

基于 LSH 和 MapReduce 的近邻模型推荐算法

陈佑雄¹, 向 阳¹, 张 骐², 潘 涛²

(1 同济大学 计算机科学与技术系, 上海 201804; 2 神华和利时信息技术有限公司, 北京 100009)

摘 要: 传统的近邻模型(k-nearest Neighborhood, KNN)是一种使用广泛的协同过滤模型,但是随着用户和项目的增加,需要计算大量用户或项之间的相似度,其时间复杂度过高. 通过结合位置敏感哈希(Locality-Sensitive Hashing, LSH)与 MapReduce,提出了一种能够在线性时间复杂度内并行计算用户或项之间相似度的近邻模型推荐算法,降低了时间和空间复杂度. 在 Tencent Weibo 数据集上进行了仿真实验,实验表明提出的模型能有效解决传统近邻模型对于大数据集时间复杂度过高的问题,显著地提高了传统近邻模型的精度和降低传统近邻模型的耗时.

关键词: 协同过滤; K-nearest Neighbor; LSH; MapReduce

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2013)12-0047-03

Nearest Neighbor Model Recommendation Algorithm Based on LSH and MapReduce

CHEN You-xiong¹, XIANG Yang¹, ZHANG Qi², PAN Tao²

(1 Department of Computer Science and Technology, Tongji University, Shanghai 201804, China;

2 Shenghua Hollysys Information Technology Limited Company, Beijing 100009, China)

Abstract: Traditional k-nearest neighborhood (KNN) model has been widely used in the recommender systems. However, with the increasing of users and items, the large scale of similarity between users or items need to be calculated and the time complexity is too high. In this paper, a nearest neighbor model recommendation algorithm combined with a locality sensitive hash (Locality-Sensitive Hashing, LSH) and MapReduce is proposed, which is a way to linear time complexity by parallel computing similarity between users or items, reducing the time and space complexity. Simulate experiments in Tencent Weibo datasets show that the proposed model can effectively solve the problem of high time complexity exists in the traditional nearest neighbor model for large data sets and significantly improve the accuracy of the traditional nearest neighbor model and reduce the time-consuming.

Key words: collaborative Filtering; K-nearest neighborhood; LSH; MapReduce

1 引言

在线社交网络取得了飞速的发展,对于庞大的社交网络平台,如何解决信息过载,向用户推荐感兴趣的其他用户,都需要推荐系统为用户提供个性化和合适的推荐或者服务. 推荐系统在电子商务系统占了很重要的一部分,另外一个应用是人们非常熟悉的个性化搜索.

一般而言,推荐系统的策略有两种:基于内容的和基于用户历史行为的^[1]. 前者需要收集额外的用户或者项资料,有时是不易获取的或者公布的,此外系统需要关于用户或者项资料的领域知识. 本文的模型是基于用户历史行为的,这种方法也叫做协同过滤(Collaborative Filtering, CF). CF 的思想是分析出用户或者项之间的关联,以便找出新的用户与项的关系^[2]. 现在很多推荐系统使用的是 CF 模型.

收稿日期: 2013-02-25; **修回日期:** 2013-04-23

基金项目: 国家自然科学基金(71171148); 国家科技计划课题(2012BAD35B01); 上海市科技创新计划(11DZ1501703); 陈家镇智慧社区和智能交通项目(11dz1210600)

CF 最常见的模型是 KNN^[3] 模型. KNN 模型的核心是兴趣相似的用户对某一项的评价会差不多, 或者比较接近的项得到的评价会差不多. 不管是基于用户还是基于项的 KNN 模型实现都很简单, 解释起来很直观.

KNN 模型需要计算用户或者项之间的相似度, 如果有 N 个集合, 求它们之间两两的相似度就需要 $N * (N-1)/2$ 次计算, 对于较小规模的数据集, 相似度的计算可以在相对较短的时间内完成, 当 N 很大时, 时间复杂度过高, 因此需要一种方法不需要遍历所有用户或者项即可以找到相似度较大的用户或者项. LSH^[4] 能够在线性的时间内将相似的用户散列到一个桶中, 从而只需要遍历用户所哈希到的桶即可以找到相似的用户, 同时通过 MapReduce 编程实现并行计算, 大幅的降低了时间和空间复杂度.

