"The final authenticated version is available online at https://doi.org/10.1007/978-3-030-21297-1_16."

Analyzing User Behavior in Search Process Models

Marian Lux[1,3], Stefanie Rinderle-Ma[1,2]

[1]University of Vienna, Faculty of Computer Science, [2]ds@univie, Vienna, Austria
marian.lux@univie.ac.at, stefanie.rinderle-ma@univie.ac.at
[3]LuxActive KG, Vienna, Austria, Vienna, Austria

Abstract. Search processes constitute one type of Customer Journey Processes (CJP) as they reflect search (interaction) of customers with an information system or web platform. Understanding the search behavior of customers can yield invaluable insights for, e.g., providing a better search service offer. This work takes a first step towards the analysis of search behavior along paths in the search process models. The paths are identified based on an existing structural process model metric. A novel data-oriented metric based on the number of retrieved search results per search activity is proposed. This metric enables the identification of search patterns along the paths. The metric-based search behavior analysis is prototypically implemented and evaluated based on a real-world data set from the tourism domain.

1 Introduction

Mapping and understanding Customer Journey Processes (CJP) has become a new trend recently. Signavio, for example, names customer journeys a *"strategic imperative"*[1]. This is underpinned by case studies in several domains including tourism [6] and entertainment [8]. In a nutshell, customer journey describes the customer touchpoints/interactions with a company's information system [9]. Search processes constitute one type of CJP as they reflect search (interaction) of customers with an information system or web platform. According to literature, *"a better understanding of user search behavior"* [2] is essential, however, has been restricted to the analysis of single events and sequences so far. We aim to bring together process-oriented analysis with the full range of patterns in a process model (cf. `http://www.workflowpatterns.com/`) and search behavior analysis.

Search process models can become complex as typically customer behavior tends to be diverse [7]. Hence, we proposed a structural metric for assessing the complexity of search process models in [7]. It was shown that together with semantic pre- and post-processing it is possible to derive search paths in the models at a structural level.

[1] `https://www.signavio.com/post/customer-journeys-as-a-strategic-imperative/`

In order to tackle the challenge of complexity, in this work, we focus on search paths in search process models and try to find out what the customer was searching for when following a certain path in the model. This is reflected by the key research question of this work:

RQ: How to assess search behavior along search paths?

Being able to answer this question yields a competitive edge for companies, for example, by providing specific offers along the search paths. Also the customers are empowered to inspect and improve their search experience.

In this work, we tackle **RQ** in a quantitative way, i.e., based on a search behavior metric. This metric requires an extension of the search process model definition provided in [7], i.e., considering the total number of search results of a search activity as data element. Search path patterns are suggested based on literature and the search behavior metric. With these patterns, search paths can be assessed with respect to the search behavior along these paths. One example, is a decreasing of search results in the search behavior which might hint at using more and more specialized search terms in this path. Also "jumping" as search behavior can yield interesting insights for the analysts.

The search path metric is prototypically implemented and the approach is applied to a real-world data set from the tourism domain. Several search paths can be identified and suggestions for the search offering of the company can be derived.

The paper is structured as follows: Section 2 repeats the structural metric and introduces the new search metric for search process models. Section 3 introduces and discusses search path patterns. Section 4 describes the evaluation and the application to a real-world data set. Section 5 discusses related work and Sect. 6 the presented approach.

2 Search Path Metrics

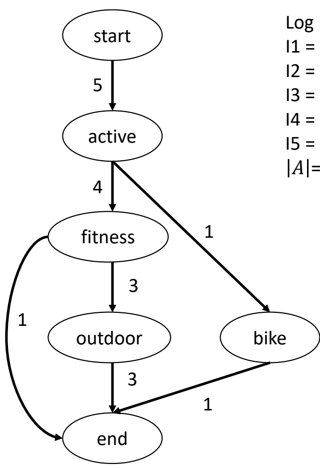
The goal of this work is to analyze (structural paths) in search process models with respect to the search behavior of the users along these paths. One parameter reflecting and influencing the search behavior is the number of search results [12]. If a user, for example, receives a too large number of results for a certain query she/her might decide to narrow down the search in order to obtain a lower number and hence a more targeted search result. Definition 1, hence, extends the definition of search process models from [7] by adding the number of search results obtained per search activity:

Definition 1 (Search Process Model with Search Results). *Let $\mathcal{S}$ be the set of all search terms. A search process model with search results is defined as directed graph $SP := (N, E, l, nsr)$ where*

