文章编号：1007-1423(2013)25-0020-04　　DOI：10.3969/j.issn.1007-1423.2013.25.005

# 使用 Hadoop 实现应用商店中的相关推荐模型

周吉寅[1]，陈　媛[2]，姚　晨[2]，冯　翱[2]

（1. 四川大学计算机学院，成都　610065；2. 联想成都研究院，成都　610065）

**摘　要：** 在应用下载网站上，当用户浏览或下载某应用时，说明他对这个应用产生一定兴趣，如果网站显示和当前应用类似的其他应用，那么可以增加用户继续下载应用的概率。将基于行为推荐和基于内容推荐结合成推荐模型，并基于 Hadoop 进行实现，可以处理较大量的数据并得到较优的推荐结果。

**关键词：** 关联推荐；余弦相似度；置信度；提升度

## 0　引　言

在电子商务网站中，为了增加应用下载量，通常会针对应用进行相关推荐，传统的相关推荐通常是基于行为数据或者基于内容相似性[1]来进行推荐，推荐算法通常采用协同过滤算法[2]，然而协同过滤算法承受不了过大的数据集[3]。本文采用了关联规则的思想，使用修正后的置信度和提升度来计算应用之间的相似性，在此基础上考虑应用本身信息。在实现过程中，使用 Hadoop 中的 MapReduce 进行数据分析，能够处理较大的数据集，提高处理效率。

## 1　相关推荐系统模型及算法

在应用商店中，当用户浏览某一应用时，系统针对该应用对用户推荐类似的应用，以此增加用户同时下载其他应用的几率。该系统采用关联规则中置信度和提升度的思想来设计基于行为数据的初步模型，即根据用户下载日志来得到应用的相似度。相似度越高，用户同时下载推荐应用的几率越大[4]。在这个初步模型的基础上，我们增加应用本身的属性特征来建立正式的应用模型，使用余弦公式来计算应用之间的相似度，以提高推荐准确率。

### 1.1　基于行为数据的初步模型

如果两个应用被用户同时下载，则认为这两个应用相似，同时下载这两个应用的用户越多，应用间相似度越大。根据分析用户下载日志可以找出所有与应用同时下载的应用，计算出它们之间的相似度，向用户推荐与相似度最高的 n 个应用。为了计算应用之间的相似度，我们采用了关联规则中的两个重要概念：置信度和提升度。在应用推荐系统中，置信度可以理解为下载了应用之后，继续下载应用的概率。提升度可以理解为下载了应用 $A$ 后，对下载应用 $A'$ 概率的提升比例。根据关联规则的思想，我们设计了初步的计算相关推荐相似度公式：

$$score(A'|A)=\frac{P(A'|A)}{\left(\frac{c+count(A')}{\sum_{i=1}^{n}count(A_i)}\right)^{\alpha}}$$

其中 $count(A)$ 为下载过应用 $A$ 的用户数。为了平衡流行度对于推荐结果的影响，避免系统频繁推荐流行应用，均衡置信度和提升度，我们设计了常数 c 和 α。c 为一个下载量的基本值，取非负整数，α 为根据应用流行度的衰减系数，取 0-1 之间的值。当 α=0 时，该

收稿日期：2013-07-11　　修稿日期：2013-08-11

作者简介：周吉寅(1989-)，女，重庆合川人，硕士研究生，研究方向为信息检索

公式为置信度公式，当 α=1，c=0 时，该公式为提升度公式。$score(A'|A)$ 为对应用 $A$ 推荐应用 $A'$ 的得分，即应用 $A$ 和 $A'$ 之间的相似度。

### 1.2 应用正式模型

经一段时间的实验，证明基于行为数据的模型能够较好的工作，但仍有以下不足：

(1)冷启动，即完全没有下载记录的时候，不管应用之间是否具有相似度，推荐结果都固定设置为流行度最高的应用。这种结果是我们不希望得到的。

(2)对于下载量较小的应用，推荐结果比较随机。

(3)假设从用户角度来说，$A$ 和 $A'$ 是两个不相关的应用，但是它们都非常流行，同时下载它们的人非常多，那么它们之间的相似度会很大，这样的推荐结果对于用户来说会显得不直观。

