ROST 内容挖掘 系统

ROST
Content Mining System
User Manual

Version 6.0

2010.9.23

武汉大学

www.fanpq.com

ROST 虚拟学习团队

http://hi.baidu.com/rostcm/blog/item/62a4b3fe1cbf69d3b58f31d7.html

目 录

一、功能性分析

（1）分词

点击功能性分析下拉列表框中的分词选项，打开分词窗口，在待处理文本框中载入待处理文件，如“虚拟学习团队 2010-8-7.txt”，则系统按照程序目录下的 User 目录下的 User.txt 文档，自动在输出文件框中生成“虚拟学习团队 2010-8-7_分词后.txt”文件，获得以空格分离的分词后文档，如果原来文档中有空格的位置保留空格。点击确定按钮，即可打开该文档。

如果需要自己增加一些词，则点击工具下拉列表框中的自定义文件→分词自定义词表，系统将自动在记事本中打开 user 目录下的 user.txt 文件，编辑后点击保存存盘，再次重新启动本软件，方可生效。

（2）字频分析

点击功能性分析下拉列表框中的字频分析选项，打开字频分析窗口，在待处理文件框中载入待处理文件，如“虚拟学习团队 2010-8-7.txt”，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7_字频.txt”文件，点击确定按钮，即可打开该文档。

（3）英文词频分析

- 文件词频统计

点击功能性分析下拉列表框中的英文词频分析选项，打开 ROST

英文词频统计和超纲单词分析窗口。点击文件菜单下的打开菜单项或点击工具栏上的打开按钮，打开要统计的英文文档，然后选择统计菜单下的统计文件词频菜单项或工具栏上的统计按钮，即可统计出文档的所有单词。点击单选按钮纲内，可统计该文档的纲内词；点击单选按钮超纲，可统计该文档中的超纲词。选择复选框全选，可全选表格所有单词；选择复选框归并单词变形，可将变形单词进行归并。

对统计出的单词，在表格上点击右键，弹出快捷菜单，可以将选择的词汇添加到常用词语表，或者将选择的词汇从常用词语表中删除。

要在文本框中高亮显示某单词，可以勾选该单词的检查框；如果取消勾选，则文本框中该单词恢复普通显示状态。

- 剪切板词频统计

如果要统计剪切板词频，则选择统计菜单下的统计剪切板词频菜单项，则剪切板上的单词会显示在打开文件框中，再点击工具栏上的统计按钮即可。

- 查看统计表格

点击查看菜单下的统计表格菜单项，即可查看空的统计表格。

- 查看大纲列表

点击查看菜单下的大纲列表菜单项，打开大纲列表窗口，即可查看大纲列表。如果要查看某大纲，双击该行即可。在大纲列表窗口，还可以自定义某个词汇表，方法是在大纲名称文本框中输入大纲名称，然后在大纲文件文本框中载入大纲文件，再点击添加按钮即可。

若要删除某词汇表，则选中该词汇表后，点击删除按钮即可。

- 描红超纲词

如果要查看所有勾选的超纲词汇在文章中的位置，则首先点击统计、然后选择超纲，再勾选全选，然后点击查看菜单中的描红选定的超纲的词汇即可。

- 查看非词表

非词表你不想统计的单词或者字符的列表，该文件位于程序目录下的 dict 子目录下的 notwords.txt。要查看非词表，点击工具菜单下的查看非词表即可。如果要启动非词表，则工具菜单下的点击启动非词表。

- 加密词表

如果要对词表加密，则点击工具菜单下的加密词表；如果要解密词表，则则点击工具菜单下的解密词表即可。

- 打开词典目录

点击工具菜单下的打开词典目录即可。

（4）汉语频度分析

点击功能性分析下拉列表框中的汉语词频分析选项，打开汉语词频统计窗口，在分词后待统计词频文件文本框中载入分词后的文件，如“虚拟学习团队 2010-8-7_分词后.txt”，则系统自动载入过滤词表，并在输出文件文本框中生成词频统计文件“虚拟学习团队 2010-8-7_分词后_词频.txt”。在归并词群表文本框中载入归并词群表，还可以对文档中的词进行归并。在保留词表文本框中载入保留词表，则可

