Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement

F. Sperrle, H. Schäfer, D. Keim, and M. El-Assady

Abstract

Mixed-initiative visual analytics systems support collaborative human-machine decision-making processes. However, many multi-objective optimization tasks, such as topic model refinement, are highly subjective and context-dependent. Hence, systems need to adapt their optimization suggestions throughout the interactive refinement process to provide efficient guidance. To tackle this challenge, we present a technique for learning context-dependent user preferences and demonstrate its applicability to topic model refinement. We deploy agents with distinct associated optimization strategies that compete for the user’s acceptance of their suggestions. To decide when to provide guidance, each agent maintains an intelligible, rule-based classifier over context vectorizations that captures the development of quality metrics between distinct analysis states. By observing implicit and explicit user feedback, agents learn in which contexts to provide their specific guidance operation. An agent in topic model refinement might, for example, learn to react to declining model coherence by suggesting to split a topic. Our results confirm that the rules learned by agents capture contextual user preferences. Further, we show that the learned rules are transferable between similar datasets, avoiding common cold-start problems and enabling a continuous refinement of agents across corpora.

1. Introduction

Mixed-initiative approaches in which human and computer agents contribute their best-suited actions at the most appropriate time [AGH99] have proven useful for the optimization of machine learning models. Due to the complex nature of many machine learning tasks, such refinements usually operate on multiple optimization objectives simultaneously. These objectives are proxy-measures for the quality of the trained machine learning model but often do not capture the user’s intuition of quality. Additionally, there often exist multiple, equally correct solutions that fulfill different user preference profiles. In particular, the desired outcome of complex analysis tasks in various domains like text analysis or crime investigation often depends on the analyst’s personal preference, the domain itself, or any downstream tasks that require certain results. A major concern of domain experts in these fields is to incorporate their nuanced domain understanding to adapt the machine learning results to their preferences, which, in turn, depend on the context and task at hand. This makes those tasks prime targets for visual analysis, where domain experts can incorporate their knowledge.

To enable more efficient mixed-initiative processes in these circumstances, visual analytics approaches should allow for co-adaptation between users and systems [SJB20]. In this paper, we
focus on interactive topic model refinement and present a guidance
technique that relies on contextualized, adaptive agents and is con-
ceptualized in Figure 2. Each of the task-specific agents is respon-
sible for a single optimization operation in the refinement process.
During the model refinement, all agents compete for the user’s satis-
faction by suggesting alternative model states based on the current
analysis context. Based on direct and indirect user feedback, each
agent aims to capture the users’ preferences by learning in which analysis
contexts to provide its suggestions. Thus, each agent strives
to deliver only suggestions that would be accepted by the user while
not making suggestions that would be rejected. This ensures that
the visual analytics process can progress with minimum feedback
for maximum gain [EASS*18] by reducing the number of unwanted
suggestions interrupting the user’s analysis process.

We instantiate the proposed technique in an application for the
iterative refinement of the Incremental Hierarchical Topic
Model [EASD*19] (IHTM). The application uses guidance agents
that suggest common topic model refinement operations and explain
and justify their suggestions using verbalizations. Ignored sugges-
tions decay over time, reducing the required amount of active sys-
tem teaching. We evaluate our technique using different approaches.
First, a qualitative expert user study to assess the usability of the in-
teractive interface. Second, three quantitative experiments to verify
the learning ability of our technique for different aspects: preference-
based topic ranking, contextualization, as well as transferability. Our
multi-faceted evaluation confirmed the effectiveness of our approach
for mixed-initiative topic refinement and suggests that agents learn
meaningful optimization rules that match user intuition but could be
improved through the addition of domain semantics.

To summarize, the main contributions of this paper are: (1) A
technique for providing agent-based, co-adaptive guidance for multi-
objective optimization problems in visual analytics. (2) An applica-
tion for the guided refinement of topic models that instantiates this
technique. (3) A multi-faceted evaluation to assess the interplay of
different aspects influencing the effectiveness of adaptive guidance
and illustrates the validity of the learned user preference models.

2. Background and Related Work

Recent years have seen a rise of interest in guidance in visual an-
alytics. Ceneda et al. characterized guidance processes in terms
of a knowledge gap encountered by the user, the available inputs
and outputs, and the degree of guidance [CGM*17]. More recently,
they also characterize guidance as an adaptive mixed-initiative pro-
cess [CGM19]. In our previous work, we extended that characteri-
ation to define co-adaptive guidance as a mixed-initiative interac-
tion process characterized by the dynamics of learning and teaching
[SJB*20]. In this paper, we present a technique that employs the
concept of co-adaptive guidance to enable users to teach a set of
guidance agents their personal preferences. Collins et al. [CAS*18]
provide a process-oriented model of guidance and state “just-
in-time-visualization” to be an important goal of guidance.

Guidance and Recommendation Interfaces – Both Ceneda et
al. [CGM19] and Collins et al. [CAS*18] survey a large num-
ber of applications that employ guidance. In particular, Ceneda et
al. [CGM19] identified twelve papers that use guidance in the model
building process. Several systems provide guidance for splitting data
into clusters or decision trees [AEK00, AAR*09]. Drucker et al. sup-
port iterative document clustering by highlighting relevant clusters
once users select an artifact [DFB11]. While these approaches pro-
vide guidance upon request, we propose a technique in which agents
learn to provide guidance automatically when appropriate. This pa-
er was loosely inspired by cooperative contextual bands [TvdS15],
a class of recommender systems that have recently been used suc-
cessfully in personalized design ideation for the creation of mood
boards [KLHO19]. However, this approach relies on partitioning the
design space of mood boards according to multiple dimensions
and assigning agents to each partition that then sample suggestions
from their partition. This is not possible for the solution space of
topic models or text-based algorithms in general. Instead, our tech-
nique uses guidance agents that each focus on presenting guidance
according to a specific, fixed optimization strategy.

