SolarMap: Multifaceted Visual Analytics for Topic Exploration

Nan Cao*, David Gotz†, Jimeng Sun†, Yu-Ru Lin‡ and Huamin Qu*
*Hong Kong University of Science and Technology
† IBM TJ Watson Research Center
‡ Harvard University and Northeastern University

Abstract—**Documents in rich text corpora often contain multiple facets of information. For example, an article from a medical document collection might consist of multifaceted information about symptoms, treatments, causes, diagnoses, prognoses, and preventions. Thus, documents in the collection may have different relations across each of these various facets. Topic analysis and exploration for such multi-relational corpora is a challenging visual analytic task. This paper presents SolarMap, a multifaceted visual analytic technique for visually exploring topics in multi-relational data. SolarMap simultaneously visualizes the topic distribution of the underlying entities from one facet together with keyword distributions that convey the semantic definition of each cluster along a secondary facet. SolarMap combines several visual techniques including 1) topic contour clusters and interactive multifaceted keyword topic rings, 2) a global layout optimization algorithm that aligns each topic cluster with its corresponding keywords, and 3) 2) an optimal temporal network segmentation and layout method that renders temporal evolution of clusters. Finally, the paper concludes with two case studies and quantitative user evaluation which show the power of the SolarMap technique.**

Keywords-**Visual Analytics, Multifaceted Information Visualization, Temporal topic visualization**

I. Introduction

Text mining techniques have been developed and applied to identifying patterns such as topics from large corpus in various domains. However, those topics need to be understood by domain users in order to be useful. In recent years, a number of visualization techniques have been developed to assist in this challenge. Topic discovery and visualization in particular has received significant attention with several systems designed to mine and render clusters of related documents. A commonly followed approach is to use some variation of spatially arranged clusters, rendered for example as a density map or an elevation map. The spatial arrangement of these maps is used to represent the relationship between clusters according to some metric, while labels or tag-clouds can be added to convey some aspect of information associated with each cluster.

While effective at showing an overview of a document collection, the conventional approach is limited in its ability to show multiple dimensions of information about the document clusters simultaneously. In addition, these techniques often make it difficult (if not impossible) to visually identify relationships between individual documents, or how a document fits within a given cluster. Unfortunately, many real-world use cases require this sort of multi-relational, multi-scale analysis.

For example, consider an analysis of a collection of articles about various diseases. It is not enough for an analyst to see which diseases fall into a given cluster. A detailed analysis requires that the visualization convey *why* two diseases may fall into the same cluster (e.g., shared symptoms or treatments) or what overlap may exist between two different yet nearby clusters.

To support the type of analysis described above, we propose SolarMap, a new interactive visualization technique that combines a labeled contour-based cluster visualization with a novel radially-oriented tag cloud technique. SolarMap enables multi-relational visualization of document collections at both the cluster and individual document scales.

As temporal dependent data such as social media, publications become upiquitous, topic evolution becomes an important problem for data mining community. There are many temporal topic discovery methods: some assume smooth topic evolution over time [1], [2]. However, less attention has made to provide visual interactive mechanisms to detect and explore the topics in the data in an effective manner.

To address these challenges, we propose SolarMap, a visual analytic technique with the following key features:
A cluster-aligned multifaceted radial tag-cloud technique. SolarMap employs a novel tag-cloud display of multifaceted textual metadata that is arranged radially around an interior cluster-based context preserving rendering of the dataset. Color coding and optimized radial alignment are used to tie tags to corresponding clusters without the need for visually distracting edges. Multifaceted information is laid out on to different radial rings of which one is shown at any given time.
Rich coordinated interaction for visual analysis. SolarMap provides a rich set of interaction tools coordinated across all visual elements of the visualization to enable detailed analysis at document and cluster scale. Dynamic highlighting and edges are used to selectively pinpoint relationships as users interact with visual objects. Controls are also provided for users to switch between radial tag rings to focus on facets of interest during the analysis of multidimensional datasets.
Optimal temporal segmentation and layout method To detect and visualize topic evolution, SolarMap detects signif-

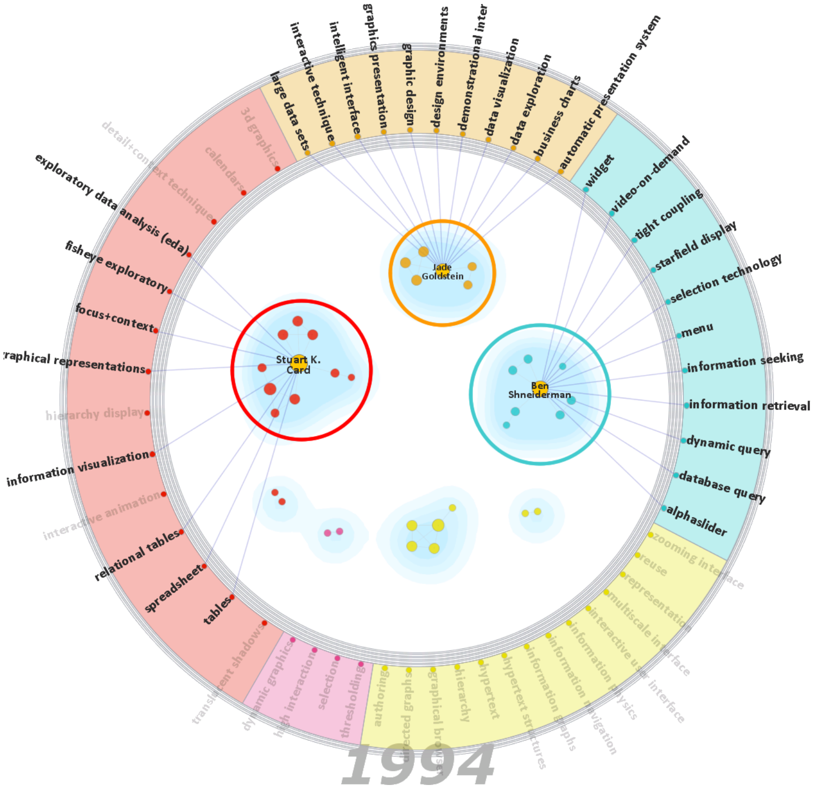


Figure 1. Overview of SolarMap. The InfoVis research communities and their research topics in the year 1994.

icant changes of the clusters across different time segments based on a hierarchical network sequence segmentation. A dynamic visualization is developed to represent topic distribution with smooth transition over time.

