北京大学学报(自然科学版) 第 52 卷 第 2 期 2016 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 2 (Mar. 2016)
doi: 10.13209/j.0479-8023.2015.113

一种地理信息检索的定性模型

高勇† 姜丹 刘磊 林星 邬伦

北京大学遥感与地理信息系统研究所, 北京 100871; † E-mail: gaoyong@pku.edu.cn

摘要 提出一种定性地理信息检索方法, 用于地理信息的定性表达、语义匹配、推理和结果排序, 可以避免目前定量地理信息检索中语义信息丢失问题。采用命题逻辑方法综合表达查询和文档中的主题信息和地理语义信息, 将文档与查询的相关性度量分为主题相似度和地理相似度。前者通过命题关键词间加权本体距离获得。后者可进一步分为概念相似度和位置相似度, 分别基于地理本体和空间语义度量。由于信息的表达形式为命题和信息单元, 采用证据理论和模糊逻辑对上述子相关性度量进行统一建模。所提方法可以基于语义检索网页中的定性地理信息, 并对相关文档进行排序。这种检索和排序方法符合人类空间认知, 因此可以有效提高地理信息检索的效率。

关键词 地理信息检索; 检索模型; 定性空间推理

中图分类号 P208; TP311

A Qualitative Method for Geographic Information Retrieval

GAO Yong†, JIANG Dan, LIU Lei, LIN Xing, WU Lun

Institute of Remote Sensing and Geographical Information System, Peking University, Beijing 100871; † E-mail: gaoyong@pku.edu.cn

Abstract A qualitative method is presented for geographic information retrieval (GIR) to support qualitative representation, semantic matching, reasoning and ranking. The novel approach can avoid semantic information lost in current quantitative GIR methods. Information in documents and user queries are represented by propositional logic, which considers the thematic and geographic semantics synthetically. The similarity between documents and queries can be divided into thematic similarity and geographic similarity. The former is calculated by the weighted distance of proposition keywords in domain ontology, and the latter is further divided into conceptual similarity and location similarity which are measured by geo-ontology and spatial semantic respectively. Represented by propositions and information units, the similarity measurement takes evidence theory and fuzzy logic to obtain a general similarity from all sub similarities. This novel method retrieves qualitative geographic information from web and ranks documents semantically, which is consistent with commonsense, and thus can improve the efficiency of geographic information retrieval technology.

Key words geographic information retrieval; retrieval model; qualitative spatial reasoning

地理信息检索(geographic information retrieval, GIR)主要关注与地理位置相关的信息源的信息提取、存储、索引、查询、排序和浏览问题[1], 可以根据用户提交的请求, 从网页文本中检索在空间语义或空间范围上匹配的信息, 而这类信息通常以地名、地址等文本形式存在。地理信息检索扩展了传统信息检索(information retrieval, IR)的方法, 侧重于文档中与特定地理位置相关信息的处理[2], 其中的关键问题包括提取地理参照、地名去歧义、模糊地理信息处理、空间和文本索引、地理相似度排序、用户接口和结果评价等 7 个方面[3]。

当前的地理信息检索系统是将文档中的地理信息和主题信息分别提取、存储和匹配。其中, 主题信息部分沿用传统信息检索的方法, 基于文本关键

国家自然科学基金(41271385)资助
收稿日期: 2014–12–24; 修回日期: 2015–01–30; 网络出版日期: 2016–01–21

字之间的共现程度评价文档与查询之间的主题相似度[4]。地理信息部分则基于传统 GIS 技术, 在地名辞典或地名库的帮助下, 将地名或地址等文本转换成以坐标表达的几何图形, 作为文档的地理范围(geographic footprint), 如中心点、外包矩形、凸包、简化多边形、点集等[5−9]。类似地, 检索请求中的地理约束条件同样被转化为定量的地理范围。在检索匹配的过程中, 主要基于文档和查询的地理范围之间的相对交叠面积或距离度量等方法[5,10]进行空间相似度计算。最终, 将主题相似度与空间相似度集成[11−12], 对检索结果进行排序。

这种表达和相似度计算方法简化了地理信息处理的复杂性, 但却存在一些问题。主题与地理信息相互割裂的表达与处理办法, 忽略了两者的内在联系, 不符合人们日常表达和决策习惯, 使检索结果与用户实际信息需求存在一定的差异。采用单一的文档地理范围表达方法, 导致对文档地理范围的低估或高估。使用基于几何图形的地理范围定量化表达和基于交叠面积或欧氏距离的相似度计算方法, 会遗失大量的空间语义信息, 降低检索精度。

