资讯类新闻套图系统①

江浩亮[1,2], 左 春[1,2,3]

[1](中国科学院软件研究所 软件工程技术研发中心, 北京 100190)
[2](中国科学院大学, 北京 100190)
[3](中科软科技股份有限公司, 北京 100190)

摘 要: 考虑到图片具有对事件诠释力强, 传播便利的特点, 研究了从大量数据密集的新闻 Web 页面中自动提取数据, 并组织成套图结构展现给用户. 基于页面模板实现动态页面抽取和解析, 处理转换为对应的套图数据结构. 基于余弦相关性对来自不同网站的新闻套图数据进行去重, 并根据相应的标准, 为数据集进行评分排序. 考虑巨大的新闻数据和用户数量, 本系统基于hadoop分布式平台, 满足系统的高可扩展性. 本文将详细描述我们的系统设计和实现, 并公布在百度资讯图片栏目上的运行结果.

关键词: Web 信息提取; 动态数据集; 高可扩展性; 个性化推荐; 套图

Web Information Extraction and Knowledge Presentation System

JIANG Hao-Liang[1,2], ZUO Chun[1,2,3]

[1](Software engineering center, Institute of Software, Chinese Academy Of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100190, China)
[3](Sinosoft, Beijing 100190, China)

Abstract: Considering the picture has the futures that a strong interpretation of events and convenient disseminating, this paper studies extraction of data from a large number of news web pages, and organizational structure chart presented to the users. It achieves dynamic pages based on page template extraction and analysis, processing converted to the corresponding sets of datastructure. Based on the news cosine correlation graph data sets from different sites are differentiated, and in accordance with the appropriate standards for data collection to score sorted. This system is based on hadoop distributed platform, considering the large number of users and imgsets. This paper will describe the design and implementation of our system in detail, and report the results of running the system on Baidu news image column.

Key words: web information extraction; dynamic data set; highly scalable; personalized recommendations; imgset

网络上从不缺少新闻, 但是在过去以文字为主题的资讯新闻常常由于大段的文字表述难以勾起用户的兴趣. 相比于文字新闻, 图片新闻有其无法比拟的传播优势, 如诠释力强, 图片对时间、变易、结果具有很强的诠释效果, 可以适合不同年龄、知识层次的读者欣赏. 正因为如此, 那些包含图片的新闻, 即资讯套图(套图是指有相同事件主题的图片及简短文字描述的数据集合)正越来越受到人们的喜爱 [1]. 现在, 无论是国内国外, 各主流媒体, 如新浪、网易、百度、CNN、BBC 等, 都有自己的新闻图片频道, 这些频道是人们浏览资讯, 了解国内外大事的重要途径. 但是每个门户网站上的图片新闻存在差异, 互有补充, 用户要想尽量多地了解新闻资讯就必须到各个网站上浏览, 这为用户带来了不便, 这使得对这些门户网站图片新闻的自动挖掘汇总[2], 统一为用户展示, 减少用户的浏览成本显得尤为重要. 我们希望设计实现一个在线资讯套图系统来解决浏览成本高的问题. 该系统面临的主要问题有:

①基金项目:"核高基"重大专项(2010ZX01045-001-006)
收稿时间:2014-02-28;收到修改稿时间:2014-03-25

可扩展性: 每天主流网络媒体将产生上百万的资讯新闻数据, 并且访问我们系统的用户也将是成千上万.

项目变动: 新闻资讯具有变化频繁的特点, 可以想象在任何时候那些用户感兴趣的故事都是发生在最近几个小时, 因此我们的系统必须能满足这种快速变化的需求.

项目推荐: 每个用户的喜好各不相同, 简单的将所有新闻咨询类套图呈现给用户, 将会浪费用户大量的时间. 所以根据用户的喜好和阅读习惯, 为用户推荐其感兴趣的套图新闻是必须要解决的问题[3,4].

基于以上原因, 我们发现现有的系统设计方案无法满足我们的需求, 逼迫我们去实现更为合理的方案. 本文介绍了我们的系统的相关算法和具体实现. 本文的其余部分安排如下: 第 1 部分概述我们系统的整体设计. 第 2 部分详细介绍系统各个部分的实现. 第 3 部分分析系统在百度平台上的运行结果, 并对结果进行分析. 最后是我们的总结.

1 系统概述

本系统总体分成三部分: Web 页面监控模块(furlspider)、Web 信息抽取和组装模块(newsimgset)、个性化推荐模块(imgrecomend). 系统总体框架如图 1 所示.