In this paper, we describe both the design and implementation of SolarMap. In addition, we demonstrate the power of our approach through a case study and formal user study. The results from our evaluation confirm the effectiveness and general applicability of SolarMap for exploring multi-dimensional relational datasets.

The rest of paper is organized as follows: Section II reviews several areas of related work; Section III presents the visualization design of SolarMap including the data model, visual encoding and layout; Sections IV and V present a case study and evaluation. Finally, the paper concludes in Section VI with a review and discussion of future work.

II. Related Work

This section highlights a number of related text visualization techniques. Topics most relevant to SolarMap include document content visualization, techniques for conveying document relationships, and multifacet visualization techniques.

A. Visualizing Document Content

Many conventional text visualizations are designed to summarize the content of a document or document collection. Perhaps most common is the tag cloud [3] which displays a set of words arranged in rows with font sizes that correspond to statistics such as term frequency. More advanced tag clouds, such as Wordle [4] and Word Cloud [5], enhance the appearance through more sophisticated layouts. In other work, document cards [6] present the content of a document collection using summary cards that include highlighted figures and content topics. The Topic Islands [7] approach uses wavelets to summarize the thematic characteristics of a single document. ThemeRiver [8] visualizes topic evolution using a temporal plot showing the ebb and flow of topic themes over time.

Some techniques attempt to augment the contents of a document with limited relationship information. For example, WordTree [9] and PhraseNet [10] focus on relational word patterns within a given text. In particular, WordTree considers the prefix relation between words at the syntax level. PhraseNet allows user defined relationships.

In contrast with this prior work, SolarMap combines content visualization techniques with visual cues that represent cross-document relationships across multiple facets.

B. Visualizing Document Relationships

Other text visualization techniques focus on displaying relationships between documents in a collection. For example, many visualizations [11], [12], [13] work by mapping a set of documents to a 2D display according to document similarity. Other projections, such as probabilistic latent semantic model [14], 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. SolarMap employs its own projection technique to create document clusters. However, these clusters are augmented with significant additional information designed to convey more clearly what each cluster represents.

In other work, Collins et al. [15] introduce parallel tag clouds (PTCs) to visualize text along multiple facets arranged as columns of words. Links across columns represent co-occurrence relationships. This technique can be very powerful but is limited to word-level relationships. In contrast, SolarMap can visualize more complex latent relationships between documents.

C. Multifaceted Text Visualization

Most recently, a number of systems targeting multifaceted text corpora have been proposed. These designs combine multiple visual techniques to depict information about both document content and inter-document relationships. For example, ContexTour [16] and FacetAtlas [17] are two systems in this category. ContexTour uses a multi-layer tag cloud design that combines clusters with their layered tag clouds which use one layer to represent the content of a cluster for each facet. However, this "content-focused" design users does not convey any information about individual documents/entities or their individual relationships. In contrast, FacetAtlas provides a query based interface which focuses specifically on visualizing complex multifacet relationships. However, FacetAtlas shows no information about the actual contents of the documents.

SolarMap captures the advantages of both ContexTour and FacetAtlas within a single integrated visualization technique. To highlight the benefits of our approach, Section V provides the results from a formal user study which compares the ContexTour and FacetAtlas techniques with SolarMap in an objective task-oriented evaluation.

III. Method

This section describes the details of the SolarMap visualization technique. We first review the SolarMap data model and describe how document corpora are transformed to fit into this model. We then define the visual encodings and layout algorithm used to render the transformed data for display.

A. Data Model and Transformation

Documents are typically unstructured in nature. Visualizing the content of a document corpus and the relationships between documents requires that these unstructured artifacts be transformed into a structured form. SolarMap uses a multifaceted entity relational data model to represent this information in a structured way. Figure 2 illustrates the processing pipeline used to transform a set of raw unstructured documents into our data model.

The first stage in the transformation pipeline is ***facet segmentation***. During this stage, each document is segmented into facet snippets. While various techniques could be used, we typically employ a topic modeling technique such LDA [18] and treat each topic as a facet. When processing documents with a well defined structure (e.g. online Google Health documents which have standard sections for symptoms, treatments, etc.), we directly use the sections to define facet snippets.

Entity extraction is the second transformation pipeline stage. In this step, a named entity recognition algorithm is applied to each facet's document snippet to generate a set of typed entities. Domain-specific ontology models are used to recognize meaningful entities for each facet. For example, in Google Health documents, entities in the symptom facet could include "increased thirst" or "blurred vision", while "type 1 diabetes" and "type 2 diabetes" are entities in the disease facet.

The third and final stage in the processing pipeline is ***relation building.*** In this stage, connections between extracted entities are established using two types of relations: internal relations and the external relations. An ***internal relation*** connects entities within the same facet. For example, the entities "type-1-diabetes" and "type-2-diabetes" are connected within the disease facet by an internal relation. An ***external relation*** is a connection between entities from different facets. For example the disease "type-2-diabetes" is connected to the symptom "increased thirst" by an external relation because "increased thirst" is a symptom of "diabetes-type-2".