2 相关工作

本文 m 为用户数, n 为项目数, 下标 u 和 v 代表用户, i 和 j 代表项. 打分值 r_{ui} 代表用户 u 对项 i 的喜好程度, r_{ui} 值越大代表喜好程度越高. 抽象来讲, CF 模型等同于缺失值估计: 给定一个带有许多缺失值的用户项矩阵 $R_{m \times n}$, 根据已有的值来填充缺失值^[3]. 一般来讲, CF 模型分为两类: 邻居模型和潜在因子模型如 SVD 模型. 后者通过抽取出用户和项的隐藏特征来描述数据的特性. 除了 CF 之外, 还有一些传统的挖掘算法加以适当的改动也能够完成打分预测的功能. 本文用到的聚类预测模型是基于传统 K-means 模型, 文献 [3-7] 对 KNN 模型, MinHash、LSH 和 Map Reduce 做了较深入的介绍.

3 结合 LSH 和 MapReduce 的 KNN 模型

LSH 是一种快速寻找相似数据的办法, 这里引入是为了整合到 KNN 模型中, 它将能够将相似的用户以一定的概率散列到相同的同一个桶中, 其基本的预测公式如下:

$$r_{us} = \sum_{u_i} w(u, u_i) \bar{I}(u_i, s),$$

式中, $w(u, u_i)$ 是用户 u 和 u_i 的相似度, $\bar{I}(u_i, s)$ 是指示向量, 当 u_i 选择了项目 s 时取值为 1, 否则为 0.

结合基准分的预测公式:

$$r_{us} = u + b_u + b_i + \sum_{u_i} w(u, u_i) \bar{I}(u_i, s)$$

对于用户和项目很大的系统, 如腾讯微博用户, 首先寻找出每一个用户所点击的项目, 构建一个用

户项目的矩阵. 这里每个用户点击的项目以及项目的个数都很可能不一样, 因此需要利用 MinHash 的思想为用户矩阵建立一个签名矩阵, 建立签名矩阵的算法如下:

SIG(i, c) 表示 signature matrix 中的第 i 行 c 列的元素值

对 r 行:

1. Compute $h_1(r), h_2(r), \dots, h_n(r)$.

2. For each column c do the following:

(a) If c has 0 in row r , do nothing.

(b) However, if c has 1 in row r , then for each $i = 1, 2, \dots, n$ set SIG(i, c) to the smaller of the current value of SIG(i, c) and $h_i(r)$.

通过 MinHash 所建立的签名矩阵保留了原来用户项目矩阵的特征, 并且将数据变得更规范, 对于维数高的向量也起到了降维的效果.

下一步利用 LSH 算法将签名矩阵中的向量哈希到不同的桶中, 其算法如下:

Preprocessing:

1. Choose L functions $g_j, j = 1, \dots, L$, by setting $g_j = (h_1, j, h_2, j, \dots, h_k, j)$, where $h_1, j, \dots, h_k, j$ are chosen at random from the LSH family H .

2. Construct L hash tables, where, for each $j = 1, \dots, L$, the j th hash table contains the dataset points hashed using the function g_j .

Query algorithm for a query point q :

1. For each $j = 1, 2, \dots, L$

i) Retrieve the points from the bucket $g_j(q)$ in the j th hash table.

ii) For each of the retrieved point, compute the distance from q to it,

iii) (optional) Stop as soon as the number of reported points is more than L .

通过这两个步骤即可以将用户哈希到 L 个桶中, 寻找某个用户的邻居只需要遍历其所在桶的用户, 并计算相似度即可以获得.