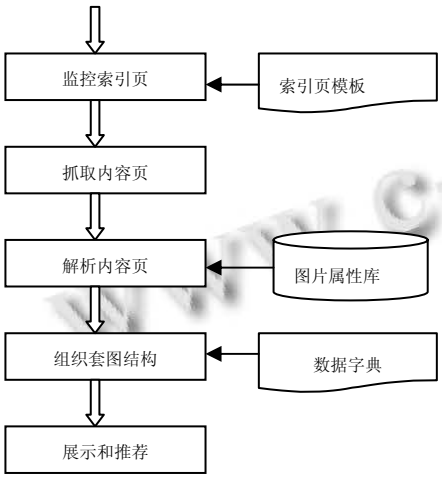

图 1 系统总体框图

Web 页面监控模块(furlspider): furlspider 模块基于配置的网站模板, 对特定的 Web 站点进行周期性的监控, 抓取新增的资讯套图页面和相关资源.

Web 信息抽取和组装模块(newsimgset): 该模块的主要功能是对抓取的套图页面进行解析, 识别翻页符, 抽取所需元素, 组装成套图数据结构, 并进行去重和评分. 最后选取评分最高的套图展示给用户.

个性化推荐模块: 基于组合访问的协同过滤推荐算法, 增加时间衰减因子, 运行于 hadoop 平台上, 保证了高可扩展性[5].

2 各模块的算法设计与实现

2.1 Web 页面监控模块(furlspider)

在时效性资讯套图挖掘系统中, 为了保证新闻套图收录的时效性和收录质量, 采用了一种定向抓取的方式. 模块的基本原理如下:

给定一批输入的图片源, 以索引页为粒度. 调度 crawler 对这些源进行定时的抓取、链接抽取, 并识别出更新的链接. 调度 crawler, 将更新的链接指向的网页抓取回来, 对页面进行简单的策略分析, 判定是否是需要收录的 fromurl, 如果是, 就将这个网页以及相关的源信息打包发送给下游进行进一步处理.

2.1.1 识别新链接

每一个索引页设置有一个权重(类似于站点评级, 从源列表读入); 整个系统设置有最大周期 Max、最小周期 Min(furlspider 配置文件读入), 根据一定计算逻辑, 初始化时先确定每一个索引页的检测周期: 权重越高的检测周期越低, 不低于 Min; 权重越低的检测周期越高, 不超过 Max.

对于输入的索引页列表, 多线程处理, 假设 M 个调度线程, N 个索引页列表, 每一个调度线程常驻内存运行, 循环处理固定的 N/M 个索引页.

处理一个索引页时, 判断当前时间与上次 check 这个索引页的时间点之差, 是否达到了这个索引页的检测周期, 如果没有达到周期间隔, 跳过该索引页, 处理下一个索引页, 如果达到了, 调度 crawler 抓取页面, 并开始 check 这个页面, check 方法: 先提取所有页面上的链接, 并根据 seed 和 banned_seed(每一个索引页有一个 seed 和 banned_seed, 从源列表读入, 是两个固定的 url 子串, 这个索引页下包含 seed 子串且不包含 banned_seed 子串的衍生链接才会收录)匹配, 筛选出有效链接. 所有的有效链接, 根据和已收录字典的对比, 进一步筛选出新链接, 放入一个全局队列中.

2.1.2 内容页抓取

模块运行多个抓取线程, 进行目标资源抓取. 抓取线程从全局队列中取出新链接, 并调度 crawler 进行抓取. 然后, 对 crawler 返回的页面进行页面类型识别, 区分内容页和索引页(newstype), 而内容页就是我们要收录的目标资源, 将内容页的网页包和相关的来源索引页信息打包成相应的数据格式发送给下游程序处理.

对于成功发送给下游的内容页链接, 将其 url 签名计算出来, 保存到已收录字典中, 用于之后判断新的页面.

2.2 Web 页面信息抽取和组装模块(newsimgset)

本模块算法流程如图 2 所示:

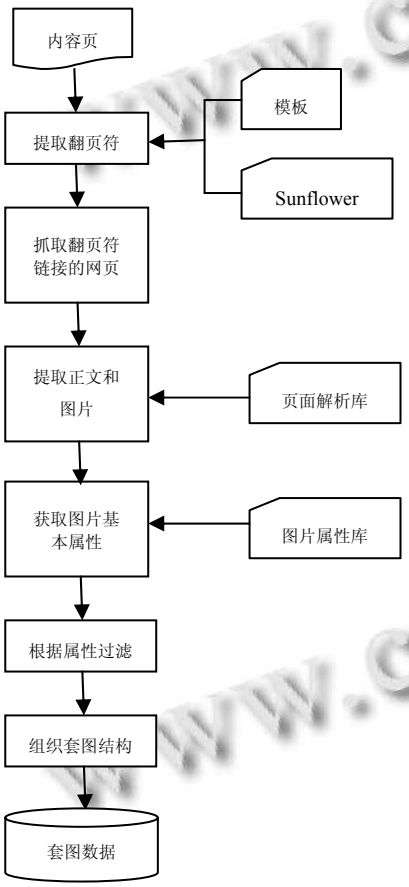

图 2 Web 信息抽取和组装流程

newsimgset 模块维护一个 ependingpool 模型接受上游连接, 获取数据包, 并将获得的数据包加入到公共数据队列中. 同时有多个工作线程, 这些工作线程操作从公共数据队列中取出数据包, 获得页面, 并使用 newshtmlparser2 页面解析抽取标题、发布时间等数据.

对于翻页的抽取和识别, 我们分为 javascript 页面和非 javascript 页面两种.

2.2.1 非 javascript 页面的翻页识别和页面解析

对于非 javascript 源衍生的数据, 识别翻页关系的策略有两种:

①该页面有本地 url pattern 模板的就按照 url pattern 模板自行模拟翻页 url(我们发现套图页面的 url 后缀有极强的规律, 数字递增或字母递增), 不知道翻页总数的情况下, 设置一个上限值, 然后轮询抓取每一个模拟的翻页 url, 直到抓取不到有效页面(死链或者 302 跳转到首页等)就跳出循环.

②本地没有存储相应的 url pattern 模板的页面, 根据 newshtmlparser2 识别的翻页符节点, 抽取出来翻页链接, 轮询抓取每一个翻页 url.

抽取到的翻页 url, 再次调用 crawler 抓取页面, 然后使用 newshtmlparser2 抽取正文和图片, 拼接到一个数组中. 对于数组中的所有图片链接, 根据图片字典中记录的重复次数, 进行广告、logo 的过滤; 然后调用 gips 进行图片处理, 根据返回的长宽, 再进行一次尺寸过滤.

最后把数组中的有效图片、正文转化成套图数据结构, 输出到本地文件中.

2.2.2 javascript 页面的翻页识别和页面解析

如果是 javascript 的源衍生来的数据, 就把 fromurl 发送给 sunflower 服务, 并且基于 fromurl 算出一个套图签名, 然后将一些套图基本信息, 例如站点、发布时间、源分类等, 写入到打上标签(套图签名 id)的本地文件, 命名为 “套图 id.head”.

同时运行多个 sunflower 工作线程以异步的方式处理 sf 返回的数据包. sunflower 服务返回的数据包, 自带有原始 fromurl、翻页总数和当前翻页序号, 使用 newshtmlparser2 进行正文和图片的抽取, 以及图片的处理(和上述图片处理类似, 图片字典过滤、gips 处理等). 根据原始 fromurl 算出套图 id, 以套图 id 和翻页序号为 tag, 生成两个本地文件:

①”套图 id.url_title.翻页序号”, 将翻页 url 和标题写入

②”套图 id.cont.翻页序号”, 将正文和有效图片写入

外围脚本控制, 对于某一个套图 id, 一旦套图 id.cont. 翻页序号的数目达到了翻页总数, 就将.head、.url_title.*、.cont.*文件合并, 输出到本地结

果文件.

2.2.3 套图去重

来源于不同网站的新闻套图可能存在重复, 所以必须对套图进行去重. 本系统采取的去重方式以新闻套图的 title 和 ct0(正文的第一句话)的余弦相关性作为去重标准.

先将套图的 title 和 ct0 进行分词, 并去掉停用词, 然后分别求得两组套图分词结果的余弦距离, 当阀值大于等于 0.3 时, 我们认为这两组套图是相同的套图, 此时必须去重, 本系统选择保留两组套图中图片数量较多的套图, 舍弃图片数量少的套图, 其原因是我们的套图都来自于定向抓取的网站, 所以图片和文字的质量较高, 都达到了我们的标准, 所以只需取数量较多的套图即可.

2.3 个性化推荐模块(imgrecomend)

个性化推荐模块的目标是根据分析登入用户的点击历史为其提供个性化推荐. 我们的系统需要接受来自世界上数以百万计的页面浏览量和用户点击. 用户的点击数有一个很大的变法范围, 对某个用户, 这个范围从 0 到 100, 甚至几千. 我们观察到在一个月的时间范围内, 新闻套图的数量达到几十万的数量级. 此外, 如前面提到的, 这些新闻报道经历不断的变化, 每分钟都有新的新闻套图被加入, 而旧的新闻套图被撤销. 因此, 我们的系统必须有较强的可扩展性, 以应付大规模的数据处理, 并且能有效地应对快速的数据变化.

