PIHVI: Online Forum Posting Analysis with Interactive Hierarchical Visualization

Jaakko Peltonen*, Ziyuan Lin*, Kalervo Järvelin*, and Jyrki Nummenmaa*
jaakko.peltonen@uta.fi, ziyuan.lin@uta.fi, kalervo.jarvelin@uta.fi, jyrki.nummenmaa@uta.fi

ABSTRACT
We introduce PIHVI: a novel interactive system for visualizing and exploring a large hierarchical text corpus of online forum postings. The main view of the visual interface shows a large-scale scatter plot, created by flexible nonlinear dimensionality reduction based on text contents of the postings, and we couple it with a coloring optimized by a second dimensionality reduction to represent the forum hierarchy. We exploit the hierarchy to provide data-driven summaries of plot areas at multiple levels of detail, allowing the user to quickly see and compare both the content-based similarity of groups of posts and how near they arise in the forum hierarchy. A user can move between hierarchy levels, mark posts or spots of interest, filter posts by content similarity and by location within the hierarchy, and inspect post contents. Experiments show the interface can reveal hidden semantic relationships between postings that would be hard to find based on the known hierarchy alone.

ACM Classification Keywords
D.2.2 Software Engineering: Design Tools and Techniques User interfaces; H.5.2 Information Interfaces and Presentation: User Interfaces User-centered design; I.2.7 Artificial Intelligence: Natural Language Processing Text analysis

Author Keywords
user interaction; visual analytics; text analysis.

INTRODUCTION
Online discussion often takes place in venues that have a hierarchical organization, such as large message boards organized in sections, subsection, sub-sub-sections and so on. While some venues such as specialized message boards may use a relatively flat organization having a small number of sections directly under a root, other venues such as general-interest message boards have an extensive multilevel hierarchy forming a deep tree of sections. Users post new threads, and post comments and replies under threads, but the hierarchy of sections is typically designed by owners of the discussion venue or administrators of the venue. Online forums are big data: as a case study we use a popular Finnish discussion forum "Suomi24" (www.suomi24.fi) spanning 16 years and 6.5 million threads. Hierarchically organized discussion occurs also for example in user commentaries on news sites whose news have a hierarchy, in online reviews of products that have a hierarchy, and so on; our work is applicable to such hierarchies as well but we focus on the discussion forum case.

In discussion forums, sections and their hierarchy are created by administrators as a simplified division of discussions intended to represent a subset of prototypical user interests, and such a division does not suffice to describe the true variety of semantic content in online discussion. Our target users are administrators and analysts of the forums: if the hierarchy does not support the content, forum owners can add subsections or shortcuts to promote content, aiming to increase user engagement with the venue. An important task in data analytics of online forums is to perform visual analytics of how discussions and their underlying topics spread across the hierarchy. When the data and the hierarchy are large this becomes a hard task, as exhaustive inspection of all data or comparisons across all sections become unfeasible: computational support is then needed. However, most visual analytics tools for social media discussion focus on other aspects such as temporal evolution or networks of conversation participants.

We propose a new system for analysis of online conversation in hierarchical forums in a large scale, by an interactive interface that scales to large data (Figure 1; details in later sections). We focus on analysis of content variation and its relationship to the hierarchical sections; we do not focus on e.g. temporal filtering or other such aspects which are extensively studied in other systems and can easily be integrated as additional views for our system. Main novelties: first work to use dimensionality reduction based linking of views through optimized coloring; among the first to use large-scale dimensionality reduction in visual analytics of discussion venues; novel multi-level inspection of a large-scale scatterplot; novel use case - analysis of suitability of existing hierarchy in a venue to describe content of the venue. We next discuss related work, introduce the system, denoted *PIHVI - Posting Analysis with Interactive Hierarchical Visualization*, then describe our machine learning solution, describe a user study and discuss conclusions.

RELATED WORK
Visual analytics for social media has been an active research topic, as social media data are rich and heterogeneous in structure and thus difficult to model. A recent survey [2] taxonomized methodologies on the field by *Data sources* – Twitter,

*All the authors are affiliated with University of Tampere, Finland. The work was supported by Academy of Finland decisions 295694 and 313748 and by Aalto Science-IT.

ESIDA '18, March 11, 2018, Tokyo, Japan

Facebook, blogs, online forums, etc.; *Entities of data* – network, geographic information, or text content; *Goal of visualization* – pattern extraction, visual monitor, anomaly detection, etc.; *Visualization techniques* – scatter plots, trees, node-link diagrams, flows, etc.; *Target users* – general public or analysts; *Applications* – journalism, politics, finance, sports, advertisements, etc. Our PIHVI system can be categorized as using interactive hierarchical scatter plots to analyze text content of general online forums for analysts, with the goals of extracting interesting patterns of threads, detecting outliers of threads like spam or novel topics, and advertising, to name a few. Below we summarize related work on using other visualization techniques to analyze aspects of online discussions.

