南京医科大学 康复医学院 林枫 peterduuslin@gmail.com
2015 年 11 月 18 日版

共现关系分析器用户指南

引用方式：
林枫, 刘云,江钟立. (2012) 汉字网络的历时性模式探析. 复杂系统与复杂性科学, 9(3): 50-61.
English Citation:
Feng L., Yun L., Zhongli J. An Exploration for Diachronic Patterns in Chinese Character Networks[J]. (in Chinese) Complex Systems and Complex Science, 2012;9(3):50-61.

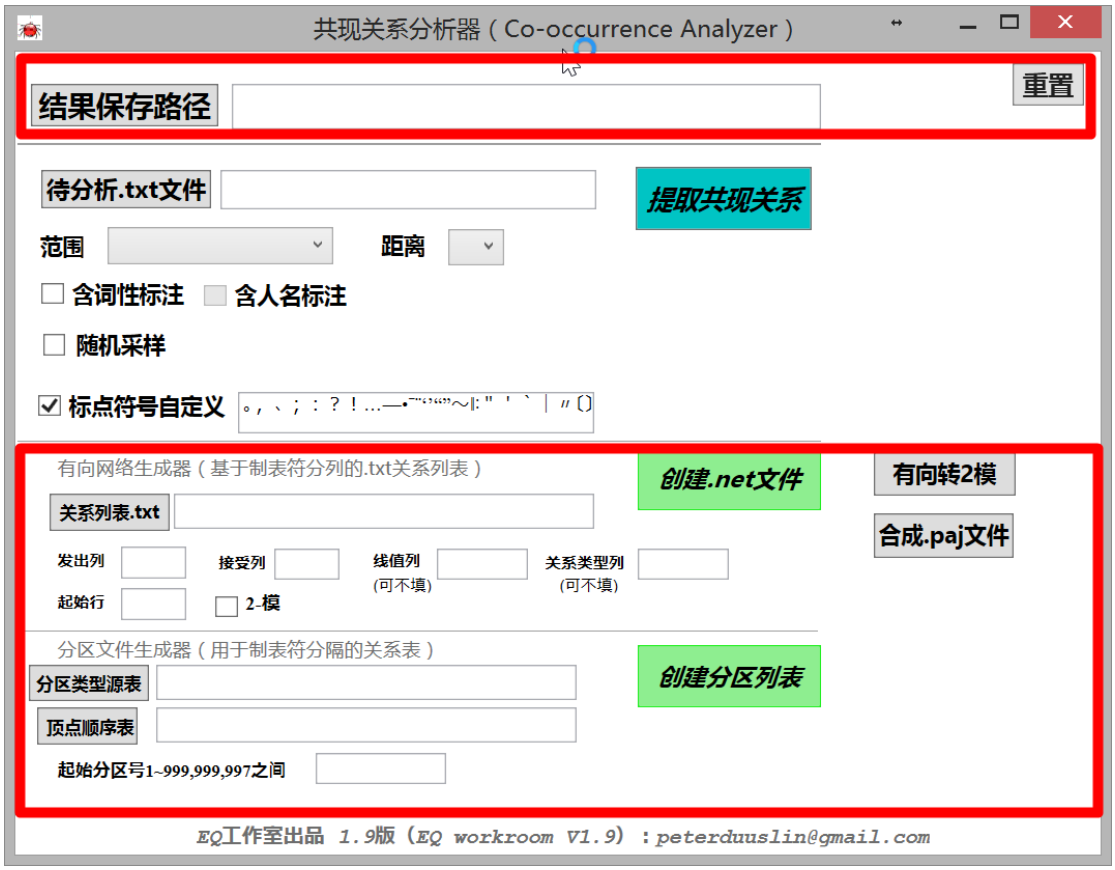
如果有任何疑问，可以按照页眉邮箱发邮件联系我，不过不能确保总是能立即有回信。如果非常非常紧张，请连续发送两封以上邮件，估计可能会收到回信(*^__^*) 。

目录

1、最简帮助：

在每个框栏和按钮上，如果光标停留，就会显示该条目的简短说明。

2、软件的主要结构和用途



软件分为三个操作区域：

1) 结果保存路径区，生成结果保存文件夹供保存所有分析结果。

2) 共现关系分析区，主要以人民日报语料库作为分析对象进行设计，可以生成词内字共现，句内词共现和段内人名共现三种网络文件。

3）网络及分区文件制备区，主要可以实现以下多种功能：

- 生成 Pajek 格式的.net 网络文件：包括有向和无向，非加权和加权，单关系类型和多关系类型，1 模和 2 模
- 生成 Pajek 格式的.clu 分区文件，创建分区列表。
- 生成 Pajek 格式的.paj 工程文件，其中可以包含巨量网络文件，其作用是可以将 Pajek 所识别的网络、分区、矢量、分层等多种文件整合

为工程文件，批量导入 Pajek 中进行分析。

该软件之所以称为**共现关系分析器**，主要是因为网络分析的对象就是关系，而在网络文件制备过程中，关系就体现为两个顶点标签共同出现在同一行上。例如：AAA 与 ZZZ 出现在同一行上，表示有 AAA 与 BBB 之间有一条连线。从 AAA 发出到 ZZZ 的连线的权重（也可以称为强度或连线值）为 1.9。BBB 发出到 UUU 的连线强度为 0.4。当然，还有一个原因，这个软件最早设计的时候就是用来做语言材料中的字词共现关系的。

AAA	ZZZ	1.9
AAA	YYY	2.1
AAA	XXX	1.3
BBB	UUU	0.4
BBB	VVV	3.6
CCC	WWW	2.3
CCC	UUU	1.1
CCC	BBB	6.5
UUU	DDD	4.2
WWW	AAA	7.2

3、设计初衷：研究语料共现关系（不关心语言研究的可略过）

语言在传递过程中，表现为语符（language symbols）的线性序列形式。无论语符单位是音节、字符、词，还是短语、句子，乃至语篇，它们都必然遵循线性序列的形式。这种线性序列，包埋了施者（sender）想要传递的绝大部分语义线索（除非在交流过程中还包括了手势等空间序列内容），通过在受者（receiver）的语义网络中形成持续递进的语义激活簇（semantic clusters），从而形成受者对语义的理解。从反证法的角度来考虑这个问题，可以发现，一段话语的意思，会在把其中的语符单元（例如词）的序列随机打乱之后消失，而这种操作却无损于语符单元的出现频率（例如词频）。因此，**在语言学研究中，基于词频的传统，并不能抓住这种语义线索。**由此可考虑的是，如何研究语符线性序列所

包埋的结构特征呢？对词法、句法、语篇的结构分析，无论是基于专家内省的例证分析，还是基于语料库的多元统计分析，都往往仍然向频率统计方向回归，而没有从宏观上把握这种结构特征。语符共现关系（co-occurrence）的研究，为此提供了新的思路。

从广义上来说，只需要设置某个框架背景，并把语符共同出现于该框架背景中就视为共现关系。这种情况下，共现可以是在语用角度上对某些交流符号的关系表述。例如，在表达承诺时一边说“放心吧”，一边做出“拍胸脯”的动作，从而在“放心吧（S）”和“拍胸脯（A）”之间形成共现关系（S 表示语音，A 表示动作）。从狭义上来说，共现性往往以上下文关系为基础。例如：在“我-爱-你-一万年”中，“我”和“爱”就构成了 0 阶共现关系（邻接关系），“我”和“你”就构成了 1 阶共现关系（间隔 1 个词“爱”）。这种共现关系，是以语符（例如词）在特定上下文范围（如句子）中共同出现为条件的。当然，在汉语中，也可以是以字为语符单位，进行共现关系构建。例如，《三国志》中有两段表述“天下将乱，非命世之才不能济也”和“天之大运，非君才力所能存也”。图 1 显示了其中汉字之间的指向关系，箭头表现的是从前字指向后字的直接相邻关系。

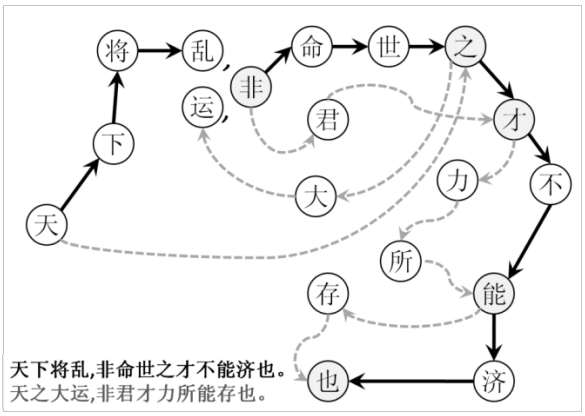


图 1 汉字邻接网络（林枫,等.复杂系统与复杂性科学, 9 (2012) 50-61.）

本分析器是作者在南京师范大学攻读应用语言学博士学位课程时学习陈小荷教授的 C#课程时的课程论文成果，其中采用了陈教授提供的类库。最初用于分析基于上下文语境的语符共现网络构建器。它主要实现以下功能：

1）输入经标注的语料

首先从单语种开始，最初提供简体中文标注语料的处理。

2）设置语符单位的上下文语境范围

首先设置三个层次的上下文语境范围：

- 字在词内共现（characters in a word，词内字共现），
- 词在句内共现（words in a sentence，句内词共现），
- 人名在段内共现（names in a paragraph，段内人名共现）。

3）设置语料共现间隔

对于词内字共现和句内词共现，可设置语符单位之间的间隔。根据目前对语言三度分割的研究结论，暂时设置 0、1、2 和 3 这四阶。对于段内人名共现网络，不考虑语符单位（人名）之间的间隔，只要两个人名在同一段中出现，即认为有共现关系。

4）提供随机抽样功能

对语境范围内拆分出来的语料，可以随机抽样。在词内字共现的分析中，可以随机抽词。在句内词共现分析中，可以随机抽句。在段内人名共现分析中，可以随机抽段。

5）输出网络文件和分区文件。

网络文件以网络分析通用的.net 格式为准。

在词内字共现网络中，由于词性可能影响到字素的共现。例如：“幸福”在做为形容词时，“幸”和“福”是在形容词中共现，“幸福”在做为名词时，“幸”和“福”是在名词中共现。“祝福”是做动词时，“祝”和“福”是在动词中共现，当“祝福”做名词时，“祝”和“福”是在名词中共现。于是在这小范围分析中发现，“福”这个语素在形容词、名词、动词中都有出现。而在整个构词结构中，最终构成的词的词性，是否受到词素的影响，可以通过这种信息来获取。因此，在词内字共现网络中，把词性作为网络内的关系类型输出到网络文件中。

在句内词共现网络，提供配套的分区文件，用于进行词频或词性分类分析。

在段内人名共现网络，今后可能提供当两人同时出现时，第三人也出现的关系。也可提供两人每次共同出现时，同时在段内每次都出现的相同名词的列表，以供进行事件判断用。此功能由于实现较为困难，暂未提供。