针对以上问题，我们在初步模型上增加了基于应用本身的属性作为特征分量，例如“关键词”、“类别”、“开发者”、“适用人群”，等等，即如果两个应用没有被任何用户同时下载，但是它们具备相同应用属性(例如关键词相同)，那么我们也认为这两个应用具有相似性。在此基础上再结合初步模型的结果，使用余弦公式来计算最终的相似度。

①特征分量的分量值计算方法

特征分量用来进行应用匹配，如果两个应用具备了相同的特征分量则配对成功。应用根据分量的标识进行配对，在这个分量上的相似度使用分量值表示。分量值计算采用修正后的 TF-IDF 算法。

②相似度计算方法

在进行相似度计算时，先分别计算对应特征的相似度，再将各分量值乘以对应权重相加，得到最终相似度。

例如，应用 $A$ 和 $A'$ 相似度计算方法为：

$$sim(\vec{A},\vec{A'})=\sum_{i=1}^{n} w_i sim_i(\vec{A},\vec{A'})$$

其中：

$$sim_i(\vec{A},\vec{A'})=\frac{\vec{A}\cdot\vec{A'}}{|\vec{A}|\cdot|\vec{A'}|}$$

即分量值。$w_i$ 是各个分量的权重。

## 2 使用 Hadoop 实现模型

在实际应用中，模型需要挖掘的数据以兆记，在某主流应用商店，具有数十万应用和千万级的用户。如果采用普通读取文件的方法，消耗的时间会超出用户期望。Hadoop 是一个开源的分布式系统基础架构，使得用户可以在不了解分布式底层细节的情况下，开发分布式应用程序，充分利用集群的威力实现高速运算和存储。Hadoop 尤其适合大数据的分析与挖掘[5]，所以我们采用 Hadoop 来实现该系统。

