FacetAtlas: Multifaceted Visualization for Rich Text Corpora

Nan Cao, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu, Member, IEEE

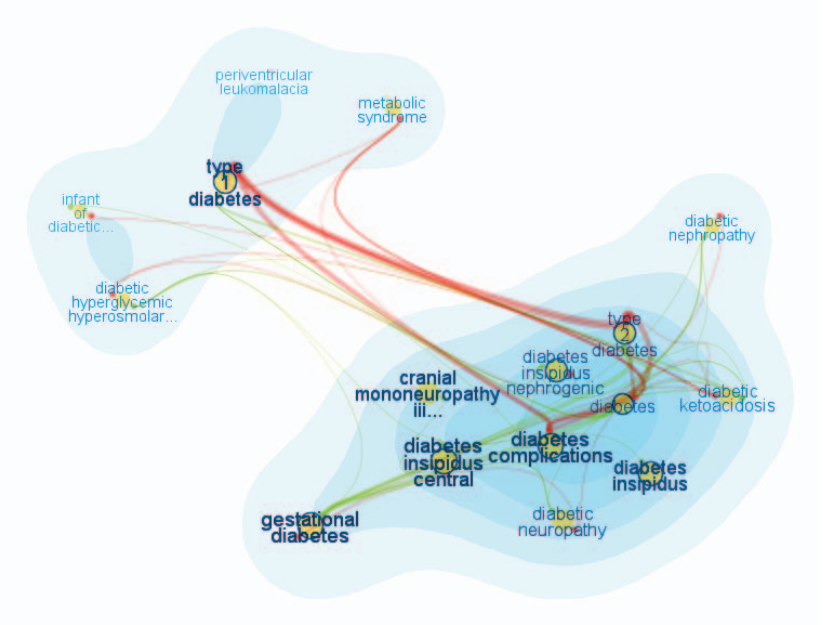


Fig. 1. FacetAtlas generates a multifaceted disease visualization on the query keyword "diabetes". The two disease clusters correspond to type-1 and type-2 diabetes, respectively. Colored links represent connections on different facets. In this figure, type-1 diabetes has similar complications as type-2 diabetes (the red links). Diseases in the type-2 cluster share similar symptoms (the green links within that cluster).

Abstract—Documents in rich text corpora usually contain multiple facets of information. For example, an article about a specific disease often consists of different facets such as symptom, treatment, cause, diagnosis, prognosis, and prevention. Thus, documents may have different relations based on different facets. Powerful search tools have been developed to help users locate lists of individual documents that are most related to specific keywords. However, there is a lack of effective analysis tools that reveal the multifaceted relations of documents within or cross the document clusters. In this paper, we present FacetAtlas, a multifaceted visualization technique for visually analyzing rich text corpora. FacetAtlas combines search technology with advanced visual analytical tools to convey both global and local patterns simultaneously. We describe several unique aspects of FacetAtlas, including (1) node cliques and multifaceted edges, (2) an optimized density map, and (3) automated opacity pattern enhancement for highlighting visual patterns, (4) interactive context switch between facets. In addition, we demonstrate the power of FacetAtlas through a case study that targets patient education in the health care domain.Our evaluation shows the benefits of this work, especially in support of complex multifaceted data analysis.

Index Terms—Multi-facet visualization, Text visualization, Multi-relational Graph, Search UI

1 Introduction

As the Internet continues to experience explosive growth, an ever increasing amount of information is becoming available through collections of rich text documents. Ranging from digital libraries to online medical references, these collections contain a wealth of multifaceted interconnected data. To navigate through this rich data, most people rely on search technologies to find relevant information. Search tools typically return a ranked list of documents whose content is highly related to a set of user-supplied keywords. This model has proven remarkably powerful for information retrieval tasks, such as locating the address of a restaurant. However, ranked lists are insufficient for more complex data exploration and analytical tasks where users try to understand an overall document corpus or relationships between complex concepts that span across multiple documents. Despite recent work on more advanced interfaces [18, 23, 27], the effective organization and presentation of search retrieval results is still largely an open problem.

This problem becomes even more challenging when considering the multifaceted nature of many documents. For instance, consider an online library of health-related articles such as Google Health. Each article in the library describes a specific disease and contains information about a number of different facets: symptom, treatment, cause,

- *Nan Cao and Huamin Qu are with the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. E-mail: {nancao, huamin}@cse.ust.hk.*
- *Jimeng Sun and David Gotz are with IBM T.J. Watson Research Center. E-mail: {jimeng, dgotz}@us.ibm.com.*
- *Yu-Ru Lin is with Arts Media and Engineering at Arizona State University. E-mail: yu-ru.lin@asu.edu.*
- *Shixia Liu is with IBM China Research Lab. E-mail:shixia@gmail.com.*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

diagnosis, prognosis, and prevention. A search engine allows users to find a page describing a specific disease, and links allow users to navigate to a small set of predefined related pages. However, answering some basic self-care questions remains very difficult. For example: *What are the general classes of diseases that can lead to the symptoms I'm experiencing? Which of those diseases have a similar prognosis? How do those diseases relate to each other in terms of treatment alternatives?*

These questions require an understanding of complex correlations across documents and across multiple facets of the contained information. To answer these questions, users need to examine both high-level overviews and fine-grained local-level relationships. For instance, a user in the scenario above would need to both explore clusters of related diseases and uncover pairwise relationships based on specific facets of information such as prognosis and treatment.

Information visualization technologies, when used in conjunction with data mining and text analysis tools, can be of great value for these sorts of tasks. For this reason, several visualizations have been designed for either high-level corpora summarizations (e.g., [13]) or low-level structure analysis (e.g., [27, 30]). See Section 2 for a more comprehensive survey of related techniques.

Although many existing techniques provide valuable insight into the visualized data, none of them offers a complete solution with the following key features: (1) interactive visualization of local data relationships within the context of global document patterns, (2) dynamic context control so that users can pivot between different facets of information, and (3) an integrated approach to multifaceted search and visualization.

To bridge this gap, we propose FacetAtlas, a new interactive visualization technique that enables users to navigate and analyze large multifaceted text corpora with complex cross-document relationships. Specifically, FacetAtlas provides the following key features:

- **Visualization of both global and local patterns.** FacetAtlas employs a multifaceted graph visualization to visualize local relations and a density map to portray a global context.
- **Integrating unstructured search with visualization.** FacetAtlas automatically converts search results from a one-dimensional list into a visual graph-based representation that is rendered within a global context. This enables an interactive exploration of multifaceted relationships.
- **Dynamic facet-based context switching.** In addition to basic interactions such as zooming, filtering and highlighting, FacetAtlas supports dynamic context switching. This allows a user to pivot the primary visualization layout arrangement across different facets while maintaining his/her analytic focus.

In this paper, we describe both the design and implementation of FacetAtlas. In addition, we demonstrate the power of our approach through a case study of FacetAtlas applied to a healthcare application[1]. Initial user feedback confirms the effectiveness and general applicability of FacetAtlas for searching and exploring rich text corpora.

The rest of paper is organized as follows: Section 2 reviews several areas of related work; Section 3 introduces the data model used in our work; Section 4 presents the visualization design of FacetAtlas and describes how users interact with the system; Section 5 includes a detailed description of our implementation; Section 6 presents a case study in the healthcare domain; Section 7 evaluates FacetAtlas through a formal user study; finally, the paper concludes in Section 8 with a review and discussion of future work.