设计界面如图 2。随机项如果选中，将显示“种子值”和“百分比”两项。

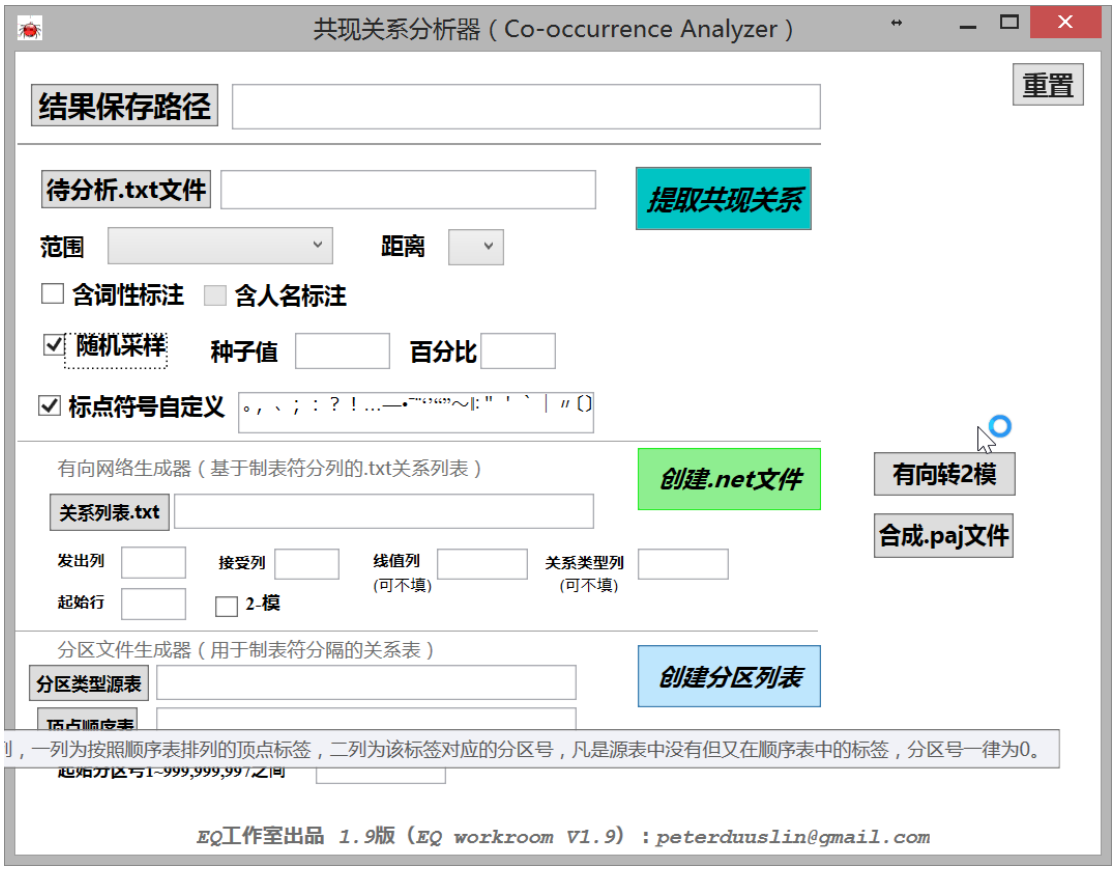


图 2 共现关系分析器界面

4、共现关系分析区的使用说明

4.1 字符的字节问题

代码中提供了四字节处理的类 String4，但实际上用于今后扩展功能用。在目前的版本中，所有字识别都采用正则表达式完成，并不涉及按字节识别汉字的问题。尽可能避免四字节字的影响。

4.2 结果保存问题

为避免结果保存的时候出现重名或覆盖，在不同类型的操作中以条件 Condition 类来传递文件名，而且可以通过点选结果保存路径（对话框打开的时候默认是桌面），可以在该路径下每次都生成一个新的文件夹，而且以精确到秒的时间来命名。例如：E:\lf\desktop\CoOccurResult20151118110517。对话框效果如图 3。值得注意的是，只要用户点击选择保存路径，就会生成新的文件夹，所

以请不要反复点击生成新保存路径。

图 3 “结果保存路径”对话框

4.3 随机抽样问题

目前版本仅提供一次简单的无放回随机抽样，今后可能附加蒙特卡洛模拟和置信区间检验功能（实际上笔者作为临床医生，实在没有精力来添加新的功能，各位用户将就点用吧 O(∩_∩)O）。设置种子值是为了能使抽样的起始条件对齐，并且能够每次都重复抽样结果。请记住自己所设置的种子值，例如 1234、138、10086 之类。

对于无放回随机抽样的方法，并没有采用通常所说的用时间作为种子来生成随机数的方法，这并不能保证在无重复。在此采用的策略是每次产生一个小于总体的随机数作为抽取个体的序号，待个体抽取后，总体相应变小，然后再次抽取，从而保证每次都只在剩余个体中随机抽取。

为了能控制抽样的规模，提供文本框供输入抽样的百分比。百分比用正则表达式限制只能输入两位数字。如果输错，会提示"请输入两位正整数"。

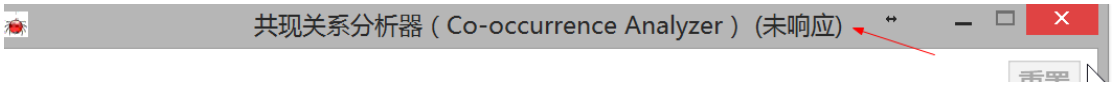
4.4 举例说明

以 98 年 1 月份人民日报标注语料为分析样本。点击待分析.txt 文件。

4.4.1 词内字共现

条件设置为距离为 0（字与字直接相邻）。随机种子值为 123，百分比为 25%，考虑词性标注（即把词性作为网络中的关系类型号）。例如以下格式表示 1 号字与 111 号字共现 7 次，它们是在编号为 1 型的关系类型中共现（NS 在人民日报语料标注中表示地名）。

```
*Arcs :1 "NS"
  1     111 7
```



由于人民日报语料较大，所以当看到未响应的提示时，要耐心等待几分钟。运算耗时 9 分 49 秒。结果界面如图 4（注意，这里的图还是我 2013 年写这个软件时的截图，所以文件夹名字前面不带有 CoOccur，但基本结果未有变化）。

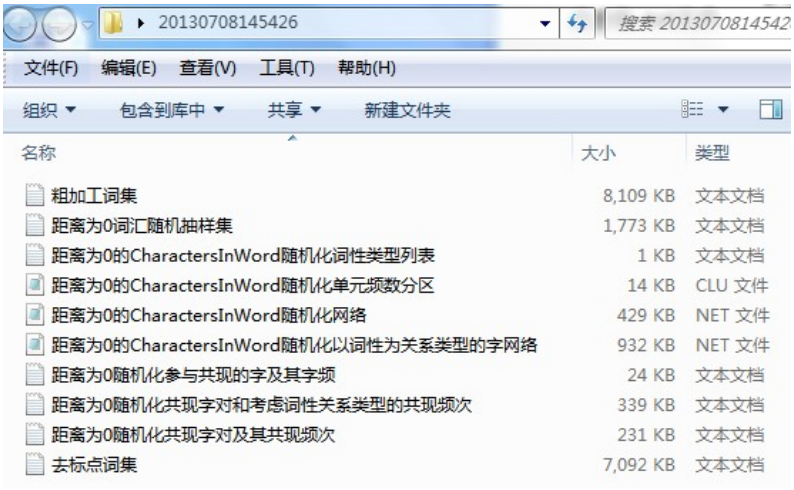


图 4 人民日报 98 年 1 月词内字共现关系分析结果（随机种子 123，抽样百分比 25%）

以 Pajek 打开.net 文件获取网络的一般特性报告，共有 3539 字，30495 条连线。提取编号为 3 的关系类型（N 名词），去除只出现过一次的共现关系，有向关系无向化（字共现只考虑组合关系，暂时不考虑方向性），并删除自环（loop）。利用字频分区“距离为 0 的 CharactersInWord 随机化单元频数分区.clu”，删除在整个抽样库中仅出现一次的字，并删除孤点（无连线的字）。最终获得 1775 个字和 5297 条连线。如图 5。

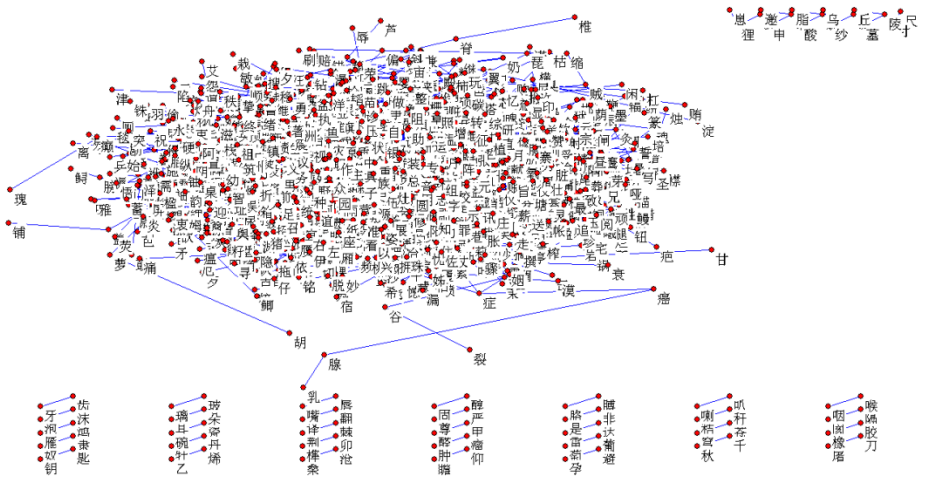


图 5 词中字共现网络

其中可以抽取的 k-核最高为 8 核，含 171 个字。即有 171 个字形成的集合，其中每个字至少可以和其中 8 个其他字组合成词。最高组合力的是“人”，可以和 1775 个字中的 114 个领接组词。

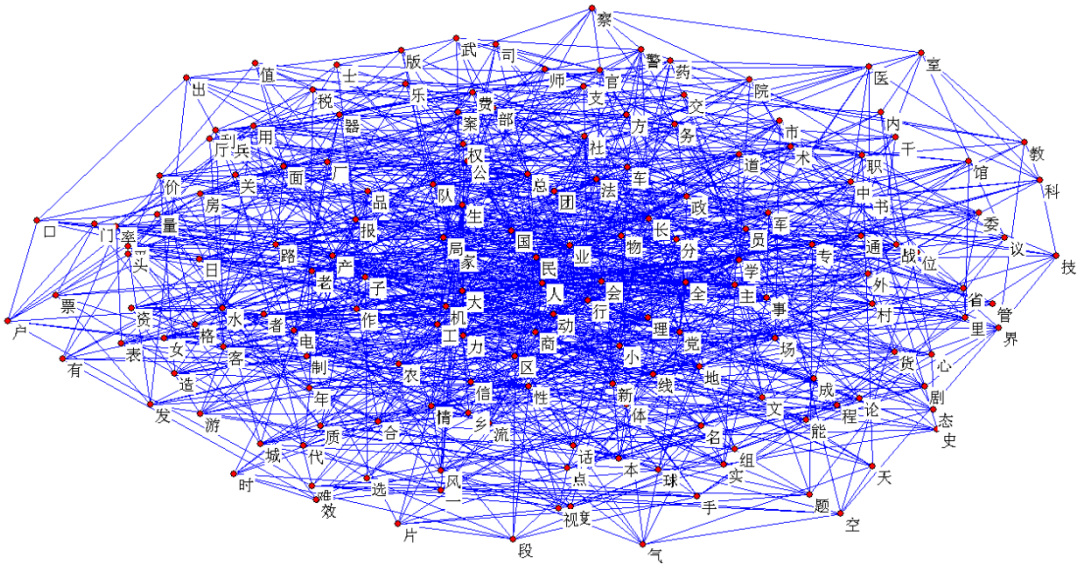


图 6 词中字共现网络的 8 核

4.4.2 句内词共现

条件设置为距离为 0（字与字直接相邻）。随机种子值为 123，百分比为 25%，勾选考虑词性标注（即专门制作一个分区文件.clu 来保存词性）。所谓句，即以句号为单位来计算，句内的其他符号均会在计算中忽略。如果两个词相隔一个逗号，在句内会就认为是两个相邻词。