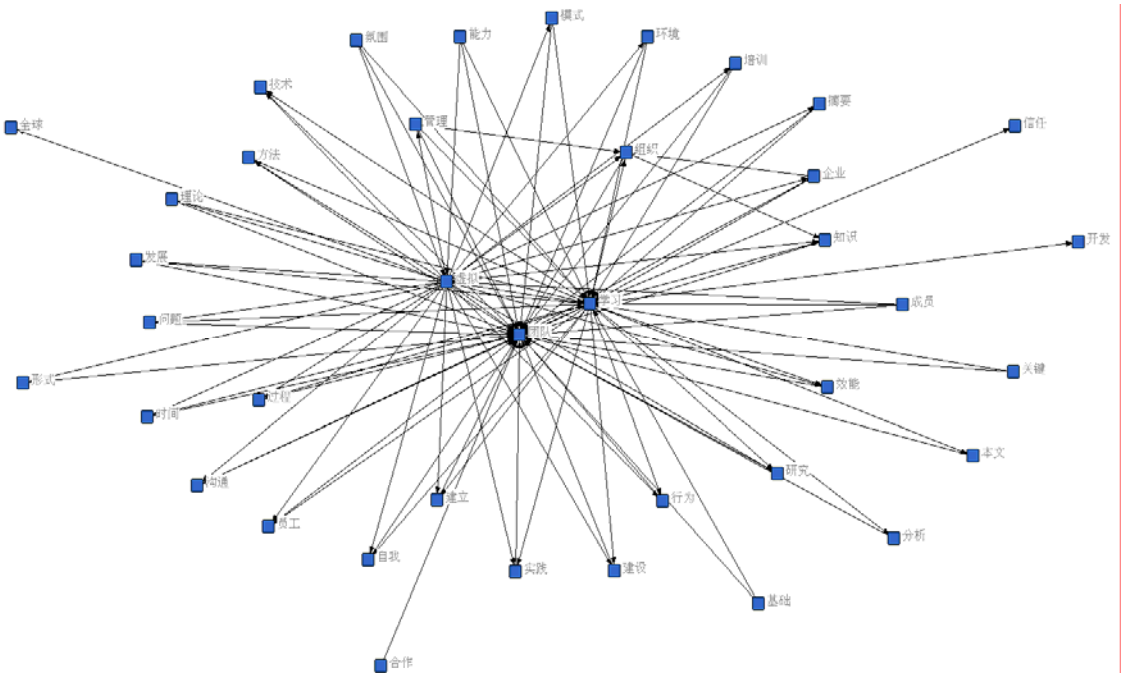
以将文档中在保留词表中的词保留下来。

（5）社会网络和语义网络分析

点击功能性分析下拉列表框中的社会网络分析选项，打开 ROST 语义网络和社会网络生成工具，在待处理文本框中载入待处理文件（待处理文件格式可以是一行一句的未分词文件，比如聊天记录，全网分析中的摘要文件等；也可以是一句若干词的已分词文件；还可以是多行有关联的已分词文件），然后点击高频词按钮，可以生成高频词表；点击过滤无意义词按钮，可以生成过滤后的高频词和共现矩阵词表；点击提取行特征按钮，可以生成行特征词；点击构建网络按钮可以生成语义网络的.VNA 文件和.txt 文件，如果进一步点击启动 NetDraw 按钮，则可以打开 NetDraw 工具，查看图形结果；点击构建矩阵按钮则可以生成共现矩阵文件。双击文件框可查看相应结果。

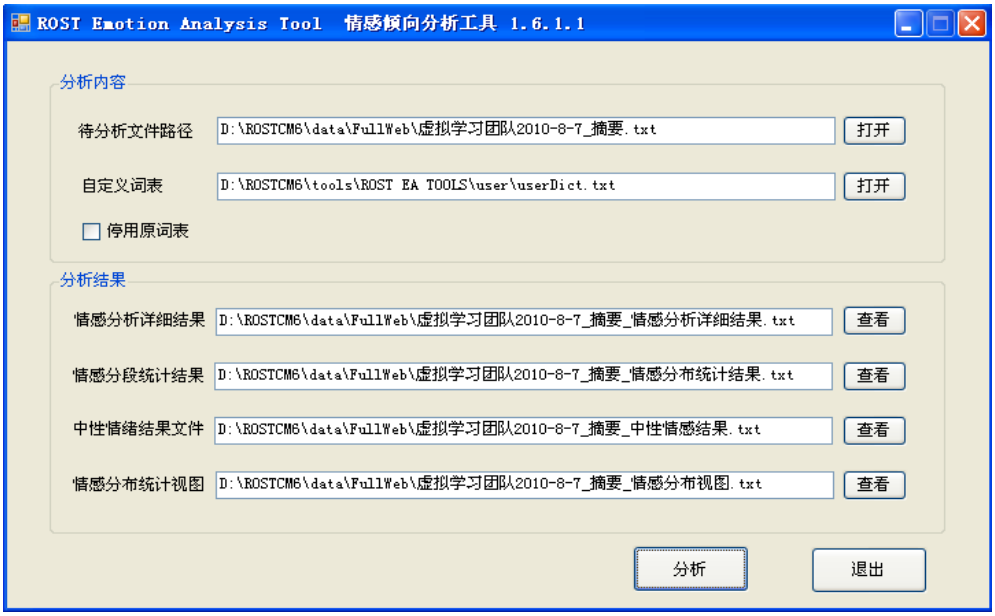
如果想进行快速分析，则载入待处理文件后，点击“快速分析”按钮，即可一次生成上述文件。可以是聊天内容文件，文件格式是

例如，以下是对“虚拟学习团队摘要文件”分析的结果：



（6）情感分析

点击功能性分析下拉列表框中的情感分析选项，在待分析文件路径文本框中载入待分析的文件，点击分析，然后双击各文本框后的查看，即可查看情感分析详细结果、情感分段统计结果、中性情绪结果文件和情感分布统计视图结果。