- *N is a set of nodes*
- *$E \subseteq N \times N$ denotes the set of control edges*
- *$l : N \mapsto S$ denotes a function that maps each node to its label, i.e., $\forall n \in N$ n is a search term.*

– $nsr : N \mapsto \mathbb{N}_0$ *maps each node to the total number of results achieved by the search.*

Note that *nsr* refers to the total number of search results and not to the number of results that are possibly shown to the user (cf. paging). Figure 1 shows an example search process model consisting of four search activities (plus explicit start and end node). The number on the edges reflect the number of instances for which a the path containing the edge has been executed. Also shown is the corresponding process execution log L that contains the execution events for five instances $I1$ to $I5$. Each log entry reflects the execution of one activity together with the number of retrieved search results, e.g., the execution of `fitness` with 50 results for instances $I1$, $I2$, $I3$, and $I4$.



Log L
I1 = <start(0), active (500), fitness(50), outdoor(600), end(0)>
I2 = <start(0), active (500), fitness(50), outdoor(600), end(0)>
I3 = <start(0), active (500), fitness(50), outdoor(600), end(0)>
I4 = <start(0), active (500), fitness(50), end(0)>
I5 = <start(0), active (500), bike(300), end(0)>
$|A|$=6, $|L|$= 23

x ∈ A	start	active	fitness	outdoor	bike	end	$\sum$
freq(x)	5	5	4	3	1	5	23
nrs(x)	0	2500	200	1800	300	0	4800
deg(x)	1	3	3	2	2	3	
spm(x)	0,95	0,84	0,80	0,83	0,39	0,84	
sbm(x)	0	0,41	0,034	0,33	0,06	0	

Fig. 1. Example Search Process with sbm and spb metrics

In our previous work [7], the search process quality metrics (spm metric) was found useful to assess the complexity of search process models and to find relevant search paths. The spm metric relates the degree and frequency of a node to the overall number of activities and number of entries in the underlying log (the one the search process model was mined from). The table in Fig. 1 contains the spm metric for each of the activities. For activity `fitness`, for example, spm turns out as $1 - \frac{3*6}{4*23} = 0.8$. One path of interest can be detected based on spm as `active` $\rightarrow$ `fitness` $\rightarrow$ `outdoor`.

To assess the user behavior along a search path in the search process model we introduce the *search behavior metric sbm* that is based on the number of search results. According to [12] search includes an iterative execution of "query formulation + reformulation" and "evaluation of the results", following certain strategies that depend on the satisfaction with the number of retrieved search results [11]. The number of search results is also considered in web search analysis [10]. The sbm metrics hence takes the number of search results into consideration and for an activity of interest puts it into relation with the overall success of the search in the search process model. Moreover, it weighs the search results by the relative number of executions of the activity that has produced the search results. Doing so enables to differentiate whether, for example, a high number of search results has been produced by a single activity execution or by several activity executions. The latter shall result in a higher value of sbm as more users have conducted the same search.

Definition 2 (Search Behavior Metric). *Search process model, L the corresponding log, and A the corresponding set of distinct activities. Let further freq:* $A \mapsto \mathbb{N}_0$ *count the occurrence of an activity* $x \in A$ *in L. Then the search behavior metric for x sbm(x) is defined as*

$$sbm(x) := \left(1 - \frac{freq(x)}{\sum_{n \in N} freq(n)}\right) * \frac{nsr(x)}{\sum_{n \in N} nsr(n)}$$

with $\sum_{n \in N} freq(n) > 0 \land \sum_{n \in N} nsr(n) > 0$.

If activity x does not produce any search results (i.e., $nsr(x) = 0$), the metric yields a value of 0. As $|L| > 0$ and $|A| \leq |L|$, $|A| > 0$, $sbm(x) \in [0; 1)$ holds.

Consider again Fig. 1 where the sbm metrics is shown for all activities. Along the search path `active` $\rightarrow$ `fitness` $\rightarrow$ `outdoor` first sbm(`active`) $= 0,41$ is achieved, followed by an obvious narrowing down of the search to sbm(`fitness`) $= 0,034$. Then interestingly, the search is again widened to sbm(`outdoor`) $=$ $0,33$. Such "jumps" in the search behavior in one path might indicate a shift in the search strategy. How this search behavior can be analyzed and interpreted will be discussed in the Sect. 3.