### 2.1 实现初步模型

我们使用初步模型来生成行为得分。具体操作如图 1。

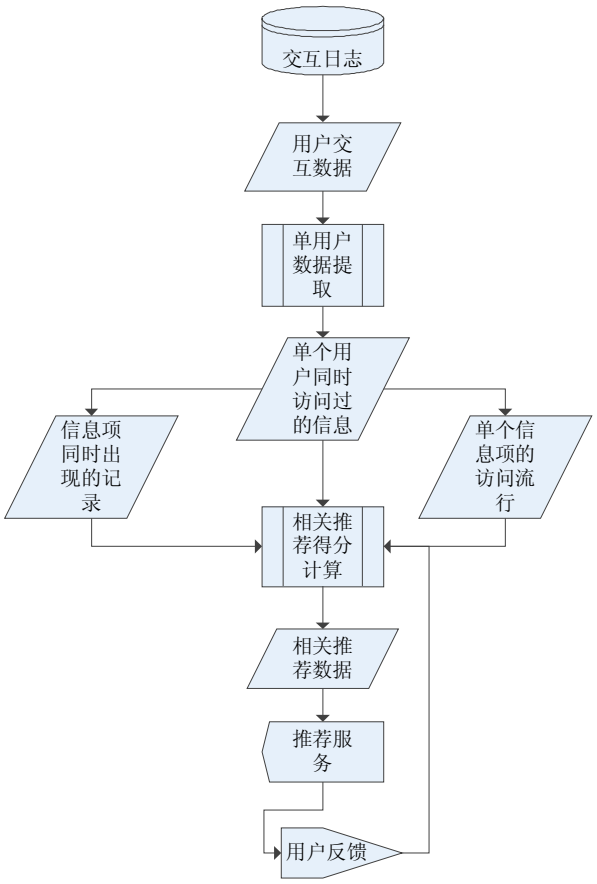


图 1 初步模型实现流程

### 2.2 实现正式模型

正式模型中加入应用本身属性，具体操作如图 2。

## 3 结果评测

### 3.1 评测方法

NDCG(Normalized Discounted Cumulative Gain)是一种对信息检索有效性的度量，它有两个假设[6]：

(1)强相关的文档出现在结果列表越靠前越有用。

(2)强相关文档比弱相关文档有用,比不相关文档有用。

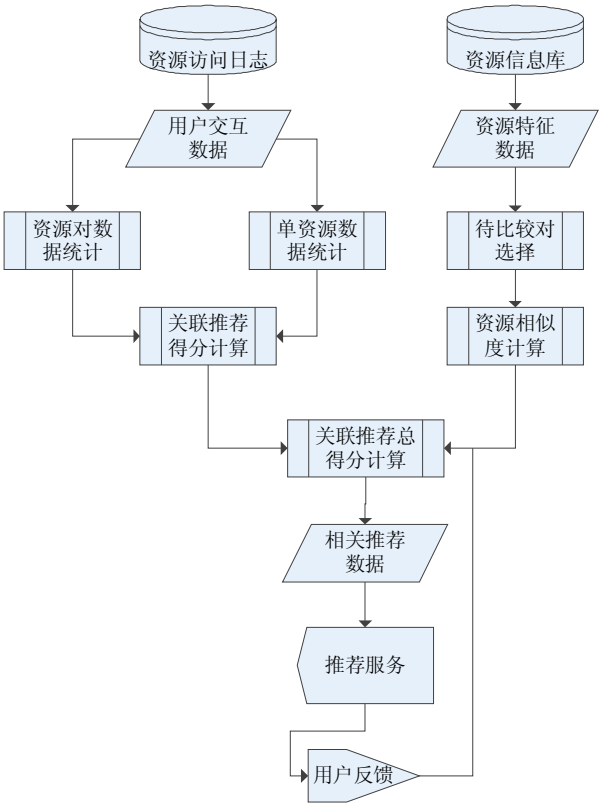


图2 正式模型实现流程图

在计算出应用的推荐序列之后,我们随机取出一组推荐结果,对每个推荐结果人工标注评分,然后对每一次推荐结果计算 NDCG 值,最后将所有 NDCG 值进行平均得到推荐结果总得分。NDCG 值越高,说明推荐结果质量越高。

### 3.2 评测结果

我们选取了一个月的下载记录作为计算数据对推荐结果计算 NDCG 值,评测结果如图 3。

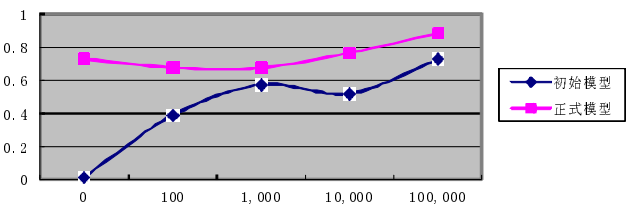


图 3 推荐结果评测

## 4 结 语

对于下载量较大的应用,初步模型的推荐结果已经较让人满意,通过正式模型的修正,能够改善频繁推荐流行应用的缺点,对推荐结果有较小提高。

对于下载量较小的应用,初步模型的推荐结果比较随机,正式模型使用应用本身信息进行修正,对推荐结果有较大提高。

对于没有下载量的应用,初步模型使用下载排行进行推荐,即推荐下载量最大的应用。这样的推荐虽然在少数情况下能带来一些下载量,但是对用户来说这样的推荐明显是不合理的。正式模型使用特征分量进行匹配产生推荐,对推荐结果质量有显著提升。

经过一段时间的运行,证明该算法可以较快得到推荐结果,并且推荐结果较优,重复较少。

### 参考文献

[1]Asim Ansari, Skander Essegaier, Rajeev Kohli. Internet Recommendation Systems [J]. Journal of Marketing Research, 2000,37(3):363~375

[2]Weiyang Lin, Sergio A. Alvarez, Carolina Ruiz. Collaborative Recommendation via Adaptive Association Rule Mining[D]. Worcester Polytechnic Institute: Dept. of Computer Science, 2000

