

一种改进的基于项目语义的推荐算法

雷 鸣, 朱 明

(中国科学技术大学 信息科学与技术学院, 安徽 合肥 230022)

摘 要: 针对传统的基于项目语义的推荐算法计算项目相似度不够精确的情况, 提出了一种改进的基于项目语义的推荐方法, 通过挖掘数据集, 提取内容特征, 利用概率统计模型, 构建物品的特征向量, 精确地计算了项目之间的相似度, 并且与基于项目语义的推荐算法的推荐结果相结合, 不但比一般的基于项目评分的协同过滤推荐算法以及基于项目语义的推荐算法的推荐效果提高不少, 而且在评分用户数量较少的情况下, 进一步提高推荐精度。

关键词: 项目语义; 概率统计模型; 特征向量; 项目相似度

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2016)12-0093-05

An Improved Recommendation Algorithm Based on Item Semantics

LEI Ming, ZHU Ming

(School of Information Science and Technology, USTC, Hefei 230022, China)

Abstract: Compared to the traditional recommendation algorithm based on item semantic with the precision of similarity calculated not high, we proposes an improved recommendation algorithm based on item semantics, through mining data sets, extracting content features, using probabilistic model and building feature vectors of items precisely, it calculates the similarity accurately between the items and combines with the results of recommendation algorithm based on semantic recommendation, not only its results is much better than the item-score recommendation algorithm for collaborative filtering and recommendation algorithm based on item semantics, but also it improves on the precision in the less users with scores.

Key words: item semantics; probabilistic model; feature vectors; item similarity

1 引言

基于项目语义的推荐算法是采用一种基于项目之间的语义相似度度量, 计算项目之间的语义相似度, 然后预测用户对待推荐项目的评分. 传统的基于项目的推荐算法利用项目评分进行推荐, 基于领域内容的推荐算法是利用语义计算项目之间的相似度, 对此我们提出了一种利用概率统计模型进行特征向量的计算, 通过挑选合适的内容特征, 以多维度的方式建立内容特征空间, 从而更加准确地计算项目的相似度. 通过预测评分, 为用户推荐他喜欢的项目所相似的项目, 并且与基于项目语义的推荐算法的预测评分结果相结合, 挑选出分数靠前的项目, 然

后完成推荐.

2 基于项目语义的领域内容推荐算法

基于项目语义的领域内容推荐算法, 一般是以项目的信息为基本单位来计算项目之间的相似度, 例如电影的导演、演员或者简介等, 项目的信息多以文本为主, 而且这些信息多以一些关键词加以描述, 文本是以关键词为单位组成的集合, 所以计算项目之间的相似度, 要以计算关键词的相似度为基础的.

词语相似度的计算方法大体上可以分为两类: 一类是利用大规模的语料进行统计, 它是将词语的上下文信息的概率分布作为词语语义相似度计算的主要指标; 一类是根据本体知识来计算, 是按照概念

间结构层次关系组织的语义词典方法,按照概念之间的关系来计算词语的相似度^[1]. 我们这里采用的是刘群等人利用知网为语义词典计算的词语相似度,用词语之间的距离来度量词语的语义相似度,它的思想是将词语的义原看作是一个层次结构的结点,词语之间的距离可以由义原在层次结构的路径长度来表示,基于 Hownet 的语义词汇相似度计算公式为

$$\text{Sim}(W_1, W_2) = \frac{\gamma}{\text{Dis}(W_1, W_2) + \gamma} \tag{1}$$

式中,Dis(W_1, W_2)代表词 W_1 和 W_2 之间的距离,其中 γ 是一个可调节的参数,默认值为1.6. 我们可以设定一个阈值 β ,当 $\text{Sim}(W_1, W_2)$ 的值大于某个阈值 β 时,就表示这两个词语语义相似性比较强.

根据上述提出的词语语义相似度的方法,本文还利用了机器学习中的两个簇之间的组平均值的方法^[2],然后提出将项目语义相似度和用户的评分相结合的方法,预测用户对于待推荐项目的评分大小.

首先我们需要进行分词预处理,这里采用结巴分词工具,因为本章实验的项目是旅游景点,描述景点的关键词是从景点的名称和景点的简介中产生,而这些关键词一般是名词或者名词短语,所以对它进行分词的同时,需要进行词性的标注,对满足要求的关键词进行保留,构成项目的关键词集合.