B. Design Principles and Visual Encoding

The visual encoding used to represent the information in the SolarMap data model is motivated by several key design principles.

Focus + Context. In SolarMap, there is one facet selected at any given time to serve as the *topic facet*. Entities in the topic facet (which we call topic entities) are considered in focus and are rendered as nodes arranged within the central region of the visualization. The topic entities are clustered by their internal relations to determine the nodes' spatial positions. Contours are then rendered to further highlight the cluster structures. The value of each topic entity is rendered on top of the node, resulting in a clustered tag cloud of topic entity labels.

All other facets in the data model are considered *keyword facets*. Keyword facets are visually encoded as surrounding rings that circle around the central topic cluster region. Entities within a keyword facet are called keyword entities. Only keyword entities from a single selected keyword facet

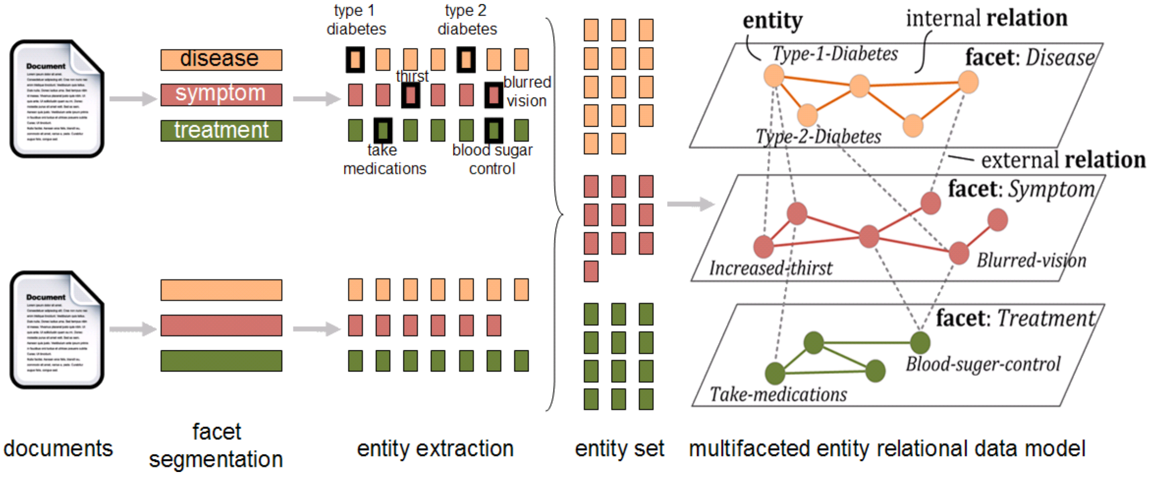


Figure 2. Data transformation process and the multifaceted entity-relational data model.

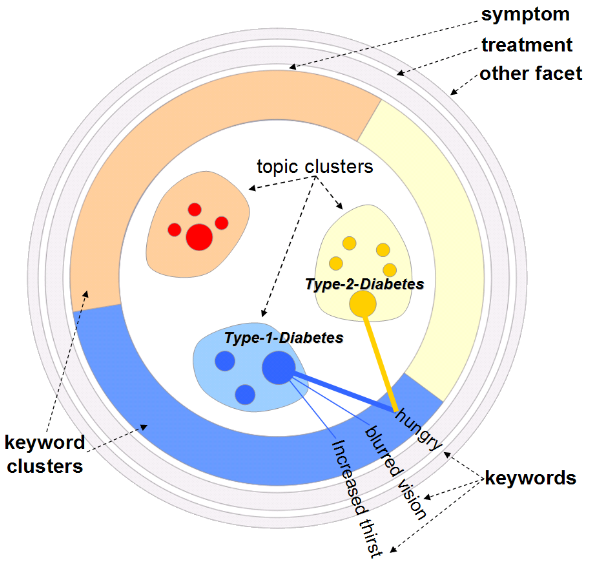


Figure 3. Visual Encoding

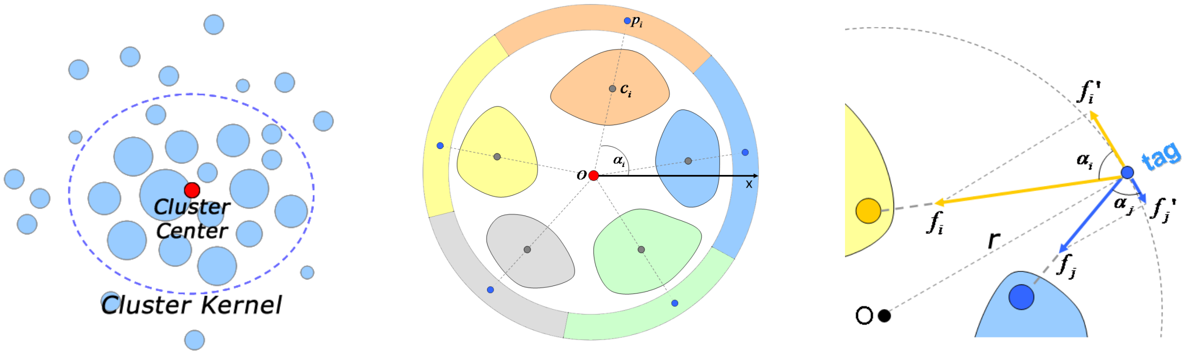
are rendered at any given time. Keyword entities are displayed as radial tag clouds and provide secondary contextual information about each cluster. The radial tags are grouped based on the clusters identified along the primary topic facet. This forms wedge-shaped sections along each ring which one wedge for each cluster. The size of each wedge indicates the size of the corresponding topic cluster, and the correspondence between cluster and wedge is captured using both color and position.