针对上述问题, 本文提出一种地理信息检索的定性模型, 基于命题逻辑和不确定性推理, 一体化建模主题信息和地理信息, 并评价文档相关性, 使其符合人类以概念化和自然语言形式使用地理空间知识的习惯, 从而可以很好地理解用户的地理信息检索请求和待检索的文档内容, 更准确地获取真正符合其检索需求的信息。

1 地理信息检索的定性表达方法

1.1 定性表达模型

地理信息检索处理文档和查询两类数据, 它们都是用自然语言表达的文本形式, 其中包含主题和地理两类信息, 前者对应一系列文本关键词, 后者主要表现为文本地名、地址、IP 地址、电话区号以及空间关系谓词或其组合。查询请求则可以表示为<theme><spatial relationship><location>三元组[3], 其中 theme 表示用户关心的主题信息, spatial relationship 和 location 共同表示相关的地理空间范围约束。

为了更好地处理传统地理信息检索所缺失的语义特征, 这里以命题逻辑为基础, 建立主题信息与地理信息一体化的定性表达方法。命题逻辑是以逻辑运算符结合原子命题构成代表命题的公式, 以及允许某些公式建构成定理的一套形式证明规则。其中, 命题是能分辨真假的陈述句, 原子命题是能指派真假值的最小项。

首先将文档经过文本切割等预处理操作, 转换成一系列独立的信息单元, 每个单元均由<主题信息, 位置信息>组成, 表述一个独立的信息内容。基于命题逻辑, 将文档表达为命题的形式。相应地, 主题信息的每个关键词表达为一个主题命题, 位置信息则表达为地理空间命题的集合, 其中每个地理空间命题由地名、空间关系谓词及其逻辑连接符组成。进一步, 两类信息以命题逻辑组合, 形成一体化的表达方法。查询也采用相同的表达模型。

更一般地, 给定文档 d, 基于主题信息与地理信息之间的关联性, 将其分割成一系列独立的信息单元, 则文档 d 的信息内容表示为

$$d=\{u_i \mid 1\leqslant i\leqslant n\}, \tag{1}$$

其中, n 为文档 d 包含的信息单元的数量, 如果 $n=0$, 则 d 是一个空文档, 不包任何信息内容。信息单元 u_i 包含主题部分 t_i 与地理部分 g_i, 表示为

$$u_i=(t_i, g_i), \tag{2}$$

相应地, 查询 q 基于信息单元表示为相同的结构。

1.2 主题信息的定性表达

将信息单元 u_i 的主题信息 t_i 表示为该信息条目中出现的主题命题的集合, 即

$$t_i=\{\mathrm{tp}_i(k) \mid 1\leq k\leq \mathrm{NK}_i\}, \tag{3}$$

其中, NK_i 是对应信息单元 u_i 中表示主题内容的关键词数目; $\mathrm{tp}_i(k)$为主题命题语句, 可以是单个关键字或者是一个由关键字与连接符组成的复合主题命题语句。

基于命题逻辑, 主题命题语句由命题、逻辑连接符 {∧, ∨, ¬}组成, 定义其合成规则如下。

1) 单个关键字是原子主题命题语句。

2) 如果 P 是主题命题语句, 那么它的逆$\neg P$ 也是主题命题语句。

3) 如果 P 和 Q 是主题命题语句, 那么合取式 $P\wedge Q$ 也是主题命题语句。

4) 如果 P 和 Q 是主题命题语句, 那么析取式 $P\vee Q$ 也是主题命题语句。

一般情况下, 文档中的一个信息单元可能包含多个主题关键词。在信息检索中, 由于文档内容经过分词等处理后得到的是词的罗列, 因此认为关键词是基于析取符号连接构成主题命题。由于查询是

关键词的布尔逻辑，所以只有在查询中存在合取和逆连接符。例如“军事新闻”是“军事”和“新闻”两个关键词的合取，“非官方消息”是“官方消息”的逆。这种情况在地理信息的表达中同样存在。

