基于内容过滤的电子商务推荐系统研究

曹 毅,贺卫红
(湖南工学院,湖南 衡阳 421008)

摘 要:个性化推荐在网络应用中能有效提高服务质量,在电子商务中的表现更加突出。论述了基于内容过滤的电子商务推荐系统,利用向量空间模型挖掘用户独特的兴趣特征,然后根据产品信息特征的量化值产生推荐序列,并根据用户的反馈信息自适应学习,以提高系统的综合性能。实验结果表明,基于内容过滤的推荐方法其总体性能随时间的推移得到了提高。

关键词:电子商务;推荐系统;个性化推荐;向量空间模型
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)06-0182-04

Research on E-Commerce Recommender System Based on Content-Based Filtering

CAO Yi, HE Wei-hong
(Hunan Institute of Technology, Hengyang 421008, China)

Abstract: The application of personalized recommendation in the Internet effectively improved its service, especially the service of E-commerce. content-based filtering E-commerce recommender system was discussed fully in this paper. Users' unique features can be explored by means of vector space model (VSM) firstly. Then based on the qualitative value of products information, the recommender lists were obtained. Since the system can adapt to the users' feedback automatically, its performance were improved comprehensively. According to the experiments result, the overall performance of the recommender based on content-based filtering was enhanced with time.

Key words: E-commerce; recommender system; personalized recommendation; VSM

0 引 言

随着互联网的普及和电子商务的发展,电子商务系统在为用户提供越来越多选择的同时,其结构也变得更加复杂,用户经常会迷失在大量的商品信息空间中,无法顺利找到自己需要的商品。在日趋激烈的竞争环境下,电子商务推荐系统能有效保留用户、防止用户流失,提高电子商务系统的交叉销售能力。研究表明,电子商务的销售行业使用个性化推荐系统后,销售额能提高2%~8%[1],尤其在书籍、电影、日用百货等产品相对较为低廉且商品种类繁多、用户使用个性化推荐系统程度高的行业,推荐系统能大大提高企业的销售额[2]。

文中提出了一个基于内容过滤的电子商务推荐系统,基本问题包括用户兴趣特征的获取、建模以及相似性计算。该系统包括数据处理和推荐处理两个阶段,数据处理阶段生成初始的推荐模板和初始阈值;推荐处理阶段,系统能够自动调整推荐模板和阈值,以获得最佳的推荐性能。

收稿日期:2008-09-17;**修回日期**:2008-12-09
基金项目:湖南省教育科学研究项目(08C234);湖南省软科学研究计划项目(04ZH6005);湖南省普通高校教学改革研究项目(2006191)
作者简介:曹 毅(1977-),男,硕士,研究方向为电子商务。

1 用户兴趣特征描述

1.1 用户兴趣

要实现个性化的推荐服务,首先必须搜集用户的个人信息,建立用户兴趣特征描述的模型。目前收集用户信息普遍采用的方式有两种:显示特征描述和隐式特征描述。

在显示特征描述方式下,直接获得用户的兴趣信息,但对用户而言这种方式使用起来比较繁琐,其准确程度取决于表单、问卷设计的水平和用户的配合力度[3];在隐性特征描述方式下,系统追踪访问者的行为,用户浏览某个网页所用的时间与该页面中字符数目的比值能有效揭示用户的兴趣[4],系统通过记录用户的IP地址、时间、查询项等浏览行为,通过Web挖掘的方式来分析用户的兴趣[5]。

1.2 向量空间模型

在本系统中,通过去除网页中与内容无关的结构,只保留与网页内容相关的文本信息,从而得到与网页对应的文本。为了让计算机处理这些文本信息,系统采用向量空间模型[6](VSM),其基本思想是假定文本中的字和词在文本中出现的概率在内容和位置上是相互独立的,在确定文本内容的类别时可不考虑文本结构和字词的出现顺序。向量空间模型可以将给定的文本转换成一个维数很高的向量,其最突出的特点是可以方便地计算出两个向量的相似度,即向量所对应的文本的相似性。