接下来我们需要计算项目之间的语义相似度,假如项目 U 和 V 的关键词集合分别为 $\{u_1, u_2, \dots, u_n\}$ 和 $\{v_1, v_2, \dots, v_m\}$,那么项目 U 和 V 之间的语义相似度为

$$\chi_{UV} = \sum_{i=1}^m \sum_{j=1}^n \frac{\text{Sim}(u_j, v_i)}{m \times n} \tag{2}$$

式中,设 $\text{Sim}(u_j, v_i) > \theta, \theta \in (0, 1), \theta$ 表示一个阈值,它将相似度低的关键词组合去掉,这里我们令 $\theta = 0.6$,用以提高项目之间的语义相似度. 我们还可以根据公式产生项目 U 和 V 之间的语义相似度矩阵,如下:

$$[\chi_{uv}]_n = \begin{bmatrix} 1 & 0.61 & 0.78 & \dots \\ 0.63 & 1 & 0.91 & \dots \\ 0.85 & 0.88 & 0.66 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

3 改进的基于项目语义的推荐算法

3.1 领域内容特征表示

本文采用的是携程网的景点信息数据,其中包括景点的名称、景点的一级类目、景点的地址、景点

的简介、用户对景点的评分等. 我们利用基于 Scrapy 的爬虫框架,爬取了携程网 8 652 个景点的名称、地址和简介,以及 956 个用户对于 8 652 个景点的评分,大约 50 000 个. 这些旅游景点包括秀山峻岭、大江湖泊、特色街区、农家乐、游乐园、温泉、动物园、植物园、公园、古镇、博物馆、展馆展览、度假村、民俗体验、海滩、古代园林、采摘、郊游、地貌奇观、养生会所 20 个类别,所有景点信息存储在 Mysql 数据表里面,作为本文算法实验的基础数据.

对于文本内容而言,里面的文字信息是计算机无法识别的形式,我们需要将实际的问题转化为计算机可以识别的形式,对于景点信息而言,它最重要的一个属性就是名称,景点名称一般比较短,属于文本内容,最常用的文本内容特征表示方法是基于向量空间模型(Vector Space Model, VSM).

对于空间向量模型,需要一个种子特征字典,我们可以根据种子特征字典的长度来定义空间向量的维度,可以从景点数据集中产生这个种子特征字典,也可以人为设定,例如图 1 中的景点名称金丝雀温泉养生度假村,种子特征字典就是[金, 黄山, 温泉, 养生, 酒店, 七天, 风景, ..., 度假村],对应的特征向量为[1 0 1 1 0 1 0 1 1].

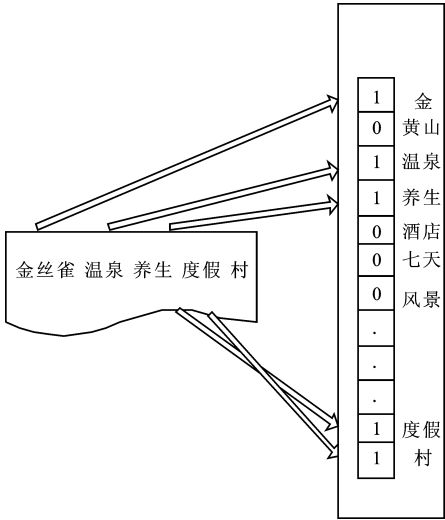


图 1 文本空间向量举例

首先对基础景点信息进行处理,使用分词工具结巴分词对数据进行分词处理和词性标注,对标注为名词的词语进行保留,作为种子特征字典的预选集.

对于种子特征字典的预选集中的每个词,我们作为表征文本的预选,需要计算每个预选特征的权值,文本特征值的计算方法有很多,比较常用的就是

布尔值和基于词频两种,利用这两种方法都可以得到文本的特征向量,利用文本特征向量组成文本数据集的布尔矩阵或者是词频矩阵.本文主要利用的数据是景点名称和景点简介,因为景点信息一般为短文本,分词后出现的词语一般出现的频率不高,所以综合考虑,选用布尔值作为特征值的计算.

