Healthcare KOL Mapping through Data Mining in PubMed

-- AN APPLICATION OF R

Yuchen Wang
School of Finance and Statistics
East China Normal University
Dec 20, 2012

1 Introduction to PubMed

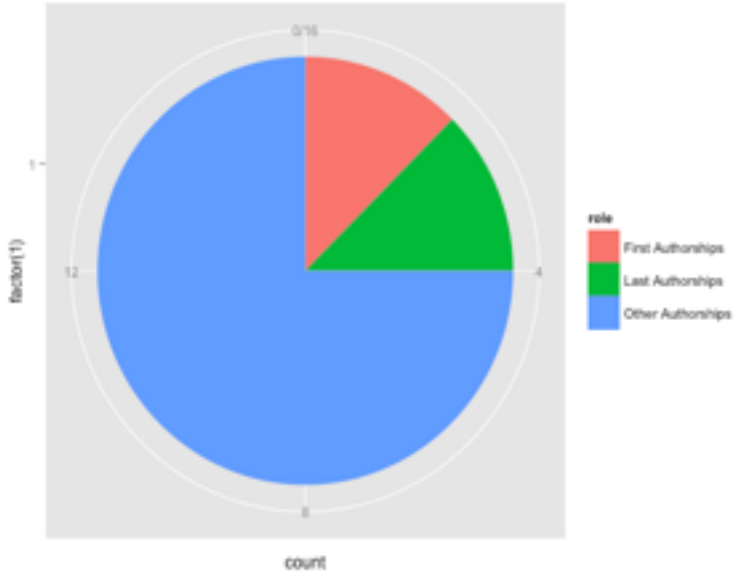
- PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics.
- The United States National Library of Medicine (NLM) at the National Institutes of Health maintains the database as part of the Entrez information retrieval system.

