Report of Project : Topic Analysis in Acemap

5130309757 程子洺

Chapter 1

Brief Introduction

Have you ever gotten lost in the large amount of papers during your research?

As new papers are written year by year, it is increasingly difficult for us to see the complete picture. That problem is what this project try to solve. We are working on generating and analyzing topics from all papers in Acemap.

It should be pointed out that this work is done by a group, four students. I will only introduce the part that is related to me.

Chapter 2

The General Model

First, I will introduce the general model of topics. It is from the paper "*understanding research themes: a probabilistic generative model for citations*".

Suppose each document d cites a subset of other documents. There are two distributions between documents and topics.

One is the doc-topic distribution. It is a probability distribution over topics conditioned on document d, and can be interpreted as the topic coverage in document d when generating citations.

The second distribution is the topic-doc distribution. It gives a reverse conditional distribution of documents given a topic, and can be interpreted as how a topic is characterized by a set of documents that are cited.

The two distributions can be got by LDA or other models. But it is too complex. Luckily, in our database, we already have much information.

Chapter 3

Database and Our Model

In our database, papers have been classified into many topics. And topics are classified into four levels, L0, L1, L2 and L3. For example, psychology, economics and mathematics are L0 topics. Computer hardware and finance are L1 topics. Digital topology is a L2 topic and data transmission is a L3 topic.

In the database, there are total 19 L0 topics, 290 L1 topics, 1500 L2 topics and almost 50000 L3 topics.

Since topics are already generated in database, it is really a relief to our work. We will use these topics to do the analysis.

So, we can construct our model from the general one.

We define the first distribution as, given a document, according to topics which its citations belong to, we can get a probability distribution over topics. The second one is, given a topic, we can get the whole documents that belong to it. So we can get the information of these documents, such as paper year, paper ID, paper rank, keywords and so on.

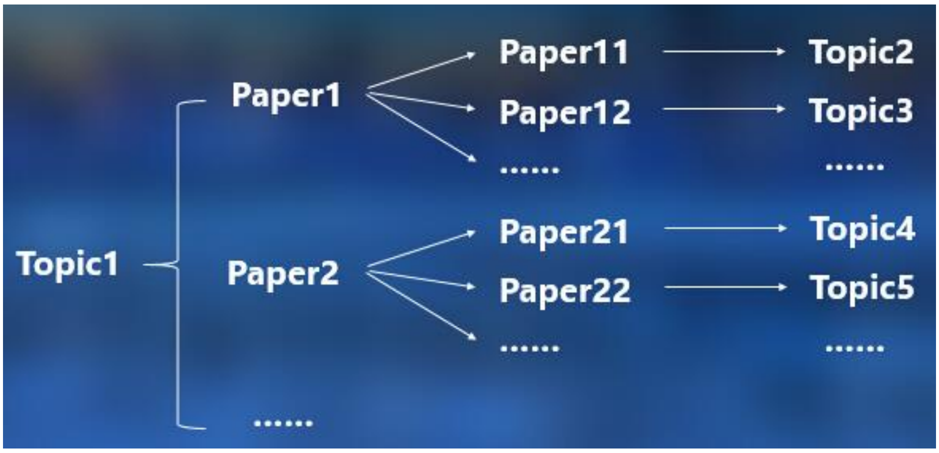
Chapter 4

Topic Analysis and Visualization

Next, using the first distributions, we can do topic analysis. In my work, I mainly analyze the relation between topics. Discovering the relation between topics provides a good guidance for users when searching for relevant topics. Also, understanding the relation through time can give us a whole picture about topic evolution.

So, how do we calculate the relation between two topics? The answer is by citations.

The algorithm is simple. Let’s see a simple graph.



For example, we have a topic1. In this topic, we have many papers, such as paper1, paper2. And these papers will cite other papers. For example, paper1 cites paper 11, paper 12. And paper2 cites paper 21, paper 22. Then, in our database, we can find which topics these papers belong to. So, we get the relation between topic1 and topic 2, 3, and so on.

Then, by some calculation and coding, we can get all the relation between topics. Then, we need to do data visualization. I use Gephi and ECharts to do this work.

Gephi is an interactive visualization and exploration platform for all kinds of graphs and networks. We can use it to plot a network graph or to get a gexf file. ECharts is a chart library based on Javascript. We can use it as a visualization tool to make a website to show our results.