在向量空间模型中,文本用 D(Document) 表示;特征项是指出现在文档 D 中且能够代表该文档内容的基本语言单位,用 t(Term) 表示。文本可以用特征项集表示为 $D(T_1, T_2, \cdots, T_n)$,其中 T_k 是特征项,$1 \leq k \leq n$。对任一特征项而言,由于在文本中出现的位置和词频不同,对文本内容的价值也是不同的,所以,对含有 n 个特征项的文本而言,应该给每个特征项赋予一定的权重表示其重要程度,即文本 D 的向量表示为 $D = D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$,简记为 $D = D(W_1, W_2, \cdots, W_n)$,其中 W_k 是 T_k 的权重,$1 \leq k \leq n$。在向量空间模型中,两个文本 D_1 和 D_2 之间的内容相关度 $\mathrm{Sim}(D_1, D_2)$ 用向量之间夹角的余弦值表示[7],公式为:

$$\mathrm{Sim}(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{(\sum_{k=1}^{n} W_{1k}^2)(\sum_{k=1}^{n} W_{2k}^2)}}$$

式中,W_{1k}、W_{2k} 分别表示文本 D_1 和 D_2 第 k 个特征项的权重,$1 \leq k \leq n$。

1.3 特征项的提取

由从文本中得到文本特征向量,要经历一个特征项提取的过程,特征项的提取就是从全部可能的词汇中抽取一个对表达文本内容有较强说服力的最优的特征项子集。这样做的目的主要有两个:一是为了提高程序的效率,简化运算,提高运行速度;二是所有几万个词汇对文本内容的意义是不同的,一些通用的普遍存在的词汇对文本内容的贡献小,为了提高推荐系统的精确度,应当去除那些表现力不强的词汇,筛选出针对文本内容的最优的兴趣特征项集合。

最优特征项就是那些与相关文本集 rel(Q)互信息量最大的词汇,词汇和相关文本集之间的对数互信息量由下式计算[8]:

$$\log MI(w_i, \mathrm{rel}(Q)) = \log(P(w_i \mid w_i \in \mathrm{rel}(Q))/p(w_i))$$

式中,w_i 表示文本中第 i 个词,$P(w_i \mid w_i \in \mathrm{rel}(Q))$ 表示词 w_i 在相关文本集 rel(Q) 中出现的比重,$P(w_i)$ 表示词 w_i 在数据处理文本中的比重。

2 系统结构

基于内容过滤的电子商务推荐系统的系统结构如图 1 所示。

整个推荐系统按功能划分为两个部分,即数据处理部分和自适应推荐部分,这两个部分对用户而言都是不可见的。数据处理部分主要对用户的个性特征进行分析和加工,利用向量空间模型提炼出用户的兴趣,建立推荐模型,并设置初始阈值;自适应推荐部分的主要功能是生成推荐访问序列,通过 Web 服务器将序列推荐给用户,并获得用户对推荐内容的反馈信息,及时对推荐模型进行自适应调整,以获得最佳的推荐质量。

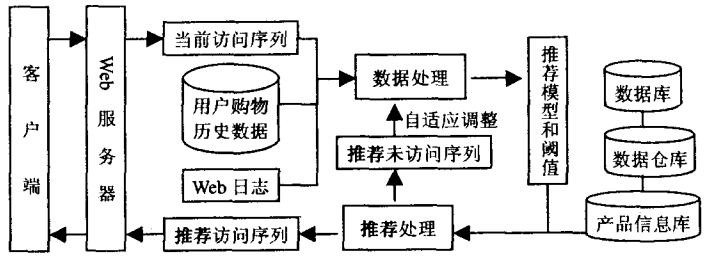

图 1　基于内容过滤的电子商务推荐系统结构

3 基于内容过滤的推荐系统算法

3.1 数据处理算法

3.1.1 数据处理流程

首先将用户当前访问序列转变为兴趣主题向量,连同从用户历史购物数据和 Web 日志中抽取的特征向量进行加权和运算,得到初始推荐模型,再计算初始向量和当前访问序列之间的相似度,最后为每一个兴趣主题设置最优的初始相似度阈值。