2 Related Work

In this section, we first review text visualizations that have focused on global patterns. Then we discuss work on visualizing local relational patterns in text. Finally, we survey related work on using visualization as a search interface.

[1]The screen capture video is available at: http://www.cse.ust.hk/∼nancao/movie/facetatlas.mov

2.1 Visualizing Global Patterns in Text Corpora

Content level: Many conventional text visualizations are designed to reveal global patterns in content from single or multiple documents. Perhaps most common is the tag cloud [12] which displays a set of words in rows with font sizes that correspond to statistics such as term frequency. More advanced tag clouds such as Wordle [29], and Word Cloud [9] enhance the appearance through more sophisticated layouts. In other work, document cards [24] present the content of a document collection using summary cards that include highlighted figures and content topics. Topic Islands [20] uses wavelets to summarize the thematic characteristics of a single document. ThemeRiver [13] visualizes topic evolution using a temporal plot showing the ebb and flow of topic themes over time. FacetAtlas goes beyond global patterns by displaying both global and local patterns.

Document level: In contrast to content-level visualization methods, document-level tools use projection-based techniques to visualize relationships between documents in a collection. Many of these visualizations [1, 7, 31] map a set of documents to a 2D display according to document similarity. Other projections, such as probabilistic latent semantic model [17], can reveal topic clusters. However, because of information lost when projecting from a high dimensional space to 2D coordinates, it is often hard for users to understand the semantic meaning of the resulting clusters. In FacetAtlas, we follow a projection-based approach to render document-level relationships. We combat the information lost due to dimensionality reduction by providing a novel multifaceted graph-based display that is integrated with an optimized density map.

2.2 Visualizing Local Relational Patterns

Visualizing local relational patterns has received significant attention in recent years, specifically in the context of text and graph visualization.

Text Visualization: Text visualizations such as WordTree [30] and PhraseNet [28] focus on relational word patterns in text. In particular, WordTree considers the prefix relation between words at the syntax level. PhraseNet allows users to define relationships. However, these systems do not focus on multifaceted relations as we do with FacetAtlas. In other work, Collins et al. [8] introduce parallel tag clouds (PTCs) to visualize text along multiple facets arranged as columns of words. Links across columns represent co-occurrence relationships. In contrast to the word-level focus of PTCs, FacetAtlas can visualize more complex latent relationships.

Graph Visualization Various network visualizations [15] have been designed to analyze relational patterns. However, many of these, such as Vizster [14], consider only one type of relationship. In order to visualize multiple types of relationships, Shen et al. [21] introduce OntoVis which uses nodes and links to represent various concepts and relations for large and heterogeneous social networks. In particular, OntoVis connects each entity in the focused concepts with its related entities from both focused and unfocused concepts. Compared to OntoVis, our FacetAtlas adopts a completely different visual design to present multifaceted relations which can easily convey both global and local patterns in one visual metaphor. In addition, it also provides several novel visual interaction to facilitate users to identify outliers and co-occurrences.

In other work, SocialAction [22] supports relational pattern detection for social networks through smart filtering of important nodes, clusters and outliers. Similarly, FacetAtlas provides users with rich interaction tools that allow them to further interpret and examine multifaced interconnected data from multiple perspectives. In addition FacetAtlas includes automated pattern detection to support the visualization of clusters, co-occurrences and outliers.

2.3 Visual Search Interfaces

Traditional search interfaces for text corpora present a ranked list of search results. Recognizing the limitations of this approach, some researchers have explored visualization-based search interfaces. For example, van Ham et al. [27] present a visual search tool to allow users

to navigate through a subgraph in a huge document network. Smith et al. [23] introduce FacetMap and FacetLens [18] which provide a visualization-based interface for multifaceted document search. Commercially, Grokker (http://www.grokker.com/) is notable for its use of a circular Treemap visualization to dynamically generate topic clusters on web search results. However, these systems do not consider multifaceted relationships as is the focus of FacetAtlas.

3 Data Model and Transformation

In this section, we introduce the multifaceted entity-relational data model used by FacetAtlas. We first define the core data model constructs. We then discuss how a set of documents is transformed from raw text to fit this model.

3.1 Multifaceted entity-relational data model

The FacetAtlas data model is a multifaceted representation that captures entities and their relationships. The model consists of the following abstract data :

- **Entities** are instances of a particular concept from the data. For example, "Type-1-Diabetes" is a disease entity.
- **Facets** are classes of entities. For example, "disease" is a facet which contains both the "Type-1-Diabetes" and "Type-2-Diabetes" entities.
- **Relations** are connections between pairs of entities. There are two types of relations. *Internal relations* are connections between entities or entity groups within the same facet. For example, "Type-1- Diabetes" has an internal relation to "Type-2-Diabetes" because both are diseases. *External relations* are connections between entities of different facets. For example, the disease entity "Type-1-Diabetes" has several external relations to symptom entities such as"increased thirst" and "blurred vision."
- **Clusters** are groups of similar entities within a single facet. For example, a group of diseases related to "Type-1-Diabetes" forms a cluster on the disease facet.
- **Life** Each entity and relations are assigned

A simple example of the data model is illustrated in Fig. 2(a). The figure shows three facets—Disease, Symptom and Treatment—each represented as a separate layer. Nodes on each layer represent entities within the corresponding facet. Edges within a layer are internal relations, while edges across layers are external relations.

3.2 Transformation

Before FacetAtlas can be used to visualize a corpus of text documents, the raw text material needs to be transformed to fit into the multifaceted entity relational data model described above. The transformation process consists of the following key steps: multifaceted entity extraction, similarity measurement, and index building.

First, we extract the multifaceted entities by either applying a standard name entity recognition (NER)[2] based on some domain-specific model, such as the medical model, or rely on topic modeling (e.g., ContexTour [19]). The former approach, NER, can easily extract entities in facets like organization, location and time. The latter one analyzes topic threads from documents that are used as facets with their keywords being used as entities in FacetAtlas.

Second, we construct a similarity graph for the extracted entities. In this step, we use either standard information retrieval measures (e.g., cosine similarity) or topic-level similarity through topic modeling. This step may be skipped when topic information is available from the text corpus such as in the Google Health case.

Finally we use Lucene[3] to build a separate search indices for each facet. FacetAtlas leverages these indices for online queries. As a result, user-supplied query keywords can be used at runtime to access targeted portions of the data model. When a query is issued, FacetAtlas retrieves and visualizes the most relevant entities and their corresponding relations.

[2]Stanford name entity recognizer, http://nlp.stanford.edu/software/CRF-NER.shtml

[3]http://lucene.apache.org/java/docs/

4 Visualization Design of FacetAtlas

Based on the multifaceted entity-relational data model we design an interactive visualization for the exploration and analysis of multifaceted interconnected data. In this section, we describe in detail how FacetAtlas visualizes such data.

4.1 Overall Visual Design

To encode both global cluster information as well as detailed pairwise relationships in multifaceted interconnected data, we combine a density map with a multifaceted graph. As show in Fig. 1, the cluster context is displayed as a density map in the background layer. In the multifaceted graph, entities are represented by circles, color-coded by their facets.

In the following sections, we introduce three key aspects of the FacetAtlas design including: (1) the visual encoding adopted to represent elements of the FacetAtlas data model, (2) the visual patterns employed to facilitate data exploration; and (3) the user interactions that allow users to examine data from multiple perspectives.

