基于贝叶斯网络模型的用户兴趣联合推送*)

欧 洁

(中山大学博士后流动站 广州510275) (广发证券博士后工作站 广州510075)

The Association Push of Interesting Information Based on Bayesian Network Model

OU Jie

(Postdoctoral R & D Base, SUN YAT-SEN University, Guangzhou 510275)

(Postdoctoral R&D Base, GF Securities Co., Ltd., Guangzhou 510075, China)

Abstract The association push of interesting information is implemented according to the information that other users with similar interests have read, therefore user can easily find high quality interesting information from the immense Web information. The association push of interesting information based on Bayesian network model is presented in this paper, it stores the conditional probability and semantic meaning between the terms, so the similarity between the users' interests and the similarity between the user's interest and document can be computed according to the semantic meaning of the terms, therefore the association push of interesting information can be implemented. The experience result indicates this method is effective.

Keywords Bayesian network model, Association push of interesting information, User interests model, Relevance terms

1 引言

1991年,WWW(World Wide Web)出现,在随后的几年中它获得了空前的发展,Internet上的信息量以指数形式飞速增长。现在,Internet已成为一个浩瀚的海量信息源,但由于Internet是一个具有开放性、动态性和异构性的全球分布式网络,其中的信息良莠不齐,非常缺乏关于信息质量和内容的描述,这就导致了用户获取所需高质量信息的困难。如何快速、准确地从浩瀚的信息资源中寻找到所需的高质量信息已经成为困扰用户的一大难题。

传统的信息提取系统将WWW上的内容尽量多地建立索引并存入检索数据库,其查询服务只是负责接收、解释用户的查询,然后根据较为简单的匹配策略(使用较多的是简单布尔模型和模糊布尔模型)在索引库中进行查找,将结果地址集反馈给用户。对于这样的一个系统,存在如下问题:首先,它要求用户把检索意图用有限的几个关键字表达出来,这经常会造成用户表达歧义和表达不全面的问题,其次,对于给定的查询项,要实现用户所要求的高检索率和高检索精度是非常困难的。高检索率即找到大部分与用户感兴趣的主题相关的文档,高检索精度即尽量不包含与主题不相关的文档。并且其搜索结果的排序并没有考虑检索结果的质量。

为了解决以上问题,实现信息提取的智能化,帮助用户快速、有效获取所需要的信息,人们将人工智能技术引入到信息提取中,已经研究出了各种智能信息提取方法,提出了一些智能型信息检索系统模型[1~3]。个性化智能信息服务即根据用户的兴趣将信息主动推荐给用户以及过滤查询结果,但是,这些智能型信息检索系统均未使用用户兴趣联合推送方法。用户兴趣联合推送是一项重要的智能信息服务,它根据其他具有相似兴趣的用户的阅读信息主动推送信息,从而使用户能够很容易地从Web浩瀚的信息海洋中发现所需的高质量内容。

本文所提出的基于贝叶斯网络模型的用户兴趣联合推送系统采用多个特征向量组成用户兴趣模型,不同的特征向量代表用户的不同兴趣,如专业、音乐等,并使用相关反馈学习算法更新用户兴趣模型,智能化适应用户兴趣的变化。它利用贝叶斯网络模型表示术语间的概念语义关系,从而预测用户兴趣模型间以及用户兴趣模型和文档间的概念语义相似度,并根据其他兴趣类似的用户的阅读信息主动推送用户感兴趣的文档,该系统把这些信息根据用户的不同兴趣聚成不同的类,向用户推送聚类后的信息。该方法在计算推送信息和用户兴趣模型间的相似度时考虑的因素有:阅读过该信息的用户的个数;这些用户和被推送用户间的兴趣相似度;该信息和被推送用户的兴趣模型间的相似度。因为阅读过该信息的用户的个数反映了该信息的质量,阅读过该信息的用户与被推送用户间的兴趣相似度以及该信息与被推送用户的兴趣模型间的相似度反映了用户对该信息感兴趣的程度,所以能推送用户感兴趣的高质量信息。

本文的第2节介绍我们所采用的用户兴趣建模方法,第3节介绍我们所采用的基于贝叶斯网络模型的相似度计算,第4节介绍基于贝叶斯网络模型的用户兴趣联合推送,在第5节,将给出这种方法的实验结果,最后是结论。