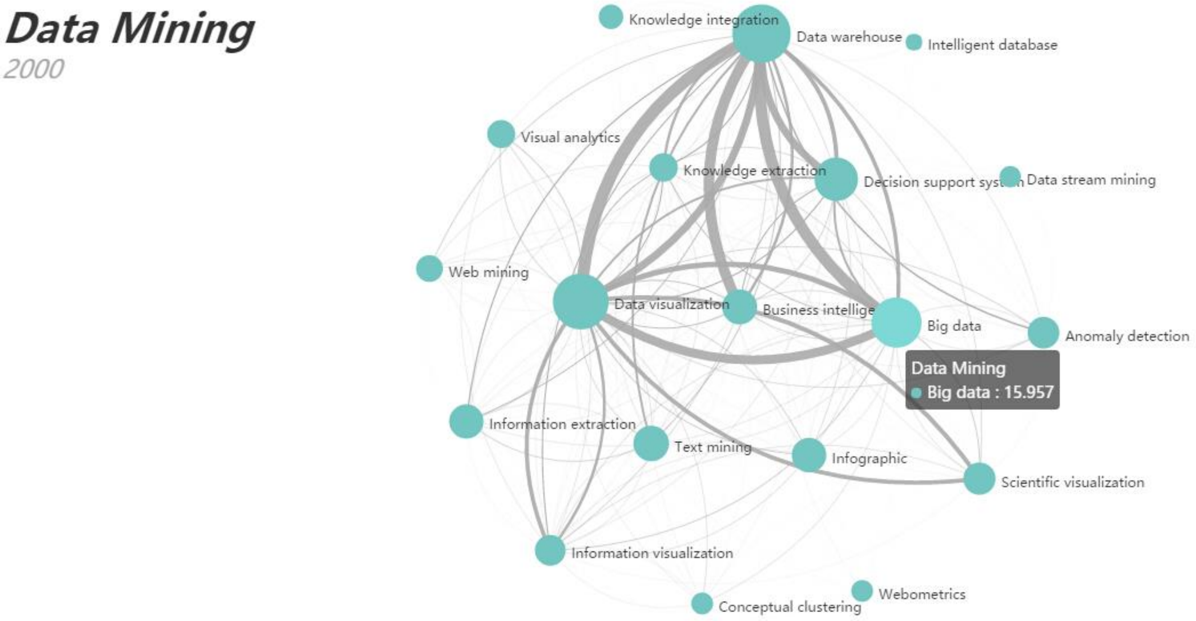
Chaper 5

Demo

Next, I will show a demo of our work. We choose a level 1 topic, data mining. There are many level 2 topics that belong to it, such as data visualization, concept mining, big data and so on. This demo is about the relation between these level 2 topics.

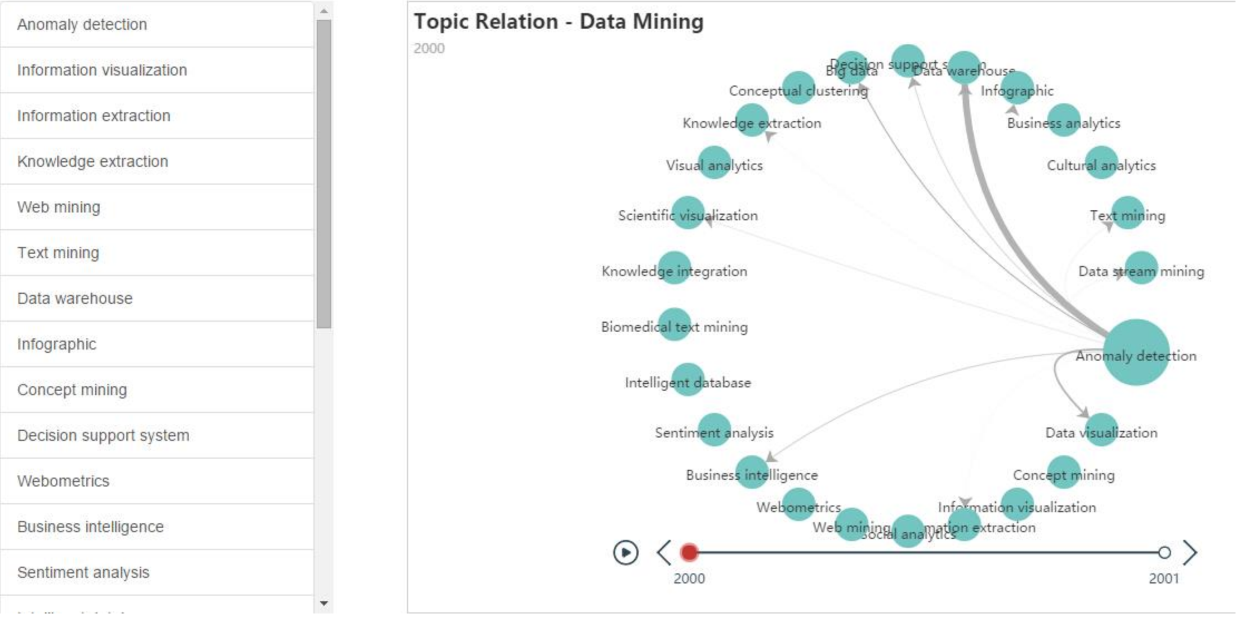
There are three websites to show.