[3]IEEE Computer Society, Technical Committee on Computational Intelligence, Web Intelligence Consortium,et al. Proceedings of the 2003 IEEEIWIC International Conference on Web Intelligence[C]. Choonho Kim, Juntae Kim. A Recommendation Algorithm Using Multi-Level Association Rules. Canada: Jiming Liu,2003:524~527

[4]Christopher D Manning, Prabhakar Raghavan, Hinrich Schtze. Introduction to Information Retrieval[D]. New York:Cambridge University Press, 2008

[5]Tom White.Hadoop 权威指南(中文第二版)[M]. 北京:清华大学出版社,2011:4~15

[6]Wikipedia. Discounted Cumulative Gain[EB/OL]. (2013-5-13).[2013-5-17]..http://en.wikipedia.org/wiki/Discounted_cumulative_gain

# Implementation of Associative Commendation Model of Application Store by Using Hadoop

ZHOU Ji-yin[1] , CHEN Yuan[2] , YAO Chen[2] , FENG Ao[2]

(1. College of Computer Science, Sichuan University, Chengdu 610065; 2. Lenovo Inst. of Chengdu, Chengdu 610065)

Abstract: When users are browsing or downloading one application on the application downloading website, it indicates that they are interested in the application. If there has other similar applications displayed for the user, there is a big possibility that the he will download them as well. A recommendation model which combined by the user-based and item-based recommendation, and implement it using the map and reduce method in Hadoop can process the big data and get the satisfying results.

Keywords: Associate Recommendation; Cosine Similarity; Confidence; Lift

(上接第 6 页)

# Evolutionary Analysis of Dynamical Systems of Travel Behavioral Decision-Making

LI Zhuo-Jun

(Department of Information Engineering, Wuhan Business University, Wuhan 430056)

Abstract: The road traffic flow evolutionary patterns of metropolitan areas evolve slowly through a complex multi-dimensional travel decision-making behavior (including travel mode, departure time and route choice joint decision-making). Aims at the general travel behavioral decision-making process, proposes a novel dynamical systems formulation of the traffic assignment problem using evolutionary game theory. The assumptions on drivers' behavior in multi-dimensional travel choice are supposed to be fairly general and reasonable. And the stable properties of this dynamical system on its equilibrium points are investigated using Lyapunov method in a general network. It shows that the evolutionary dynamical system exist only one solution on the condition that the traveler population satisfies some hypotheses which individual's trip payoff satisfy some constraint conditions. These mean that there maybe exist inherent motive power which drive the traffic flow evolve to some stable patterns from long run view point. It can improve our understandings to urban traffic flow evolution process and provide significant reference for relevant management section.

Keywords: Multi-Dimensional Travel Choice; Travel Behavioral Decision-Making; Evolutionary Game; Dynamics System Model; Stability Analysis

# 使用Hadoop实现应用商店中的相关推荐模型



作者： 周吉寅， 陈媛， 姚晨， 冯翱， ZHOU Ji-yin， CHEN Yuan， YAO Chen， FENG Ao

作者单位： 周吉寅,ZHOU Ji-yin(四川大学计算机学院,成都,610065)， 陈媛,姚晨,冯翱,CHEN Yuan,YAO Chen,FENG Ao(联想成都研究院,成都,610065)

刊名： 现代计算机（专业版）

英文刊名： Modern Computer

年，卷(期)： 2013(17)

## 参考文献(6条)

1. Asim Ansari;Skander Essegaier;Rajeev Kohli Internet Rec-ommendation Systems 2000(03)

2. Weiyang Lin;Sergio A. Alvarez;Carolina Ruiz Collaborative Recommendation via Adaptive Association Rule Mining 2000

3. IEEE Computer Society;Technical Committee on Computa-tional Intelligence;Web Intelligence Consortium Pro-ceedings of the 2003 IEEEIWIC International Conference on Web Intelligence 2003

4. Christopher D Manning;Prabhakar Raghavan;Hinrich Schtze Introduction to Information Retrieval 2008

5. Tom White Hadoop权威指南(中文第二版) 2011

6. Wikipedia Discounted Cumulative Gain 2013