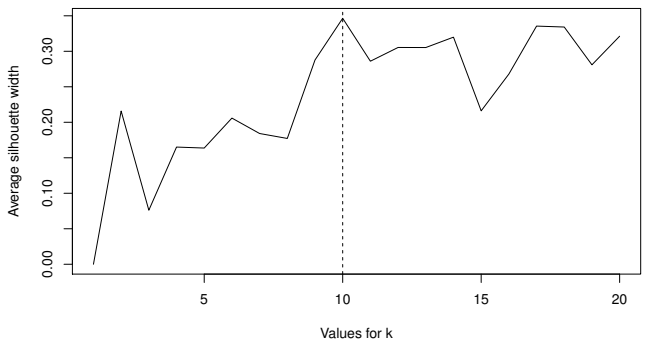


**Figure 1: Average silhouette widths for k in March**

the following values: "local", "national", one of the former Dutch colonies ("Indonesia", "Suriname", or the "Antilles"), or, in a few cases, "unknown". The available newspaper item types are: news articles, advertisements, announcements (relating to births, marriages, deaths) or images (illustrations or photographs, search on the caption text). Table 2 shows the percentages for each metadata value in the collection.

The search interface combines full-text search with facets. The facets are filters based on the metadata attributes of the collection, and include time facets, indicating the publication date, item type facets, and distribution zone facets. In addition, users may change the relevance ranking of the results on a results page to alphabetical or chronological ordering. From a results page, a user can click on a document and, after viewing a document, download it.

The logs used in this experiment were collected between October 2015 and March 2016 (raw data 200M records). In addition, we received the full text digitalization and metadata records of the historical newspaper collection (103M documents at the time), making it possible to link the clicked documents in the logs to the metadata records of all the documents in the collection.

## 4.1 Session Identification and Representation

The user logs contain all HTTP requests to the server. This includes the requested URL, the referrer URL (the origin of the request), the IP address of the client, the browser agent and a timestamp.

We identified sessions from these logs using a *clickstream* model, following the navigational path of a user on the search platform. We removed all logs stemming from web crawlers based on browser agents or a request for robots.txt, redirects, and the loading of style sheets and images. Sequential requests for the same URL right after each other are removed as well, as these are likely reloads of the browser and do not represent a new user interaction.

We group the records by IP address, using the referrer URL to link subsequent requests. Since we are interested in search behavior in relation to the metadata of facets used and documents clicked, we kept only sessions where the sequence consists of more than one search interaction or clicked document in the newspaper collection. This brings the total number of sessions to 255,175 in six months.

For the clustering we create a feature set relating to the metadata values of (i) the clicked documents and of (ii) the facets used in search (Table 1), proportional to the number of clicks or search interactions in that session. For the time facets and publication dates of clicked documents, we split the values into four bins based on equal proportions over all clicked documents and rounded to decades. This leads to a single bin for the period before 1900 and three bins in the 1900-1995 period (Table 2 for the distribution of these values within the collection). The values for the time facets are based on dates within the indicated years, using the same bins as for clicks. We add an extra time facet for the period 1900-1995 to capture those facets that cross the boundaries of the bins in the period 1900-1995.

We define additional session variables influenced by the user interactions in the search interface, and not used for clustering; these are the duration of a session, the number of search interactions and clicks, the use of facets or multiple facets in an interaction, the