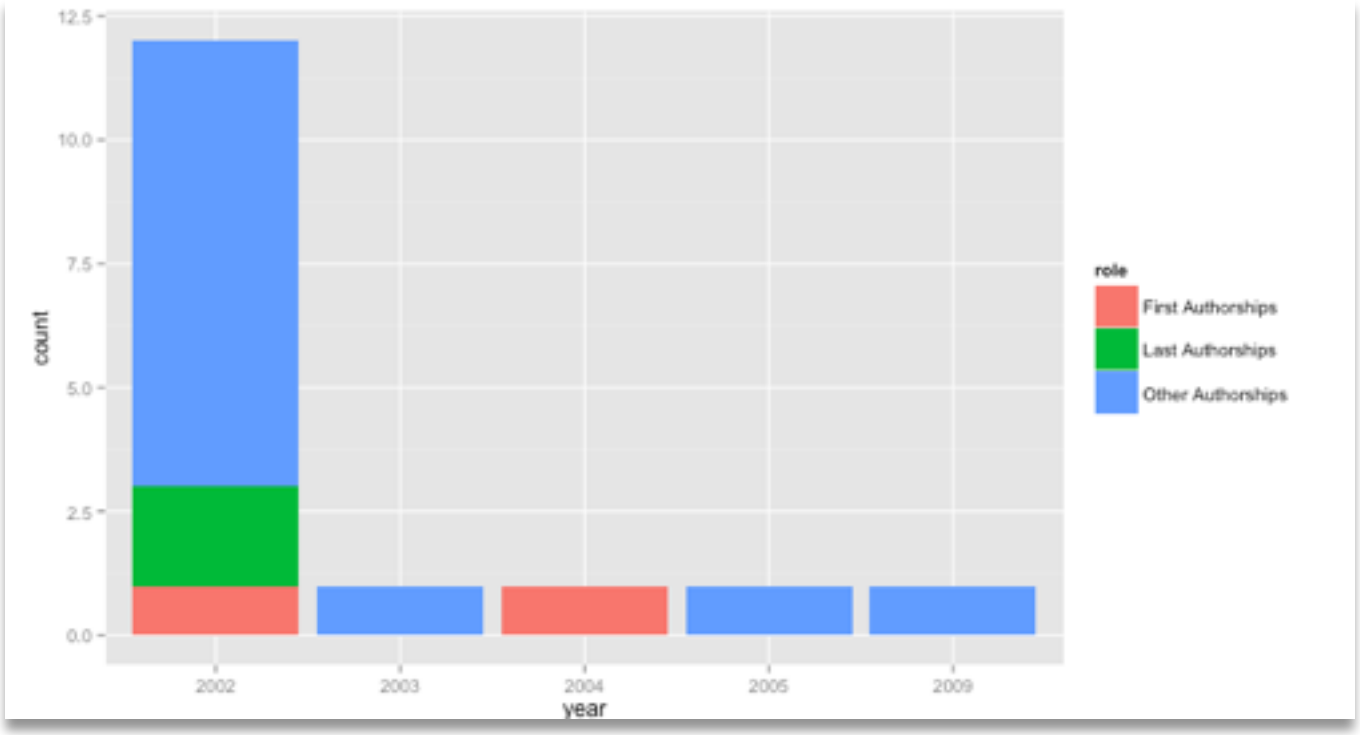
2 Data Mining in PubMed

- Simple searches on PubMed can be carried out by entering key aspects of a subject into PubMed's search window.
- This kind of web-based searches can be carried out in R using the RCurl and XML packages too.
- All information obtained were stored in special R data structures for later data mining. Quantitative and Graphical results of R can be inserted into HTML web pages.
- An automatic work flow can be established with a key aspect (e.g. an author) as input and web pages as output.

2.1 Publication Role Analysis

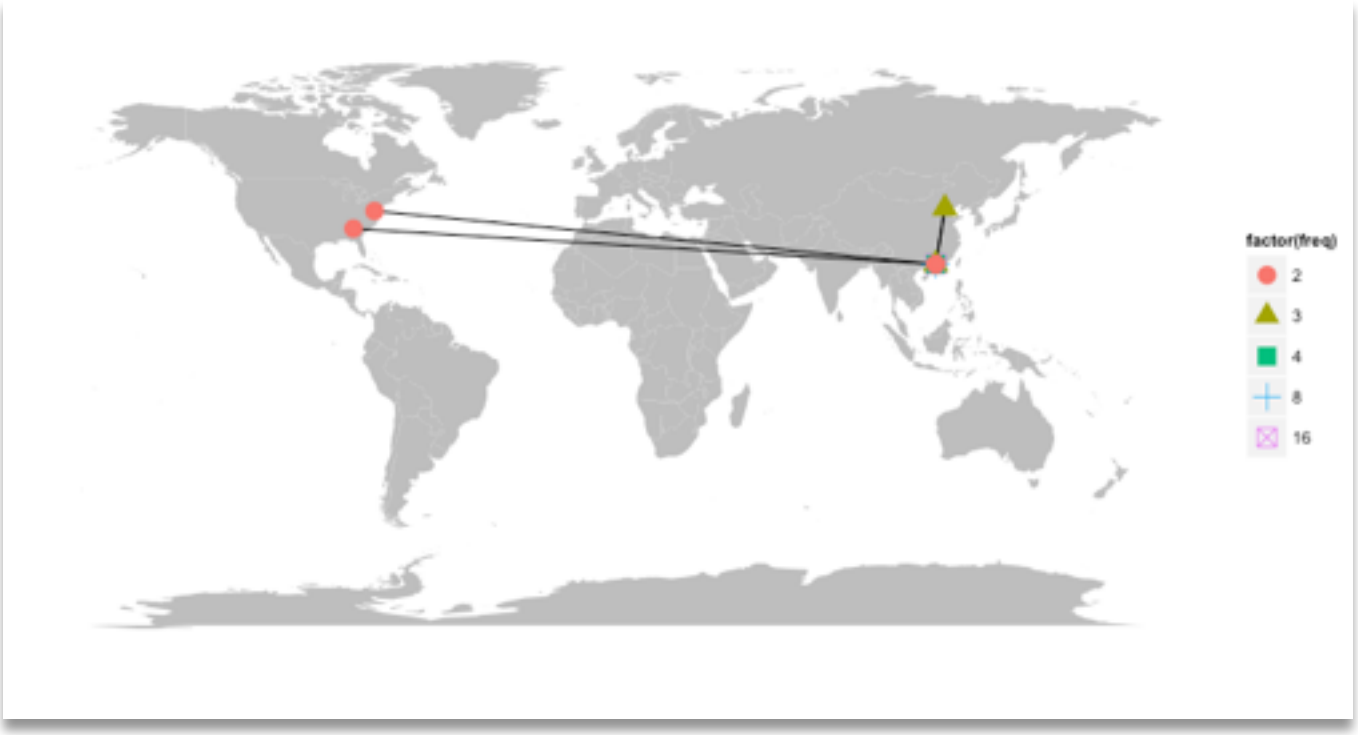
- PubMed search results of an author contain all publications of that author and each co-author's information.
- By analyzing all the publications, one author's publication role can be obtained.





2.2 Publication Role with Year

Integrating publication year into publication role, the change in publication role was visualized.

