ReportViz: Interactive Visualization and Exploration of Topics and Keywords in Public Health Reports

Wei Zhuo*

Graphics, Visualization and Usability Center
College of Computing
Georgia Institute of Technology

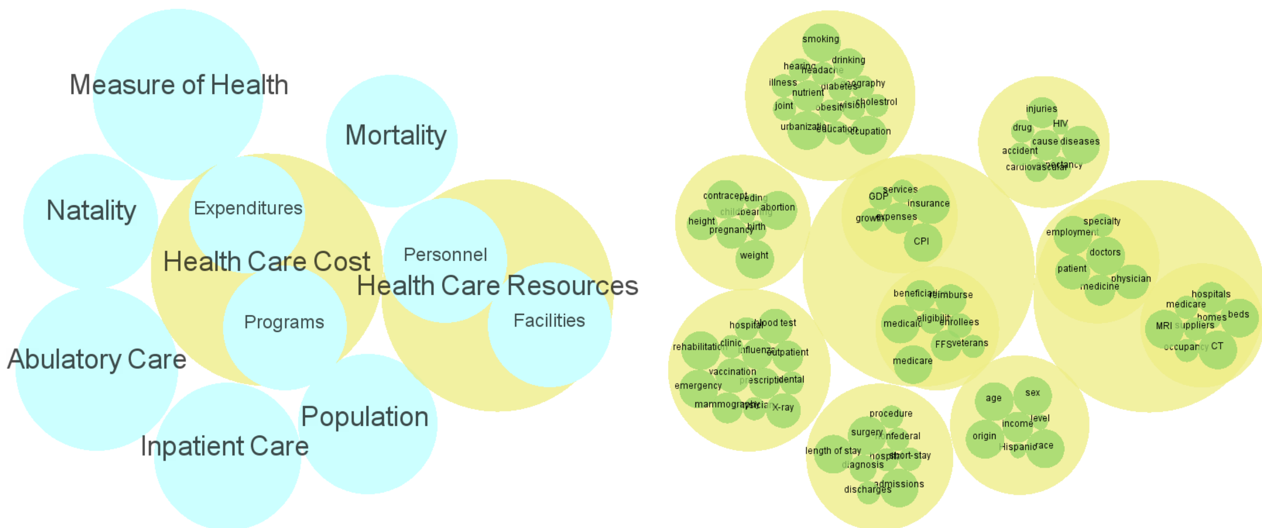

Figure 1: We show topics in a health report as gathered bubbles on the left. Expanding all topic bubbles spawns groups of key word bubbles within each topic bubble (right). The bubble radius encodes the term weight.

Abstract

Public health documents contain a rich set of topics and keywords that cover various aspects of a nation's health system. Each topic is characterized as a distribution over vocabulary from which we extract the keywords. Visualizing this topic structure allows the reader to have an intuitive and convenient overview of a usually lengthy report. In this paper, we describe *ReportVis*, an interactive utility that uses nested bubbles to represent topics, key words and their weights. We use a collection of US Health Reports issued by CDC from 2009 to 2012 as our exemplar corpus to demonstrate a interactive exploration of topics and their keywords in public health. We further introduce an animation mode to allow a dynamic view of topic evolution.

Keywords: Topic Visualization, Key Words in Health

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Animations; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques

*wzhuo3@cc.gatech.edu

1 Introduction

Health organizations regularly issue reports that contain a rich set of subjects covering various aspects of a health system. As the collection continues to grow, it becomes more challenging to mine topics of interest. We search with keywords to find target texts in a document. In addition, we might want to ① explore texts associated with keywords within the same topic and ② have an overview of the topic structure in this document. All these tasks call for an intuitive interface to visualize the topics/keywords and their weights mined from a collection of reports.

We build ReportViz that aim to address these tasks. We choose a collection of comprehensive health reports issued yearly by Centers for Disease Control and Prevention. A nice feature of this collection is its consistency: the set of topics are well-defined. Although different reports may have different focuses and sometimes multiple topics are nested under a broader theme, the total number of topics is fixed and is shown in Tab. 1. We first extract, for each topic, a list of key words with their weights, and then visualize the topic structure. We restrict our focus to visualization in describing ReportViz. Our work claims the following contributions. First, we represent each topic as a tree node and keywords as leaf nodes. This allows convenient operations for topic merge and split. Second, we use physics-inspired layout and interaction techniques that provide engaging experiences for topic exploration and an animation mode for visualizing topic evolution. Last but not least, we show lists of extracted key words as well as the revision made by medical students. The result could serve as future references or set the priors

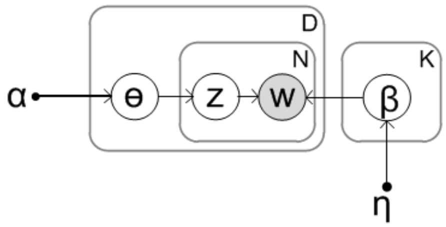

Figure 2: graphical model of Latent Dirichlet Allocation

for topic modeling in health.

