

多层的贝叶斯网络检索模型

白彦霞, 云彩霞, 李珊, 张秋菊, 杨鹏

BAI Yan-xia, YUN Cai-xia, LI Shan, ZHANG Qiu-ju, YANG Peng

北京化工大学 北方学院, 河北 廊坊 065201

North College of Beijing University of Chemical Technology, Langfang, Hebei 065201, China

E-mail: 163byx@163.com

BAI Yan-xia, YUN Cai-xia, LI Shan, et al. Multi-layered Bayesian network model for information retrieval. Computer Engineering and Applications, 2009, 45(16): 165-168.

Abstract: To quantify the degree of similarity between synonyms by term similarity accurately. Combining this quantification, the simple Bayesian network for information retrieval is improved, a four-layered Bayesian network retrieval model based on the quantified term relationships is proposed. The topology of the new model, probability estimation of all nodes and the whole inference process are described. Experimental results show that the new model in which term similarity is used to encode term relationships behaves better than others, realizing semantic retrieval to some extent, which is the inevitable tendency of information retrieval.

Key words: Bayesian networks; term similarity; information retrieval; synonyms

摘要: 利用术语相似度将同义词间的相似程度数量化, 以此量化关系对用于信息检索的简单贝叶斯网络进行若干改进, 构造一个四层贝叶斯网络检索模型。给出新模型的拓扑结构、各层节点详尽的概率估计以及文档检索与推理过程。最后, 对新模型进行评估, 结果表明该模型可以有效地提高检索性能, 在一定程度上实现基于语义的信息检索, 这正是目前信息检索发展的必然趋势。

关键词: 贝叶斯网络; 术语相似度; 信息检索; 同义词

DOI: 10.3778/j.issn.1002-8331.2009.16.048 **文章编号:** 1002-8331(2009)16-0165-04 **文献标识码:** A **中图分类号:** TP18

1 引言

在信息检索领域, 语义相同或相近的术语, 如同义词、近义词、高度相关词等, 从语义上表达了用户的查询意图, 所以包含这些术语的文档在一定程度上也满足用户的要求。贝叶斯网络作为人工智能领域处理概率问题的主要方法, 在过去十几年里已经通过不同的方式应用到了信息检索领域^[1], 其灵活的拓扑结构, 能表示术语间的条件概率和概念语义, 从而为更准确地进行语义检索提供了保障。

文献[2]所描述的简单贝叶斯网络检索模型只能检索到那些与查询术语本身相关的文档, 而检索不到那些与查询术语语义相同或相似的文档。当使用给定文档集合所包含的术语间的关系时, 信息检索系统的性能一般可以提高^[3]。文献[3]利用术语聚类技术对简单贝叶斯网络检索模型进行扩展, 扩大了相关信息的检索, 但是术语之间的强度关系主要依赖于被测试的集合, 只能在两个术语共现的情况下计算其强度, 不能很好地处理数据稀疏问题。文献[4]利用同义词扩展查询, 在一定程度上提高了检索性能, 但是没有对同义词间的相似程度数量化, 无法区别查询术语与其同义词间的相似程度, 不能准确而有选择地确定用于扩展查询的同义词, 也没有挖掘文档之间的相似关系来进一步提高语义检索的性能。

本文针对上述模型的缺点, 在文献[4]所述扩展模型的基础上进行若干改进, 不仅将基于术语相似度的同义词间的量化关系应用于术语之间, 同时也应用于文档之间进行层层递进的推理与检索, 从而更好地提高语义检索系统的性能。

2 两层贝叶斯网络检索模型(BNR-2)

