基于信度网的网站在线智能导航*)

Build the Navigation System for the Web Site Based on the Belief Networks

邢永康　沈一栋

(重庆大学计算机科学与工程学院　重庆400044)

Abstract　A Web site always provides a large number of topics, so the browser is prone to lost the Way. In this paper we build a Navigation system for the Web site by using the model of Belief networks. It can help the browser to find the topics that he is interesting to quikly. It also can be used to optimize the structure of the Web site.

Keywords　Belief network, Navigation system, Intelligent super link

一、引言

随着计算机网络的发展，网站提供的信息越来越丰富。一般将那些围绕一个中心思想的信息组织在一起，称为一个主题(Topic)。各个相关主题通过超级链接互相联系。如果用结点代表一个主题，两个主题之间的超级链接用一条有向边表示，则一个网站的信息结构可以抽象表示为一个复杂的有向图，称为网站的信息结构图。如图1就是一个信息结构图。

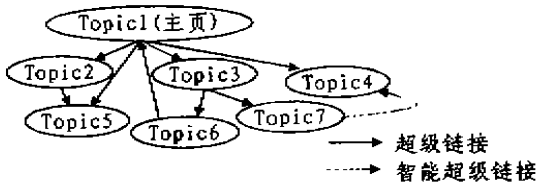

图1　网站信息结构图

浏览者在网站上的浏览过程，就是沿着该信息结构图中的有向边进行搜索的过程。因此，网站的信息结构图是一种简单的网站导航系统。网站设计者根据自己对各个主题之间相关性的理解，通过超级链接来引导浏览者在网站上浏览。然而，这种简单的网站导航方式存在以下的缺点：

1　当网站的主题较多时，网站的信息结构图将非常复杂。在这样复杂的结构图中浏览，浏览者仍然常常会迷失方向。

2　超级链接体现了网站设计者对各个主题之间相关性的理解，而对于某些主题之间潜在的相关性，网站设计者无法预先知道，所以就无法为其建立超级链接。如在图1中，当浏览者经过主页、Topic3到达 Topic7时，浏览者可能对 Topic4感兴趣。但对于 topic7与Topic4之间的这种潜在相关性，网站设计者预先并未发现，所以没有提供超级链接。因此，浏览者很可能浏览不到 Topic4。

如果能够根据浏览者的喜好，猜测出浏览者的兴趣爱好，动态地提供 Topic7和 Topic4之间的超级链接，则可以方便用户的浏览。为了区别于传统的超级链接，我们将这种具有智能性的动态超级链接称为智能超级链接。

定义1　智能超级链接不同于传统的超级链接，它存在于浏览者的一次浏览对话中，是通过猜测浏览者的兴趣爱好所产生的一个临时超级链接。利用智能超级链接，可以帮助浏览者快速找到自己感兴趣的浏览主题。

智能超级链接具有以下特点：

·智能超级链接具有动态性。它只存在于一个浏览者的一次浏览对话中。对于不同的浏览者，或者一个浏览者的两次浏览，由于浏览的目的不同，其智能超级链接也应不同。

·智能超级链接具有潜在性。网站的设计者在开始设计网站的信息结构时，无法发现主题间的这类相关性。

发现智能超级链接，并在线为浏览者同步提供引导，就称为在线智能导航。网站通过提供在线智能导航，可以使浏览者快速找到自己感兴趣的主题。在电子商务网站中，如果以每一个具体的商品作为一个主题，

*)国家自然科学基金及教育部跨世纪优秀人才基金资助项目。邢永康　博士生，研究方向：人工智能、知识工程。沈一栋　教授、博士生导师，研究方向：人工智能。

则这种智能导航就是一种智能导购过程。因为智能超级链接是根据用户的购买倾向产生的，所以符合用户的兴趣爱好，更容易满足用户的购买喜好，使用户主动购买推荐的商品。

因此，在线智能导航的主要任务是发现并提供智能超级链接。根据定义可知，智能超级链接体现了主题之间的潜在相关性。对于这种相关性，网站的设计者无法预先设定，只能通过分析网站的历史浏览数据来获得。

