不精确推理检索方法探讨

张　玉　峰

一、引　言

通常，情报系统中包含有大量不精确信息，例如：各种自然语言文本和用户情报提问中存在"相关"、"相似"、"大约"等模糊概念，文献标引中存在一些模糊处理。在检索过程中，人们的思维过程和查找处理也常涉及不精确的概念和量度。目前，多数情报检索系统不能处理模糊信息，因为它们是建立于传统集合论和两值逻辑的基础上的。为了处理不精确的信息，提高检索效率，应用不精确推理(或称模糊推理)的知识检索方法，是一种有效途径。

以模糊集合论和Dempster—Shafer证据理论为基础的不精确推理检索方法，不仅能处理模糊类型不精确信息，还可处理部分概率信息及不完整信息，并已被证明可应用于许多领域。这种检索方法可以处理用户的模糊提问，实现智能检索。

二、检索模型

检索模型基于模糊集合论和不精确推理方法，它涉及情报检索各方面的模糊特性，例如提问描述、文献描述、概念表达及检索方法。检索模型RM可定义为一个四元组：

RM = (C, D, Q, R)。

1. $C=\{C_1, C_2, \cdots, C_m\}$，表示概念集合，概念类似于标引词、关键词或主题词，通用于专业领域。一个概念 $C_i \in C$，可以是一个或一组关键词，如"自然语言处理"，表示领域中有意义的主题内容。在实际情报系统中，专业领域的概念和概念之间的模糊关系可被描述和存贮于知识库（或称概念库）。概念之间的等同关系，包括同义关系及组代关系，形如：$\text{Concept}_i \overset{S_{ij}}{\rightleftharpoons} \text{Concept}_j$，$S_{ij} \in [0, 1]$，表示概念$\text{Concept}_i$与$\text{Concept}_j$之间的等价度或相互代换度，若$S_{ij}=1$表示两个概念可完全相互代换。概念之间的蕴含关系是指等级关系（属种关系、包含关系等），形如：$\text{Concept}_K \xrightarrow{I_{ke}} \text{Concepte}$，$I_{ke} \in [0, 1]$，表示两个概念之间的蕴含度。为使用方便，用符号RL统一表示等同关系和蕴含关系。一般来说，专家提供的概念关系是不完整的，可以利用关系的传递性和等同关系的对称性推导新的关系。这些概念和概念之间的关系可以用语义网络或产生式规则表示。在实际应用中，概念库可独立构造，也可与模糊索引集合合并，生成扩展模糊词表(或词典)。

2. $D=\{d_1, d_2, \cdots, d_n\}$，代表文献描述集合。文献知识的获取和表示应是文献内容的语

义分析和表示，然而根据目前的技术条件和语言处理能力，实现较困难。人们通常用较简单的自然语言处理方法来描述文献。在各种情报中，概念起基本实体作用，在模糊检索系统中，可将每篇文献$d_i \in D$，看作概念集合C上的一个模糊子集，用隶属函数描述：$\mu d_i: C \longrightarrow [0, 1]$，其中$\mu d_i(C_j)$是概念$C_j$在文献描述$d_i$中的隶属度或权（或用$W_{ij}$表示）。因此，每篇文献描述$d_i$可表示为概念——权对的向量：

$\{(C_1, \mu d_i(C_1)), (C_2, \mu d_i(C_2)), \cdots, (C_m, \mu d_i(C_m))\}$。不出现于某篇文献中的概念的隶属值为零

设模糊标引的文献描述样例如下：

(1 (模糊检索 1.0) (模糊词表 0.7))
(2 (出处研究 0.9) (聚类分析 0.6))
(3 (模糊词表 1.0) (算法 0.8) (检索试验 0.8))
(4 (模糊逻辑 1.0) (模糊提问 0.8))
(5 (语义分析 0.9) (不精确推理 0.8) (正向推理 0.6))
(6 (检索试验 0.8) (检索效率 0.6))