4.2 Visual Encoding

The facet, entity, and relation are the abstract elements in our data model. In this section, we describe in detail how FacetAtlas employs visual elements (e.g., point, link, area, and color [4, 6]) to encode these abstract elements.

Facet Encoding. Facets are encoded by different colors. Categorical colors are selected for facets based on the CIELAB model so that facets can be easily differentiated. The facet colors are consistently used for both points and links, with the colors remaining constant across views as users navigate the visualization. In addition, the visualization differentiates between a single *primary* facet and other *secondary* facets by sizes and colors. The interactive facet legend displays the primary facet as the leftmost circle and uses entities from this facet to build the base graph of the visualization.

Entity Encoding. An entity is represented as circles colored by its facet. For primary entities (entities that belong to the primary facet), circle size is used to represent an entity's degree-of-interest (DOI) [25]. More specifically, DOI determines to what extent a user will be interested in a certain entity. In our search-oriented application area, users' interests become clear when they issue a query. Therefore, DOI is defined as the relevancy of an entity to a user's query.

Secondary entities (entities that belong to one of the secondary facets) are rendered together with primary entities as *compound nodes*. Each compound node contains a single large circle (representing a primary entity), surrounded by small nodes (representing secondary entities invisibly connected by external relations). We call these nodes entity node and facet nodes, respectively. For example, in Fig. 2(b), the disease "Diabetes-Type-1" has a symptom facet node drawn in red which corresponds to a set of symptom entities such as "increased-thirst" and "blurred-vision" that have external relations to the disease.

This design also collapses multiple secondary entities into a single facet node to reduce visual clutter. As a result, the visualization becomes more consistent since only primary entities are displayed in detail. We argue this simplification is one of the key design elements that enables the successful display of both global and local patterns in a clear fashion.

Relation Encoding Two different visual encodings are used to encode relations within FacetAtlas, one for each of the two relation types: internal relations and external relations.

Internal relations are encoded using links between corresponding facet nodes of two different compound nodes. Once again, color coding is used to illustrate which facet the link represents. For example, Fig. 2(b) shows a red link representing an internal relation between the symptom nodes of two different diseases. The thickness of a link indicates how related two entities are along a specific facet.

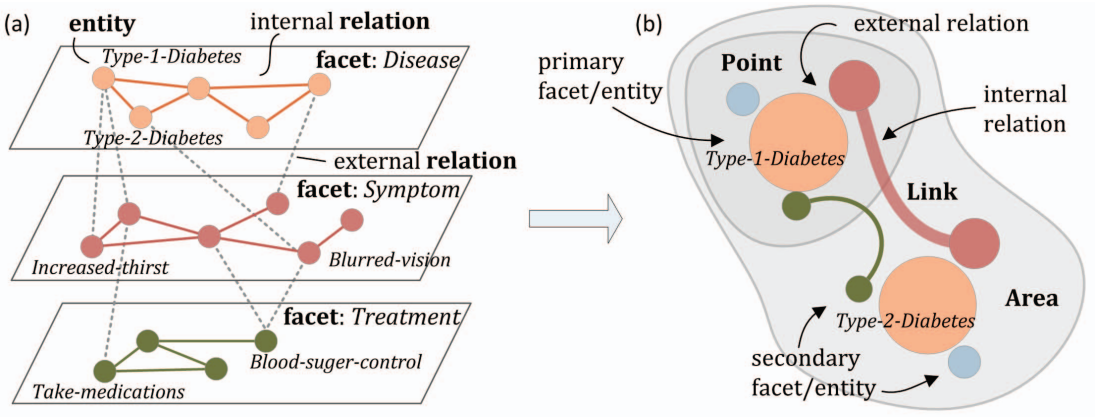


Fig. 2. (a) The FacetAtlas multifaceted entity-relational data model. Concepts in a complex text corpus are transformed into *facets*, *entities* and *relations*. (b) The data model is visually encoded using a spatial arrangement of color-coded nodes and edges.

In Fig. 2(b), "Type-1-Diabetes" and "Type-2-Diabetes" share a lot of common symptoms (shown in red) but fewer common treatments (shown in green). Therefore, the red link is thicker than the green link.

Based on this encoding, the links between two compound nodes A and B indicate the overall similarity of their primary nodes in multiple facets. More specifically, the thickness of a single link in Facet F that connects two facet nodes respectively in A and B is encoded to the similarity of A's primary entity and B's primary entity in F. This similarity is calculated by the overlap ratio of the secondary entities inside two connected facet nodes. For example, A's primary entity is "diabetes" and B's primary entity is "cancer". If there is an internal link on the symptoms facet that connects A and B, it indicates that "diabetes" and "cancer" have some similar symptoms. The number of the symptoms in common over all their symptoms is the overlap ratio that indicates the similarity of these two diseases on the symptom facet. It is encoded by the thickness of the link.

External relations are encoded implicitly through the construction of compound nodes. When a primary entity is displayed through a compound node, only facet nodes with external relations are included. Moreover, the size of a facet node is proportional to the number of external relations on that facet. For example, Fig. 2(b) shows three facet nodes for "Diabetes Type-1." Among the three, the red symptom node is the largest, signifying the most external relations between "Diabetes type-1" and different symptoms.

4.3 Visual Patterns

FacetAtlas provides several visual patterns to facilitate user exploration.

Clusters Groups of similar entities are represented using an optimized density map. Our design is similar to [3, 10], but using a completely different algorithm. Intuitively, a clustering process divides entities into groups by their internal relations. For clarity, we defer our detailed description of the algorithm behind this capability until Section 5.2.1. Visually, we use areas with boundaries to encode the calculated clusters. When visualizing large text corpora with too many entities to display at once, only the entities with highest DOI in each cluster will be shown on the screen. To provide navigation queues regarding hidden entities, a cluster density metric is computed and mapped to the color intensity of the bounded areas. This conveys the overall distribution of entities within the clusters.

Co-occurrences Co-occurrence patterns occur when two or more entities have very strong internal relations across several facets. Such a set of nodes often implies a tight cluster on secondary facets. For example, if a set of diseases share the same symptoms, treatments and prognoses, the relations across these facets will form a co-occurrence pattern. This pattern signifies that the set of diseases are deeply related. Visually, we represent co-occurrence patterns across multiple facets using parallel links between the associated entities as shown in Fig. 6(a).

Outliers Outlier patterns represent entities with internal relations that cross cluster boundaries. Visually, we represent outlier patterns by highlighting links cross the cluster boundaries via opacity adjustment as shown in Fig. 6(b). The algorithmic details of this process are provided in Section 5.4.

4.4 Interaction

To support interactive exploration of the visualized data, FacetAtlas provides several rich interactions.

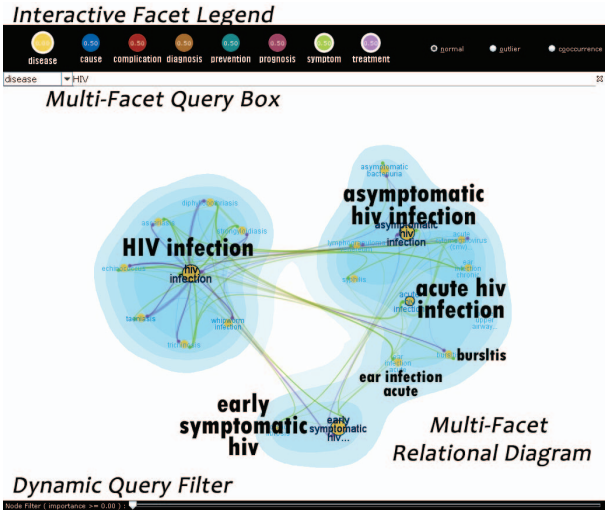


Fig. 3. FacetAtlas applied to Google Health data. This figure shows a disease diagram for the search term "HIV." FacetAtlas contains four main components: (1) an interactive facet legend, (2) a query box, (3) a canvas for rendering multifaceted relational diagrams, and (4) a dynamic query filter to control that amount of information being displayed.