Document Exploration and Topic Model Optimization – A large
number of visual analytics applications for topic model optimization
have been proposed. Some low-level approaches operate directly
on the level of keyword constraints [CLRP13, HBGSS14]. More
high-level approaches enable the user to select one of multiple pro-
vided refinement operations [CLRP13, HC15, DYW*13]. These
operations inspired the refinement operations offered by our applica-
tion. However, the respective approaches do not provide any active
guidance in the modeling process. Recently, Kim et al. [KDEP20]
presented ArchiText, a scalable approach to iterative topic modeling
that tightly integrates a large set of operations proposed in previous
work with a bespoke user interface. While this system enables itera-
tive refinement, it only has minimal guidance capabilities.

Rather than working with topics directly, Park et al. [PKL*18]
present a system in which users iteratively refine a set of concepts
that capture semantics and can be used to describe documents. Se-
matic Concept Spaces [EAKC*20] bridges the gap to topic model-
ing by allowing users to refine perceived similarities between con-
cepts and using the learned similarity information to adjust a topic
model. The system uses semantic interactions [EFN12] to suggest
additional concept space changes to the user based on the adapted
similarity relations, thus indirectly guiding them in refining the topic
model. In contrast to this approach of learning the user’s perceived
concept similarity, our approach is content-agnostic. It operates on a
vector space defined by topic model quality metrics, where it learns
which optimizations are preferred by users in which contexts.

The work closest to our approach is that of Speculative Execution
for Topic Model Optimization [EASD*19]. Our approach builds
on the same framework, using an iteratively-created model and the
same refinement operations. In contrast to this work, which focused on explaining the model’s decision-making processes, we present a technique for learning and generating adaptive, contextualized guidance, which was identified as a limitation in the previous study.

3. Guided Topic Model Refinement

Among others, scholars of the humanities use topic modeling as a means to gaining an understanding of text corpora before using the obtained results as a basis for downstream research efforts. Optimizing models to align with their understanding incorporates several tasks like understanding and diagnosing models, re-assigning documents, and refining topic hierarchies [EASD*19, EAKC*20]. Often, this presents a challenge due to a lack of technical knowledge and the amounts of analyzed data. Co-adaptive guidance in which system and user simultaneously learn from each other and teach one another [SJB*20] can make this process more efficient: systems teach the user model optimizations, and users teach the system which types of suggestions are useful in which contexts. To that end, we target three design goals concerning the guidance process:

G1: Enable contextualized preference learning. To refine in which contexts guidance should be provided, the system should gather implicit and explicit user feedback to suggestions and learn which suggestions are relevant in which contexts. As the user’s time is the limiting factor in this learning process, the system should maximize what it can learn from a minimal amount of interactions.

G2: Provide contextualized suggestions tailored to the current analysis state. To achieve this goal, we employ guidance agents that are responsible for a distinct action and aim to only take the initiative in situations where their suggestion would be meaningful.

G3: Users must remain in control of the refinement process. As mixed-initiative interfaces aim to combine the strengths of humans and machines, they strike a balance between manual labor and automation. However, while systems might aim to teach users relevant optimizations, users should remain in control of the process. This also implies that the system’s guiding suggestions should not interfere with the ongoing refinement that users are performing.

To achieve these goals and enable effective guidance, we want to avoid overwhelming users with a complex, unintuitive interface. Thus, we follow these design rationales and interaction principles:

I1: Simple to use. The system should avoid complex interaction patterns or model visualizations. To assess the current model state, we provide an initial overview first and provide access to both details and the concrete documents in a topic on demand. To flatten the learning curve of the system, we employ consistent visual representations of all major entities (topics, documents, keywords) and make the same interactions available on all occurrences of an entity. Further, we connect all views using linking and brushing and use similar colors for semantically similar topics.

I2: Intelligible and Trustworthy. Each agent enables a detailed inspection of its suggestions and provides a verbal summary and explanation. In uncertain cases, this explanation also reveals evidence considered by agents that would contradict making a suggestion. By including those factors, we aim to foster trust in users that the agents consider a holistic view of the analysis process. Throughout the system, uncertainty is visually encoded through opacity.

I3: Focus on Guided Refinement. The guidance suggestions provided by the agents take a prominent role in the refinement process, and a large part of the interface is allocated to their presentation. Nonetheless, manual refinement of topics with drag and drop is still possible to resolve situations not readily addressed by the agents.

I4: Provide Model History. A model history is particularly important in mixed-initiative processes as it can highlight the collaborative nature of the refinement and allows users to revisit previous agent suggestions at a later point in time.

4. Topic Model Refinement Interface

Before we present our technique for contextualized guidance generation in section 6 we introduce our system that instantiates this technique in the following two sections. It is depicted in Figure 3 and consists of four main views: the topic icicle shows the current modeling state, a development Sankey highlights model changes over time, the document log enables close reading, and six agent cards present the agents’ guidance suggestions. The system enables users to observe the incremental nature of the model building process: over time, documents are first added to the document log before updating the topic icicle and triggering a new model state captured in the Sankey.

4.1. IHTM: Incremental Hierarchical Topic Model

The implemented system is built around IHTM, an Incremental, Hierarchical Topic Model [EASD*19]. In contrast to other models like LDA [BNJ03], IHTM is deterministic and avoids frequent issues with probabilistic models where users cannot replicate their previously obtained results. In addition to being deterministic, IHTM is an incremental model that builds a hierarchical topic structure by sequentially adding new documents to a tree structure. Leaf nodes in IHTM’s hierarchical topic structure represent documents, and inner nodes represent (sub)topics. To insert a new document, the algorithm searches the most similar tree node through recursive, breadth-first descend. Documents are represented as tf-idf vectors and tree nodes as the average of all contained document vectors. In uncertain cases, this explanation also reveals evidence considered by agents that would contradict making a suggestion. By including those factors, we aim to foster trust in users that the agents consider a holistic view of the analysis process. Throughout the system, uncertainty is visually encoded through opacity.
process and identify potential modeling problems early. Second, fixing modeling issues early in the process directly influences the created topic tree, removing the need for multiple modeling iterations.