为了实现有效的查找，领域专家常根据概念集合C，将文献向量空间划分为m个子集合，$\{D_{C1}, D_{C2}, \cdots, D_{Cm}\}$，每个子集合$D_{Ci}$对应C中一个概念结点$C_i$。

3．$Q = \{q_1, q_2, \cdots, q_k\}$，表示自然语言模糊提问的有限集合。一个情报提问$q_i \in Q$，是一个自然语言提问或模糊逻辑表达式，表示用户检索有关文献的要求。情报提问可包含一些主题概念(如一个或多个关键词)，描述用户感兴趣的主题。提问还可包含一些模糊词，如"相关的"、"重要的"，用以修饰概念，并限制概念的相关模糊运算。逻辑运算符AND、OR和NOT连接提问中的概念，在检索处理第一阶段的布尔检索中，它们作为布尔逻辑运算符处理；在模糊检索中，它们作为模糊逻辑运算符，运算公式如下。设情报提问中包含模糊概念C_1和C_2，它们与某篇文献的相关测量分别是$m(C_1)$和$m(C_2)$，那么

$$m(\text{NOT } C_1) = 1 - m(C_1) \quad (1)$$

$$m(C_1 \text{ AND } C_2) = \min(m(C_1), m(C_2)) \quad (2)$$

或 $$m(C_1 \text{ AND } C_2) = \sqrt{m(C_1) * m(C_2)} \quad (3)$$

$$m(C_1 \text{ OR } C_2) = \max(m(C_1), m(C_2)) \quad (4)$$

同理，可运用于多个概念的连接运算。

用户提问中可能出现的相关性模糊概念，均定义相应的模糊函数存贮于模糊知识库，并产生相关性阈值，用于判断检索信息是否满足提问要求。用户提问中的模糊量词，如"大多数"、"几篇"，表示用户的检索数量要求，可使用模糊逻辑方法将其转换为数量阈值，以决定满足提问要求的文献输出量。

4．R代表模糊检索函数，可定义为笛卡尔积$Q \times D$上的模糊关系，$R: Q \times D \longrightarrow [0, 1]$。

R给每对(q, d)赋值，R(q, d)是闭区间[0, 1]上的一个数字，其中$q \in Q$，$d \in D$，这个数字表示文献d与提问q的相关测量，或文献d对提问q的支持度。

对于一个情报提问q，$A = \{d \mid R(q, d) > \lambda\}$，$A \in D$，A是检索结果，是一个文献子集合。A的大小依赖于检索函数R和阈值λ，λ是相关性或数量方面的阈值，可由系统或用户定义，或由系统从用户提问中导出。

检索函数可基于不同的理论，使用不同的运算方法，包括提问与文献特征的匹配和组合运算。模糊检索以模糊集合论为基础，应用不精确推理方法，进行相关性计算。

应用不精确推理，是人工智能型检索方法与传统检索方法的主要区别。

目前，在人工智能领域中，不精确推理理论主要有：确定性理论、贝叶斯概率论、模糊集合论和Dempster—Shafer证据理论。MYCIN证实模式基于确定性理论，并已用于大多数专家系统；贝叶斯概率论的推理检索模式虽有多种，但只能处理概率信息，不能处理模糊信息，尚未应用到实际中去；模糊集合论和Dempster—Shafer证据理论提供描述和处理事物的方法，更为接近客观真实世界，有很强的生命力。模糊集合论提供表达领域中模糊概念和概念之间关系的模式。证据理论是概率论的一般形式，当先验概率很难获得的时候，证据理论恰恰能解决这种由不知道引起的不精确性。下面简要介绍Dempster—Shafer不精确推理理论。