For example, in Figure 3, *Disease* is selected as the topic facet with “Type-1-Diabetes” being one topic entity. *Symptoms* and *Treatments* are both keyword facets. In this example, *Symptoms* is the selected keyword facet resulting in keyword entities such as “blurred vision” and “increased thirst” being visualized along the corresponding ring. These entities appear in the blue wedge of the symptom ring because they are common symptoms for diseases in the blue cluster found in the center of the figure.

Content + Relations. SolarMap is designed to provide a unified visualization of both content entities and the relationships between them. As mentioned above, topic entities and keyword entities are rendered as clustered tag clouds and radial tag clouds, respectively. Internal relations in the topic facet are encoded by screen distance between primary entities. External relations are encoded as lines that each primary entity with related keyword entities in the selected facet ring. Each line is colored by its topic entity’s cluster and line thickness represents the number of topic entities related to the same keyword entity.

Rich Interaction. SolarMap includes a number of interactive features to enable rich data exploration. In addition to traditional tools like dynamic query and filtering, two more sophisticated interactions are supported. First, SolarMap’s *context switch* capability allows users to change both the center topic facet and the surrounding keyword facets. Users can change the facet assigned to be the topic facet by double-clicking on any keyword facet ring. Users can change the selected keyword facet by single-clicking on a facet ring.

The other powerful interactive feature provided by SolarMap is *relation highlighting*. By default, the lines representing relations are not rendered to limit visual complexity. Moving the mouse over any entity selectively displays the lines representing its external relations. The textual tags for connected entities are also highlighted. Multiple selection, via mouse clicks, is also possible to highlight relations across multiple entities simultaneously. This technique is very effective at supporting entity comparison across various keyword facets.



(a) (b) (c)

Figure 4. SolarMap Layout. (a)cluster center detection, (b)keyword wedge reordering, (c)optimized cluster alignment

C. Layout

The design outlined above introduces several constraints on the SolarMap layout. Fortunately, some of the constraints are well studied problems where existing techniques can be leveraged. For example, the radial tag cloud layout can utilize prior designs such as TextArc[1]. However, there are also some new layout challenges. In particular, we must align topic clusters with their corresponding keyword ring wedges to help users map between these two facets of information.

Generally speaking, the SolarMap layout algorithm has two major steps. In the first step, we arrange topic entities in the central area of the visualization using a stabilized graph layout algorithm. The positions are then used to generate contours using a kernel density estimation technique. In the second step, keyword clusters are positioned on the surrounding ring within wedges that are ordered to reduce line crossings and positioned align with their corresponding topic clusters.

1) Topic Cluster Layout: The set of topic entities are connected via internal relations to form a graph as illustrated in Figure 2. During topic cluster layout, a stabilized graph layout algorithm [19] applied to this graph. It minimizes the following energy metric:

$$min(\sum_{i<j}\frac{1}{d_{ij}^2}(||X_i - X_j|| - d_{ij})^2 + \sum_{i<j}||X_i - X_i'||^2) \qquad (1)$$

The first term in this equation places pairs of strongly-connected entities next to each other by minimizing the difference between screen layout distance ($||X_i - X_j||$) and graph distance (d_{ij}). The second part of the equation is a smoothness term which minimizes the change in distance between an entities position at sequential time-steps during animation.

After laying out the entities, we render contours to highlight clusters using kernel density estimation [20]. This algorithm places a Gaussian kernel over each entity and uses the joint distribution $f(x,y)$ of these kernels as the approximated information density. We adjust the bandwidth of each kernel to get distribution with a high degree of smoothness. Finally, contour lines are generated using a contour plotting algorithm [21]. The details of this approach are described in [17].

2) Keyword Cluster Layout: After the topic clusters are positioned, this step positions the color-coded keyword wedges on the surrounding facet ring next to their corresponding topic clusters. The wedges within the ring are first reordered based on the centroid of each topic cluster. This reduces line crossings when external relations are displayed. Then, a force based optimization model is used to rotate the ring such that the distances between the wedges and their related topic clusters is minimized.

Cluster Center Detection. Center detection for each topic facet cluster C_i begins by first extracting its kernel set C_i'. Using the kernel set we detect and remove any outlier entities that are far away from other cluster members. Then, the convex hull P of C_i' is computed and used as cluster boundary. Finally, a center of mass is computed by considering the joint kernel density distribution $f(x,y)$ within the boundary P using the following formula:

$$C_x = \frac{\int x f(x,y) dx}{\int f(x,y) dx}, \; C_y = \frac{\int y f(x,y) dy}{\int f(x,y) dy} \qquad (2)$$

To accelerate the layout process, we treat the density distribution as a constant. This approach reduces the above formulas to the following:

$$\begin{aligned} C_x &= \frac{1}{6A}\sum_{i=0}^{N-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \\ C_y &= \frac{1}{6A}\sum_{i=0}^{N-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) \end{aligned} \qquad (3)$$

where A is the area of P, (x_i, y_i) is the ith vertex of polygon P.

[1] http://www.textarc.org/

Keyword Wedge Ordering. To reduce line crossings and minimize the distances between keyword wedges and their associated topic clusters, we organize the wedges based on the angular position of the topic clusters using a projection line technique. We first project the center of each topic cluster's contour C_i out to the surrounding ring by using a projection line that starts at the center of the visualization canvas. The projection line for C_i intersects the facet ring at point p_i as shown in Figure 4(b). The radial order of these positions are then used to order the keyword wedges.

Optimized Cluster Alignment. After ordering the wedges, the final step is optimized cluster alignment which rotates the keyword facet ring to an angle that best aligns each wedge with its corresponding topic cluster. The alignment is accomplished through the force-based optimization model defined below.

$$min \sum_i (f_i \times r \times \cos(\alpha_i)) \tag{4}$$

The model minimizes the sum of the computed forces for all external relations i between the topic entities and the displayed keyword entities. The force equation is based on the moment of force where f_i is a spring-force equation based on the distance between the pair of related entities, r is the radius of the ring, and α_i is the angle of the edge representing the relation. These terms are illustrated in Figure 4(c). This model will rotate the facet until the sum of the forces is minimized, resulting in a ring that is optimally aligned with the interior topic entities.