4.2. Model Visualizations

The topic tree produced by the IHTM is visualized (and continuously updated) as a rotated icicle plot in which the hierarchical topics are depicted from left to right. Each node’s width corresponds to its level in the hierarchy, and higher-level topics take up more space. The rightmost nodes in each cluster represent the documents contained in the respective (sub)topics. Users can directly incorporate their domain knowledge into the modeling result by dragging and dropping subtrees from the icicle plot to make isolated changes (I3). To visually distinguish document nodes from topics, they are shorter and lighter in color. All topic rectangles contain the most important topic descriptors keywords. The number of descriptors is shorter and lighter in color. All topic rectangles contain the most important topic descriptors keywords. The number of descriptors is scaled linearly with the topic level to show more information about more influential higher-level topics. A tooltip for each topic reveals additional details (I1) and shows all of its extracted descriptor keywords and their associated colors (see below) and up to ten representative sentences. Tooltips for document nodes show the document’s text. Clicking on a topic node opens a popup window that re-iterates the topic descriptors and representative sentences and adds a text view of all documents contained in the topic. Furthermore, users can zoom the icicle by control-clicking on topics, enabling the system to scale to large models.

To generate the colors for all topics and documents, we adapt the idea of Semantic Concept Spaces [EAKC20] in which similar keywords are mapped to similar colors in a 2D color map. We first enumerate all keywords from the corpus and remove stop words. Additionally, the system allows the removal of user-defined, context-specific stop words. For the remaining keywords, we obtain DistilBERT [SDCW20] embedding vectors that we project to 2D using UMAP [MHM18]. We then map the resulting projection space to a 2D color map [BLBS11] and assign each keyword a color. To obtain the color for a topic, we calculate the average DistilBERT embedding of its keyword descriptors, weighting each descriptor by its relevance score produced by the keywords extraction algorithm YAKE [CMP20]. We again project this average embedding using UMAP and assign the topic the associated color. Removing keywords contained in the aforementioned keyword blacklist has two benefits. First, it ensures that keywords with more descriptive content can be projected with less distortion. Second, it keeps these keywords from taking up room in the color map and leaves more colors available for other keywords and concepts.

To show the temporal development of top-level topics (I4), we employ the model development Sankey diagram shown in Figure 1. We place the Sankey adjacent to the icicle, and each node represents the topic directly next to it in the icicle. Each new document inserted into the model adds a column to the right of the Sankey diagram. Older model states are pushed out of view but can be reached by scrolling the visualization. The Sankey links show how many documents are shared between two (states of) topics and can be used to track how both manual refinements and agent suggestions moved documents between topics. In addition to showing document movement, the Sankey also shows the development of topic descriptors. Whenever the top-ranked descriptor of a topic changes, the new descriptor is displayed at the corresponding model state’s position in black. The font size is scaled to the number of documents in the topic. Changing descriptors typically lead to a change in topic color. We display outdated descriptors in their topic’s old color, making them stand out less. Together with descriptors only being replaced when they change rather than at fixed intervals, this shifts user focus to currently developing topics. The same topic interactions (hovering for a tooltip, clicking for a popup) that users are familiar with from the icicle plot are also available in the Sankey.

The document log shows all documents inserted into the model. New documents are placed on top of the list; older ones remain accessible by scrolling. The most important keywords in each document are highlighted to aid in quickly scanning a document. All documents are colored using the same process as topics in the icicle plot.

5. Agent-Based Guidance Interface

The system provides six different guidance agents that perform distinct optimizations like splitting or merging topics. All agents observe the modeling process and provide suggestions only when they expect them to be helpful. To avoid interrupting users, the suggestions do not stop the stream of documents from being inserted into the model. Instead, agents automatically keep their suggestions updated to reflect newly added documents. All available agents and the underlying technique for learning contextualized user preferences will be introduced in section 6. This section focuses on the interaction possibilities that the guidance agents afford (I3). We decided against representing each agent as a virtual avatar and instead show them as agent cards in the interface for simplicity and clarity. Two examples of agents providing suggestions are shown in Figure 4. Figure 3 also shows agents without current suggestion.

![Figure 4: Agents review their suggested changes in a small icicle and verbalize them. They also explain their classification and state their confidence and certainty.](Image)
Previous work has successfully applied verbalization to explain machine learning models [HSD19] and presented a design space for verbalization [SBE18]. We employ template-based overview and detail verbalization [SBE18] to provide a summary of agent suggestions at the bottom of the card. The summary (see Figure 4) contains information on existing and new topics, as well as the number of affected documents (I2). Additional information on those topics is available through the tooltip and popup introduced in the previous section (I1). In addition to the summary, each agent also states its confidence in the suggestion based on its internal classifier (I2) and adds a warning when its uncertainty goes above 50%.

The model preview icicle on the right-hand-side of the agent card shows how the model would look if the suggestion was accepted; the changed topics mentioned in the summary and their documents are highlighted with a grey border. The preview’s opacity is determined by the agent’s uncertainty, with more uncertain suggestions having a higher opacity and visually fading in comparison to other suggestions. More detailed information on how the model would change and which specific documents are affected is visualized in the model diff Sankey (see Figure 5) that can be accessed in a popup via the review button . Since agent suggestions do not stop the iterative modeling process to avoid interrupting users, new documents iteratively added to the model are also added to each suggestion, leading to potentially large differences between model states. The model diff Sankey reveals those differences by matching states its previous section (I2) and reinserts at a new position using the original IHTM algorithm. Our previous work has introduced all optimizations in more detail and shows that they provide meaningful model alternatives [EASD19].

As each agent’s refinement operation is fixed, its task in the guidance process is simplified from having to provide the right guidance at the right time to providing its associated optimization in the right model analysis context. To that end, each agent builds and maintains a binary context classifier. As a result, our implementation trains six independent classifiers in parallel and allows all agents to make suggestions independent of the other agents. As there is (theoretically) no limit to the number of agents that can be used, limiting each agent to performing only one operation does not manifest as a limitation of the technique. Instead, it enables the use of simpler classification models that can be used to explain each agent’s decisions.

