分布式系统期末 PJ 报告

—Wikipedia 索引建立

林士翰 15307130120

1. 概述

本次 PJ 我做的是建立 Wikipedia 索引。先用 hadoop 集群计算出 TF 值(TF File)、TF 文件的倒排索引(Inverted Index)和页面索引(Page Index)，然后将索引存入 MySQL 数据库，并用 Python Flask 搭建了一个简单的服务器，提供 Web 界面的搜索功能。

2. hadoop

我使用 hadoop 集群计算得到了三份文件，分别是页面索引(Page Index)，TF 文件(TF File)，以及基于 TF 文件的倒排索引(Inverted Index)。

2.1. 页面索引(Page Index)

页面索引记录的是每个页面(page)在 XML 中的起始位置(offset)和长度(length)。

由于 hadoop 默认 Mapper 输入为按行输入，对于 XML，我们需要将 XML 切分为许多小的页面作为 Mapper 的输入，所以我们需要自定义 hadoop 的 InputFormat 类。我使用了网上的一份能够切分 XML 作为 Mapper 输入的代码[1]。

MapReduce 过程：

Mapper 的输入为每一份 XML 格式的页面，key 为该页面在 XML 中的起始位置(offset)，value 为该页面的 XML。

Mapper 的输出为每一个页面的 XML 中的起始位置(offset)和长度(length)，key=offset, value=length。

建立页面索引无需 Reducer 和 Combiner。

页面索引的大小总共为 49.4M。

2.2. TF 文件(TF File)

TF 文件记录的是每个单词在每个页面中的 TF 值，以及对应页面的标题。记录页面的标题是为了让用户在查询关键词时，能够同时看到相关文章的标题，提升用户体验。此外，对于某一个单词，包含它的页面有很多，对于页面的优先级，我在 TF 之外还考虑了标题的影响，即如果某一个单词出现在页面的标题当中，则适当提高它的权重。所以，一个单词在某个页面中的得分(scores)为：

$$\text{scores} = 0.8 \cdot \text{TF} + 0.2 \cdot \text{isTitle}$$

展示给用户搜素结果时，scores 越大的排在越前面。

XML 中的页面内容为 Wiki Markup 语法，不好统计 TF，于是我使用了 Wikiclean 这个第三方库[2]，它可以把 Wiki Markup 转化纯文本。

MapReduce 过程：

[1] http://blog.csdn.net/doegoo/article/details/50401080
[2] https://github.com/lintool/wikiclean

Mapper 的输入为每一份 XML 格式的页面，key 为该页面在 XML 中的起始位置(offset)，value 为该页面的 XML。

Mapper 的输出为每一个单词及其在该页面中的得分(Scores)、页面编号(pageid)、标题(title)。即 key 为单词，value 为一个(scores, pageid, title)的三元组。

Reducer 和 Combiner 的输入为单词及其对应的三元组列表，即 key=word，value= list[(scores, pageid, title)]。

Reducer 和 Combiner 应当对同一个单词的得分进行排序，所以输出为单词和排好序的三元组列表。

但是把该算法放到集群上运行时会一直崩溃，无法得到相应的输出。最后查出原因是单词的三元组列表有可能非常大，导致内存不足，所以需要进行修改。

利用 MapReduce 中对 key 值的排序是外部排序的特点，可以得到另一个 MapReduce 算法：

Mapper 的输入为每一份 XML 格式的页面，key 为该页面在 XML 中的起始位置(offset)，value 为该页面的 XML。

调整 Mapper 的输出格式，key 为单词和得分组成的二元组，value 为页面编号和标题组成的二元组，即 key=(word, scores)，value=(pageid, title)。

该算法无需 Reducer 和 Combiner。

利用该算法即可得到排好序的 TF 文件，大小为 19.81G。

2.3. 基于 TF 文件的倒排索引(Inverted Index)