**Table 1: Session features based on metadata**

| | |
|---|---|
| | **Publication date clicks** |
| 1 | percentage of clicks published between 1600 and 1899 |
| 2 | percentage of clicks published between 1900 and 1929 |
| 3 | percentage of clicks published between 1930 and 1949 |
| 4 | percentage of clicks published between 1950 and 1995 |
| | **Item types clicks** |
| 5 | percentage clicked articles |
| 6 | percentage clicked family announcements |
| 7 | percentage clicked advertisements |
| 8 | percentage clicked images |
| | **Distribution zone clicks** |
| 9 | percentage of clicks with a local distribution zone |
| 10 | percentage of clicks with a national distribution zone |
| 11 | percentage of clicks with an Indonesian distribution zone |
| 12 | percentage of clicks with a Suriname distribution zone |
| 13 | percentage of clicks with a Antilles distribution zone |
| 14 | percentage of clicks with an unknown distribution zone |
| | **Search with time facets** |
| 15 | percentage of search with time facets |
| 16 | percentage of search with time facets within 1600 and 1899 |
| 17 | percentage of search with time facets within 1900 and 1995 |
| 18 | percentage of search with time facets within 1900 and 1929 |
| 19 | percentage of search with time facets within 1930 and 1949 |
| 20 | percentage of search with time facets within 1950 and 1995 |
| | **Search with item facets** |
| 21 | percentage of search with item facets |
| 22 | percentage of search with article facets |
| 23 | percentage of search with family announcement facets |
| 24 | percentage of search with advertisement facets |
| 25 | percentage of search with image facets |
| | **Search with distribution zone facets** |
| 26 | percentage of search with distribution zone facets |
| 27 | percentage of search with local facets |
| 28 | percentage of search with national facets |
| 29 | percentage of search with Indonesian facets |
| 30 | percentage of search with Suriname facets |
| 31 | percentage of search with Antilles facets |
| 32 | percentage of search with unknown facets |

**Table 2: Collection metadata**

| **Publication date** | **Percentage** |
|---|---|
| between 1600 and 1899 | 12% |
| between 1900 and 1929 | 27% |
| between 1930 and 1949 | 26% |
| between 1950 and 1995 | 35% |
| **Item type** | |
| articles | 67% |
| family announcements | 2% |
| advertisements | 29% |
| images | 2% |
| **Distribution zone** | |
| local | 40% |
| national | 45% |
| Indonesia | 12% |
| Suriname | 1% |
| Antilles | 2% |
| unknown | 0.05% |

use of quotes in queries and the reranking of the results by time. We compute these variables – except the total duration and length (number of interactions) – proportional to the length of the session or the number of search interactions.

## 4.2 Clustering Sessions

We applied principal component analysis on the metadata features in each month separately, in March this led to 15 principal components with an explained variance in the data of 75%. These 15 principal components are used for the clustering, reducing the number of dimensions for the clustering from 32 to 15. We have chosen the number of clusters $k$ based on the average silhouette widths for this month, the highest average silhouette width under twenty is for $k$ equals 10 with a value of 0.35, the first silhouette width above 0.3 (not a high silhouette width but this is not unexpected considering the 15 dimensions – the principal components – used to cluster). We have clustered the sessions from the month March (45,845 sessions) into ten clusters, and using the same value for $k$ as for March we have also clustered the sessions from the previous months to evaluate the stability of the patterns found in March.

# 5 RESULTS

We describe the resulting ten clusters from March ($k = 10$ based on the average silhouette widths as mentioned in section 4.2) in terms of the original values of the metadata features used for clustering, and investigate the stability of this clustering over time. Then, we analyze the search behavior within the clusters.

## 5.1 Clusters

We have labeled the clusters using the most distinctive values of the session features present in a cluster (Table 3), and provided short descriptions of the clusters. The clusters show focused sessions centered around dedicated metadata categories.

For example, one of the larger clusters, the *recent national* cluster (16%), is exclusively centered around the recent national documents in the collection. In most sessions in this cluster, all the clicked documents are published between 1950-95 and have a national distribution zone. Similarly, most sessions in the *recent local* cluster (16%) contain only clicked documents with a local distribution zone of which the large majority is published between 1950-95. This indicates that users searching in the recent parts of the collection are mainly searching for documents with either a local or a national distribution zone and not both, resulting in two separate clusters.

