ThemeRiver: In Search of Trend and Relationships among Field

Jingchao Su
School of Electronic Infomation and Electical Engineering
Shanghai Jiaotong University
Shanghai, China 200240

***Abstract*—ThemeRiver is a prototype system that visualizes thematic variations over time across a set of data. The river changes its width by flowing through time, depicting the strength of the theme during a period of time. ThemeRiver is also represented as StreamGraph or colored currents, since different themes are represented by different color. As a result, the variations of different theme is visualized within a graph, providing with comparations and trend of different themes. In this project, ThemeRiver is applied in the acadamic search engine AceMap in order to manifest the trend and relationship of fields. Besides, we include 8 kinds of graphs in a dynamic graph that runs in a smooth way, in which more information could be explored**

1. Introduction

ThemeRiver is designed to facilitate the identification of trends, and unexpected occurrence of themes or topics. In particular, ThemeRiver provides users a macro-view of thematic changes in a corpus of documents over a serial dimension. In our prototype, we use time as the serial dimension. We extract the number of academic papers of each subfield in a certain field and apply the data to ThemeRiver. In this way, We provide information of a field through a timeline. Users interact with the visualization to explore the information and to discover trends, patterns, and other features of interest. Apart from ThemeRiver, there are also other more common forms to visualize the data. For instance, single lines changing with timeline demonstrates the variation of a single field. Bar chart is used to demonstrate the basic quantity of papers in different fields. Donut chart is inclined to indicate the percent of each fields. Stacked bar chart acts like themeriver in someway but the data is not continuous along the timeline. For the sake of the diversity of the visualization, several forms of graph (including what I have introduced above) are combined and made into a dynamic graph. Each kind of graph emerge for about 2 second and transit to the next one with smooth transition.

2. Design

2.1. Considerations in Design

The design considerations for stacked graphs fall into two categories. First, as with any information graphic, legibility of the data is critical. Second, , aesthetics seem to play an important role in the popularity of this type of graphic.

2.2. Data preparation

The first step is to extract data from the database. In AceMaps database, the fields are hierarchically distributed. In our demonstration, We collect the data of the number of each subfield(L1) in field Computer Science(L0) in the year from 1980 to 2014 each, the database returns to 34 subfields in which contains 35 entries containing year from 1980 to 2014. We choose this time period because the papers before 1980 are few and display little variation and trend.

2.3. Graph Design

Firstly we assemble the data by generating 35 stacks indicating the year from 1980 to 2014. Then

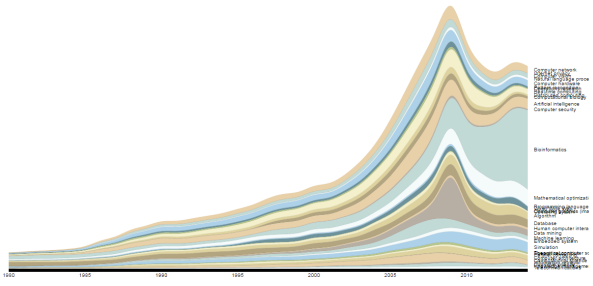

Figure 1. A stacked graph with baseline $g_0 = 0$

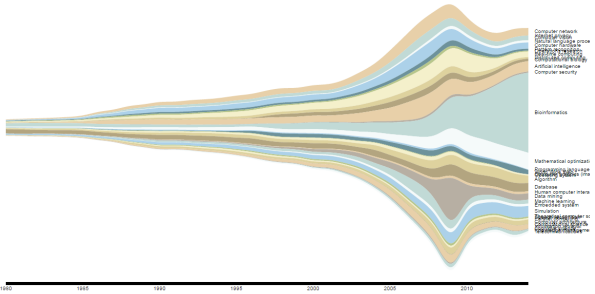

Figure 2. A themeriver with "wiggle"

we place the data of each subfield into all the stack. In this way, the information in each subfield contains x, y_0, y, where x indicates the year, y_0 indicates the baseline, y indicates the ending line. This data arrangement state is manipulated by the d3.layout.stack function in D3.js library. After the preparation of the data, we create the area between y0 and y using distinguished color for each subfield. Besides, we attach the label indicating the name of the field. We call this kind of graph "stacked graph",making a distinction between ThemeRiver. The example is shown in Fig 1.

In this layout, the size of each subfield is easy to read, and each field acts as a current in the river. However, the baseline $g_0 = 0$ makes it not really a river.

Another alternative layout is also applied in our design. It is a symmetric layout around the x-axis. Achieved by the set of offset "wiggle" in the stack function, the graph looks more like a river. Aside from a certain aesthetic quality, this layout has the effect of minimizing some important quantities. In particular, at each point, the silhouette is as close as possible to the x-axis, and in addition the slopes of the top and bottom of the silhouette are in a sense as small as possible. This directly addresses design issue by making the overall graph much less spiky. The example is shown in Fig 2.

2.4. Data Analysis

The goal of our project is to provide a direct variation between themes through a period of time. ThemeRiver is a great layout that shows the trend of all the subfields in a certain field. In our demostration, we can see from Fig 2 about the Computer Science Field.

For instance , we can conclude that something occurs in 2009 in Computer Science Field that cause the sharp increase of the number of its paper, especially in Data Mining Area. Another arresting fact is the field of Bioinformatics's fast development since 2009 and it takes up a large percentage in Computer Science in recent years. We can also conclude that the fields with wider width such as Computer Network, Bioinformatics, Mathematical optimization, Simulation are studied more compared with other fields.

However, the analysis in our example above are based only on the visualization themeriver graph. Those are direct and superficial analysis without verification. But this graph provide us with some obvious facts among different fields, which lead us to further data mining and reasoning.