6. Adaptive Guidance Generation

To generate adaptive guidance suggestions, we introduce a flexible technique for learning contextualized user preferences. This technique is not specific to topic modeling and generalizes to any metric-based model optimization task. Existing guidance systems have often focused on content-based guidance [SSKEA19] that presents users with suggestions that are semantically similar to their previous interactions. Our technique is content-agnostic and does not rely on knowing specific (semantic) interaction sequences. Instead, it builds on the temporal changes observed in metrics that describe the model.

6.1. Contextualized Guidance Agents

At the core of our technique, we employ guidance agents that each target one distinct optimization strategy. Over time, agents build a user preference model by having their suggestions accepted or rejected in different situations. Six complementary agents support various refinement operations: the Combine Similar Topics agent identifies and merges the most similar topics. In contrast, the Split agent splits a topic into two new topics. The Remove Topic Chains agent removes so-called topic chains, modeling artifacts produced by IHTM in certain conditions [EASD19]. The remaining three agents, Reinsert Small Topics, Reinsert Outliers, and Reinsert Worst Topics, each identify a certain subset of topics to be removed from the tree and reinserted at a new position using the original IHTM algorithm. Our previous work has introduced all optimizations in more detail and shows that they provide meaningful model alternatives [EASD19].
In our implementation, we use the five established topic quality metrics: coherence, separation, variance, certainty, and distinctiveness [EASD 19]. A sixth dimension captures the HTM events new topic, new subtopic, and document added that can influence the quality metrics.

To learn satisfactory classifiers on this six-dimensional model analysis context, agents would require large amounts of training and interaction data. Even if enough training data could be obtained for offline pretraining, the classifiers would likely not be able to adapt to the users' preferences during the analysis session. To tackle this issue, we significantly reduce the space by transforming the model analysis context into vectors of metric development shapes that we call context vectors. Rather than individual metric values, the context vector contains each metric’s development between two points in time. We use seven linear shapes to vectorize the metric development over time with a window size of 1: low flat \(\hat{\downarrow}\), medium flat \(\downarrow\), high flat \(\downarrow\), low raise \(\uparrow\), medium raise \(\uparrow\), high lower \(\hat{\uparrow}\), and medium lower \(\hat{\downarrow}\). The categories low, medium, and high are determined by the first, second, and third tertile, respectively. The direction of change is determined bycomparison with the metric value at the previous time step. Metrics with a change of less than one percent are considered to be flat. The window size of 1 means that a new vector is created after each document insert. Different instantiations of the technique can freely choose window sizes to consider smaller or larger contexts. We also note that the shapes or shapelets that are useful in a context vector depend on the analysis context. While we rely on seven linear shapes, our proposed technique is generic and works with any number of arbitrary shapes or shapelets.

Although the simplification of the analysis context to context vectors reduces the state space drastically, the number of possible context vectors remains large. To avoid overfitting on a few specific context vectors and not being able to provide meaningful guidance in the majority of states, agents also use partial context vectors for the classification. Each n-dimensional context vector \(S = \langle s_1, \ldots, s_n \rangle\) has \(2^n - 1\) partial context vectors defined by the power set \(\mathcal{P}(S) \setminus S\) of the elements contained in the context vectors. By not considering one or multiple dimensions \(s_i\), these partial context vectors are broader and more general, allowing agents to improve their user preference model to meaningfully cover the model analysis context with significantly fewer visited states. Furthermore, avoiding the overfitting of user preference models to specific datasets enables continuous agent learning and allows users to refine their agents’ preference models over many visual analytics sessions instead of starting from scratch every time. Our evaluation showcases the transferability of preference models across datasets.

6.3. Context Classification

To determine suitable contexts for providing guidance, each agent builds a classifier that accepts (partial) context vectors and returns a binary decision on whether to provide a suggestion or not. The used classifier should enable continuous refinement and be intelligible to foster trust in an agent’s suggestions. We chose a rule-based classifier that we introduce in the following.

As partial context vectors are based on the power set of analysis context dimensions, longer partial context vectors frequently comprise the information encoded in shorter partial context vectors. To avoid over-emphasizing this duplicated information during classification, we adapt the concept of “closed patterns” from sequential pattern mining [ZH02]: the support of a pattern is defined as its number of occurrences, and a pattern is considered closed if it is not included in a longer pattern with the same support. The agents store the support of each partial context vector and only use closed context vectors during classification. Further, they store how often suggestions provided in a given context \(c_i\) were accepted (\(a_i\)) or rejected (\(r_i\)). When a new model analysis context is encountered during the optimization process, the support of all associated partial context vectors is incremented. As a result, the closedness of a partial context vector can vary over time.

After each document insert into the topic model, each agent performs the steps connected with orange arrows in Figure 6: It vectorizes the analysis context, filters out non-closed contexts, and makes a guidance trigger decision based on all closed context vectors \(C\) using a linear additive model. For a context vector \(c_i \in C\), let \(w_i = \frac{|C \setminus \{c_i\}|}{C} - 0.5\) its weight. Weights are in \([-0.5, 0.5]\), and larger absolute weights show that suggestions in the respective contexts have received more consistent user feedback. Longer, more specific context vectors are expected to have a higher weight as they appear less frequently and are typically less controversial. To keep the classifier from overfitting on those contexts, we log-normalize all weights by the support of the context, leading to lesser weight differences between short and long partial contexts. While the agents are learning, a significant number of contexts will have low weights close to zero, indicating uncertainty. As those contexts can distort the classification result, we exclude all contexts with an absolute weight below 0.25 from the classifier. This boundary and the normalization parameters introduced above have been chosen experimentally and limit the impact of very frequent contexts. They might require fine-tuning for different instantiations of the technique.

