浙江科技学院学报,第 17 卷第 4 期,2005 年 12 月
Journal of Zhejiang University of Science and Technology
Vol. 17 No. 4, Dec. 2005

基于相似性的商品搜索算法研究

陈明晶,王　衍
(浙江财经学院 信息学院,浙江 杭州 310018)

摘　要:传统的电子商务系统中,使用关键字匹配的算法实现商品搜索的功能,只能得到与顾客输入精确匹配的商品。通过引入模糊系统的概念和信息增益的 ID3 算法,对商品搜索算法进行改进,使得在顾客搜索商品时,不仅显示精确匹配的商品,而且可以提供与其要求相似的商品。这对于商家增加交易机会、发掘潜在顾客、提高个性化服务水平都有很大的促进作用。笔者在文中以手机销售网站为例,介绍了该算法的实现。

关键词:相似性;隶属度;信息增益;搜索算法
中图分类号: TP311.13　　**文献标识码**: A　　**文章编号**: 1671-8798(2005)04-0273-04

Research on searching algorithm of products based on imilarity

CHEN Ming-jing, WANG Yan
(Information School, Zhejiang University of Finance and Economic, Hangzhou 310018, China)

Abstract: In traditional E-C System, we search merchandise by matching keywords. This can only get the records which include these keywords. We can improve the algorithm in a cell-phone store on-line by using fuzzy system and ID3 algorithm of information gaining. When we use it, we can not only display the exact matching merchandise, but also those which are approximated to our demand. To the their business men, it can increase their business, develop potential customer, and also enhance are service for individuation.

Key words: similarity; degree of menbership; information ganining; searching algorithm

随着 Internet 的普及和信息技术的发展,电子商务越来越得到人们的重视,无论是企业还是顾客都开始应用这种新的交易方式。

电子商务最大的优势就是能方便地使用搜索功能来查找商品,而在常见的电子商务网站中,网页开发者们使用 SQL 语句查询并显示符合条件的记录,这种方法的优点是容易实现,并且查找速度快,缺点是只简单罗列那些满足顾客需要的信息,并没有对他们的兴趣作分析。随着计算机系统各方面性能的提高和企业对市场及消费者研究的深入,越来越多的企业开始关注顾客的购买需求和兴趣偏好,以向其推销企业的产品,增加潜在的交易机会。

一般来说,商品有很多属性,而相似性商品就是在这些属性上存在共同点。笔者在本文中以手机销售网站为例,提出了一个能实时分析顾客兴趣并推荐相似商品的算法。

收稿日期: 2004-08-01

作者简介: 陈明晶(1978—　),男,江苏句容人,助教,主要从事电子商务系统研究。

1 相关概念

1.1 属性和隶属度

设 $T=\{t_1,t_2,\cdots,t_m\}$ 是一个数据库，t_j 表示 T 的第 j 个记录，$I=\{i_1,i_2,\cdots,i_m\}$ 表示属性集，属性可以是数量型属性、布尔型属性和类别型属性[1]，第 j 个记录在属性 i_k 上的取值为 $t_j(i_k)$。首先，将记录在属性上的取值划分成若干个模糊集。设布尔型属性的2个取值为 A_1 与 A_2，则可以被划分成2个模糊集，仍记为 A_1 与 A_2。模糊集 A_1 的隶属度定义为：当取值为 A_1 时为1，取值为 A_2 时为0；模糊集 A_2 的隶属度定义类似。类别型属性只有少量的几个取值，也可以采用与布尔型属性类似的方法划分。如果是数量型属性，则需要采用模糊 c - 均值(FCM)算法[2]划分。

1.2 熵与信息增益

基于信息论的方法中常常利用信息增益(information gaining)如ID3算法[3]，来选择属性；或者利用信息增益率，如C4.5算法。另外，还有Gini-index、距离度量(distance measure)等方法，并且不同的方法对于不同的多值属性有不同的效果。这里以ID3算法为例。任意样本分类的期望信息：

$$I(s_1,s_2,\cdots,s_m)=-\sum p_i\times\log_2(p_i) \quad (1)$$

式(1)中：$i=1,2,\cdots,m$；S 是 s 个数据样本的集合。类别属性具有 m 个不同值 C_i。s_i 是类 C_i 中的样本数。P_i 是任意样本属于 C_i 的概率，$Pi\approx\frac{|S_i|}{|S|}$。