设Ω是样本空间，研究领域中的命题或假设用Ω的子集表示。对任一个属于Ω的子集A，命它对应一个数m(A)∈〔0，1〕，函数m：$2^{\Omega}\longrightarrow$〔0，1〕称为基本概率分配函数（basic probability assignment），它满足：

$$m(\phi)=0 \qquad (5a)$$

$$\sum_{A\subset\Omega} m(A)\leqslant 1 \qquad (5b)$$

m(A)为A的基本概率值，表示对A的精确信任程度。(5a)式表示对空集合的不信任，(5b)式表示总的信任值小于或等于1。

定义 命题的信任函数(belief function)Bel：$2^{\Omega}\longrightarrow$〔0，1〕为

$$Bel(A)=\sum_{B\subseteq A} m(B) \qquad (6)$$

对所有的A∈Ω，其中2^{Ω}表示Ω的所有子集。

由m的定义，很容易导出 $Bel(\phi)=m(\phi)=0$ (7a)

当(5b)式中总的信任值小于1时，则信任函数

$$Bel(\Omega)=\sum_{A\subset\Omega} m(A)+m(\Omega)=1 \qquad (7b)$$

其中m(Ω)表示未知量度，即这个数不知如何分配。

定义 似然函数（plausibility function）pl:$2^{\Omega}\longrightarrow$〔0，1〕为

pl(A)=1－Bel(～A)，对所有的A∈Ω。

Bel(A)和pl(A)分别称为命题A的下限和上限。

所谓Demster—Shafer证据组合规则，就是对于所给的若干不同证据，组合信任函数计算新的信任函数。这种组合规则是贝叶斯规则的一般形式，可通过定义组合基本概率分配函数的规则来实现。

设m_1和m_2是二个基本概率分配函数，表示相同样本空间中不同的不确定证据，Dempster—Shafer证据组合规则定义m_1和m_2的正交和$m=m_1\oplus m_2$如下：

$$m(C)=\begin{cases}0 & \text{当}C=\phi\\ \sum\limits_{A\cap B=C} m_1(A)\cdot m_2(B) & \text{当}C\neq\phi\\ -1\sum\limits_{A\cap B=\phi} m_1(A)m_2(B)\end{cases} \qquad (8)$$

三、检索算法

不精确推理检索算法，其匹配功能应用不精确的蕴含规则进行正向和逆向推理，由已知事实推得检索概念的相关值。Dempster—Shafer证据理论用于组合推理规则，确定文献与

提问之间的相关值。在推理检索过程中，必须应用专业领域和检索专家的知识。例如专家利用概念之间的关系组织和使用词汇的知识。这种情报系统就能够以类似于专家的方法理解用户提问，执行高效的检索处理。

为便于描述推理检索算法，有必要对概念知识库中的知识描述加以说明。概念之间的关系，如概念C_i等同或蕴含概念C_j，即$C_i \xrightarrow{RL_{ij}} C_j$，等价于产生式规则：if C_i then C_j (RL_{ij})。其中C_i为规则的前提，D_j为结论，RL_{ij}为C_i与C_j之间的关系强度，可看作规则的可信度。

1．推理检索算法

推理检索算法包括两个步骤：

步骤1　产生检索候选集。本步骤产生一个与情报提问相关的文献集合。首先，扫描用户提问中的检索概念，以正向推理方式搜索概念库，分别推导出与每个检索概念相关的文献集合。然后，对这些集合执行布尔逻辑运算，求出与提问相关的候选文献集合D_q。

步骤2　产生检索结果集。对于候选集D_q中的每篇文献，重新扫描用户提问，应用知识库中的规则，进行逆向推理，推导出每个检索概念基于文献证据的信任值(相关值)，然后应用Dempster—Shafer 证据组合规则和模糊逻辑运算，求出文献与提问之间的相关值。最后，按照这种相关值分类文献，并根据用户的数量要求，产生和输出结果文献集合。

关于候选集D_q中每篇文献处理的子算法描述如下。给出已知事实文献$d_i \in D_q$，文献中的每个主题概念作为证据，提问q中每个检索概念作为假设，目标是求出提问基于文献d_i的信任值(即提问与文献的相关值)。