2 用户兴趣建模

用户兴趣联合推送系统所实现的个性化信息服务为:根据其他具有相似兴趣的用户的阅读信息主动推送信息,从而使用户能够很容易地从Web浩瀚的信息海洋中发现所需的

*)本文工作得到中科院计算所创新课题:"高性能媒体服务器"和国家863计划课题"中国数字图书馆示范系统"的支持,课题编号:863-306-ZD11-03。

高质量内容。它的实现过程包括：(1)用户兴趣建模并存储其阅读信息。(2)将其它用户阅读过且被推送用户未阅读过的信息与被推送用户的兴趣模型相比较，把与被推送用户兴趣相符合的信息主动推送给用户。(3)根据用户的兴趣模型对推送结果进行排列，考虑的因素有：阅读过该信息的用户的个数；这些用户和被推送用户间的兴趣相似度；该信息和被推送用户的兴趣模型间的相似度。

因为这里所指的个性化信息服务是由计算机来完成的，所以信息和用户兴趣都必须以一定的方式在计算机内表示出来，这就是目标表示。

2.1 目标表示

目标表示是指从目标信息(如用户兴趣或文档)中选取一些特征项(如术语等)，用这些特征项及其在目标信息中的重要性来代表目标信息。目标表示模型有多种，我们采用的是向量空间模型(VSM)法，其实现步骤为：(1)选取一组适合于表示目标信息的术语($T_1,T_2,\cdots,T_n$)。(2)对于目标信息 F，根据术语 T_i 在 F 中的重要程度求出权值 $W_i,i=1,\cdots,n$，然后把目标信息 F 用术语特征矢量($W_1,W_2,\cdots,W_n$)表示出来，W_i 表示 T_i 在 F 中的权重，$i=1,\cdots,n$。我们根据术语在文档中的属性及其文件频度和反频度将文档表示为术语特征矢量。如果术语 T_i 出现在文档的标题和作者中，那么其权重为1。如果术语出现在文档 F 的正文中，则根据术语的文件频度和反频度计算其权重，计算公式如下[4]：

$$W_i=\frac{(0.5+0.5\frac{tf(i)}{tf_{\max}})(\log\frac{n}{df(i)})}{\sqrt{\sum_{T_j\in F}((0.5+0.5\frac{tf(j)}{tf_{\max}})^2(\log\frac{n}{df(j)})^2)}}$$

其中，tf(i)为术语 T_i 在文档 F 中的出现次数，df(i)是包含术语 T_i 的文档数，n 是全部文档数目，$tf_{\max}$ 是文档 F 中出现最多的术语的出现次数。

通过用户兴趣建模方法将用户兴趣表示为多个术语特征矢量，用户兴趣模型中的术语特征矢量也称为个性化向量，我们采用的用户兴趣建模方法请见2.2—2.6节。当我们将文档和用户兴趣都表示为术语特征矢量以后，采用贝叶斯网络模型预测用户兴趣间以及用户兴趣模型和文档间的概念语义相似度，具体请见第3节。

2.2 用户兴趣建模

本系统所采用的用户兴趣建模的方法为：1. 通过用户主动提供自己的兴趣来得到用户的初始兴趣模型。2. 通过用户对搜索结果的反馈信息来更新用户的兴趣模型。3. 在用户没有明确参与的情况下，系统通过观察用户行为来更新用户的兴趣模型。并将用户阅读过的信息的元数据保存起来。因为用户阅读的信息与用户的兴趣模型中的某一个个性化向量相关，所以我们将其元数据存储到与这个个性化向量对应的阅读信息中。

为了让用户主动提供自己的兴趣，我们让用户回答一些问题，如：(1)从事的专业、研究兴趣和研究方向。(2)除了自己的专业外经常阅读哪些专业和研究方向的资料。(3)对哪些种类的音乐、体育、小说、电影感兴趣?等等。因为用户的回答一般比较简短，所以对其进行目标表示的方法为：从用户的回答中提取术语，将其初始权重设为1。对用户输入的答案进行目标表示后，将不同问题的答案形成不同的用户个性化向量，表示用户的各种兴趣。这样就可以避免将各种不同兴趣表示成一个用户个性化向量的缺陷，如只与某一项兴趣有关的文档与用户的兴趣模型的相关性小等。因此，用户的兴趣模型由多个个性化向量组成。