3.1.2 初始推荐模型的建立

初始推荐模型向量是由兴趣主题向量、从用户购物历史数据中抽取的特征向量以及从 Web 日志中抽取的特征向量进行加权和运算得到的,设权重分别为 a,b 和 c,则有:

$$Pf_0(Q) = a * P_0(Q) + b * P_1(Q) + c * P_2(Q)$$

式中 Q 表示兴趣主题,$Pf_0(Q)$ 表示兴趣主题 Q 的初始推荐模型向量,P_0、P_1 和 P_2 是它的 3 个分向量。

对于主题向量 $P_0(Q)$,有 $P_0(Q) = (p_{01}, p_{02}, \cdots, p_{0w})$,$W$ 表示词汇的总数,p_{0i} 表示第 i 个词 w_i 的权重。根据 Buckley 等人提出的 Smart 系统中的 ltc 公式有:

$$p_{0_i}=\begin{cases}\log(N/df(w_i)) & \text{if } w_i\in Q\\ 0 & \text{otherwise}\end{cases}$$

式中 N 表示文本总数，$df(w_i)$ 表示词汇 w_i 在文本中出现的数量，若 w_i 没有在主题 Q 中出现，则权重为 0。

同样的，对于从用户购物历史数据中抽出的特征向量 $P_1(Q)$，有 $P_1(Q)=(p_{11},p_{12},\cdots,p_{1w})$，$p_{1i}$ 表示 w_i 的权重：

$$p_{1i}=\begin{cases}\log MI(w_i,\text{rel}(Q)),\text{if } \log MI(w_i,\text{rel}(Q))\geq 3\\ 0, \quad \text{otherwise}\end{cases}$$

对于从 Web 日志中抽出的特征向量 $P_2(Q)$，有 $P_2(Q)=(p_{21},p_{22},\cdots,p_{2w})$，$p_{2i}$ 表示 w_i 的权重：

$$p_{2i}=\begin{cases}\log MI(w_i,\text{pseudo}-\text{rel}(Q)),\\ \qquad \text{if } \log MI(w_i,\text{pseudo}-\text{rel}(Q))\geq 3\\ 0, \quad \text{otherwise}\end{cases}$$

3.1.3 *初始阈值的设置*

相似度阈值的确定是十分困难的，在本系统中采用预定初始值，然后给出测试数据进行推荐，再根据推荐的准确程度调整初始值。阈值一旦设立，那些与模型向量的相似度大于或等于阈值的数据就被认为是相关兴趣主题数据，应进行推荐处理。

本推荐系统仿照信息过滤的 T9P[9]评价指标，可以通过计算模型向量和预处理数据之间的相似度，对任意的阈值水平，计算在这个阈值下的性能评价指标 T9P 值，从中选择能导致最佳性能的阈值作为初始阈值。而模型向量和预处理数据之间的相似度可以用向量之间夹角的余弦值表示：

$$\text{Sim}(d,pf)=-\frac{\sum_{k=1}^{m}d_k*pf_k}{\sqrt{(\sum_{k=1}^{m}d_k^2)(\sum_{k=1}^{m}pf_k^2)}}$$

式中 d 表示要处理数据文本，pf 表示模型向量，m 表示特征向量的维数，d_k 表示 d 中第 k 个词的权重。

3.2 推荐处理算法

3.2.1 *推荐处理流程*

当初始的推荐模型已建立并且阈值也设置好之后，对产品信息库等数据库中的关于产品介绍的数据，都可计算它和用户某个兴趣主题的模型向量的相似度。若相似度大于等于阈值，就被认为是与用户兴趣相关，形成推荐序列，通过 Web 服务器推荐给用户，然后由用户判断推荐是否有效，再根据判断的结果自适应地修改模型向量或调整阈值，使得推荐系统的性能不断提高以更好地为用户服务。