运行耗时 17.26 秒。文件类表如下图。

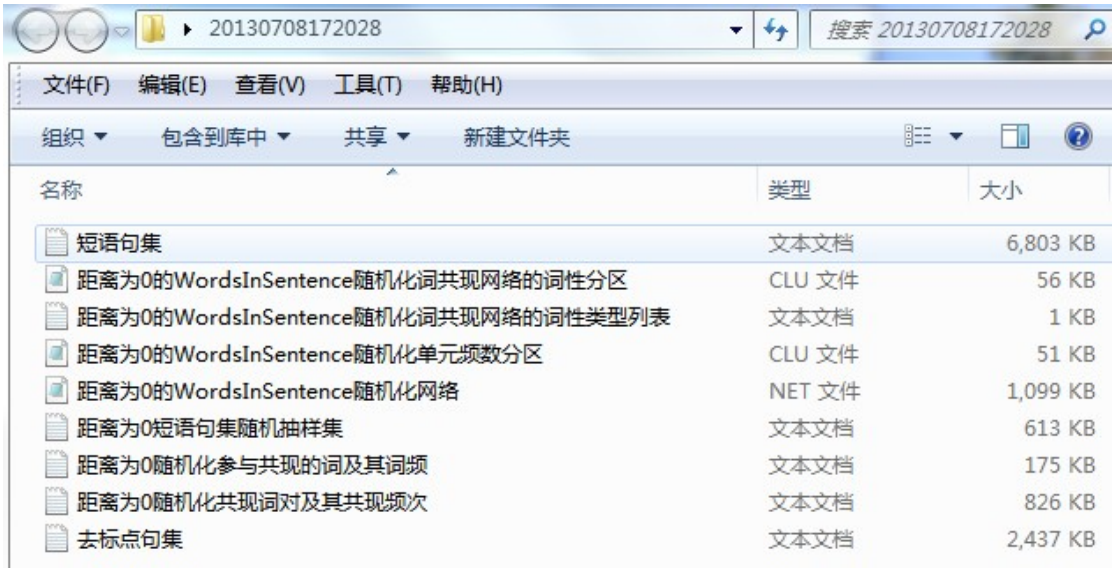


图 7 人民日报 98 年 1 月句内词共现关系分析结果（随机种子 123，抽样百分比 25%）

共有 16436 词，51518 条弧。删除在文本中词频为 1 的词，删除网络中的环，

删除出现频次仅为 1 次的弧，删除孤点，共剩下 5706 个词和 12453 条弧。如果从 16436 词开始，提取名词（分区号 5）和动词（分区号 10），则 8345 词。从中删除仅出现一次的词，剩余 6762 词。然后再删除仅出现 1 次的连线和没有连线的孤点，删除环，最后共剩余 1824 词和 2177 条连线，并且在此基础上通过 Pajek 操作可以得到这 1824 个词的词性分区（名词和动词），其中有 1378 个词构成最大连通子网络。最后以粉红色表示名词，绿色表示动词，顶点大小表示点所发出的弧线数量。如图 8。

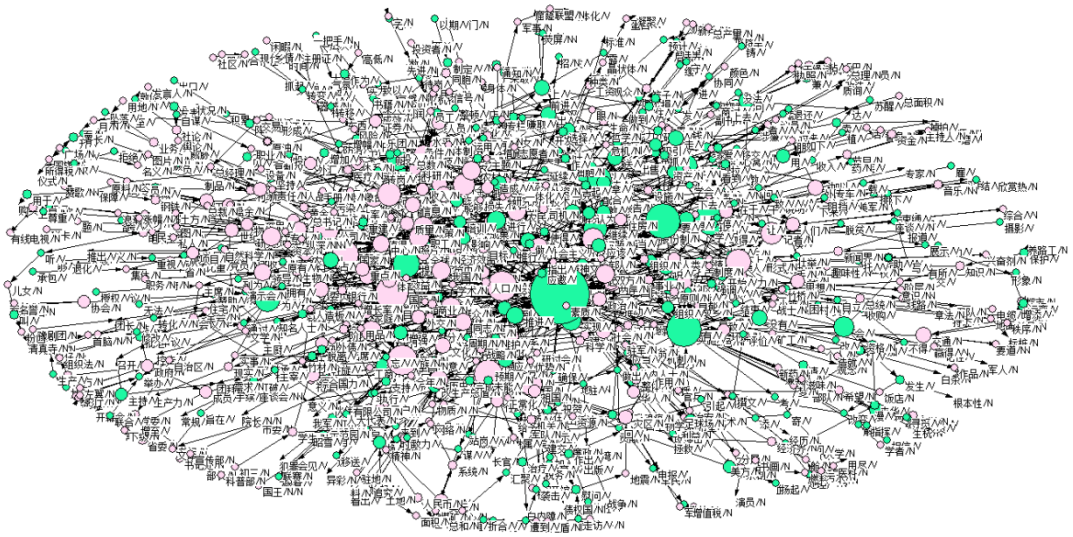


图 8 句内词共现网络（名词粉红色，动词绿色，大小与点所发出的弧线数呈正比）

还可以分析该网络中最紧密的核心成分。此处析出 40 个词，如图 9。可供分析名词动词的搭配整体结构使用。这种整体结构关系，是通常的词频分析无法得到的。

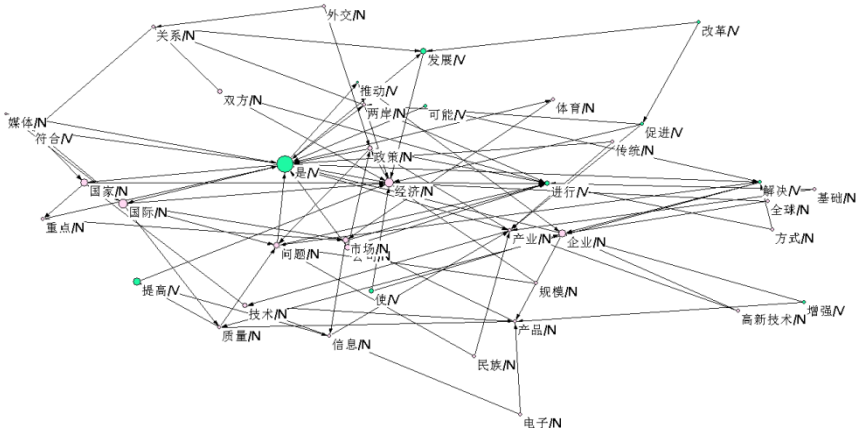


图 9 句内词共现名词和动词 2 核

4.4.3 段内人名共现

人民日报 98 年 1 月语料库。条件设置：随机种子值 123，抽样规模 25%。耗时 0.88 秒。请注意：一定要选中“含词性标注”和“含人名标注”这两项，才会在“范围”框栏内出现段内人名共现的选项。

共有 2044 个不同人名。去除仅出现 1 次的人名。共有 709 人。去除未与其他人名共现的独立人名，去除仅共现 1 次的关系，去除自环，剩余 378 人和 324 条线，如图 10。

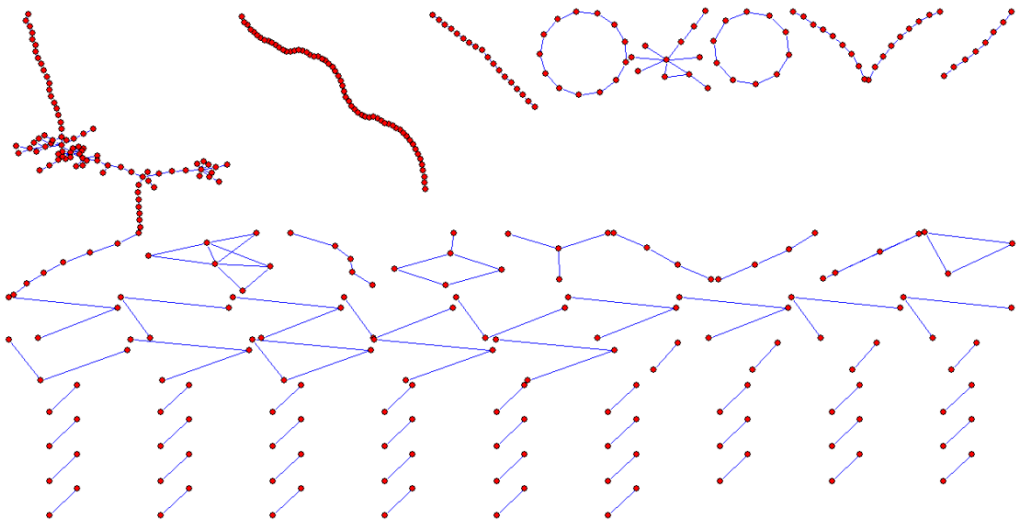


图 10 人民日报 98 年 1 月语料段内人民共现关系网络概览

提取其中最大的连通子网络，共有 83 个人，如图 11。

图 11 人民日报 98 年 1 月语料段内人名共现网络最大连通子网络（点的大小与其所拥有的连

线数量呈正比）

4.5 如何灵活运用

该分析器实际上只提供了文本归类整理和输出网络文件列表的功能。用户需要掌握网络分析的基本逻辑，同时会使用以下功能，才能灵活运用该分析器。

1) .txt 文件的基本操作，包括了解什么事制表符分隔的.txt 文件；
2) EXCELL 的基本操作，包括：如何从 EXCELL 列表导出制表符分隔的.txt 文件、如何从.txt 文件导入列表至 EXCELL。
3) Pajek 的基本文件格式：网络.net 文件、分区.clu 文件、矢量.vec 文件、工程.paj 文件。关于 Pajek 的使用，请参看作者翻译的教程《蜘蛛：社会网络分析技术》一书。关于 Pajek 的基本数据格式，也可参看豆瓣链接：http://www.douban.com/group/topic/79439307/

