32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

DiseaseAtlas: Multi-facet Visual Analytics for Online Disease Articles

Jimeng Sun, PhD[1], David Gotz, PhD[1], Nan Cao, MCS[2]
[1]IBM Research; [2]HKUST

Abstract— **Online health information portals provide valuable content to casual consumers. However, the page-oriented nature of these resources makes it difficult for users to understand the overall information space and navigate the complex relationships between various diseases. We have developed a visual analytic system named DiseaseAtlas that helps users navigate a large set of disease-related documents and understand multi-dimensional relationships for key semantic concepts such as symptoms and treatments. This paper describes several unique aspects of DiseaseAtlas and demonstrates its capabilities through a case study.**

I. Introduction.

Patients have access to a wide variety of online health information portals. Sources such as WebMD, MayoClinic.com, and Google Health provide provide concise, easy-to-understand descriptions of diseases with information on symptoms, diagnosis, prevention, and treatment. Unfortunately, while studies have shown that non-experts turn to online information to explain undiagnosed symptoms [16], the organization of these document-based resources is best suited for situations where patients are searching for a specific disease after receiving a diagnosis. In this work, we aim to assist patients during self-diagnosis. In these cases, patients must navigate a web of inter-related disease properties (e.g., symptoms, treatment, or causes) to identify which diseases may be most relevant. Therefore, our visualization is designed to help provide decision support for patients [6] prior to diagnosis.

Today, these portals rely on traditional 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. However, ranked lists are not ideal for self-diagnosis use cases where users must explore relationships between complex medical concepts that span across multiple disease-related articles. Despite recent work on more advanced search interfaces [7], [11], [15], the effective organization and presentation of search retrieval results within the medical domain is still largely an open problem.

This problem becomes even more challenging when considering the multi-facet nature of many medical documents. A search engine, for instance, 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?*

To bridge this gap, we propose DiseaseAtlas, a new interactive visual analytic system that enables users to navigate and analyze large multi-faceted disease text corpora with complex cross-document relationships. Specifically, DiseaseAtlas provides the following key features:

Visualization of both global and local patterns. DiseaseAtlas employs a multi-faceted graph-based display to visualize local relations and a density map to portray a global context.

Integrated unstructured search and visualization. DiseaseAtlas 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 multi-facet relationships.

Dynamic facet-based context switching. In addition to basic interactions such as zooming, filtering and highlighting, DiseaseAtlas supports dynamic context switching. This allows a user to pivot the primary visualization layout arrangement across different facets while maintaining his/her analytic focus.

II. Related Work

In this section we survey several areas of related work.

A. Medical Text Visualization

Visualization techniques have been applied in the past to medical document collections. For example, Synnestvedt et al. [12], [13] have visualized AMIA proceedings to look for trends. In other work, the PubGraph system [1] has been designed to provide a visual interface to PubMed documents. Our work similarly targets the visualization of medical documents. However, DiseaseAtlas provides both greater interactivity and a number of advanced features not found in these tools.

B. Graph Visualization

Various network visualizations (e.g., [5]) have been designed to analyze relational patterns. However, many of these, such as Vizster [4], consider only one type of relationship. In order to visualize multiple types of relationships, Shen et al. [9] introduce OntoVis which uses nodes and links to represent various concepts and relations for large and heterogeneous social networks. At a high level, DiseaseAtlas uses a similar approach but adopts a completely different visual design for presenting and interacting with multi-facet relationships. In other work, SocialAction [10] supports relational pattern detection in social networks through smart filtering of important nodes, clusters and outliers.

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

DiseaseAtlas provides similar interaction tools but focuses on healthcare applications.