2 Prior Art

2.1 Statistics and Text Analysis

Probabilistic topic modeling aims to automatically extract topics and keywords from a collection of documents. The Latent Dirichlet Allocation (LDA [3]) assumes prior distributions (θ, β) over topics and vocabulary, and estimates these parameters from the observed text (w), as shown in Fig. 2.

Tractable implementations includes Gibbs sampling [11] and variational inference [1]. Dynamic Topic Models [2] are a family of probabilistic time-series models developed to analyze the time evolution of topics in a collection of documents ordered by time. Its state space models inspire us to design a dynamic view of the time-varying weights evaluated from texts.

2.2 Visualization

Off-the-shelf visualization models such the tag cloud in ManyEyes [12] can provides web-based visualization of word frequencies: the bubble chart displays a set of numeric values as circles, hence can be used to represent key terms with their frequencies. For compactness, we use physics-inspired techniques to layout the bubbles.

Note that the CVT energy function can be exploited for general icon layout [4]. Cui et al. [5] point out that topic could merge or split over time, and present a static view of topic evolution as a flow graph. Liu et al. [6] use stacked graph to visualize topic evolution by summarizing emails in different times with the output of LDA on the whole corpus.

3 Visualization

3.1 Health Document

Our text corpus consists of yearly health reports issued by CDC from 2009 to 2012 [7] [8] [9] [10]. There are about 2300 pages presenting analysis and tables by topics. The major topics are: *Population*, *Fertility and Natality*, *Mortality*, *Measure of Health*, *Ambulatory Care*, *Inpatient Care*, *Personnel*, *Facilities*, *Expenditures* and *Coverage and Programs*. In addition to lists of automatically extracted keywords, we ask two medical students to input their revision. The revision tasks are:

(1) **Selection**: within each topic and subtopic, we let our colleagues choose 3-10 terms which they think are most related to the topic.

(2) **Addition**: we also ask them to add any term which they think is as relevant as those selected in (1), but is not in the list of extracted key terms.

The resulting group of topics and lists of key terms are shown in Tab. 1.

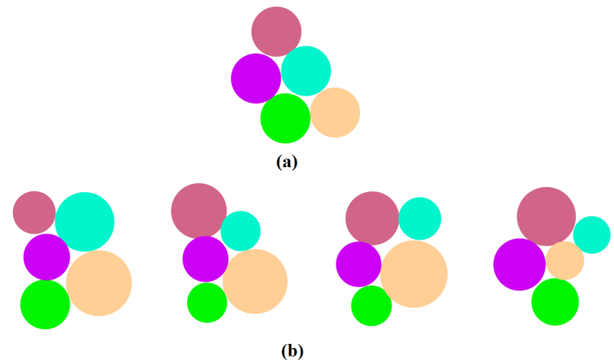

Figure 3: (a) disks are initially of the same radius. (b) disks are changing their radius while remain repelled and gathered.

3.2 Representation and Layout

Each topic has a group of keywords. Also, each topic may have subtopics or belong to a broader theme in health reports. Motivated by these observations, we use a tree-node to depict each topic. The root node denotes the whole document and has a array of child nodes corresponding to a group of topics in the document. A leaf node represents a key word and has no child nodes. The root node is not visualized. Each non-root node is visualized as a circle with its weight encoded as the radius. To keep a group of circles together without overlapping, we compute the disk centers iteratively depending on the following geometric relations:

- **Repel**: For each pair of disks (e.g. D_1 and D_2) within a group, if the distance between their centers is smaller than the sum of their radii, we compute a scaled vector V from D_1's center to D_2's center and translate D_2 by V and D_1 by $-V$. This would prevent overlapping among a group of disks.

- **Sink**: we translate each disk D towards the center of D's parent disk by a scaled amount. This would keep the group of disks together.

We apply the "sink" and "repel" steps at each frame update until convergence. The scaling factor in "repel" is larger than that in "sink" as overlapping is less desirable than being off-centered.