网站的历史浏览数据可表示为如图2的形式，称为浏览信息数据库。它是一个以所有的需要进行导航的主题为列标题的表，表中的每一个纪录对应一个浏览者的一次浏览信息。在这次浏览中，如果浏览者浏览过这某一主题，则该主题对应的字段值为1，否则为0。

序号	Topic1	Topic2	…	TopicN
1	1	0	…	1
2	0	0	…	1
m	1	1	…	1

图2 浏览数据库

分析浏览数据库来发现智能超级链接，是一个数据挖掘问题。常用的数据挖掘方法很多，这里我们采用信息网模型，因为它具有以下优点：

1. 智能超级链接是主题之间的一种潜在关系，具有统计性。而信度网模型是一种基于概率理论的模型，该模型的建立就是对实例数据的统计过程。

2. 信度网具有灵活多样的推理能力，为在线智能导航提供了方便。

利用信度网模型进行在线智能导航，包括两个步骤：一是通过学习浏览数据库，建立信度网模型；二是利用该信度网模型，在线为用户提供导航。

二、信度网模型的建立

建立信度网模型，就要通过对浏览数据库的学习，来建立以所有的主题为结点的信度网。它包括建立信度网的结构，以及学习对应结构的条件概率表两个过程。在信度网研究中，该问题称为信度网学习问题。分析浏览数据库可以看到，该数据库中的每一个纪录都是完整的，即每个字段都有取值，所以它是一个完整的学习数据库。基于完整学习数据库的信度网学习问题比较简单，已有一些成熟的学习算法。

我们首先学习建立信度网结构。根据不同的学习指导思想，信度网的结构学习算法可以分为两类：一类是基于测度的模型选择法；另一类是基于独立性测试的学习法。

基于测度的模型选择法认为，从一批实例数据中学习信度网，可供选择的信度网结构数目巨大（由全体变量构成的有向无环图都可以作为信度网结构），学习的目的就是要找出最符合实例数据的信度网结构，所以这类方法首先根据学习的具体要求，选用一种测度，用来衡量一个结构对实例数据的适合程度，再根据该测度从大量的模型中找出测度最优的结构作为学习所得的信度网结构。由于结构空间太大，计算每个结构的测度几乎不可能，所以一般都采用启发式搜索算法，以该测度为指导，逐步构造出该测度最优的结构。这类方法常用的测度有：贝叶斯测度[1]、最小描述长度测度[2]、贝叶斯信息测度等。已经证明，在实例数据库很大时，贝叶斯测度、最小描述长度测度和贝叶斯信息测度是相互等价的。

基于独立性测试的方法主要着眼于信度网的结构语义，即信度网的结构表达了变量之间的条件独立性关系。通过对实例数据的分析，用不同的条件独立性测试方法（如互信息测试等）可以获得变量之间的条件依赖关系，从而可以根据这种依赖关系来构造信度网结构。这类方法有：CL 算法、三阶段学习算法[3]等。CL 算法只能求得树形结构的信度网，三阶段学习算法对其进行了改进，可以求得任意结构的信度网。

这两类方法在进行信度网学习时各有优缺点。从获得联合概率分布的角度来看，基于测度的模型选择方法明显优于基于独立性测试的方法，因为前者就是以此作为优化测度的，而后者主要考虑的是信度网中表现出的条件独立性。然而，第一类方法的计算比较复杂，即使在实际计算中采用了启发式搜索算法，其计算的时间复杂性仍然很高。

尽管这些方法也可以学习得到信度网的结构，但其计算都比较复杂，尤其当网站所包含的主题较多时，学习信度网的结构其计算量相当大。为了简化，我们可以利用网站的信息结构图来构造信度网的结构。但信度网的结构是一个不包含有向环的图形结构，所以不能直接使用网站信息结构图，而必须对其进行修改。算法1通过分析网站中的超级链接关系，可以建立一个关于这些主题的树形信度网结构。如图3是对图1所示的网站建立的树形信度网结构。