The decision on whether to provide guidance for a filtered and normalized set of context vectors \(C_\downarrow\) is determined by summing up the weights of all context vectors: \(g_{C_\downarrow} = \sum_{c \in C_\downarrow} g_c\). Agents will make a suggestion if \(g_{C_\downarrow} > 0\). The confidence of this decision is defined as the ratio of all scores to all absolute scores: \(c = \frac{\sum_{c \in C_\downarrow} g_c}{\sum_{c \in C_\uparrow} |c|}\). The uncertainty is given by the number of ignored contexts below the weight threshold: \(u = 1 - \frac{|C_\downarrow||C_\uparrow|}{|C_\downarrow| + |C_\uparrow|}\). Both confidence and uncertainty are displayed in the system to justify each agent’s decisions.

To avoid a cold-start of the classification model at the beginning of the first analysis session, we generate initial contexts for each agent. We compiled three subsets of the 20news corpus [Lan], each containing 15 documents each from eight newsgroups for a total of 360 documents. For each of the subsets, we build an IHTM and request a suggestion from all agents after each document insert. We then automatically evaluate the quality of both the suggestions and the original models using normalized mutual information [MGH13] (NMI) and extract the context vectors based on the original model. Depending on whether the suggestion improved or worsened the NMI score, we initialize the acceptance or rejection scores of the context vectors to 1, respectively. If an analysis context is encountered multiple times, the values for support, acceptance, and rejection are increased for the respective existing context vectors in the user preference model.
6.4. Learning from User Interaction

Agents learn from both implicit and explicit relevance feedback by adapting either their acceptance or rejection score ($a_i$ or $r_j$) for a given context $c_i$, in turn influencing all future classification results. This section presents how much the scores of which contexts are updated depending on the feedback type. As the closedness of a partial context vector can change over time, agents do not filter by this property in the learning phase and pre-emptively update the scores for all relevant partial contexts.

In our implementation for topic model refinement, each agent remembers which suggestions it has made before and does not suggest its operation on the same topic(s) more than twice. To avoid decaying to a state where agents cannot make suggestions anymore, this memory is wiped whenever a suggestion from the agent is accepted, or no valid suggestion candidates can be found. In addition, users can always manually request guidance from an agent, teaching them that guidance might be needed in a given context.

Types of Feedback – We distinguish between explicit and implicit feedback. Explicit feedback is provided whenever users accept or reject a given agent’s suggestion by using the respective buttons √/×. Implicit feedback is provided automatically by the system when suggestions are rejected because they have been ignored for more than 10 document inserts.

Contexts to Learn From – To maximize the amount of knowledge extracted from a single user interaction, agents consider two distinct points in time: the trigger point and the interaction point of a given suggestion. The trigger point is the point in time at which the guidance agent decided to trigger a new suggestion. Figure 6 shows that the model at the trigger point is the common ancestor between the original model and the suggested model. As the context vectors at the trigger point are responsible for an agent making a suggestion, updating them directly influences whether the agent will provide a suggestion in a similar analysis state again. The interaction point is the point in time at which the suggestion was accepted (or rejected). While the context vectors at the trigger point were responsible for initially proposing the suggestion, the suggestion was still relevant at the point where the user accepted it. Consequently, suggestions similar to that made at the trigger point could also become relevant in the future, leading us to update the respective context vectors as well.

Weighting of Updates – Depending on the type of point in time and the type of feedback that was provided, we update the context vectors with different weights. The weights are outlined in the adjacent table and have been determined experimentally, based on experience from previous work [EASS’18].

Weights for explicit feedback are twice as high as weights for implicit feedback. This ensures that agents primarily learn from explicitly provided feedback but do not ignore the associated implicit information. Furthermore, it gives users the possibility to manually overwrite previously learned rules when necessary, e.g., after changing the analysis direction. Intermediate points encountered between the trigger and acceptance or rejection are not updated.

7. Evaluation

Since the evaluation of mixed-initiative systems depends on a variety of interlinked aspects, a systematic assessment of our technique as a whole is challenging. Therefore, we provide a multi-faceted evaluation. We first report the results from a qualitative expert study to showcase the usability of the provided system and its guidance agents. In our quantitative evaluation, we utilize three approaches to assess the effect of our technique. We determine the effects on result quality in a topic ranking task, verify whether the contexts identified by the agents correspond to human-selected optimizations, and confirm that contexts are transferable between similar datasets.

7.1. Qualitative Evaluation: Expert User Study

This evaluation aims to gain insights into how co-adaptive guidance impacts the topic modeling process. We investigate three main questions: (1) What is the domain expert’s preconception of guided analysis processes? (2) How does the visual analytics interface support interactive model optimization? (3) What is the impression on the utility of the provided co-adaptive guidance? We have since used this study’s feedback to improve our approach. In particular, we optimized the layout by displaying the Sankey next to the icicle, visually simplified the agent cards, and replaced the log-likelihood ratio driven topic descriptor extraction with YAKE [CMP’20].

Methodology – Each session of our pair analytics study [KF14] was a recorded video call with a domain expert and lasted 120-150 minutes. All sessions were split into three phases: introduction and elicitation of expectations (20 minutes), system use with intermittent questionnaires (45-70 minutes), and a semi-structured interview.

Participants – We recruited four participants (2F/2M) from political science (P1–P3) and computational linguistics (CL1). They are Ph.D. students (P1, P3, CL1) and postgraduate researchers (P2) and have previous experience optimizing topic models. While P1 and P3 were interested in getting a broad and detailed overview of the data, respectively, CL1 was interested in which topic was occupied by which speaker in order to analyze framing effects later.

Dataset – All participants refined an IHTM of the second 2012 Obama-Romney debate, as they all had experience with analyzing political debates, but not with this particular debate prior to the study.