由非类别属性 A 划分为子集的熵：

$$E(A)=\sum(s_{1j}+s_{2j}+\cdots+s_{mj})/s\times I(s_{1j},s_{2j}\cdots,s_{mj}) \quad (2)$$

非类别属性 A 具有 v 个不同值 $\{a_1,a_2,\cdots,a_v\}$。利用 A 将 S 划分为 v 个子集 $\{S_1,S_2,\cdots,S_v\}$；

其中 S_j 包含 S 中在 A 上具有值 a_j 的样本。S_{ij} 是子集 S_j 中类 C_i 的样本数。

信息增益：$\mathrm{Gain}(A)=I(s_1,s_2,\cdots,s_m)-E(A)$ (3)

在ID3算法中，就是用 $\mathrm{Gain}(A)$ 来选择属性，最大信息增益的属性将作为主要的测试属性。

2 基础数据

2.1 数据库设计

当描述商品时，会用到很多的属性，如品牌、式样、网络类型、颜色、功能、价格等，每个属性的类型不一定相同，而且也有不同层级的分类，使得可以对商品进行相关各种层级的分类。为此，以商品表为事实表，其他的属性表为维度表，构成的星形结构如图1所示。商品表中的品牌、外形等均为外键。

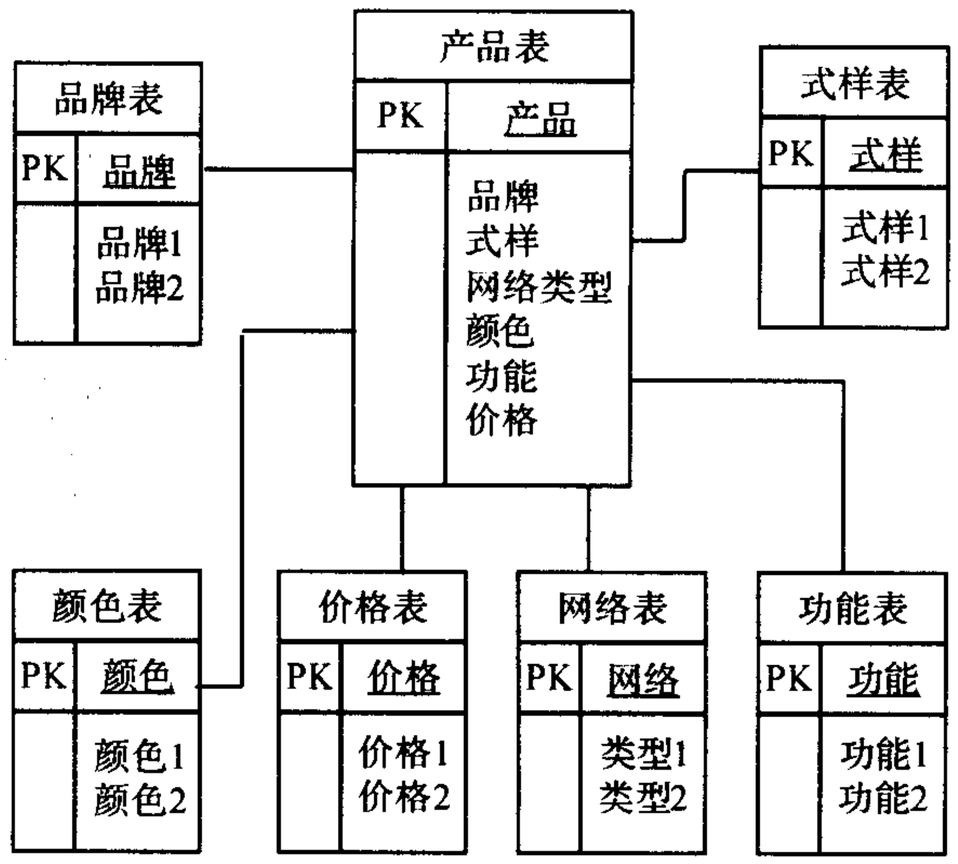

图1 产品的星形结构图

2.2 隶属度的计算

这里设 $T=\{t_1,t_2,\cdots,t_M\}$ 是商品的集合，其中 t_m 表示第 m 个商品($m=1,2,\cdots,M$)。将上文中列出的各个属性定义为一组属性序列：$I=\{i_1,i_2,\cdots,i_N\}$。对于一个给定的商品 $t_m(m=1,2,\cdots,M)$，其在属性 $i_n(n=1,2,\cdots,N)$ 上有一个具体的属性描述 D_{mn}，如表1所示。

表1 商品表

ID	型号	品牌	式样	颜色	网络类型	功能	价格/元
1	V600i	MOTOROLA	翻盖	冰岛银	GSM/GPRS	…	3 100
2	E808	SAMSUNG	滑盖	银雪白	GSM/GPRS	…	4 000
3	1100	NOKIA	普通	淡蓝	GSM/GPRS	…	700
4	i519	SAMSUNG	翻盖	银色	CDMA	…	5 000
5	F1	波导	翻盖	红色	GSM/GPRS	…	1 300
6	…	…		…	…	…	…