3. Other Graph

Besides themeriver, the data can be shown in many other forms of graphs, including lines, areas, stacked bars, dount chart, bar chart etc. Each kind of graph has its own emphasis. In our project, eight kinds of graphs are combined in a dynamic way for the sake of a diversity of overview of the information in the fields.

3.1. Line and Area

In this section, each subfield is firstly represented by a circle of different colors located at the left of the screen. And each circle draws a line

Figure 3. Line Graph Drawing Process

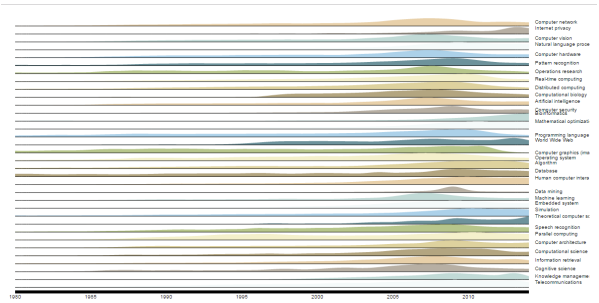

Figure 4. Seperated Area Graph

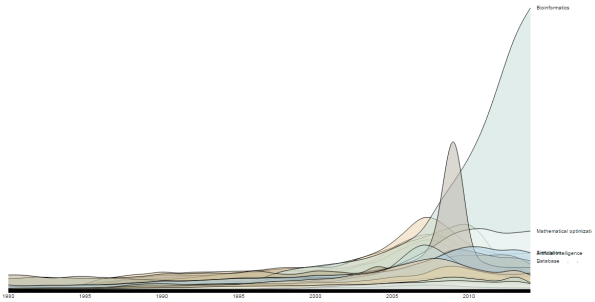

Figure 5. Overlapping Graph

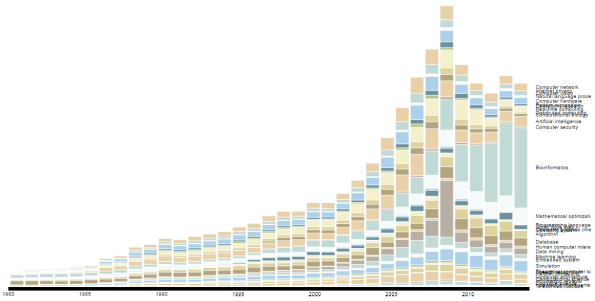

Figure 6. Stacked Bar

smoothly on the timeline to indicate the variation in each field. When the drawing process ends, the lines become areas with corresponding color gradually. The process is shown in Fig 3 and Fig 4.

3.2. ThemeRiver

The seperated areas gradually become conpacted and adjust their width under the same y-axis. It's a stacked graph shown in Fig 1. Afterwards, the stacked graph changes its baseline to make a symmetric layout around the x-axis. This is what we have introduced above–ThemeRiver, shown in Fig 2.

3.3. Overlapping Graph

Then each current of the ThemeRiver becomes seperated under the same y-axis, transiting into an overlapping graph, which is shown in Fig 5.

3.4. Stacked Bar Graph

After the overlappping graph comes the Stacked Bar, which appears by the transition from the Grouped Bar. The Stacked Bar has the same data structure as themeriver, but the data along x-axis is discreted without interpolation, making it looks like a bar. The graph is shown in Fig 6.

3.5. Traditional Histogram

All kinds of graphs introduced above are graphs with timeline as x-axis. However, we then assemble each color of every year together from Stacked Bar, making it a traditional histogram, classified by the name of subfields. The transition process is like the assembling of colored bar. The traditional histogram is shown in Fig 7.

3.6. Donut

In order to present the percentage in all the subfields in a field, donut is used as the last frame

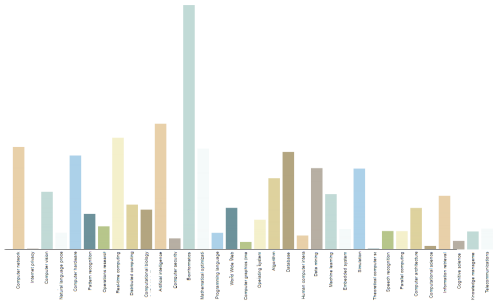

Figure 7. Stacked Bar

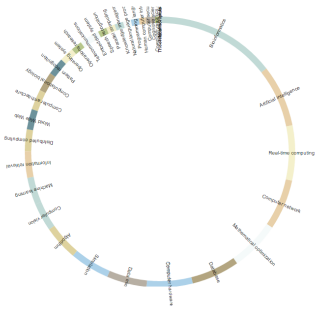

Figure 8. Stacked Bar

in our dynamic graph. The graph is shown in Fig 8.

4. Further Work

There exists a controdiction between the dynamic model and the interaction with human. Since our project is a dynamic graph in which each frame has several second to show, it is hard to add the extra information exhibition when the mouse has some action. There is a choice that the graph be made into static form where people could choose which kind to show.

Some details need optimazing such as the overlapping of the adjacent text label. The order of the subfields is under considerarion in order to make the graph both legible and aesthetic.

5. Conclusion

Through this project, I am able to master tools including SQL and D3 library in Javascript. I have studied the structure of several kinds of graphs and work them out by D3. The data are extracted from the database of AceMap. The innovation among the project is that several kinds of data graphs are attempted in a dynamic form, thus a diversity of visualization are presented for people so that they are probably to get what they desire in just one graph.

References

[1] Susan Havre, Beth Hetzler, and Lucy Nowell, *ThemeRiverTM*: In Search of Trends, Patterns, and Relationships* , Washington USA.

[2] Lee Byron and Martin Wattenberg, *Stacked Graphs Geometry and Aesthetics* .