3) Temporal Sequence Layout: In the presence of time-evolving topics, we propose a visualization technique that displays topic cluster changes in a smooth and continuous fashion so that users can easily follow the underlying topic shifts. Recall that the layout algorithm is based on graph visualization. Therefore, when we visualize an evolving graph over time, the stability over time need to be enforced. To encode the temporal constraints, we propose a recursive temporal segmentation method to partition a sequence of graphs into a set of graph segments. Such a segmentation can be done recursively to further group graph segments into longer segments as shown in Figure 5.

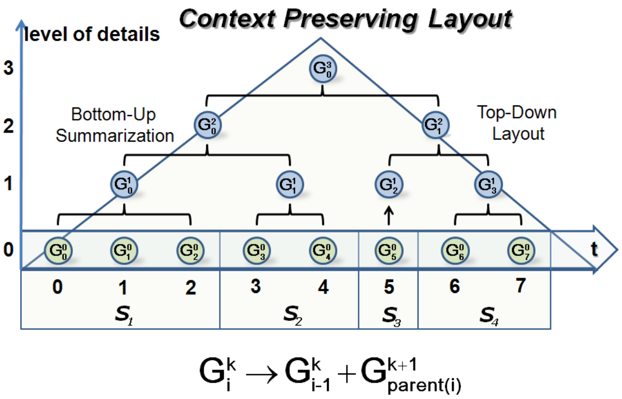


Figure 5. Graph sequence segmentation and the context preserving layout algorithms.

Formally, given a graph sequence $GS = \{G_1, G_2, \ldots, G_T\}$, the k-partitioning of GS is to partition the graph sequence GS into k segments $S_1, S_2, \ldots, S_k$, each has size T_i. And $T = \sum_{1 \leq i \leq T} T_i$. Specifically, segment S_i consists of $G_{s_i}, \ldots, G_{e_i}$ for $1 \leq i \leq k$, where $s_i(e_i)$ is the starting(ending) index for segment S_i. A graph segment S_i can be approximated by averaging all the graphs within S_i, denoted by S_i. The approximation error for S_i is

$$sse(S_i) = \sum_{s_i \leq j \leq e_i} ||G_j - S_i|| \tag{5}$$

where $||G_j - S_i||$ defines as the number of inconsistent edges. Then through dynamic programming, we can formulate the problem of finding the best k-partitioning of GS such that the total error $\sum_{1 \leq i \leq k} sse(S_i)$ is minimized. We define the cost function $F(T,k)$ as the minimal cost of partitioning graph sequence $GS = \{G_1, \ldots, G_T\}$ into k segments. The following recursion defines the optimal substructure for the dynamic programming:

$$F(T,k) = \min_{t<T}(F(t,k-1) + sse(S_k)) \tag{6}$$

where $S_k = \{G_{t+1}, \ldots, G_T\}$. This recursion says the minimal cost for partitioning graph sequence GS into k segments is the optimal sum of the minimal cost for partitioning a subset of GS into $k-1$ segments and the error of the last segment S_k.

By solving such dynamic programming problems for different k, we can construct a multi-level Directed Acyclic Graph (DAG). The resulting structure is not necessarily a tree in general. Following a topological order of DAG, we can layout the graph sequence in a top-down fashion. In this process, the position of node x of graph G depends on both the structure of G as well as the positions of x in G's parent graphs.

The layout method will preserve the stability of topic clusters over time, which can be used for visualizing any dynamic topic modeling results.

IV. Case Studies

To demonstrate the utility of our approach, we applied SolarMap to two use cases. First, we developed a healthcare application to analyze the 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. Second, we developed a tool to visualize research community evolution using DBLP data. The DBLP dataset spans from 1992 through 2002, and includes over 800 researchers, 614 papers and 2000 keywords.

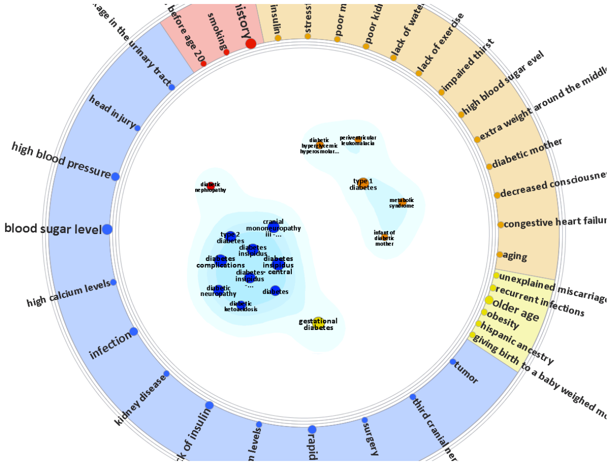


Figure 6. SolarMap visualization of the Google Health library.