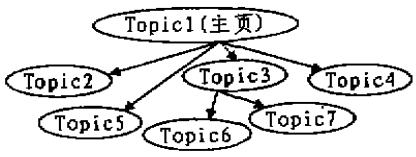

图3 树形信度网结构

算法1 建立树形信度网结构算法。

输入：网站信息及主题集合 E，E 中的每一个元素对应一个不用于导航的主题。

输出：树形信度网结构。

过程：

1 初始化。建立一个空集合 $T1$，及空队列 $T2$。

2 将主页作为一个主题放入队列 T2中，为主页对应建立一个结点。

3 反复循环，直到队列 T2为空：

(1)按顺序从 T2中取出一个主题 Topic。

(2)打开 Topic 对应的页面，对其中的每一个超级链接引进的主题 $Topic1$。

if $Topic1 \notin T1$ 且 $Topic1 \notin E$ *then*

为 $Topic1$ 对应建立一个结点，并用一条有向边与 $Topic$ 相连。

将 $Topic1$ 加入到集合 $T1$ 中，即：$T1 = T1 \cup \{Topic1\}$；

将 $Topic1$ 加入到队列 $T2$ 后面；

endif

4 返回建立的图形结构。

从分析可以看出，算法1将每一个超级链结点作为一个主题，这种简单处理具有某种盲目性，因为网站中的某些超级链结并不一定对应一个可用于导航的主题，或者管理者并不希望将这些主题参与导航，如网站上的临时通知，以及某些指向其他站点的友好链结等。为了避免将这些主题作为信度网的结点，算法1设置了一个集合 E，用于过滤掉这些超级链接。

采用树形信度网结构具有如下的优点：

1 信度网结构的建立比较简单，通过搜索网站的超级链接结构，就可以建立信度网结构，不需要复杂的计算。

2 在利用信度网模型进行在线导航时，需要快速进行信度网推理，从而为浏览者提供同步导引(参见第三节)。在一般的信度网结构上的推理问题是一个 NP 问题，但对于树形结构的信度网，许多推理算法都可以在多项式时间内完成。所以，采用树形结构这一特殊结构，可以简化信度网的推理，从而能为浏览者及时提供在线导航。

建立了信度网结构，下一步就要为该结构中的所有结点指定条件概率表。从前面的分析我们知道，用于信度网学习的浏览数据库是一个完整的实例数据库，所以计算结构中各个结点的条件概率表比较简单，通过对浏览数据库的统计就可以完成。如对于图3中的结点 $Topic6$，它的条件概率 $P(Topic6|Topic3)$ 的各项可以这样计算：

$$P(Topic6=1|Topic3=1)=\frac{P(Topic6=1 \wedge Topic3=1)}{P(Topic3=1)} \quad (1)$$

浏览数据库的纪录总数用 n 表示，$Topic6=1$ 且 $Topic3=1$ 的纪录数目表示为 $m_{Topic6=1 \wedge Topic3=1}$，则可将式(1)转换为

$$P(Topic6=1|Topic3=1)=\frac{m_{Topic6=1 \wedge Topic3=1}/n}{m_{Topic3=1}/n}=\frac{m_{Topic6=1 \wedge Topic3=1}}{m_{Topic3=1}} \quad (2)$$

从(2)式可以看出，通过统计浏览数据库，就可以计算出条件概率表的这一项。同理，可以求出条件概率表中的其它各项。

三、在线智能导航

通过对浏览数据库的学习，建立的关于主题的信度网模型，包含了大量浏览者的网站上的浏览中所体现出的对各个主题的相关性。利用这些信息就可以进行在线导航。基于信度网模型的在线导航是通过对信度网的推理来完成的，其过程大致如下：