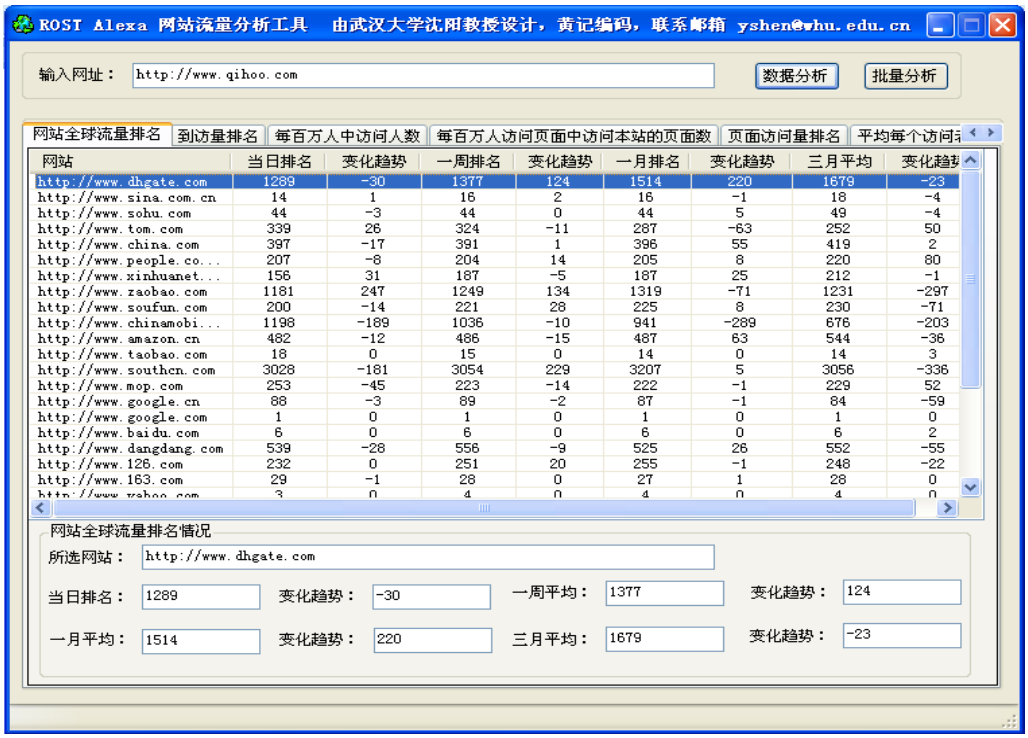


（7）流量分析

点击功能性分析下拉列表框中的流量分析选项，打开 Rost Alexa 网络流量分析工具，在输入网址文本框中输入要进行流量分析的网址，点击数据分析按钮即可。

还可以在该工具中进行批量分析，这时只需要点击批量分析按钮，导入需要进行批量分析的网页链接表，即可得到批量分析结果。

（8）TF/IDF 批量词频分析



点击功能性分析下拉列表框中的 TF/IDF 批量词频分析选项，打开 TF/IDF 批量词频分析窗口，点击批量打开文件按钮，选择需要打开的文件夹，即可在工具栏的下方打开所选文件夹中所有的.txt 文件。勾选文件前面的复选框，选中文件（可同时勾选多个文件），点击计算批量文件 IDF，窗口的左下方即可出现所选文件的 IDF 值。在已计算完 IDF 值的文件中选择一个文件，

然后点击计算当前所选文件 TFIDF 值，则在窗口的右下方出现所选文件的 TFIDF 值。

（9）相似分析

点击功能性分析下拉列表框中的相似分析选项，打开文档相关性监测工具，首先点击打开按钮，在待查文章选项卡下可以打开要检测的文档，点击检测按钮，即进行文档相关性检测，并可在结果查看选项卡下查看检测结果。点击停止按钮，即可停止检测。点击结果按钮，可以查看分析统计数据。点击目录按钮，可以打开相似度分析目录。点击退出按钮，即可退出检测系统。

（10）网站信息分析

点击功能性分析下拉列表框中的网站信息分析选项，打开 ROST 网络环境分析窗口，点击分析按钮，即可完成网络的环境分析。

（11）聚类分析

点击功能性分析下拉列表框中的聚类分析（测试模块）选项，打开聚类分析窗口，在待处理文本框中载入待类聚文件，然后填上类别数量，点击开始聚类即可对所选文件进行聚类分析。

（12）分类分析

点击功能性分析下拉列表框中的分类分析（测试模块）选项，打开短文本分类工具窗口，在待处理文本框中载入待分类文件，然后填上按第几字段分类，点击分析即可对所选文件进行分类分析；双击输出文件框中的文件目录即可打开分类后文件；双击特征词表中的文件目录即可看到特征词表。

二、文本操作

（1）字段抽取

点击文本操作下拉列表框中的字段抽取选项，打开抽取字段窗口，在待处理文本框中载入待处理文件，如“虚拟学习团队 2010-8-7.txt”，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7_抽取.txt”文件，然后在抽取出字段文本框中输入需要抽取的一个或两个字段，并在下面的复选框中选择抽取条件(注意:只有当抽取两个字段时,才选择抽取出两个字段都不为空的行复选框,否则,抽取无结果)。点击确定按钮，即可打开抽取结果文档。

（2）一般性行处理

点击文本操作下拉列表框中的一般性行处理选项，打开一般性行处理窗口，在待处理文本框中载入待处理文件，如“虚拟学习团队 2010-8-7.txt”，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7_一般性行处理.txt”文件，然后在处理条件单选框中，点击所需的处理条件，再点击确定按钮，即可打开按要求处理后的文档。