Physics-based vs. Physics-inspired

An important difference of our approach from a force-directed layout is that we directly manipulate on the positions instead of velocities. In an typical physically-based approach, each frame update requires time-integration[1]: ① computing forces based on geometric relations, ② updating velocities w.r.t. forces, ③ updating positions w.r.t. velocities. We found that in general, directly modifying the positions converges faster and hence produces more visually stable layout with less oscillations. Hence, we make a distinction from alternative physically-based approaches and refer to ours as *physics-inspired*.

3.3 Changing Radius Over Time

As the topic/keyword weight may change over time, we allow the circles representing terms to inflate or shrink in an animation mode. Specifically, the animation requires T N-tuple vectors as inputs, where T is the number of time slices. These are weight vectors for all N terms over time.

[1]E.g. explicit Euler method

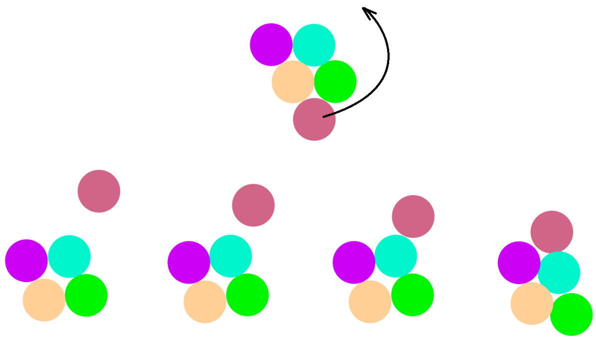

Figure 4: Drag a disk around and release it: the disk join the group in a different spot.

An issue is that the number of time slices may not match the number of animation frames. Usually, even an short animation of 5 seconds (with a framerate of 30 fps) requires far more frames than slices available. Therefore, we use closed, cubic interpolation to fill the gap and produce a smooth, periodic animation.

4 Interactions

ReportViz aims to provide engaging experiences for topic exploration. Currently, it allows a viewer to:

Drag and release: The viewer can interfere with the layout by dragging a disk around and releasing it. Then the disk joins the group in a different spot driven by the physics-inspired algorithm, as shown in Fig. 4.

Expand and collapse: The viewer can select a tree node to expand or collapse it. Expanding a topic node spawns a group of key term nodes, as shown in Fig. 1.

Also the user can switch to the animation mode to play a synthesized topic fluctuation where disks are changing their radius periodically while they remain repelled and gathered. (Bias can be adjusted in a default range for faster or slower convergence of the layout).

5 Feedbacks

Preliminary experiments with ReportViz suggest it as a utility providing an overview of topics and key words in public health. We summarize feedbacks from users of ReportViz as follows:

Topic complexity: As shown in Fig. 5, a topic could split into subtopics, or multiple topics could be nested in a broader theme. For example, in Health 2011 [10], "Expenditure" and "Programs" are bundled as "Health Care Cost", while they remain separated in Health 2008 [7], 2009 [8] and 2010 [9]; In particular, the topic "Expenditure" is splitted into "National Expenditure" and "State Expenditure" in Health 2008 [7].

Word choices: words represent the same concept are used across different topics. For example, "cancer" is a key word in the topic "Measure of Health" while "maglignant neoplasm" is used in the topic "Mortality", as shown in Fig. 6.

6 Conclusion

In this paper we present ReportViz, a utility that integrates text mining and topic visualization. We propose to use the tree data structure for topic merge and split, physics-inspired algorithms for the layout, and animation for incorperating time-varying features. In the future, we would like to build ReportViz as a full-fledged navigation tool for health documents. Specifically, we would like to encode more attributes, such as word-to-topic specificity and topic-to-topic correlation into the visualization. It would also be interesting to design use cases to evaluate the effectiveness of ReportVis for understanding topic complexity and word choices in public health.

Figure 6: different word choices in the two topics "Measure of Health" and "Mortality".

Acknowledgements

I wish to thank Professor Jarek Rossignac, Jacob Esenstein, PhD students Zhuhong (Jonah) Chen, Zhen Chen, and Yangfeng Ji for their discussions.

References

[1] D. M. Blei. Variational inference. Technical report, Princeton University, 2011.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2581–2590, 2011.

[5] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2412–2421, 2011.

[6] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol.*, 3(2):25:1–25:28, Feb. 2012.

[7] National Center for Health Statistics. *Health, United States, 2008: With Special Feature on the Health of Young Adults*, 2009.