首先，当一个浏览者进入网站时，他处于网站的主页上。由于此时浏览者尚未进行任何浏览，导航系统没有浏览者的任何信息，因此只能依靠大量浏览者的浏览特点来为其提供导航。这种特点可以通过在信度网上进行最可能解释(Most Probable Explanation)的计算来获得。最可能解释又称为最大可能配置(Maximal Probable Configuration)，是指在没有证据信息时，确定信度网中各个变量的取值，使整个联合概率分布的值最大。对于具体的导航任务来说，获得信度网的最可能解释后，那些取值为1的主题，就可以解释为浏览者在该网站上最常浏览的主题。因此，可以将这些主题推荐给浏览者，让其从中选择自己喜欢的主题。需要注意的是这些主题与主页之间可能存在传统的超级链接，也可能不存在链接，后者就是一个智能超级链接。

其次，当用户选择了一个超级链接(可能是一个智能超级链接)进入对应的主题浏览时，导航系统就获得了关于用户的信息(他选择了某个主题)，利用该信息就可以猜测浏览者最可能浏览的下一个或多个主题。该过程可以通过对各个主题的信度计算和比较来获得。在信度网中，一个结点(变量)的信度定义为给定证据时该变量的后验概率。在导航系统中，将浏览者正在浏览的主题作为证据来计算其它各个主题的信度，每一个主题的信度就可以理解为浏览者下一步将浏览该主题的概率。如在图3中，假设浏览者通过主页，进入了 Topic3浏览，则主题 Topic5的信度为 P(Topic5 = 1 | Topic3 = 1)，它表示浏览者下一步浏览主题 Topic5的可能性。因此，可以将 Topic3 = 1作为证据输入信度

网，计算出其余结点的信度，然后再通过比较，从中选择信度最大的几个主题，推荐给浏览者，这些主题将是浏览者下一步最可能浏览的主题，从而为浏览者提供导航。

随着浏览者的进一步浏览，他浏览过的主题越多，导航系统获得的浏览者的信息也就越多。在每一次导航时，可以将这些已浏览过的主题作为证据加入，从而提高导航的准确性。如在上面的例子中，当浏览者浏览主题 Topic3时，此时的证据集合为{Topic3＝1}，接着浏览者进入了主题 Topic5进行浏览，那么此时的证据集合为{Topic3＝1，Topic5＝1}。随着证据的增加，其推理的正确性也将提高。

从以上的导航过程可以看出，基于信度网的导航中，主要依靠信度网的推理计算。这里对其作简单介绍。信度网的推理算法可以分为两类：一类称为精确推理，即精确地计算假设变量的后验概率。另一类称为近似推理，即在不影响推理正确性的前提下，通过适当降低推理精度来达到提高计算效率的目的。精确推理一般用于结构较简单的信度网，主要算法有：直接计算法、消息扩散与汇聚算法、关联树算法[4]、消元法等。对于结点数量大、结构复杂的信度网常常采用近似推理，近似推理算法可以分为两类：一类是随机仿真法，即通过统计一个事件在一系列仿真实验中所发生的次数来计算事件的概率。另一类是基于搜索的近似计算法。这类算法首先由信度网构造一个对应的搜索空间树，每个节点对应一些变量的取值组合及其概率，根据概率的扩展公式，每个中间节点的概率等于其所有后代叶节点的概率之和。于是，信度网的推理计算就变成了在该树上的搜索过程。

在线导航需要为浏览者及时地提供导航信息，因此，它对于信度网的推理速度要求较高。这里我们结合在线导航的具体特点，对实际中较常用的联合树算法做一些优化。联合树算法可以参见文[4]。在分析在线导航过程中，可以看到，它用于推理的证据集合逐步增加，即在下一次导航时，所使用的证据集合包含了上一次导航时的证据。我们将其称为证据的单调递增性。这个特性可以用来简化联合树算法的计算。

