基于加权 XML 模型的个性化产品推荐方法*

李树青
(南京财经大学信息工程学院信息管理系　南京 210046)

【摘要】提出一种基于加权 XML 数据结构的用户兴趣模型构建方法,对于其中的每个 XML 节点都增加了表达用户个性化信息的权值,并据此设计一种对加权 XML 模型进行相似度比较的算法。最后,详述基于此用户兴趣模式的个性化产品推荐系统的实现方法。

【关键词】个性化　XML 检索　用户模式表达　信息推荐
【分类号】TP391

The Personalized Product Recommendation Method Based on Weighted XML Model

Li Shuqing
(Department of Information Management, College of Information Engineering,
Nanjing University of Finance & Economics, Nanjing 210046, China)

【Abstract】This paper puts forward a new method for constitution of user preference model based on weighted XML data structure, with each node appends weight value for representing users' personalized information. It also designs a new arithmetic to compare similarity of weighted XML model. Finally, this paper discusses the implementation of personalized product recommendation system based on this user preference model at detail.

【Keywords】Personalization　XML retrieval　Expression of user profiles　Information recommendation

1 引　言

个性化推荐技术已经成为现代电子商务领域的一个重要研究方向。它的主要目的在于减少提供给用户的无关信息,消除信息负载现象。很多研究表明,采用个性化技术可以大幅度提高用户满意度[1]。目前的研究角度主要有以下几个方面:

(1)研究如何提高现有算法的精确度;

(2)研究用户与推荐系统间的交互方法,以便设计更为合理和有效的用户信息采集机制,同时也能兼顾用户隐私的保护;

(3)研究诸如用户特征和产品特征对推荐系统性能的影响因素等[2]。

用户个性化推荐服务的关键内容在于准确和有效的表达用户兴趣模型,并基于此模型来计算与推荐客体的相关度。因此,如何有效和准确表达用户的个性化特征成为所有个性化推荐方法的基础和重要影响因素。目前,常见的方法主要分为两大类:

收稿日期:2008 - 12 - 22
收修改稿日期:2009 - 03 - 25
*本文系 2007年江苏省省属高校自然科学基础研究面上项目"基于 Web个性化推荐服务的 C to C电子商务平台框架"(项目编号:07KJD520074)和江苏省教育厅"青蓝工程"基金资助项目的研究成果之一。

(1)基于关键词表达 (Keyword - based)的用户兴趣模型,这种方法出现较早,简单易行,它主要使用与当前用户个性化特征相关的关键词序列来表达用户兴趣模型;

(2)基于语义表达 (Semantic - based)的用户兴趣模型,该方法基于关键词表达的方法,利用词语概念和彼此之间的语义联系来构造较为完整的语义层次模型或者语义网络模型[3,4]。

由于该方法可以更好地处理一词多义和多词一义的各种语言现象,因此近年来逐渐受到学者的广泛关注。对于如何表达这种语义特征,伴随着用户本体理论的研究和发展,很多学者都尝试使用 XML数据结构来表达用户个性化信息,并以此构造语义信息更为丰富的用户兴趣模型[5,6]。

目前的研究大多基于传统的 XML标准数据模型,这些方法通常仅利用 XML数据结构中存储的节点语义信息和节点结构信息来表达用户兴趣特征。值得注意的是,这些用户兴趣模型往往直接使用用户注册时或者使用时产生的个性化信息来单独构造层次数据模型,反映不同信息需求的用户兴趣模型通常具有不一样的层次结构。这样做比较直观,却需要对用户兴趣模型和相应的产品信息模型进行相应的数据转换和结构转换,才能进行有效的相似度计算,所以系统的计算开销和复杂度较大。

本文提出和设计了一种基于加权 XML模型的用户个性化推荐系统,可以有效的表达不同用户对不同产品的个性化信息需求,而且利用 XML层次数据结构本身所体现的语义信息,可以很方便的实现语义相关检索和产品推荐服务。