The first one is about the relation between all L2 topics under data mining in 2000.



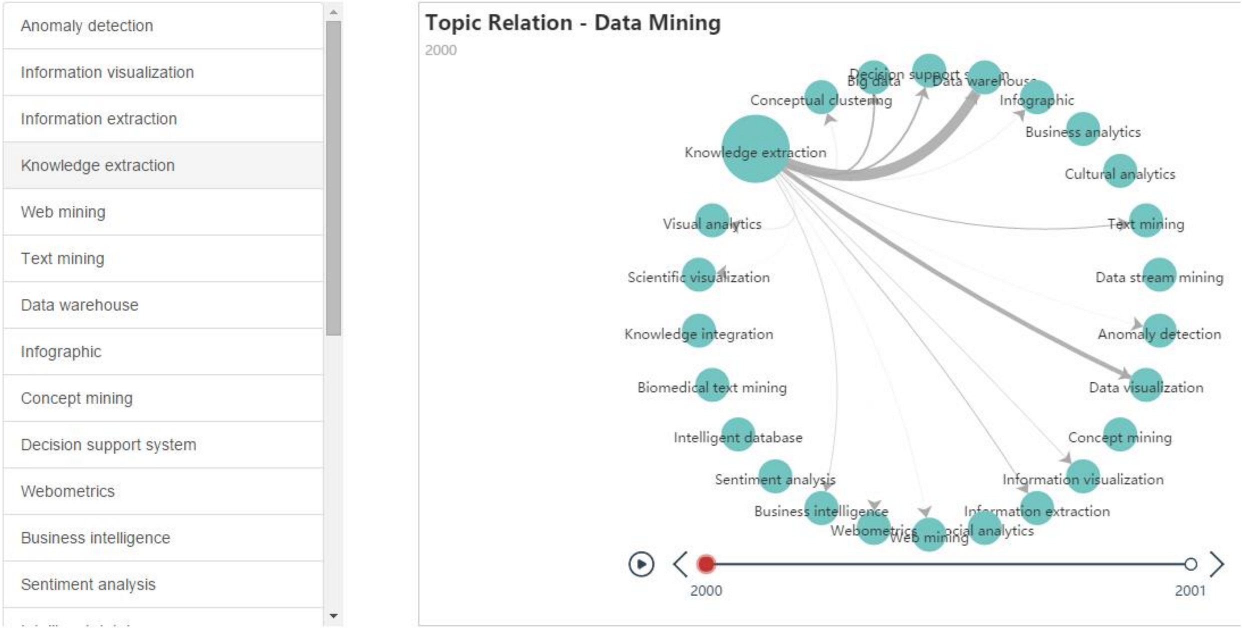
Each points represents a topic. The size of it represents the strength of this topic. Each edge shows the relation between two topics. The thickness of it shows how relevant the two topics are. Thicker the edge is, more relevant the two topics are.

The second graph shows the relation between one topics and other topics in 2000.