Expectations towards Guidance – Participants noted that the expected model results depend on the envisioned downstream task (P1 & P3) and decided to optimize for broader (P1 & P2) or more fine-grained (P3 & CL1) results. CL1, who had previously manually op-
timized topic models, was sure that the results she could achieve using visual analytics “will surely be better and get closer to what my corpus is actually like”. P3 appreciated guidance but noted the importance of human control in the process. He explained that guided analysis should only make “testable” suggestions – a requirement that our system fulfills with preview and comparison views. P3 further expected that he would get annoyed quickly if the agents did not do what he wanted. In such a case, he would prefer for them to “just let me do it on my own”.

Participants quickly grasped the agents’ adaptive nature, comparing them to their changing feeds on social media platforms. Drawing from this experience, P3 was worried that the system would not “recommend new options” and put too much emphasis on the initial interactions without providing controls to override this shortcoming or reverse a previous interaction. He appreciated that our system only uses certain rules above a weight threshold as an approach to preventing this behavior. P3’s concern raises the question of balancing exploration of the potential model state-space with the exploitation of the expressed preferences. Most participants were aware of the risk of creating a “bubble” (P3) in which models would also learn the analyst’s inherent biases (P1).

**Design and Usability** – All participants frequently work with larger amounts of text semi-manually. P1 and P2 expressed that they would typically “spend a lot of time reading documents” (P1). P2 felt that it was difficult for her to make well-grounded decisions based on the set of keywords and instead fell back to reading the underlying text before making a decision. In general, all participants found the topic labels not always immediately obvious, with CL1 saying “the keyword extraction is bad”. We had initially aimed to mitigate this issue by providing more details on demand but have since exchanged the keyword extraction algorithm. During the study, visual analytics novice P2 adapted her mental model from being overwhelmed (“hard to grasp all this”) to perceiving the system as “more convenient and intuitive” and having “the curiosity to dig deeper”.

**Guidance with Agents** – When an agent re-suggested a rejected optimization, P2 noted that the suggestions would not provide “enough diversity”. We are now preventing agents from immediately repeating a rejected suggestion. P1 felt “there are no heuristics” to distinguish the quality of simultaneous suggestions since he had not noticed the certainty-based opacity changes to the preview icicle. A future version of the system could display which agents align best with user preferences and which ones are more exploratory. P2 proposed introducing agent-specific “training sessions” to update the agents in batches rather than training them during the model optimization.

As participants were aware that the agents learned from their interactions, they often traded off between teaching the agents and progressing their refinement. CL1 debated how much time to spend reviewing unexpected suggestions, trading off her immediate analysis goal and the model profiting from her feedback when spending time to review the suggestion carefully. Similarly, P1 kept rejecting unwanted suggestions, saying that providing feedback was “not helpful for me, but helpful for the model.” When P3 encountered a suggestion that did not fit his analysis plan, he instead ignored it and let it decay over time to avoid teaching something wrong and later spending much time on re-training.

**Guidance Timing and Helpfulness** – While using the system, participants were asked to rate the timing and helpfulness of suggestions on a scale from 1 (bad) to 5 (good) three times in fixed intervals. On average, they rated the timing 3.75, 3.75, and 4.25 at the three time points (overall average 3.83, SD 0.72). Helpfulness was rated 3.75, 4, and 3.75 (average 3.92, SD 0.67). While we observe an upwards trend in timing ratings as the analysis progresses, more participants in a more controlled study are needed to confirm the significance of this trend. Both P1 and CL1 stated that they found it difficult to judge the extent to which the system had learned from their interaction. However, the quantitative evaluation suggests that agents, in fact, learn tangible knowledge from interactions. Consequently, further research is needed to investigate how context adaptation of the agents can be translated to more readily noticeable system behavior changes.

### 7.2. Quantitative Evaluation: Model Quality & Transferability

The expert user study showed the applicability of our approach and found that users liked interacting with their agents to train them. To evaluate the agents’ impact on the resulting topic model, we performed three quantitative experiments targeting different aspects.

**Datasets** – Throughout all evaluations, we use two datasets: the first (DB1) and the second (DB2) presidential debate of Obama and Romney in 2012 [CNNb, CNNa]. We chose the second debate because we had used it in the expert user study and the first debate to verify that the learned contexts are transferable to similar datasets.

**7.2.1. Topic Ranking Task**

Our previous work on the automatic refinement of topic models has shown that optimizing models towards improved quality metrics actually worsens their perceived quality when compared to a baseline [EASD’19]. To verify the validity of the agents’ learnings, we repeated our previous perception experiment and let study participants rank optimized and unoptimized model results.

**Methodology** – We first prepared a baseline IHTM result for both the first and the second debate. We then trained the agents through the guided refinement of the second debate (supervised result). The agents’ suggestions were manually reviewed by a topic modeling expert and accepted or rejected following our previously presented methodology of aiming to match an established, expected topic distribution [EASD’19]. We extracted the agents’ learned user preference models and used them to perform automatic optimizations on both debates, where we accepted all presented suggestions. Whenever two agents made a suggestion at the same time, we selected the agent with higher confidence and resolved draws using random picks. We use the five obtained modeling results in two independent ranking tasks between the three results for the first debate and the two results from the second debate. Participants were asked to rank the results (given as lists of topic descriptors) on a scale from 1 (best) to 3 (worst) and 1 to 2, respectively, and to justify their decision.

**Participants** – In addition to three participants from the qualitative evaluation, we recruited five new ones (overall: 5F/3M) with various backgrounds (computational linguistics, political science, computer science) to avoid biased results based on previous involvement.

**Results** – The results from 40 ranking annotations are summarized in Table 1, and show that, on average, automatic refinement

© 2021 The Author(s)

Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.
was received better than the baseline with a rank of 1.9 and 1.1, respectively. This constitutes an improvement over our previous work where the baseline clearly outperformed automatic refinement [EASD’19] and suggests that the contexts learned by agents are better suited for suggesting refinements than simply optimizing towards quality metrics. From the participants’ comments, we conclude that they either preferred results with few topics (supervised result) or with many topics (baseline result) for the second debate. As a result, we observe a high standard deviation in the ranking, and no model gains a clear advantage. For the first debate, the ranked results had similar numbers of topics, and participants focused more on content-features (mentioning keywords in their justifications). As justifications for their preferences, participants particularly pointed out the better separation of topics and their cohesive granularity.