该分析器在结果文件夹中生成了许多过程文件，其中一些句集和词集是用来存档，而制表符分隔的.txt 文本能够用于导入 EXCELL 生成列表，或者用软件下部的网络生成器和分区生成器进行处理，生成用户自定义的网络和分区文件。

4.6 关于“标点符号自定义”的说明

除了人民日报语料库以外，任何以空格分词的文本都可以进入共现关系分析器处理，但是如果做词性标注，必须以人民日报语料库的标注标准进行标注。之所以设置标点符号自定义，是因为在人民日报语料库中是以/w 来标注标点，如果用户需要分析自制的语料，那么首先需要自己用空格来分词。分完词之后，如果用户不想再做词性标注，就会出现一大段诸如以下样子的文字：

妈妈，我 想 吃 猪头肉。

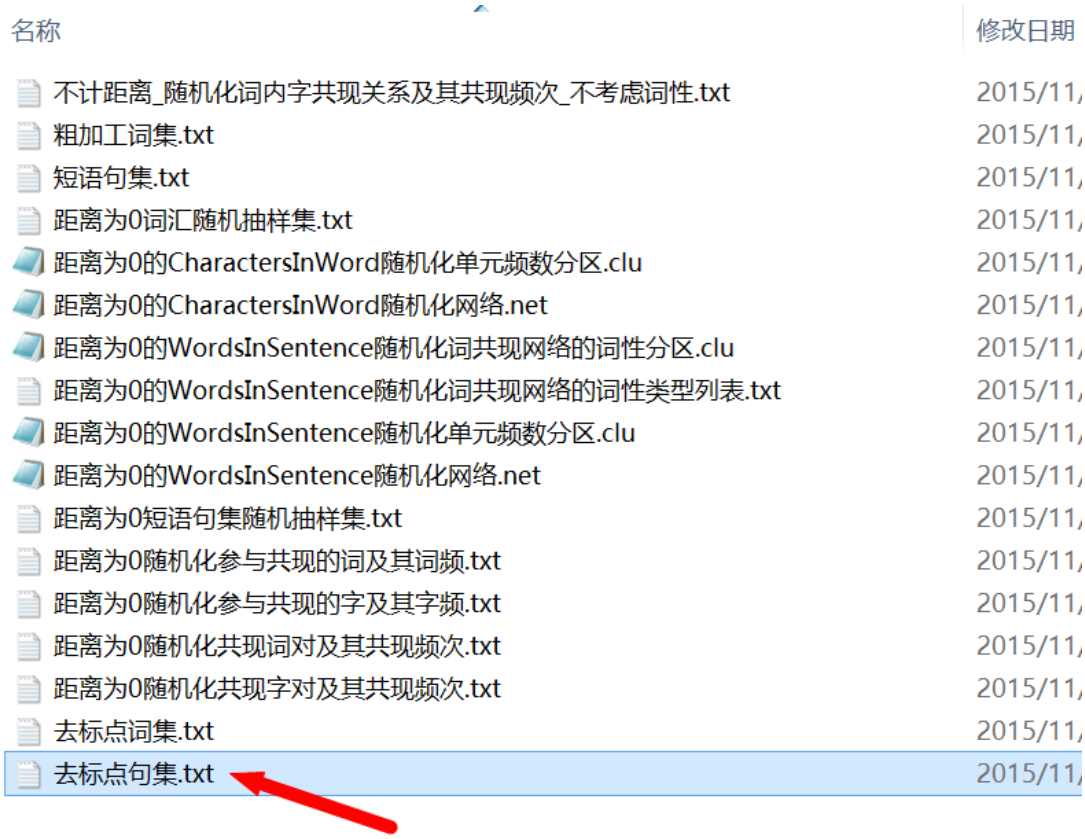
在分析句内词共现时，软件的操作首先是根据句号分出句子，然后删除句子中的其他任何标点，然后再进行词共现关系的分析。这种操作在已经标注了标点符号的人民日报语料库是很容易实现的，如果用户自己分词，但没有对词性和标点做过标注，要分析句内词共现就有一定的困难。因此，使用自制语料库分析的前提是：

① 自己已经用空格分了词。

② 句子已经用回车键分好，也就是说每个句子实际上已经分为段。

③ 向软件说明句子中用到了哪些标点符号，这些标点符号在稍后的分析中会首先删除。

如下图所示，在分析结果中，会有一个叫做“去标点句集”的文件，这就是在处理过程中会首先产生的文件。删除标点的过程依据的就是/w 或者自定义的标点符号。当然，在该框栏中已经填入了默认的标点符号集，基本上都全了。用户可以自己查看以下是否要增减。