（3）基于正则的特定信息抽取

点击文本操作下拉列表框中的基于正则的特定信息抽取选项，打开基于正则的特定信息抽取窗口，在待处理文本框中载入待处理文件，如“虚拟学习团队 2010-8-7_域名表.txt”，然后在正则表达式文本框中右键点击所需行抽取条件的正则表达式，这里选择域名正则表达式，则当前表达式文本框中自动显示所选的正则表达式。点击确定按钮，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7_域

名表_正则抽取词.txt”文件，同时打开该文档。

（4）基于字段特征的行处理

点击文本操作下拉列表框中的基于字段特征的行处理选项，打开基于字段特征的行处理窗口，在待处理文本框中载入待处理文件，如“虚拟学习团队 2010-8-7.txt”， 则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7_一般性行处理.txt”文件，然后在抽取条件单选框中，选择所需抽取条件，再点击确定按钮，即可打开按要求处理后的文档。

（5）基于辅助词群的行处理

点击功能性分析下拉列表框中的基于辅助词群的行抽取及处理选项，打开基于辅助词群的行抽取及处理窗口。在待处理文本框中载入待处理文件，并在辅助文件文本框中载入辅助文件，然后在抽取条件单选框中，点击所需的单选按钮。点击单选按钮抽取出包含词群的行，则将在待处理文件中选出包含辅助文件中词语的行输出；点击单选按钮抽取出不包含词群的行，则将在待处理文件中选出不包含辅助文件中词语的行输出；点击单选按钮按照给定的批量行号提取行，则此时的辅助文件中只输入需要输出的行号（若需要输出多行，则辅助文件中输入一个行号后换行后再输入另一个行号。），则将在待处理文件中选取辅助文件中指定的行。

（6）文本的替换和增补

点击文本操作下拉列表框中的文本的替换与增补选项，打开文本的替换与增补窗口，在待处理文本框中载入待处理文件，如“虚拟

学习团队 2010-8-7.txt”，然后在处理条件单选框中，点击所需的单选按钮。点击单选按钮替换字段间隔符号，再点击确定按钮，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7 --替换隔离符号.txt”文件，即可获得按要求处理后的文档，即将文档中字段间的空格键替换为 Tab 键；点击单选按钮补行号(例如：将 1 补到 2)，再点击确定按钮，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7 --补行号.txt”文件，打开该文档，即可以看到该文档中只保留了源文档中的前两行，而且每行首部增加了相应的行号；点击单选按钮字段位置互换（例如：将 1 换到 2），再点击确定按钮，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7 --字段位置互换.txt”文件，即可以获得源文档中两个字段互换后的文档（例如：源文档中第 1 个字段与第 2 个字段进行了互换）；点击单选按钮批量词群替换，然后在辅助词群下的文本框中输入替换词和被替换词（被替换词应该是待处理文件中包含的词），两个词之间用空格键隔开，再点击确定按钮，则系统自动在输出文件框中生成“虚拟学习团队 2010-8-7 –词群替换.txt”文件，即可以获得源文档中某个词被另一个词替换后的文档。

三、 可视化

（1）标签云

点击可视化下拉列表框中的标签云选项，打开标签云窗口，点击工具栏上的打开按钮，打开已经分频后的频度文本文件，则在工具栏下方右边的输出窗口内自动显示打开的频度文本文件里的内容，在左边的输出窗口内将显示频度文本文件里的字或词（即生成的标签云），

而且这些字或词按照频度大小确定了自己的大小和颜色。即相同频度的字或词将以同一颜色和大小显示。调节工具栏上的最大字体，则可以调节标签云字体整体的大小。点击工具栏上的保存按钮，则可以将生成的标签云以 JPG 图片的形式保存下来。

四、工具

（1）剪贴板控制器

点击工具下拉列表框中的剪贴板控制器选项，打开剪贴板数据采集窗口，在文本框中可看到剪贴的数据，点击右键可进行复制、清空等相应操作；点击目录按钮，即可打开剪切版目录，选择该目录下面的若干文件，在弹出菜单中点复制，然后在自动采集工具中点击粘名按钮，就可以获得这些文件的文件名。选中清空前次内容的复选框，即可清空前次剪切内容；选中监控剪切板的复选框，即开始监视，本项默认选中，用户可根据需要适时取消；点击退出按钮，即可退出。

（2）域名排名查询器

点击工具下拉列表框中的域名排名查询器选项，打开网站排名查询工具窗口；点击打开按钮，打开一个域名文件，在总共查询_网址文本框中看到文件中的网址；点击开始按钮，可开始排名查询，相应结果会在右边得到_个排名网址的文本框中显示；若在查询期间已经查询到自己要的结果，可点击中止按钮；点击排名按钮，排名结果会保存到自定义的文件中；点击退出按钮，即可退出该工具。