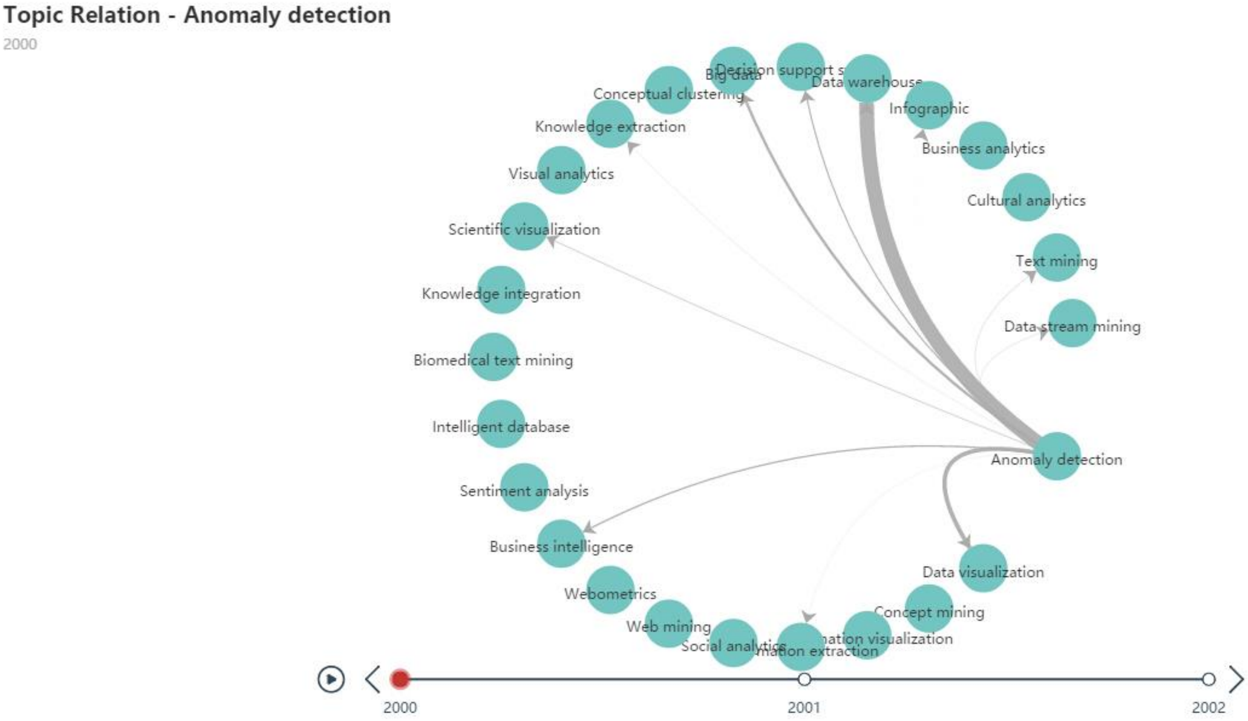


We can see many buttons on the left. When click each one, we can see the relation between this topic and other topics. In this graph, the size of each edge has no meaning, just to distinguish the specific topic. And the edge has the same meaning.

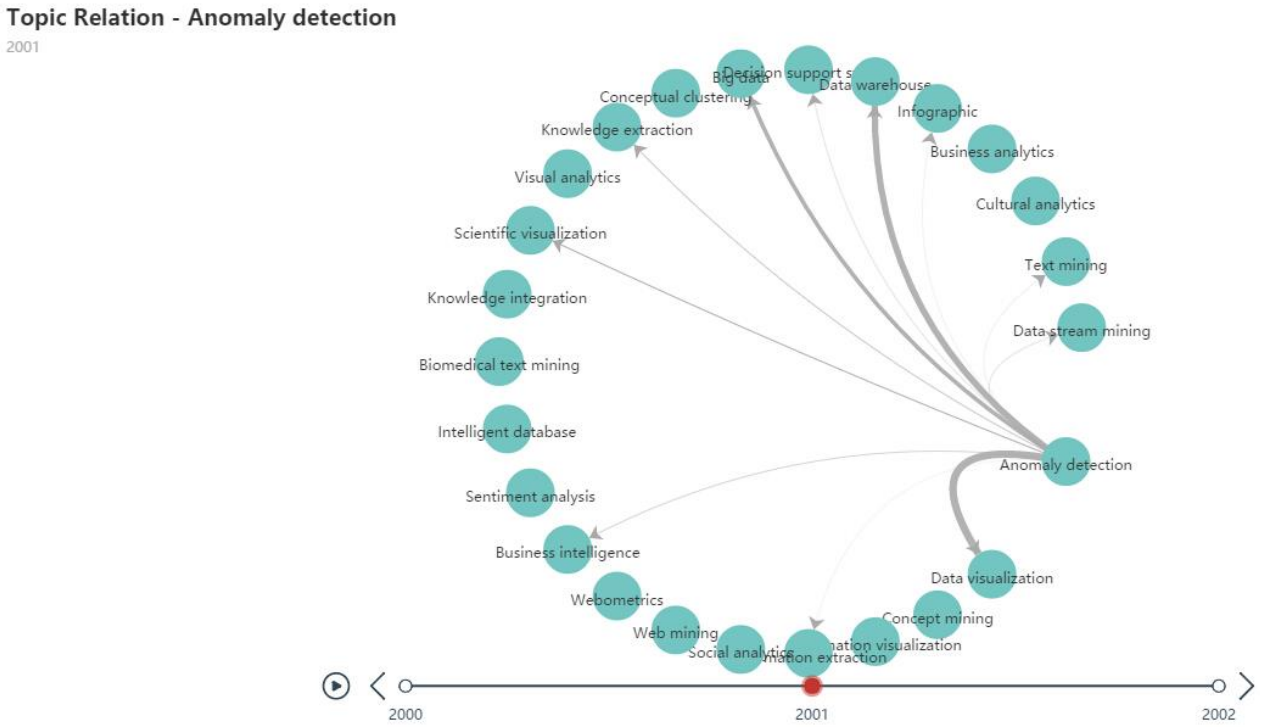
We can see another example.



