文章编号:1001 - 9081(2006)11 - 2628 - 03

基于同义词扩展的贝叶斯网络检索模型

徐建民[1,2],白彦霞[1],吴树芳[1]
(1. 河北大学 数学与计算机学院,河北 保定 071002; 2. 天津大学 系统工程研究所,天津 300072)
(yy. csi@ mail. hbu. edu. cn)

摘 要:利用同义词挖掘术语间的关系,对用于信息检索的简单贝叶斯网络进行若干改进,得到一个包含术语间直接关系的扩展模型。实验结果表明通过进一步调节扩展模型中的参数,可以获得良好的检索效果。

关键词:贝叶斯网络;同义词;信息检索;强度关系
中图分类号: G354.4 **文献标识码**:A

Extended Bayesian network retrieval model based on synonyms

XU Jian-min[1,2], BAI Yan-xia[1], WU Shu-fang[1]
(1. *College of Mathematics and Computer Science, Hebei University, Baoding Hebei* 071002, *China*;
2. *Institute of Systems Engineering, Tianjin University, Tianjin* 300072, *China*)

Abstract: To capture relationships between terms by means of synonyms and introduce several modifications to the simple Bayesian network for information retrieval, an extended retrieval model that included direct relationships between terms was proposed. Experiment results show that good retrieval effectiveness can be achieved by adjusting the parameter of the Bayesian networks used in our model.

Key words: Bayesian networks; synonyms; information retrieval; strength relationship

传统的信息检索技术一般将用户查询和文档进行精确匹配[1],无法满足语义概念上的匹配,因而检索不到与查询术语语义相似或相关的文档。当使用给定文档集合所包含的术语间的关系时,可以提高信息检索系统的性能[2]。然而,如何准确获取术语间的关系,并在检索过程中合理使用就成为提高性能的关键。贝叶斯网络作为人工智能领域处理概率问题的主要方法,在过去的 15 年里已经通过不同的方式应用到了信息检索领域[3]。其灵活的拓扑结构,能表示术语间的条件概率和概念语义,从而为更准确的检索信息提供了保证。本文利用同义词对简单贝叶斯网络检索模型进行扩展,解决了术语间语义概念的匹配问题,提高了检索性能。

1 相关知识

1.1 信息检索中的同义词

在信息表示和信息检索领域,同义词的概念并不等同于语言学和日常生活中的同义词,它不考虑感情色彩和语气,主要是指能够相互替换、表达相同或相近概念的词汇[4]。用于信息检索的同义词主要分为四类:1)等价词和等义词或词组,即意义完全相等的词,如电脑 - 计算机、自行车 - 脚踏车等。2)准同义词和准同义词词组,即意义基本相同的词和词组,如边疆 - 边境、住房 - 住宅等。这类词在同义词中占很大的比例。3)某些过于专指的下位词。例如在词表中只使用“球类运动”,而没有在下面列举出“门球”、“毽球”、“网球”等词,这些过于专指的下位词也被看作同义词。4)极少数的反义词。这类词描述相同的主题,但所包含的概念互不相容,如平滑度 - 粗糙度等。

信息检索的实践表明,由于自然语言中存在大量的同义词、近义词,用户检索时很难全部列举出表示同一概念的不同词汇,因而在检索时易造成漏检。利用同义词扩展查询,可以解决检索系统的此类漏检问题,提高检索性能。信息检索中识别同义词的义类词典和词汇分类体系资源包括 Roget's Thesaurus、WordNet 以及《同义词词林》、《知网》等[4]。本文利用哈尔滨工业大学信息检索实验室刘挺教授等对《同义词词林》扩展后的版本《同义词词林(扩展版)》来获取术语的同义词。