Dynamic Query FacetAtlas allows users to provide a text query through a multifaceted query box. This feature is illustrated in Fig. 3 where a user has searched for the term "HIV." In response to a search, the main visualization space shows a depiction of the facets, entities, relations and clusters relevant to the query. In addition, users can use the dynamic query filter slider (located at the bottom of Fig. 3) to filter entities based on their DOIs.

During the dynamic query process, animated transitions as well as a context-preserving layout algorithm [5] are applied to maintain a

user's mental map. This approach balances layout stability against overall readability to provide an optimized dynamic visualization experience with minimum changes and maximum aesthetics.

Semantic Zoom This feature allows users to select a particular compound node to zoom in for more details about related entity nodes. For example, when a user zooms in from the disease "Asymptomatic HIV Infection" shown in Fig. 3, they arrive at the new view shown in Fig. 7(a). Similar to dynamic queries, we also use animated transitions and context-preserving layouts to maintain a user's mental map during semantic zoom.

Context Switching Users perform a context switch to change the "primary facet" around a focal point. As an extension of semantic zoom, context switching allows users to first focus on a specific compound node and then switch view point to other facet. For example, as depicted in Fig. 7, when users switch context from *Disease* to *Symptom* view on "HIV Infection", all its related symptoms will be shown as entity nodes in the resulting visualization. Diseases, meanwhile, collapse into facet nodes surrounding the symptom nodes.

Highlight Two types of highlighting interactions are supported by FacetAtlas: link highlights and pattern highlights. Link highlights provide contextual information for an entity. More specifically, when a mouse-over occurs over a compound node x, the tooltip with a summary of x is shown. In addition, all other entity nodes that are directly connected to x are also highlighted.

Pattern highlights are designed to help filter out trivial connections and enhance more meaningful patterns such as outliers and co-occurrences. Users can select radio button on the top right of the user interface to control the pattern highlighting feature.

Other Interactions In addition to the interactions described above, several standard interactions are also provided:

Power Buttons. These buttons allow users to turn on/off certain facets. In Fig. 3, the legend bar on top of the main view shows color-coded facet buttons. The leftmost and largest button shows the primary facet, followed by smaller secondary facet buttons. The secondary facet buttons can be clicked to turn on/off the corresponding facet and its relations.

Links To Documents. At any given time, users can connect back to the original documents by double clicking on the nodes and links. A popup window will be shown to illustrate the summarization of the contents. FacetAtlas generates summarization in the following ways.

First, for the entities that have corresponding documents such as the entities in disease facet, we return the list of such documents. If only a single such document exists, we will directly connect to this document.

Second, for the entities that occur in multiple documents such as the symptoms entities, we use MEAD [4] to summarize all the related documents into a temporary text file which contains both the content summarization of all the original documents and the hyperlinks pointing to them. This is similar to the search results generated by traditional search engines such as Google.

5 Implementation

This section describes the implementation of FacetAtlas based on the design outlined in the previous section. It first presents an overview of the FacetAtlas system architecture. It then provides detailed descriptions for several key algorithms employed in support of the overall design, including *cluster layout*, *relation layout*, and *pattern enhancement*.

5.1 Architecture

The FacetAtlas architecture, shown in Fig. 4, consists of three primary components. First, the *Data Transformation* module transforms a collection of text documents into the entity-relational data model through text mining and entity extraction. The transformation process

[4] A public domain portable multi-document summarization system. http://www.summarization.com/mead/

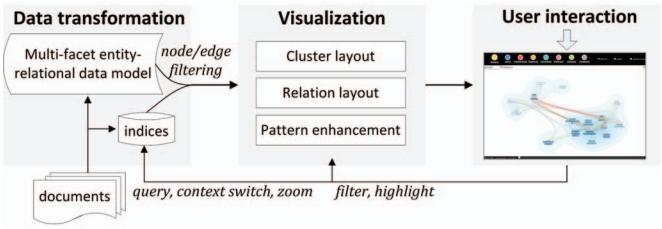


Fig. 4. The FacetAtlas architecture consists of three primary components: (1) a data transformation module, (2) a visualization rendering module, and (3) a user interaction module.

also constructs a set of indices over the data model for online querying. A description of the data model and transformation process can be found in Section 3.

The *Visualization* module maps the indexed entities and relations to a multifaceted visual display according to the visual design outlined in Section 4. It employs custom algorithms for laying out clusters of nodes and relationships between those nodes. It also includes pattern enhancement capabilities that improve the overall appearance and legibility of the visualization.

The *User Interaction* module enables rich interactions for users to explore the data through operations such as filter, query and context switch. These operations feed back into the data transformation and visualization modules to enable user-driven data exploration.

5.2 Cluster Layout

Cluster layout is performed within the FacetAtlas visualization module. Given a set of entities retrieved from an online query, we render a density-map in the background to depict the overall distribution of matching entities. Unfortunately, it is impossible for us to know the real distribution of the data. Therefore we render density map by first performing a *Kernel Density Estimation*. This process finds the optimized smooth kernel density estimator to simulate entity distributions over the entire dataset. After that, these samples are laid out within the display space. Then, we put the optimized estimator over all samples and compute joint distributions for intermediate locations within the space. Finally, we trace the contours within the estimated values to generate a density map with contour lines at multiple levels.

5.2.1 Kernel density estimation

Kernel density estimation (KDE) is a non-parametric way of estimating the probability density function of a random variable[5]. To approximate and visually illustrate the global cluster context in multifaceted interconnected data, we use all the underlying entities to model the density distribution. To learn an optimal KDE, we extend the traditional KDE technique [26] to determine the density distribution around entities. More specifically, to well represent all the interconnected nodes by the sampled nodes, we assign each of the hidden nodes to one of the closest sample nodes by performing a reverse-kNN search based on graph topology. Mathematically, given a kernel function $K(\cdot)$ and a positive number h as its bandwidth, the n-sample kernel density estimator on the k-th facet is defined as:

$$f_n^k(v) = \frac{1}{n} \sum_{i=1, i \neq v}^{n} \frac{m_i}{h} K\left(\frac{d_{iv}}{h}\right) \quad (1)$$

where $v \in V$ is an entity in the raw data; d_{iv} is the length of the shortest path between the entity v and the sample entity i on facet k; and m_i is the kernel mass of the i-th sample. The bandwidth h controls the amount of smoothing. When h is small, $f_n(x)$ gives a set of spikes. When h is large, $f_n(x)$ becomes a uniform distribution. To find the best estimator that is closest to the real distribution, the following loss function between the unknown real distribution f(x) and its n-sample estimator $f_n(x)$ is minimized by choosing an optimal bandwidth h^*:

[5] http://en.wikipedia.org/wiki/Kernel_density_estimation

$$L(h) = \int (f(v) - f_n(v))^2 dv = \int f_n(v)^2 dv - 2\int f_n(v) f(v) dv + \int f(v)^2 dv \quad (2)$$

Considering that $\int f(x)^2 dx$ is not dependent on h, we can reformulate Eq. 2 as

$$J(h) = \int f_n(v)^2 dv - 2\int f_n(v) f(v) dv \quad (3)$$

Empirically, this can be trained using leave-one-out cross-validation [26] over the training samples. Thus we effectively put a smooth mass over all data points through the best n-sample estimator.