Semantic features of threads in online forums. Bag-of-word (BoW) vectors like TF-IDF are typical to represent text content for analysis [9, 13]. Topic facets to extract sentence level sub-topics are considered [6], The facets are shown to reflect programming concepts for questions in programming discussion forums, visualized in a prototypical system. An alternative hierarchical semantic model for sentences incorporating discrete Fourier transform with a mixture of topic models was proposed [4]. In this first publication on PIHVI, we focus on the main goal of studying content relationships to the known hierarchy within the venue, and in order not to muddle that focus we use simple BoW-based features which already work well in capturing thread similarities in our experiments. Other derived features like topical features are possible, the rest of our method works for any choice of features.

Showing posts from individual threads. In PIHVI, we emphasize exploration based on the content-based hierarchical scatter plot and the related tasks, thus we use a simple detail view showing plain texts from selected threads. However, we are aware of works that visually display thread content. This line of research usually focuses on visualizing the structure of the thread under inspection. For example, thread structures can be encoded with indented rectangles [3], where each indentation corresponds to one level of nested reply to a post. Such thread visualization is orthogonal to PIHVI novelties and can be easily integrated into PIHVI.

Dimensionality reduction in visual analytics. Dimensionality reduction has been brought into the pipeline of visual analytics by providing a 2D representation of threads, to help users to gain an overview. It is related to semantic feature construction discussed above, if obtained features are given to dimensionality reduction algorithms. For example, t-distributed Stochastic Neighbor Embedding (t-SNE) [7] has been considered [16]. In PIHVI, we build our hierarchical exploration functionality on top of a t-SNE based embedding. See Section 4 for details. Another branch of studies in this line is to let a user personalize the representation by interaction. See, e.g., our work in [11, 10, 12] and references in [14]. We point out the novel goal of PIHVI is to study relationships of large-scale content to a large-scale hierarchy, hence the novelty is in incorporation of the hierarchy into the content visualization and in hierarchical exploration of content through multiple views.

Forum analysis system combined with different visual analytics. There have been integrated systems with views of various visual analytics for inspecting online forums. Simple text search does not suffice when users aim to broadly comprehend the information space with multiple exploratory searches and information needs [17]. For example, a set of visualization designs showing different aspects from MOOC forums, including posts, users, and threads at different scales is considered, with an aim to support online education with the forums [5]. A recent work [19] focuses on revealing anomalous information spreading occurring in social media along the time-line with coordinated views, which particularly contains a scatter plot of threads created from Multidimensional scaling (MDS) [1] without the hierarchical functionality. We provide views that help gain comprehensive understanding of ongoing discussion and how it relates to a multi-level hierarchy of the whole discussion venue, rather than on relationships within individual threads; moreover our focus is on mostly anonymous discussion where most users are unregistered so user or user-group views are less useful.

OVERVIEW OF THE SYSTEM: PIHVI - POSTING ANALYSIS WITH INTERACTIVE HIERARCHICAL VISUALIZATION

We now describe the proposed system *PIHVI - Posting Analysis with Interactive Hierarchical Visualization*, on an overview level, describing the design principles and how the system operates. In the next section we describe how the system is implemented through dimensionality reduction methods.

Design Principles. The design principles of the PIHVI system are: **1.** The system should *illustrate, in a compact way, the overall variety of discussion over the online forum*, both in terms of semantic content, and the hierarchical variety of discussion sections where threads can be posted. This answers the question: "Which kinds of topics are users discussing, and in which sections is the discussion happening?" **2.** The system should let analysts *efficiently browse semantic content of threads across the forum*, on an overall level of trends and on more detailed levels of individual thread groups, without restricting browsing to boundaries of specified sections. This answers the question: "Which discussion trends are individual threads of interest related to?" **3.** The system should *compactly show relationships of threads* in terms of semantic content and position in the hierarchy. This answers the question: "What other similar threads exist for a thread of interest?"

The above design principles should hold even when the number of threads, the number of sections in the hierarchy, and the depth of the hierarchy, all become large. The resulting system for visual analytics should support key tasks in Shneiderman's [15] taxonomy of tasks for information visualizations, including the visual information seeking mantra *overview first, zoom and filter, details on demand, and relate*.

System Components. Following the design principles in Section 3, we devise the following components for large-scale interactive analysis on hierarchical forum posts, and to support our user tasks in PIHVI. Screen shots of the system are shown in Figure 1. The system has five linked views, where views 1-3 below are always shown and 4-5 are on demand:

1. *Content-based map*: An interactive scatter plot of the entire collection of threads from the online forum, which can be zoomed to several detail levels. The map shows a spatial