1.2 贝叶斯网络

贝叶斯网络可有效表示和处理 n 维的概率分布图[5]。贝叶斯网络由定量和定性两部分组成:1) 定性部分为有向无环图 DAG, $G = (V, E)$,其中 $V = \{X_1, X_2, \cdots, X_n\}$, $X_i (i = 1, 2, \cdots, n)$ 表示节点,即所要解决的随机变量;E 为 DAG 中的弧组成的集合, E 中的弧表示变量间的条件依赖关系。2) 定量部分是根据有向无环图得到的条件概率分布集合,每一个变量 $X_i \in V$ 都对应一张条件概率分布表 $P(X_i \mid pa(X_i))$,其中 $pa(X_i)$ 是 $Pa(X_i)$(G 中 X_i 的父节点集合)中每个变量取值后的一个组合,这些概率值的大小反应了变量间的依赖程度。

1.3 简单贝叶斯网络检索模型

简单贝叶斯网络 G_S 的变量集合 V_S 由两个不同的变量集组成,即 $V_s = T \cup D$。其中 $T = \{T_1, T_2, \cdots, T_M\}$ 是 M 个索引术语组成的集合, $D = \{D_1, D_2, \cdots, D_N\}$ 是 N 篇文档组成的集合。文中的符号 $T_i(D_j)(i = 1, 2, \cdots, M; j = 1, 2, \cdots, N)$ 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

收稿日期:2006 - 05 - 11;**修订日期**:2006 - 06 - 27
基金项目:国家自然科学基金资助项目(70471049);河北省科学技术研究与发展计划项目(04213534)
作者简介:徐建民(1966 -),男,河北馆陶人,教授,博士研究生,主要研究方向:信息检索、不确定性信息处理; 白彦霞(1979 -),女,河北晋州人,硕士研究生,主要研究方向:信息检索; 吴树芳(1980 -),女,河北磁县人,硕士研究生,主要研究方向:信息检索.

$\{\bar{t}_i, t_i\}$ 和 $\{\bar{d}_j, d_j\}$。$\bar{t}_i$ 和 t_i 分别表示“术语 T_i 不相关”和“术语 T_i 相关”;$\bar{d}_j$ 和 d_j 分别表示“文档 D_j 与给定的查询不相关”和“文档 D_j 与给定的查询相关”。

图 1 给出了简单贝叶斯网络检索模型的拓扑,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。这意味着术语相互边缘独立,文档在给定其所包含术语的情况下相互条件独立。这样便得到了仅包含一层术语和一层文档的简单贝叶斯网络检索模型,该模型由一个术语子网和一个文档子网组成。

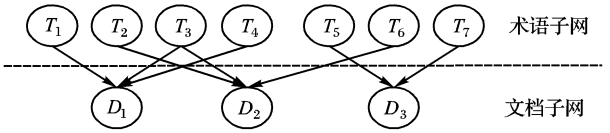

图 1　简单贝叶斯网络检索模型的拓扑

2　基于同义词扩展简单模型

2.1　扩展术语子网

在简单模型中,如果一篇文档 D_j 不包含查询 Q 中的任何术语,那么可以肯定即使索引该文档的术语与查询术语语义相同或相似也检索不到该文档,因为文档节点仅通过共同的术语节点相关。利用同义词扩展术语子网,加入模拟术语节点间直接关系的弧,就可以检索到那些与查询术语语义相同或相似的文档。

图 1 仅包含一层术语和一层文档,即仅包含模拟术语和文档间直接关系的弧,并不包含模拟术语间直接关系的弧。在扩展模型中,复制原始术语层 T 中的每个术语节点 T_i 得到术语节点 T'_i,形成一个新术语层 T',因此扩展模型 G_E 的变量集合 $V_E = T' \cup T \cup D$。T' 中的术语变量 T'_i 也是二进制的随机变量,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”和“术语 T'_i 相关”。这样就建立了两个完全相同的术语层 T 和 T',并通过在术语节点间加入弧来模拟术语间的关系,以此提高检索性能。