3.2 领域内容特征选择

根据上节的步骤,我们对所有文本数据集处理以后,会得到一个种子特征字典集,但是我们发现,这个字典集合里面有很多词对于要表征文本来讲,是不重要的,没有多大的帮助,例如金、丝雀、温泉、养生、度假、村中,金和丝雀两个词对于判断景点名称的类别属性没有任何作用,需要删除.

对于文本分类而言,一般会面临着高维度特征空间的问题,高维度的特征空间不但会增加分类的时间复杂度和空间复杂度,而且还会降低分类的准确率.解决高维度问题的方法有很多,其中特征选择就是其中一个,特征选择的思路就是通过删除一些所含信息比较少,对分类结果影响小的特征,选取其中对分类能力比较强的特征,来减小特征维度,提高分类的准确率^[3].这里我们采用的是信息增益的特征选择方法,通过计算每个特征词的 IG 值,选择 IG 值大小排名靠前的前 N 个特征词(前 20%),我们选取排名前 50 的特征词,这些特征词与景点类别具有较强的关联性,我们也可以称它们为种子特征词典,记作 $T = \{t_1, t_2, \dots, t_{|V|}\}$,这里 $|V| = 50$.

这里使用了基于统计的特征选择方法,也可以使用基于领域知识的规则方法,因为本文是针对短文本的特征选择问题,短文本有时可能只包含景点的名称,景点名称一般是名称核心词和类别属性词的组合,而名称核心词对于判断景点的类别标签没有帮助,而且有时会起到误导的作用,金丝雀温泉养生温泉度假村,金丝雀是它的名称核心词,我们在对这个景点选取类别标签时,很可能将它归为动植物的类别,从而得到错误结论.在我们得到类别特征词之后,可以对景点名称中的名称核心词进行去除,保留剩余部分,实验表明这种方法的评测准确率约有 4% 的提高,缺点是可能会降低了算法覆盖率.

3.3 基于 Bayes 的概率模型构建

文本分类算法一般分为判别型和生成型,常见的算法有 K 近邻(KNN)、贝叶斯(Bayes)、支持向量机(Support Vector Machine, SVM)、逻辑回归(Logistic Regression, LR)以及神经网络算法等.这里采用的是基于朴素贝叶斯的分类算法,因为贝叶斯

算法可以发现数据之间的因果关系,特别是对数据量不大的数据集分类时,时间花销不是很大,并且可以综合先验信息和样本信息.

已知定义类别集合为 $C = \{c_1, c_2, \dots, c_{|C|}\}$,训练的文本文档集为 $D = \{d_1, d_2, \dots, d_{|D|}\}$,由上节得到的种子特征集为 $T = \{t_1, t_2, \dots, t_{|V|}\}$,对于文档 d 可以由向量表示为 $d = \{w_1, w_2, \dots, w_{|V|}\}$,贝叶斯分类器预测文档 d 属于哪个类别的依据是通过估计条件概率 $P(c_i | d)$ 来确定的,为了减小分类的误差,一般我们会将文档 d 分到 $P(c_i | d)$ 较大的类别中 c_i 去,一开始我们并不知道 $P(c_i | d)$,但是我们可以根据贝叶斯公式进行计算^[1]:

$$P(c_i | d) = \frac{P(c_i)P(d | c_i)}{P(d)} \quad (3)$$

这里可以通过计算 $P(c_i)$ 、 $P(d | c_i)$ 和 $P(d)$ 的值对 $P(c_i | d)$ 进行估计,其中 $P(c_i)$ 表示类别为 c_i 的文档所占的比例, $P(d)$ 表示从训练文本数据集中随机抽取一个文档是概率,是个常数,计算的时候常将它省略,而对 $P(d | c_i)$ 进行计算,因为文档 d 可以由向量表示,所以我们可以将它分解为

$$P(d | c_i) = \prod_{k=1}^n P(w_k | c_i) \quad (4)$$

将式(4)带入式(3)中,可得最终的 $P(c_i | d)$ 的计算公式:

$$P(c_i | d) = \frac{P(c_i) \cdot \prod_{k=1}^n P(w_k | c_i)}{P(d)} \quad (5)$$

由于我们通过计算文档 d 属于不同的类别 c_i 的概率,从而概率值最大的类别作为文档 d 所属的类别.故我们有公式:

$$\hat{c}(d) = \operatorname{argmax}\{p(c_i) \cdot p(d | c_i)\} \quad (6)$$

贝叶斯分类算法是一种有监督的学习方法,这里我们采用的是朴素贝叶斯算法,朴素指的是假定文本中的特征是相互独立的,和上下文没有关系.它有两种常见的模型:一种是伯努利模型(Bernouli Model),一种是多项式模型(Multinomial Model),出于简便运算考虑,我们这里采用伯努利模型.