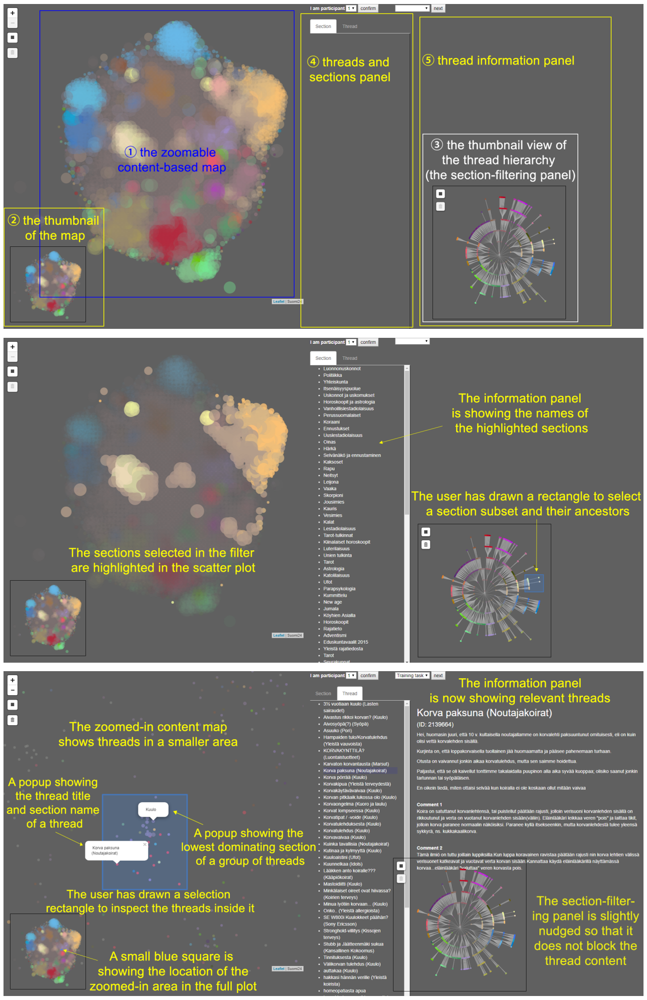

Figure 1: The PIHVI interface. Top: the initial state, where the content-view zoomed out to show an overview of the entire thread collection. Middle: section-filtering state. An analyst selects a subset of sections in the section panel, circles of corresponding sections are then highlighted. Bottom: zoomed-in state for an analyst exploring a cluster of hearing-related threads from different sections across the forum; the analyst has selected a group of threads with a bounding box and is viewing their titles and content of an example thread.

organization of threads by content similarity: similar threads are shown nearby. The map is created by machine learning based dimensionality reduction. The map is overlaid with colors representing sections and their similarity, again created by dimensionality reduction. At detailed zoom levels, each thread is shown as a dot, and colors of the dots indicate the section of the threads. At less detailed zoom levels, groups of threads are shown as dots with diameters scaling with the numbers of threads in the groups, and colors of the dots indicate a *dominating common parent section* of threads in the group, as detailed later. We draw the different-sized dots with moderate transparency, yielding color blending when there are overlaps between the dots, which *approximates the content and section distributions*: higher opacity reflects a higher density of threads and vice versa; color mixing suggests mixtures of dominating sections sharing similar content. Both patterns may interest a forum analyst/administrator. For interactivity, each dot, from different zoom levels, can be inspected by clicking to show a pop-up: the pop-up may contain either the thread title and the name of its section, if the dot represents a single thread, or the name of the dominating common parent section, if the dot represent a group of threads. To avoid occlusions of multiple pop-ups, the analyst can bring a pop-up to front or back by clicking or dragging it; also, when a pop-up is mouse-overed, the remaining pop-ups without the cursor hovered will turn half-transparent. Buttons on this view allow the analyst to drag a rectangle over an area of the map: all threads in the area will be described in the Threads and Sections panel.

2. *Thumbnail view of the map*: This view shows the entire collection and a rectangle denoting which portion of the map an analyst is currently viewing; this lets the analyst maintain *focus and context* while browsing the collection.

3. *Thumbnail view of the thread hierarchy*: This view shows a radial plot of the tree hierarchy of discussion sections, and shows how optimized colors of the sections vary across the hierarchy. Due to space constraint, showing all 2434 section names at once is not feasible: we show section names on demand when the analyst mouse-overs the tree nodes of sections. The window size can be enlarged by standard zoom controls to help mouse-over the nodes. This view allows the analyst to relate where each thread or group of threads seen in the content view arises from in the hierarchy. Buttons on this view allow the analyst to drag a rectangle over a subset of sections: those sections will be highlighted in the content-based map.

4. *Threads and Sections panel*: When the analyst has selected an area of the content map, all threads in that area will be listed here as a scrollable list. The panel has two tabs: the Thread tab lists the threads by title and section, and the Sections tab lists all unique sections from which the threads arise.

5. *Thread Information panel*: When the analyst clicks any individual thread in the content-based map (individual dot that represents a single thread in the scatter plot) or clicks the title of any individual thread in the Thread tab of the Threads and Sections panel, the full text of that thread is shown here.

Walkthrough example. Here we depict a typical usage of PIHVI for thread discovery. Suppose an analyst is interested in sport and fitness related discussion but does not yet know how such discussion is related to other discussion within the Suomi24 forum. This can be a possible sequence of findings and interactions from the analyst when discovering the threads of interest: **1.** Initially the analyst views the content view zoomed out to show all Suomi24, and notices several clusters of color, indicating discussion similar in content and arising from tight branches of the section hierarchy.

2. The analyst clicks some of these clusters to find a starting point for analysis. After a time the analyst find a green cluster (at the bottom in Figure 1 top) whose pop-up label shows the dominating section in the cluster is Sport and Fitness ("Urheilu ja kuntoilu"). Alternatively, the analyst could have browsed the section tree to find a Sport and Fitness section.