Other clusters are likewise focused, either on a specific period, such as the *1930-49* cluster (15%) with the clicks on documents published during the Great Depression and World War II in the Netherlands, the *1900-29* and the *historical* cluster; or on a specific item type, such as the *family* cluster, where in addition to a majority of announcement clicks most sessions also include announcement facets, and the *article* cluster. For the two smallest clusters, based on a distribution zone, the *Suriname* cluster and the *Antilles* cluster, most sessions include the distribution zone facet next to a majority of clicks from the distribution zone.

The largest cluster (19%), however, is the cluster with sessions without distinct metadata, labeled *no metadata*. Despite leaving out the sessions of length 1 in the data preparation, there is still a relatively large cluster of sessions where hardly any facets are used or documents clicked, leading to a sessions without any representative metadata values.

## 5.2 Cluster Stability

To evaluate the clustering, we check the stability of the clusters, matching the sessions in the clusters to the cluster centers of the previous five months. Table 4 shows, per cluster and for all clusters combined, the percentage of sessions in the clusters of March that falls in the highest matching cluster of each of the previous months.

We observe that overall the clustering is stable, with an average stability of 73%. In particular, the *recent national*, *historical* and *family* clusters are stable every month, as is the *no metadata* cluster. (Note that, even while the percentage of family announcements in the collection is low at 2% (Table 2), there is stable user interest in this part.) Nevertheless, not all clusters in March can be traced back in the previous months. For example, the two smallest clusters in March, *Suriname* and the *Antilles*, do not match well in most of the previous months. Furthermore, the *1930-49* and *1900-29* clusters match well in most but not all months.

The silhouette widths (measuring the consistency within and between the clusters) of the clusters show no direct connection to whether a cluster is stable over time. The *family* cluster, for example, has a relatively high silhouette width of 0.62, but the *historical* cluster, similarly stable, has a lower silhouette width of 0.19. On the other hand, the *Suriname* cluster also has a relatively high silhouette width of 0.64 but a low stability, as for the *Antilles* cluster, both the silhouette width and the stability are low. This can be explained by the fact that cluster stability and silhouette width measure different things: consistency between clusterings over time and consistency within a clustering respectively.

To better understand the stability measurements in detail, we show a single month of the stability results in Figure 2, comparing March 2016 to February 2016. The clusters in February (on the rows) have been labeled in the same manner as the clusters in March.

**Table 3: Clusters March**

| Number of sessions | Silh. width | Label | Description |
|---|---|---|---|
| 8667 (19%) | 0.66 | no metadata | Sessions with little to no metadata values. |
| 7238 (16%) | 0.42 | recent national | At least half the sessions include 100% clicks between 1950-95 with a national distribution zone. About 25% sessions additionally include other clicks. |
| 7549 (16%) | 0.25 | recent local | At least half the sessions include 100% local clicks, of which 85% or more are between 1950-95. About 25% sessions additionally include other clicks. |
| 6837 (15%) | 0.13 | 1930-49 | At least half the sessions in this cluster include 100% clicks between 1930-49. About 25% of sessions additionally include clicks after 1949. |
| 5537 (12%) | 0.02 | 1900-29 | At least half the sessions include 100% clicks between 1900-29. In the sessions more clicks have a national distribution zone than a local one. About 25% of sessions additionally include a minority of clicks between 1930-49s. |
| 4156 (9%) | 0.19 | historical | At least 75% sessions include facets or (a majority of) clicks from before 1900. About half the sessions also include clicks on adverts, about 25% clicks on announcements. There are more clicks in the sessions on documents with a local distribution zone than with a national one. |
| 2701 (6%) | 0.62 | family | At least 75% sessions include announcement facets and a majority of clicks on announcements. In addition, more clicks in the sessions are local than national, and published in the 20th century. About 25% of sessions additionally include clicks on adverts; 25% include clicks on Indonesian documents; 25% clicks on pre-1900 documents; and 25% include time facets or distribution zone facets. |
| 2101 (5%) | 0.37 | article | All sessions include item facets. At least 75% include article facets, 25% advertisement facets. Most of the sessions include a majority of article clicks; some sessions additionally include advertisement clicks. About 25% include time facets between 1900-95; and 25% include national distribution zone facets. |
| 850 (2%) | 0.64 | Suriname | At least 75% sessions include a majority of Suriname clicks, and about half the sessions include Suriname facets. Additionally, about half the sessions include clicks between 1950-95; and about 25% sessions include announcement facets; 25% include some announcement clicks; and 25% include some advertisement clicks. |
| 208 (0.5%) | 0.0 | Antilles | All sessions include Antilles facets; at least 75% sessions also a majority of Antilles clicks. |