在该形式化表达的过程中，需要引入领域本体，对分词处理后的文档，利用领域本体术语表对关键词汇进行识别、匹配和替换。

1.3　地理信息的定性表达

信息单元 u_i 中包含的地理信息内容 g_i，表示为地理空间命题的集合，即

$$g_i = \{\mathrm{gp}_i(m) \mid 1 \le m \le \mathrm{NG}_i\}, \tag{4}$$

其中，地理空间命题 $\mathrm{gp}_i(m)$是一个描述地理位置或范围的表达式，NG_i 是对应信息单元 u_i 中地名的数目。地理空间命题可以继续细分，直至地理空间元命题。每个地理空间元命题由一个空间关系谓词和一个地名组成。地理空间命题之间通过逻辑连接符相互关联、嵌套，组成对地理范围的定性描述。

基于命题逻辑，地理命题语句包括命题、逻辑连接符{∧，∨，¬}和空间关系算子 $\wp$，定义其合成规则如下。

1) 单个地名，或由一个空间关系谓词修饰下的单个地名，或由一个¬操作符修饰下的单个地名，是原子地理空间命题语句，例如“北京”。

2) 如果 $\wp$是一元空间关系算子，p 是地名，那么 $\wp p$ 是地理空间命题语句，例如“北京南部”、“上海周边”等。

3) 如果 $\wp$是二元空间关系算子，p 和 q 是地名，那么 $p \wp q$ 是地理空间命题语句，例如“一环路和二环路之间”。多元空间关系算子依此定义，例如“被太平洋、印度洋和大西洋包围”。

4) 如果 P 是地理空间命题语句，那么它的逆 $\neg P$ 也是地理空间命题语句，例如“北京”和“非北京”。

5) 如果 P 和 Q 是地理空间命题语句，那么合取式 $P \wedge Q$ 也是地理空间命题语句。

6) 如果 P 和 Q 是地理空间命题语句，那么析取式 $P \vee Q$ 也是地理空间命题语句。

值得注意的是，本文引入的空间关系算子$\wp$的操作元数量比通常所指的空间关系元数少 1，例如“……与……邻接”是二元空间关系，而作为空间关系算子“与……邻接”是一元的。另外，由于空间关系在不同场景下的模糊性不同[13]，本文所提出的表达和推理方法仅针对文本表示的地理信息检索领域。

在前人的研究中，确定了地理空间中的空间关系谓词表达及其推理演算方法[14–15]，可以基于此建立空间关系算子集合。尽管空间关系算子同时支持对度量和方位关系的定量表达，但在自然语言或网页文本中仍以定性描述为主要形式。

2　定性检索模型

2.1　相似度计算方法

在地理信息检索中，判断候选文档是否满足用户的检索请求，需要进行文档和查询的相似度计算，并据此排序检索结果，这其中需要同时从主题和地理范围两个方面评价相似度。需要指出的是，相似度的评价具有方向性，即文档满足查询的程度与查询满足文档的程度并不完全相同。例如，当查询“西餐厅”时，检索结果返回“纽约披萨”是符合要求的；反之，检索“纽约披萨”时，检索结果返回“西餐厅”则不那么符合要求。从推理的角度看，相似度推理是方向相关的，是 $d \rightarrow q$ 的推理过程置信度。

基于上述定性表达模型，文档 d 由 n 个信息单元组成，查询 q 包含 m 个子查询，子查询是原子的、相互独立的，并由逻辑连接符相连。据此提出基于证据理论和模糊逻辑的语义相似度计算模型，将相似度求解过程分解为 3 个步骤。

1) 计算文档的每个信息单元 u_i 与子查询 q_j 的相似度，记为 $\mathrm{Ru}(u_i, q_j)$，是 u_i 与 q_j 主题相似度和地理相似度的组合。

2) 将文档 d 中所有信息单元与子查询 q_j 的相似度合成，计算文档 d 满足子查询 q_j 的程度，记为 $\mathrm{Rq}(d, q_j)$。

3) 将查询 q 的所有子查询的相似度合成，形成最终文档 d 对查询 q 的满足程度，记为 $R(d, q)$。

在上述过程中，涉及几个不同层次的相似度的计算，可以将其区分为两个类别：信息单元的主题信息与地理信息、查询与子查询等是基于命题逻辑表达，因此其相似度的合成函数采用模糊逻辑推理的方法；文档是信息单元的集合表达，因此将信息单元作为文档内容的证据，基于证据理论合成信息单元相似度。下面将详细论述上述相似度的计算过程。