联合树算法的推理过程包括以下几个步骤：

1　将信度网转化为一个联合树；2　初始化该联合树；3　加入证据；4　执行消息传递过程，使联合树达到全局一致；5　利用边际化操作，计算结点的信度。

当获得证据 e 时，对 e 中的每一个结点 V＝v1，寻找它的一个父团结点，并将该团结点的分布表中的那些 V≠v1的项置为0。然后再进行上述第3和第4步计算。

完成上述计算后，假设要计算证据 e′，此时一般要重新对联合树进行初始化，这是因为在证据 e′中可能包含一个与 e 相同的结点 V，但此时 V＝v2。按照证据的加入方法，首先找到 V 的父团结点，并将其分布表中 V≠v2的结点置为0，但其余的项不变，而在上一次推理时，加入证据 e 后，使其分布表中所有的 V≠v1的项变为了0。要恢复这些项的内容，就必须重新对联合树进行初始化。所以每次对不同的证据进行计算时，都要执行上述的2、3、4、5步。但由于在进行导航时，证据的增加具有单调递增性，所以不会出现上面的情况。为了提高计算效率，每次推理时，可以直接从第3步开始，而免去第2步的初始化过程，从而提高计算速度。

四、在线智能导航器的实现与总结

根据在线智能导航系统的功能特点，它与网站及浏览者之间的数据流向关系如图4所示。当浏览者点击某个主题(topic)进行浏览时，该信息分别流向两个方向：一个方向是进入智能导航系统的在线导航模块，用于产生导航信息；另一个方向是进入浏览数据库，用于更新该数据库中的历史数据。在线导航模块接到浏览者的浏览信息后，进行推理计算，产生实时导航信息，传送给浏览者。当浏览数据库发生变化时，数据预处理模块首先对这些数据进行预处理，处理后的数据进入导航器建立(更新)模块，由该模块对用于导航的信度网模型进行更新。

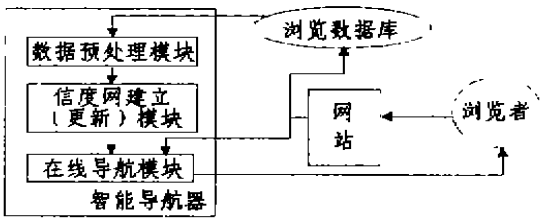

图4　数据流程图

基于信度网模型的在线智能导航系统，能够通过分析浏览者在网站上的浏览纪录，并结合浏览者的现有浏览状态，来及时准确地为浏览者提供导航。根据信度网模型的特点，该系统还可以用于网站的结构优化及其它用途。如通过最大可能配置计算获得的结果，表示了浏览者对网站的各个主题的兴趣，利用这些信息来重新调整网站的结构设计，就可以设计出符合浏览者爱好的主题结构。另外，对于商业网站，通过对这些热点主题的发现，可以找到最佳的广告插入点，从而提高广告的访问量。

随着新的浏览数据的加入，必须及时更新信度网

(下转第126页)

序，本文中将不再赘述。

下面将阐述在树的全局遍历过程中处理结点运算时涉及到的回溯操作，在具体的处理结点的计算流程中，按分治策略把对'叶结点'和'根结点'的处理分开来，以实现处理时的高效率。在处理叶结点时，关键是找到当前报告日，并计算此报告日的报告进度率。具体报告日定位和报告日进度率的计算策略由于涉及很多软件的琐碎细节，在此不赘述。当树的遍历回到该叶结点的上层父结点时，需要回溯以便取得此时计算的结果。在早期的 Sisfurd 版本中，所有计算所得的子项目进度报告日报告率保存在全局数据树中，回溯时通过对当前内结点的子树再次实行遍历而实现。对一棵有 N 个结点的树，我们假设每个内结点平均有 k 个子女，则这棵树的层数 $n=\log_k(k-1)N+1$，如果在每个内结点上都需要对其所有直接子女进行回溯，设在每个内结点上的回溯运算量为 O(k*s)(s 是回溯访问一个子结点时的运算量)，那么对于高度为 n 的树回溯运算量就为：

$$T(n)=O(\sum_{i=0}^{n-1}k^i(k,s))=O(k^{n+1},s)$$

当 s 的值比较大时，由上述公式可知 T 值将以几何级数增长。当然 T(n)的值还与 k、n 有关，但要实现树的遍历运算，k、n 的值是没办法作优化的。要使 T(n)的值控制在能容忍的范围内，关键在于设法控制 s 的值，使之不会增长过快。