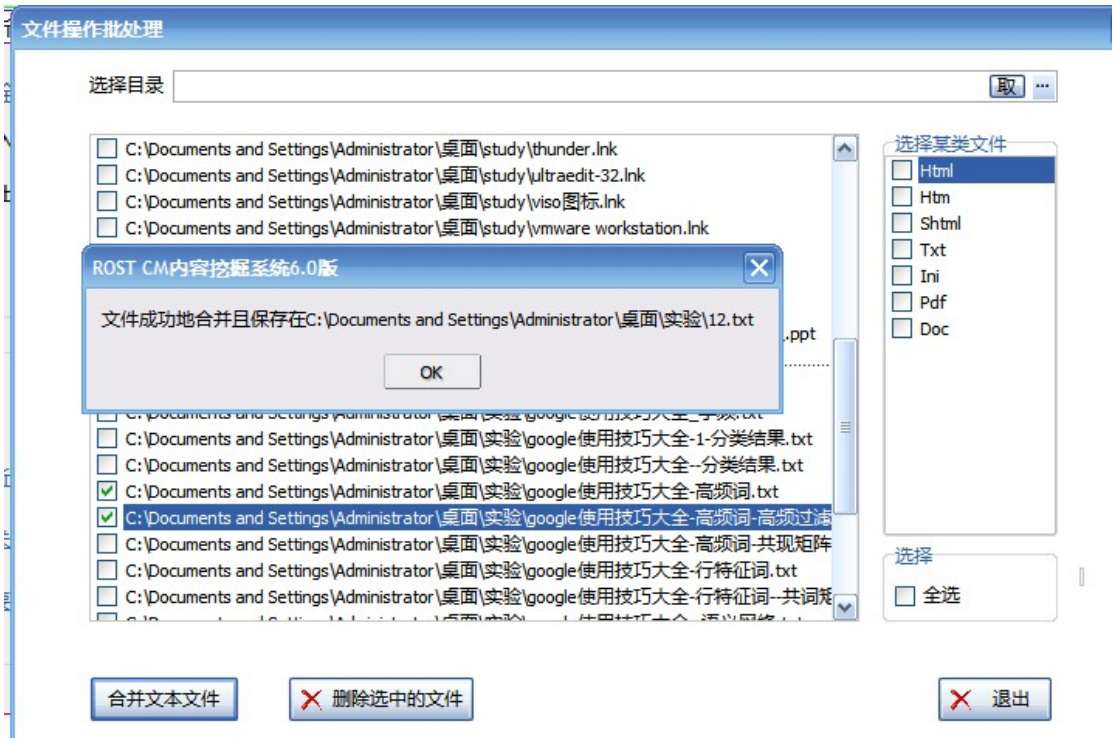
（3）批量文件格式转换器

点击工具下拉列表框中的批量文件格式转换器选项，打开各类文件转换为 Txt 文件窗口；在输入目录文本框中选择相应输入目录，输出目录默认和输入目录所选一样，用者可自行选择；然后点击确定按钮，及开始转换，转换完毕会有如下提示：

双击输出目录中的文件目录即可打开相应分析目录。

（4）批量文件处理器

点击工具下拉列表框中的批量文件处理器选项，文本操作批处理窗口；在选择目录的文本框中选择相应目录，然后选择所需文件类型前面的复选框，若全选，可直接选中全选的复选框；点击合并文本文件按钮，选择文件存储目录，填写文件名称，即可开始合并。



选中相应文件，点击删除选中文件按钮，即可删除该文件。删除成功有如下提示：

（5）浏览网页文本实时抓取器

点击工具下拉列表框中的浏览网页文本实时抓取器选项，打开实时浏览数据抓取窗口；选中监控网页前的复选框（默认选中），即可开始实时监控，当在浏览器中打开一个网页后，在浏览器事件日志文本框中显示浏览事件日志,在网页正文文本框中会显示网页正文，在网页所含链接及标签文本框中显示该网页所含链接及标签，点击合并按钮，即可生成合并链接.txt,合链接与标签.txt 和合并正文.txt 件 3 个

合并文件；点击目录按钮，即可打开浏览网址中数据文件所在目录.

（6）NetDraw

点击工具下拉列表框中的 NetDraw 选项，打开 NetDraw 软件，点击 file->open ,开一个.VAN 文件,可以生成语义网络图.