3 Revealing Search Behavior on Paths in Search Process Models

In the following, we suggest search path patterns suggested in literature and investigate whether and how these patterns can be applied to analyzing search behavior along paths in search process models. The authors in [12] identify two iteratively executed search steps "query formulation + reformulation" and "evaluation of the results" in user search. They further name two basic search strategies applied during these phases, i.e., narrowing and broadening. Narrowing is applied if the number of search results is perceived as too high. It uses more keywords, conjunction (AND), or negation (NOT). Broadening is employed if the number of search results is perceived too low; it uses less keywords, disjunctions, or text processing techniques such as stemming. Also the use of ontologies, resolving synonyms/homonyms, and adding/removing constraints can support both of the strategies.

We "unroll" the steps "query formulation + reformulation" and "evaluation of the results" and their strategies according to [12] into search paths in the search process model reflecting one line of search that was possibly shared by multiple users. The following patterns are based on a full combination of the strategies "narrowing/no narrowing" and "broadening/no broadening"[2]. Aside the description of the pattern it is discussed how the strategy can be revealed based on the development of the sbm metric of the nodes in the path. The interrelation between the number of search results and the frequency of a node is further elaborated after the pattern descriptions.

Search Path Pattern 1 – Decreasing corresponds to the strategy of narrowing and no broadening, i.e., applying different strategies on the search terms in order to decrease the number of search results along the activities in a path. The manifestation of this pattern in a search process model is a path for which the contained activities unfold a decreasing sbm metric (cf. Fig. 2a).

Search Path Pattern 2 – Increasing corresponds to the strategy of broadening and no narrowing, i.e., increasing the number of search results along the path of interest. This pattern can be deduced from the search process model based on decreasing sbm metric values along a path (cf. Fig. 2b).

Search Path Pattern 3 – Jumping reflects search behavior that jumps between narrowing and broadening along a path, i.e., search results increase and decrease reflected by an increasing and decreasing sbm along a path. This can be caused by using different search terms within one path. We denote the points in the path where the sbm changes from decreasing to increasing and vice versa as jumping points (cf. Fig. 2c).

Search Path Pattern 4 – Constant Search reflects search behavior that produces a similar number of search results along a path, i.e., the variability in the sbm metric is low (cf. Fig. 2d). This pattern can result from using different search terms resulting in a similar amount of search results: ***a)*** By accident or if narrowing / broadening strategies are not effective, e.g., a specification of the search term does not result in any narrowing. ***b)*** If entered terms hardly limit search results because the provided search functionality follows the strategy not to limit search results but sorting them by relevance, e.g., during the search, ontology support helps to broaden and sort the results by considering the original entered search terms first, followed by its synonyms, specializations and finally its generalizations. ***c)*** The underlying overall quantity of possible search results in an information system is small, e.g., an information systems contains merely 10 substantially different documents to search for and most search terms return generally 1 search result. Therefore, most frequent search results contain 1 or 0 search results.

Let us dig a bit deeper into the behavior of the sbm metric for nodes in a path. Let $x_1, x_2 \in N$ be two nodes in a path of a search process model SP=(N, D, l, nsr) where x_2 is a direct successor of x_1 (see Fig. 2). Then:

$$sbm(x_2) > sbm(x_1) \text{ if } nsr(x_2) > \frac{\sum_{n \in N} freq(n) - freq(x_1)}{\sum_{n \in N} freq(n) - freq(x_2)} * nsr(x_1) \tag{1}$$

[2] Note that we only use conjunction in this work.

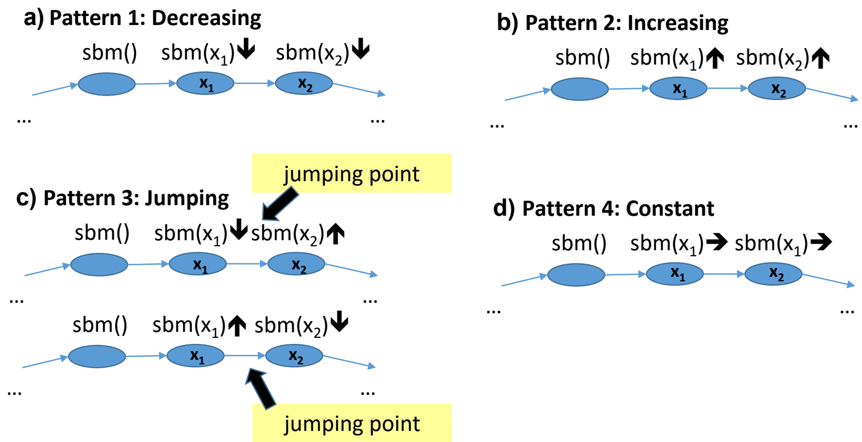


Fig. 2. Overview Search Patterns and Search Metric sbm

Equation 1 holds accordingly for sbm decrease by replacing $>$ by $<$. We fathom Equation 1 by analyzing corner cases.