2.1 拓扑结构

贝叶斯网络 $G=(V, E)$ 是一个有向无环图(DAG), V 中的节点表示所要解决的随机变量, E 中的弧表示这些变量之间的依赖关系。两层贝叶斯网络的变量集合 V_2 由两个不同的变量集组成, 即 $V_2=T \cup D$ 。其中, $T=\{T_1, T_2, \dots, T_M\}$ 是 M 个索引术语组成的集合, $D=\{D_1, D_2, \dots, D_N\}$ 是 N 篇文档组成的集合。文中的符号 $T_i(D_j)$ ($i=1, 2, \dots, M; j=1, 2, \dots, N$) 既表示术语(文档), 也表示与其相关的变量和节点。术语变量 T_i 和文档变量 D_j 都是二进制的随机变量, 取值集合分别为 $\{i, t_i\}$ 、 $\{d_j, d_j\}$ 。 i 和 t_i 分别表示“术语 T_i 不相关”、“术语 T_i 相关”, d_j 和 d_j 分别表示“文档 D_j 与给定的查询不相关”、“文档 D_j 与给定的查询相关”。两层贝叶斯网络的拓扑结构如图 1 所示, 其中的弧是由术语节点指向包含这些术语的文档节点, 术语节点之间或文档节点之间不存在弧。

作者简介: 白彦霞(1979-), 女, 助教, 主要研究方向为信息检索; 云彩霞(1981-), 女, 助教, 主要研究方向为无线通信与计算机应用; 李珊(1980-), 女, 助教, 主要研究方向为数据挖掘及系统安全; 张秋菊(1982-), 女, 助教, 主要研究方向为信号与信息处理; 杨鹏(1981-), 男, 助教, 主要研究方向为计算机应用。

收稿日期: 2009-01-21 **修回日期:** 2009-03-30

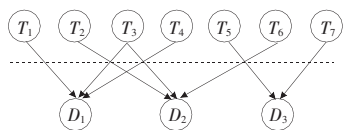


图1 两层贝叶斯网络检索模型

2.2 概率分布与推理

根据拓扑结构中弧的指向,可知术语相互边缘独立,文档在给定其所包含术语的情况下相互条件独立。由此可以估计节点的概率分布:

(1)术语节点:所有术语都是根节点,每个术语 T_i 的先验概率为 $P(t_i)=1/M, P(\bar{t}_i)=1-P(t_i)$,其中 M 为集合中术语的数目。该定义为所有的根术语节点赋予了相等的先验概率。

(2)文档节点:设 $Pa(D_j)$ 表示文档 D_j 的所有索引术语集合,即文档 D_j 的父节点; $pa(D_j)$ 表示文档 D_j 的所有索引术语取值(相关或不相关)后的一种组合。由此定义 D_j 的条件概率如下:

$$p(d_j|pa(D_j)) = \sum_{T_i \in D_j, t_i \in pa(D_j)} w_{ij} \quad (1)$$

其中,公式(1)中 w_{ij} 为文档 $D_j \in D$ 的索引术语 $T_i \in D_j$ 的权重, $0 \leq w_{ij}, \forall i, j$, 且 $\sum_{T_i \in D_j} w_{ij} \leq 1, \forall D_j$ 。公式(1)中 $t_i \in pa(D_j)$ 表示仅将组合 $pa(D_j)$ 中取值为 t_i 的术语 T_i 的权重相加。因此 $pa(D_j)$ 中相关术语越多, D_j 的相关概率值就越大。权重 w_{ij} 有多种定义方法,本文定义如下:

$$w_{ij} = \alpha^{-1} \frac{tf_{ij} \times idf_i^2}{\sqrt{\sum_{T_k \in D_j} tf_{kj} \times idf_k^2}}$$

其中, α 为规格化常数,用来保证 $\sum_{T_i \in D_j} w_{ij} \leq 1, \forall D_j \in D$ 。 tf_{ij} 为术语频度,即术语 T_i 在文档 D_j 中出现的次数。 idf_i 为逆文档频度,定义为 $idf_i = \lg(N/n_i) + 1$,其中 N 为测试集中的文档数量, n_i 为包含术语 T_i 的文档数量。