[8] National Center for Health Statistics. *Health, United States, 2009: With Special Feature on Medical Technology*, 2010.

[9] National Center for Health Statistics. *Health, United States, 2010: With Special Feature on Death and Dying*, 2011.

[10] National Center for Health Statistics. *Health, United States, 2011: With Special Feature on Socioeconomic Status and Health*, 2012.

[11] P. Resnik and E. Hardisty. Gibbs sampling for the uninitiated. Technical report, University of Maryland, 2010.

[12] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, Nov. 2007.

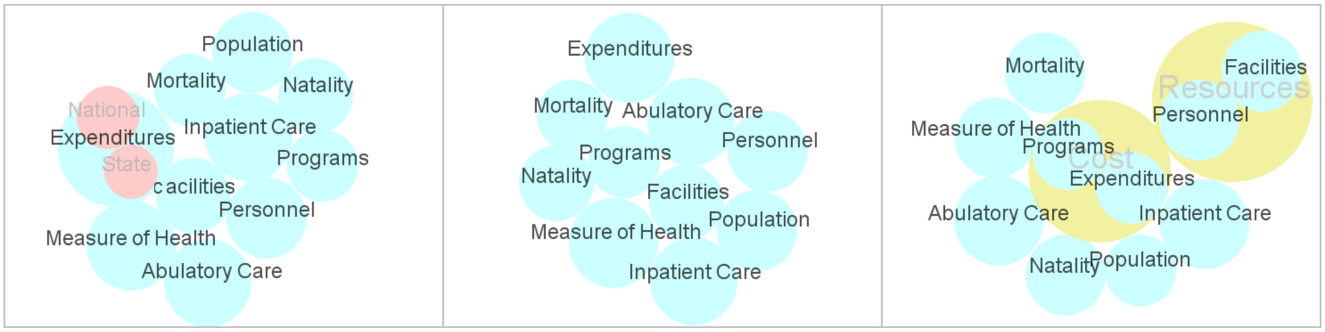

Figure 5: The topic “Expenditure” is splitted into “National” and “State” on the left. On the right, “Expenditure” and “Programs” are bundled as one theme.

TOPICS	LISTS OF KEY TERMS
Population	level, **age**, **sex**, **race**, **resident**, poverty, Hispanic, **income**, **origin**
Fertility and Natality	**weight**, **height**, childbearing, prenatal, birth, breastfeeding, **pregnancy**, **abortion**, **contracept**, marital
Mortality	**cause**, **injuries**, homicide, suicide, **cardiovascular**, **malignant neoplasm**, trachea, bronchus, breast, HIV, drug, expectancy, fetal, **diseases**, **accidental death**
Measure of Health	**occupation**, illness, industry, condition, geography, survival, heart, stroke, diabetes, headache, pain, joint, activity, cancer, limitation, vision, hearing, assess, disability, **urbanization**, distress, **smoking**, education, **drinking**, hypertension, cholesterol, **nutrient**, leisure, obesity, dental
Ambulatory Care	**prescription**, urbanization, access, visit, clinic, influenza, **vaccination**, coverage, pneumococcal, dental, mammography, pap smears, procedures, **emergency**, **X-ray**, physician, hospital, outpatient, primary, dietary, supplement, **blood tests**, **rehabilitation**
Inpatient Care	hospital, **admissions**, discharges, nonfederal, short-stay, **diagnosis**, **length of stay**, procedure, **surgery**, **specialty**
Personnel	**physician**, **patient**, **doctors**, medicine, primary, **specialty**, dentists, **employment**, wages, enrollment, graduates, schools
Facilities	**hospitals**, **beds**, occupancy, ownership, organization, treatment, community, nursing, homes, medicare, certified, providers, suppliers, **MRI (Magnetic resonance imaging)**, **CT (computed tomography)**
Expenditures	**GDP (gross domestic product)**, national, **CPI (Consumer Price Index)**, growth, services, annual, **expenses**, payment, out-of-pocket, **insurance**
Coverage and Programs	Insurance, private, **medicaid**, **medicare**, enrollees, FFS (fee-for-service), **beneficiaries**, eligibility, veterans, state, poverty, fiscal, **reimbursement**

Table 1: We show in the left column a list of topics addressed in health reports. On the right column we show lists of key words associated with each topics. The terms shown in black are summarized from text. The words highlighted in bold are selected by medical students as terms with higher specifity. Words shown in red are considered related, but not reported from text analysis.