$$sbm(x_2) > sbm(x_1) \text{ if } \begin{cases} nsr(x_2) > nsr(x_1) & freq(x_1) \approx freq(x_2) \\ nsr(x_2) > \frac{nsr(x_1)}{\sum_{n \in N} freq(n) - 1} & freq(x_1) \gg freq(x_2) \\ nsr(x_2) > nsr(x_1) * (\sum_{n \in N} freq(n) - 1) & freq(x_1) \ll freq(x_2) \end{cases}$$

In the first case, if the frequencies of x_1 and x_2 are roughly the same, the development of sbm follows the development of the search results, i.e., we can directly interpret the results on the strategies narrowing and broadening as described for the search patterns. The second case illustrates the case where $freq(x_1) \gg freq(x_2)$. Hence we put $freq(x_1) = \sum_{n \in N} freq(n) - 1 \wedge freq(x_2) = 1$ as extreme values. Then the number of search results for x_2 has to exceed the number of search results for x_1 divided by the overall number of node frequencies. For high overall frequency, this can mean a drop in search results for x_2 when compared to x_1, i.e., x_2 has to produce less results than x_1. If in the third case $freq(x_1) \ll freq(x_2)$, i.e., we set extreme values $freq(x_1) = 1 \wedge freq(x_2) = \sum_{n \in N} freq(n) - 1$, the number of search results for x_2 has to be higher than the number of search results for x_1 multiplied by the overall frequency. For corner cases two and three this can mean quite a difference in how the number of search results evolves for x_2, however, typically it can be assumed that the number of search results will exceed the number of node frequencies which will cushion the effect. Altogether the sbm metric provides a balanced tool of evaluating the effects of higher or lower search results, but considering the node frequencies that contributed to the search results.

Note that for the abstract example in Sect. 2 the evaluation of the sbm metric does not allow any conclusion as the search results have been synthesized in an arbitrary manner. The conclusiveness of the proposed search behavior path patterns will be evaluated on a real-world data set in Sect. 4.

The above set of patterns covers all combinations of "narrowing", "no narrowing", "broadening", "no broadening". However, further patterns are conceivable. **Search Behavior Pattern 5 – Homogeneous Search**, for example, is not tied to a path in a search process model, but results from finding cluster of activities with similar (homogeneous) search behavior. However, due to space restrictions, Pattern 5 will not be investigated in this work.

4 Evaluation

We evaluate the assessment of search behavior in search process models based on a log from a real-world application called *oHA*. This log comprises the data of four instances executed over the period of *1.5* months. The application is a commercial tourism platform which provides touristic information, called *activities*, from Austria (e.g., points of interest, events, tours, etc.) to tourists and as well locals[3]. The platform is accessible for users as progressive web app[4] and contains about *294,000* activities. The keyword based search functionality for activities has a location based filter method with the option to define an individual search radius around a current or a selected position. An online ontology support guides users through the search functionality by showing suggestions for search terms based on their previous entered search terms and the same ontology is used to broaden, and as well to sort, the results by taking into account, synonyms, generalizations, and specializations.

The search logs are accessible through a *PostgreSQL* database and used as input to calculate sbm results on a mined search process model from these logs. One log record contains the following fields: The field *case device* contains a unique id which is automatically generated per user device. It is used to trace a users' search path. *action time* contains the time stamp when a particular log entry was generated for representing the order of occurred activities inside a search path and is therefore used to calculate the edges in a process model. The *search string* represents the activity which results as node in a process model. Finally, the *search result count* shows the number of returned results and is used to calculate the number of search results *nsr*.