伯努利模型又称为二项独立模型,是我们经常使用的一种模型^[4-5],它是使用一种二值向量来表征一个文档,已知种子特征集为 $T = \{t_1, t_2, \dots, t_{|V|}\}$,对于文档 d 有 $d = \{w_1, w_2, \dots, w_{|V|}\}$, $w_k \in \{0, 1\}$,其中 $|V|$ 表示种子特征集的长度,假如 $w_k = 1$ 则代表的是种子特征 t_k 在文档 d 中出现,反之则代表不出现,那么 P_{ki} 表示 $P(w_k=1 | c_i)$,那么对于已

知类别 C_i , 特征 w_k 的先验条件概率为

$$P(w_k | c_i) = P_{ki}^{w_k} (1 - P_{ki})^{1-w_k} \quad (7)$$

$$= \left(\frac{P_{ki}}{1 - P_{ki}} \right)^{w_k} (1 - P_{ki}) \quad (8)$$

对于文档 d 而言, 它的特征向量 w_k 和 w_1 是相互独立的, 根据式(5)、(6) 和(8), 根据最大似然估计的原理得出它们的估计值, 对于训练文本集 $D = \{d_1, d_2, \dots, d_{|D|}\}$, 可以得出类别 c_i 的概率:

$$P(c_i) = \frac{n_i}{|D|} \quad (9)$$

式中, n_i 表示训练文本集中类别为 c_i 的文档数量。

对于给定类别 c_i 向量 w_k 的条件概率, 假如一个样本中的特征在某条件下没有出现, 那么会认为它产生的概率为 0, 这是不科学的, 所以我们对这种问题的解决方法是加 1 平滑:

$$P(w_k = 1 | c_i) = \frac{n_{ki} + 1}{n_i + 2} \quad (10)$$

式中, n_i 表示训练文本集中类别 c_i 的文档数, n_{ki} 表示训练文本集中含有 t_k 并且类别为 c_i 的文本数。

已知种子特征集为 $T = \{t_1, t_2, \dots, t_{|V|}\}$, 类别集合 $C = \{c_1, c_2, \dots, c_{|C|}\}$, 对于文档 d , 我们可以估算出文档的属于类别集合 C 中每个类别中的概率集合为 $P(d) = \{p(c_1 | d), p(c_2 | d), \dots, p(c_{|C|} | d)\}$, 这个概率集合是表征文档 d 的特征向量, 我们是通过计算每个文档的特征向量, 从而利用相似性度量公式计算文档间的相似度。

3.4 基于项目评分的个性化推荐

基于项目评分的协同过滤算法主要的思想是: 如果对于某些项目, 一些用户对于这些项目的评分相似, 那么这些项目就是相似的。基于项目的协同过滤算法通过相似性度量方法找到目标用户有过评分记录的项目集合相似的项目, 然后通过已评分的项目去计算待推荐项目的评分, 从中挑选评分靠前的项目, 产生对目标用户的推荐列表^[6]。

下面通过相似性公式度量用户间的相似性, 如式(14)所示:

$$\text{Sim}(a, b) = \frac{\sum_{i \in Q} |R_{ai} \times R_{bi}|}{\sum_{i \in Q} |R_{ai} - R_{bi}| + 1} \quad (11)$$

式中, a, b 代表项目 a 和 b , Q 代表对项目 a, b 之间有过共同评分的用户集, R_{ai} 和 R_{bi} 分别表示用户 i 对项目 a, b 的评分。得到项目之间的相似度后, 就可以得到用户有过记录的项目的相似的项目, 我们预测它们的评分, 从而推荐分数较高的项目。

