export LANG="C"

cat weibo.top10wan |cut -f 5,9 > weibo\_user

##### 去重 #####

cat weibo\_user | sort -k2rn | awk -F"\t" 'BEGIN{last="0";}{if($2!=last)print$0;last=$2;}' > weibo\_user\_id\_uniq

##### Split成4个小文件（我的机器是4核，所以分成4个文件） #####

split -l $(expr $(expr $(cat weibo\_user\_id\_uniq | wc -l) / 4) + 1) weibo\_user\_id\_uniq split\_weibo\_user\_

##### 并发sort小文件 #####

cut -f1 split\_weibo\_user\_aa | sort | uniq -c| awk -F" " '{print $2"\t"$1;}' > result\_aa &

cut -f1 split\_weibo\_user\_ab | sort | uniq -c| awk -F" " '{print $2"\t"$1;}' > result\_ab &

cut -f1 split\_weibo\_user\_ac | sort | uniq -c| awk -F" " '{print $2"\t"$1;}' > result\_ac &

cut -f1 split\_weibo\_user\_ad | sort | uniq -c| awk -F" " '{print $2"\t"$1;}' > result\_ad &

wait

##### 合并 #####

cat result\_aa result\_ab result\_ac result\_ad > result

##### 将小文件中相同的userid合并，其weibo count 相加 #####

cat result | sort -k1rn | awk -F "\t" 'BEGIN{

lastid="0";

lastcount=0;}

{if($1!=lastid){

if (lastid!="0") {

print lastid"\t"lastcount;}

lastid=$1;

lastcount=$2;

}else{

lastid=$1;

lastcount+=$2;

}

}END{print lastid"\t"lastcount;}' > result\_merge

##### 按照发文量重新sort #####

cat result\_merge | sort -k2rn -k1 > finalresult

head -100 finalresult

Tips：

1. 在我的测试中，将文件分成越多份，运行的速度越快，我最多分到了10份，运行速度是最快的。
2. 如何测试最后得到的weibo count是否正确，我写了一个简单的代码测试：

cut -f5,9 <PATH\_TO\_FILE> | grep <USER\_ID> | cut -f2 | sort | uniq | wc -l  
跑出来后与得到的结果的count比较，相同即正确。