04 推荐引擎

推荐引擎也是一种专家系统，常用于电子商务（Electronic Commerce，EC）网站等的评价系统。下面笔者就来介绍它。

要点
- 推荐引擎是一种预测缺失信息并将其推荐给用户的专家系统
- 常用于电子商务网站和媒体
- 简单的填充示例：根据共现关系推导相关性
- 基于协同过滤的个性化推荐

预测并推荐相似内容的推荐系统

专家系统除在根据质谱数据推测物质化学结构的程序中使用之外，还应用于现在被广泛使用的推荐引擎。

用户在电子商务网站上查看某件商品时，网站会提示浏览过该商品的用户购买了哪些产品。这个向用户推荐相似商品的系统就是推荐引擎。推荐引擎也是一种专家系统，用来将用户的浏览信息作为关键词显示相似的内容。

推荐引擎可以分为两种类型，一种是基于内容的推荐，另一种是基于用户浏览记录和购买记录等个人信息的推荐。

基于内容的推荐

基于内容的推荐引擎只通过物品信息（电子商务网站的商品信息、新闻网站的报道信息等）进行计算，从而得到相似的内容。这类推荐引擎不使用任何用户的个人信息。

知识库内除包含标题、种类等信息的构成要素以外，还包含通过计算推导出的其他的数据表现形式。我们把信息的构成要素和通过计算推导出的数据表现形式统称为特征，把通过计算推导特征的处理过程称为特征提取。

例如，A 先生正在浏览关于熊本地震的新闻报道。推荐引擎需要解决的问题是接下来推荐哪些报道给 A 先生（图 2-12）。假设每篇报道都设置了关键词，我们可以利用这些关键词创建特征。

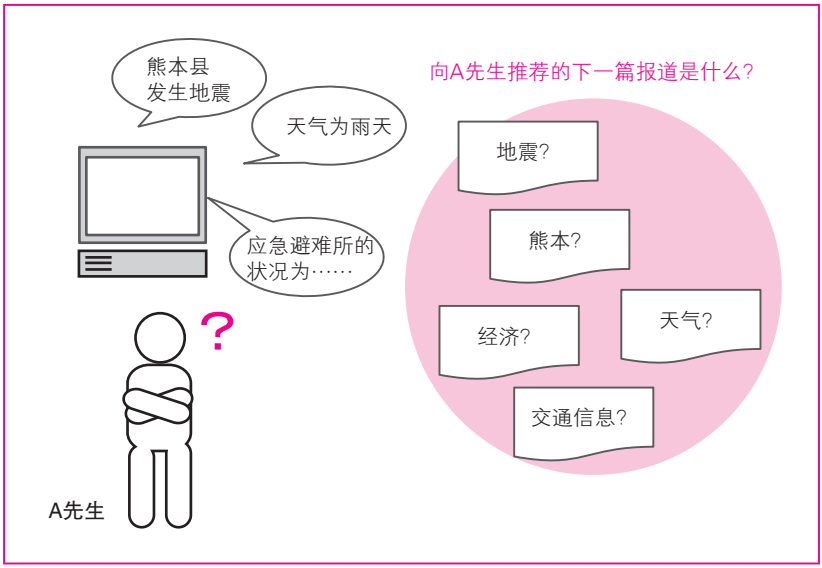

图 2-12　A 先生正在浏览熊本地震的新闻报道，下一篇该显示什么报道？

一些关键字或关键词等信息构成要素频繁出现在多篇报道或文章中的状态称为共现。共现状态的表达形式称为共现模式或共现关系（表 2-6）。

表 2-6　报道和关键词的关系表

	报道 a	报道 b	报道 c	报道 d
熊本	1	1	0	1
地震	1	0	1	1
地层	0	0	1	0
断层	0	1	1	1
下雨	1	0	0	0
停运	0	1	0	1

得到上述共现关系的数据后，我们就可以计算报道间的相关性了（表 2-7）。假设报道 a 和报道 b 的相关性由共同的关键词占二者关键词总数的比例来决定，这时我们可以循环计算出报道间的相关性。

表 2-7　表示新闻报道之间相关性的表

报道 a	1.000			
报道 b	0.333	1.000		
报道 c	0.333	0.333	1.000	
报道 d	0.571	0.857	0.571	1.000
	报道 a	报道 b	报道 c	报道 d

通过这个处理，我们可以按照内容相似度由高到低的顺序将和报道 a 相似的报道排列出来。上表的结果可排列为报道 d ＞报道 b= 报道 c。

上述例子的前提是每篇报道都设置了关键词，当然我们也可以通过计算来实现对文本的特征提取。笔者会在第 11 章简单介绍一下文本特征提取的相关内容。

不过，推荐系统如果只是单纯地把相似的报道放在一起，就会出现推荐内容雷同的问题，所以我们需要采用一些方法来防止过度推荐。

基于协同过滤的个性化推荐

协同过滤算法可以根据用户的浏览记录和购买记录等个人信息为用户推荐更合适的信息。亚马逊公司就使用了协同过滤推荐系统。

前面介绍的基于内容的推荐，是通过推导报道间的关键词的共现关系来定义相关性并以此来提取相似报道的。而个性化推荐是根据用户个人的历史信息与其他用户的信息之间的共现关系来进行相关分析，从而实现个性化推荐的。也就是说，协同过滤基于这样一个假设：如果某些用户对某些项目的行为和评分相似，则这些用户对其他项目的行为和评分也相似。