2　用户兴趣模型的基本表达方法

在一个典型电子商务站点的产品信息推荐服务系统中,有大量事先已经建立好的产品分类,借助此分类体系,产品信息可以组织成一个完整的层次数据模型。该模型中的上层节点通常代表着较大的产品类别,而下级节点则代表着具体的类别属性或者具体的产品型号等,如图 1展示了"手机数码"类别下的一些产品相关信息。

可以利用一个 XML层次数据结构来表达这些产品信息。如上述"手机数码"产品信息所对应的 XML数据模型可以分为 5层:根节点层是"手机数码",第二层是大类,可以划分为"手机"、"配件"、"播放录音"和"数码摄像"等,第三层是小类,根据每个大类再次将类别细分,如"播放录音"层次下有"MP3"、"MP4"和"录音笔"等,第四层是品牌类别,对每个小类产品按照生产厂商和品牌的不同来划分,最后一层是具体的产品信息,可以唯一确定一个产品型号。

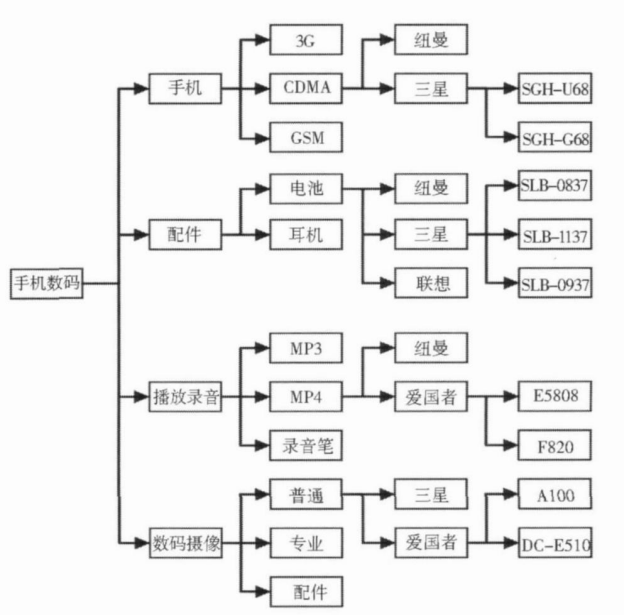

图 1　具有层次结构的"手机数码"产品信息

根据上述产品信息,可以得到相应的产品信息模型,这是一种 XML数据结构。值得注意的是,这些不同产品之间存在着密切的相关关系,如手机产品和电池配件,MP3和耳机等都有密切的相关性。为了表达这种相关关系,可以在对应节点上增加单向引用关系,如在手机节点上增加标记属性: <手机 ref = "电池" > < /手机 >。如上述"播放录音"部分产品的对应产品信息模型例 1所示:

例 1:产品信息模型的基本表达方法

```
<播放录音 ref = "电池" >
  <mp3 > < /mp3 >
  <mp4 >
    <纽曼 > < /纽曼 >
    <爱国者 >
      < E5808 > < /E5808 >
      < F820 > < /F820 >
    < /爱国者 >
  < /mp4 >
  <录音笔 > < /录音笔 >
< /播放录音 >
```

基于这种产品信息模型,可以得到一种用户兴趣模型的新型表达方法,即直接使用产品信息模型来表达用户兴趣模型,通过对特定节点赋予不同的权值来表达不同用户的个性化信息需求,如在每个产品叶节点中添加一个可以反映用户兴趣度的量化属性weight。该属性可以在每个节点中定义,其初始值为0。由于所有的完整用户兴趣模型都具有相同的数据结构,因此为了节省存储和加快计算速度,只需存储用户兴趣模式中的非零节点权值信息。在以后的用户使用过程中,可以通过调整该属性值来表达特定用户的兴趣特征,如例2所示:

例2:用户兴趣模型的基本表达方法

```
<播放录音 ref="电池" weight="1">
  <mp3 weight="0"></mp3>
  <mp4 weight="2">
    <纽曼 weight="2"></纽曼>
    <爱国者 weight="2">
      <E5808 weight="4"></E5808>
      <F820 weight="0"></F820>
    </爱国者>
  </mp4>
  </录音笔 weight="0"></录音笔>
</播放录音>
```