连接两个术语层的弧的指向:1) 任意术语 T'_i 与其本身 T_i 之间存在由 T'_i 指向 T_i 的弧,即 $T'_i \to T_i$;2) 若术语 T_i 与 T_j 互为同义词,则存在由 T'_i 指向 T_j 的弧和由 T'_j 指向 T_i 的弧,即 $T'_i \to T_j$, $T'_j \to T_i$。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。

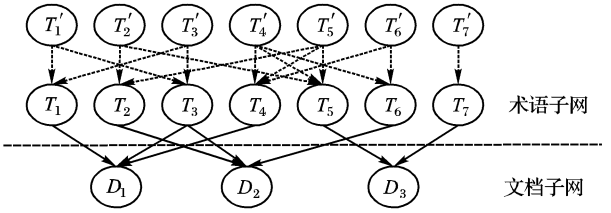

图 2　扩展模型的拓扑结构

扩展后的术语子网,定义任意根术语节点 T'_i 相关的边缘概率[6] 为 $P(t'_i) = 1/M$(M 为给定集合中的术语数量),其不相关的概率[6] 为 $P(\bar{t}'_i) = 1 - P(t'_i)$。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[5] 可得:

$$P\left(t_i \mid pa(T_i)\right) = \sum_{T'_j \in Pa(T_i),\, t'_j \in pa(T_i)} v_{ij} \tag{1}$$

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:

$$v_{ij} = \begin{cases} \beta, & 0.5 \leqslant \beta \leqslant 1.0 \text{ 且 } i = j \\ \dfrac{1-\beta}{|\,Pa(T_i)\,| - 1}, & i \neq j \end{cases} \tag{2}$$

若 T_i 只有一个父节点 T'_i,则权重定义为:$v_{ij} = 1.0$。

公式(2)中 $|\,Pa(T_i)\,|$ 表示术语节点 T_i 的父节点个数,β 为调节权重影响程度的参数。若 $\beta = 0.5$,则对于只有一个同义词的术语而言,术语本身和其同义词就具有了等同的重要性,但实际上术语本身要比同义词重要,故取值一般应大于 0.5;若 $\beta = 1.0$,则术语的同义词对其不产生影响,这种情况等价于简单贝叶斯网络检索模型。除此之外,这样定义既保证了 T'_i 对 T_i 的最大强度关系,又保证了每一个同义词 T'_j 对 T_i 有相同的强度关系,即挖掘了术语间的强度关系。

2.2　文档子网

文档子网中的弧由索引该文档的术语节点指向文档节点,文档节点 D_j 的父节点集合由该文档的所有索引术语节点组成,即 $Pa(D_j) = \{T_i \in T \mid T_i \in D_j\}$。令 $pa(D_j)$ 为 $Pa(D_j)$ 中每个术语变量取值(相关或不相关)后的一个组合,同公式(1)类似,定义文档 D_j 相关的条件概率:

$$P(d_j \mid pa(D_j)) = \sum_{T_i \in D_j,\, t_i \in pa(D_j)} w_{ij} \tag{3}$$

公式(3)中 w_{ij} 为文档 $D_j \in D$ 的索引术语 $T_i \in D_j$ 的权重,$w_{ij} \geqslant 0 \,\forall i, j$,且 $\sum_{T_i \in D_j} w_{ij} \leqslant 1.0 \,\forall j$。$t_j \in pa(D_j)$ 意味着只将 $pa(D_j)$ 中相关术语的权重相加,所以 $pa(D_j)$ 中相关术语越多,D_j 的相关概率值就越大。权重定义为 $w_{ij} = \alpha^{-1} \dfrac{tf_{ij} \times idf_i^2}{\sqrt{\sum_{T_k \in D_j} tf_{kj} \times idf_k^2}}$,其中 α 为规格化常数(用来保证 $\sum_{T_k \in D_j} w_{ij} \leqslant 1.0 \;\; \forall D_j \in D$)。$tf_{ij}$ 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 为测试集合中的文档数量,n_i 为包含术语 T_i 的文档数量。当然也可以使用其他的权重公式。