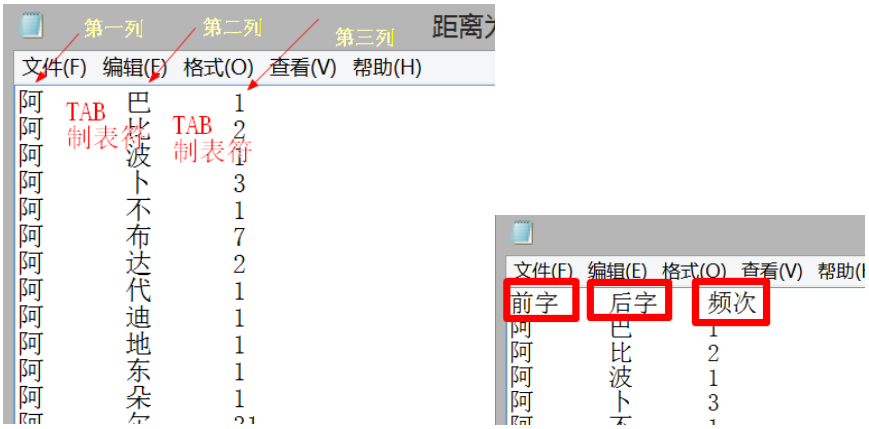


5、网络和分区文件制备区的使用说明

5.1 有向网络文件制备

本节采用上述制备网络制备结果中的“距离为 0 随机化共现字对及其共现频次.txt”（上图“去标点词集”的上一个文件）作为例子，说明网络和分区文件制备区的功能。

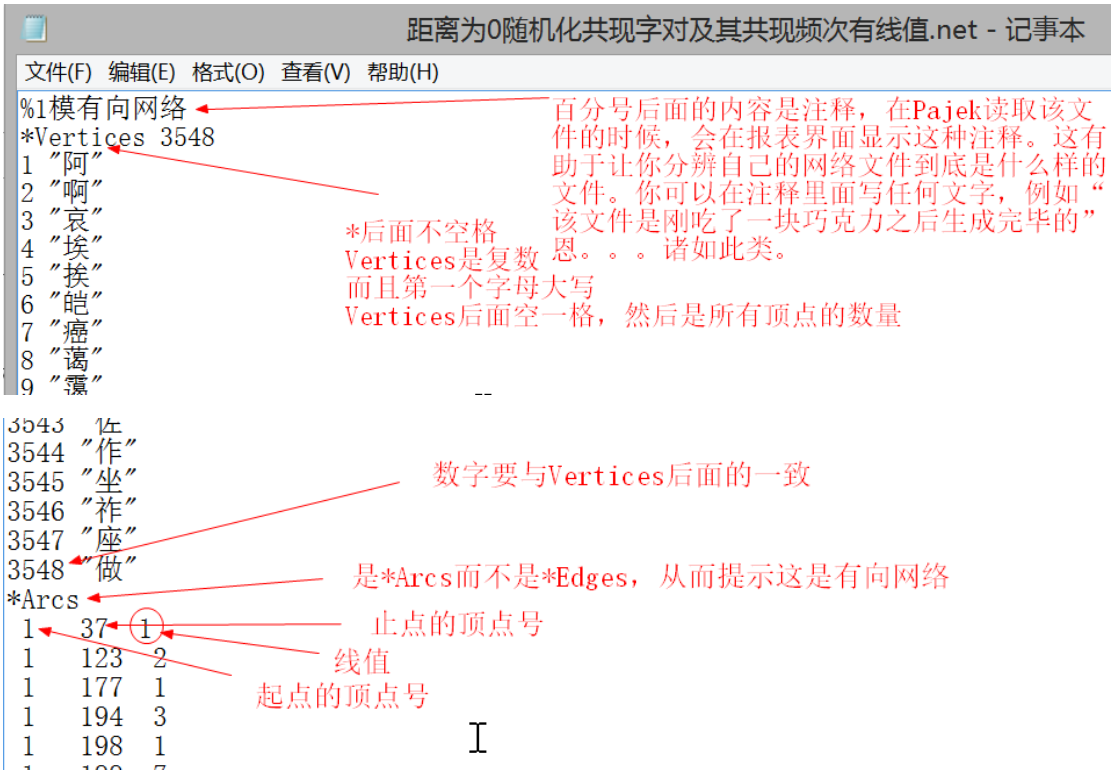
打开“距离为0随机化共现字对及其共现频次.txt”，可以看到下图所示结构：

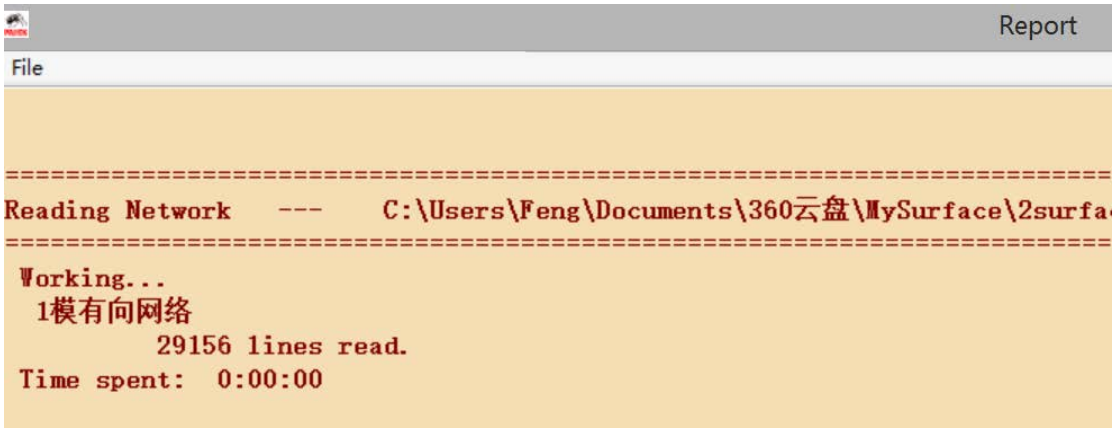


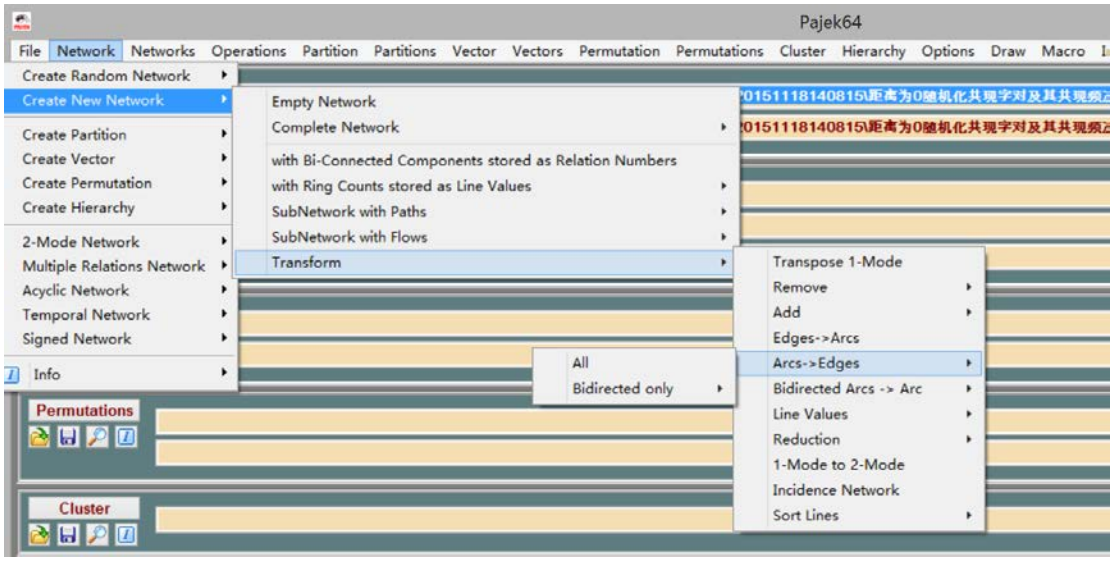
这种制表符分隔的文本文件，每一列之间都是制表符（你如果按一下 TAB 键，就会出现一个制表符而不是一个空格）。这样就确保可以把文件分成一列一列的数据结构。在这里可以看到，这个文件是没有列名的。也就是说，在第一行上没有诸如“第一个词”、“第二个词”这样的名称。因此数据的起始位置就是第 1 行。

如上图所示，点击“关系列表.txt”按钮，选择所要的“距离为 0 随机化共现字对及其共现频次.txt”文件，在发出列填入 1 表示，在接受列填入 2，线值列填入 3。这样就表示需要生成从第一列发出箭头指向第二列，这个箭头的线值或者说权重由第 3 列来规定。点击“创建.net 文件”按钮，就会在指定的结果文件夹下面生成“距离为 0 随机化共现字对及其共现频次有线值.net”这么一个文件。请注意，这个文件的文件名的前半部分是数据所来源的关系列表的文件名，后半部分“有线值”表示在“线值列”的框栏内指定了线值数据来源。如果不指定线值列，也可以生成网络文件。如果用户需要修改线值或者做一些计算操作，诸如所有线值乘以 0.001 之类的计算，那么需要在 Pajek 里面进行，请参看相关教程。

请注意：如果不设置起始行，则默认从第一行开始生成数据文件。但如果有列名的时候怎么办呢？很简单，只要在“起始行”框栏中输入 2，表示从第二行开始读取数据。这个起始行功能还可以用于某些特殊情况。例如，如果有一个很大的数据集，在第 1~999 行是亚洲的数据，第 1000 行以后开始时非洲的数据。如果只想制备非洲网络怎么办呢？指定“起始行”为 1000 即可。那么如果只想制定亚洲网络怎么办呢？暂时这个分析器还不能同时指定起始行和结束行，但是用户可以在文本文件中把从第 1000 行开始的数据都剪贴到其他文件里面，只保留 1~999 数据用来建立亚洲网络。





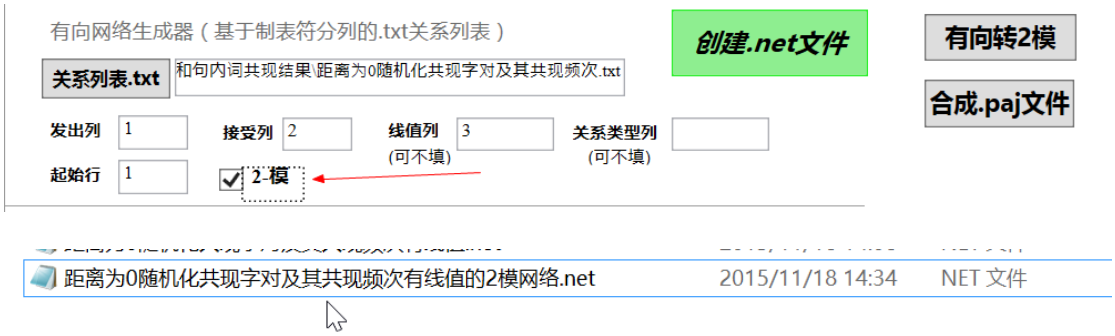


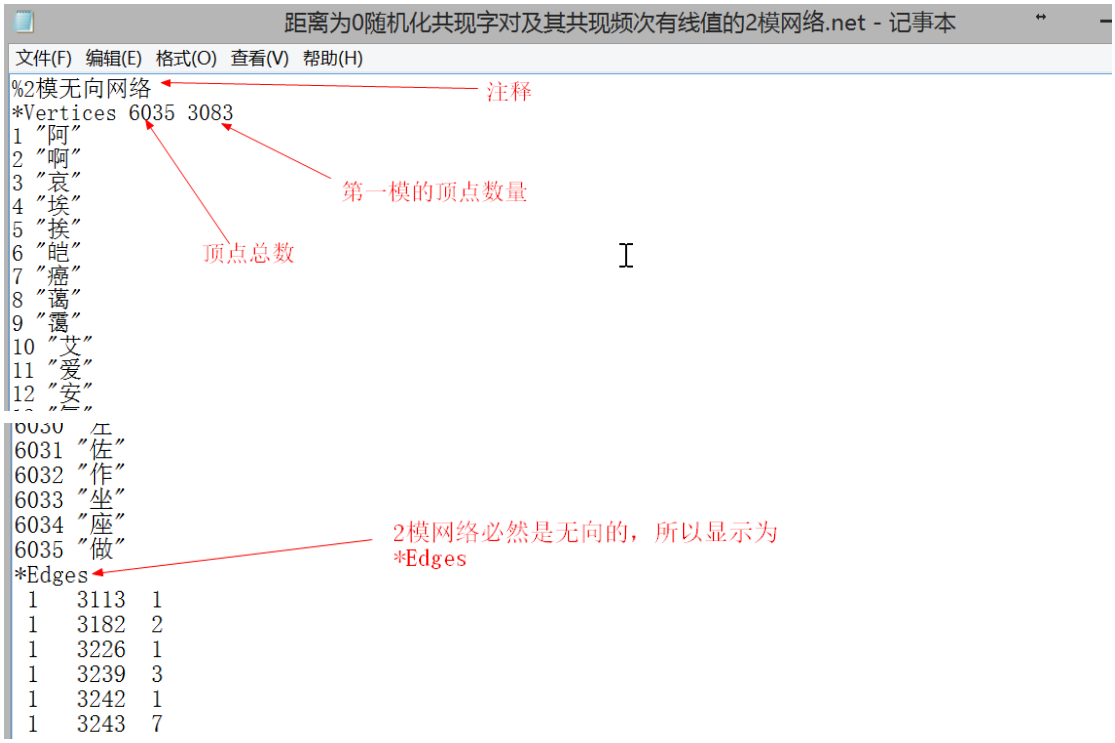
如上图所示，前两张图是.net 文件的结构截图。生成的是 1 模有向网络，在 Pajek 打开时可以看到报表界面直接显示百分号之后的注释内容。

如果想要生成无向网络怎么办？请到 Pajek 里面使用转换功能（如上图最后一张所示）。

5.2 二模网络文件制备

如果勾选“2-模网络”这个选项，那么会生成这样的网络文件“距离为 0 随机化共现字对及其共现频次有线值的 2 模网络.net”。可以发现，同样的数据集 2-模网络顶点数要比 1-模网络多，这是因为同样标签的顶点，在 1 模中算是同一个点，而在 2 模中则算是两个不同的点。

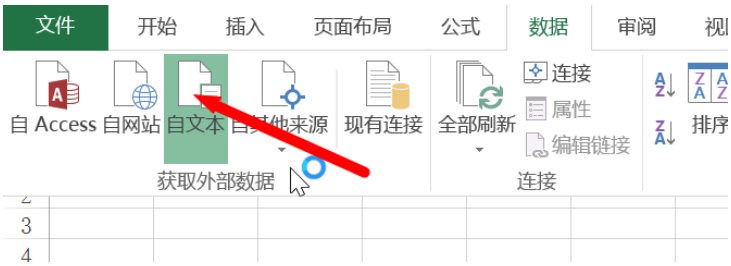


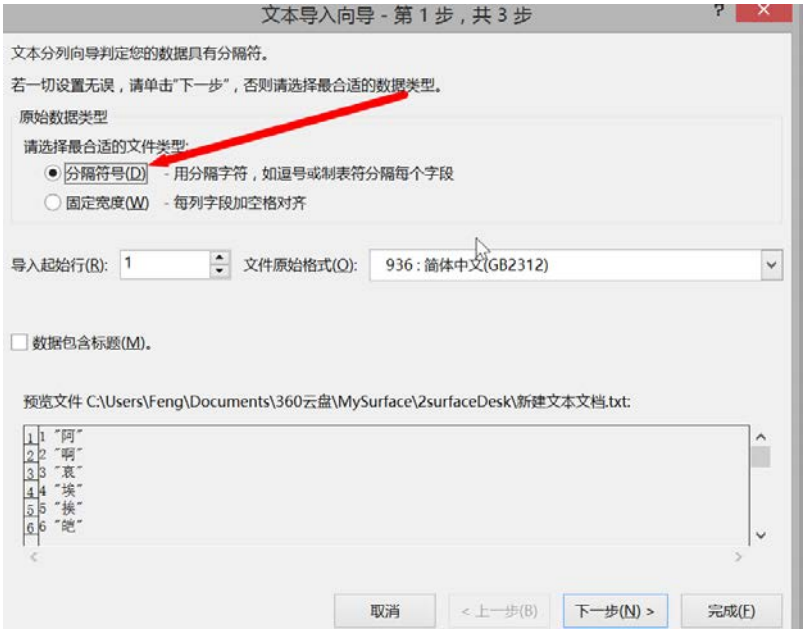


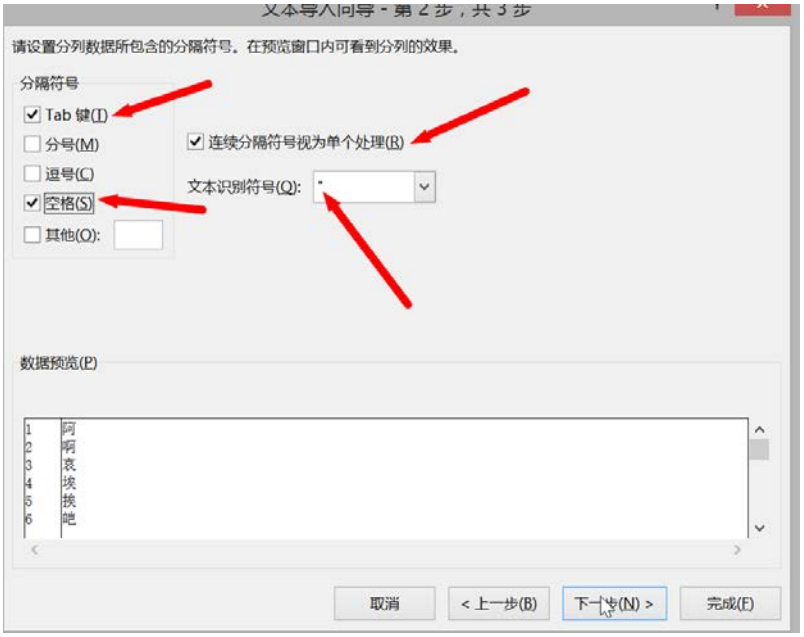
5.3 从 1-模有向网络转制成 2-模网络

2-模网络的有一种特殊性。如果在第一模块和第二模块中有相同名字的顶点，Pajek 网络会把它们认为是两个不同的顶点。但是如果是 1-模有向网络中，相同名称的顶点只属于一个模块，Pajek 只会把它看做是同一个顶点。因此，如果要转换网络文件，从 1-模有向转为 2 模是可以的。在有向网络中发出箭头的顶点就是第 1 模，而接受箭头的顶点就是第 2 模。