以上是传统的 LSH 与 MinHash 的实现, 虽然能够在线性的时间内完成计算, 但如果单机运行时间复杂度还是比较高, 于是引入 MapReduce, 将 LSH 与 MinHash 的计算放到 MapReduce 的编程框架下实现. 另外, 原算法中每个用户被 hash 到 L 个桶中, 当用户量比较大时, 遍历 L 个桶的时间复杂度依然很高. 于是将 L 个哈希值进行连接操作,

最终连接的值为该用户所在桶的编号,通过连接操作达到了进一步提炼用户的效果,并且能够提高用户分类的精确度,改进算法描述如下:

```

For each input vector  $v$ 
  Map input(userid, itemids)
  choose  $d$  Hash Functions  $h_1, h_2, \dots, h_d$ 
  for each element in  $v$ 
    compute the hash value and maintain the
    min value
  the output  $v'$  is a vector in
  Signatures Matrix
  For each  $v'$ 
    choose  $L$  functions  $g_j, j = 1, \dots, L$ , by setting
     $g_j = (h_1, j, h_2, j, \dots, h_k, j)$ 
    concatenate the Hash value of  $g_j$ , the results
    represents the clusterid of each vector  $v$ 
  Map output(userid, clusterid)
  for all  $i \in I(u)$  do
     $r_{ui} \leftarrow \mu + b_u + b_i + p_u q_i^T$ 
     $e_{ui} \leftarrow r_{ui} - r_u$ 
     $b_u \leftarrow b_u + \gamma(e_{ui} - \lambda_1 b_u)$ 
     $b_i \leftarrow b_i + \gamma(e_{ui} - \lambda_1 b_i)$ 
  for all  $i \in I(u)$  do
     $r_{us} = u + b_u + b_i + \sum_{u_i} w(u, u_i) \bar{I}(u_i, s)$ 

```

4 实验分析

本文基于 Tencent Weibo(<https://www.kddcup2012.org/c/kddcup2012-track1>)数据集对改进之后的算法精度和时间复杂度进行了实验. 该数据集是由腾讯微博所提供的,里面有 200 多万用户对 6 000 多个项目的选择情况. 本文选取了 80% 作为训练集,20% 作为测试集. 本实验使用的度量指标除了耗时之外,另外一个指标就 MAP(Mean Average Precision): 单个主题的平均准确率是每篇相关文档检索出后的准确率的平均值. 主集合的平均准确率(MAP)是每个主题的平均准确率的平均值. MAP 是反映系统在全部相关文档上性能的单值指标. 系统检索出来的相关文档越靠前(rank 越高), MAP 就可能越高. 如果系统没有返回相关文档,则准确率默认为 0. 计算公式如下:

$$AvP = \frac{\sum_{r=1}^N (P(r) \times \text{rel}(r))}{M}$$

$$\text{MAP} = \frac{\sum_{q=1}^Q \text{AveP}(q)}{Q}$$

式中, M 是项目的总数.

试验中通过设置 L 函数的个数可以控制同一个桶内用户的个数,通过调节偏移量 ω 的大小可以改变能够找到邻居的数目,计算结果便更加准确,图 1 所示为是在不同的偏移量大小 ω 下的结果:实验过程中选择 $L = 2$.

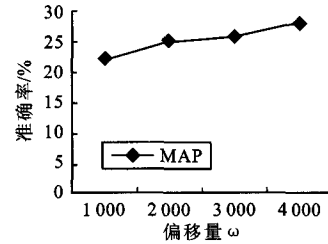


图1 不同偏移量 Y-准确率的变化

为了检验 MapReduce 算法与单机的算法的时间复杂度,图 2 中列出了 MapReduce 程序和单机程序在不同的数据量大小下运行的耗时,实验的硬件环境为: Pentium(R) Dual-Core CPU E5200, 2.50 GHz 主频, 2 GB 内存.