The search functionality has some characteristics, which may influence the resulting process model and are described in the following: There exists a special activity * which signals that the user just hit the search button without considering a search term as query input. The activity * is also automatically executed by the system, if a user enters the search functionality the first time, after manually selecting a new the search position or after deleting the whole search query by pressing a clean button. This results in many *-activities in the logs which are naturally reflected in the mined process models. As mentioned before, the system also shows recommendations for search terms which

[3] `https://austria.myoha.at`

[4] `https://developers.google.com/web/progressive-web-apps/`

can be selected by users. After such a term is selected, a new search will be automatically performed by the system with the added search term (e.g., the first query had *tours* and a user selects *mountain* as recommendation, an automatic search query *tours mountain* will be performed). Therefore the search results will be predominantly refined through the before described behaviour. In addition to the search query, the number of returned results is influenced by the user-selected position and distance radius. Therefore, the same search query – and thus activity – can return a different number of results.

Figure 3 depicts the process model mined in Disco (`https://fluxicon.com/disco/`).

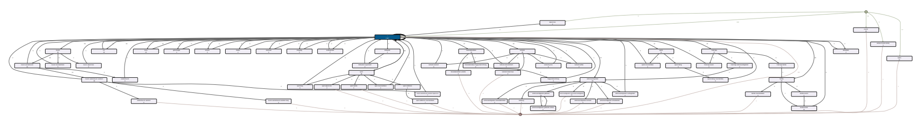


Fig. 3. *oHA* search process model with 10% of the most frequent activities

With a prototypical implementation in Java the spm and sbm metrics were calculated and annotated to the nodes in the search process model. Nodes with high spm results (which are closer to 1) indicate a clear path. Hence, we filter the model depicted in Fig. 3 for nodes with a spm metric, for example, of at least 0.96 (result see Fig. 4).

Then we visually inspect the search behavior based on sbm results of the contained nodes. Since the activity * appears predominantly in every search path, at least at the beginning as described before, we ignore this activity for the pattern recognition and treat it as "outlier". We can identify all four search behavior path patterns introduced in Sect. 3. Note that the logs are produced by the platform with a German and an English user interface. Most users used the German user interface. Therefore, the vast majority of the search paths contain German search terms, which are translated into English in the following for a better comprehensibility: *"familie"="family", "Spielplatz" = "playing area", "touren" = "tours", "wandertouren" = "hiking tours", "tipp" = "advice", "sehenswuerdigkeiten" = "sights", "kulinarik" = "cuisine", "kinder" = "children"*. The following paths are selected for further discussion regarding the identified pattern for the path, e.g., set P1 contains selected paths for which Search Behavior Path Pattern 1 – Decreasing can be revealed:

