关于生成的统计结果文件的说明

目　录

1 基本说明

1.1 爬取范围(爬取页面)

指从起始 url 开始爬取，当 url 符合界面上设置的限定域范围时，该 url 属于爬取范围(爬取页面)。只有当 url 属于爬取范围时，该页面才会被继续跟踪爬取。

1.2 抓取范围(下载条目)

url 在属于爬取范围的基础上，若同时符合界面上设置的条目抓取规则，则该 url 属于抓取范围(下载条目)。

2 基于抓取条目的分析

2.1 运行结果

本文档记录了爬虫的运行结果，包括：运行时间、请求页面数、响应页面数、响应字节数、响应成功页面数(200)、抓取条目数、成功下载条目数。

2.2 基于抓取条目的链接列表

2.2.1 所有链接的列表

针对属于抓取范围内的条目，若该页面上存在链接(链接可以属于抓取范围内外、爬取范围内外，即页面上的全部链接)，则列举出来。

2.2.2 抓取范围内链接列表

针对属于抓取范围内的条目，若该页面上存在链接(链接必须属于抓取范围内)，则列举出来。

2.2.3 抓取范围外链接列表

针对属于抓取范围内的条目，若该页面上存在链接(链接必须属于抓取范围外，可以属于爬取范围内外)，则列举出来。

2.3 基于起始 url 对应站点的链接统计

2.3.1 各类链接统计

针对属于抓取范围内的条目所进行的统计。统计表格中，纵向为起始 url 对应的站点，横向为各类统计指标。其中，Pages_1 指对应站点所包含的爬取范围内的页面数；Pages_2 指对应站点所包含的抓取范围内的页面数；KnownLinks 统计链接，该类链接指向抓取范围内的页面；UnknownLinks 统计链接，该类链接指向抓取范围外的页面(包括爬取范围内外)；InterLinks 统计链接，该类链接指向对应站点内的页面(范围为爬取范围内)；OutLinks 统计

链接，该类链接指向对应站点外的页面(范围为爬取范围内外)。

对于每个爬取范围内的页面，若该页面属于起始 url 对应的站点，则相应的 Pages_1 指标加 1；如该页面同时属于抓取范围，则相应的 Pages_2 指标加 1。

对于每个抓取到的条目，若它不属于纵向中的任一站点，则跳过；否则，针对每项指标逐一统计：若条目对应的页面中存在一条指向抓取范围内页面的链接，则 KnownLinks 加 1；以此类推。

2.3.2 互链统计

针对属于抓取范围内的条目所进行的统计，即链接两端均属于抓取范围。统计表格中，第一列、第二列为起始 url 对应的站点；对于统计的链接对象，第一列指链接的起始端，第二列指链接的目标端，第三列为该类链接的统计结果。

对于每个抓取到的条目，若它不属于起始 url 对应的站点，则跳过；否则，查找该条目所对应的页面中是否存在链接指向起始 url 对应的站点，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于抓取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

2.3.3 互链统计矩阵

针对属于抓取范围内的条目所进行的统计，即链接两端均属于抓取范围。统计表格中，纵向和横向为起始 url 对应的站点；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个抓取到的条目，若它不属于起始 url 对应的站点，则跳过；否则，查找该条目所对应的页面中是否存在链接指向起始 url 对应的站点，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于抓取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

2.4 基于限定域的链接统计

2.4.1 各类链接统计

针对属于抓取范围内的条目所进行的统计。统计表格中，纵向为设置的限定域，横向为各类统计指标。其中，Pages_1 指对应域所包含的爬取范围内的页面数；Pages_2 指对应域所包含的抓取范围内的页面数；KnownLinks 统计链接，该类链接指向抓取范围内的页面；UnknownLinks 统计链接，该类链接指向抓取范围外的页面(包括爬取范围内外)；InterLinks 统计链接，该类链接指向对应域内的页面(范围为爬取范围内)；OutLinks 统计链接，该类链接指向对应域外的页面(范围为爬取范围内外)。

对于每个爬取范围内的页面，该页面必然属于限定域，相应的 Pages_1 指标加 1；如该页面同时属于抓取范围，则相应的 Pages_2 指标加 1。

对于每个抓取到的条目，它必然属于纵向中的某一域，针对每项指标逐一统计：若条目对应的页面中存在一条指向抓取范围内页面的链接，则 KnownLinks 加 1；以此类推。