由上节中的文本 d , 我们得到了由概率组成的

特征向量 $P(d)$, 可以把 d 作为项目, 对于文本 u 和 v 之间的相似度, 可以通过利用相似性度量公式计算, 对于不同领域的商品, 例如旅游景点, 需要考虑的因素有很多, 有景点距离、景点价格, 淡季和旺季等, 这里我们忽略了这些因素, 只是根据文本内容计算相似度, 进行推荐。为了减小流行度偏见, 提高推荐的准确率, 覆盖率和流行度, 我们需要修正项目之间的相似度计算公式, 修正后的公式:

$$W_{u,v} = \frac{\sum_{i=1}^{|C|} |p_{ui} \times p_{vi}|}{\log(1 + \sum_{i=1}^{|C|} |p_{ui} \times p_{vi}|)} \quad (12)$$

$$\frac{\sum_{i=1}^{|C|} |p_{ui} - p_{vi}| + 1}{\sum_{i=1}^{|C|} |p_{ui} \times p_{vi}| + 1}$$

式中, $|C|$ 表示是文档集合类别个数,

$\frac{1}{\log(1 + \sum_{i=1}^{|C|} |p_{ui} \times p_{vi}|)}$ 是对热门商品的惩罚度。

已知用户 u 评过分的项集合, $D = \{d_1, d_2, \dots, d_n\}$ 对应的评分集合为 $C = \{c_1, c_2, \dots, c_n\}$, 待推荐项目为 B , 项目 B 与项目集合 D 的相似度集合 $S = \{s_1, s_2, \dots, s_n\}$, 则用户对项目 B 的评分公式为

$$\text{Score}(B) = \frac{\sum_{i=1}^n c_i \cdot s_i}{\sum_{i=1}^n s_i} \quad (13)$$

式(13)成立的假设条件是: 假如用户对于相似的物品有相同的喜好(例如同系列的电影, 或者同类型的手机等), 那么我们可以用待预测物品和已评分物品之间的相似度作为权值, 以已评分的项目的分数的加权平均数作为待预测物品的分数, 而且也不考虑其他用户对于评分的影响。

3.5 改进的基于项目语义的推荐算法

利用上述基于项目语义的相似度模型和 Bayes 概率模型的相似度计算方法相结合, 用户 u 对项目 B 的最终预测分数由这两种方法共同决定, 如下:

$$\text{Score}^\wedge(B) = \begin{cases} \lambda \text{Score}(B)^{(1)} + (1 - \lambda) \text{Score}(B)^{(2)} \\ \min(\text{Score}(B)^{(1)}, \text{Score}(B)^{(2)}) > 0 \\ \max(\text{Score}^{(B)(1)}, \text{Score}^{(B)(2)}), \text{其他} \end{cases} \quad (14)$$

式中, $\text{Score}(B)^k, k = 1, 2$, 分别表示基于 Bayes 概率模型和基于项目语义两种方法得出用户 u 对项目 B 的预测得分, λ 表示这两种方法在最终预测分数中所占的比重, 可以根据对项目 B 评过分的用户占总

用户人数的百分比进行调节,也可以人为的进行设定,这里将其设为 $\lambda = 0.7$.

改进的基于项目语义的推荐算法的算法步骤是:

- (1) 将训练集中的用户、项目和评分数据和测试集中的用户、项目数据相结合,按照第 2 节的步骤,计算出训练集和测试集中全部项目之间的语义相似度矩阵 $[\chi]_n$;
- (2) 按照式(13)计算出测试集中项目基于语义相似的预测分数;
- (3) 利用基于 Bayes 概率模型的推荐算法,得出训练集和测试集中项目的特征向量;
- (4) 按照式(13),计算出测试集项目基于概率模型的预测分数;
- (5) 按照式(14),得出最终的预测分数;

4 实验设计及结果分析

本文实验所用的数据集是采用的是携程网中的数据,爬取了携程网大约 10 000 个景点的名称、地址和简介,以及这 956 个用户对于这 8 652 个景点的评分,约 50 000 个,我们以每个景点的名称和简介作为一个文档,作为领域内容. 本文所涉及的算法是在 Spark 平台上面进行运行测试的,其中 Spark 采用的是 Python 版本的 PySpark.