3. The analyst then zooms in to view individual threads within the cluster, and draws a rectangle around them to quickly view

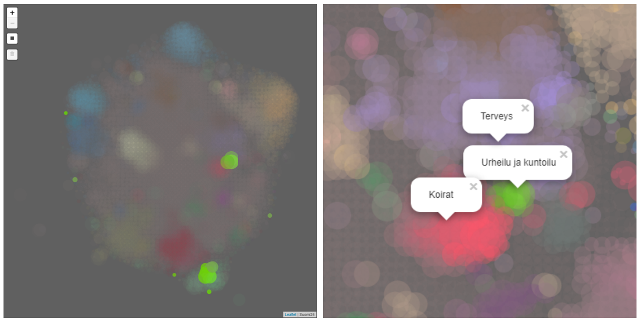

Figure 2: Possible interactions when an analyst is discovering threads of interest. **Left**: the analyst has used section filtering to discover several locations where Sport and Fitness ("Urheilu ja kuntoilu") is discussed (green highlighted clusters). **Right**: the analyst zoomed in towards the middle-right green cluster, and inspected nearby other clusters: near the green Health and Fitness cluster ("Urheilu ja kuntoilu") are two clusters, Health ("Terveys"; violet color) and Dogs ("Koirat"; red color).

their titles. The analyst discovers many of them are about ice hockey, a popular sport in Finland, such as threads about Finnish ice hockey teams and leagues and about related sports like "Jääpallo" (Bandy in Finnish).

4. The analyst draws a rectangle around the Sport and Fitness section and its nearby sections, to detect where else such threads occur in the content map: Figure 2 (Left) shows that besides the ice hockey related cluster at the bottom, there are small isolated clusters (light green dots) at edges of the map, and a large cluster at center right in the map. Thus the analyst discovered a new cluster of Sport and Fitness discussion.

5. The analyst removes the section filtering since he/she is interested in all sections in the area. The analyst zooms towards the discovered cluster, and notices it is located at the border of two larger clusters (Figure 2 Right): clicking the clusters reveals their dominating sections as Health ("Terveys"; violet colored cluster) and Dogs ("Koirat"; red colored cluster).

6. Since the three clusters are close-by, this indicates they may have common topics. The analyst is then interested what commonalities there are in threads arising from Sport and Fitness and in threads arising from Health, so the analyst zooms in more on the plot, and draw a rectangle around the boundary between "Terveys" and "Urheilu ja kuntoilu".

7. The topics of the selected threads are listed in the detail panel; the analyst finds several threads on weight-loss in the information panel.

8. The analyst could continue investigating Sport and Fitness (e.g. moving to view some of the other small clusters of Sport and Fitness threads), or could e.g. investigate commonalities between Health and Dogs in the same window–this would turn out to be discussion of allergies when living with dogs.

The above described interaction can go on as long as desired by the analyst. For an analyst, e.g., in the forum administration team, such exploration may help understand the distribution of the threads, which can inspire, for example, reorganization of forum sections, or a digest on hot discussion topics.

Properties of the System. As desired in *Design Principles*, the proposed system directly supports several key tasks in Shneiderman's [15] taxonomy, in particular the visual information seeking mantra is supported as follows:

Overview: The two linked views (content-based map and the tree of the section hierarchy) together show an at-a-glance overview of threads in the entire online forum. The radial tree view shows an at-a-glance overview of the section organization: sections with numerous child sections are shown as wide branches, and sections with multiple levels of subsections under them are shown as long branches. On the other hand the content-based map shows an at-a-glance overview of semantic content across the whole forum, organized so that similar content is shown nearby in the map, and groups of similar threads appear as clusters; moreover, section colors overlaid on the map provide an easy overview of which sections have similar content, which sections have varied content appearing in multiple places on the map, and what the lowest common hierarchy section is for some cluster of content.

Zoom and filter: the analyst can zoom into the levels of detail in the content-based map with the mouse wheel, and can filter threads from both linked views with selection boxes, e.g. to select threads in a particular location in the content map and highlight threads from a particular branch of the section tree.

Details on demand: clicking on threads in the content based map shows their content, clicking on thread-clusters gives the name of their lowest common section, and creating a selection box over an area of the content map opens a linked view listing all threads in the area and all sections of those threads; and clicking any thread in that list shows the thread content.

Relate: the content-based map relates threads by their content similarity by placing similar threads nearby in the map; the section tree relates sections by showing their parent-child relationships; and the overlay of section colors onto the map relates similarity of thread content to similarity of their sections. Inspecting these two linked views lets analysts quickly study how threads and sections at different branches in the hierarchy relate to each other in terms of content, and study how a cluster of semantically similar content is related to multiple sections where discussion of the content happens in the forum.

As noted above, PIHVI includes filtering/highlighting of the content map by selected sections. Highlighting could also be done by results of text lookup searches; we do not focus on simple text search for several reasons: 1. Our focus is not on searching individual documents but on overall comprehension of the forum where information needs of the user evolve as he/she learns about the data; such comprehension would not be well served by individual text lookups. 2. Naive text search does not take into account complexity of natural language and could miss related threads if they did not share search terms (due to synonyms etc.), whereas the content map is organized by *overall similarity of all words* in the documents, and serves exploration of discussion content better than simple search would. Nevertheless, where text lookup is needed it is trivial to highlight lookup results like we do to selected sections.