C. Visual Search Interfaces

Traditional search interfaces for text corpora present a ranked list of search results. Recognizing the limitations of this approach, others have explored visualization-based search interfaces. For example, van Ham et al. [15] present a visual search tool to allow users to navigate through a subgraph in a huge document network. Smith et al. introduce FacetMap [11] and FacetLens [7] which provide a visualization-based interface for multi-faceted document search. Commercially, Grokker (http://www.grokker.com/) is notable for its use of a circular Treemap visualization of dynamically generated topic clusters on web search results. However, these systems do not consider multi-faceted relationships as is the focus of DiseaseAtlas.

III. System Architecture

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

The *Visualization* module maps the indexed entities and relations to a multi-faceted visual display according to the visual design outlined in Section V. 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.

IV. Data Model and Transformation

In this section, we first define the core constructs of our data model. We then discuss how a set of disease documents are transformed from raw text to fit this model.

A. Multi-facet entity-relational data model

The DiseaseAtlas data model is a multi-faceted representation that captures entities and their relationships. The model consists of the following logical constructs:

- **Entities** are instances of a particular concept from the data. For example, "type-1 diabetes" is a disease entity as shown in Figure 1.
- **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 within the same facet. For example, "type-1 diabetes" has an internal relation to "type-2 diabetes" because they are related 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 examples, a group of diseases related to "type-1 diabetes" forms a cluster on the disease facet.

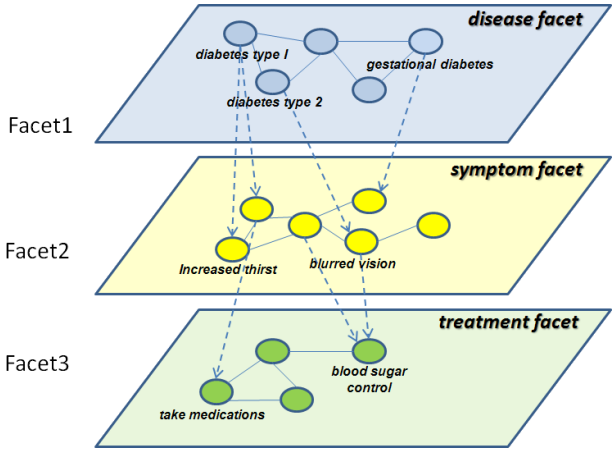


Fig. 1. The DiseaseAtlas multi-facet entity-relational data model. Concepts in a complex text corpus are transformed into *facets*, *entities* and *relations*.

B. Transformation

Before DiseaseAtlas can be used to visualize a corpus of disease text documents, the raw text material must be transformed to fit into the multi-facet entity relational data model described above. The transformation process consists of the following stages: text segmentation, entity extraction, similarity measurement, and index construction.

First, for each document we partition the text into consecutive phrases. We then segment the text into sections based on document formatting structures such as font size. For example, a Google Health article is split into several sections such as *Overview*, *Symptoms* and *Treatments* which have special formatting. After segmentation, we extract keywords from each section. We can either apply a standard text mining approach or rely on a domain-specific parser based on a customized ontology such as UMLS[1]. Next, we construct a similarity graph for the extracted keywords. For this phase, we use either standard information retrieval measures (e.g., cosine similarity) or topic-level similarity through topic modeling (e.g., ContexTour [8]). The initial topics categories are based on the standard ICD-9 classification[2].

Finally we build separate search indices for each facet. DiseaseAtlas 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