(1) 基于文献d_i中的一个证据，用逆向推理方式，求出提问中每个检索概念的信任值。给出概念$C_j \in d_i$，$C_k \in q$，首先用检索概念C_k匹配概念库中规则的结论部分。若匹配成功，存在规则 if C_j then C_k (RL_{jk})。那么C_j为支持假设C_k的一个证据。然后从文献库中取出C_j在文献d_i中的权值W_{ij}。C_k基于证据C_j的信任值$m(C_k)$可用下式求得：

$$m(C_k) = W_{ij} * RL_{jk} \tag{9}$$

(2) 组合文献的所有证据，求出每个检索概念基于文献d_i的信任值。利用证据组合公式(8)，组合由文献的每个证据所求出的各检索概念的信任值，求得每个检索概念基于文献所有证据的信任值。

(3) 求出整个提问基于文献的信任值。首先，应用相关性阈值评估文献与提问中每个检索概念之间的相关度，然后应用模糊逻辑运算公式(1)--(4)，对所有符合要求的检索概念的信任值进行运算，求出提问与文献的相关值。

2．检索实例

设用户的自然语言提问是："请给一些关于专家系统和自然语言处理的文献"。(注：下面用ES代表提问中的专家系统，用NLP代表自然语言处理)。

系统将提问转换为内部形式，主要包括数量表和概念表。数量表描述用户希望检索的资料数量，由数量阈值控制。概念表包含提问中的主题概念(或检索概念)、相关性阈值与逻辑运算符。以上提问可转换为以下形式：数量表：　(3)

概念表：((F 0.3)(C ES)(OP AND)(F 0.4)(C NLP))。其中F指示后继概念的模糊相关性阈值；OP表示逻辑运算符；C表示检索概念。

推理检索过程如下：

步骤1　根据提问中检索概念ES 和NLP，正向搜索概念库，分别推得对应的相关文献集 合(用文献号表示)为

ES：{5，8，12，15，20，21，25}

NLP：{5，12，15，20}

根据布尔逻辑运算符“AND”，以上两集合进行交运算，得到集合{5，12，15，20}，就是所求候选文献集合。本步骤暂不考虑F型模糊元素。

步骤2　求出候选集合中每篇文献与情报提问的相关程度。文献信息中每个主题概念都是一个证据，情报提问中每个检索概念作为一个假设，应用逆向推理，求出假设基于每个证据的信任值。然后，应用 Dempster 证据组合规则及模糊逻辑运算，确定文献与情报提 问的相关程度。现以文献 5 为例说明执行过程。文献 5 的描述为：

(5 (语义分析　0.9)(不精确推理　0.8)(正向推理　0.6))。

①　建立初始信任值

m(ES) = 0，m(NLP) = 0，m(Ω) = 1．因为开始时，无证据支持 2^{Ω}中的任何子 集 合，根据公式(7b)，置未知量度m(Ω)为 1 。

②　文献 5 中第一个证据是概念“语义分析”，由知识库中等同和蕴含知识推出：语义分析 $\xrightarrow{1.0}$NLP，新的信任值为

m(NLP) = 0.9 * 1.0 = 0.9　{由公式(9)}

m(ES) = 0　　m(Ω) = 1 − 0.9 = 0.1　　{由公式(7b)}

③　第二个证据是概念“不精确推理”，同理由知识库推出：不精确推理 $\xrightarrow{0.8}$ES，新的信任值为 m(ES) = 0.8 * 0.8 = 0.64　　m(NLP) = 0　　m(Ω) = 1 − 0.64 = 0.36