通过这种方法，我们可以快速地得到用户的兴趣，而且实现简单。但是，由于用户的兴趣并不是一成不变的，让用户每次访问网站时都输入这些内容会使用户觉得很繁琐，而且用户一般不可能提供所有的兴趣以及感兴趣的程度。所以本系统还采用了另外两种方法来更新用户兴趣模型。

本系统还通过用户已访问过的页面去挖掘用户兴趣。一般来说，用户只会访问自己感兴趣的页面，所以，我们认为用户访问过的页面中包含用户感兴趣的信息，可以通过对这些页面进行分析来更新用户的兴趣模型。这种方法的实现要用到页面访问挖掘，详细描述请见2.3—2.5节。

本系统所采用的另外一种更新用户兴趣模型的方法为：根据用户对推送信息和检索结果的评价来更新用户的兴趣模型。这种方法实现简单，但需要用户对推送页面做出评价。详细描述请见2.6节。

本系统结合这三种方法，因此可以快速、精确地发现用户兴趣。

2.3 从日志文件中发现用户的访问信息

用户与 Web 服务器间交互的过程信息(包括用户的登录信息、用户的浏览记录)都以日志文件的形式存在，分析这些日志文件可以发现用户已浏览过的页面集和浏览这些页面的时间长度。这个过程要完成的任务为：将日志文件中的信息按用户进行分类整理，得到每一个用户的访问情况库，库中的记录表示页面的访问信息，包括的内容有：页面的 URL 地址、页面的访问日期和访问页面的时间长度等。这些记录按访问时间的先后进行排序。访问情况库的内容可用以下三元组来描述[5]：

$$t=\langle ip_i,uid_i,\{(l'_1.url,l'_1.time,l'_1.length),\cdots,(l'_m.url,l'_m.time,l'_m.length)\}\rangle$$

其中，ip_i 表示用户的 IP 地址，uid_i 表示用户的 ID，对于任意的$0<k<m+1$，l'_k 表示用户在该事件中访问的第 k 个页面，$l'_k.url$ 表示页面 l'_k 的 URL 地址，$l'_k.time$ 表示用户访问 l'_k 的日期，$l'_k.length$ 表示用户访问 l'_k 的时间长度，$l'_1.time<l'_2.time<\cdots<l'_m.time$。

2.4 发现内容页面

用户浏览 Internet 上的信息的主要方法为：打开信息所在站点的主页面，点击一系列链接来打开信息所在的页面。这样，用户为了访问自己需要的信息所在的页面(内容页面)，一般都需要经过一系列其它的页面(辅助页面)。为了发现用户的兴趣，只需要对内容页面进行挖掘，所以要区分用户访问的页面哪些是内容页面，那些是辅助页面。我们根据访问时间的长短来区分内容页面和辅助页面。实现方法为：设定一个阈值 a，访问时间超过 a 的认为是内容页面，访问时间小于 a 的认为是辅助页面。对内容页面进行目标表示的方法是 VSM 方法，具体见2.1节目标表示中的文档目标表示方法。

2.5 根据用户的页面访问信息更新用户库的方法

我们提出了如下根据用户的页面访问信息更新用户库的方法。

设 $M_1,\cdots,M_n$ 表示用户兴趣模型中的所有个性化向量，向量 $V_1,\cdots,V_m$ 表示用户已访问过的所有内容页面。那么，如果页面 V_i 与用户的某一项兴趣 M_j 最接近，且 $V_i\cdot M_j/|M_j|$ 的值大于阈值 a(我们取 a=0.15)，则用该页面对应的向量加到

用户的兴趣 M_j 中，即：$M_j=M_j+\frac{m^2*V_i*VI(i)*T(i)}{\sum_{k=1}^{m}VI(k)*\sum_{k=1}^{m}T(k)}$，$VI(i)$表示用户访问页面 V_i 的频度，$T(i)$表示用户访问页面 V_i 的时间长度，$i=1,\cdots,m$。因此，用户访问页面的信息加强了用户的相应兴趣。如果页面 V_i 不与用户的任何一项兴趣接近的话，则将分量值对应为 $M_j=M_j+\frac{m^2*V_i*VI(i)*T(i)}{\sum_{k=1}^{m}VI(k)*\sum_{k=1}^{m}T(k)}$ 的向量作为用户的新兴趣，即增加一个个性化向量。系统用页面 V_i 更新用户兴趣模型后，将页面 V_i 的元数据存储到与被更新的个性化向量对应的阅读信息中。当然这样可能会导致用户的兴趣种类大量增加，所以设定一个阈值 b，如果用户的某一项兴趣对应向量的模小于 b 时，在推送相关页面时不考虑该项兴趣。我们取 b=0.2。