3　推理和检索

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'_{iQ} 的状态为 t'_{iQ}(相关);然后,据此在整个网络中推理,计算出每篇文档 D_j 与查询 Q 的相关概率 $P(d_j \mid Q)$;最后,文档以概率递减的顺序呈现给用户。

由于扩展模型复制了一层术语节点,这样贝叶斯网络中就有大量的术语节点且许多节点又有多个父节点。因此,即使对于小的文档集合,一般的推理算法其效率也不够理想。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,其相应的推理过程可分两步进行:

1) 估计术语层 T 中任意术语 T_i 的后验概率 $P(t_i \mid Q)$:

$$P(t_i \mid Q) = \sum_{T'_j \in Pa(T_i)} v_{ij} P(t'_j \mid Q) \tag{4}$$

因为术语层 T' 中的术语相互边缘独立,所以 $T'_j \in Q$ 时,则(4)式中的 $P(t'_j \mid Q) = 1.0$;否则 $P(t'_j \mid Q) = 1/M$。注意,对于只有一个父节点 T'_i 的术语 T_i 而言 $v_{ij} = 1.0$,若 $T'_i \in Q$,由(4)式可得 $P(t_i \mid Q) = P(t'_i \mid Q) = 1.0$,否则 $P(t_i \mid Q) = P(t'_i \mid Q) = 1/M$。若术语 T_i 有多个父节点,v_{ij} 用公式(2)代替,则:

$$P(t_i \mid Q) = \sum_{T'_j \in Pa(T_i),\, i \neq j} \frac{1-\beta}{|\,Pa(T_i)\,| - 1} P(t'_j \mid Q) + \beta P(t'_i \mid Q) \tag{5}$$

公式(5)考虑了 T_i 的所有父节点对其产生的影响,为了

有效地计算 $P(t_i \mid Q)$ 的值,可以简化为如下两种情况:

情况 1 对于任意术语 T_i,若 $Pa(T_i)$ 中的所有术语都未在查询 Q 中出现,(5) 式的最终结果为:$P(t_i \mid Q) = 1/M$。

情况 2 对于任意术语 T_i,若 $Pa(T_i)$ 中的术语在查询 Q 中出现,则分下列四种情况讨论:

① $Pa(T_i)$ 中的所有术语都在查询 Q 中出现,这种情况比较少见,(5) 式的最终结果为 $P(t_i \mid Q) = 1.0$;

② 只有 T'_i 在查询 Q 中出现,(5) 式转化为 $P(t_i \mid Q) = \frac{1-\beta}{M} + \beta$;

③ 只有 T_i 的部分或全部同义词在查询 Q 中出现,(5) 式转化为 $P(t_i \mid Q) = \frac{1-\beta}{\mid Pa(T_i) \mid - 1} \sum_{T'_j \in Pa(T_i),\ i \neq j} P(t'_j \mid Q) + \frac{\beta}{M}$;

④ T'_i 和 T_i 的部分同义词在查询 Q 中出现,这种情况也比较少见,(5) 式转化为 $P(t_i \mid Q) = \frac{1-\beta}{\mid Pa(T_i) \mid - 1} \sum_{T'_j \in Pa(T_i),\ i \neq j} P(t'_j \mid Q) + \beta$。

2) 基于以上推理,计算文档 D_j 的最终后验概率:

$$P(d_j \mid Q) = \sum_{T_j \in Pa(D_j)} w_{ij} P(t_i \mid Q) \qquad (6)$$

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。

4 实验与分析

表 1 SBN 与 EBN-β 的 Recall-Precision 对照

Recall	Precision					
	SBN (EBN-1.0)	EBN-0.9	EBN-0.8	EBN-0.7	EBN-0.6	EBN-0.5
0.1	0.7576	0.7687	0.8223	0.8871	0.9190	0.9504
0.2	0.6807	0.7049	0.7921	0.8532	0.8724	0.9375
0.3	0.6704	0.6599	0.7506	0.7873	0.8042	0.8372
0.4	0.6261	0.6572	0.7145	0.7523	0.7739	0.7578
0.5	0.5664	0.6481	0.7021	0.7188	0.7199	0.7566
0.6	0.5275	0.6290	0.6743	0.6770	0.6695	0.7232
0.7	0.4991	0.6027	0.6426	0.6479	0.6240	0.7007
0.8	0.4479	0.5936	0.6243	0.6208	0.5876	0.6315
0.9	0.3872	0.5441	0.5971	0.5864	0.5619	0.5820
1.0	0.2220	0.3767	0.4502	0.4648	0.4284	0.4177