[1]www.nlm.nih.gov/research/umls
[2]http://icd9cm.chrisendres.com/

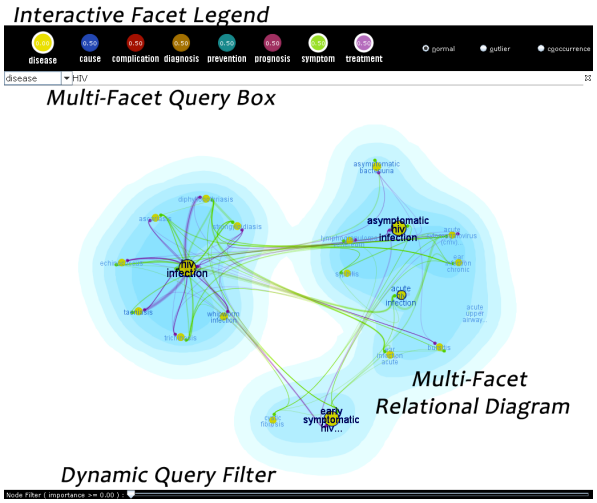


Fig. 2. DiseaseAtlas applied to Google Health data. This figure shows a disease diagram for the search term "HIV." DiseaseAtlas contains four main components: (1) an interactive facet legend, (2) a query box, (3) a canvas for rendering multi-facet relational diagrams, and (4) a dynamic query filter to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

is issued, DiseaseAtlas retrieves and visualizes the most relevant entities and their corresponding relations.

V. DiseaseAtlas Visualization

DiseaseAtlas combines a background-layer density map with multi-faceted graph-based diagrams rendered in the foreground to display both global cluster information as well as detailed pairwise relationships. An overview of the user interface is shown in Fig. 2. This section reviews several key aspects of the DiseaseAtlas design.

DiseaseAtlas employs four basic visual constructs for visually encoding data within the visualization. These include three elementary graphical marks (points, links, and areas) and one visual feature (color) [2], [3]. This section describes how we use these simple visual elements to represent the logical constructs in the DiseaseAtlas data model.

Facet Encoding. Facets are encoded using color. We maximize the color differences in hue based on the HSL color model[3] so that facets can be easily separated. The facet colors are consistently used for both points and links, and the colors remain constant across views as users navigate the visualization. In addition, the visualization differentiates between a single *primary* facet and other *secondary* facets. The interactive facet legend displays the primary facet as the leftmost circle and uses entities from this facet to structure the visualization.

Entity Encoding. An entity is represented using color-coded circles. An entity's color corresponds to the entity's facet. For primary entities (entities that belong to the primary facet), circle size is used to represent an entity's degree-of-interest (DOI) [14]. More specifically, DOI corresponds to the extent that 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 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 connected by external relations). We call these nodes entity nodes and facet nodes, respectively.

Relation Encoding. Two different visual encodings are used within DiseaseAtlas, 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.

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.

VI. Evaluation

A. Setup

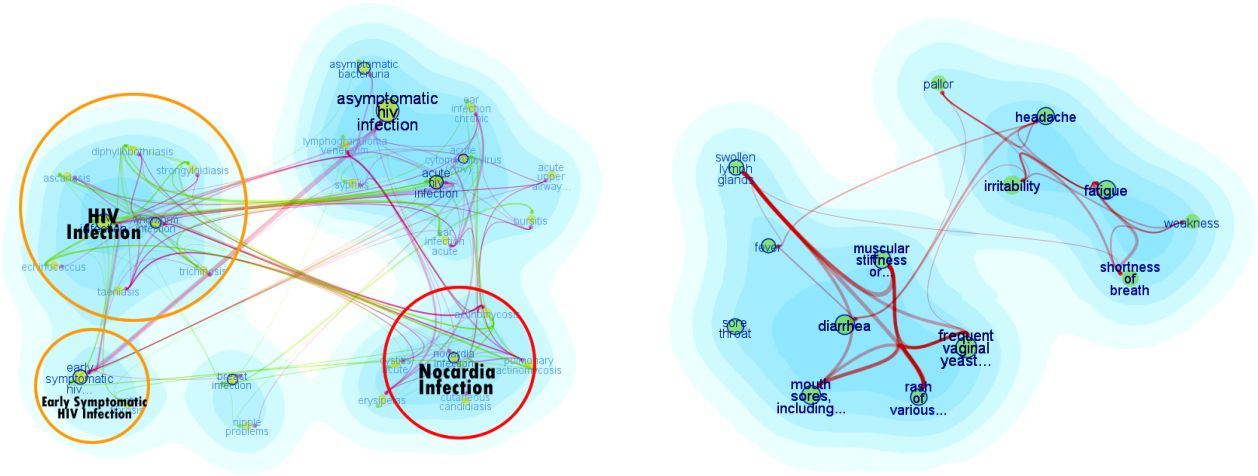
Our case study application is based on the online Google Health library which contains over 1,500 online articles. With data transformation, we converted the articles into a multi-facet entity-relational data model. The transformed dataset contains 8 facets and around 25,000 entities with more than 50,000 internal links.