由于 TF 文件达到 19.81G，尽管是有序的，但直接在它上面检索仍然十分耗时，因此，我们需要对 TF 文件建立索引，也就是单词的倒排索引。倒排索引记录的是每个单词在 TF 文件中的起始位置(offset)和长度(offset)。

MapReduce 过程：

Mapper 的输入为 TF 文件中的每一行，key 为该行在 TF 文件中的起始位置(offset)，value 为这一行的文本。

Mapper 的输出为这一行所代表的单词和及其起始位置、长度组成的二元组，即 key=word，value=(offset, length)

Reducer 的输入为单词及其对应的二元组列表，即 key=word，value= list[(offset, length)]。

Reducer 要求 offset 的最小值，并对 length 进行求和，所以输出为 key=word，value = $(\min\{offset\}, \sum length, count)$。count 为该单词在 TF 文件中的行数，用于方便计算$idf = \log\frac{|D|}{count+1}$。

TF 文件的倒排索引的大小总共为 73.1M。

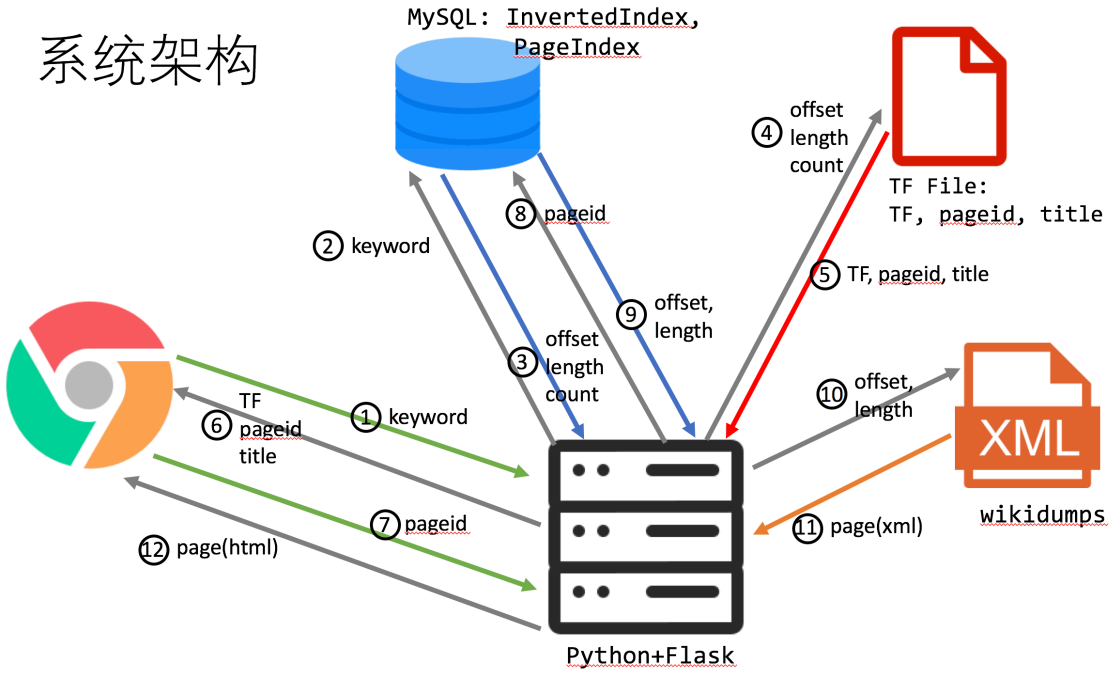
3. 服务器搭建

得到了 TF 文件和索引后，可以搭建服务器和前端来提供用户检索了。由于页面索引和 TF 文件的倒排索引较小，可以将他们存入 MySQL 当中加速检索。另外，我用 Python Flask 搭建了一个简易的服务器，使得能够 Web 界面上进行检索。