7.2.2. Contextualization of Agents

The ranking task confirms that agents can learn context rules that can be used to refine models automatically. In this experiment, we aim to assess to what extent the contextualizations learned by agents matches user expectations in a given modeling scenario.

Methodology – We sampled nine intermediate modeling states from the automatic refinement processes of the first and second debate in which either a combine, split, or reinsert worst optimization was suggested by the respective agents, and one state in which none of the agents suggested an optimization. For each state, participants were provided with a screenshot of the icicle plot (annotated with topic IDs) and a list of topic descriptors and asked to specify which optimization (combine, split, reinsert worst, or no action) they would have performed (on which topics), justify their decision, and give their certainty on a scale from 1 (uncertain) to 5 (certain). We only selected optimizations from the three most active agents to avoid overwhelming users.

Participants – The task was performed by the same eight participants that produced the result ranking annotations.

Results – Participants produced 70 context annotations with an average certainty of 2.62 (SD 0.58). Most frequently, they chose to reinsert topics (23); all other operations were almost equally distributed (15 combine, 16 split, 16 no action). There were minimal differences between the agent’s prediction accuracy for the three optimizations. On average, agents successfully predicted the operation that participants wanted to perform in 35% of cases. For the three participants from computer science, this value increased to 56%. Considering that a computer scientist trained the agent rules, those results suggest that different user groups might need substantially different agents. As one of our participants pointed out, a common issue in modeling discourse data is the creation of a so-called moderation-topic. She noted that “this topic does not seem to have any specific content, and many of the descriptors are only expressions for discourse transition.” In many cases, our participants focused on trying to optimize this topic. This is particularly apparent in participants from computational linguistics, where 47% of responses are targeted at splitting or removing this specific topic. On the other hand, computer scientists consistently ignored the moderation topic and focused on other optimization targets.

The evaluation also shows that all but two participants had very skewed preference distributions for splitting or reinserting topics, primarily picking one or the other. However, users from both groups refer to the same topics in their justification as both operations intuitively have a similar effect. Artificially removing the difference between operations would increase the overall accuracy to 52%. We suspect that users from both groups would train similar contexts for the split or combine agent, respectively, and achieve similar final modeling results. These initial findings highlight the importance of personalized and adaptive guidance and suggest that different user groups have differing preferences.

7.2.3. Contexts Learned and Rule Transferability

Our last experiments investigate the transferability of the learned user preference models between datasets. Such transferability is the foundation for continuous agent training across users and corpora as it shows that agents can learn generalizable rules rather than overfitting on specific analysis contexts.

Methodology – Through an exploratory analysis, we aim to verify the finding that learned user preference models can be transferred between datasets. Before presenting the results, we define two core terms: A context is seen if it is closed and considered for classification by the agent at least once. A context is applied if it is seen and has been used in the classification at least once. Recall that only those contexts with a sufficient weight are used in this decision.

Results – First, we investigated how many contexts were seen and applied during agent training. Table 2 shows that agents saw up to 142 contexts and, on average, applied 24.72% of contexts at least once. Agents with a larger impact on the result (e.g., combine and split) are accepted much more frequently and learn significantly more contexts. The table also shows that comparable amounts of contexts are seen when automating refinement on the same or similar datasets. Here, agents apply 18.66% and 15.99% of seen contexts.

Next, we were interested in verifying that the similar numbers actually manifested in the same contexts being encountered. Table 3 shows the percentage of encountered contexts during automation were also encountered during training. All agents saw at least 48.54% of training contexts (split agent). This number increases with an increasing ratio of seen and applied contexts in training, highlighting that confident rules are more likely to be transferred. As a result, the remove chains agent that learned many unconfident contexts was not able to make a suggestion in automation. In contrast, both the

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Type</th>
<th>Rank (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Second Debate</td>
<td>baseline</td>
<td>2.0 (.93)</td>
</tr>
<tr>
<td></td>
<td>automatic supervised</td>
<td>1.9 (.64)</td>
</tr>
<tr>
<td>First Debate</td>
<td>baseline</td>
<td>1.9 (.35)</td>
</tr>
<tr>
<td></td>
<td>automatic supervised</td>
<td>1.1 (.35)</td>
</tr>
</tbody>
</table>

Table 1: Average annotator rank for different models from two corpora. Users consistently preferred models automatically optimized through agents over the baseline.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Type</th>
<th>Rank (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training DB2</td>
<td>120 (27)</td>
<td>103 (16)</td>
</tr>
<tr>
<td></td>
<td>10 (7)</td>
<td>98 (12)</td>
</tr>
<tr>
<td>Automation DB2</td>
<td>139 (19)</td>
<td>100 (25)</td>
</tr>
<tr>
<td></td>
<td>13 (1)</td>
<td>120 (35)</td>
</tr>
<tr>
<td>Automation DB1</td>
<td>118 (15)</td>
<td>100 (14)</td>
</tr>
<tr>
<td></td>
<td>9 (1)</td>
<td>110 (29)</td>
</tr>
<tr>
<td></td>
<td>7 (0)</td>
<td>107 (24)</td>
</tr>
</tbody>
</table>

Table 2: Counts for contexts seen (applied) are similar during training, automation on the same dataset, and transfer to unseen data.
We have presented a technique for the adaptation of guidance based on transferable contexts across two target corpora. As a result, they were not applied during the training session. As a result, they applied up to 2.4 times as many contexts during automation.

Finally, we use the \( G^2 \) [Dun93] metric to determine the most characteristic contexts for each agent. Table 2 shows that up to 45% of characteristic contexts learned in training remain characteristic during automation, suggesting that they capture distinctive, generalizable situations. These findings illustrate that we can obtain refined user preference models by continuously training agents over long periods of time and across datasets. However, future research and long-term evaluations are needed to confirm this finding.