我们基于组合访问的协同过滤方法[6], 并引入时间衰减因子来解决第二个问题. 对于第一个问题, 我们使用 hadoop 分布式平台, 来增强处理大规模数据的能力[6,7].

2.3.1 推荐算法的打分机制

我们的问题与 Netflix, MovieLens 等不同, 在亚马逊这样的网站, 用户给出 1-5 级的准确打分或者将购买行为看做是正面的投票, 而我们的系统没有如此明确的正面投票, 数据噪音更多[8,9].

我们采用如下模型对用户的评价进行建模:

$$r(c,s)=\frac{w_c}{w_s}-\alpha \tag{1}$$

其中, w_c 为用户 c 访问套图 s 的图片数量; w_s 为套图 s 中图片的总数; $r(c,s)$ 为用户 c 对套图 s 的评价. α 为兴趣系数, 在我们的设置中, 一个用户点击一套套图中图片数量占图片总数的比例超过 α 将被视为该用户对这篇文章的积极投票, 而小于这个比例, 则被认为是负面投票, 该比例系数为随机选取系统运行中的 10 万套实际套图资源的点击数据的中位数,该值为 0.532. 这样模型就可以有效地表示用户的正面兴趣和负面兴趣[10], 正数为正面兴趣, 负数为负面兴趣.

2.3.2 引入时间衰减因子的协同过滤

我们的套图集合变化非常迅速, 并且点击主要发生在新闻套图刚出现的几天甚至几个小时, 而之后的点击量逐渐减少, 甚至消失. 因此, 不同于常规的协同过滤算法, 这些算法多基于静态的项目集用户集. 我们的应用场景无论是用户集还是项目集都是动态变化的. 考虑到这个棘手的问题, 我们引入时间衰减因子来解决.

$$W=\frac{e^{-\lambda}\times\lambda^{t}}{t!} \tag{2}$$

其中, λ 是通过拟合实际系统中 1 万套套图的访问数据均值的变化曲线, 学习获得的, 其系数初值为 5.6, 随着系统运行时间的延长, 以及实际数据集合的变化, 通过学习获得的该值将持续变化; W 为套图 s 在发布 t 时刻时的时间衰减系数.

假设用户集为 $C=\{c_1,c_2,...,c_n\}$, 套图集为 $S=\{s_1,s_2,...s_m\}$, $C*$ 为用户 c 的邻居集, 即集合 C 减去用户 c 的子集:

$$r_{c,s}=(\bar{r_c}+k\sum_{c'\in C*}sim(c,c')\times(r_{c',s}-\bar{r_{c'}}))\times W \tag{3}$$

其中, k 为标准化因子, 此处设为 $1/\sum_{\bar{c}\in C*}|sim(c,c')|$; $r_{c',s}$ 为用户 c' 对新闻套图 s 的评价; $r_{c,s}$ 为用户 c 对新闻套图 s 的评价; $\bar{r_c}$ 为用户 c 对所有套图评价的均值; $\overline{r_{c'}}$ 为用户 c' 对所有套图评价的均值; $r_{c',s}$ 为用户 c' 对新闻套图 s 的评价; $r_{c,s}$ 为用户 c 对新闻套图 s 的评价; $sim(c,c')$ 为用户 c 和用户 c' 的相似度.

式(2)衰减因子是一种泊松分布, 其中 λ 的值会随采样数据的规模和类型不同而不同, 系统将离线迭代参数.

考虑到不同用户在不同情况下作的评价可能有不同的尺度, 式(3)进行平均归一化的操作以消除这种尺度影响, 并引入了时间衰减因子.

3 系统运行分析

本系统运行于百度图片频道资讯平台, 定向抓取包括新浪, 网易, 搜狐, 人民网等 602 个网站的套图(收录的网站数随着时间不断地增加), 每天要处理数万套套图项目数据, 系统对定向网站套图数据抓取的准确率为 99.5%, 召回率为 97.2%. 系统 7*24 小时运行, 稳定运行 6 个月, 用户访问量呈稳定增长, 6 个月用户数增加一倍, 如图 3 所示:

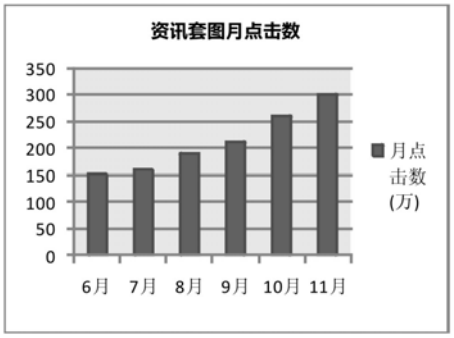

图 3 资讯新闻套图月点击数变化

在套图数据去重上, 准确率达到 95%, 召回率达到 92.3%, 由于本系统以文本相关性作为去重标准, 而套图资源具有短文本的特点, 在实际系统中, 这样的准确率和召回率是可以被接受的.