(3)推理:根据两层节点的概率分布情况,计算出每篇文档 D_j 与查询 Q 的相关概率 $p(d_j|Q)$:

$$p(d_j|Q) = \sum_{T_i \in D_j} w_{ij} p(t_i|Q) \quad (2)$$

因为术语节点边缘独立,由条件独立性可得:如果 $T_i \in Q$, 则 $p(t_i|Q)=1$; 如果 $T_i \notin Q$, 则 $p(T_i|Q)=1/M$ 。将其带入公式(2)得:

$$p(d_j|Q) = \sum_{T_i \in D_j \cap Q} w_{ij} + \frac{1}{M} \sum_{T_i \in D_j \setminus Q} w_{ij} = \frac{1}{M} \sum_{T_i \in D_j} w_{ij} + \frac{M-1}{M} \sum_{T_i \in D_j \cap Q} w_{ij} \quad (3)$$

注意:文档 D_j 相关的先验概率为 $p(d_j)=1/M \sum_{T_i \in D_j} w_{ij}$, 因此 $p(d_j|Q)$ 总是大于等于 $p(d_j)$ 。

3 四层贝叶斯网络检索模型

3.1 同义词与相似度

在信息表示和信息检索领域,同义词主要指能够相互替换、表达相同或相近概念的词汇,如电脑-计算机、边疆-边境等^[5]。本文利用哈尔滨工业大学信息检索实验室刘挺教授等编制的《同义词词林(扩展版)》来获取查询术语的同义词。但是这样得到的同义词可能有多个,而这些同义词之间的相似程度又存在着区别,究竟使用哪些同义词来扩展查询直接影响着检索

结果的质量。术语相似度是术语间语义相似紧密程度的一种定量度量,也就是术语间语义关系的数量化^[6]。它可以从一定程度上来进一步辨别这些同义词之间的近似程度,从而可以更准确地确定用于扩展查询的同义词。目前国内外对术语相似度的计算方法大体可分为两类:(1)基于语义词典的术语相似度计算方法;(2)基于语料库的术语相似度计算方法^[6]。本文采用第一类方法,即基于《知网》^[7]的术语相似度计算方法,确定同义词之间语义相似的程度。

同义词可以扩展查询,术语相似度又进一步将同义词间的相似程度数量化,所以将《同义词词林(扩展版)》与《知网》相结合既可以扩大相关信息的检索,又可以准确地检索到相关文档。

3.2 术语子网的扩展

复制原始术语层中的每一个术语节点 T_i 得到另外一个术语节点 T'_i , 形成一个新的术语层 T' 。连接两个术语层的弧的指向为:(1)每一个术语 T'_i 与其本身 T_i 之间的相似度 $Sim(T_i, T'_i)$ 为 1.0, 故总存在由 T'_i 指向 T_i 的弧, 即 $T'_i \rightarrow T_i$; (2)若术语 T_i 与 T_j 互为同义词且其相似度 $Sim(T_i, T_j)$ 大于某一个阈值, 则存在由 T'_i 指向 T_j 的弧和由 T'_j 指向 T_i 的弧, 即 $T'_i \rightarrow T_j, T'_j \rightarrow T_i$; (3)对于只有一个同义词的术语而言,为了扩展查询,则不管其相似度多大,总存在这样的弧。因此,术语节点 $T_i \in T$ 的父节点集合 $Pa(T_i)$ 由术语节点 T'_i 及 T_i 的同义词节点 T'_j 组成。

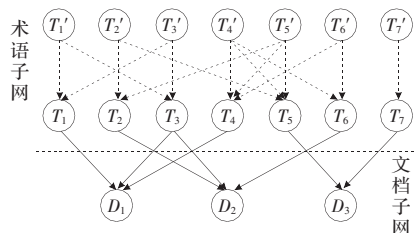


图2 术语子网扩展后的模型

术语子网扩展后,变量集合 $V_3 = T' \cup T \cup D$, 新术语层 T' 中的术语变量 T'_i 也是二进制随机变量,取值集合为 $\{\bar{t}'_i, t'_i\}$, \bar{t}'_i 和 t'_i 分别表示“术语 T'_i 不相关”、“术语 T'_i 相关”。