5.2.2 Density map layout and estimation

We generate the density map based on the sample node locations within the display space, and use the optimized density estimator to simulate entity distribution of the entire data corpus.

A base layout is first computed by mapping samples to the display space. To stabilize the layout of visible nodes during animated transitions, we use a context preserving stress majorization algorithm [5]. In this algorithm, we balance between the readability and stability of the density diagram by taking the dynamics of data exploration into account. Furthermore, we preserve layout relationships between individually connected components by adding virtual connections among the most related nodes in different components respectively. The relatedness is computed using internal relations on the secondary facets.

After the base layout is obtained, the optimized density estimator is applied over all the samples. Joint distributions are computed in real-time according to the sample locations. Contour lines of the density map are then generated by tracing the gradient of the joint densities across the display space. To accelerate the density map generation, we grid the screen into a low resolution density matrix. Finally we use the estimated density values to determine the color transparency of areas in the density map to complete this portion of the visualization.

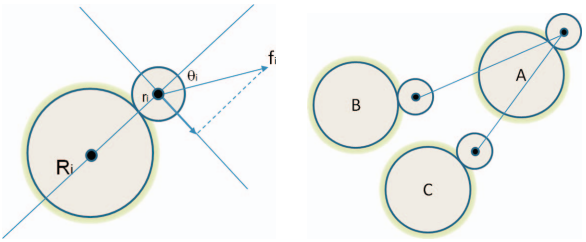
5.3 Relation Layout

We implement two types of link layouts to represent both internal and external relations between facets simultaneously. As described in Section 4.2, we represent external relations with facet nodes by placing them around a central entity node. This arrangement forms a single compound node. Internal links are represented as edges that connect two facet nodes from different compound nodes. To reduce line crossings and facilitate relational pattern search, a custom layout algorithm is used to arrange the visual presentation of these elements.

Our algorithm first reduces line crossings by adjusting the position order of facet nodes. It repeatedly swaps adjacent pairs within each compound node as long as thees swaps result in a lower number of crossings. This process is repeated until it reaches a pass with no swaps. Using the refined order, we then apply a global spring force model across all compound nodes. The objective is twofold: (1) to minimize the average edge length, and (2) to refine compound node orientations to facilitate edge bundling. As demonstrated in various systems [11, 16, 32], edge bundling can reduce visual clutter and improve the clarity of displays. The spring force is applied to each internal link. The model is defined as (see Fig. 5(a)):

$$min \sum_k \sum_i \omega_k f_i^k sin(\theta_i^k)(r_i^k + R_i) \quad (4)$$

where ω_k is the importance of the k-th facet. i is the index of the entity node. R_i and r_i^k are radii of the i-th entity node and its k-th facet node, respectively. θ_i^k is the orientation of the edge with an endpoint of the k-th facet node of the i-th entity node. The new objective balances the force moment on each compound node. Thus, it avoids unnecessary link-node overlapping as depicted in Fig. 5(b). This is achieved by rotating the compound node to adjust link orientations.



(a) (b)

Fig. 5. The force model on compound nodes. (a) The force momentum resolution of one single spring force in the model, (b) link-node overlapping. The internal links overlap with entity node A in a bad layout of the facet nodes.

Using the layout produced by the algorithm above, we further perform a graph partitioning process to divide the clusters before applying hierarchical edge bundling [16] to bundle links through the center of the each cluster containing their end points.

5.4 Pattern Enhancement

As outlined in Section 4.3, FacetAtlas is able to automatically detect two detailed link patterns: *outlier* patterns and *co-occurrence* patterns. In a complex text corpus, identifying these patterns is challenging. Filtering alone does not help because such patterns can only be found when all connections are shown. Therefore, FacetAtlas applies an automated algorithm to adjust link color opacities to enhance these patterns. The result is illustrated in Fig. 6(a) and (b).

The adjustment of color opacities is based on two similarity measurements: *semantic similarity* and *layout closeness*. Semantic similarity sim_{ij} between any pair of entity nodes i and j is calculated by considering all internal connections of these two nodes:

$$sim_{ij} = \sum_{k=1}^{M} sim_k(i,j) \quad (5)$$

where $sim_k(i,j)$ computes the similarity between entity nodes i and j on facet k; M is the number of facets. In our implementation, the similarity is calculated by summing the weights of the corresponding internal connections.

Layout closeness d_{ij} between two entity nodes i and j measures how close the two nodes are in the layout. In our implementation, a hierarchical clustering metric is applied. In this metric, we first cluster the entity nodes in a hierarchy by considering their own similarities or based on an expert ontology. Given the n-level cluster hierarchy, we assign each primary entity i a cluster vector $c_i[1...n]$ where $c_i[k]$ is a cluster ID number in the k-th level in the hierarchy. Then, d_{ij} is calculated by:

$$d_{ij} = 1 - \frac{< c_i, c_j >}{\|c_i\| \|c_j\|} \quad (6)$$

where $< c_i, c_j >$ is the inner product between vector c_i and c_j; and $\|c\|$ is the L2 norm of the vector c.

We enhance the co-occurrence pattern by using sim_{ij} to encode the color opacities of the internal relation links and their related entities. Thus the entities that have connections on multiple facets are automatically highlighted as in Fig. 6(a).

Enhancing the outlier pattern as shown Fig. 6(b) requires a combination of both semantic similarity and layout closeness metrics. More specifically, we use Eq. 7 to adjust the color opacity for all internal relations between i and j.

$$Opacity(i,j) = \sqrt{d_{ij} * sim_{ij}} \quad (7)$$

The rationale behind this formulation is that we will highlight the links that connect nodes topologically far away (d_{ij} is large) that are semantically similar (sim_{ij} is large). In our implementation, both sim_{ij} and d_{ij} are normalized to the range [0, 1].

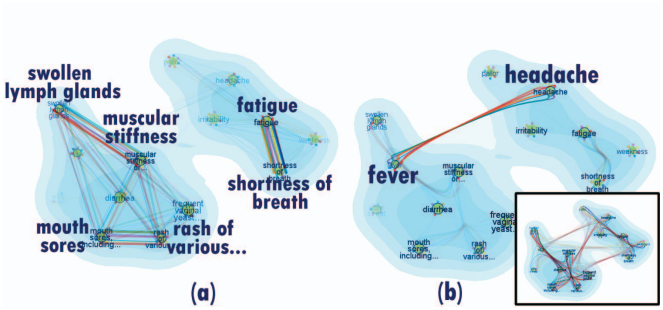


Fig. 6. Automatic Opacity Pattern Enhancement. This visualization depicts the symptoms related with "Asymptomatic HIV Infection" shown in Fig. 3. The right bottom view shows the visualization without pattern enhancement. (a) Enhancement of co-occurrence patterns. This view indicates strongly correlated symptoms using parallel links connecting the correlated facets. (b) Enhancement of outlier patterns. In this case "Fever" and "Headache" are two symptoms in different clusters with a strong connection through secondary facets.