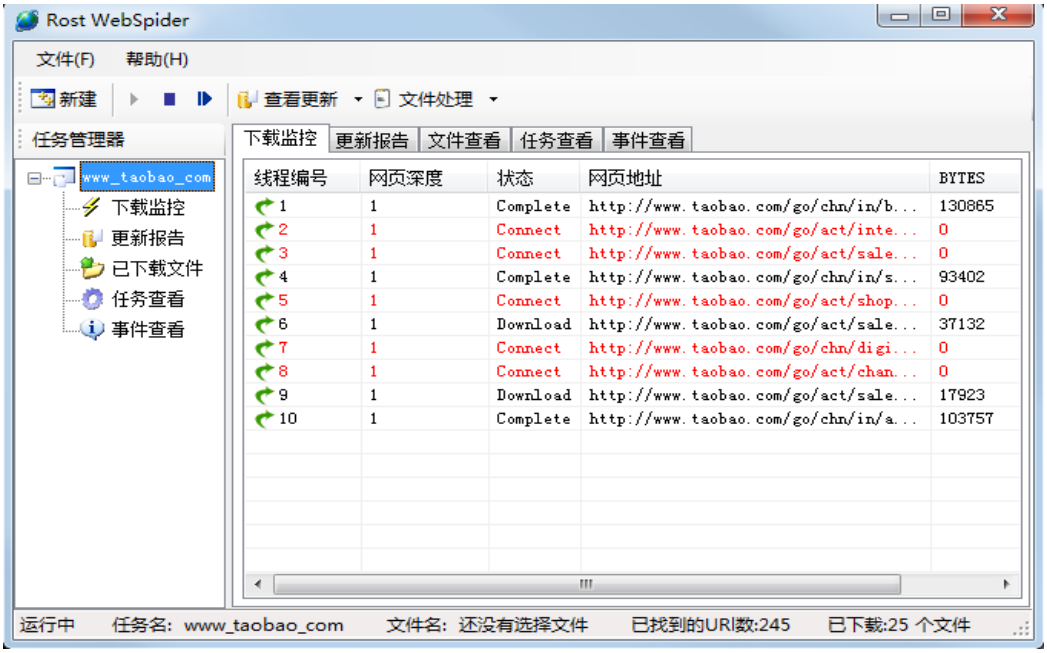
（7）ROST WebSpider

点击工具下拉列表框中的 ROST WebSpider 选项，打开 ROST WebSpider 窗口；在文件菜单下点击新建任务菜单项，打开新建任务窗口，该窗口包含任务目录、下载类别、连接设置、文件类型和网址过滤 4 个选项卡。如果进行任务目录设置，则点击任务目录选项，输入任务名称，并设置存放网站数据的位置；如果下载类别设置，则点击下载类别选项，让后选择下载类别选项卡，如果是整站下载，则点击整站下载选项卡，输入网站入口 URL；如果是指定 URL 下载，则点击指定 URL 下载选项卡，并将要下载的 URL 添加到 URL 列表中；如果是指定目录下载，则点击指定目录下载选项卡，输入入口 URL；最后点击跨站下载选项卡，并添加 URL 入口或从文件导入 URL 到 URL 入口列表中即可。

如果进行连接设置，则点击连接设置选项，即可对下载的线程数、连接超时时间、抓取网页最大深度、URL 队列为空时线程等待时间、两个连接之间的停顿时间、以及超链接的最大长度进行设置。此外还可以选择是否同一 TCP 连接要抓取多个网页。

如果要对下载的文件类型进行设置，则点击文件类型选项卡，对允许下载的文件类型进行设置。

还可以对下载的内容进行限制。点击内容限制选项卡，可以限制下载包含某些域名的网页、包含某些文件扩展名的网页或指定链接需要包含的字符串。



此外，还可以在窗口中进行下载监控、更新报告、文件、任务和事件的查看。

（8）调试用

（9）程序目录

点击工具下拉列表框中的程序目录选项，则可看到程序目录、工具目录、用户目录、样例目录。

（10）数据目录

点击工具下拉列表框中的数据目录选项，即可看到：全网分析数据目录、微博分析数据目录、网站分析数据目录、浏览分析历史数据目录、浏览分析实时数据目录、期刊分析数据目录、剪切板数据目录、新浪评论目录。

（11）第三方工具

点击工具下拉列表框中的第三方工具选项，可进行：RSS 阅读数据采集器、全文检索、导入数据库数据、 测试第三方工具操作。

（12）自定义文件

点击工具下拉列表框中的自定义文件选项，可进行：分词自定义词表、重载自定义词表、分词过滤词表、词频统计过滤词表操作。

五、聊天分析

要分析聊天记录，首先必须从QQ消息管理器的导入导出菜单下的导出消息记录导出消息的文本文件（.txt文件），然后点击在待处理文件文本框后的...，载入要处理的消息文本文件，然后点击导入按钮，使之格式化，即完成用户数据的整理。然后再点击分析按钮，进行分析。分析完成后，可点击分析框中的发言频度文件、口头禅文件、总词频文件和聊天内容文件超链接，查看相应结果。

启动情感分析模块，载入格式化后的聊天记录文件（不是刚刚导出的聊天记录原始文件），点击分析按钮，还可得到情感分析详细结果、情感分段统计结果、中性情绪结果文件和情感分布统计视图等情感分析结果。

六、全网分析

在输入搜索词文本框中输入要搜索的关键词，点击搜索按钮，则搜索引擎根据该关键词搜索并返回的所有网页结果默认存放在程序目录下的 data 目录下的 fullweb 目录中，类似这样命名：虚拟学习团队 2010-8-7.txt。双击输出文件文本框，即可查看结果。也可以进一

步点击分析按钮，待分析完毕，即可分别点击相关词频表、网页链接表、域名表和摘要超链接，查看相应结果。该结果也默认存放在上述目录中。

通过搜索引擎得到的全网数据还可做以下分析：

（1）全网数据中的摘要或标题数据中的词语、机构的共现关系。方法是在社会网络分析工具中载入全网分析结果的摘要文件，点击“快速分析”按钮，即可双击文件框查看结果，或启动 NetDraw 查看图形结果。