2.6 根据用户反馈的信息更新用户兴趣模型的方法

如果用户对被推荐页面进行了评价的话，可以根据用户的评价信息来更新用户兴趣模型。例如用户对页面的评价标准可以设为：正好符合要求、相关、无关，对应的用于更新用户个性化向量的系数为：正好符合要求：+0.2，相关：+0.1，无关：−0.2。设 $M_1,\cdots,M_n$ 表示用户兴趣模型中的所有个性化向量，向量 $V_1,\cdots,V_n$ 表示用户评价过的所有内容。那么，如果 V_i 与 M_j 最接近，且 $V_i\cdot M_j/|M_j|$ 的值大于阈值 a（我们取 $a=0.15$），则将 V_i 乘上用户对它的评价系数后累加到用户的兴趣 M_j 中，即：$M_j=M_j+e_i\cdot v_i$，e_i 表示用户对页面 V_i 的评价。通过上式可知，用户对页面的评价反映到了用户的相应兴趣中。如果页面 V_i 不与用户的任何一项兴趣接近的话，则将分量值对应为 $e_i\cdot v_i$ 的向量作为用户的新兴趣。系统用页面 V_i 更新用户兴趣模型后，将页面 V_i 的元数据存储到与被更新的个性化向量对应的阅读信息中。

2.7 用户兴趣联合推送服务

用户兴趣联合推送包括直接推送和间接推送。直接推送即推送与自己兴趣相近的用户所阅读的文档，间接推送即在对文档进行排序时考虑该文档的点击数和推送点击数高的文档。由于用户的兴趣模型由多个个性化向量组成，所以兴趣相近的判断准则是某一个个性化向量比较接近，推送与这个个性化向量对应的阅读文档，不同的个性化向量对应的推送信息形成不同的类，同一类中的推送信息按其与用户兴趣模型间的相似度大小排序。我们通过贝叶斯网络模型预测用户兴趣间以及用户兴趣模型和文档间的概念语义相似度。其描述请见第3节。

3 基于贝叶斯网络模型的相似度计算

贝叶斯网络是一个带有概率注释的有向无环图，图中的点表示所要解决的问题中的变量。这种概率图模型能表示变量之间的联合概率分布，分析变量之间的相互关系，利用贝叶斯定理揭示的学习和统计推断功能，可以实现预测、分类、聚类等任务[6,7]。在用户兴趣联合推送系统中，主要是利用贝叶斯网络模型表示术语间的关系，预测用户兴趣模型间以及用户兴趣模型和文档间的概念语义相似度，从而实现基于语义概念的信息主动推送服务[8,9]。我们采用贝叶斯网络模型表示术语间的语义关系，并将此关系反映在术语的权重中。为了减少计算复杂度，我们使用向量内积的方法计算用户兴趣模型间以及用户兴趣模型和文档间的相似度。

因为用户的兴趣模型由多个个性化向量组成，该系统依次取出这些个性化向量，采用贝叶斯网络模型计算该个性化向量与其它用户兴趣模型间的相似度以及该个性化向量与文档间的相似度，依此推送与该个性化向量接近的高质量信息，这些信息形成一个类。用户的兴趣模型包含多个个性化向量，因此，多个个性化向量对应的推送信息形成多个类，系统将这些类推送给用户。个性化向量与文档间的相似度的计算方法见3.1—3.4节。个性化向量与用户兴趣模型间的相似度定义为：该个性化向量与用户兴趣模型的多个个性化向量间的相似度中的最大值，其计算方法见3.5节。

3.1 基于贝叶斯网络模型的用户个性化向量与文档间的相似度计算

用于计算用户个性化向量和文档间的相似度的贝叶斯网络模型中包含两种类型的节点，一种节点代表术语，一种节点代表文档。每一个节点的取值范围为[0,1]，它代表术语的权重或文档与用户个性化向量的相似度。由代表文档的节点组成的集合称为文档子集，类似地，由代表术语的节点所组成的集合称为术语子集。