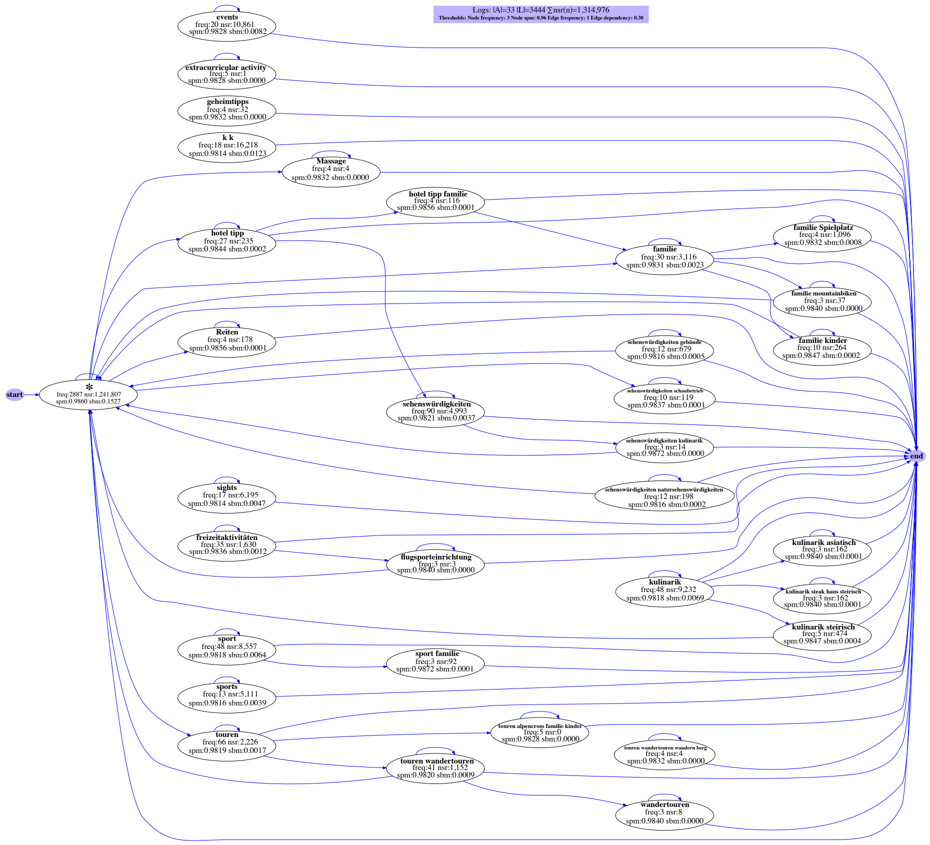


Fig. 4. *oHA* search process model filtered with spm $\geq$ 0.96, produced by implemented prototype

$$
\begin{aligned}
P1 = \{ & \langle start, *(sbm = 0.1527), familie(sbm = 0.0023), familie\, Spielplatz(sbm = 0.0008), end\rangle, \\
& \langle start, *(sbm = 0.1527), sport(spm = 0.0064), sport\, familie(sbm = 0.0001), end\rangle, \\
& \langle start, *(sbm = 0.1527), touren(spm = 0.0017), touren\, wandertouren(sbm = 0.0009), \\
& wandertouren(sbm = 0.0000), end\rangle\} \\
P2 = \{ & \langle start, *(sbm = 0.1527), hotel\, tipp(sbm = 0.0002), sehenswuerdigkeiten(sbm = 0.0037), \\
& end\rangle\} \\
P3 = \{ & \langle start, *(sbm = 0.1527), hotel\, tipp(sbm = 0.0002), sehenswuerdigkeiten(sbm = 0.0037) \\
& sehenswuerdigkeiten\, kulinarik(sbm = 0.0000), end\rangle, \\
& \langle start, *(sbm = 0.1527), hotel\, tipp(sbm = 0.0002), hotel\, tipp\, familie(sbm = 0.0001), \\
& familie(sbm = 0.0.0023), familie\, kinder(spm = 0.0002), end\rangle\} \\
P4 = \{ & \langle start, *(sbm = 0.1527), hotel\, tipp(sbm = 0.0002), hotel\, tipp\, familie(sbm = 0.0001), end\rangle\}
\end{aligned}
$$

P1 refers to Pattern 1 – Decreasing and is most frequently revealed for the given use case. The sbm results are decreasing along the search path because users refined their search results to find useful results. We illustrate one of the paths as model snippet in Fig. 5.

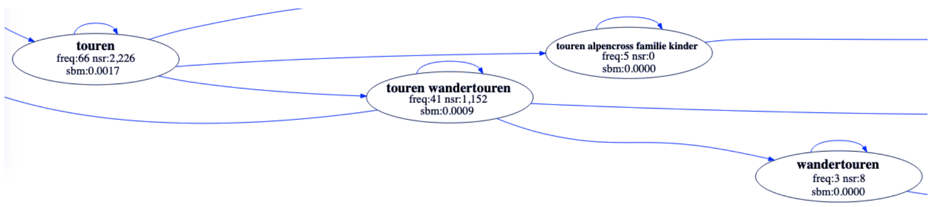


Fig. 5. Search path P1, snippet of search process model, produced by implemented prototype

P2 reflects an increase in the sbm results. Apparently, users tried to broaden their search results because not enough results appeared at the beginning of their search. This pattern appears rarely in the search process model. *P3* shows jumping points (cf. sehenswuerdigkeiten, familie) where users changed their search strategy during the search path. These jumping points can be of interest, e.g., in marketing, by investigating correlations between these jumping points and their previous search terms in their particular paths, for identifying e.g., new user recommendations. *P4* reflects a constant search path. In this example we assume that the users performed the broadening strategy and it was not successful. This pattern appeared only once in the discovered search process model.
Interpretation: The analysis of the search behavior shows that in several search paths users, who are tourists, started with “hotel tipp” which means some advise by the hotel they are staying in. From there, the search continued either with more specialized advises by the hotel (“hotel tipp familie”) or by choosing new, probably connected topic such as sights (“sehenswuerdigkeiten”). We can also see that from the more specialized search, users go back to a connected, but more general search term, e.g., from “hotel tipp familie” to “family” (in English “hotel advise family” to “family”). The interpretation could be that users in general try the advises offered by the hotel and from there apply different search strategies (narrowing and/or broadening) along topics they are specifically interested in, e.g., family or sights. Hence, we could suggest to a hotel owner with such paths to further invest in the hotel advises, particularly targeting certain topics such as family. Also, if the system provides an artificial intelligence function for recommendations of search terms, like the present tourism platform does as described before, the system could replace the suggestion ”hotel tipp” with ”sehenswuerdigkeiten” and ”family”.