3.2.2 *阈值的调整*

阈值的选取是影响该推荐系统模型性能的关键因素。阈值可以根据用户对推荐访问序列的反馈信息进行调整，提高或降低阈值。如果阈值设置较低，则大于或等于阈值的信息数将增多，这样可大大提高系统的召回率；反之，如果阈值设置较高，则满足推荐条件的信息数将减少，这样可以提高准确率。因此，阈值的调整有以下原则：

(1)当推荐的信息多于必要时，就提高阈值，提高准确率降低召回率；

(2)当推荐的信息少于必要时，就降低阈值，降低准确率提高召回率。

3.2.3 *模型的修改*

推荐访问序列若被用户判断为与自己兴趣相关，就会浏览其相关信息，这时推荐访问序列就变成了当前访问序列。调整模型向量时，可从当前访问序列中抽取兴趣主题向量，从用户购物历史数据和 Web 日志(这时 Web 日志也发生了相应的变化)中抽取特征向量。新的模型向量就是由主题向量与特征向量进行加权和运算得到的，设权重分别为 a'，b' 和 c'，则有：

$$pf'(Q)=a'*P_3(Q)+b'*P_1(Q)+c'*P_4(Q)$$

式中 $P_3(Q)$ 表示从当前访问序列中提取的兴趣主题向量，$P_3(Q)=(p_{31},p_{32},\cdots,p_{3w})$，$P_4(Q)$ 表示从 Web 日志中提取的用户特征向量，$P_4(Q)=(p_{41},p_{42},\cdots,p_{4w})$。

4 实验结果与分析

为测试推荐系统，将某计算机杂志 2000～2003 年 4 年中每年选用期刊中 800 篇科技论文共 3200 篇作为实验数据。在实验中，把科技论文的摘要模拟成对商品信息的介绍，将科技论文模拟成商品，用户下载论文的记录模拟成用户购物历史数据，该记录和 Web 日志均可从提供下载论文的服务器上得到。由于科技论文的内容比较明确，所以能够获得比较清晰的结果。

实验时将根据用户当前访问序列、用户下载论文的历史记录和 Web 日志建立初始推荐模型，并设置阈值，然后根据每年科技论文的摘要进行推荐，并将产生的推荐序列按年份提供给用户进行确认；系统将根据用户的反馈信息对模型进行修改，然后依据修改后的模型对下一年的论文再进行推荐处理。实验结果如图 2 所示。

系统推荐质量的衡量一般采用信息检索领域的评价标准，用正确率(precision)和召回率(recall)衡量：

$$\text{precision}=\frac{\text{正确推荐的项目数}}{\text{所有推荐的项目数}}$$

$$\text{recall}=\frac{\text{正确推荐的项目数}}{\text{所有应推荐的项目数}}$$

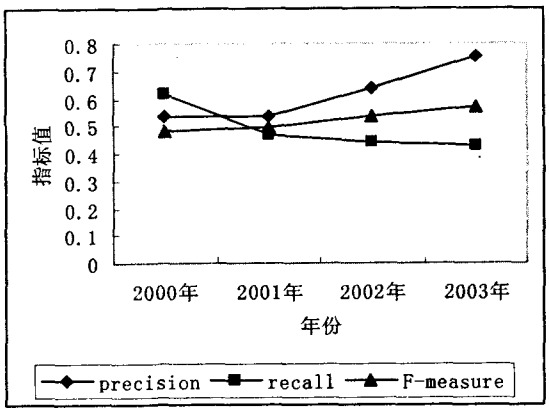

图2　系统性能曲线图

正确率和召回率在一定程度上是一对相对矛盾的指标,正确率高意味着召回率低。为了平衡两者,通常采用综合评价指标 F－measure:

$$F-measure=\frac{2*precision*recall}{precision+recall}$$

从图2可以看出,随着时间段的推移,推荐系统的正确率有一定幅度的上升,而召回率有所下降,但系统的综合评价指标 F－measure 值随着时间段的推移有所上升,这说明推荐系统的总体性能随时间段的推移有所提高。