贝叶斯网络模型是有向图，其中有两种类型的节点，所以在这两种类型的节点间的连线可以有两种方向，一种是从文档子集到术语子集，一种是从术语子集到文档子集。我们认为在给定了用户个性化向量后再计算文档与术语间的相似度更为直观，而且由用户个性化向量中的术语出发进行统计推断所需要的计算复杂度较小。因此，我们采用的连线方向为从术语子集到文档子集。

我们认为文档间的关系可以从它们所包含的术语中推导出来，所以在我们的贝叶斯网络模型中文档之间是没有连线的。这种假设表示在给定了文档中的术语后，文档是条件独立的。因此，在我们的贝叶斯网络模型中只有两类连线，一类是从术语到术语，表示术语之间的关系，连线所对应的值为术语间的条件概率。一类连线是从术语到文档，它代表文档和术语间的包含关系，连线所对应的值为术语在该文档中的权重。如果术语 A 在文档 B 中的权重大于0，那么从术语 A 到文档 B 有一条连线。

我们的贝叶斯网络模型借鉴了文[10,11]，它分为三层，最上层的节点代表术语，表示用户的兴趣项及其权重，第二层的节点也代表术语，它们与第一层中的术语相关，其值代表相关程度和权重。第三层的节点代表文档，节点中的值表示文档与用户个性化向量的相关程度。其示意图如下：

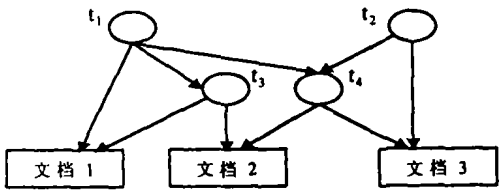

图1 贝叶斯网络模型的示意图

最上层的节点及其权重通过用户个性化向量可以直接得到，下面我们主要讨论如何求得从最上层到第二层的节点的连线及其对应值，通过这些连线求出第二层的节点的权重，以及通过第一层、第二层的节点求第三层的文档的相关度。

3.2 计算术语间的条件概率

我们认为经常同时出现于一个文件中的术语具有某种相关关系，如上下位关系、所属领域相同关系等，因此，我们将之称为相关术语。因为文档的标题是文档内容的概括，而且通常

由一些与领域有关的术语组成，所以通过对文档标题的挖掘可得术语间的相关关系。我们挖掘相关术语的依据为经常同时出现在文档的标题中的术语。根据关联规则得到的相关术语不一定满足同义或近义关系，所以我们还使用同义词典的方法定义同义或近义的术语。

我们运用关联规则发现的方法和同义词典的方法挖掘出术语之间的相关关系，这种相关关系表示了术语间的概念语义关系。我们的贝叶斯网络模型利用了这种概念语义关系，所以能实现基于概念语义的查询。

下面首先介绍如何运用数据挖掘中的关联规则发现方法挖掘术语之间的关系，运用同义词典的方法挖掘术语之间的关系将在3.3节中介绍。

在运用数据挖掘中的关联规则发现方法挖掘术语之间的关系时，术语对应于项目(Item)，文档的标题对应于交易(Transaction)。如果项目集的频度(frequency)大于2，我们认为它是频繁项目集。用关联规则发现的方法挖掘出项目数目为2和1的频繁项目集，存储这些频繁项目集及其文件频度，依此计算出连线对应的值和节点的权重。

为了动态生成贝叶斯网络模型，对于所有可能的用户个性化向量，为了计算出第二层节点的权重，有一种方法是存储所有贝叶斯网络模型中由第一层节点到第二层节点的联合条件概率，这个数目为：$n+C_n^2+C_n^3+\cdots+C_n^m$，其中n是术语的个数，m是所有用户的个性化向量中术语的数目。显然这个数目很大，保存这个数目需要很多存储空间，而且在其基础上动态生成贝叶斯网络模型也很费时。因此，我们采用了另外的方法来计算术语之间的联合条件概率。