4. 总体框架

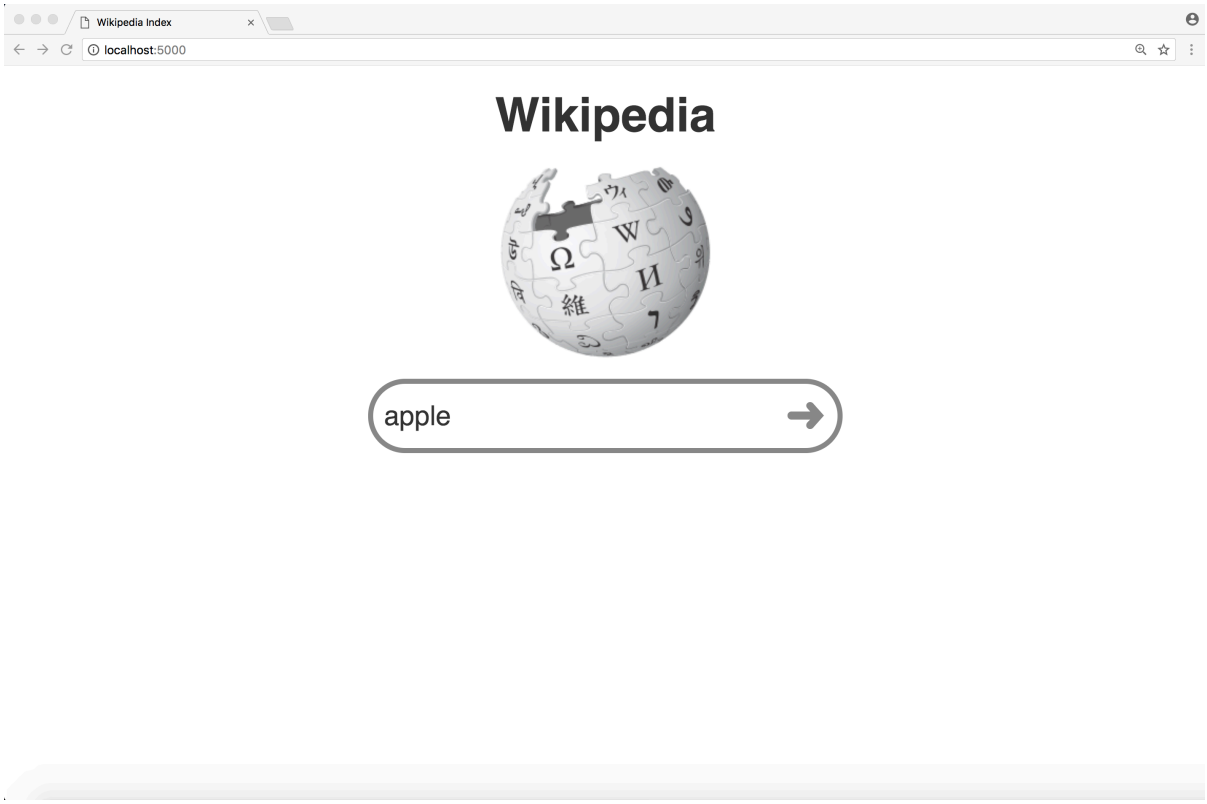
如图所示，当用户在浏览器输入一个单词进行搜索时，流程如下：

1. 浏览器发送请求给服务器，服务器得到关键字。
2. 服务器向 MySQL 查询该关键字在 TF 文件中的偏移(offset)和长度(length)，以及用于计算 idf 的 count(包含该关键字的页面数)。
3. 服务器根据 offset、length、count，从 TF 文件中读取包含该关键字的页面的信息，包括 TF 值(scores)、pageid、title。由于已经按照 scores 排好序，服务器返回前 20 条给用户。
4. 用户查看到相关的页面，点击相应页面的 pageid，浏览器发送 pageid 给服务器，请求相应的页面。
5. 服务器向 MySQL 查询该页面在 XML 文件中的偏移(offset)和长度(length)。
6. 服务器根据 offset 和 length，从 XML 中获得相应的页面，渲染成 HTML 后返回给用户。我使用了 smx 这个第三方库把 XML 中 Wiki Markup 转化为 HTML 页面，提升用户体验。