We will compare PIHVI to a baseline system we implemented to represent a more traditional interactive section-based browsing interface, described in the *User Experiment* section.

DIMENSIONALITY REDUCTION FOR THE PIHVI SYSTEM

We describe the dimensionality reduction solution for creating the two main views – the content-based scatter plot and its color-based link to the section tree view, and then describe how zooming is implemented efficiently and using the hierarchy for summarization of thread groups.

Content-based Representation of Textual Data. PIHVI is initiated by a content-based two-dimensional scatter plot of threads on the online forum (here Suomi24), followed by other interface components facilitating exploratory data analysis on the forum. In PIHVI, dots are discussion threads in the "content-based" scatter plot. By "content-based", we mean to pursue a two-dimensional representation such that 1. if two threads are similar in content, they should be placed nearby; and 2. if two threads are dissimilar in content, they should be placed faraway. We use BoW models to capture (dis)similarities between content of threads: the more words two threads share, the more similar they are in content. We extract thread features based on BoW as follows, and perform a nonlinear dimensionality reduction with neighbor embedding.

BoW-based features in high-dimensional space. We start with lemmatized tokens of Suomi24, provided by the Finnish Language Bank service ("Kielipankki", http://www.kielipankki.fi). The resulting data set contains the lemmatized version of each token in each thread, along with other meta-data of the token. We first calculate the TF-IDFs for each token, constituting raw features of each thread. We further process the raw features as follows: **1.** Remove stop-words based on a Finnish stop-word list; **2.** Only keep tokens that appear in at least k threads, to discard incidental tokens that do not carry shared semantic content, such as rare URLs or non-word character strings. **3.** For each thread, keep only l tokens with highest TF-IDF values. With stop-words already removed, we assume these kept tokens are more informative than others in each thread. **4.** Normalize each obtained feature vector by dividing vector entries with their sum. We feed the final features into t-SNE to obtain the two-dimensional content-based representation for PIHVI. Considering the size of our corpus, t-SNE is a suitable choice since 1. t-SNE has been shown to work well with high dimensional data; 2. scalable t-SNE variants exist. We use a tree-based implementation [18] to achieve a $O(N \log N)$ complexity for N threads.

Section Distance based Coloring for Linking the Views. In the thumbnail view of the section hierarchy, we show a radial tree plot of the hierarchy and draw section nodes in different colors so the analyst can distinguish the sections and link them with the content-based map. There are a huge number of sections, and giving each a completely different color is not possible, thus assigning colors to successfully link the two main views is nontrivial. We assume the analyst wishes to notice similar content from far-apart sections of the hierarchy, and highly different content from close-by in the hierarchy, hence color differences should correspond to distances in the section hierarchy. We use dimensionality reduction to optimize colors so that the most important differences are encoded with the most different colors, and close-by sections in the hierarchy (having small path length from one section to the other in the section tree) get close-by colors. This color encoding is a visual cue helping users distinguish sections based on their locations in the hierarchy. We do not require users to perceive section distances exactly from colors; it is enough to notice similar colors (close-by sections) when browsing the content. We take path lengths between sections in the section tree as input distances, and give weights to atomic paths – paths connecting a section to its direct parent or child section with $w_{i \leftrightarrow j} = (C_i + C_j)/2$, where C_i is the number of child sections of section i. Intuitively, if one end of the path has a large number of child sections, the parent is not connected to any particular one of its child sections, implying a looser connection with the other end. The distance between two sections is then the sum of weights of atomic paths connecting them. The distances are reduced to coordinates in a 3D output space by multi-dimensional scaling (MDS): the 3D output coordinates are then normalized to the unit cube where they can be used as RGB color coordinates (more sophisticated color encodings from perception theory are possible [8]). As a result, sections close to the root get mild colors, with deepening saturation for more specialized lower-level sections. We empirically find MDS can yield a more continuous coloring than non-distance-preserving dimensionality reduction algorithms like t-SNE.

Zooming in the Content-based Plot. At the most detailed level in the content plot, threads are individual dots and the user can easily see their content similarities from their spatial relationships and their section similarities from their colors. When the user zooms out, he/she should similarly be able to see content and section similarities of larger groups of threads. However, naive zooming of a large scatter plot would not accomplish this: firstly, in large discussion forums it would not be feasible to redraw huge numbers of threads into a scatter plot quickly; secondly, in areas where threads arise from multiple sections the zoomed-out view would become a clump of different-colored dots, making it hard to notice overall similarities in sections of such threads. We solve both problems by an *intelligent zoom*. The plot area is divided into a grid (the higher the zoom-out factor, the rougher the grid). Instead of plotting individual threads we plot one circle per grid cell, summarizing all threads in it: the circle radius indicates number of threads in the cell, and circles are drawn with slight translucency to allow smooth overlap instead of occlusion. Grid cells are indexed by Hilbert curves in each zoom level for better performance in querying points in cells. For each grid cell the color of the circle must be chosen intelligently as the grid cell contains threads from multiple sections: we compute the color by starting from each thread and traveling up the section hierarchy until we find the lowest section covering at least 40% of threads in the grid cell, denoted the *lowest dominating section*. Thus the color of a circle indicates *diversity of content* in it: the less saturated (i.e., the closer to the color of the root section), the more diverse the content. Such adaptive coloring lets the user quickly browse forum content at multiple detail levels. At each level, circles, their colors, and their overlap yield an at-a-glance view of content and section variation. As the user zooms out, areas of mixed sections become large diffuse circles colored ever closer to the root color; as the user zooms in, such circles break into subclusters with more vibrant colors, indicating thread groups arising from tight branches of the section hierarchy.