3.3 文档子网的扩展

在该子网中复制原始文档层中的每一个文档节点 D_j 得到另外一个文档节点 D'_j , 形成一个新的文档层 D' 。所以,在四层贝叶斯网络模型(BNR-4)中,变量集合 $V_4 = T' \cup T \cup D \cup D'$ 。复制的任意文档节点 D'_j 的父节点集合 $Pa(D'_j)$ 定义为与文档 D_j 最相似的文档集合,则连接两个文档层的弧由 $D_i \in Pa(D'_j)$ 指向 D'_j , 如图3所示。

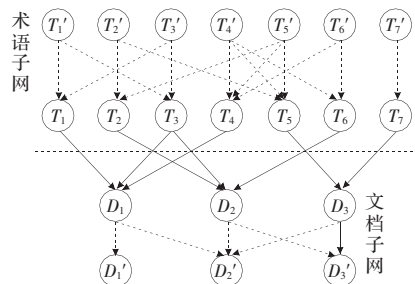


图3 四层贝叶斯网络检索模型

在模型 BNR-4 中加入模拟文档间直接关系的弧,类似于信息检索模型中的聚类技术。利用文档间的术语相似关系度量文档间的相似度,即假定另外一篇文档相关的条件下,通过估

计每一篇文档与其相关的条件概率来实现文档相似度的度量。为了计算文档相似度, 令 $e(D_i)$ 表示文档 D_j 相关的某类证据。对于一篇给定的文档 D_j , 如果计算概率 $p(d_j|e(D_i))$, $\forall D_i \in D$, 那么产生 $p(d_j|e(D_i))$ 最大值的文档 D_i 就是与 D_j 最相关的文档。对于每一篇文档 D_j , D 中所有的文档以概率 $p(d_j|e(D_i))$ 递减的顺序排列, 根据此概率即可确定与文档 D_j 最相似的文档集合, 即文档节点 D_j' 的父节点集合 $Pa(D_j')$ 。关于 $p(d_j|e(D_i))$ 的具体计算参考 4.2 节。

4 概率分布

4.1 术语节点

BNR-4 术语子网中, 对于任意根术语节点 T_i' , 其相关的边缘概率 $P(t_i')$ 和不相关的概率 $P(\bar{t}_i')$ 的定义与 BNR-2 模型相同。对于任意非根术语节点 T_i , 令 $pa(T_i)$ 为 $Pa(T_i)$ 中每个术语变量取值 (相关或不相关) 后的一个组合, 利用一般正则模型的概率函数^[2]可得:

$$p(t_i|pa(T_i)) = \sum_{\substack{T_j' \in Pa(T_i) \\ t_j' \in pa(T_i)}} v_{ij} \quad (4)$$

其中, v_{ij} 为衡量每个术语 $T_j' \in Pa(T_i)$ 对术语 T_i 影响程度的权重, $t_j' \in pa(T_i)$ 意味着只将 $pa(T_i)$ 中相关术语的权重相加。若 T_i 只有一个父节点 T_i' , 则定义 $v_{ij}=1$; 若术语 T_i 有多个父节点, 则权重 v_{ij} 定义如下:

$$v_{ij} = \begin{cases} \beta, & (0.5 \leq \beta \leq 1.0, \text{ 且 } i=j) \\ \frac{1-\beta}{\sum_{T_j' \in Pa(T_i), i \neq j} Sim(T_i, T_j')} Sim(T_i, T_j'), & (i \neq j) \end{cases} \quad (5)$$

该定义既保证了术语本身对 T_i 的最大强度关系, 又可以辨别相异的同义词对 T_i 的强度关系, 即挖掘了术语间精确的强度关系。

4.2 文档节点

文档层 D 中的文档节点 D_j 的条件概率 $P(d_j|pa(D_j))$ 的计算与公式(1)相同。在此重点介绍文档层 D' 中的文档节点 D_j' 的条件概率 $p(d_j'|pa(D_j'))$ 的计算, 与公式(4)类似, 利用一般正则模型的概率函数^[2]对 $p(d_j'|pa(D_j'))$ 定义如下:

$$p(d_j'|pa(D_j')) = \sum_{\substack{D_i \in Pa(D_j') \\ d_i \in pa(D_j')}} u_{ij} \quad (6)$$