The opacity calculated above is rendered by alpha-blending, i.e., a link or node with high opacity is rendered with low transparency. In this way it provides a soft-cut with the display context and visually illustrates the patterns in a smooth way.

6 Case Study

To demonstrate the capabilities and usefulness of FacetAtlas, we apply it to a health care application. In this section, we first introduce the application setup. Then we describe in detail two specific use cases that revealed interesting patterns. Note that both cases were suggested by the medical experts who used our system.

6.1 Setup

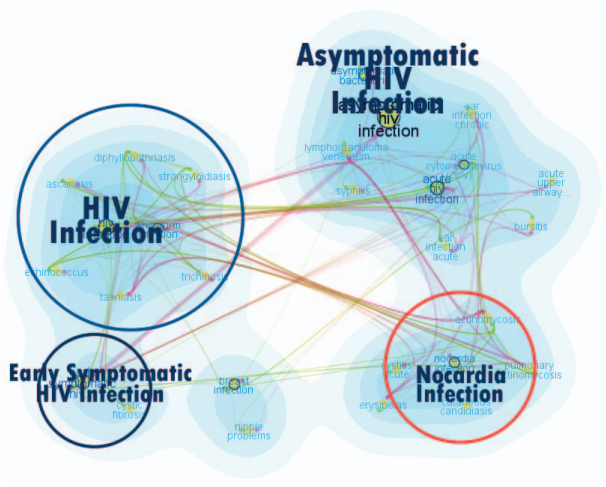
Our case study application is based on the online Google Health library which contains over 1,500 online articles. Each article describes a single disease in multiple sections such as *disease overview, treatment, symptoms, cause, diagnosis, prognosis, prevention and complications*. To prepare the data for FacetAtlas, we transformed these online articles to fit into our multifaceted entity-relational data model. Each document section is mapped to a facet in our model. For each section, there are often several bulleted lists. Each of those bullets becomes an entity in the corresponding facet. For those sections that do not contain any bullets, name entity recognition with a medical text model is applied to extract medical entities. Furthermore, we leveraged the standard ICD-9 classification[6] to group the disease entities into clusters. With this transformation, we converted the articles into a multifaceted entity-relational data model. The transformed dataset contains 8 facets and around 25,000 entities with more than 50,000 internal links.

6.2 Study on HIV Infection

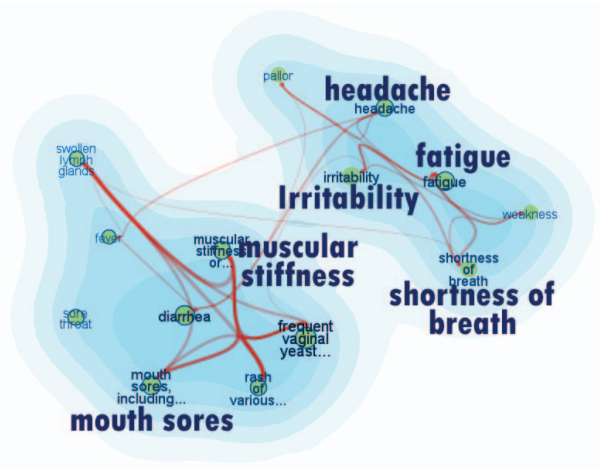
To help users study "HIV infections", we choose "HIV" as the query for further exploration. As shown in Fig. 3, FacetAtlas generates the disease diagram initially as a density map without any links. Three cluster patterns were clearly shown. Each of the three clusters represents a different stage of "HIV infection". We turned on the symptom and treatment facets by clicking the corresponding buttons at the top. Interestingly, we found that all three clusters share similar symptoms (as illustrated by the green symptom links that cross cluster boundaries) while each cluster has relatively distinct treatments (purple treatment links are within clusters).

To learn more about the disease, we double-clicked on each of the three center diseases: "HIV Infection", "Asymptomatic HIV Infection" and "Early Symptomatic HIV Infection". Their Google Health articles were shown. Then we learned that the "Asymptomatic HIV infection" is a very dangerous stage since there are no obvious HIV

[6]http://icd9cm.chrisendres.com/



(a)



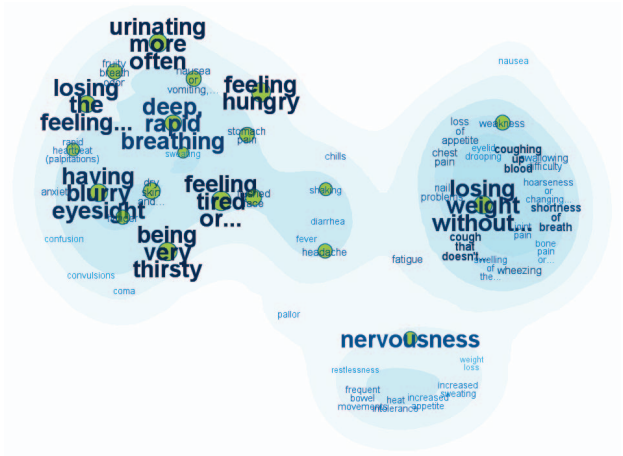
(b)

Fig. 7. Case study on HIV infection. (a) Semantic zoom. When zooming in on "Asymptomatic HIV Infection" from the initial view shown in Fig. 3, more related diseases are shown (highlighted in the red circle). The initial context is preserved and represented by the black circles. (b) Context switch. After switching from a disease view to a symptom view for "Asymptomatic HIV Infection", two prominent symptom clusters are shown. These symptoms share similar complications within each cluster as indicated by the red links.

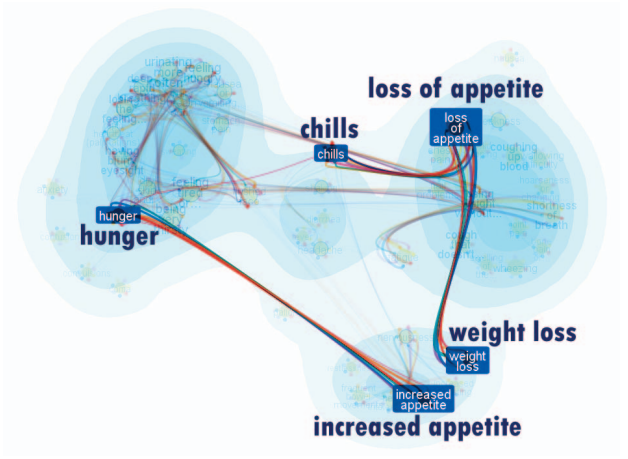
symptoms. When we performed a semantic zoom in to this disease, we confirmed that it has strong symptom connections with some other infections (see Fig. 7(a)). Furthermore, we switched to the symptom view to explore all its related symptoms in detail. Two symptom clusters were visible as shown in Fig. 7(b). After exploring the multifaceted relationships of those two clusters, we found that these symptom clusters led to different complications as shown in Fig. 7(b). More interestingly, when switching to the co-occurrence view in Fig. 6(a), we found some symptoms always co-exist with each other. This is evident through the parallel lines that are used to highlight co-occurrence patterns. Similarly, when switching to the outlier view, some cross-cluster symptoms like "fever" and "headache" were highlighted. This means that they commonly occurred in many other diseases but not in "Asymptomatic HIV infection".