2.4.2 互链统计

针对属于抓取范围内的条目所进行的统计，即链接两端均属于抓取范围。统计表格中，第一列、第二列为设置的限定域；对于统计的链接对象，第一列指链接的起始端，第二列指

链接的目标端，第三列为该类链接的统计结果。

对于每个抓取到的条目，它必然属于限定域，查找该条目所对应的页面中是否存在链接指向限定域，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于抓取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

2.4.3 互链统计矩阵

针对属于抓取范围内的条目所进行的统计，即链接两端均属于抓取范围。统计表格中，纵向和横向为设置的限定域；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个抓取到的条目，它必然属于限定域，查找该条目所对应的页面中是否存在链接指向限定域，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于抓取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

2.5 基于抓取页面的链接统计

2.5.1 各类链接统计

针对属于抓取范围内的条目所进行的统计。统计表格中，纵向为设置的抓取范围内的条目，横向为各类统计指标。其中，KnownLinks 统计链接，该类链接指向抓取范围内的页面；UnknownLinks 统计链接，该类链接指向抓取范围外的页面(包括爬取范围内外)；InterLinks 统计链接，该类链接指向本条目对应站点(不一定为起始 url 对应的站点)内的页面(范围为爬取范围内)；OutLinks 统计链接，该类链接指向本条目对应站点(不一定为起始 url 对应的站点)外的页面(范围为爬取范围内外)。

对于每个抓取到的条目，针对每项指标逐一统计：若条目对应的页面中存在一条指向抓取范围内页面的链接，则 KnownLinks 加 1；以此类推。

2.5.2 互链统计矩阵

针对属于抓取范围内的条目所进行的统计，即链接两端均属于抓取范围。统计表格中，纵向和横向为抓取范围内的条目；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个抓取到的条目，查找该条目所对应的页面中是否存在链接指向抓取范围内的其它页面，每查找到一条符合条件的链接，则相应的统计条目的值设置为 1(即两页面间若存在链接，则统计值设置为 1；若不存在链接，则统计值设置为 0)。

2.5.3 互链统计矩阵(去除全零行)

针对属于抓取范围内的条目所进行的统计，即链接两端均属于抓取范围。统计表格中，纵向和横向为抓取范围内的条目；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个抓取到的条目，查找该条目所对应的页面中是否存在链接指向抓取范围内的其它页面，每查找到一条符合条件的链接，则相应的统计条目的值设置为 1(即两页面间若存在链接，则统计值设置为 1；若不存在链接，则统计值设置为 0)。

对于统计结果，若存在某行的值全为零(即该页面内不存在指向抓取范围内的其余页面的链接)，则去除该行。以此，得到最终的互链统计矩阵。

3 基于爬取页面的分析

3.1 运行结果(同 2.1)

本文档记录了爬虫的运行结果，包括：运行时间、请求页面数、响应页面数、响应字节数、响应成功页面数(200)、抓取条目数、成功下载条目数。

3.2 基于爬取条目的链接列表

3.2.1 所有链接的列表

针对属于爬取范围内的条目，若该页面上存在链接(链接可以属于抓取范围内外、爬取范围内外，即页面上的全部链接)，则列举出来。

3.2.2 爬取范围内链接列表

针对属于爬取范围内的条目，若该页面上存在链接(链接必须属于爬取范围内)，则列举出来。

3.2.3 爬取范围外链接列表

针对属于爬取范围内的条目，若该页面上存在链接(链接必须属于爬取范围外)，则列举出来。

3.3 基于起始 url 对应站点的链接统计

3.3.1 各类链接统计

针对属于爬取范围内的条目所进行的统计。统计表格中，纵向为起始 url 对应的站点，横向为各类统计指标。其中，Pages 指对应站点所包含的爬取范围内的页面数；KnownLinks 统计链接，该类链接指向爬取范围内的页面；UnknownLinks 统计链接，该类链接指向爬取范围外的页面；InterLinks 统计链接，该类链接指向对应站点内的页面(范围为爬取范围内)；OutLinks 统计链接，该类链接指向对应站点外的页面(范围为爬取范围内外)。

对于每个爬取范围内的页面，若该页面属于起始 url 对应的站点，则相应的 Pages 指标加 1。

对于每个爬取到的条目，若它不属于纵向中的任一站点，则跳过；否则，针对每项指标逐一统计：若条目对应的页面中存在一条指向爬取范围内页面的链接，则 KnownLinks 加 1；以此类推。

3.3.2 互链统计

针对属于爬取范围内的条目所进行的统计，即链接两端均属于爬取范围。统计表格中，第一列、第二列为起始 url 对应的站点；对于统计的链接对象，第一列指链接的起始端，第二列指链接的目标端，第三列为该类链接的统计结果。