为了便于计算和比较，这里将只有两个相反情况的属性，如品牌(本土品牌、国外品牌)、网络类型(GSM/GPRS、CDMA)等，前者记为 -0.5，后者记为 $+0.5$(确保其区间长度为1)，而将其他属性量化至区间[0，1]，如在颜色(或式样)属性上，由于颜色(或式样)种类是有限的，可建立一个颜色表(或式

样表)，将常见颜色(或式样) 作简单划分，并给定一个隶属度值；在功能属性上，将其划分为基本功能和时尚功能，并将商品的功能数目进行统计；在价格属性上，可使用如下公式进行计算：

$$\mu(p)=\begin{cases}\dfrac{p}{5000} & p\in(0,5000)\\ 1 & p\in[5000,\infty)\end{cases} \tag{4}$$

通过以上方法，可得到“t_m 在属性 i_n 上的隶属度”，记为 μ_{mn}。计算结果见表 2。

表 2　商品隶属度表

ID	品牌	式样	颜色	网络类型	功能	价格	结果
1	0.5	0.8	0.15	−0.5	0.8	0.62	3.15
2	0.5	0.95	0.1	−0.5	0.85	0.8	5.05
3	0.5	0.1	0.5	−0.5	0.4	0.14	5.85
4	0.5	0.9	0.15	0.5	0.95	1	6.95
5	−0.5	0.7	0.7	−0.5	0.3	0.26	8.55
6	…	…	…	…	…	…	…

可见，有 $t_a\in T$、$t_b\in T$，如果对于 $\forall i_n\in I$，有 $\mu_{an}\approx\mu_{bn}$，则 t_a 与 t_b 是相似的可互为替代商品。一般情况下，当用户对 t_a 有购买兴趣时，可以将 t_b 也展示出来，供用户选择。

2.3　属性优先级

表 2 中的数值只显示出了商品在各个属性上的隶属度，并没有区分出各个属性的重要程度；事实上，商品的各种属性对其销售的贡献并不是相等的，即顾客在购买商品时，往往更注重某方面的属性。由于不同顾客的偏好不同，即使是同一顾客，在不同时期的购买偏好也不尽相同，因此，无法制定出适合大多数顾客的方案。这里我们计算出每个属性的信息增益，并按降序排序。具有最高信息增益的属性就是具有最高区分度的属性。

以某一时间段的销售数据为例，比较品牌与价格两个属性的优先级。表 3 是销售业绩表(数据来源于天极网[4])。

表 3　销售业绩表

行	品牌	价格	销售份额 /%	销售业绩
1	本土	低档	20.4	好
2	本土	中档	10.2	差
3	本土	高档	3.4	差
4	国外	低档	39.6	好
5	国外	中档	19.8	好
6	国外	高档	6.6	差