**Table 4: Stability testing over time (March)**

| clusters | freq | Oct | Nov | Dec | Jan | Feb |
|---|---|---|---|---|---|---|
| combined | 100% | 74% | 68% | 75% | 72% | 75% |
| no metadata | 19% | 93% | 96% | 95% | 96% | 97% |
| recent national | 16% | 80% | 78% | 81% | 75% | 80% |
| recent local | 16% | 60% | 62% | 63% | 59% | 66% |
| 1930-49 | 15% | 72% | 49% | 77% | 50% | 48% |
| 1900-29 | 12% | 58% | 26% | 79% | 67% | 81% |
| historical | 9% | 82% | 82% | 68% | 81% | 83% |
| family | 6% | 88% | 86% | 86% | 84% | 85% |
| article | 5% | 61% | 69% | 19% | 67% | 66% |
| Suriname | 2% | 49% | 55% | 54% | 38% | 49% |
| Antilles | 0.5% | 35% | 29% | 27% | 30% | 33% |

Here we observe good matching scores on the diagonal for the *no metadata*, *recent national*, *recent local*, *1900-29*, *historical*, *family* and *article* clusters. The *1930-49* cluster, however, does not match to a single cluster, but to two with 48% in one and 28% in another cluster of February. A closer inspection shows that in February the period 1930-49 is split up into two separate clusters, one with mainly local clicks, and a second cluster with mainly national clicks within the same time period. On the other hand, the smallest clusters, the *Suriname* and *Antilles* clusters, have no good match in February at all. The highest matches here are with the *no metadata* cluster. This is because frequently for these sessions the Manhattan distance to the *no metadata* cluster is smaller than to the other clusters, resulting in these cases in an assignment to the *no metadata* cluster.

## 5.3 Search Behavior

We observe a split between the first five clusters in March (*no metadata*, *recent national*, *recent local*, *1930-49* and *1900-29*), and the last five clusters (*historical*, *family*, *article* and *Suriname* and *Antilles*) in Table 5. The first five clusters are shorter, use fewer advanced search techniques, and – with the exception of the first cluster – are more click-oriented; the last five clusters are much longer in time spent and pages visited, and use more advanced search techniques such as facets or reranking of results.

Among the first five clusters, the *no metadata* cluster is different. The sessions in this cluster are the shortest, with the majority less than 2 minutes, and consist of only search interactions, no clicks. Nevertheless, users do spend time and effort (median of 5 interactions), possibly we observe users that completed their search using only the snippets on the results page, or these might be


**Figure 2: Stability matrix, tracing how the sessions in the clusters of March (columns, next to the label the size of the cluster is given) are matched to the cluster centers of February (rows, next to the label the size of the February cluster is given). The percentages in the columns (each column totaling 100 percent) signify the percentages of sessions in the March cluster closest to a February cluster in the rows (distance measured using the Manhattan distance of the metadata features for each session in March to the median values of the metadata features of the clusters of February).**