对于每个爬取到的条目，若它不属于起始 url 对应的站点，则跳过；否则，查找该条目所对应的页面中是否存在链接指向起始 url 对应的站点，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于爬取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

3.3.3 互链统计矩阵

针对属于爬取范围内的条目所进行的统计，即链接两端均属于爬取范围。统计表格中，纵向和横向为起始 url 对应的站点；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个爬取到的条目，若它不属于起始 url 对应的站点，则跳过；否则，查找该条目所对应的页面中是否存在链接指向起始 url 对应的站点，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于爬取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

3.4 基于限定域的链接统计

3.4.1 各类链接统计

针对属于爬取范围内的条目所进行的统计。统计表格中，纵向为设置的限定域，横向为各类统计指标。其中，Pages 指对应域所包含的爬取范围内的页面数；KnownLinks 统计链接，该类链接指向爬取范围内的页面；UnknownLinks 统计链接，该类链接指向爬取范围外的页面；InterLinks 统计链接，该类链接指向对应域内的页面(范围为爬取范围内)；OutLinks 统计链接，该类链接指向对应域外的页面(范围为爬取范围内外)。

对于每个爬取范围内的页面，该页面必然属于限定域，相应的 Pages 指标加 1。

对于每个爬取到的条目，它必然属于纵向中的某一域，针对每项指标逐一统计：若条目对应的页面中存在一条指向爬取范围内页面的链接，则 KnownLinks 加 1；以此类推。

3.4.2 互链统计

针对属于爬取范围内的条目所进行的统计，即链接两端均属于爬取范围。统计表格中，第一列、第二列为设置的限定域；对于统计的链接对象，第一列指链接的起始端，第二列指链接的目标端，第三列为该类链接的统计结果。

对于每个爬取到的条目，它必然属于限定域，查找该条目所对应的页面中是否存在链接指向限定域，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于爬取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

3.4.3 互链统计矩阵

针对属于爬取范围内的条目所进行的统计，即链接两端均属于爬取范围。统计表格中，纵向和横向为设置的限定域；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个爬取到的条目，它必然属于限定域，查找该条目所对应的页面中是否存在链接指向限定域，如不存在，则跳过；若存在，判断该链接指向的目标页面是否属于爬取范围；每查找到一条符合条件的链接，则相应的统计条目加 1。

3.5 基于爬取页面的链接统计

3.5.1 各类链接统计

针对属于爬取范围内的条目所进行的统计。统计表格中，纵向为设置的爬取范围内的条目，横向为各类统计指标。其中，KnownLinks 统计链接，该类链接指向爬取范围内的页面；UnknownLinks 统计链接，该类链接指向爬取范围外的页面；InterLinks 统计链接，该类链接指向本条目对应站点(不一定为起始 url 对应的站点)内的页面(范围为爬取范围内)；OutLinks 统计链接，该类链接指向本条目对应站点(不一定为起始 url 对应的站点)外的页面(范围为爬取范围内外)。

对于每个爬取到的条目，针对每项指标逐一统计：若条目对应的页面中存在一条指向爬取范围内页面的链接，则 KnownLinks 加 1；以此类推。

3.5.2 互链统计矩阵

针对属于爬取范围内的条目所进行的统计，即链接两端均属于爬取范围。统计表格中，纵向和横向为爬取范围内的条目；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个爬取到的条目，查找该条目所对应的页面中是否存在链接指向爬取范围内的其它页面，每查找到一条符合条件的链接，则相应的统计条目的值设置为 1(即两页面间若存在链接，则统计值设置为 1；若不存在链接，则统计值设置为 0)。

3.5.3 互链统计矩阵(去除全零行)

针对属于爬取范围内的条目所进行的统计，即链接两端均属于爬取范围。统计表格中，纵向和横向为爬取范围内的条目；对于统计的链接对象，纵向指链接的起始端，横向指链接的目标端，表格内数据为对应链接的统计结果。

对于每个爬取到的条目，查找该条目所对应的页面中是否存在链接指向爬取范围内的其它页面，每查找到一条符合条件的链接，则相应的统计条目的值设置为 1(即两页面间若存在链接，则统计值设置为 1；若不存在链接，则统计值设置为 0)。

对于统计结果，若存在某行的值全为零(即该页面内不存在指向爬取范围内的其余页面的链接)，则去除该行。以此，得到最终的互链统计矩阵。