为了控制 s 的值，必须设法降低对树的回溯时的运算量。在早期版的 Sisfurd 中，回溯是通过遍历所有直接子女的方法实现的。树的回溯操作是基于链表的，对于图3所示的数据树，在后根遍历到某个父结点时，可通过指向其子女链表的指针取得子女信息，从而对子女的内容进行回溯访问。这个操作包括两个过程，首先取得指向子女链表的指针，然后再根据子女序号取得子女结点指针。因为在实现子女链表时用的是系统定义的数组式链表[1]，而这个操作相对而言是比较费时的。考虑到对子女结点的回溯访问是按子女的序号随机访问的，并且在对某个结点处理时所有回溯仅限于其直接子女，在回溯完成后，被回溯访问过的结点不会再在其他回溯过程中被访问，所以我们在改进程序中把原来的保存在子树中的进度报告日报告率临时保存在一个动态数组中，每次回溯不用再到数据树中去查找，直接在动态数组中通过下标映射实现随机存取。

3.3 算法分析

下面分析一下 s 的值是如何有效地被控制的。先定义两个数据结构，第一个是长度为 n 的单链表 L，另一个是有 n 个元素的一维数组 A，在程序中，对 L 和 A 的访问完全是随机的，假设处理 L 和 A 中的某个元素的时间是相同的，设为 t，则由文[1]可知，遍历 L 所需时间 T(L)=t*n*(n+1)/2，遍历 A 所需时间 T(A)=n*t。

在 Sisfurd 早期版本中，由于要实现树的动态生长，很直观地就采用单链表作为计算数据的中间存储池。但在实际使用中，树的每一层上结点数目一般总是在500以内，对于超出500的部分，我们仍用链表作为存储池，但这种情况在实际应用中几乎不出现。对于500个结点随机遍历，单链表花费时间 T(L)=125250t；而一维数组只需时间 T(A)=500t，因而核心计算部分的效率的提高是很显著的。

结束语 本文所论述的具体问题的算法抽象及带有回溯的树的运算的算法优化在新版本的 Sisfurd 中得到运用，在日本富士通公司与南大富士通软件技术有限公司的评测中，证明其具有较高的效率，获得了用户满意的效果。

参考文献

1 Microsoft MSDN
2 陈本林，陈佩佩. 数据结构. 南京大学出版社，1999
3 余祥宣，崔国华，邹海明. 计算机算法基础. 华中理工大学出版社，2000
4 Kruglinski D J 著，潘爱民，王国印 译. Visual C++技术内幕(第四版). 清华大学出版社，1998
5 Sisfurd 项目文档. 富士通公司，1996

(上接第104页)

模型，包括对信度网结构的修改及对条件概率表的调整。最简单的方法是利用浏览数据库，脱机对信度网进行更新，称为批量更新。当然，也可以进行在线更新，但由于信度网的更新所需的计算量比较大，因此可能会对在线导航造成影响。

本文中的在线导航系统存在的一个问题是，它没有利用浏览者在网站上的浏览顺序这一信息。可以设想，通过分析用户浏览各个主题的先后顺序，可能会产生更准确的导航信息。对此，还可做进一步的研究。

参考文献

1 Heckerman D. Learning Bayesian Network: The Conbination of Knowledge and Statistical Data. [Technical Report MSR-94-09]
2 Rissanen J. Stochastic Complexity in Statistical Inquiry. World Scientific, River Edge, NJ, 1989
3 Chen J, Bell D A, Liu W. Learning Belief network from data. An information thoery based approach. In Proc of ACM CIKM'97
4 Jensen F V, Lauritzen S L, Olesen K G. Bayesian updating in causal probabilistic networks by local computations, Comp Stat Quart, 1994, 4: 269～282