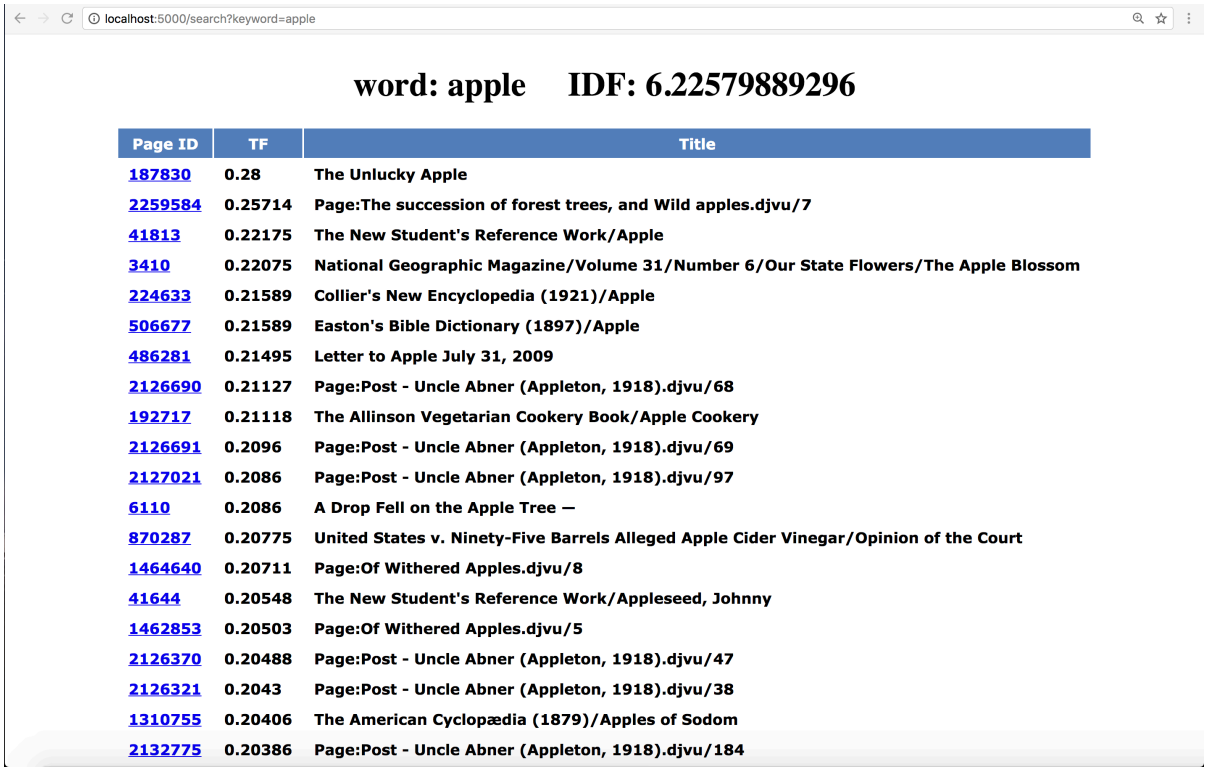


5. 效果展示

搜索页面：



相关结果页面：



word: apple IDF: 6.22579889296

Page ID	TF	Title
187830	0.28	The Unlucky Apple
2259584	0.25714	Page:The succession of forest trees, and Wild apples.djvu/7
41813	0.22175	The New Student's Reference Work/Apple
3410	0.22075	National Geographic Magazine/Volume 31/Number 6/Our State Flowers/The Apple Blossom
224633	0.21589	Collier's New Encyclopedia (1921)/Apple
506677	0.21589	Easton's Bible Dictionary (1897)/Apple
486281	0.21495	Letter to Apple July 31, 2009
2126690	0.21127	Page:Post - Uncle Abner (Appleton, 1918).djvu/68
192717	0.21118	The Allinson Vegetarian Cookery Book/Apple Cookery
2126691	0.2096	Page:Post - Uncle Abner (Appleton, 1918).djvu/69
2127021	0.2086	Page:Post - Uncle Abner (Appleton, 1918).djvu/97
6110	0.2086	A Drop Fell on the Apple Tree —
870287	0.20775	United States v. Ninety-Five Barrels Alleged Apple Cider Vinegar/Opinion of the Court
1464640	0.20711	Page:Of Withered Apples.djvu/8
41644	0.20548	The New Student's Reference Work/Appleseed, Johnny
1462853	0.20503	Page:Of Withered Apples.djvu/5
2126370	0.20488	Page:Post - Uncle Abner (Appleton, 1918).djvu/47
2126321	0.2043	Page:Post - Uncle Abner (Appleton, 1918).djvu/38
1310755	0.20406	The American Cyclopædia (1879)/Apples of Sodom
2132775	0.20386	Page:Post - Uncle Abner (Appleton, 1918).djvu/184

点击某个 PageID 后获取相应页面：

Template:NSRW

Apple, the name of a tree and of the king of gay fruits, the most important commercial pomological fruit in the world. It will grow in a variety of climates and soils; in the Old World its range is from Scandinavia south to the mountainous portions of Spain; in the New World, from Hew Brunswick to the mountains of Georgia, from British Columbia down to the mountains of Mexico. And in New Zealand and Tasmania the apple thrives. It has been in cultivation since prehistoric times. Notable reference is made to it in ancient literature; it is mentioned several times in the Bible; in the tale of Troy's fall the apple played a part; names and other evidence shows its extensive cultivation by the Romans; the folk-lore of Scandinavia and Germany abounds with stories of apple trees and golden apples.