**Table 5: Search behavior in metadata clusters March**

| clusters | duration | length | search | clicks | facets | | multiple facets | | quotes | reranking results |
|---|---|---|---|---|---|---|---|---|---|---|
| | median | median | median | median | median | q3* | median | q3* | q3* | q3* |
| combined | 00:13:22 | 16 | 81% | 18% | 0% | 46% | 0% | 3% | 0% | 0% |
| no metadata | 00:01:43 | 5 | 100% | 0% | 0% | 33% | 0% | 0% | 0% | 0% |
| recent national | 00:13:30 | 17 | 68% | 29% | 0% | 7% | 0% | 0% | 0% | 0% |
| recent local | 00:13:16 | 17 | 71% | 25% | 0% | 16% | 0% | 0% | 0% | 0% |
| 1930-49 | 00:18:04 | 20 | 75% | 23% | 0% | 39% | 0% | 0% | 0% | 0% |
| 1900-29 | 00:17:17 | 16 | 71% | 25% | 0% | 31% | 0% | 0% | 0% | 0% |
| historical | 00:44:02 | 36 | 84% | 15% | 43% | 67% | 0% | 13% | 0% | 3% |
| family | 46:07:41 | 70 | 83% | 17% | 52% | 77% | 9% | 36% | 29% | 19% |
| article | 01:14:03 | 41 | 82% | 17% | 48% | 85% | 13% | 59% | 0% | 17% |
| Suriname | 00:39:27 | 30 | 80% | 19% | 38% | 68% | 0% | 29% | 0% | 0% |
| Antilles | 01:03:30 | 40 | 81% | 18% | 61% | 87% | 11% | 48% | 0% | 28% |

* q3, or third quartile, is the middle value between the median and maximum

examples of failed search. Of the four more click-oriented clusters, all focus on documents published in the 20th century, with the *recent national* on average the highest percentage of clicks per session. The majority of sessions in these clusters does not make much use of the facets or other more advanced search techniques, but show a more "browsing" behavior where users click through results instead of refining their search. This could in part be explained by the collection, these clusters represent larger parts of the collection (Table 2), the digitization of these documents is likely better (the paper of the newspapers are not aged as much, the language in the documents easier to digitize), and fewer search techniques may be needed to find the desired document.

Next, we have five clusters where users spend a long time and visit many pages. The sessions in these clusters contain a lower percentage of clicks, and the majority of the sessions uses facets. Note that, apart from the *article* cluster, these clusters correlate with smaller parts of the collection (Table 2), and thus likely require more effort from the user. Of these, the *family* cluster contains on average the longest sessions, the majority is longer than a day and the number of interactions is by far the highest, in line with previous research into genealogists and family historians [4], and this cluster likely represents in large part this user group. (Sessions longer than a day are unlikely to be sessions where a user continuously searches, but sessions where a user returns to the same search a day later.) This cluster contains just 6% of the number of the sessions in the month, but the number of interactions is high with a median of 70, resulting in a lot of traffic on the search platform even while the percentage of announcements in the collection is just 2%, and suggesting the users in this cluster are highly engaged in their search. In this cluster we also observe the most frequent use of quotes for the queries, this is not unexpected as search within the family announcements are likely to include search for personal names with respect to genealogy and family histories.

## 6 DISCUSSION

Our results demonstrate that patterns of user behavior can be correlated with document metadata in a way that provides clusters that can be described in a meaningful way to collection curators.

*Metadata Dependency.* Our clustering using the metadata of search and clicks is dependent on by the existing metadata categories the curators have given; however, this is inherent to any curated online collection. It is possible to (additionally) use query analysis and link the query to the metadata of the collection, for example by using a relevant ontology or thesaurus as was done in [10, 11]. Query analysis, however, suffers from several disadvantages: queries can be ambiguous as they form an uncontrolled vocabulary, and queries may include privacy-sensitive information.