在使用这个转制功能时，需要在分区文件生成器的操作区针指定“顶点顺序表”。到哪里找顶点顺序表呢？只需要打开.net 网络文件，把*Vertices 这一行之下，到*Arcs 这一行之上的所有内容，都复制黏贴到另外一个.txt 文件中，然后用 EXCELL 的导入功能。如下图所示。



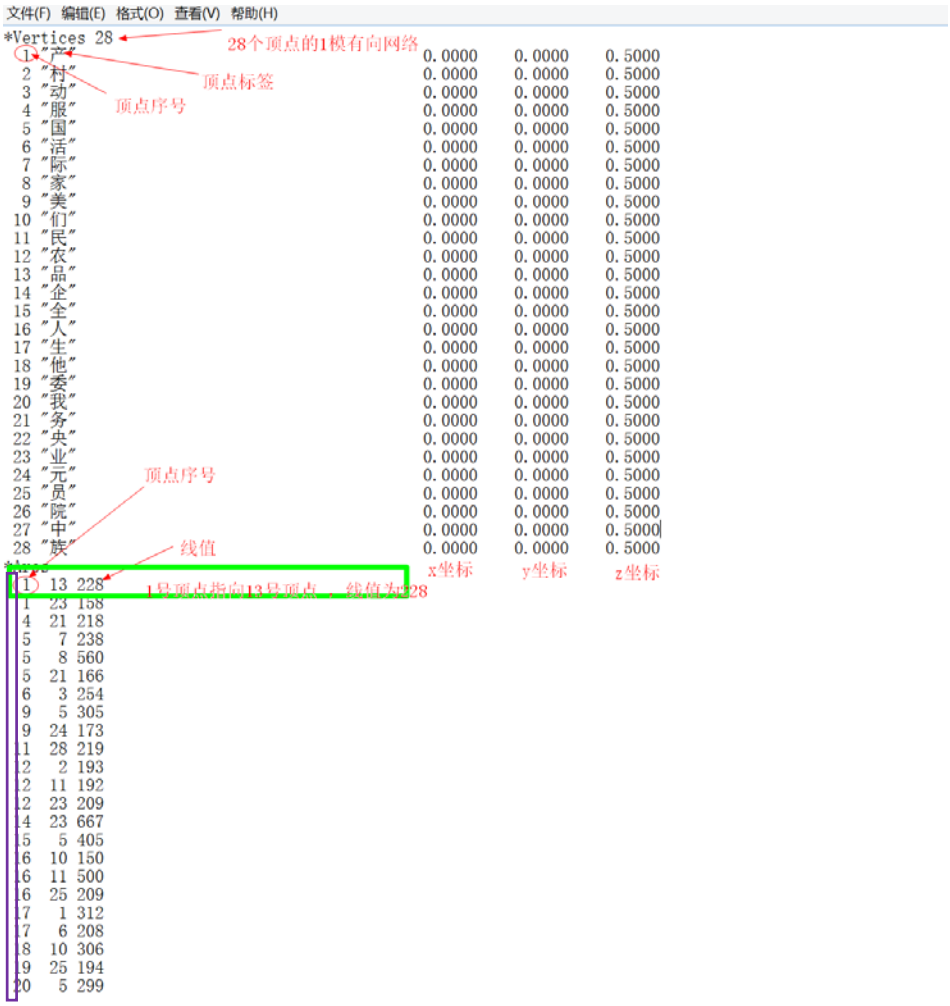


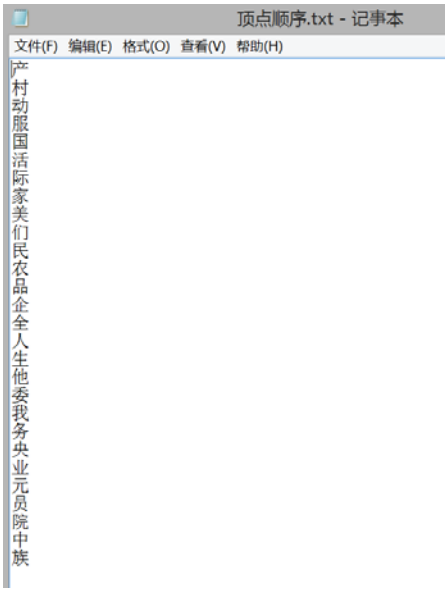


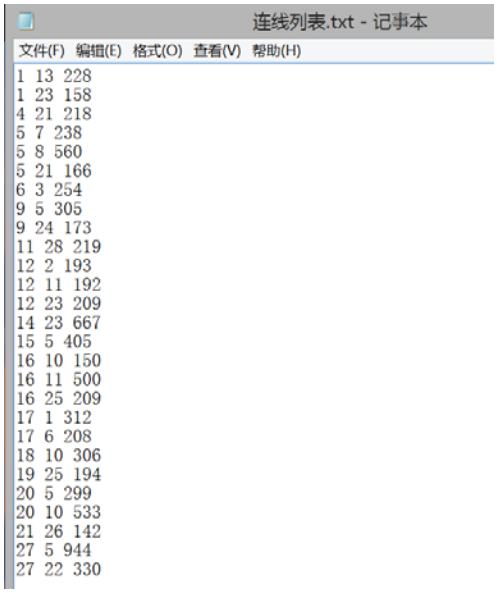
如果把鼠标停留在“顶点顺序表”按钮上，会得到如下提示：

因此，复制黏贴 EXCELL 表格里面的顶点名称列，而不要复制顶点编号列。把这一列复制黏贴到一个 txt 文件中，就制备成功了顶点顺序表。接下来把鼠标放到“有向转 2 模”按钮上。可以看到如下提示。也就是说。需要把.net 文件中*Arcs 以下的内容都黏贴到另一个文本文件中制成关系列表。值得注意的是，这种关系列表中的各列之间是用空格（SPACE）分隔的。这是因为在.net 文件中绝对不允许出现制表符，所有的空白区域都是空格。所以在使用“有向转 2 模网络”时，在“关系列表”这个框栏中指定的关系列表文件，实际上是有别于生成 1 模有向网络时所使用的制表符分隔关系文件的。

在指定了有向转 2 模用的关系列表和顶点顺序表之后，就可以点击有向转 2 模按钮进行转换了。在以下图片中使用一个简单的网络片段（在压缩包中名为“28 顶点网络文件.net”）来演示操作：



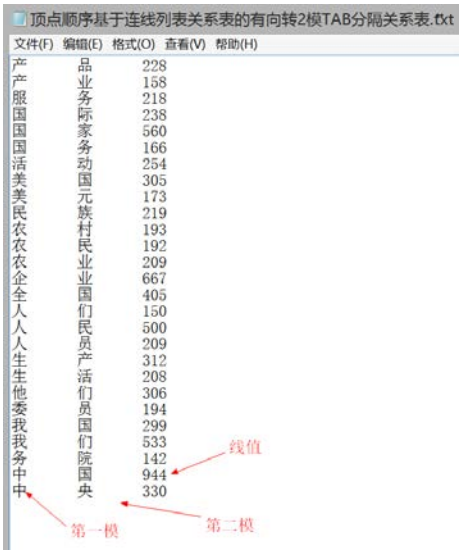


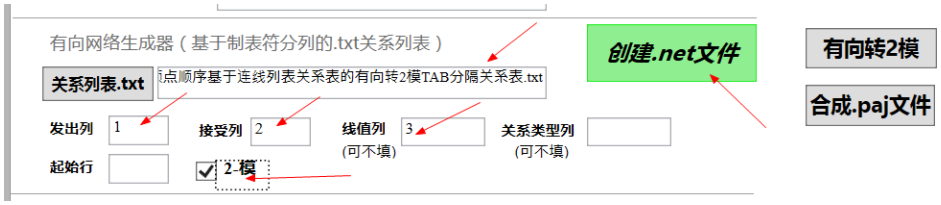


请注意：这里有一个程序 BUG。如果直接从.net 文件中复制黏贴连线关系列表的话，如上图紫色框所示，每一行的前面都会有一个空格。这会影响到分析器的识别。请采用前述 EXCEL 导入方法，把连线列表导入 EXCELL，另存为制表

符分隔的文件后，再把制表符替换成空格。由于在记事本中的“查找替换”界面上如果你按动 TAB 键是无法输入制表符作为查找目标的，所以你需要先在文本界面中找到一个制表符，Ctrl-c 复制它，再回到“查找-替换”界面中用 Ctrl-v 黏贴到“查找内容”框栏中，然后在“替换为”框栏中输入一个空格。最终替换成上图中“连线列表”的样子。最终转换界面如下。

经过转换后，实际上最后生成的不是一个 2 模网络，而是一张制表符分隔的列表。可以把该表重新指定到关系列表框中，勾选“2 模网络”单选框，生成网络。之所以要这么复杂地操作，主要是因为 1 模到 2 模，对于同名顶点来说，涉及查重问题和配到哪个模块的问题。操作如下图所示。当然，你也可以再次生成 1 模网络，并且用 Pajek 来与原来那个 1 模网络比较一下（Networks>Cross-Difference），看是不是相同，以验证软件的正确性。