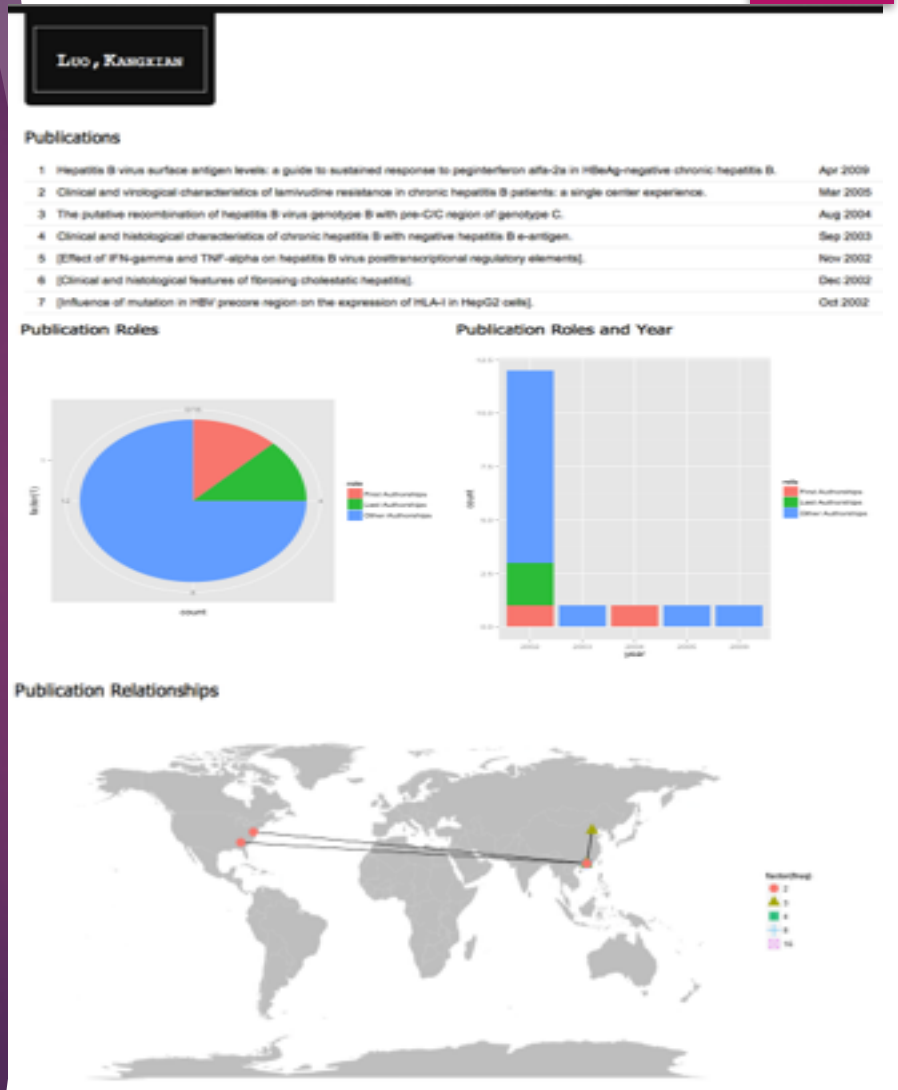


2.3 Publication Relationship

By locating every author and coauthor on a map, the publication relationship can be displayed.

2.4 Author Profile

Taking advantage of R and HTML, a web page of author profile was generated.



3.1 RCurl and XML

- RCurl::getForm() is used to combine query terms (such as author name) with base query URL and download the result page.
- XML::xmlTreeParse() is used to parse the HTML page into a tree structure.
- XML::getNodeSet() can search specific tags within the tree, and XML::xmlValue() can return the data value within the tags.
- Inspecting the HTML page, tag names are always self-explanatory such as Title and Year.

3.2 Geocoding

- With those tools, we can get all article’s data related to an author.
- Address data are related to articles too. It's text-based so it's hard to compare such data to get an unique address for an author.
- Geocoding is Google's service for transforming address data into coordinates. The same as searching data in PubMed, we can search address data in Google Map's database, and get numeric coordinates.
- Then the most frequent coordinates in one's articles is defined as his address.

3.3 knitR

- knitR is an R package for dynamic report generation. It supports different formats including HTML. We can write an RHTML document with HTML for format markups and R code for contents. Here is an example:

```
<div class="title">
  <!--begin.rcode echo=FALSE, comment=''
  load("./luokangxian.RData")
  cat(author)
  end.rcode-->
</div>
```

- knitR will evaluate R code within the gray chunk and return it's output in an HTML page, in which the HTML tags will markup the output as page title.

Thank You

- See page demo at http://wangyuchen.github.com/demo
- Contact me by email:

  ycwang0712@gmail.com