USER EXPERIMENT AND RESULTS

We perform a user experiment to show the advantage of PIHVI compared to a baseline, in controlled information seeking scenarios representing typical analytics subtasks. We describe the experiment design (the forum and thread corpus, task design, baseline system we designed&implemented, participants, and evaluation criteria) and then the results.

Details of the Online Forum and Corpus. We run our user experiment on Suomi24, one of Finland's largest online communities: it is mostly in Finnish with several Swedish sections; in 2015 it had 0.8 million unique visitors per week[1]. We consider the realistic scenario where analysts are analyzing data of a specific time frame: here we filter data to the subset of thread from 2005; our methods can naturally be used for data of any other years or for all of Suomi24. We set $k = 20$ and $l = 200$ in the preprocessing told in *BoW-based features in high-dimensional space*. We then have 314871 threads and 80123 tokens in the corpus, arising from 2434 sections.

Task Design. In order to measure the success of our interface for visual analytics of hierarchical online discussion, we evaluate performance in a typical subtask that would often be present as part of an analytics session: **given a thread of interest, find semantically related threads across the online forum**. We point out that given a thread of interest, it is trivial to browse other threads in the same section, and therefore we focus on the nontrivial task of finding related content from *other sections across the online forum*. The ability to find related content comprehensively across an online forum is crucial for comprehensive analytics of online discussion and how topics of interest are spread across the forum.

We choose four cases where discussion on a semantic topic spreads across several Suomi24 sections: for each task, we indicate to the user one thread as a starting point, and ask to find threads on the same topic from other sections. Users carry out a training task to learn the interfaces: *Kuulo* ("hearing" in Finnish; the task is to find threads about hearing outside of the discussion section dedicated to hearing). We pick three tasks for the actual experiments: 1. *Matti Nykänen* (a Finnish ski jumper; the task is to find threads about him outside of the discussion section dedicated to him), 2. *Svenskt* ("Swedish" in Swedish; the task is to find discussions in Swedish outside of discussion sections dedicated to Swedish-language discussion), and 3. *Hiukset* ("hair" in Finnish; the task is to find discussion about hair/hair loss/hair care/hair fashion outside of the section dedicated to hair). Tasks involve several words of interest due to informal names, loose language, etc., and are not fully solvable by naive text search. All the tasks correspond to the research questions we ask in Section 3.

The Baseline System. We compare the proposed system in terms of task performance and usability against a baseline system. The baseline is an interactive system for browsing threads by interacting with a radial visualization of the section hierarchy. It is comparable to the proposed PIHVI system, but lacks the dimensionality reduction based content view and focuses on the section hierarchy only, allowing the user to browse it in a large resolution. This baseline represents a system that would be readily implementable without having access to a dimensionality reduction based visualization of thread content and thread relationships. In the baseline system there are four linked views: **1.** Scrollable large-scale radial plot of the section tree hierarchy: each node in the tree corresponds to one section of the forum.
2. A thumbnail of the whole tree, with a square control for navigating inside the tree; moving the square over the thumbnail changes the corresponding position in the large-scale plot of the tree. This thumbnail view is similar to the one in PIHVI, but focuses on navigation instead of filtering a content-based plot which is not available in the baseline system.
3. A panel that appears when the user clicks on any section in the tree. It shows the thread titles from the selected section. This panel is similar to the threads tab in the PIHVI system.
4. A window for full-text content of threads; clicking any thread in the title panel shows the full-text content, again similar to the corresponding panel in PIHVI.
The thumbnail window and the window containing the thread-list panel and full-text panel are movable by dragging. Figure 3 shows a screen shot of the baseline system. Essentially, the baseline corresponds to PIHVI without a content map, thread content of sections is then naturally accessed through the section hierarchy tree. This makes the two systems comparable.

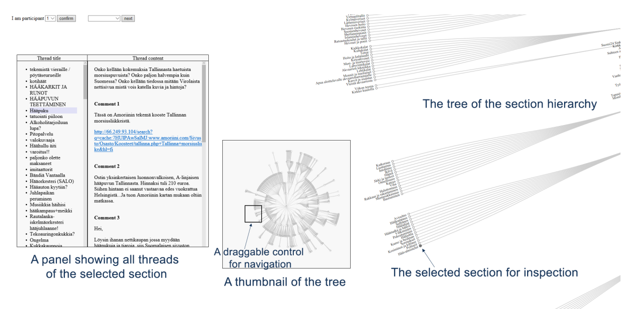

Figure 3: A state of the baseline system. Components are marked with labels: the tree of the section hierarchy, a thumbnail of the full tree with a draggable control for navigation, and a thread information panel showing thread titles and content. In the figure, the threads in the information panel are from the selected section "Häävalmistelut" ("Wedding preparation" in Finnish, marked in black in the section tree).