我们通过项目数为2和1的频繁项目集及其文件频度来建立贝叶斯网络模型。术语之间的连线表示术语间的关联规则，连线对应的值为这个规则的置信度，其取值范围为[0,1]，它反映了术语间的相关程度。我们采用以下计算公式来计算术语间的条件概率(节点A的权重为通过所有与A相关的节点计算出来的权重及节点A的原始权重中的最大值。)。

$$P(T_j|T_{i1},\cdots,T_{im})=\max(w(T_j),P(T_j|T_{i1})*w(T_{i1}),\cdots,P(T_j|T_{im})*w(T_{im}))$$

其中，术语$T_{i1},\cdots,T_{im}$是用户的兴趣项，它们对应于贝叶斯网络模型中的第一层的节点。$P(T_j|T_{ik})$表示规则$T_{ik}\rightarrow T_j$的置信度，即$n(T_j,T_{ik})/n(T_{ik})$，$n(T_j,T_{ik})$是项目集(T_j,T_{ik})的文件频度，$n(T_{ik})$是项目集(T_{ik})的文件频度。$w(T_{ik})$表示术语T_{ik}的权重。如果项目集(T_j,T_{ik})不是频繁项目集，则$P(T_j|T_{ik})=0$，$k=1,\cdots,m$。因为$T_{i1},\cdots,T_{im}$是用户的兴趣项，所以$P(T_j|T_{i1},\cdots,T_{im})$就是术语$T_j$的权重。

为了用这个方法进行联合条件概率计算，只需要存储所有项目数为2和1的频繁项目集及其文件频度，大大减少了存储量。

3.3 通过同义词典计算术语的权重

显然，语义相同或相近的术语从概念语义上表达了用户的兴趣，因此，我们还通过同义词典来挖掘术语之间的相关关系，同义词典是将术语按照语义相同或相近的关系进行分类后形成的词典。我们认为语义相同的术语应该对于用户兴趣具有相同的权重，其值为所有同义术语的权重中的最大值。用3.2节的方法计算出了术语的权重后，用同义词典计算术语的权重的方法如下：

①将所有术语标识为未通过同义词典的方法计算其权重。②对于任意权重不为0的术语A且A未通过同义词典的方法计算其权重，通过同义词典的方法求出所有与术语A语义相同的术语，将之记为$\{A_1,A_2,\cdots,A_m\}$。③将术语A，A_1，$A_2,\cdots,A_m$的权重赋值为$\max(w(A),w(A_1),w(A_2),\cdots,w(A_m))$，其中$w(A)$代表术语A的权重，$w(A_i)$代表术语$A_i$的权重，$i=1,\cdots,m$。④将术语A，$A_1,A_2,\cdots,A_m$标识为已通过同义词典的方法计算其权重。重复步骤②，③，④，直到所有权重非零的术语均被标识为已通过同义词典的方法计算其权重为止。

因为同义关系是定义在由所有术语组成的集合上的一个关系，显然同义关系是自反的、对称的和传递的，所以同义关系是等价关系。上面的计算方法即依次计算由术语所形成的同义关系等价类中的术语的权重，遍历所有同义关系等价类后，所有的术语权重均已通过同义词典的方法计算出来。

3.4 计算术语与文档间的条件概率

为了计算术语与文档间的条件概率，我们采用了文[12]的计算方法，即将术语和文档间的条件概率表示为术语向量和文档向量间的向量点积。

用3.2和3.3节的方法对用户个性化向量进行基于贝叶斯网络模型的扩展后，可以求出所有术语的权重，从而将用户个性化向量用术语向量表示出来。对于文档D，根据术语T_i在D中的重要程度求出权值W_i，$i=1,\cdots,n$，这样就可以把文档D用术语特征矢量$(W_1,W_2,\cdots,W_n)$表示出来，W_i表示T_i在D中的权重，$i=1,\cdots,n$。我们用这两个向量的内积表示文档D与用户个性化向量的相似度。

3.5 计算个性化向量与用户兴趣模型间的相似度

在3.2和3.3节中我们介绍了用贝叶斯网络模型对用户个性化向量进行基于概念语义扩展的方法。本节中我们介绍在运用3.2和3.3节的方法的基础上计算用户个性化向量与其他用户兴趣模型间的相似度的方法。我们将个性化向量与用户兴趣模型间的相似度定义为：该个性化向量与用户兴趣模型中的多个个性化向量间的相似度中的最大值，从而将个性化向量与用户兴趣模型间的相似度计算转化为两个用户个性化向量间的相似度计算。