在新闻套图个性化推荐方面, 我们采用的验证数据是运行在百度资讯套图 6-11 月这 6 个月期间用户数据的一部分, 随机选取 10 万点击数据. 评价标准模型为:

$$P=\frac{q}{l} \tag{4}$$

其中, q 为向用户推荐的套图数; l 为用户点击推荐的套图并浏览超过 α 比例的套图数. 随着用户点击日志数据的不断增加, 推荐的准确率在 6 个月期间逐渐提高, 见图 4.

运行结果表明, 本系统拥有如下优点:

1) 系统基于 hadoop 分布式系统, 具有良好可扩展性, 在对大规模数据的处理上表现出了良好的性能.

2) 本系统对于无结构的 Web 页面信息抽取, 包括 javascript 动态页面的信息抽取有较高的准确率和召回率.

3) 在个性化推荐方面, 系统在足够的用户行为数据的基础上, 表现出了较好的性能. 并引入套图评价模型和时间因子, 很好地解决了特定领域的需求和问题.

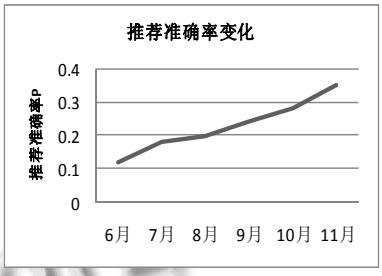

图 4 推荐准确率变化

4 结论

针对多媒体和图片新闻需求的日益增加, 本文我们提出并实现了一个可扩展的实时套图挖掘系统, 并给出了在在百度上的运行和评估结果. 该项目高流失性和大规模的数据集的特点使我们的系统有别与其它系统. 我们在 hadoop 分布式平台上, 基于 Web 页面解析技术, 动态挖掘定向网站的套图新闻数据.

在个性化推荐方面, 我们根据具体的应用场景, 提出了新的方法, 使用我们提出的套图评价模型和加入时间衰减因子的协同过滤策略, 并运行在 Mapreduce-Hadoop 框架之上, 在我们的动态数据集上进行实验. 我们在现实生活中的数据的试验结果也表明, 这种高并发的可扩展性没有损害推荐的质量.

我们的试验, 通过现实中的数据, 用户点击情况和质量度量表明我们推荐引擎的质量. 对于实际情况的评价是在百度资讯图片一段时间内大数据的一部分上实现的. 这清楚地表明了我们系统的可收缩性.

我们的方法是基于协同过滤的, 内容是无关的, 因此很容易被扩展到其它领域. 未来的发展方向是增加内容相关性模型, 以及探讨包括使用适当的学习技术, 以确定从不同的算法获得组合的权重, 并探索在组合访问统计中使用高阶的(和定向的)成本效益权衡.

参考文献

1 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展.自然科学进展,2007,19(1):35–45.

2 于满泉,陈铁睿,徐洪波.基于分块的网页信息解析器的研究与设计.计算机应用,2005,25(4):974–976.

3 许海玲.互联网推荐系统比较研究.软件学报,2009,20(2): 350–362.

4 王巧荣,赵海燕,曹健.个性化信息服务中的用户建模技术.小型微型计算机系统,2011,32(1):39–46.

5 曾理,王以群.Hadoop 集群和单机数据处理的耗时对比试验.信息科学,2009,19:55–56.

6 Das A, Datar M, Garg A, Rajaram S. Google news personalization: Scalable online collaborative filtering. Industrial Practice and Experience, 2007, 28(14): 53–63.

7 Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state–of-the-art and possible extensions. IEEE Trans. on Knowledge and Data Engineering, 2005, 17(6).

8 Zhang Z, Lu L, Liu J G, et al. Empirical analysis on a keywork based smantic system. The European Physical Journal B, 2008, 66(4): 557–561.

9 Adomavicius G, Kwon Y. Improving aggregate recommendation diversity using ranking-based techniques. Proc. of the IEEE Trans. on Knowledge and Data Engineering. Piscataway, NJ, USA, IEEE, 2011. 1–15.

10 Cremonesi P, Koren Y, Turrin R. Performance of recommender calgorithms on TopN recommendation. Recommender Systems, USA , ACM. 2010. 39–46.