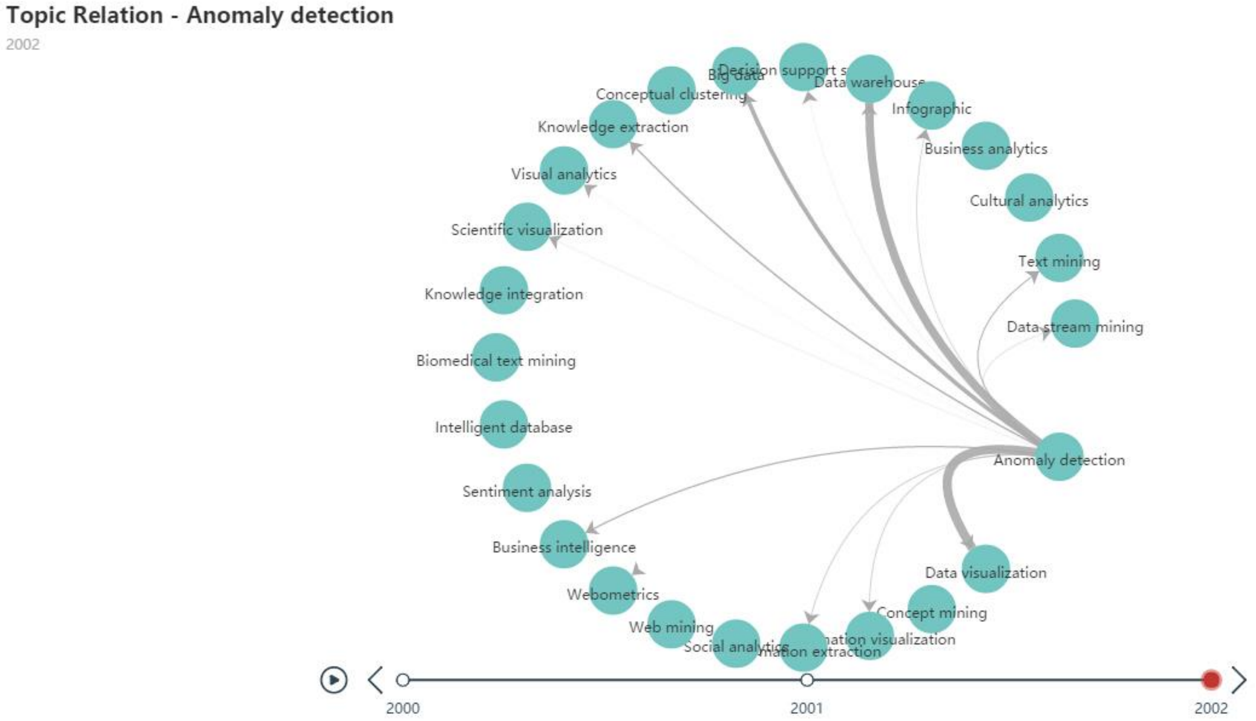
The third graph is about one topic called anomaly detection.



There is a timeline under the graph. It shows how the relation between topics changes with time. In 2001, it changes like that.



And in 2002, it becomes like that.



Chapter 6

Summary

We start to do this project when this semester begins. At first, we discuss together about what we are going to do and design the algorithms. This project includes two topic analysis —— topic strength and topic relation. Then, two students write the codes and get the resulting data. One student and I do the data visualization.

During the project, I learn about how to operate data in database and how to use tools to do data visualization. Also, I learn more about the academic search engine Acemap. This experience will benefit me a lot.