5　结束语

文中提出的基于内容过滤的电子商务推荐系统,利用向量空间模型挖掘用户独特的兴趣需求,再根据产品信息特征的量化值来进行商品的推荐,并根据用户的反馈信息自适应学习,以提高系统的综合性能。

基于向量空间模型的方法根据特征项在文本中出现的位置和频率来判断其对文本内容的贡献,其本质只是一种浅层次的文本统计方法。基于内容过滤的推荐需要积累足够数量的评价才能构建出一个可靠的分类器,对于新用户和新产品就比较难处理;并且当用户的兴趣特征发生转变时,推荐系统模型不易更新,系统最终还要通过在实践中应用而不断优化。

参考文献:

[1] Lawrence R D, Almasi G S, Kotl Y V, et al. Personalization of supermarket product recommendations [R]. US: IBM, 2000.

[2] 吴辉娟,袁　方.个性化服务技术研究[J].计算机技术与发展,2006,16(2):32－34.

[3] 李　鹏,汪东升,陈　康.一个基于 VSM 的个性化信息推荐系统[J].计算机工程与设计,2003,24(10):19－22.

[4] 周文刚,马占欣.基于代理的 Web 页访问语义过滤与内容重现[J].计算机技术与发展,2007,17(4):120－124.

[5] 李石君,李　洲,余　军,等.基于 URL 过滤与内容过滤的网络净化模型[J].计算机技术与发展,2006,16(1):5－7.

[6] Buckley C, Salton G, Allan J, et al. Automatic query expansion using SMART[R]. [s.l.]:[s.n.],1995.

[7] 曹　毅,贺卫红.基于向量空间模型的信息安全过滤系统[J].计算机工程与设计,2006,27(2):224－227.

[8] 黄萱菁,夏迎炬,吴立德.基于向量空间模型的文本过滤系统[J].软件学报,2003,14(3):435－442.

[9] Robertson S, Hull D. The TREC－9 filtering track final report [C]//In: Voorhees E M, Harman D K, eds. Proceedings of the 9th Text Retrieval Conference (TREC－9). Gaithersburg: NIST Special Publication, 2001:25－40.

(上接第181页)

树检测效率的一个重要因素。通过及时对训练数据进行补充更新,以便将新的入侵特征更新到决策树模型中。

参考文献:

[1] 凌　宇,徐　雄,石林安. NIDS 中协议分析和模式匹配的研究[J]. 信息技术, 2006(5):82－84.

[2] 蔡　敏,叶　震,徐吉斌. 协议分析技术在入侵检测中的应用[J]. 计算机技术与发展,2007,17(2):240－241.

[3] Lee W. A data mining framework for constructing features and models for intrusion detection systems[D]. New York: Columbia University, 1999:250－310.

[4] 李庆华,赵延喜,蒋盛益. 基于数据挖掘的协议分析检测模型[J]. 计算机工程与设计,2005,26(7):1701－1703.

[5] 王丽萍,孙　蕾. 基于 Ethereal 开源代码构建协议解析器的方法研究[J]. 计算机技术与发展,2007,17(10):27－30.

[6] 杨学兵,张　俊. 决策树算法及核心技术[J]. 计算机技术与发展, 2007,17(1):43－45.

[7] Han Jian－wei, Kamber M. Data mining concepts and techniques[M]. Beijing: China Machine Press, 2000:180－200.

[8] Lee W, Stolfo S J, Mok K W. Data mining in workflow environments: experiences in intrusion detection [C] // ACM SIGKDD International Conference on Knowledge Discovery & Data Mining(KDD－99). USA:[s.n.],1999:114－124.

[9] Kruegel C, Toth T. Using decision trees to improve signature based intrusion detection[C]//Proc of the 6th International Workshop on the Recent Advances in Intrusion Detection (RAID). USA: Springer－Verlag, 2003:173－191.

[10] 佘石泉,周肆清. 正则表达式在编程题自动阅卷中的应用[J]. 计算机技术与发展, 2007,17(7):244－246.