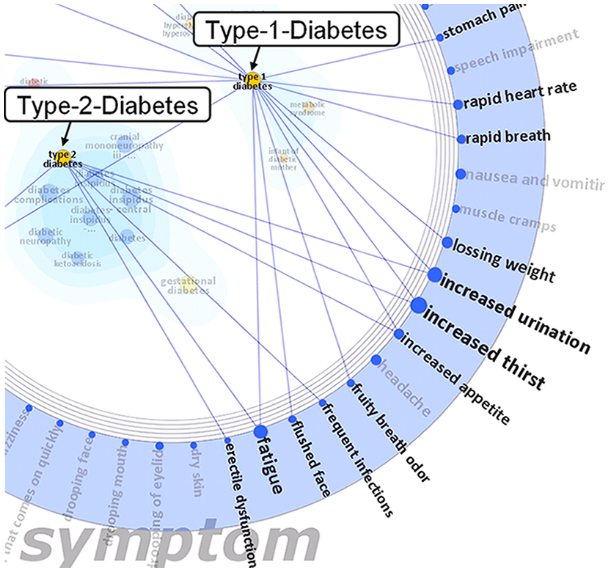


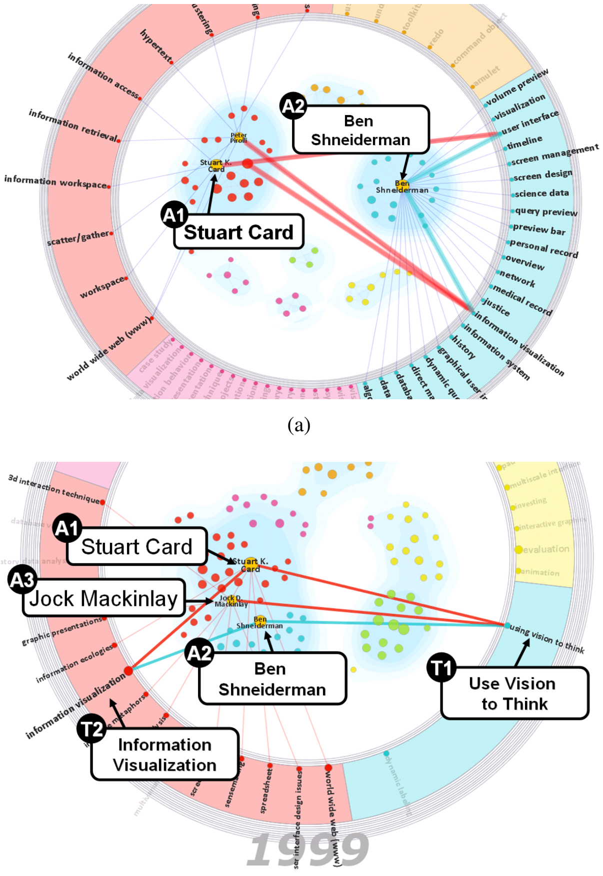
Figure 7. Case Study on Diabetes

A. Case Study One: Healthcare

The Google Health data has several facets. For our initial exploration, we selected disease name as the topic facet. The diseases appear as topic clusters in the center of Figure 6. Other facets are visualized using the surrounding keyword rings.

A key strength of SolarMap is the ability to explain relations between entities. For example, SolarMap can easily explain how two diseases are related to each other. To do that, we double click on the diseases we wish to compare to select them. This highlights the external relations for the selected diseases as as shown in Figure 7. By switching through different keyword facets (e.g, symptom, complication, and cause), we can easily observe that Type-1-Diabetes and Type-2-Diabetes are related because they share similar symptoms such as "increased urination" and "fatigue", as well as similar complications such as "kidney disease " and "Stroke". However, they do not share any common causes. This case study demonstrates the capability of SolarMap to explain clusters through the links between topic clusters and keyword clusters.

B. Case Study Two: DBLP



(a)

(b)

Figure 8. Case study on DBLP data. (a) In 1996, visualization communities were rather isolated but began to study similar topics. (b) By 1999, the communities were collaborating more closely and had even more research topics in common.

Time plays an important role in the DBLP dataset as it captures the evolution of research topics and teams over several years. This case study examined changes in the InfoVis community from 1992 through 2002. As illustrated in Figure 1 and Figure 8, we use author names as the topic facet, and paper keywords grouped by year as the keyword facets. The years are ordered allowing easy navigation through time using the keyword rings.

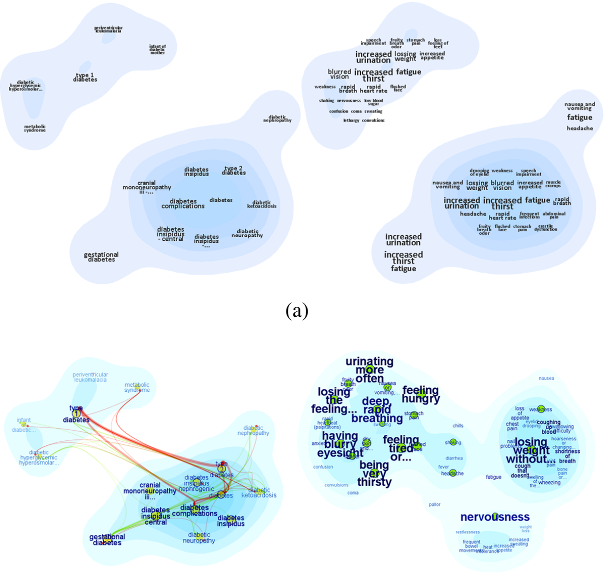
Exploring the data year by year, we found some interesting evolution patterns. In the first years, such as 1994 (see Figure 1), several isolated author clusters emerged. The largest were led by Ben Shneiderman and Stuart K. Card.

Shneiderman's cluster focused most on interaction designs such as "dynamic query" and information exploration such as "information seeking and retrieval". In contrast, Card's group focused more on "graphical representation" and "explorative data analysis".

In 1996, researchers in both clusters began working on a few similar topics such as "Information Visualization" and "User Interface", as indicated by the common links to those keywords on Figure 8(a). However, as shown by the author clusters, the research communities were still not directly collaborating. However, by 1999, the clusters begin to merge. This merger, as shown in Figure 8(b), occurs around the time that Card and Shneiderman join as two of the co-authors on the book "Using Vision to Think".

V. Evaluation

In addition to the case studies described above, our system was evaluated quantitatively through a formal user study. This section describes the study's design and presents both objective and subjective results.