第一步　计算样本分类所需的期望信息，即熵：

$$I(s_1,s_2,\cdots,s_M)=\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$$

第二步　计算每个属性的熵及信息增益：

$$E(\text{价格})=\frac{2}{6}\times(-\frac{2}{2}\log_2\frac{2}{2})+\frac{2}{6}\times(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})+\frac{2}{6}\times(-\frac{2}{2}\log_2\frac{2}{2})=0.33$$

$$\text{Gain(价格)}=I-E(\text{价格})=0.67$$

$$E(\text{品牌})=\frac{3}{6}\times(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})+\frac{3}{6}\times(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})=0.92$$

$$\text{Gain(品牌)}=I-E(\text{品牌})=0.08$$

第三步　比较排序：

因为价格的信息增益大于品牌的，所以，价格的优先级高于品牌。也就是说，在用户购买产品时，对价格的要求比对品牌的要求更重要，在分析用户提出的要求时，要先考虑价格方面的要求。同理可计算求出商品所有属性的优先级排序。在对大多数人的购买兴趣进行分析后，得出以上属性的优先级：价格、品牌、式样、功能、颜色、网络类型。

3　应用实例分析

通常情况下，顾客可以通过搜索功能查找自己想要的商品，因此，需要先输入一部分信息，而我们的工作就是分析他们的输入内容，进而了解顾客的兴趣，并提供相应的商品信息。

3.1　顾客输入页面

在顾客输入的页面中，列出常见的属性字段，为了便于计算和分析，除了商品型号之外，其他字段都采用单 / 复选按钮或下拉列表式，因此，需要对各个属性进行分类，分类规则的好坏，对后期的计算和分析至关重要。可以采用至顶向下的方式进行分类。如价格属性，可以得出如图 2 的分类树。

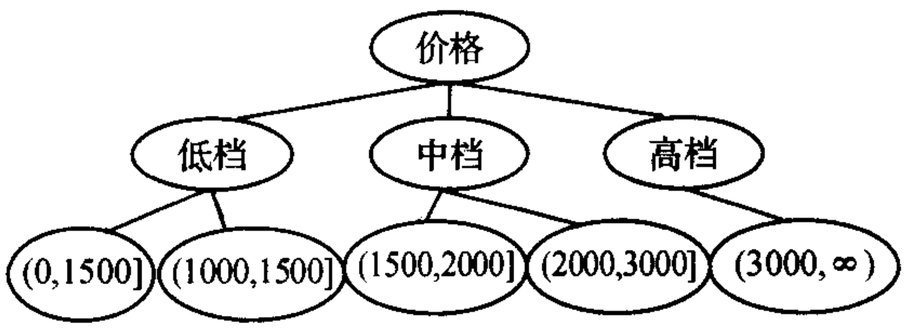

图 2　价格分类树

3.2　分析顾客输入

当顾客对列出的属性作选择后，将各个属性转换成隶属度值，转换方法与商品隶属度的算法类似。这里仅对品牌和价格的转换方法作描述。在品牌属性中，允许顾客对其钟爱的品牌作多项选择，根据顾客选择的本土品牌和国外品牌的数目，用公式(5)

计算其隶属度：

$$\mu(t) = \frac{t_f - t_n}{2 \times (t_f + t_n)} \tag{5}$$

式(5)中：t_f 为国外品牌数；t_n 为本土品牌数。

在价格属性中，根据所选价格分类中上下限的平均值，用式(4)计算。如果某一属性未做选择，则其隶属度为0，表明顾客对这个属性没有任何限制。因此，如果顾客未作任何选择，则 $v_k = \{0, 0, \cdots, 0\}$，表明顾客没有显示出兴趣偏好，无法作进一步分析。

设顾客选择了3个国外品牌和1个本土品牌，颜色为浅色，并选择了(2000，3000]的价格，则计算得到的结果为：$v_k = \{0.25, 0, 0.25, 0, 0, 0.5\}$。