Participants. We perform a user study with 12 participants, including 1 female and 11 males; all participants are computer science PhD students or post-doctoral researchers who know both Finnish and Swedish. Each participant first performs a training stage on both systems by using the systems to perform the training task. Next, the participant performs all three experiment tasks on both systems (PIHVI and baseline). The order of the task and system combinations (which order of the three tasks; which system first in each task) was randomized and counterbalanced over the participants to avoid bias.

[1]As reported in https://goo.gl/Ukt49K.

Evaluation Measures. We evaluate the experiment results in four ways. *First*, we inspect the results qualitatively: is the PIHVI system generally better in the opinions of the participants? *Secondly*, we quantitatively measure task performance: how many relevant threads was the user able to find in each task, compared to the amount of time spent? We measure this as a proportion of total relevant threads to time spent, averaged over users, for each task. *Thirdly*, we measure user experience for both systems by a questionnaire asking users to rate their agreement to statements about the system on a 5-point Likert scale from strong disagreement to strong agreement; the questions are based on the standard ResQue questionnaire with some modification for the online forum domain, and are listed in Table 3. *Fourthly*, after the tasks we also ask each user which system they preferred overall.

	Baseline	PIHVI
Mean (Std)	1.045 (1.222)	4.137 (1.359)

Table 1: Means (over tasks and users) and standard deviations of quality measures R calculated from retrieved threads. A larger mean suggests better relevance.

	System	Participant	System×Participant
p-value	6.82×10^{-13}	0.938	0.440

Table 2: Two-way ANOVA for relevance of retrieved threads. We have $p < 0.05$ in the system effect; the 95% confidence interval of the PIHVI system coefficient is $[0.379, 4.819]$, fully above zero. Thus PIHVI performs significantly better than the baseline in terms of relevance of retrieved threads.

Results

Qualitative analysis. We collect qualitative descriptions on the systems by allowing participants to give free-form written or oral feedbacks. Overall, the participants agree the tasks are clear. Besides good quantitative performance in tasks, participants made discoveries, hence PIHVI helped them comprehend the discussion venue. One participant stated he discovered Swedish is sometimes used for trolling in Suomi24 when he was doing the task "Svenskt" on PIHVI. In task "Matti Nykänen", one participant noted PIHVI can put threads on Matti Nykänen together, though people in the threads call him different names. Another participant said discussions on Matti Nykänen in marriage related sections reminded him of the ski jumper's series of short-lived marriages. Participants commented PIHVI can give them threads from diverse sections, compared to the baseline. Participants remarked that the baseline system is reasonable, and one pointed out it is similar to how people usually browse the forum; the only difference is that in the baseline we organize sections radially, not linearly. As for the section filter, one participant stated it helps understand section distributions when the map is zoomed-out. Lastly, though some of them agreed that PIHVI is slightly more complicated than the baseline, the learning curve of PIHVI is not steep, so they can learn to use it fairly quickly.

Task performance. We evaluate task performance of participants in terms of relevance of retrieved threads and time spent. To evaluate relevance of threads retrieved by the participants, a native Finnish speaker assessed all retrieved threads without knowing which system they were from, rating them on a three-point scale from 0 to 2 for non-relevance, partial relevance, and full relevance. For each task on each system performed by each participant, we use the quantity $R = \sum_i r_i / T$ as the quality measure for this (task, system, participant) combination, where T is the number of minutes taken to finish the task, r_i is the rated relevance of the i-th thread, with i going from 1 to the number of retrieved threads in the combination. The measure R can be interpreted as the amount of "retrieved relevance in unit time". A larger R indicates that a larger number of more relevant threads are retrieved in a shorter time. We collect the R values from all results of the participants on the two systems, and perform a two-way analysis of variance (two-way ANOVA; "participant"× "system"). Means and standard deviations of the R values are shown in Table 1. And the ANOVA results are in Table 2. We observe the system effect is statistically significant in this ANOVA, suggesting that PIHVI is significantly better than the baseline system when measuring the relevance of the retrieved threads over time spent.

User responses to post-task questionnaire. Table 3 shows the distribution of answers to the post-task questionnaire. The questions can be divided into two groups, 13 positive questions (higher user agreement is better) and 5 negative questions (lower user agreement is better). Roughly, positive questions cover overall user experience and ability to perform tasks with the system, negative questions focus on ease of use. Half of the 18 questions yielded statistically significant differences. All statistically significant differences are in favor of the PIHVI system. Table 3 lists in detail the p-values showing statistical significance (Q1, Q8, Q9, Q13–Q18) at the level of 0.05. Users wanted to use PIHVI frequently more than the baseline (Q1), found PIHVI less cumbersome than the baseline (Q8), felt more confident using PIHVI than the baseline (Q9), found it more easy with PIHVI to find what they were interested in (Q13), found it more easy to find a thread to read with PIHVI (Q14), would use PIHVI more than the baseline to find threads to read (Q15), would use PIHVI again more than the baseline (Q16), would use PIHVI frequently more than the baseline (Q17), and would prefer PIHVI over other analytics systems more than they would prefer the baseline over other systems (Q18). Thus overall, participants report they can find what they are interested in easily with PIHVI, and would like to use it to do exploratory data analysis frequently (Q1, Q13-Q18). Also, PIHVI is easier to use and gives the participant more confidence during the exploration compared with the baseline (Q8-Q9). The confidence may arise because PIHVI's layout reflects thread similarity better so that in PIHVI the relevance between the found threads and the tasks is clearer than in the baseline. For brevity we simply list the questions that did not yield significant differences: Q2 I found the system unnecessarily complex; Q3 I thought the system is easy to use; Q4 I think that I would need the support of an analytic person to be able to use this system; Q5 I found the various functions in this system are well integrated; Q6 I thought there was too much inconsistency in this system; Q7 I would imagine that most people would learn to use this system very quickly; Q10 I needed to learn a lot of things before I could get going