(a)

(b)

Figure 9. Baseline systems in our user study. (a) ContexTour, the disease tag clouds (left) and related symptoms tag clouds (right). (b) FacetAtlas, the disease view (left) and Type-1-Diabetes' symptoms view (right).

A. Study Setup

To evaluate the effectiveness and efficiency of SolarMap in support of multifaceted data analysis, we conducted a comparison study. Our study compared SolarMap with two baseline systems: ContexTour [16] and FacetAtlas [17]. See Section II-C for a comparison of features in these two baselines with the features of SolarMap.

Tasks. We applied all three systems to the same Google Health dataset and and users in our study to perform a series of analysis tasks. The tasks in our study were as follows:

- **T1:** *Identify all clusters of diseases that match the query term "diabetes".* This task tests a tool's ability to convey clusters.
- **T2:** *Identify the top 3 symptoms for a specified disease cluster.* This task tests how well a tool allows users to interpret of clusters.
- **T3:** *Identify the top 3 symptoms shared between two specified disease clusters.* This task tests a tool's ability to compare clusters across specific facets.

These tasks increase in complexity from relatively simple (T1) to complex (T3). The tasks were chosen to simulate a common self-care education scenario for chronic diabetic patients. Together, these three tasks represent a concrete use-case of analyzing the Google Health library for self diagnosis.

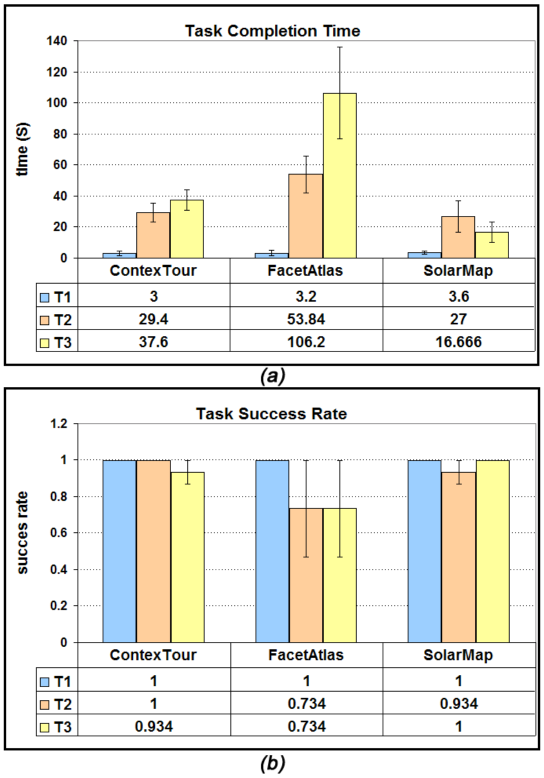
Participants and Methodology. We recruited 15 participants for our study (9 researchers and 6 students majoring in computer science, psychology and mathematics). Inspired by the repeated-measures study methodology [22], we divided the participants into three groups of 5. The first group used FacetAtlas for task T1, ContexTour for task T2, and SolarMap for task T3; the second group used the ContexTour for task T1, SolarMap for task T2 and FacetAtlas for task T3; finally the third group used SolarMap for task T1 and FacetAtlas for task T2 and ContexTour for task T3. At the beginning of each user session, We gave a brief tutorial of all three systems. The participants were then asked to complete the three assigned tasks.

We recorded two objective measures: task completion time (the time spent on each task, measured in seconds), and task success rate (the percentage of assigned tasks completed successfully). We computed the mean and standard deviation of task completion time and task success rate across all users and tasks.

We also recorded subjective measures via user surveys. We compared SolarMap with the two baseline systems on aesthetics, ease of use and usefulness. To further evaluate the design of SolarMap, we asked users to score specific aspects of the visualization in terms of (1) usefulness (how useful a system is for solving a specific task) and (2) usability (how easy the system was to use for a specific task).

B. Objective Results

As shown in Figure 10(a), the baselines systems (ContexTour and FacetAtlas) both exhibit an increasing trend in task completion time from T1 to T3. This confirms the increasing complexity across tasks. For the relatively simple task T1, all three visualizations perform equally well with less than 4 seconds spent on the task on average. For the medium difficulty task T2, SolarMap shows a small advantage when compared to baseline methods while FacetAtlas requires significantly more time. This is because FacetAtlas requires an context switch to view individual symptoms,



	ContexTour	FacetAtlas	SolarMap
T1	3	3.2	3.6
T2	29.4	53.84	27
T3	37.6	106.2	16.666

(a)

	ContexTour	FacetAtlas	SolarMap
T1	1	1	1
T2	1	0.734	0.934
T3	0.934	0.734	1

(b)

Figure 10. Study results comparing (a) task completion time and (b) task success rate.

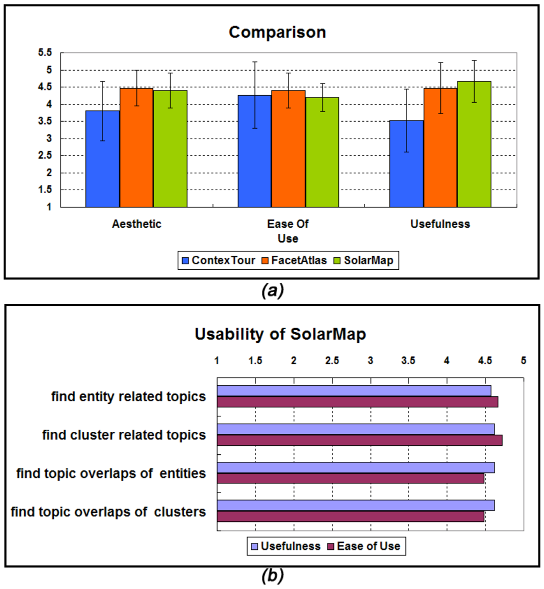
while ContexTour and SolarMap can directly present the related symptoms on the same view.