The apple belongs to the rose family of plants, and is a native of southwestern Asia and adjacent Europe. The common apples are all modifications of a single species; while the crab apples have all been derived from another species. The number of varieties actually on sale in America during any year is not far from 1,000. North America is the greatest apple country of the world, and a full crop for the United States and Canada is said to be not less than 100,000,000 barrels.

Apples were early introduced in this country, and at first prized specially for cider. In the United States the apple is adapted to all portions save Florida, the lands immediately bordering the Gulf and the warmer localities of the southwest and Pacific coast. The most perfect apple region, Bailey considers, begins with Nova Scotia and extends to the west and southwest to Lake Michigan; other important regions are the Piedmont country of Virginia and the highlands of adjacent states, the plains region, the Ozark and Arkansas regions and the Pacific region.

While the apple thrives in a variety of soils, it reaches its best in a clay-loam. It is propagated both by budding and by grafting the sort desired on young seedling trees. Apples grown from seeds are very apt to revert to the wild type. Dread enemies of the apple are apple worm and apple scab. Spraying with poison is the means used to check their work of ruin.

There are several species of crab apple native to North America; the prairie, the wild (*Coronaria*), the narrow-leaved, and the oregon crab. The blossom of the wild crab-apple is of exquisite beauty and fragrance, and thickets of these trees now have place in many of our city parks. There is no wild flower more highly prized in this country, and for every region there is a crab-apple tree.

The common apple tree is rightly valued for its beauty as well as its utility. In the spring, when the rugged, sturdy trunk bears aloft its huge bouquet of fragrant bloom and freshest leaves, all pay homage— and here may be made declaration that beauty is excuse for the wealth of flowers, for not one tenth of the blossoms is needed to "set" all the fruit the tree could mature. Aummer orchard, too, is very attractive, and decidedly attractive is the orchard in the season of ripened fruit. In winter the spreading bare branches and leaning tree stand out in full picturesqueness.

> "Health to thee, good apple-tree,
> Well to bear pocketfulls, hatfulls,
> Peckfulls, bushel-bagfulls."

See Bailey's *Cyclopedia of American Horticulture* and Bailey s *Field Notes on Apple Culture*; Thomas: *The Book of the Apple*; McFarland: *Apples*.