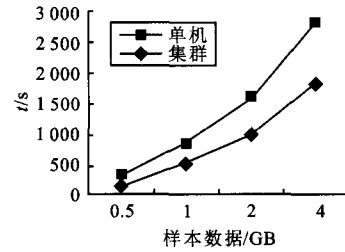


图2 不同数据量 Y 运行耗时

5 结束语

本文提出了利用 LSH 算法对用户快速聚类并计算用户之间相似度的方法,解决了传统的 KNN 模型需要遍历所有用户计算相似度所带来的高时间复杂度的问题. 通过 MAP 和耗时分析,有效地验证了算法的性能,说明了算法能够很好地解决对大规模数据集系统推荐耗时间空间过大的问题. 然而对于用户推荐,需要结合用户的背景信息和其他一些信息才能做出更准确的推荐,所以在后续中要考虑加入其他的因素. 此外,本文主要考虑了 LSH 和 KNN 模型的结合,可以考虑结合一些其他的模型如矩阵奇异值分解(Singular Value Decomposition, SVD), 概率潜在语义分析(probabilistic Latent Semantic Analysis, pLSA)等,这些都是以后研究的主要工作内容.

(下转第 53 页)

- CMOS VCO and frequency divider design for quadrature signal generation[C]// Microwave Symposium Digest. Taipei, IEEE, 2004(3):1493-1496.
- [8] Shi Chunqi, Zhang Runxi, Chen Lei, et al. A low Noise VCO with quadrature prescaler for UHF RFID reader[C]// Networks Security, Wireless Communication and Trusted Computing. Wuhan, 2009 (1): 357-360.
- [9] Zhang X, Yan Y, Du Z, et al. Wideband RC polyphase network for IEEE 802.11a quadrature signal generation[J]. Electronics Letters, 2010, 46(10): 730-732.
- [10] 张新旺. 适用于无限局域网及数字电视收发机的频率综合器设计[D]. 北京:中国科学院研究生院,2010.
- [11] Jayaraman K, Khan Q A, Chiang P. Design and analysis of 1-60GHz, RF CMOS peak detectors for LNA calibration[C]//VLSI Design, Automation and Test. Hsinchu, 2009:311-314.
- [12] Alberto V G, Radhika V, Jose S M. A broadband CMOS amplitude detector for on-chip RF measurements[J]. IEEE Transactions on Instrumentation and Measurement, 2008, 57(7):1470-1477.
- [13] Galal S H, Ragaie H F, Tawfik M S. RC sequence asymmetric polyphase networks for RF integrated transceivers[J]. IEEE Transactions on Circuits Systems, 2000(46):8-27.
- [14] Xie J, Do M A, Yeo K S, et al. Analysis and design of RC polyphase network for quadrature signal generation in the 2.45 GHz ISM band[C]// International Symposium on International Circuits. Singapore, 2007:470-473.

作者简介:

刘荣江 男,(1984—),博士研究生.研究方向为射频集成电路设计.

郭桂良 男,(1981—),博士.研究方向为射频集成电路设计.

阎跃鹏 男,(1963—),研究员,博士生导师.研究方向为集成电路与系统设计.

(上接第49页)

参考文献:

- [1] Das A, Datar M, Garg A. Google news personalization: scalable online collaborative filtering[C]// ACM 2007 Article. Bibliometrics Data Bibliometrics. New York, 2007:271-281.
- [2] Koren Y. Factor in the neighbors: scalable and accurate collaborative filtering[J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(1):1-11.
- [3] Bell R, Koren Y. Improved neighborhood-based collaborative filtering[C]// KDD-Cup and Workshop. New York:ACM press, 2007.
- [4] Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions [C]//Proceedings of the twentieth annual symposium on Computational geometry. New York, 2004: 253-262.
- [5] Chen P L, Chen T, Chen Y N, et al. A linear ensemble of individual and blended models for music rating prediction [C]// KDD-Cup and Workshop. New York:ACM press, 2011.
- [6] Rajaraman A, Ullman J. Mining of massive datasets [M]. USA:Stanford Univ,2011:62-65.
- [7] Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions [J]. Communications of the ACM - 50th anniversary, 2008 (51):117-122.

作者简介:

陈佑雄 男,(1989—),硕士研究生.研究方向为数据库、数据挖掘、云计算.

向阳 男,(1962—),教授.研究方向为数据库、数据挖掘、数据仓库、电子商务等.

张骐 男,(1964—),硕士.研究方向为信息管理科学.

潘涛 男,(1975—),博士,助理教授.研究方向为矿井移动通信、煤矿电磁兼容.