基于该种用户兴趣模型,还需从两个方面来进一步构建完整的个性化信息推荐系统,下文分别对此做出详细说明。

3 用户兴趣模型的权值定义方法

3.1 基本方法

用户兴趣模型中的节点属性权值可以采用多种赋值方法,如直接利用用户访问对应节点产品信息的次数等。为了实现不同用户兴趣模型之间的可比性,此处的节点属性应该进行必要的规范化处理,以消除不同用户访问次数总量不同而产生的绝对值差异。目前,有多种现成的归一化公式可以使用,本文采用了一种较为简单的方法,即使用该用户最大的节点权值去除所有的非零节点权值,如:

$$weight_{nodei}=\frac{weight_{nodei}}{\max(weight_{nodei})} \quad i\in[1,\ldots,n] \qquad (1)$$

n为节点总数。

3.2 反映时变特征

为了准确反映用户兴趣随时间变化的特点,对那些很长时间没有访问过的节点属性权值应该予以适当的降值处理。相应计算方法为:

$$weight_{nodei}=weight_{nodei}\times\left(1-\frac{Time_{Now}-Time_{LastAccess}}{Time_{Now}-Time_{Register}}\right) \quad i\in[1,\ldots,n] \qquad (2)$$

$Time_{Now}$表示当前时间,$Time_{LastAccess}$表示最近一次访问的时间,$Time_{Register}$表示用户注册的时间。显然,如果最近一次用户访问的时间越远,则节点权值的衰减程度越大。

3.3 结合兴趣扩散方法

结合语义扩散方法(Semantic Expansion Approach)可以更为有效的表达用户兴趣模型,该种方法基于认知心理学中的扩散激活模型(Spreading Activation Model),主要应用于基于内容的推荐服务中。它在构建和使用用户模型时,利用已有的信息语义结构对具有相关概念的信息也计算相应的权值,而不是简单的予以忽略。它由两个主要成分组成:扩散激活网络和激活扩散机制。权值可以随着扩散过程的进行而不断地在各个相关节点上累积,从而形成最终稳定的节点权值。每一个节点代表着一个概念,而每一个连接则代表着两个概念间的联系[7]。

扩散可以分为两种方向:垂直方向的扩散,即对上下层次概念的扩散;水平方向的扩散,即对其他层次下的相关概念的扩散。其中,基于垂直方向的扩散过程可以分为概化(Generalization)和细化(Specialization)两个方向。概化主要是指沿着概念层次的上行方向进行信息扩散,而细化则相反,它是沿着概念层次的下行方向进行信息扩散。在本系统中,用户最终选择的都是产品信息,初始权值也针对产品节点来设置,所以基于垂直方向的扩展只有概化一种。而对于水平方向的扩散,可以被称为相关性扩展(Relevance Expansion)[8]。在上述的XML文档结构中,存在着很多具有相互关系的节点,如使用ref属性建立单向连接的节点等。用户在选择此处的某一产品时,权值应该在这些相关节点中自动扩散。在本系统中,水平扩展主要基于节点的ref属性来进行。

在权值扩散中,有三个非常重要的处理内容:

(1)计算扩散值,一般而言,不同情况下的权值扩散值都可以按照当前节点值和扩散系数来得到,其中当前节点值往往为反映用户兴趣的客观数值,如用户访问产品信息的次数,而扩散系数则可以由系统和用

户来自行设定,如为 0.5。

(2)如何决定扩散的范围,这里仍然需要考虑两个子问题:需要指定最大扩散距离和需要指定最小扩散阈值,也就是说,如果某些节点的权值小于该阈值时,扩散过程将会在此节点终止。在系统中,由于用户选择的产品元素都是叶节点,所以采用了只向上的概化扩散策略,同时考虑到层次结构不深和后续计算的要求,对每个节点都将当前节点值的 50%向上层节点扩散,直至抵达根节点为止。可以看出,一个节点的最终权值应该是自己的初始权值加上所有扩散过来的权值之和,即:

$$weight_{nodei}^{final} = weight_{nodei}^{preceding} + \sum_{k} weight_{nodek}^{final} \quad (3)$$