P/N	ID	Question	Baseline mean (std)	PIHVI mean (std)	p (Welch's t-test)
+	Q1	I think that I would like to use this system frequently	1.917 (0.862)	3.750 (0.433)	9.866×10^{-6}
-	Q8	I found the system very cumbersome to use	3.167 (1.280)	2.000 (0.816)	0.0198
+	Q9	I felt very confident using the system	3.083 (1.187)	4.000 (0.816)	0.0480
+	Q13	Using this system to find what I am interested in is easy	1.750 (0.829)	4.083 (1.115)	1.780×10^{-5}
+	Q14	Finding a thread to read with the help of this system is easy	2.750 (1.090)	4.083 (1.037)	7.604×10^{-3}
+	Q15	I will use this system to find threads to read	2.083 0.954)	3.333 (0.850)	3.757×10^{-3}
+	Q16	I will use this system again	2.083 (0.862)	3.167 (1.143)	0.0206
+	Q17	I will use this system frequently	1.833 (0.799)	2.917 (1.115)	0.0165
+	Q18	I will use this system rather than other systems of exploratory data analysis for online forums	2.167 (0.898)	3.583 (0.640)	3.852×10^{-4}

Table 3: The questions yielding statistical significance in the post-questionnaire, with mean sand standard deviations of user agreement with the questions. For "positive" questions (Q1, Q3, Q5, Q7, Q9, Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18) higher agreement is better, for "negative" questions (Q2, Q4, Q6, Q8, Q10) lower agreement is better.

with this system; Q11 The layout of this system interface is attractive; Q12 The layout of this system is adequate.

User preference. Users were also asked which system they overall prefer: all users preferred the proposed PIHVI system.

CONCLUSIONS

We presented PIHVI – Posting Analysis with Interactive Hierarchical Visualization, a novel visual analytics system for hierarchical online forums based on multiple linked views. The main view (interactive thread content scatterplot) and the coloring that links it to another prominent view (radial tree of sections) are created by nonlinear dimensionality reduction from high-dimensional similarities between threads and between sections, respectively. A user study on a large Finnish online forum shows significant advantages compared to a baseline without dimensionality reduction based views: users found more relevant results with PIHVI, and preferred it over the baseline, with statistical significance. The system helps analysts find and relate discussion content across large hierarchies, an important need as forums keep growing.

REFERENCES

1. I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.
2. S. Chen, L. Lin, and X. Yuan. Social media visual analytics. *Computer Graphics Forum*, 36(3), 2017.
3. K. Dave. Flash forums and forumreader: Navigating a new kind of large-scale online discussion. In *CSCW 2004*.
4. J. Duan, J. Zeng, and S. Zhang. Hierarchical semantic model for objectionable web text content detection. In *IEEE ASID 2012*.
5. S. Fu, J. Zhao, W. Cui, and H. Qu. Visual analysis of mooc forums with iforum. *IEEE TVCG*, 23:201–210, 2017.
6. I.-H. Hsiao and P. Awasthi. Topic facet modeling: Semantic visual analytics for online discussion forums. In *LAK 2015*. ACM.
7. L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.
8. T. Munzner and E. Maguire. *Visualization analysis and design*. CRC Press, 2015.
9. A. I. Obasa, N. Salim, and A. Khan. Hybridization of bag-of-words and forum metadata for web forum question post detection. *Indjst*, 8(32), 2016.
10. J. Peltonen, K. Belorustceva, and T. Ruotsalo. Topic-relevance map: Visualization for improving search result comprehension. In *ACM IUI 2017*, 2017.
11. J. Peltonen, M. Sandholm, and S. Kaski. Information retrieval perspective to interactive data visualization. In *Eurovis 2013 short papers*, 2013.
12. J. Peltonen, J. Strahl, and P. Floreen. Negative relevance feedback for exploratory search with visual interactive intent modeling. In *ACM IUI 2017*, 2017.
13. Z. Qu and Y. Liu. Finding problem solving threads in online forum. In *IJCNLP 2011*.
14. D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. North, and D. A. Keim. Visual interaction with dimensionality reduction: a structured literature analysis. *IEEE TVCG*, 23(1):241–250, 2016.
15. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE VL 1996*.
16. A. J. Soto, R. Kiros, V. Kešelj, and E. Milios. Exploratory visual analysis and interactive pattern extraction from semi-structured data. *ACM TiiS*, 5(3):16:1–16:36, 2015.
17. J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *CHI 2004*.
18. Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In *ICML 2013*.
19. J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE TVCG*, 20(12):1773–1782, 2014.