因为待计算相似度的两个用户个性化向量均可以通过贝叶斯网络模型进行基于概念语义扩展，所以我们先将它们扩展成术语向量后，再用这两个术语向量间的内积表示这两个用户个性化向量间的相似度。

4 基于贝叶斯网络模型的用户兴趣联合推送

用户兴趣联合推送系统根据其他具有相似兴趣的用户的阅读信息主动推送信息，从而使用户能够很容易地从Web浩瀚的信息海洋中发现所需的高质量内容。本文提出了基于贝叶斯网络模型的用户兴趣联合推送系统，它利用贝叶斯网络模型表示术语间的条件概率和概念语义关系，从而预测用户兴趣间以及用户兴趣模型和文档间的概念语义相似度，依此进行用户兴趣联合推送。

本文所提出的基于贝叶斯网络模型的用户兴趣联合推送系统的实现过程如下：设目标用户为A，对于用户A没有阅读过且被其它用户阅读过的任一文件B，设阅读过文件B的用户为：$A_1,\cdots,A_k$，用户$A_1,\cdots,A_k$的兴趣模型中与文件B对应的个性化向量为：$C_1,\cdots,C_m$。我们设用户A对文件B的感兴趣的程度为：$C_1,\cdots,C_m$与用户A的兴趣模型相似度之和，其计算方法见3.5节，再乘上文档B和用户A的兴趣模型中的多个个性化向量间的相似度中的最大值，并且文件B属于

与其相似度最大的这个个性化向量所形成的类。这样，我们在计算目标用户对文档的感兴趣程度时考虑了阅读该文档的用户数目、阅读过该文档的用户与目标用户的兴趣相似度以及文档与目标用户的兴趣模型间的相似度。因而可以分类推送用户感兴趣的高质量信息。

设用户 A 的兴趣模型中与文档 B 最接近的个性化向量为 D，则将文档 B 聚类在个性化向量 D 对应的类中，将所有用户 A 没有阅读过且被其他用户阅读过的文件聚类，同一类中的文件按用户 A 感兴趣的程度大小排序，因而可以分类推送用户感兴趣的高质量信息。

5 实验结果

我们所选的实验数据为786篇信息提取和数字图书馆方面的文章，数据来源主要为 CIIR 所发表的文章和数字图书馆杂志中的文章。这些文档的标题中包括术语1816个。

我们选择了10个用户，其兴趣包括基于贝叶斯网络模型的信息检索、基于概率模型的信息检索、分布式信息检索、文档聚类、分布式数字图书馆、文档分类、语义检索等。

我们衡量用户兴趣联合推送系统的性能的度量标准为：用户阅读了的推送文档的个数，以及这些被用户阅读了的文档在被推送时的排列位置及其排列位置之和。

因为其它的智能信息检索系统没有使用用户兴趣联合推送方法，所以本文比较了基于贝叶斯网络模型的用户兴趣联合推送系统和不基于贝叶斯网络模型的用户兴趣联合推送系统的实验效果。从实验结果可以看出，基于贝叶斯网络模型的用户兴趣联合推送系统优于不基于贝叶斯网络模型的用户兴趣联合推送系统，并且用户阅读了的文档个数较多，说明我们的用户兴趣联合推送方法取得了很好的效果。其实验结果如下：

在基于贝叶斯网络模型的用户兴趣联合推送系统中，用户阅读了的推送文档的个数平均为：5.1，其排列位置之和平均为：30。在不基于贝叶斯网络模型的用户兴趣联合推送系统中，用户阅读了的推送文档的个数平均为：4.6，其排列位置之和平均为：30.2。

结论 本文所提出的基于贝叶斯网络模型的用户兴趣联合推送系统采用多个特征向量组成用户兴趣模型，并使用相关反馈和浏览信息更新用户兴趣模型，智能化适应用户兴趣的变化。它利用贝叶斯网络模型表示术语间的概念语义关系，从而预测用户兴趣模型间以及用户兴趣模型和文档间的概念语义相似度，依此根据其他兴趣类似的用户的阅读信息主动推送用户感兴趣的文档。该方法在计算推送信息和用户兴趣的相似度时考虑的因素有：阅读过该信息的用户数；这些用户和被推送用户间的兴趣相似度；该信息和被推送用户的兴趣模型间的相似度。因为阅读过该信息的用户的个数反映了该信息的质量，阅读过该信息的用户和被推送用户间的兴趣相似度以及该信息和被推送用户的兴趣模型间的相似度反映了用户对该信息感兴趣的程度，所以可以推送用户感兴趣的高质量信息。