本文用一般的基于项目评分的推荐算法(ItemCF)、基于项目语义的推荐算法和本文改进的基于项目语义相似的推荐算法进行对比实验. 为了很好地体现对比实验的效果,我们首先将这 50 000 个评分按 4 : 1 的比例随机抽取构成 7 组训练集和测试集,另外我们再挑选出 3 组训练集和测试集,其中抽取条件是这 3 组的测试集中的评过分的用户数占总用户数的比例不能超过 15%.

实验中对预测推荐结果进行评估验证的方法是应用比较广泛的平均绝对误差(Mean Absolute Error, MAE):

$$MAE(f) = \frac{1}{R_{test}} \sum_{r_{ui} \in R_{test}} |f(u,i) - r_{ui}|$$

(15)

式中, $f(u,i)$ 是用户 u 对于项目 i 的预测的分数, r_{ui} 是测试集 R_{test} 中项目的实际分数,通过用户对于新物品的预测评分来和实际的评分的偏差度来度量预测准确率,MAE 的值越小的话,说明推荐效果越好.

本文将改进的基于项目语义的推荐算法和与基于项目语义的推荐算法(Semantic Similarity Col-

laborative Filtering, SSCF)以及基于项目评分的推荐算法(ItemCF)进行实验对比,得到的实验结果如图 2 所示.

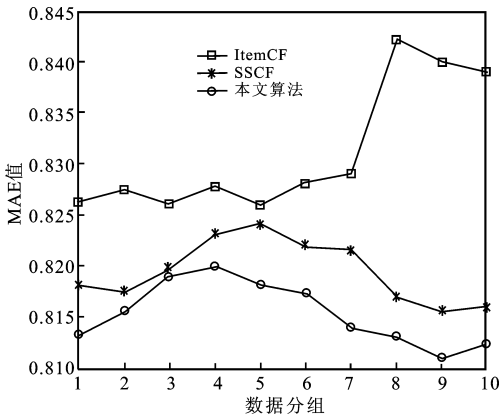


图 2 实验结果

图 2 中横坐标代表的是实验中所用的分组数据,一共 10 组数据,纵坐标代表的是每组数据分别用两种算法得出的 MAE 值.

前 7 组数据使我们随机挑选的数据,从实验结果看,本文提出的改进的基于项目语义的推荐算法的 MAE 值比一般的基于项目语义的推荐算法以及基于项目评分的推荐算法有所减小,证明了本文提出的算法的优越性,特别是实验的最后 3 组数据,数据中的评过分的用户数量较少,可以看出 ItemCF 算法与其他两种算法的 MAE 值差距很大,实验结果表明,基于项目评分的推荐算法的推荐精度容易受到对候选推荐项目评过分的用户数量的影响. 而基于项目相似度的领域内容推荐算法,在保证算法性能的前提下,能够在一定程度上面减小这种影响,从而提高了算法的推荐精度,而本文提出的改进的基于项目语义的领域内容推荐算法在这种情况下,效果更好.

5 结束语

本文针对基于项目评分和基于项目语义的领域内容推荐算法的推荐精度不高的情况,提出一种改进的基于项目语义的推荐算法,通过利用概率模型,多维度的描述景点特征,精确地计算出景点间的相似度,对用户已有评分景点做出预测. 实验结果表明改进的基于项目语义的推荐算法比一般的基于项目语义和项目评分的推荐算法,在预测精度上进一步提高,而且特别是在评分用户所占比例比较小的情况下,改进的基于项目语义的推荐算法的优势更加明显.