5 Related Work

In web search analysis, e.g., [10], search terms and the number of search results are considered, but not the underlying search process models and the search behavior along paths. This observation is underpinned by current work in the Information Retrieval community stating that "[e]xisting work focuses on modeling and predicting single interaction events, such as clicks" [2] where the very same paper proposes the prediction of click sequences. The work at hand, however, is not restricted to click sequences, but considers process models with all kinds of structural patterns (cf. `http://www.workflowpatterns.com/`).

We can understand search processes as a special type of Customer Journey Processes (CJP). CJP comprise of all interactions and touch points of users – such as searches – with a company's information system or (web) platform [5, 6]. It seems that currently commercial tools and systems such as Signavio are at the forefront to develop CJP maps and models. CJP mining has been recently discovered as promising in different application domains [8, 6]. [1] suggests clustering and merging CJP maps with process trees in order to detect representative CJP models from event logs. If a merge happens will be decided by the analyst. In [4] an alternative method to detect relevant CJP based on Markov models is introduced. [7] presents pre- and post-processing methods, together with a structural path metric in order to detect paths in CJP models. The proposed metric is employed in the work at hand. None of the mentioned approaches addresses the search behavior.

For mining search process models, metrics like the spm can be used for filtering nodes to show only paths of interest for analysts. For example the process mining framework ProM[5] offers several metrics as well as edge and node filtering, particularly in the context of the Fuzzy Miner [3].

6 Discussion and Outlook

This work proposes structural and search behavior metrics for discovering and visually inspecting search process models. This enables the detection and analysis of search behavior along paths and facilitates suggestions for improving the search offer in the sequel. For a data set from the tourism domain, for example, it is possible to discover search paths and to derive that the tourism provider (e.g., a hotel) could improve its search suggestions. The novelty of the approach is to employ data and data values into metrics. Using search results is a first step, particularly suited for search process models, but further data can become relevant in explaining customer behavior, e.g., weather or location. Accordingly, new metrics and enhanced mining, filtering, and inspection techniques become necessary to yield the most valuable insights from the data.

References

1. Bernard, G., Andritsos, P.: Cjm-ab: Abstracting customer journey maps using process mining. In: CAiSE Forum 2018. pp. 49–56 (2018)

[5] `http://www.promtools.org/doku.php`

2. Borisov, A., Wardenaar, M., Markov, I., de Rijke, M.: A click sequence model for web search. In: Research & Development in Information Retrieval. pp. 45–54 (2018)
3. Günther, C.W., van der Aalst, W.M.P.: Fuzzy mining - adaptive process simplification based on multi-perspective metrics. In: Business Process Management. pp. 328–343 (2007)
4. Harbich, M., Bernard, G., Berkes, P., Garbinato, B., Andritsos, P.: Discovering customer journey maps using a mixture of markov models. In: Symposium on Data-driven Process Discovery and Analysis. pp. 3–7 (2017)
5. Lemon, K.N., Verhoef, P.C.: Understanding customer experience throughout the customer journey. Journal of Marketing 80(6), 69–96 (2016)
6. Lux, M., Rinderle-Ma, S.: Problems and challenges when implementing a best practice approach for process mining in a tourist information system. In: Business Process Management, Industry Track. pp. 1–12 (2017)
7. Lux, M., Rinderle-Ma, S., Preda, A.: Assessing the quality of search process models. In: Business Process Management. pp. 445–461 (2018)
8. Pmig, Y., Yongil, L.: Customer journey mining. Tech. rep., LOEN Entertainment (2018)
9. Richardson, A.: Using customer journey maps to improve customer experience. Harvard Business Review 15(1), 2–5 (2010)
10. Silverstein, C., Henzinger, M.R., Marais, H., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33(1), 6–12 (1999)
11. Stelmaszewska, H., Blandford, A.: Patterns of interactions: user behaviour in response to search results. In: JCDL Workshop on Usability of Digital Libraries. pp. 29–32. UCL Interaction Centre (UCLIC) (2002)
12. Sutcliffe, A.G., Ennis, M.: Towards a cognitive theory of information retrieval. Interacting with Computers 10(3), 321–351 (1998)