根据以上两个证据，应用Dempster组合规则产生新的信任值

	ES(0.64)	Ω(0.36)
NLP(0.9)	φ(0.576)	NLP(0.324)
Ω(0.1)	ES(0.064)	Ω(0.036)

$$m(ES)=\frac{\Sigma ES(\ )}{1-\Sigma\phi(\ )}=\frac{0.061}{1-0.576}=0.151$$

$$m(NLP)=\frac{\Sigma NLP(\ )}{1-\Sigma\phi(\ )}=\frac{0.324}{1-0.576}=0.764$$

$$m(\Omega)=\frac{\Sigma\Omega(\ )}{1-\Sigma\phi(\ )}=\frac{0.036}{1-0.576}=0.085$$

④　文献 5 中另一个证据是概念“正向推理”，由知识库推出：正向推理 $\xrightarrow{1.0}$ES，可得m(ES) = 0.6 * 1.0 = 0.6　m(Ω) = 0.4将③中的结果作为一个证据，组合这两个证据产生新的信任值

	ES(0.6)	Ω(0.4)
ES(0.151)	ES(0.091)	ES(0.06)
NLP(0.764)	φ(0.458)	NLP(0.306)
Ω(0.085)	ES(0.051)	Ω(0.034)

$$m(ES)=\frac{0.091+0.06+0.051}{1-0.458}=0.373$$

$$m(NLP)=\frac{0.306}{1-0.458}=0.565$$

$$m(\Omega)=\frac{0.034}{1-0.458}=0.063$$

⑤　根据概念表中的相关性阈值，文献 5 与提问中两个检索概念 ES和NLP 的相关度满足要求。两个概念之间有“AND” 运算符，进行模糊逻辑运算，文献 5 与整个提问的相关度为

$m(ES \quad AND \quad NLP) = \sqrt{m(ES) * m(NLP)} = \sqrt{0.373 * 0.565} = 0.459$

⑥ 利用处理文献5的同样方法，可以求出文献号为12、15、20的文献的信任值，并根据数量表中的阈值，将信任值较高的文献输出给用户。

四、结 论

不精确推理检索同其它检索方法相比较，具有下述基本特点：

1．这种检索方法基于模糊集合论，文献的模糊标引和检索能够充分表达和有效地检索文献的知识内容。还可进一步探讨新型索引和书目信息分析的方法，如模糊词表的自动产生和文献聚类。模糊集合论是情报检索理论的最好框架，它将促进情报检索领域的发展。

2．Dempster—shafer证据理论的不精确推理方法有光明的应用前景，可以处理等级知识结构中的多元继承问题，通过组合证据的模糊推理计算，处理多元相关概念，能较客观准确地确定文献与提问的相关度。通过概念知识库的查找，不仅检索包含用户检索词的文献，还可检索包含检索词相关概念的文献。若系统适当组合概念知识库的查找和检索处理，可执行反馈检索。因此，不精确推理检索方法可实现文献的智能检索，最大可能地提高检全率和检准率。

3．系统能处理用户提问中的模糊信息，有助于建立友好用户接口。

4．系统结果可以分等排序输出，用户可通过阈值方便地控制输出。

（本文责任编辑　江平）

（上接121页）

存和发展。因此，民族参予国际竞争的意识和能力之强弱，会直接影响到国家、民族的历史命运。这是一个很有见地和富有现实意义的见解。最后，作者还提出，作为整体的近代世界史，其历史阶段性分期的标准，除依据社会形态的更迭外，还应考虑反映世界历史整体发展的过程，并提出了自己的三段分法的主张。

综观全书，作者对世界体系的论述，实际上已涉及到了近代世界史的概念、结构、基本内容、各种世界性的历史运动及历史发展的动力、历史发展总趋势和各地区在其中的地位和作用、历史阶段性的分期等重要问题，这对于重构一个具有中国特色和体现时代新精神的世界史新学科体系，也是很有裨益的。诚然，作为史学研究中的一家之言，我们不能苛求本书尽善尽美和完整无缺。重要的是作者把近代世界史研究中遇到的问题和困惑，集中地提到了读者面前，并提出了能引人共鸣或深思的见解。但愿本书能获得史学界应有的注意与评价。