**Table 3: Ratios of transferable contexts across two target corpora.**

<table>
<thead>
<tr>
<th>Context</th>
<th>Corpus</th>
<th>Combine</th>
<th>Split</th>
<th>Small</th>
<th>Outlier</th>
<th>Chains</th>
<th>Worst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seen</td>
<td>DB2</td>
<td>86.67%</td>
<td>48.54%</td>
<td>90.00%</td>
<td>93.88%</td>
<td>53.83%</td>
<td>64.08%</td>
</tr>
<tr>
<td></td>
<td>DB1</td>
<td>58.33%</td>
<td>58.25%</td>
<td>90.00%</td>
<td>112.24%</td>
<td>53.83%</td>
<td>75.35%</td>
</tr>
<tr>
<td>Applied</td>
<td>DB2</td>
<td>48.15%</td>
<td>75.00%</td>
<td>14.29%</td>
<td>216.67%</td>
<td>0.00%</td>
<td>116.67%</td>
</tr>
<tr>
<td></td>
<td>DB1</td>
<td>55.56%</td>
<td>87.55%</td>
<td>14.29%</td>
<td>241.67%</td>
<td>0.00%</td>
<td>133.33%</td>
</tr>
<tr>
<td>( G^2 )</td>
<td>DB2</td>
<td>30.00%</td>
<td>15.00%</td>
<td>45.00%</td>
<td>10.00%</td>
<td>15.00%</td>
<td>30.00%</td>
</tr>
<tr>
<td></td>
<td>DB1</td>
<td>30.00%</td>
<td>15.00%</td>
<td>30.00%</td>
<td>0.00%</td>
<td>5.00%</td>
<td>15.00%</td>
</tr>
</tbody>
</table>

8. Discussion

We have presented a technique for the adaptation of guidance based on learned user preference profiles, and implemented and evaluated it in the context of topic model refinement. Incorporating Semantics – Our evaluation of agent contextualization has shown that participants frequently aimed to optimize one specific topic. As context vectors in our implementation only contain information about metrics and modeling events, they are not ideally suited for capturing those particular interaction patterns. As a first step to addressing semantically different expectations, our implementation provides alternative suggestions for several agents. However, future work could investigate the integration of semantic representations into the context vectorization, thus enabling a more personalized contextualization of guidance.

Teaching Agents – Participants in our expert user study showed a noticeably high motivation to teach their domain knowledge to their agents and asserted that optimizing such a guidance model, even over the span of several days, is a worthwhile investment. Even though our system represents the agents as simple cards and not, e.g., animated avatars, the concept of training their personal set of agents seemed to increase engagement in our participants. However, we observed that despite explicitly aiming to teach their agents, users became more susceptible to confirmation bias after a short time of using the system: while some were initially skeptical of the agents’ quality, they rapidly grew comfortable in accepting guidance without strict quality assessment. This finding from our study reinforces uncertainty communication and the mitigation of confirmation bias as central challenges for guidance in visual analytics.

Generalizability – In contrast to other existing approaches, our technique relies on a vectorization of the analysis context that captures the temporal development of metrics and events rather than similarity to previous interactions. In particular, it is not specific to topic model refinement but could be generalized to other model building tasks in visual analytics that are supported by quality metrics and existing refinement strategies. A considerable challenge to tailor our approach is the identification of a suitable context vectorization. The presented implementation uses a sliding window and computes the difference between subsequent metric values (i.e., window size = 1).

A suitable vectorization for other domains should consider the trade-off between context complexity and both the required amount of training data and the perceived agility of agents during interaction.

Classification Algorithm – In our implementation, the context vectorization relies on diverse topic model quality metrics and HMT-specific events to capture the model development and encodes this information at different levels of granularity in full and partial context vectors. Agents learn at all granularities and filter context vectors by closedness when classifying to avoid duplicated information biasing the result. As a result, a rule-based classifier and a simple approach to relevance feedback proved effective. It is likely that more complex machine learning techniques might enable even better performance, potentially at the cost of intelligibility, cold-start issues, and implementation complexity.

Limitations – The proposed technique currently requires an initialization with (synthetic) pretraining data to be effective. Even with proposed simplifications like metric shapes, the state space spanned by context vectors is large, and training a classifier from scratch is not feasible with the typical number of interactions in a visual analytics session. The technique’s design aims to combat this issue by learning (more) generalizable knowledge on partial contexts. Our evaluation suggests that rules are transferable between datasets, with automatic optimization outperforming the baseline. While future research is needed to confirm this finding, transferable rules would enable continuous training of agents, overcoming this limitation.

In this paper, we provide four distinct evaluation approaches to validate our technique from both human-centered and algorithm-centered perspectives. However, the paper only instantiates the technique for the iterative refinement of topic models. Further investigations are necessary to evaluate how the technique performs on different tasks and over longer periods of time. Additionally, the evaluation can only establish single-objective, context-dependent agents as one possible way to provide guidance. In particular, it does not provide a comparative evaluation against other guidance techniques.

9. Conclusion

We have presented a visual analytics technique for co-adaptive guidance through contextualized preference learning and instantiated it in an approach to topic model refinement. The technique is centered around guidance agents that, over time, learn user preferences to provide distinct guidance suggestions in specific contexts. The flexible nature of the context vectorization enables system designers to choose the metrics, events, or semantic representations that are most relevant to their application scenario. Our initial evaluation within a qualitative user study shows that the technique is accepted and valued by domain experts. The quantitative results show that context rules learned by agents can be used in automatic topic model optimization, outperforming previous optimization approaches. We plan to extend our technique on contextualizing guidance to enable agents to personalize the content of their suggestions in the future. Our system prototype is available at topic-model-guidance.lingvis.io.

Acknowledgements – This work has been partially funded by the
References


© 2021 The Author(s)
Computer Graphics Forum © 2021 The Eurographics Association and John Wiley & Sons Ltd.