6.3 Study on Diabetes

As indicated by our medical collaborators, diabetes is one of the leading chronic diseases in the United States. Following a similar analysis process as in the HIV case study, we also explored "Diabetes"



(a)



(b)

Fig. 8. Case study on diabetes. (a) Cluster pattern of symptoms related with Type-1-Diabetes. (b) Outlier patterns on symptoms.

using FacetAtlas. As shown in Fig. 1, two clusters are visible, corresponding to "Type-1-Diabetes" and "Type-2-Diabetes". Through the multifaceted links, we easily learned that both type-1 and type-2 have similar complications (red links) and symptoms (green links).

After switching to the symptom view of "Type-1-Diabetes" (illustrated in Fig. 8(a)) all of the disease's related symptoms were automatically grouped into three clusters according to their disease co-occurrence. The cluster on the left had the most common symptoms with "Type-1-Diabetes". The cluster on the right had some less common but still relevant symptoms like "losing weight without trying" and "nervousness". We also found strong outlier patterns in this symptom view as illustrated in Fig. 8(b). Some of these patterns were expected and others were more interesting. For example, we found that for diabetes "hunger" is related to "increased appetite"; "loss of appetite" related to "weight loss" and "chills". These visual patterns can help explain how different symptoms are connected with each other and why.

7 Evaluation

In addition to the case studies above, our system was evaluated quantitatively through a formal user study and expert interviews.

7.1 Study Setup

To evaluate the effectiveness and efficiency of FacetAtlas in support of multifaceted data analysis, we designed and conducted a comparison study. Our study compared the usefulness and usability of FacetAtlas for a health care application with a baseline system that we developed. The baseline used an improved graph visualization for representing multiple relations. We used this approach as our baseline for two major reasons. First, there are no other visualizations beyond FacetAtlas that are especially designed to visualize the multifaceted entity-relational data. Second, the baseline system is implemented based on the FacetAtlas framework which provides exactly the same information and similar interactions as FacetAtlas to enable a fair comparison. The baseline system used a force-based method to cluster related nodes together. However, it doesn't have other visual cues, such as contour map used in FacetAtlas for clusters.

To quantitatively and qualitatively evaluate the main design objectives of FacetAtlas (i.e., soft clustering and multifaceted relations), we designed the following tasks on the healthcare data. T1: identify the major clusters for a test query; T2: identify the representative members in the clusters for the query; T3: identify the facet with the most within-cluster connections; T4: identify the facet with the most cross-cluster connections; T5: identify the facet of the two most connected members on the symptom facet; T6: identify the facet with the most overall connections across entities. We use the test query "diabetes" based on our experts' recommendations in order to simulate a common self-care education scenario for chronic diabetic patients. The overall scenario demonstrates a concrete use-case of a patient using FacetAtlas to understand the underlying relations of different diabetic diseases and complications. From the evaluation perspective, T1-2 evaluate the features on soft-clustering visualization (global patterns), and T3-6 evaluate the features on multifaceted connections (local patterns).

We recruited 20 participants for our study (8 researchers and 12 students majoring in computer science, psychology and mathematics). None had any prior medical or InfoVis background. Inspired by the repeated-measures study [2], we divided the participants into two groups of ten. The first group used our baseline tool for tasks T1, T3, and T5 while the second group used the FacetAtlas for these same tasks. The tool assignment was reversed for tasks T2, T4, and T6 with group one using the FacetAtlas. At the beginning of each user session, we gave a brief tutorial of both systems. The participants were then asked to complete the six tasks.

We recorded two objective measures: task completion time (the time spent on each task), and task success rate (percentage of the successful tasks). We computed the mean and standard deviation of task completion time and task success rate across all the users and questions. We also recorded subjective measures via user surveys: usefulness (how useful a system is for the specific task), usability (how easy the system can be used to complete a specific task), and satisfaction (the number of complaints and favorites on the system features).

7.2 Results and Analysis

The main findings are (1) FacetAtlas and the baseline system lead to similar task success rates across all tasks while (2) FacetAtlas requires significantly less time than the baseline to complete the tasks. The mean (M) and standard deviation (SD) of correctness and time cost are shown in Fig. 9.

When compared on task success rate, both FacetAtlas and the baseline achieve similarly high accuracy levels. This is because both visualization methods display the same underlying information.

The benefit of FacetAtlas becomes evident in the time required to complete the tasks. The standard independent t-test shows that when compared with the baseline system, FacetAtlas has a significant efficiency improvement when visualizing the soft-clusters ($t(18) = 2.26$, $p = 0.0064 < .05$, two tails), showing an overview of multiple connections across clusters ($t(18) = 2.26$, $p = 0.001 < .05$, two tails), as well as representing the details of multifaceted connections between entities ($t(18) = 2.26$, $p = 0.003 < .05$, two tails). All of these aspects are key design objectives of FacetAtlas.

The study results also show that FacetAtlas leads to a slight improvement in finding the most connective facet within a cluster, although the difference is not statistically significant ($t(18) = 2.26$, $p = 0.34 > .05$, two tails). We believe this is due to the use of a

Task Completion Time

	TASK1		TASK2		TASK3	
	M	SD	M	SD	M	SD
FacetAtlas	4.5	0.92	15.6	1.97	65.5	6.84
Baseline	10.7	1.79	37.6	5.68	70	10.9
	TASK4		TASK5		TASK6	
	M	SD	M	SD	M	SD
FacetAtlas	31	5.26	36.6	4	83.9	6.54
Baseline	127.1	21.39	45.1	6.71	149.6	15.63

Task Success Rate

	TASK1		TASK2		TASK3	
	M	SD	M	SD	M	SD
FacetAtlas	1	0	1	0	0.89	0.11
Baseline	1	0	1	0	0.77	0.14
	TASK4		TASK5		TASK6	
	M	SD	M	SD	M	SD
FacetAtlas	0.78	0.14	0.89	0.32	0.76	0.04
Baseline	0.67	0.16	1	0	0.78	0.05

Fig. 9. The user study results of time cost and correctness.

more complex edge visualization when displaying multiple facets in the FacetAtlas approach. Nevertheless, the results are still better than the results obtained from the baseline, which suffers from edge crossings that reduce comprehensibility. FacetAtlas also had no clear advantage over the baseline on task 5 ($t(18) = 2.26$, $p = 0.31 > .05$, two tails). Though statistically insignificant, the baseline performed better in terms of accuracy while the FacetAtlas resulted in faster performance times. We believe that the accuracy advantage was likely due to the use of straight lines for edges which were easier to read when the graph was sparse. In summary, this user study quantitatively confirmed the effectiveness and efficiency of FacetAltas over the baseline in solving several key multifaceted exploration tasks.

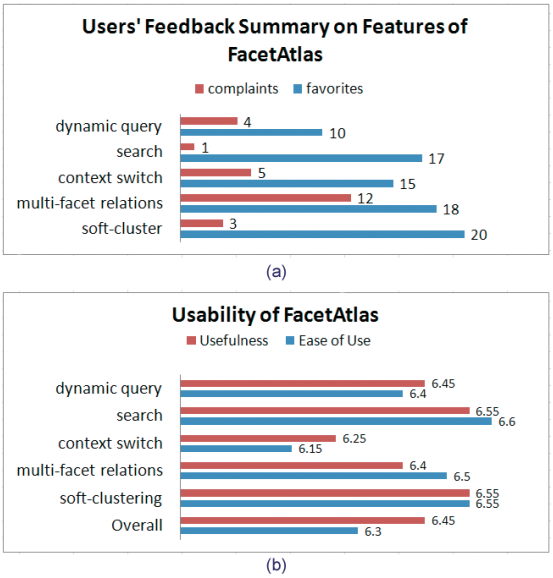


Fig. 10. Users' feedback on FacetAtlas. (a) Summarization of users feedback on FacetAtlas key features. (b) The evaluation of ease of use and usefulness. The scores range from 1 (lowest) to 7 (highest).