参 考 文 献

1 Joachims T, Freitag D, Mitchell T. Web Watcher: A Tour Guide for the WWW. In: Proc. of IJCAI97, Aug. 1997, 1～6

2 汪晓岩，胡庆生，李斌，庄镇泉. 面向 Internet 的个性化智能信息提取. 计算机研究与发展，1999，36(9)：1039～1046

3 Chen L, Sycara K. WebMate: A Personal Agent for Browsing and Searching. In: Proc. of the 2nd Intl. Conf. on Autonomous Agents and Multi Agent Systems, May 1998. 132～139

4 Balabanovic M, Shoham Y, Yun Y. An Adaptive Agent for Automated Web Browsing. Journal of Visual Communication and Image Representation, 1995, 6(4)

5 Mobasher B, et al. Web Mining: Pattern Discovery from World Wide Web Transactions: [Technical Report TR96-050]. Department of Computer Science, University of Minnesota, 1996

6 林士敏，田凤占，陆玉昌. 贝叶斯学习、贝叶斯网络与数据采掘. 计算机科学，2000，27(10)：69～72

7 Callan J, Lu Z, Croft W. Searching distributed collections with inference networks. In: Proc of ACM SIGIR Conf. Seattle, 1995. 21～28

8 Fung R, Del Favero B. Applying Bayesian Networks to Information Retrieval. Communication of ACM, 1995, 38(3): 42～57

9 Indrawan M, Ghazfa D, Srinviasan B. Using Bayesian Networks as Retrieval Engines. In: Proc. of 6th Text Retrieval Conference. NIST Special Publication 500-238, 1996. 437～444

10 de Campos L M, Fernandez J M, Huete J F. Building Bayesian Network-Based Information Retrieval Systems. In: Proc. of the 11th Intl. Workshop on Database and Expert Systems Applications (DEXA'00)

11 de Campos L M, Huete J F. Document Instantiation for Relevance Feedback in the Bayesian Network Retrieval Model. ACM SIGIR' 01 Workshop on Mathematical/Formal Methods in IR, New Orleans, Louisiana, 2001

12 Rbeiro B, Muntz R. A belief network model for IR. In: Proc. of 19th ACM-SIGIR Conf. 1996

（上接第68页）

了理论推导公式，最后还对 RS6000服务器群中接入控制的实现进行了说明。

在后继的工作中，我们将进一步探讨自相似性理论[9]对接入控制的影响。此外，基于策略的接入控制以及无线、移动系统的接入控制也是值得研究的方向。

参 考 文 献

1 Cleary K. Video on Demand - Competing Technologies and Services. In: Proc. of International Broadcasting Convention, Sept. 1995. 432～437

2 http://www.nCube.com

3 Jamin S, Shenker S, Danzig P. Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service. In: Proc. of INFOCOM '97, April 1997

4 Perros H, Elsayed K. Call Admission Control Schemes: A Review. IEEE Magazine on Communications, special issue on Congestion Control, Nov. 1996. 82～91

5 Chang E, Zakhor A. Cost Analyses for VBR Video Servers. IST/SPIE Multimedia Computing and Networking, San Jose, Jan. 1996. 381～397

6 Ng T R, Yang Jinhai. Maximizing buffer and disk utilization for news-on-demand. In: Proc. of 20th Intl. Conf. on Very Large Data Base, Santiagp, Chile, Sept. 1994. 451～461

7 周笑波，谢立. Rembard Lueling. 视频服务器网络中影像对象映射问题的研究. 软件学报，Vol. 11(12)：1620～1627

8 华兴. 排队论与随机服务系统. 上海：上海翻译出版社，89～112

9 Leland W, Taqqu M, Willinger W, Wilson D. On the Self-Similar Nature of Ethernet Traffic. IEEE/ACM Transactions on Networking, Feb. 1994. 1～15