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 701 篇文档作为文档测试集合,经处理后这些文档被 1083 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 18 个查询。为了准确比较简单模型和扩展模型的性能,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,分别比较它在 10 个标准的查全率(Recall)值所对应的平均查准率(Precision)值。实验结果如表 1 所示。从实验数据可以看出:扩展模型(EBN-β) 的检索性能明显优于简单模型(SBN),而且通过调节参数 β 的取值改变扩展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,扩展模型的检索效果最佳,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;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,扩展模型等价于简单模型。

5 结语

文章利用同义词表示术语间关系的拓扑结构,提出一个扩展的贝叶斯网络检索模型,并通过实验将新模型和原模型的检索性能进行分析与比较。结果表明:新模型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,尤其是检索非专业类文档,这主要是因为本实验所用的同义词识别工具——《同义词词林(扩展版)》,目前收录的词汇大部分是一般意义上的同义词而非专业领域的同义词,随着同义词识别技术的不断完善以及各种义类词典所收录的词汇不断扩充,所提模型会具有更好的应用价值。

致谢:本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] 殴洁,林守勋. 基于贝叶斯网络模型的信息检索[J]. 微电子学与计算机, 2003, 20(5): 83 - 87.

[2] DE CAMPOS LM, FERNANDEZ-LUNA JM, HUETE JF. Clustering terms in the Bayesian network retrieval model: a new approach with two term-layers [J]. Applied Soft Computing, 2004, 4(2): 149 - 158.

[3] DE CAMPOS LM, FERNANDEZ-LUNA JM, HUETE JF. Bayesian networks and information retrieval: an introduction to the special issue[J]. Information Processing and Management, 2004, 40(5): 727 - 733.

[4] 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.

[5] ACID S, DE CAMPOS LM, FERNANDEZ-LUNA JM, *et al.* An information retrieval model based on simple Bayesian networks[J]. International Journal of Intelligent Systems, 2003, 18(2): 251 - 265.

[6] DE CAMPOS LM, FERNANDEZ-LUNA JM, HUETE JF. The BNR model: foundations and performance of a Bayesian network-based retrieval model[J]. International Journal of Approximate Reasoning, 2003, 34(2/3): 265 - 285.

(上接第 2627 页)

为便于比较,设计了 2 个基准测试:1)位置抽取方法。提取每篇文章第一语句产生摘要。2)随机选取语句。这里的“随机”指的是随机从句子集合里面挑选语句的办法。经过 5 次随意选择后,挑选中值作为最终结果,实验结果见表 1。

结果证明对于多文档文本摘要,在 HITS 框架下结合启发规则和内容特征是一种有效的摘要算法。另一个方面,Authority 中的词汇能作为关键词来阐明一些文档中的主题。

参考文献:

[1] LIN CY, HOVY EH. The potential and limitations of sentence extraction for summarization[A]. Proceedings of the HLT/NAACL Workshop on Automatic Summarization[C]. Edmonton, Canada, 2003.

[2] ERKAN G, RADEV D. LexPageRank: Prestige in Multi-Document Text Summarization[A]. Proceedings of EMNLP 2004[C]. Barcelona, Spain, 2004.

[3] KLEINBERG JM. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604 - 632.

[4] WU J, KHUDANPUR S. Building a topic-dependent maximum entropy language model for very large corpora[A]. Proceedings of ICASSP[C], 2002, 1. 777 - 780.