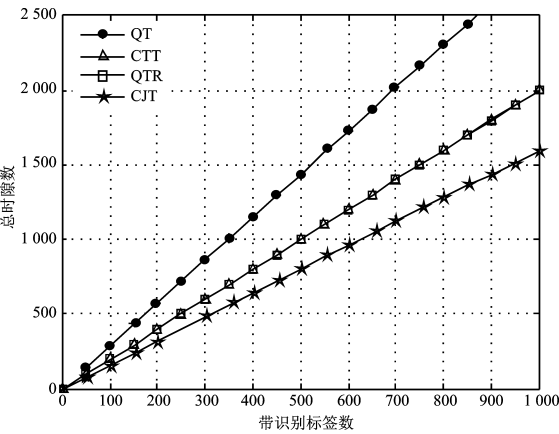


图 4 总时隙数

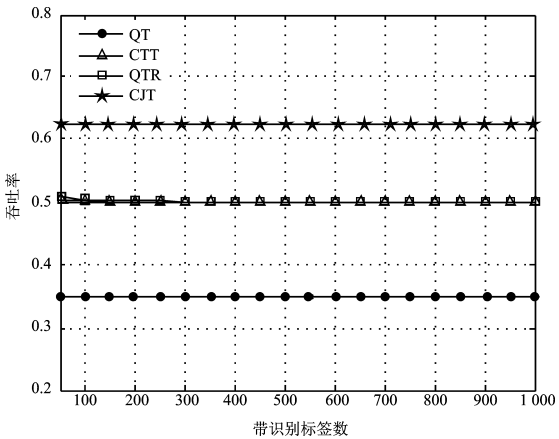


图 5 吞吐量

度,提高效率 and 吞吐率,在特殊场合的应用中,更能发挥其优点.进一步的研究将充分考虑标签信息存储与安全,以及电路实现等实际应用问题,推动 RFID 技术在物联网中的应用.

(上接第 97 页)

参考文献:

[1] 葛斌,李芳芳,郭丝路,等.基于知网的词汇语义相似度计算方法研究. 计算机应用研究, 2010, 27(9): 3329-3333.

[2] 王毅,楼恒越.一种改进的 Slope one 协同过滤算法. 计算机科学, 2011, 38(10): 192-194.

[3] 范小丽,刘晓霞,文本分类中互信息特征选择方法的研究. 计算机工程与应用, 2010, 46(34): 123-125.

[4] LEWIS D D, Naïve (Bayes) at forty: The independence assumption in information retrieval[C] // Proceedings of ECML-98, 10th European Conference on Machine Learning. Chemnitz, Germany: IEEE, 1998;

参考文献:

[1] 吴博,周铜,王栋. RFID 防碰撞算法分析与研究[J]. 微电子学与计算机, 2009, 26(8): 237-242.

[2] 李青青,刘洪武,张小林.一种基于不等长时隙的射频识别防碰撞算法[J]. 电子与信息学报, 2011, 33(11): 2628-2633.

[3] 熊伟,滕培俊,梁青.一种基于冲突跟踪的 RFID 防冲突算法[J]. 空军工程大学学报(自然科学版), 2009, 10(3): 68-72.

[4] 白乐强,杨晰.并行识别 RFID 自适应多叉树防碰撞算法[J]. 微电子学与计算机, 2015, 32(1): 136-139.

[5] Huang Y, Chen X. Dynamic slots collision tracking tree based RFID anti-collision algorithm[C]// 10th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2014). Beijing, China: IET, 2014: 662-669.

[6] Zhang L, Zhang J, Tang X. Assigned tree slotted Aloha RFID tag anti-collision protocols[J]. IEEE Transactions on Wireless Communications, 2013, 12(11): 5493-5505.

[7] Hush D R, Wood C. Analysis of tree algorithms for RFID arbitration[C]//IEEE International Symposium on Information Theory. Cambridge: IEEE, 1998: 107-116.

作者简介:

张荣华 女, (1987-), 硕士. 研究方向为射频识别、嵌入式系统. E-mail: zrh3536@163.com.

杨大志 男, (1976-), 硕士, 副教授. 研究方向为视频识别、嵌入式系统.

廖红华 男, (1972-), 博士, 教授. 研究方向为系统级建模.

柳忠彬 男, (1972-), 博士, 教授. 研究方向为机械设计.

4-15.

[5] McCallum A, Nigam K. A comparison of even models for naive Bayes text classificationIN AA1-98 Workshop on learning for text categorization[R]. 1998.

[6] 刘芳先,宋顺林.改进的协同过滤推荐算法. 计算机工程与应用, 2011, 47(8): 72-75.

作者简介:

雷鸣 男, (1989-), 硕士研究生. 研究方向为文本数据挖掘. E-mail: lm946<lm946@mail.ustc.edu.cn.

朱明 男, (1963-), 教授, 博士生导师. 研究方向为数据挖掘.