把目标用户 X 先生和 A 先生～E 先生浏览 10 种商品后是否购买了该商品的信息用 0 和 1 表示（表 2-8）。没有数据的地方填入连字符。

表 2-8　包含网站用户和商品购买记录的相似度矩阵

		商品										相关系数
		1	2	3	4	5	6	7	8	9	10	
用户	X	–	1	0	–	–	–	–	0	0	1	
	A	1	1	1	–	–	–	–	0	0	0	
	B	–	–	–	0	0	0	1	1	1	0	
	C	0	1	0	0	–	1	1	0	0	1	
	D	0	–	–	0	1	1	0	0	1	1	
	E	–	1	0	–	1	0	–	0	0	0	
推荐度												

这里的问题是根据 X 先生的购买记录，预测最适合推荐给 X 先生的商品，即计算商品的推荐度。

首先，计算 X 先生浏览过的 2、3、8、9、10 这 5 种商品，与其他 5 人共同浏览过的商品间的 0 和 1 的相关系数。在这种情况下，通常计算的是皮尔逊相关系数。例如，X 先生和 A 先生的相关系数可通过图 2-13 上端的公式计算出来。计算结果如图 2-13 下端所示。用同样的方法可计算出 X 先生与其他 4 人的相关系数。

X先生的购买记录 $\{x_1, x_2, x_3, x_4, x_5,\} = \{1, 0, 0, 0, 1\}$

A先生的购买记录 $\{y_1, y_2, y_3, y_4, y_5,\} = \{1, 1, 0, 0, 0\}$

相关系数

$$r = \frac{\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left(\sum_{i=1}^{5}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{5}(y_i-\bar{y})^2\right)}}$$

$$= \frac{\sum_{i=1}^{5}(x_i-0.4)(y_i-0.4)}{\sqrt{\left(\sum_{i=1}^{5}(x_i-0.4)^2\right)\left(\sum_{i=1}^{5}(y_i-0.4)^2\right)}}$$

$$= \frac{0.6\times0.6+(-0.4)\times0.6+0.4\times0.4+0.4\times0.4+0.6\times(-0.4)}{\sqrt{(0.6^2+0.4^2+0.4^2+0.4^2+0.6^2)(0.6^2+0.6^2+0.4^2+0.4^2+0.4^2)}}$$

$$= \frac{0.36+0.16\times2-0.24\times2}{0.36\times2+0.16\times3} = \frac{0.2}{1.2} \cong 0.167$$

图 2-13　相关系数的计算

计算结果中有 3 人（C 先生、D 先生和 E 先生）的相关系数大于 0.5，这表明他们与 X 先生之间是正相关的（购买趋势相同）（表 2-9）。

表 2-9　包含网站用户、商品购买记录和相关系数的相似度矩阵

		商品										与 X 先生的相关系数
		1	2	3	4	5	6	7	8	9	10	
用户	X	–	1	0	–	–	–	–	0	0	1	1.000
	A	1	1	1	–	–	–	–	0	0	0	0.167
	B	–	–	–	0	0	0	1	1	1	0	–1.000
	C	0	1	0	0	–	1	1	0	0	1	1.000
	D	0	–	–	0	1	1	0	0	1	1	0.500
	E	–	1	0	–	1	0	–	0	0	0	0.612
推荐度												

本来应该另行讨论如何选择目标对象，但接下来我们就直接以这 3 人为对象，介绍如何选择推荐给 X 先生的商品。

X 先生未浏览过的商品包括 1、4、5、6、7。我们把 C 先生、D 先生和 E 先生 3 人浏览这 5 种商品的数据平均值作为推荐度。不使用总值的原因是考虑了缺失值的影响。这样就能根据与 X 先生购买行为相似的 3 个人的数据，通过计算找到 X 先生未浏览过的但最有可能购买的商品。这里商品 5 的推荐度最高，为 1.00，所以接下来要向 X 先生推荐的是商品 5。

表 2-10 中的示例用 0 和 1 表示购买记录，而目前一些常用推荐引擎中使用的是五级评分法。

表 2-10　包含网站用户、商品购买记录、相关系数和推荐度的相似度矩阵

		商品										相关系数
		1	2	3	4	5	6	7	8	9	10	
用户	X	–	1	0	–	–	–	–	0	0	1	1.000
	A	1	1	1	–	–	–	–	0	0	0	0.167
	B	–	–	–	0	0	0	1	1	1	0	–1.000
	C	0	1	0	0	–	1	1	0	0	1	1.000
	D	0	–	–	0	1	1	0	0	1	1	0.500
	E	–	1	0	–	1	0	–	0	0	0	0.612
推荐度		0.00			0.00	1.00	0.67	0.50				