*Session Identification.* To identify the sessions to be clustered we have chosen a clickstream model, as it can help to split possible multiple users behind a single IP address, and to find complete searches. This approach leads in some cases to shorter or longer sessions than when a timeout is used, think for example when a user continues their search in a new tab thereby breaking off a clickstream-based session, or the opposite case when a user continues the next day with their search, this would lead to a break in a timeout-based session. For example, the sessions in the *family* cluster last for longer than a day in the clickstream-based sessions, when using the timeout-based session definition these sessions would be broken up into multiple sessions. An alternative to a purely clickstream-based session definition could be a combination of clickstream and query-term overlap, even though query analysis can introduce another sort of bias, and also for this reason we have chosen to keep the session definition simple.

*Exploring k.* We have clustered the sessions into ten clusters based on the best average silhouette width under twenty, however, the number of clusters $k$ can also be used as a parameter of how fine-grained the analysis of the user interests is going to be. As the average silhouette widths for $k$ values under twenty illustrate (Fig. 1), higher values of $k$ can have similar average silhouette widths, making it possible to first set $k$ low for an overview, and then higher to investigate more detailed user interests. The extent to which the value of $k$ should be manipulated is dependent on, among other things, the existing metadata categories and the level of detail deemed appropriate by curators. Also, a higher value for $k$, might solve the disappearance of clusters like the *Suriname* and *Antilles* clusters in previous periods, which in the stability matrix merged into the *no metadata* cluster in the month of February (Fig. 2).

*Fuzzy Clustering.* Even though the large majority of sessions in each cluster are highly focused, we do find sessions on the edges of the clusters that are a bit more "mixed" with respect to user interests, such as the *1900-29* cluster where some of the sessions also include a minority of clicks from between 1930-49 (Table 3). The clustering algorithm we applied, however, is binary, in the sense that a session belongs to a single cluster, even if in some cases it is possible that it has characteristics matching more than one. For future work, it could be interesting to look into more fuzzy or soft clustering techniques, where a session can belong to multiple clusters.

*Clustering Search Behavior.* It is possible to cluster the same sessions using interaction features describing search behavior, such as session duration or number of clicks. These “behavior” clusters can then be mapped to the identified user interests, as opposed to a simple statistical summary, making it possible to find more than a single search pattern for each user interest. However, a first attempt using the same clustering method but with interaction features based on the search interface did not lead to more detailed insights than the statistical analysis provided: the overall overview remained the same. Possibly a search task analysis, such as presented in [8] is more effective here.

## 7 CONCLUSION

By applying a clustering algorithm we were able to identify user interests and investigate the relation between them and search behavior within the historical newspaper collection of the National Library of the Netherlands. The user interests we identified are stable over a six-month period. Our approach can be used to find relations between user interests and behavior in any collection described by metadata, such as digital libraries and archives.

Using the clustering based on the metadata features of search and clicks, we were able to observe users focusing on specific parts of the collection, in some parts spending less time and few search techniques, in other parts spending a large amount of time and a variety of search techniques. This method can help to find and investigate these highly-focused users. These findings can inform the design of more targeted user interfaces providing better access to specific parts of the collection, or help to improve search systems or collection management.

## ACKNOWLEDGMENTS

We would like to thank the National Library of the Netherlands for their support. This research is partially supported by the VRE4EIC project, a project that has received funding from the European

Union's Horizon 2020 research and innovation program under grant agreement No 676247.

## REFERENCES


[1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. *Database Theory âĂŞ ICDT 2001* (2001). https://doi.org/10.1007/3-540-44503-X{_}27

[2] Olivier Chapelle and Ya Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/1526709.1526711

[3] Hui Min Chen and Michael D Cooper. 2001. Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology* 52, 11 (2001), 888–904. https://doi.org/10.1002/asi.1159

[4] Paul Darby and Paul D Clough. 2013. Investigating the information-seeking behaviour of genealogists and family historians. *J. Information Science* 39, 1 (2013), 73–84. https://doi.org/10.1177/0165551512469765

[5] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/2556195.2556217

[6] Daniel Grech and Paul Clough. 2016. Investigating Cluster Stability when Analyzing Transaction Logs. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, New York, NY, USA, 115–118. https://doi.org/10.1145/2910896.2910923