In addition to the quantitative results presented above, we collected qualitative user feedback on the key features of FacetAtlas (dynamic query, search, context switch, multifaceted relations, soft-clustering and the overall visualization) through a survey. In addition to asking users for their most and least favorite features, we also asked users to provide numeric rankings for each feature where 1 was the worst score and 7 the best score. Users could also provide free-form feedback. The subjective results are summarized in Fig. 10.

As depicted in Fig. 10(a), the density map based soft-cluster representation is the users' favorite feature whereas the dynamic query is the least favorite. More interestingly, the multifaceted relational visualization (including related interactions) is a strongly appealing feature. Many users rated it as one of their most favorite features.

Fig. 10(b) provides the ratings statistics for each feature (the scores range from 1 to 7). All features have fairly high scores on both usefulness and ease of use, though search and soft-cluster visualization were rated the highest in both aspects.

7.3 Expert Interview

Based on the Google Health data, we also performed 30-minute one-on-one interviews to three medical doctors who have very strong domain expertise. The first doctor is an emergency physician with over 30 years of hospital-based experience. He has published multiple articles and book chapters on both clinical and management subjects. The second doctor is a well-respected health care and biotechnology executive who has more than 30 years of expertise in sophisticated managed care organizations, strategic planning, and operations management. The third doctor is a young medical professional in a hospital.

All of them were very impressed by the interactive visualization that FacetAtlas provides. The first physician was amazed by FacetAtlas. He considered FacetAtlas *"...extremely creative, and has great potential for clinical therapeutic usage and diagnosis decision support."* We asked him to elaborate on how he believes FacetAtlas can help with diagnosis support and why he think so. He believes that the outlier visual patterns can *" ... enhance the current thought process of physicians, and help create the subtle associations between different concepts."* After we explained the patient education scenario, the first physician confirmed by saying *"this will be very helpful for nurses who run the self-care education activities to better engage patients."* Furthermore, the second physician believes that *"this tool has great potential as an education tool for interns and residents who have just started their medical career"*. All three physicians believed that FacetAtlas can also be useful as an alternative interface for many medical resources such as PubMed, UpToDate and even classic medical textbooks. They believed the visual search exploration capabilities of FacetAtlas can help medical researchers and MD students explore the medical literature more effectively. Two of them expressed a strong interest in being the first users of a professional version of FacetAtlas based on authoritative sources of medical literature.

8 Conclusion

In this paper, we presented FacetAtlas, a multifaceted visualization for entity-relational text documents. FacetAtlas is able to visualize both global and local relations of complex text document collections. In particular, global relations are displayed through the use of a density map; and local relations are conveyed through compound nodes and edge bundling techniques. FacetAtlas also provides rich interactions such as filtering, visual pattern highlighting, and context switching. These interactions enable users to examine a text corpus from multiple perspectives. We performed an in-depth case study on a patient education application in the health care domain. The feedback from both regular users and medical professionals was extremely positive and confirmed our main design objectives. In future work, we plan to apply FacetAtlas to more applications, to incorporate the time dimension in the visualization, and to conduct more thorough user studies.

Acknowledgments

This work was supported in part by grant HK RGC GRF 619309 and an IBM Faculty Award. The authors would like to thank all the user study participants and doctors for their contributions to the system evaluation, and the anonymous reviewers for their valuable comments.

REFERENCES

[1] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3/4):166–181, 2002.

[2] R. Bakeman and B. Robinson. Understanding statistics in the behavioral sciences. pages 246–247, 2005.

[3] M. Balzer and O. Deussen. Level-of-detail visualization of clustered graph layouts. In *the 6th International Asia-Pacific Symposium on Visualization.*, pages 133–140, 2007.

[4] J. Bertin. *Semiology of graphics*. University of Wisconsin Press, 1983.

[5] N. Cao, S. Liu, L. Tan, and X. Zhou. Interactive Poster : Context-Preserving Dynamic Graph Visualization. In *IEEE Symposium on Information Visualization*, 2008.

[6] S. Card, J. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[7] Y. Chen, L. Wang, M. Dong, and J. Hua. Exemplar-based Visualization of Large Document Corpus. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1161–1168, 2009.

[8] M. W. Christopher Collins, Fernanda B. Viegas. Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 91 – 98, 2009.

[9] J. Clark. http://neoformix.com/. Neoformix Blog, March 2009.

[10] P. Cortese, G. Di Battista, A. Moneta, M. Patrignani, and M. Pizzonia. Topographic Visualization of Prefix Propagation in the Internet. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):725–732, 2006.

[11] W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li. Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1277–1284, 2008.

[12] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*, pages 25–28, 2006.

[13] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Vizualization*, pages 115–123, 2000.

[14] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization*, pages 32–39, 2005.

[15] I. Herman, G. Melancon, and M. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[16] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[17] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 363–371. ACM, 2008.

[18] B. Lee, G. Smith, G. G. Robertson, M. Czerwinski, and D. S. Tan. Facetlens: exposing trends and relationships to support sensemaking within faceted datasets. In *CHI 09: the 27th international conference on Human factors in computing systems*, 2009.

[19] Y.-R. Lin, J. Sun, N. Cao, and S. Liu. Contextour: Contextual contour visual analysis on dynamic multi- relational clustering. In *SIAM Data Mining Conference(accepted)*, 2010.

[20] N. Miller, P. Wong, M. Brewster, and H. Foote. TOPIC ISLANDS - a wavelet-based text visualization system. In *IEEE Visualization*, pages 189–196, 1998.

[21] Z. Shen, K. Ma, and T. Eliassi-Rad. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1427–1439, 2006.

[22] O. Side, H. Store, V. Us, H. Page, P. Alumni, H. Studying, V. Scholars, C. Websites, and A. Sponsorship. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700, 2006.

[23] G. Smith, M. Czerwinski, B. Meyers, D. Robbins, G. Robertson, and D. Tan. FacetMap: A Scalable Search and Browse Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):797–804, 2006.

[24] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. Keim, and O. Deussen. Document Cards: A Top Trumps Visualization for Documents. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1145–1152, 2009.

[25] J. Thomas and M. Schneider. *Human factors in computer systems*. Ablex Pub, 1984.

[26] B. Turlach. Bandwidth selection in kernel density estimation: A review. *CORE and Institut de Statistique*, pages 23–493, 1993.

[27] F. van Ham and A. Perer. Search, Show Context, Expand on Demand : Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):953–960, 2009.

[28] F. van Ham, M. Wattenberg, and F. Viégas. Mapping text with phrase nets. *IEEE transactions on visualization and computer graphics*, 15(6):1169–1176, 2009.

[29] F. Viégas, M. Wattenberg, and J. Feinberg. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, 2009.

[30] M. Wattenberg and B. Fernanda. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008.

[31] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow, and W. Richland. Visualizing the Non-Visual: Spatial analysis and interaction with information from text documents. In *IEEE Symposium on Information Visualization*, page 51, 1995.

[32] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. *Comput. Graph. Forum*, 27(3):1047–1054, 2008.