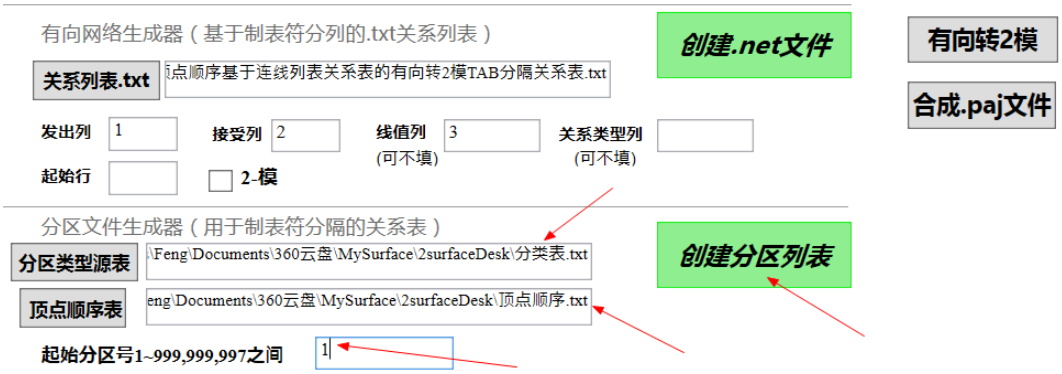


5.4 创建分区列表

Pajek 的分区文件是用于把顶点分成不同类别，以便进行类别的比较、合并或进一步细分之类的操作。事物总是可以分为各种不同的类别。仍以上面的 28 个顶点为例。如果有以下“分类表”，并且也没有可以按照网络的顶点顺序排序：

委	A 类
他	A 类
生	A 类
人	A 类
全	WU 类
企	ZO 类
品	Z 类
农	Z 类
民	Z 类
们	000 类
美	YES 类
家	000 类
际	YES 类
活	YES 类
国	A 类
服	000 类
动	abc 类
村	000 类
产	abc 类
族	CCC 类
中	000 类
院	A 类
员	A 类
元	WU 类
业	ZO 类
央	YES 类
务	YES 类
我	A 类

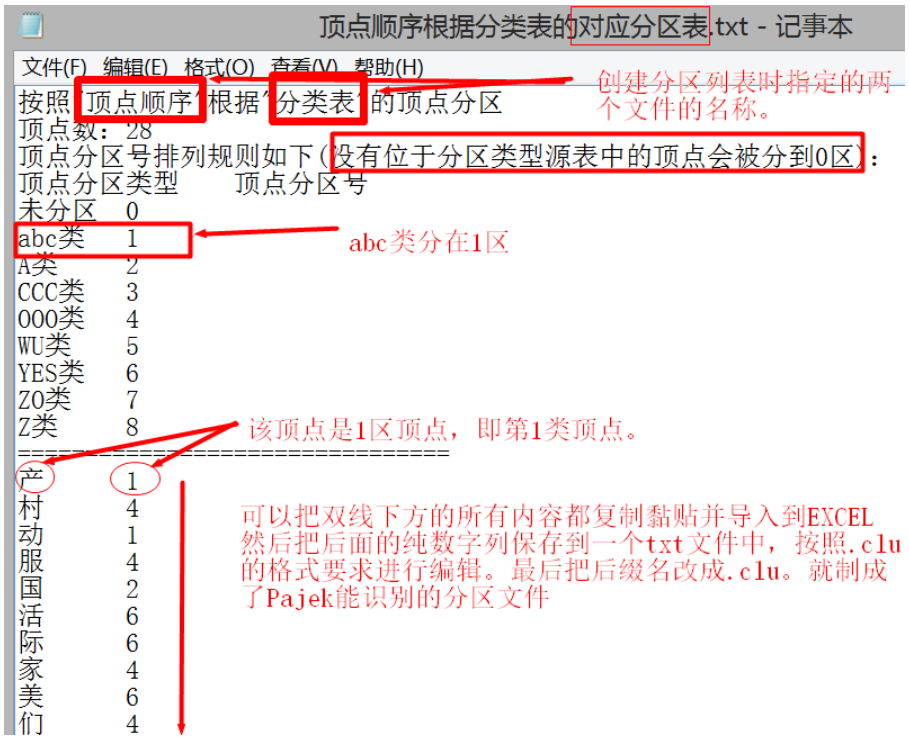
那么，上表可以存为一张制表符分隔的“分区类型源表”。无表头（即无列名），一列为未经排序的顶点标签。另一列为分区类型标签，例如上述“YES 类”。



之所以要指定起始分区号，是因为有些用户可能希望使用不同的分区系列，比方

说第一种分类方法分区号为 1~100 区间，而另外的分类法则使用 101~200 区间。便于在一个分区文件中管理所有分区类型。

请注意，这里的创建分区列表的结果不是一个.clu 文件，而是一张“分区列表”。该分区列表名为“…的对应分区表.txt”，其内容结构如下图所示：

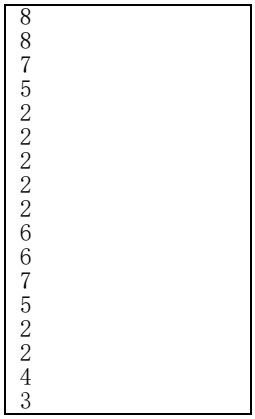


Pajek 的分区文件实际上就是一个文本文件，只是后缀名不是.txt，而是.clu。该文件的第一行是*Vertices 紧跟一个或多个空格再接着一个数字，数字指定了所对应的网络的顶点数，也就是说，网络中有多少个顶点。从第二行开始，每行一个数字，对应于所有顶点的分区号。也就是说，有多少个顶点，在*Vertices 下就有多少行。

在上图所示的分区列表文件中，双线下方的顶点标签实际上已经按照“顶点顺序”框栏中指定的顶点顺序进行了排序。因此后面的数值列就可以直接导入到 EXCEL 中后再复制黏贴回.clu 文件中，从而建立分区文件。

在文件包中有“28 顶点网络分区.clu”可供查看，结构如下：

```
*Vertices 28
1
4
1
4
2
6
6
4
6
4
8
```



这个分区文件构建器提供了一种有效的归纳整理顶点的方法，用户不需要在 EXCEL 里面再做顶点排序了，只需要根据标签在它的旁边一列输入其分类标签，就可以最后自动按照指定的顺序排好顶点。事实上，“顶点顺序表”中的标签数量可以多于或少于“分类表”中的标签，只要没能在分类表中找到相应的顶点标签，该顶点就会被分入 0 区。因此，在起始分区号的框栏中，是不允许填入 0 的。

该功能的另一个缺点是无法随意指定特定的分区类型号。但是用户可以通过 EXCELL 的查找替换功能来自己手工制作。

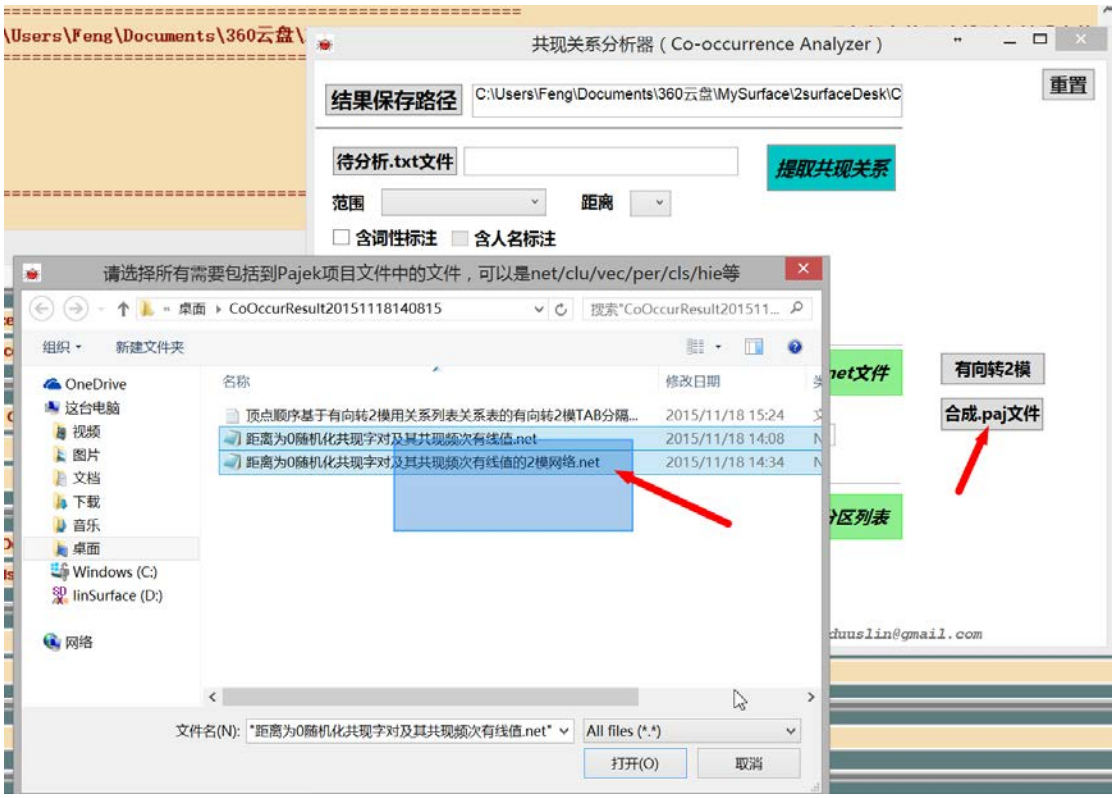
另外，该功能还生成另一个结果文件“…的分区类型名和用户定义起始值的分区号.txt”，其中有顶点分类名称和分区号。

5.4 创建工程文件

如果你同时创建了 100 个网络文件，加上 100 个相应的分区文件以及 100 个矢量文件。如何批量导入 Pajek 呢？

第一种方法是你用 Shift+左键或全选文件，并且点击按住第一个文件，拖动到 Pajek 相应的框栏（比方说网络文件就拖动到网络文件框栏）。那么鼠标所点中的那个文件，就会排列为 Pajek 所导入的第一个网络文件，其他所有文件依次批量导入。

第二种方法，就是创建工程文件.paj。



该功能最终会形成两个结果文件：

- paj 合成来源文件清单.txt
- 合成.paj

请注意：合成工程文件前，最好把所有来源文件放到同一个文件夹下，并且避免重名，最好用"名称_编号"的形式来命名。所有文件名要能够在 Pajek 中一目了然，以免批量操作时乱了次序。

有关工程文件的使用，请参看《蜘蛛：社会网络分析技术》。

6、常见问题：

关于结果文件的覆盖问题（重要!! 请一定阅读）：

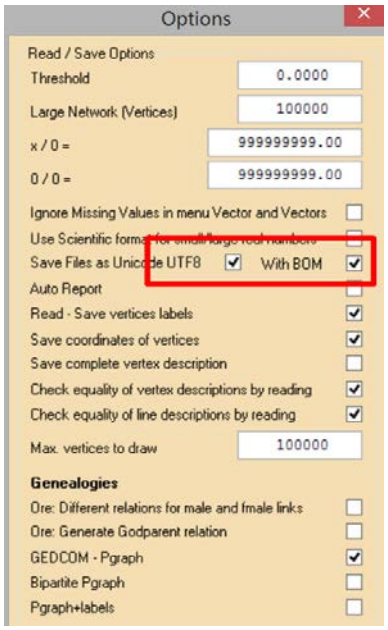
所有在结果文件夹中创建的.txt 文本，为了避免多次操作覆盖以往数据而导致数据损失，特地设置成为如果操作同一个功能，生成的文本在第二和第三次操作时会把新的结果添加在文件末尾。这是为了避免数据损失的一种特意设置。如果用户不希望出现这种添加发生，没有其他选项可设置，只有一个方法，就是再