B. Study on HIV infection

In our case study, we first search for "HIV" using the query box. In response, DiseaseAtlas generates a 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 then turned on the symptom and treatment facets by clicking the corresponding buttons at the top. Interestingly, as seen in Figure 2, 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 symptoms. When we performed a semantic zoom in to this disease, we confirmed that it has strong symptom connections with several other infections (see Figure 3(a)). Furthermore, we context switched to the symptom view to explore its related symptoms in detail. Two symptom clusters were visible. After exploring the multi-facet relationships of those two clusters, we found that these symptom clusters lead to different complications as shown in Figure 3(b).

[3]http://en.wikipedia.org/wiki/HSL_and_HSV



(a) (b)

Fig. 3. Case study on HIV infection. (a) Semantic zoom. When zooming in on "Asymptomatic HIV Infection" from the initial view shown in Figure 2, more related diseases are shown (highlighted in the red circle). The initial context is preserved as highlighted in the yellow 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.

C. Feedback from domain experts

We also performed demo sessions to medical doctors. All of the physicians were very impressed by the interactive visualization that DiseaseAtlas provides. One physician was amazed by DiseaseAtlas. He considered DiseaseAtlas *"...extremely creative, and has great potential for clinical therapeutic usage and diagnosis decision support."* Another physician believes DiseaseAtlas can help with diagnosis support, stating that the cross-cluster links can *" ... enhance the current thought process of physicians, and help create the subtle associations between different concepts."* After we explained the patient education scenario, one 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 are just starting their medical career"*.

VII. Conclusion

We present DiseaseAtlas, a multi-facet visual analytic system for exploring medical documents. DiseaseAtlas is able to visualize both global and local relations of complex disease document collections. DiseaseAtlas also provides rich interactions such as filtering and context switching. These interactions enable users to examine a medical text corpus from multiple perspectives. We performed an in-depth case study on a patient education application in the health care domain. The feedback was extremely positive and confirmed our main design objectives. In future work, we plan to apply DiseaseAtlas to other applications such as visualizing patient records, and to incorporate time dimension in the visualization, and to conduct more thorough user studies.

References

[1] UCLA Center for Cognitive Phenomics. PubGraph. http://www.pubgraph.org/, April 2010.

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

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

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

[5] I. Herman and M. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2000.

[6] H. B. Jimison, P. P. Sher, and J. J. B. Jimison. *Clinical Decision Support Systems: Theory and Practice*, chapter Decision Support for Patients, pages 249–261. Springer, 2nd edition, 2007.

[7] 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: Proceedings of the 27th international conference on Human factors in computing systems*, pages 1293–1302, New York, NY, USA, 2009. ACM.

[8] 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*, 2010.

[9] 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.

[10] 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*, 12:693–700, 2006.

[11] 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*, 12:797.

[12] M. B. Synnestvedt and C. Chen. Visualizing amia: a medical informatics knwledge domain analysis. In *AMIA Annual Symposium Proceedings*, 2003.

[13] M. B. Synnestvedt, C. Chen, and J. Holmes. Visual exploration of landmarks and trends in the medical informatics literature. In *AMIA Annual Symposium Proceedings*, 2005.

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

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

[16] R. W. White and E. Horvitz. Experiences with web search on medical concerns and self diagnosis. In *AMIA Annual Symp Proc*, 2009.