节点 k为当前节点 i包含的所有直接下级一层节点。

(3)迭代计算问题,一般而言,需要在多轮迭代计算后才能得到所有节点的最终值。由于本系统所采用权值扩散主要是从叶节点向上层节点单向进行,因此这种迭代计算并非必须,本系统没有使用这种方法。

3.4　结合正负反馈

不同的用户具有不同的兴趣需求特点,对于非常明确自身需求的用户而言,他们希望可以明确的指定对哪些产品有兴趣或者没有兴趣。这些信息的提交往往发生在用户注册期间,或者在使用期间由很多电子商务站点通过用户评价和反馈的方式来收集此类信息。

系统在此用户兴趣模型中增加了相应的属性 preference。由于这种偏好的设定反映用户较为具体和明确的需求,因此没必要将每个节点都增加此属性,相反,本文所述的方法是在产品信息中增加此属性。如果用户明确的表明对此产品有兴趣,则 preference属性值为“yes”,如果表示没有兴趣,则为“no”,默认情况下此属性值为“none”,反映一种没有确定的状态,如例 3所示:

例 3:增加正负反馈的用户兴趣模型

```
<播放录音 ref="电池" weight="1" >
  <mp3 weight="0" > </mp3 >
  <mp4 weight="2" >
    <纽曼 weight="2" > </纽曼 >
    <爱国者 weight="2" >
      <E5808 weight="4" preference="no" >, /E5808 >
      <F820 weight="0" preference="yes" > </F820 >
    </爱国者 >
  </mp4 >
  <录音笔 weight="0" > </录音笔 >
</播放录音 >
```

为了方便使用和灵活设置,这些正负反馈信息并不直接用于对产品兴趣相关度的计算当中,而是在最终得到用户的推荐产品时,根据当前用户的正负反馈信息,有选择地进行提升和过滤。这样做的主要目的在于一旦用户对此类反馈信息进行了调整,就不需要重新计算产品兴趣的相关度。

4　推荐服务的实现

本系统所采用的推荐服务有两种:

(1)基于单个用户兴趣模型的信息推荐,将最有可能被访问的其他相关未访问产品推荐给用户;

(2)基于协同过滤的推荐内容,通过计算不同用户之间的兴趣模型相似度,进而形成具有相似访问行为和内容的用户组,以组中其他用户的感兴趣内容作为当前用户的推荐内容。

对于第一种情况而言,可以利用当前用户已有的兴趣模型来得到推荐结果。首先可以在当前用户的兴趣模型中,找到访问次数较高的 n个产品项(n为事先定义的常量值)。在产品信息模型中,如果沿着具有最高访问频率的产品逐渐上移,就会得到诸如具有最高访问频率的品牌、小类和大类等信息,这些类别下面的其他未访问产品就可以作为当前用户的推荐产品,最终可以得到一个完整的个性化产品推荐列表。当然,随着上移的层次增多,对该用户的个性化兴趣表达能力和推荐效果也就越差。

对于第二种情况而言,是利用其他相似用户的访问产品来作为当前用户的推荐内容,这种方法充分挖掘了用户群的访问信息,在推荐能力上要比使用单一用户信息更为有效,适合发现那些用户从未访问过但是极有可能被访问的新产品。而且,利用该方法还能识别出具有特定产品需求的用户群体,这有助于电子商务站点开展一些有针对性的产品营销活动。为了实现该种功能,需要采用下面几个基本步骤:计算不同用户访问模型的相似度;对不同用户进行聚类,并得到类中心的兴趣模式;以用户所在类别进行推荐。下面具体予以说明:

(1)计算不同用户访问模型的相似度

在本系统中,可以采用两种方法进行。

(1)提取用户兴趣模式中具有较高访问频率的 XML 节点及其所在路径,形成一棵并不完整的 XML 子树,然后对这些子树进行相似度比较。现有的文献方法多半基于此类方法,大都采用节点内容和 XML 层次结构的双重信息来计算,常见的方法有元素比较法、边集比较法和编辑距离法等[9]。

(2)主要基于本系统中 XML 数据结构的特点。在本系统中,任何用户都使用同样的产品信息模式来表达用户兴趣模式,所以用户兴趣模式的相关语义信息和基本结构形态都是一致的,主要区别则体现在 XML 节点元素的权值上。

所以,在计算这些用户兴趣模式的相似度时,可以设计一种结合加权 XML 模型的相似度计算方法。该方法需要遍历每个用户的非零权值叶节点,也就是说,最终获取的所有节点都是产品节点,此时每个用户可以得到一组所有非零权值叶节点的完整路径。如果两个用户存在具有相同的完整叶节点路径,则反映出具有一定的相似性。如果这种相同路径的数量越多,对应节点的权值越接近,则两个用户的兴趣相似性越大,反之,如果不存在此类路径或者节点权值差异较大,则说明两个用户的兴趣相似性较低。相应的计算公式为:

$$Similarity(User_i, User_j) = \frac{\sum_{k=1}^{n}(1 - |weight_{leafnodeik} - weight_{leafnodejk}|)}{n} \quad (4)$$

式(4)计算了 $User_i$ 和 $User_j$ 两个用户之间的兴趣相似度,其中的 n 为节点总数,对于用户兴趣模型的每个叶节点路径,该公式可以累加它们之间的相似程度。只有具有相同非零权值叶节点并且权值接近的两位用户,才可以获取较高的相似度值,如果两位用户的兴趣模型完全相同,则公式(4)的值为 1。显然,通过这种方式,可以在各种概念层次上对不同用户之间的兴趣进行相似度的比较和测度。

(2)对不同用户进行聚类

基于 XML 数据结构的用户兴趣模型可以采用多种聚类方法,考虑到用户访问行为之间的差异性较大,所形成的类别数量可能较多,因此本文采用了聚合聚类法(Agglomerative Cluster Algorithm)[10]。

它的基本步骤描述如下:

①将所有的用户各自归入一个类 Cuseri,这些类都只具有唯一的成员,整体上形成了一个完整的类别集合:C = {Cuser1, Cuser2, …, Cusern}(假设有 n个用户)。此时,每个类都使用类中用户的非零权值叶节点完整路径集合来表征类用户的兴趣模式。

②使用上述的用户兴趣相似度计算方法分别得到类别集合中两两类别之间的相似度:$Similarity(User_i, User_j)$。选择具有最大相似度的类对并将其合并形成一个新的聚类,该类中就包含原先类中的所有用户成员。此时,可以将合并前两个类的用户兴趣模式进行加权平均以得到这个新类的用户兴趣模式。

③反复迭代进行上述计算,终止条件有两个:一个是 C 类集合只有一个类别元素,在一般情况下,这种条件难以形成而且没有实际意义;另一个是采用类别相似度判断方法,如果在当前一轮的类别相似度计算过程中,所有的相似度都小于事先设定好的一个阈值,即可结束计算。

(3)信息推荐

以用户所在聚类的用户兴趣模式为依据,得到该组用户成员中访问频率最高的 k种产品(k为事先定义好的数值),并以此作为当前用户的推荐产品候选集合。在得到的候选产品集合中,具体的推荐内容有两个:一个是剔除用户已经访问过和对应节点 preference 属性为“no”的产品信息,按照权值的降序排列即可得到最终的推荐结果;另一个是根据当前用户所在组,得到具有最大权值的产品类别节点,将这些产品所在类别的一些诸如销售量最高的产品作为当前用户的推荐产品。

5 结 语

基于以上分析,可以得到一个完整的个性化产品推荐系统原型,其基本结构和推荐流程如图 2所示:

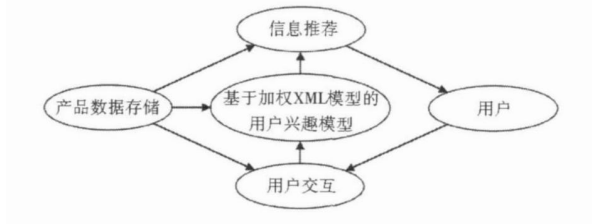

图 2 基于加权 XML 模型的个性化产品推荐流程

其中,“产品数据存储”主要使用关系型数据库来存储所有的分类目录和产品详细信息;“基于加权 XML 模型的用户兴趣模型”利用产品数据信息和用户访问信息以表达用户兴趣特征,也采用关系型数据库来存储,由于不同用户的兴趣模型结构相同,因此在存

储时只保存反映用户兴趣的非零节点权值;“用户交互”模块展示了系统的主要界面和功能,可以收集用户的相关访问信息;“信息推荐”模块是向用户产生推荐内容的最终接口。

笔者发现很多需要改进和进一步研究之处,主要表现为下面三个方面:

(1)在目前的产品信息模型中,水平方向的扩展除使用语义扩展外,还可以考虑更多的扩展形式,如基于购物车分析的扩展,将那些最经常被共同销售的产品建立扩展联系。事实上,整个系统的推荐效果非常依赖于用户兴趣模型的准确性和有效性,笔者准备在下一步的实验中对此进行完善。

(2)部分计算方法需要通过更多的实验加以对比,以期找到更为有效的处理方法,包含节点权值的规范化处理方法和用户兴趣模型的相似度比较方法等。同时,笔者还准备进一步完善用户兴趣模型中的节点权值表示方法,本系统采用的是基于产品信息访问次数的规范化权值,在下一阶段的研究中,计划加入对不同权值定义方法的比较和测试分析,以期找到更为合理的表示形式。

(3)目前采用的推荐算法是聚合聚类法,不可否认,基于 XML数据结构的聚类方法还有很多,有必要对这些不同方法的实验效果做出测试对比,从而得出更有效的聚类算法。

个性化推荐技术是现代 Web网络站点用以提高用户访问满意度的有效方法。本系统的设计和实现旨在探索一种新颖的方法,使用加权 XML模型来改进传统用户兴趣模型的表达能力,从而改善系统的最终推荐效果。

参考文献:

[1] Kobsa. Privacy - enhanced Web Personalization: From the Adaptive Web: Methods and Strategies of Web Personalization Volume 4321 of Lecture Notes in Computer Science[M]. Springer - Verlag, Berlin Heidelbergs, New York, 2007.

[2] Ting P L, Yung F Y, Deng N C. A Semantic - expansion Approach to Personalized Knowledge Recommendation[J]. *Decision Support Systems*, 2008, 45(3): 401 - 412.

[3] Gonzalez R A, Chen N, Dahanayake A. Personalized Information Retrieval and Access[M]. IGI Global, 2008.

[4] Henrik B S. Ontology - based Information Retrieval Computer Science Section[D]. Denmark: Roskilde University, 2006.

[5] Trajkova. Improving Ontology - Based User Profiles[D]. USA: University of Kansas, 2003.

[6] Sure Y, Staab S, Studer R. On - to - Knowledge Methodology (OTKM)[M]. Handbook on Ontologies, Berlin, Heidelberg, 2003: 117 - 132.

[7] Crestani. Application of Spreading Activation Techniques in Information Retrieval[J]. *Artificial Intelligence Review*, 1997, 11(6): 453 - 482.

[8] Middleton, De R, Shadbolt N: Ontology - based Recommender Systems[M]. Handbook on Ontologies. Berlin, Heidelberg, New York, 2004.

[9] 潘有能. XML文档自动聚类研究[J]. *情报学报*, 2006(4): 215 - 220.

[10] 郑仕辉,周傲英,张龙. XML文档的相似测度和结构索引研究[J]. *计算机学报*, 2003, 26(9): 1116 - 1127.

(作者 E - mail: leeshuqing@163.com)