其中, 权重 u_{ij} 衡量文档 D_i 与文档 D_j 的相似程度, 定义为 $u_{ij} = p(d_j|e(D_i))/S_j$, $S_j = \sum_{D_k \in Pa(D_j')} p(d_j|e(D_k))$ 。为了说明 $p(d_j|e(D_i))$ 计算方法, 首先给出事件 $e(D_i)$ 的含义, 然后说明如何有效地计算 $p(d_j|e(D_i))$ 的值。

假设 $e(D_i)$ 表示事件 $[T_k=t_k, \forall T_k \in D_i]$, 即假设文档 D_i 的索引术语都相关; 那么这时计算 $p(d_j|e(D_i))$ 就等价于引入了一个仅包含索引文档 D_i 的所有术语的查询, 也就是说, $e(D_i)$ 相当于查询 Q_i , 且包含了 D_i 所有相关术语。所以可得 $p(d_j|e(D_i))$ 的计算公式:

$$p(d_j|e(D_i)) = p(d_j|Q_i) = \sum_{T_k \in D_i} w_{ij} p(t_k|Q_i) \quad (7)$$

因为模型 BNR-4 中具有两层术语节点, 此时的术语 T_i 已不再是根节点, 所以关于 $p(t_k|Q_i)$ 的值由公式(9)计算, 这样术语之间的相似关系便应用于了文档之间。

5 推理和检索机制

基于模型 BNR-4 的检索过程, 也就是计算出每篇文档 D_j' 与查询 Q 的相关度 $p(d_j'|Q)$; 而对于 BNR-2, 最终计算出每篇文档 D_j 与查询 Q 的相关度 $p(d_j|Q)$ 。当查询 Q 提交给系统时, 便开始了检索过程: 首先, 假定查询 Q 的每个术语 T_{i_0}' 的状态为 t_{i_0}' (相关); 然后, 据此在整个网络中层层递进推理。最后, 文档以相关度递减的顺序呈现给用户。

模型 BNR-4 复制了一层术语节点和一层文档节点, 这样贝叶斯网络中就有大量节点且许多节点又有多个父节点。因此, 即使对于较小的文档集合, 一般推理算法的效率都不够理想。为了解决该问题, 本文综合利用网络拓扑结构、术语节点和文档节点的概率函数, 其相应的推理过程为:

(1) 术语层 T 中的任意术语 T_i 的后验概率 $p(t_i|Q)$, 可按下式估计:

$$p(t_i|Q) = \sum_{T_j' \in Pa(T_i)} v_{ij} p(t_j'|Q) \quad (8)$$

因为术语层 T' 中的术语相互边缘独立, 所以 $T_j' \in Q$ 时, 式(8)中的 $p(t_j'|Q)=1.0$, 否则 $p(t_j'|Q)=1/M$ 。注意, 对于只有一个父节点 T_i' 的术语 T_i 而言 $v_{ij}=1$, 若 $T_i' \in Q$, 由式(8)可得 $p(t_i|Q)=p(t_i'|Q)=1.0$, 否则 $p(t_i|Q)=p(t_i'|Q)=1/M$ 。若术语 T_i 有多个父节点, v_{ij} 用公式(5)代替, 则 $p(t_i|Q)$ 为:

$$p(t_i|Q) = \frac{1-\beta}{\sum_{T_j' \in Pa(T_i), i \neq j} Sim(T_i, T_j')} \sum_{T_j' \in Pa(T_i), i \neq j} Sim(T_i, T_j') p(t_j'|Q) + \beta p(t_i'|Q) \quad (9)$$