[7] Fan Guo, Chao Liu, and Yi Min Wang. 2009. Efficient Multiple-click Models in Web Search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*. ACM, New York, NY, USA, 124–131. https://doi.org/10.1145/1498759.1498818

[8] Jiyin He, Pernilla Qvarfordt, Martin Halvey, and Gene Golovchinsky. 2016. Beyond actions: Exploring the discovery of tactics from user logs. *Information Processing & Management* 52, 6 (2016), 1200 – 1226. https://doi.org/10.1016/j.ipm.2016.05.007

[9] Daniel Hienert and Dagmar Kern. 2017. Term-Mouse-Fixations As an Additional Indicator for Topical User Interests in Domain-Specific Search. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17)*. ACM, New York, NY, USA, 249–252. https://doi.org/10.1145/3121050.3121088

[10] Vera Hollink, Theodora Tsikrika, and Arjen P de Vries. 2011. Semantic search log analysis: A method and a study on professional image search. *Journal of the American Society for Information Science and Technology* 62, 4 (6 2011), 691–713. https://doi.org/10.1002/asi.21484

[11] Bouke Huurnink, Laura Hollink, Wietske Den Van Heuvel, and Maarten De Rijke. 2010. Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis. *Journal of the American Society for Information Science and Technology* (2010). https://doi.org/10.1002/asi.21327

[12] Bernard J. Jansen and Amanda Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. (2006). https://doi.org/10.1016/j.ipm.2004.10.007

[13] Steve Jones, Sally Jo Cunningham, Rodger Mcnab, and Stefan Boddie. 2000. A Transaction Log Analysis of a Digital Library. *International Journal on Digital Libraries* 3, 2 (2000), 152–169. https://doi.org/10.1007/s007999900022

[14] L Kaufman and P J Rousseeuw. 1987. Clustering by means of medoids. (1987).

[15] Fang Liu, Clement Yu, and Weiyi Meng. 2002. Personalized Web Search by Mapping User Queries to Categories. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02)*. ACM, New York, NY, USA, 558–565. https://doi.org/10.1145/584792.584884

[16] Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten de Rijke. 2011. Mapping queries to the Linking Open Data cloud: A case study using DBpedia. *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 4 (2011), 418 – 433. https://doi.org/10.1016/j.websem.2011.04.001

[17] Raymond T Ng and Jiawei Han. 2002. CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Trans. Knowl. Data Eng.* 14, 5 (2002), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

[18] Xi Niu and Bradley Hemminger. 2015. Analyzing the interaction patterns in a faceted search interface. *Journal of the Association for Information Science and Technology* (2015). https://doi.org/10.1002/asi.23227

[19] Xi Niu and Bradley M. Hemminger. 2010. Beyond text querying and ranking list: How people are searching through faceted catalogs in two library environments. In *Proceedings of the ASIST Annual Meeting*. https://doi.org/10.1002/meet.14504701294

[20] Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, Supplement C (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[21] Robert Tibshirani and Guenther Walther. 2005. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics* 14, 3 (2005), 511–528. http://www.jstor.org/stable/27594130

[22] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. 2016. Unsupervised Clickstream Clustering for User Behavior Analysis. In *Proceedings of the 2016 CHI Conference on Human Factors in ComputingSystems, San Jose, CA, USA, May 7-12, 2016*, Allison Druin and, Cliff Lampe and, Dan Morris and, Juan Pablo Hourcade and, and Jofish Kaye (Eds.). ACM, 225–236. https://doi.org/10.1145/2858036.2858107

[23] Ryen W White, Peter Bailey, and Liwei Chen. 2009. Predicting User Interests from Contextual Information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 363–370. https://doi.org/10.1145/1571941.1572005

[24] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing Personalized Search by Mining and Modeling Task Behavior. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 1411–1420. https://doi.org/10.1145/2488388.2488511