次点击“结果保存路径”按钮，在指定文件夹下生成一个新的结果文件夹。由于结果文件夹用精确的时间数字命名，可以确保不会出现文件夹重名和覆盖的现象。好处是你避免了多次文件覆盖，不好的是你会在特定文件路径中产生大量的结果文件夹。所以建议用户一开始就设置一个诸如“分析结果文件夹”这样名字的母文件夹，把所有结果文件夹作为它的子文件夹。

如果操作的不是同一个功能，不需要更换结果文件夹，因为每种不同功能所生成的结果文件名字都不一样，不会产生覆盖现象。

Pajek 另存为网络文件，为什么再打开时汉字是乱码？

请勾选以下选项再另存文件。With BOM 是指在文件中加上一段数据代码，这些数据代码并不现实在文本文件阅览器上，而是隐藏的代码，这种代码可以让文字处理器知道要处理的是什么编码规格的文件。



转制文件的时候为什么总是报错？

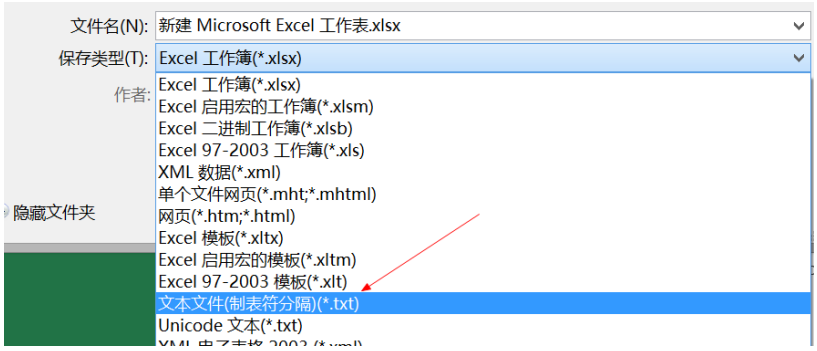
最常见的原因是在输入的制表符分隔文件中存在异常的空格。例如在某一行上只有起点列，没有止点列，或者没有线值列但却选了线值。所以在制备制表符分隔的文档时需要仔细检查。有一些用户试图从网页或其他文件上复制黏贴一些资料到 EXCEL 中，然后就直接制备制表符分隔的文本文件。这是很偷懒而且很不负责的方式，每一行数据都要经过检查，确保没有遗漏空格、制表符或者其他

异常符号，例如意外的单引号之类。

要在顶点标签中杜绝空格，而使用下划线代替空格。英文最好都大写。例如“America Company”，需要改成“AMERICA_COMPANY”。

请务必使用 EXCEL 的另存为制表符分隔文件的功能，而不要频繁黏贴复制，否则容易在文本文件中形成意外的空格或制表符，从而导致文件制备失败。

如下图所示：



为什么不提供矢量文件生成器？

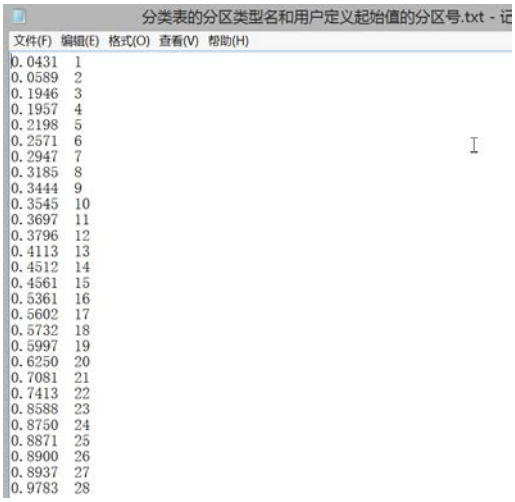
真是 sorry 了，当时做这个软件的时候忘记做进去了。

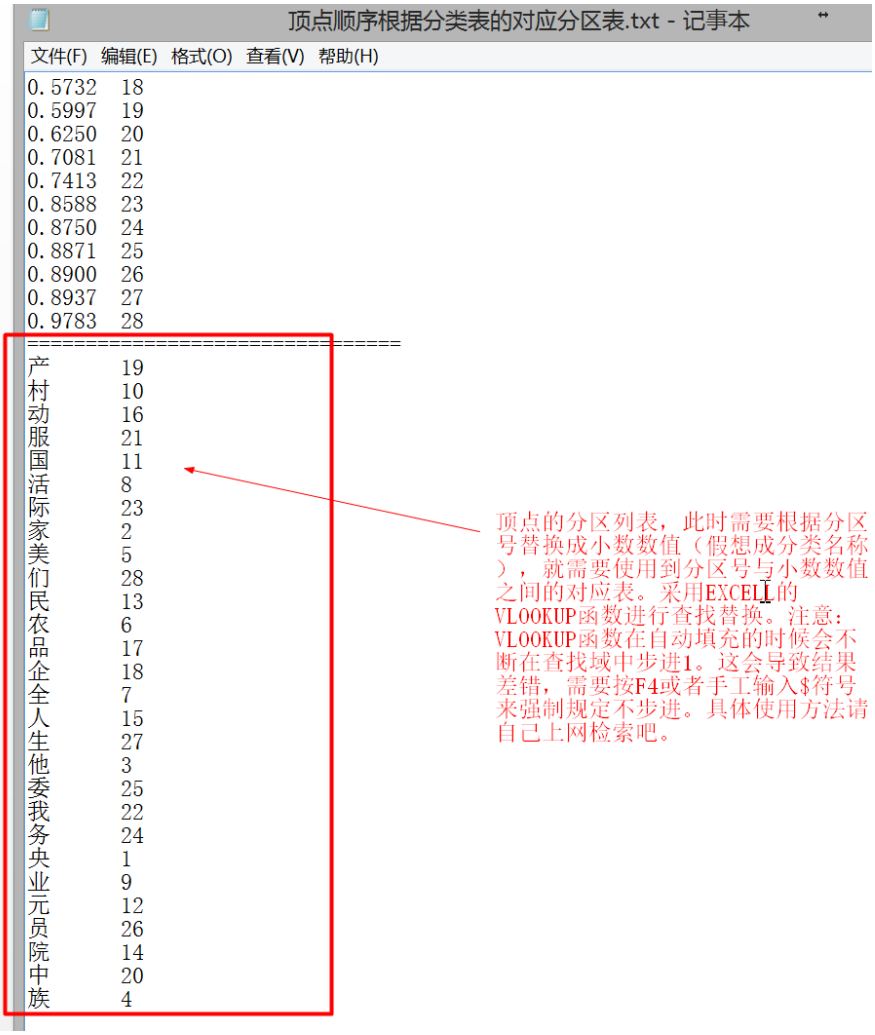
如果大家一定要生成矢量文件，也有一个折衷的办法。举例而言，如果顶点是公司，需要建立对应于公司营业额的带小数点的矢量（如 1.39 万元）时，可以先把这个是带小数点的数值直接就视为某种分类，即把它当做分区源文件来用。这样仍然用分区文件功能进行操作。这样在生成的分区结果中，会有相应的信息。如下例所示：

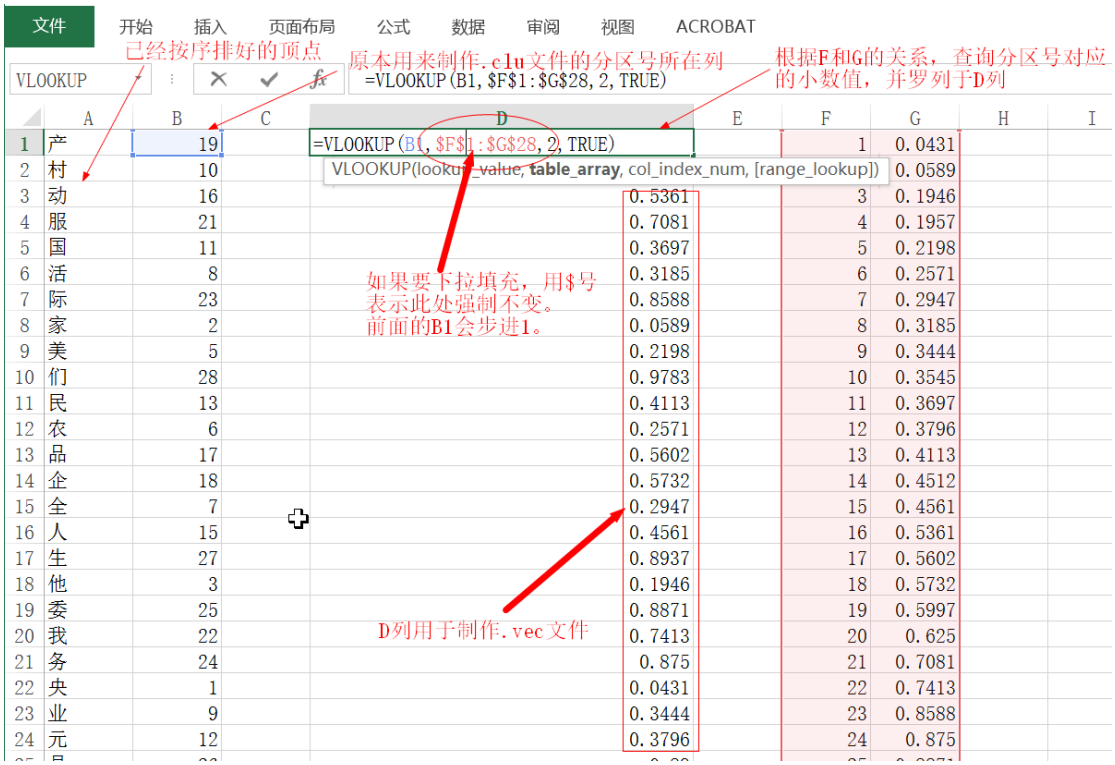
委	0.8871
他	0.1946
生	0.8937
人	0.4561
全	0.2947
企	0.5732
品	0.5602
农	0.2571
民	0.4113
们	0.9783
美	0.2198
家	0.0589
际	0.8588
活	0.3185
国	0.3697
服	0.7081
动	0.5361
村	0.3545
产	0.5997
族	0.1957

中	0.6250
院	0.4512
员	0.8900
元	0.3796
业	0.3444
央	0.0431
务	0.8750
我	0.7413

经过操作后你可以得到这么两个内容：







最终的.vec 文件如下，其数据排列逻辑实际上与.clu 文件相同。这种操作虽然较为繁琐，但是熟练后应当能满足所有网络分析文件制备需求。

```
*Vertices 28
0.5997
0.3545
0.5361
0.7081
0.3697
0.3185
0.8588
0.0589
0.2198
0.9783
0.4113
0.2571
0.5602
0.5732
0.2947
0.4561
0.8937
0.1946
0.8871
0.7413
0.875
0.0431
0.3444
0.3796
0.89
0.4512
0.625
0.1957
```