For the most complex task T3, SolarMap outperforms both baselines significantly. Interestingly, SolarMap requires less time on T3 than T2, despite T3 being the most complex task for both baseline systems. We believe that the major reasons for this are that (1) the relation highlighting feature of SolarMap helps users to quickly identify common symptoms, and (2) the shorter list of shared symptoms in T3 (compared to the longer list of symptoms in T2) makes it easier to identify the top thee symptoms.

These results show that SolarMap provides a strong reduction in task completion time for more complex tasks. In particular, a two-way repeated measures ANOVA analysis shows that when compared with the FacetAtlas system on T2, both SolarMap and ContexTour yield a significant efficiency improvement (T2, SolarMap $p = 0.014 < .05$, ContexTour $p = 0.018 < .05$). Similarly, the performance improvement on T3 is also significant (T3, SolarMap $p = 0.005 < .05$, ContexTour $p = 0.014 < .05$). In both cases SolarMap performs best, and for task T3 SolarMap is significantly better than ContexTour ($p = 0.003 < .05$).

We also compared task success rates as shown in Figure 10(b). Both SolarMap and the ContexTour achieve similarly high accuracy levels. Only one error was observed for each of these systems. In contrast, the FacetAtlas yielded lower success rates on both tasks T2 and T3. We believe the drop in accuracy was due to the need for a context switch between diseases and symptoms when using FacetAtlas which may have forced users to lose the context of the original clusters. This is further evidenced by the fact that these least accurate tasks (FacetAtlas T2 and T3) were also the ones that users spent the most time completing.

C. Subjective Results



(a)

(b)

Figure 11. (a) Comparison of ratings for aesthetics, ease of use, and usefulness of the three designs. (b) Usability feedback for SolarMap's key features.

In addition to the quantitative results presented above, we gathered subjective feedback through user surveys. All participants were asked to compare SolarMap with ContexTour and FacetAtlas in terms of aesthetics, ease of use and usefulness. The survey asked users to score each method from 1 (lowest) to 5 (highest) in each of these dimensions. The results are shown in Figure 11(a). In terms of aesthetics, users liked both FacetAtlas and SolarMap more than ContexTour. Ease of use scores were relatively even across all three tools. However, when considering usefulness, users felt that SolarMap was the most useful visualization among three.

Finally, we collected qualitative user feedback on the key capabilities of SolarMap. For each of four capabilities (find entity related topics, find cluster related topics, find topic overlaps of entities, and find topic overlaps of clusters) we had users provide scores (from 1 to 5) for usefulness and ease of use. Figure 11 summarizes the results. All features exhibited fairly high scores for both usefulness and ease of use. This confirms that users felt comfortable with SolarMap and were confident in its ability to support the assigned analysis tasks.

VI. Conclusion

This paper presents SolarMap, a multifaceted visual analytic technique for visually mining and exploring topics in temporally evolving multi-relational data. SolarMap simultaneously visualizes the topic distribution of the underlying entities from one facet together with keyword distributions that convey the semantic definition of each cluster along a secondary facet, and also provides smooth visual transition of temporal evolution of topic clusters. As described in this paper, SolarMap combines several visual techniques including 1) topic contour clusters and interactive multifaceted keyword topic rings, 2) a global layout optimization algorithm that aligns each topic cluster with its corresponding keywords and 3) optimal temporal segmentation of evolving topic sequences.

We also described two use cases where SolarMap can be applied and conducted a formal user study to compare our new technique with two competing baseline systems. Both the objective and subjective results from our study show that SolarMap outperforms the baseline systems in many areas. In future work, we plan to apply SolarMap to more applications, to conduct more thorough user studies.

References

[1] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML '06, 2006, pp. 113–120.

[2] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *KDD*, 2007, pp. 687–696.

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

[4] F. Viégas, M. Wattenberg, and J. Feinberg, "Participatory Visualization with Wordle," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1137–1144, 2009.

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

[6] 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*, vol. 15, no. 6, pp. 1145–1152, 2009.

[7] N. Miller, P. Wong, M. Brewster, and H. Foote, "TOPIC ISLANDS - a wavelet-based text visualization system," in *Visualization'98. Proceedings*, 1998, pp. 189–196.

[8] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: Visualizing theme changes over time," in *Proceedings of the IEEE Symposium on Information Vizualization*, 2000, pp. 115–123.

[9] M. Wattenberg and B. Fernanda, "The word tree, an interactive visual concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1221–1228, 2008.

[10] F. van Ham, M. Wattenberg, and F. Viégas, "Mapping text with phrase nets." *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, p. 1169, 2009.

[11] 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*, vol. 3, no. 4, pp. 166–181, 2002.

[12] Y. Chen, L. Wang, M. Dong, and J. Hua, "Exemplar-based Visualization of Large Document Corpus (InfoVis2009-1115)," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1161–1168, 2009.

[13] 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," *Proceedings, information visualization: October 30-31, 1995, Atlanta, Georgia, USA*, p. 51, 1995.

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

[15] 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)*. IEEE, 2009, pp. 91 – 98.

[16] 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.

[17] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "Facetatlas: Multifaceted visualization for rich text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 1172–1181, 2010.

[18] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

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

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

[21] C. Singh and D. Sarkar, "A simple and fast algorithm for the plotting of contours using quadrilateral meshes," *Finite Elements in Analysis and Design*, vol. 7, no. 3, pp. 217 – 228, 1990.

[22] R. Bakeman and B. Robinson, *Understanding statistics in the behavioral sciences*. Lawrence Erlbaum, 2005.