（2）情感分析。只需要将全网数据中的摘要数据载入情感分析工具，点击分析按钮即可。

（3）域名的批量流量分析。只需将网页链接表载入到流量分析模块中，即可进行该网页链接表对应的域名批量流量分析。

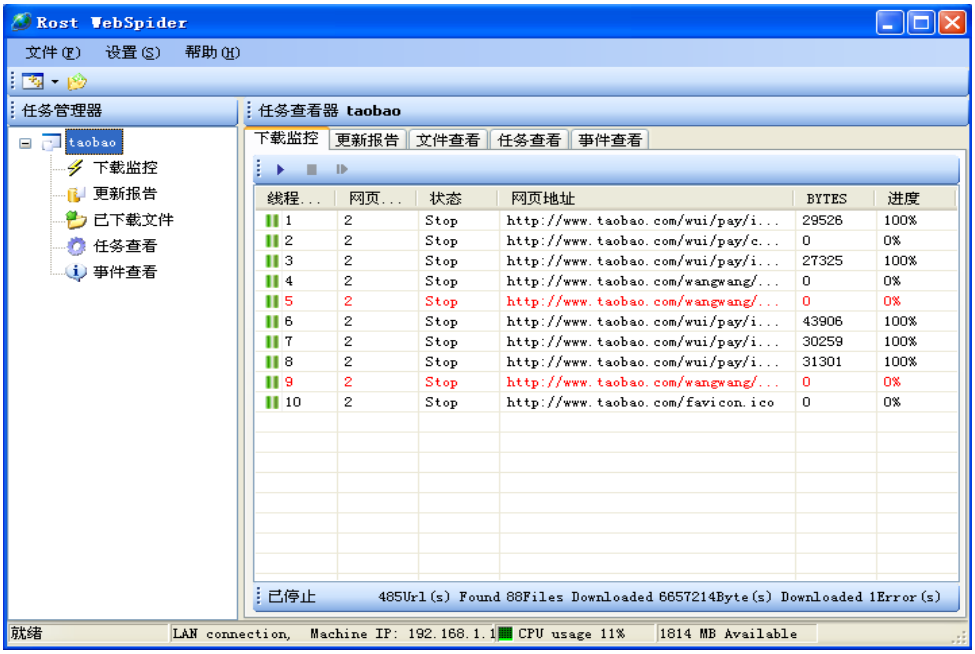
（4）将网址列表载入到迅雷中进行下载。

七、网站分析

（1）获得网站数据

有两种方法获得网站数据，一是直接启动网站抓取，抓取下来的网页保存在程序目录的 data\website\网站名\webPage 目录下。

另一个获得网站数据的方法是启动高级网站抓取功能，即启动 Rost　WebSpider 抓取工具，如下图。



在文件菜单下点击新建任务菜单项，打开新建任务窗口，该窗口包含地址设置、连接设置、文件类型和内容设置 4 个选项卡。如果进行地址设置，则点击地址设置选项卡，输入任务名称，如果是整站下载，则点击整站下载选项卡，输入网站入口 URL；如果是指定 URL 下载，则点击指定 URL 下载选项卡，并将要下载的 URL 添加到 URL 列表中；如果是指定目录下载，则点击指定目录下载选项卡，输入入口 URL；最后点击跨站下载选项卡，并添加 URL 入口或从文件导入 URL 到 URL 入口列表中即可。

注意为了将下载的网站数据放到指定的位置，可以点击设置菜单项的设置任务文件夹菜单项，设置存放网站数据的位置。

如果进行连接设置，则点击连接设置选项卡，即可对下载的线程数、连接超时时间、抓取网页最大深度、URL 队列为空时线程等待时间、两个连接之间的停顿时间、以及超链接的最大长度进行设置。此外还可以选择是否同一 TCP 连接要抓取多个网页。

如果要对下载的文件类型进行设置，则点击文件类型选项卡，对允许下载的文件类型进行设置。

还可以对下载的内容进行限制。点击内容限制选项卡，可以限制下载包含某些域名的网页、包含某些文件扩展名的网页或指定链接需要包含的字符串。

此外，在任务查看器中可以进行下载监控、查看更新报告、查看文件、任务和事件。

（2）分析

点击分析按钮对抓取的网页文件即可做进一步的分析，生成网页的文本文件和全站合并文件。点击分析框中的网页的文本文件和全站合并文件超链接，即可查看结果。这些结果分别默认存放在 data\website\网站名\webPage\analysis 目录下。

八、浏览分析

首先点击获得历史浏览数据按钮，然后点击分析按钮，即可得到分析结果。点击标题文件、URL 文件和标题词频文件超链接，即可查看结果。

点击获得实时阅读数据按钮，打开 ROST 实时浏览数据抓取工具，即可获得实时阅读数据。

九、微博分析

（1）扫描数据

在微博分析前首先要登录自己的微博，然后在词或微博地址文本

框中输入要搜索的关键词或者微博地址，点击分析按钮，则搜索引擎根据该关键词或微博地址搜索并将返回的所有结果默认存放在程序目录下的 data 目录下的 Mblog 目录中，类似这样命名：虚拟学习团队 2010-9-15-21-31-47.txt。双击输出文件文本框，即可查看结果。

若输入的词或网址有误，会出现如下提示:

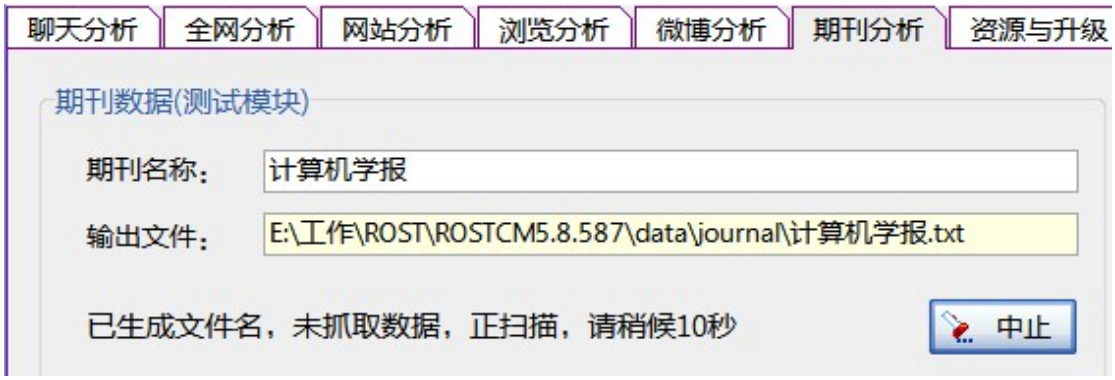
双击 OK，重新输入正确的微博地址即可。若选择连续采集，则可以连续的搜索相关的数据。

（2）分析

点击分析按钮对抓取的网页文件即可做进一步的分析，生成网页的文本文件和全站合并文件。点击分析框中的网友网名关系表和微博和微博文词频文件超链接，即可查看结果。这些结果分别默认存放在 data\Mblog\网站名\Mblog\analysis 目录下。

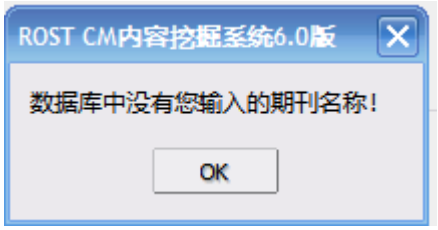
十、期刊分析

首先在期刊名称文本框中输入相应的期刊名称，然后点击扫描按钮，出现如下图提示：



此时，若想中止，可点击中止按钮，待扫描完毕，所有扫描的期刊数据结果默认存放在程序目录下的 data 目录下的 journal 目录中，类似这样命名：计算机学报.txt。双击输出文件文本框，即可查看结果。

若数据库中没有所输入的期刊，则会出现如下提示：



，双击 OK 即可关闭。

注：由于在教育网中由于各个图书馆做了地址跳转，拿不到任何数据，所以该模块只能非教育网使用。

基于内容挖掘的人文社会科学数字化研究平台，是一组功能联系紧密，可相互智能协作，无缝互操作的软件及插件包，最终形成能够依据一定范式进行人文社科智能化学术研究的数字化研究平台。人文社会科学数字化研究平台（以下简称“数字人文平台”）的构建和升级能够为研究者提供一个高效、有针对性的人文知识的获取、分析、集成和展示的数字化研究平台。能够对目前海量的数字化人文资料进行组织、标引、检索和利用，以保证人文研究的海量性、智能性和客观性，可节省大量的人力物力，提高研究效率，并可通过定量分析和定性分析的结合，从中归纳出具有说服力的普遍性结论。

在人文科学研究活动中融入了现代信息技术，整合人与计算机的优势研究复杂问题，这不仅仅只是传统研究范式向新范式的转移，也是自然科学研究范式与人文研究范式、定量研究范式与质性研究范式地整合。

软件的构造为插件型整合体系，即整个软件由多个小软件构成，它们各自实现不同的功能，相互联系又相互独立。应用于网络数据采集的小软件有 ROST WebSpider、ROST SeaT 和 ROST MicroBlog。其中，利用 ROST WebSpider 采集网页信息；利用 ROST SeaT 采集搜索引擎信息，并能够支持批量监控；利用 ROST MicroBlog 获取微博客信息。通过这些软件根据用户输入关键词对该类数据进行采集，采集对象包括特定主题网页、特定主题网站、某些网站的特定网页和特定内容、微博客、博客圈、论坛、社会网络、语料库、带有公开密码的数据库内容、搜索引擎内容

解析、公开的 QQ 群记录、学生上网上机数据、个人上网信息、邮箱数据、各类人员名单以及机构名单等。

ROST CM（数字人文辅助研究平台），可分析论文、微博、博客、论坛、网页、书籍、聊天记录、电子邮件、本地文本类格式文件、数据库中各类文本字段，分析方法目前支持：分词、字频统计、词频统计、聚类、分类、情感分析（含简单和复杂）、共现分析、同被引分析，依存分析、语义网络、社会网络、共现矩阵等分析方法。

ROST CM 目前的下载量超过 7000 次，使用者遍布国内外 100 多个高校，包括 Cambridge University（剑桥大学）、Loughborough University、Texas A&M University、日本北海道大学、北京大学、清华大学、浙江大学、诺基亚、武汉大学、南开大学、厦门大学、四川大学、天津大学、东北大学、东北师范大学、中南大学、中央民族大学、中山大学、北京科技大学、南京农业大学、南京航空航天大学、山东大学、广州大学、武汉理工大学、江西师大、江西理工大学、河南大学、河海大学、泰山学院、西南交通大学、长沙理工大学、澳门大学等。

本平台由武汉大学信息管理学院、计算机学院沈阳教授博导设计，编码。

其他编码参与人员有：洪婧倞、付晴川、寇文波、沈劲枝、李舒晨、田晨耕、任晓东、吴尚儒等、王鹏、涂龙。