(2) 计算文档层 D 中的文档 D_j 的最终后验概率, 即文档 D_j 与查询 Q 的相关度 $p(d_j|Q)$:

$$p(d_j|Q) = \sum_{T_i \in Pa(D_j)} w_{ij} p(t_i|Q) \quad (10)$$

(3) 基于以上推理, 计算文档层 D' 中的文档 D_j' 的最终后验概率, 即文档 D_j' 与查询 Q 的相关度 $p(d_j'|Q)$:

$$p(d_j'|Q) = \sum_{D_i \in Pa(D_j')} u_{ij} p(d_i|Q) = \frac{1}{S_j} \sum_{D_i \in Pa(D_j')} p(d_i|e(D_i)) p(d_i|Q) \quad (11)$$

最后, 文档以概率递减的顺序呈现给用户, 这样就完成了整个信息检索过程。

6 模型性能评测与分析

实验所用文档来源于中国学术期刊网全文数据库。从该数据库共下载 741 篇文档作为文档测试集合, 经处理后这些文档被 1 113 个代表文档主要内容特征的术语索引, 针对这些文档共构造 10 个查询。对于不同模型, 检索性能的评测标准仍采用该领域常用的查全率 (Recall)/查准率 (Precision)。因为扩展模型中的参数 β 应该是通过大量实验得到的一个经验常数, 所以并不能确定 β 的最终取值。为了准确评测不同模型的检索性能优劣, 参数 β 取 6 个不同的值 (0.5, 0.6, 0.7, 0.8, 0.9, 1.0) 进行实验, 分别比较其在 10 个标准查全率级所对应的平均查准率值。实验数据如表 1 所示。

由表 1 的实验数据可以看出: BNR-4 的检索性能明显优于 BNR-2, 而且通过调节参数 β 的取值改变模型中术语间的强度关系可以获得更理想的检索效果。 $\beta=0.5$ 时, BNR-4 的检索效果最佳, 但是对于只有一个同义词的术语而言, 缺乏辨别同义词的能力, 即查询术语与其同义词等同对待了, 而实际上查询术语本身比其同义词更重要一些, 所以 β 一般应大于 0.5;

表1 基于贝叶斯网络检索模型的查全率-查准率对照表

Recall f(%)	Precision/(%)						
	BNR-2	BNR-4-1.0	BNR-4-0.9	BNR-4-0.8	BNR-4-0.7	BNR-4-0.6	BNR-4-0.5
10	66.20	76.96	80.75	86.15	90.36	96.33	98.12
20	58.78	69.28	74.64	82.28	87.15	93.48	96.51
30	57.22	67.22	73.86	80.36	82.91	86.57	89.50
40	57.13	68.57	72.59	77.67	79.83	83.34	85.42
50	47.54	64.39	68.95	75.58	74.66	80.87	81.38
60	43.80	62.06	68.53	73.77	74.12	76.84	77.65
70	39.84	60.27	66.59	71.54	71.91	72.20	74.17
80	36.18	56.84	63.42	64.73	65.69	66.97	68.98
90	29.63	53.41	61.83	61.28	62.73	63.10	63.76
100	11.90	26.49	38.76	43.26	45.99	38.97	37.85

对于 β 取其他值的情况,如 $\beta=0.6$ 和 $\beta=0.7$,检索效果比较理想; $\beta=1.0$ 时,BNR-4和BNR-2的术语子网等效,但是由于模型BNR-4扩展了文档子网,其检索性能依然明显优于模型BNR-2。由上述分析可知,无论参数 β 取何值,在查全率相等的条件下BNR-4更能有效提高查准率的值。

BNR-4的检索性能明显提高,主要原因在于:(1)利用同义词扩展查询可以检索到更多的相关文档;(2)通过术语相似度将术语间的相似程度数量化,对用于扩展查询的同义词做了一定的限制与选择,实验中规定相似度小于0.4的同义词都被滤掉,这样就可以避免检索到大量的无关文档。