3.3 显示结果

3.3.1 *精确匹配* 虽然根据顾客的输入能够分析出他们的兴趣范围，但也需要首先使用精确匹配的方法对商品表进行查找。如果顾客输入了商品的型号，则优先显示商品型号中包括顾客输入的记录，然后根据顾客输入的条件，显示符合要求的其他记录。精确匹配用SQL语句很容易实现。如：

```
Select * from 商品表
where 型号 like '%user_input%'
Select * from 商品表
where 品牌 in('MOTOROLA', 'NOKIA') and 价格 between 1000 and 1500
```

3.3.2 *推荐相似商品* 为了能增加潜在的销售机会，给用户更多的选择空间，需要在列出部分符合条件的商品之后(一般是10条记录)，再显示出不一定完全符合用户要求，但与其要求相似的商品。从前面分析可知，相似的产品在各个属性上的隶属度应该是相近的，两者之间的差比较小；同理，如果商品符合用户的条件，则两者隶属度之间的差也比较小，因此，可以通过比较其差异来选择符合条件的产品。考虑到各个属性优先级的因素，需要不同程序地扩大某些属性的影响，可设 $\delta = (7, 5, 2, 1, 4, 10)$，因此，商品的相似性算法如下：

$$D = \sum d_n \tag{6}$$

式(6)中：

$$d_n = \begin{cases} |\mu_n - v_n| \times \delta_n & v_n \neq 0 \\ 0 & v_n = 0 \end{cases}$$

计算结果如表2中“结果”列所示。

对于商品 m 和用户的搜索条件 k，如果 m 和 k 在 N 个特征上相似，则其各个特征的隶属度数值相近，对应两数之差的绝对值比较小，因此 d_n 也比较小。而如果 m 和 k 不相似，则它们在某个特征上相差较大，甚至符号相反，对应两数之差的绝对值必然较大，则 d_n 比较大。考虑两种极端的情形，当商品 m 和搜索条件在各个特征上隶属度均相同，则 $D = 0$，视为 m 和 k 完全相似；而当 m 和 k 在各个特征上隶属度分别为上限和下限，则 $D = 29$，视为 m 和 k 完全不相似。因此，对于搜索条件 k，可以通过选择 D 较小的项来推荐相应的商品。

3.3.3 *测试结果分析* 从表2“结果”可知，即使顾客在选择品牌时未选择商品1的品牌，但他选择的国外品牌较多，说明他对国外品牌较偏好，而他选择的价格为(2000，3000]。显然，商品1的价格也不在其选择范围内，但在各个属性上，商品1与顾客的选择条件都非常接近，因此，在计算时其差异最小，可被选为推荐商品；而其他产品在某些属性上都与条件相去甚远，因此，差异较大，一般不予推荐。而一些在属性上差异较大的商品，也可能最后结果相近，很可能导致结果的不准确。由于商品功能的多样化，顾客对不同功能的偏好也存在差异，因此，在其功能描述的量化上还需改进；另外，对属性优先级的划分，考虑到大多数人的选择偏好，实际上也存在着其他情况，无法面面俱到，因此属性优先级难以把握。

4 结束语

文中采用商品相似性作为商品搜索与推荐的重要依据，不仅实现了根据顾客的要求进行精确匹配的功能，而且也向顾客展示与其搜索意图相近的商品。这对于企业来说增加了销售机会，对于顾客来说也增加了选择的空间，体现了商家的个性化商业智能服务。由于不同商品的属性和描述方式不同，因此，本文的算法在广泛应用时还需要进一步的研究。

参考文献：

[1] Hathaway R J, Davenport J W, Bezdek J C. Relational dual of the c-means algorithms[J]. Pattern Recognition, 1989, 22(2): 205-212.

[2] 陆建江，徐宝文，邹晓峰. 模糊规则发现算法研究[J]. 东南大学学报(自然科学版)，2003，33(3)，271-274.

[3] 周世雄，韩永生. 数据挖掘技术在产品相似性上的应用研究[J]. 计算机工程与应用，2005，41(1)，207-209.

[4] 天极网. 2005年1月手机销售TOP10排行榜[R/OL]. http://www.yesky.com/Fashion/73747603539361792/20050201/1908204.shtml. 2005-02-01.