7 结束语

文章详细阐述了利用术语相似度表示同义词间量化关系的拓扑结构,提出一个四层贝叶斯网络检索模型,并通过实验将不同模型的检索性能进行比较与分析。实验证明:模型BNR-4均具有良好的检索性能,因为术语相似度将同义词间的相似关系数量化,依此选择最优的同义词扩展查询,这样可以在不偏离用户检索目标的前提下扩大相关信息的检索。模型BNR-4不仅挖掘了术语间的相似关系,同时也挖掘了文档间的相似关系,故在查全率相同的情况下进一步提高了查准率,也避免了 β 等于1.0时BNR-4与BNR-2等价的不足。

(上接129页)

状拓扑上加入的融合点,仿真得到的网络拓扑和原网络拓扑是一样的。

6 结束语

本文通过分析现有多源网络层析成像方法可辨识性存在的原因,提出了一种基于层次特征提取的2-by-2网络拓扑信息提取方法,并用可辨识任意网络的融合算法来把所有2-by-2子网的拓扑信息进行整合,对整个网络的拓扑进行估计,发现提取出任意2-by-2网络的详细拓扑信息,并用好的融合算法来整合这些信息,能够有效地提高多源网络层析成像的可辨识性。从理论上论证了本文的方法具有更强的可辨识性。从仿真中可以看出,对于现有方法不可辨识的网络,本文的方法能够估计出唯一的正确解。

参考文献:

[1] Ratnasamy S, McCanne S. Inference of multicast routing trees and bottleneck bandwidths using end-to-end measurements[C]//IEEE

致谢 实验所用的同义词识别工具——《同义词词林(扩展版)》,由哈尔滨工业大学信息检索实验室刘挺教授提供,在此表示感谢!

参考文献:

[1] de Campos L M, Fernandez-Luna J M, Huete J F. Bayesian networks and information retrieval: An introduction to the special issue[J]. Information Processing and Management, 2004, 40(5): 727-733.
 [2] Acid S, de Campos L M, Fernandez-Luna J M, et al. An information retrieval model based on simple Bayesian networks[J]. International Journal of Intelligent Systems, 2003, 18(2): 251-265.
 [3] de Campos L M, Fernandez-Luna J M, Huete J F. Clustering terms in the Bayesian network retrieval model: A new approach with two term-layers[J]. Applied Soft Computing, 2004, 4(2): 149-158.
 [4] 徐建民, 白彦霞, 吴树芳. 基于同义词扩展的贝叶斯网络检索模型[J]. 计算机应用, 2006, 26(11): 2628-2630.
 [5] 陆勇, 侯汉青. 用于信息检索的同义词自动识别及其进展[J]. 南京农业大学学报: 社会科学版, 2004, 4(3): 87-93.
 [6] 张涛, 杨尔弘. 基于上下文词语同现向量的词语相似度计算[J]. 电脑开发与应用, 2005, 18(3): 41-43.
 [7] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 计算语言学及中文信息处理, 2002, 7(2): 59-76.

INFOCOM 1999, New York, NY, Mar 1999.

[2] Duffield N G, Horowitz J, Lo Presti F, et al. Multicast topology inference from measured end-to-end loss[J]. IEEE Trans Inf Theory, 2002, 48: 26-45.
 [3] Coates M, Castro R, Nowak R. Maximum likelihood network topology identification from edge-based unicast measurements[C]//ACM SIGMETRIC, Marina Del Rey, CA, Jun 2002.
 [4] Shih Meng-Fu, Hero A O. Hierarchical inference of unicast network topologies based on end-to-end measurements[J]. IEEE Transactions on Signal Processing, 2007, 55(5).
 [5] Rabbat M G, Coates M J, Nowak R D. Multiple source Internet tomography[J]. IEEE Journal on IEEE Selected Areas in Communications, 2006, 24(12).
 [6] Rabbat M, Nowak R, Coates M. Multiple source, multiple destination network tomography[C]//Proc IEEE INFOCOM, Hong Kong, March 2004.
 [7] Coates M, Rabbat M, Nowak R. Merging logical topologies using end-to-end measurements[C]//Proc ACM SIGCOMM Conf on Internet Measurement, Miami, FL, October 2003.