2.2　信息单元与子查询的相似度计算

由于子查询可以表达为一个简单的信息单元，因此信息单元 u_i 与子查询 q_j 的相似度 $\mathrm{Ru}(u_i, q_j)$可

以归结为两个信息单元的相似度计算。每个信息单元均由主题部分和地理部分组成, 因此其相似度包括主题相似度、地理相似度及其组合 3 个部分。

在定性表达模型中, 各要素都是基于命题逻辑表达, 由命题变元基于逻辑连接符{∧, ∨, ¬}连接。命题变元相似度的合成可以基于模糊逻辑推理完成。模糊逻辑是用于近似推理的逻辑[16], 是基于规则的, 规则的前提是基于逻辑算子建立的模糊集的组合, 规则的结论是一个具有相应隶属函数的模糊集。在匹配规则的过程中, 命题变元的逆是其模糊集合的补, 而命题变元的合取和析取分别采用最小算子和最大算子进行合并。

2.2.1 主题相似度

两个信息单元在主题上的相似程度, 依赖于其传达的相同信息内容的程度, 是基于主题关键词之间语义相近程度的评价, 与主题词共现频率的评估并不相同。设 a 和 b 是两个主题关键词, 它们之间的主题相关性算子用⊕表示, 那么,

$$\begin{cases} a \oplus b = h(a,b), \\ a \oplus \neg b = 1 - h(a,b), \end{cases} \tag{5}$$

其中 $h(a, b)$为 a 和 b 的概念相似度函数。由于相似度的方向性, ⊕不满足交换律, 即 $a \oplus b \neq b \oplus a$。

对于文档中的一个信息单元, 其主题信息 X 由关键词集合$\{x_1, x_2, \ldots, x_n\}$基于析取连接符构成。设 a 是一个关键词, 则 X 与 a 的主题相似度计算为

$$\begin{cases} X \oplus a = \max\{x_i \oplus a \mid i = 1,2,\ldots,n\}, \\ X \oplus \neg a = 1 - \min\{x_i \oplus a \mid i = 1,2,\ldots,n\}。 \end{cases} \tag{6}$$

基于模糊逻辑的最大–最小方法, 两个主题信息 $X=\{x_1, x_2, \ldots, x_n\}$和 $Y=\{y_1, y_2, \ldots, y_k\}$的主题相似度计算为

$$X \oplus Y = \min\{X \oplus y_i \mid i = 1,2,\ldots,k\}。 \tag{7}$$

设 $H=\{h_1, h_2\}$是一个检索查询的主题命题部分, 由两个关键词组成, ⊎是连接子命题的逻辑连接符, 可以为∨或∧, 那么文档信息单元与查询的主题信息相似度计算为

$$X \oplus H = \begin{cases} \max\{X \oplus h_1, X \oplus h_2\}, & \uplus = \vee, \\ \min\{X \oplus h_1, X \oplus h_2\}, & \uplus = \wedge。 \end{cases} \tag{8}$$

任何复杂的查询命题, 总能分解为上述形式, 因此核心的问题是关键词之间语义相似度 $h(a, b)$的计算。

领域本体中概念之间的相互关系提供了度量两个关键词之间语义相似度的途径, 因此基于概念加权最短距离[17]提出主题关键词语义相似度的计算方法。本体中的概念以及概念之间的联系形成一个树状或网状结构(概念语义网), 其中概念为节点, 概念之间的关系为连接。概念之间比较重要的关系有 BT (广义词)关系、NT (狭义词)关系和 RT (相关词)关系。两个概念之间最短路径定义为: 在概念语义网中, 从一个概念出发, 经过最少中转节点到达另一个概念的路径。加权最短距离就是根据最短路径上概念两两间直接连接关系的强弱不同, 为对应连接设置不同的权重, 然后将最短路径上每条连接的加权距离值相加, 即得到两个概念之间的加权最短距离。

设 a 和 b 为两个主题关键字, 对应于本体中的概念, 其加权最短路径 TD(a, b)为

$$\mathrm{TD}(a,b) = \left(\frac{C_{a,x_1}}{L_{x_1}} + \frac{C_{x_1,x_2}}{L_{x_2}} + \ldots + \frac{C_{x_{n-1},x_n}}{L_{x_n}} + \frac{C_{x_n,b}}{L_b} \right), \tag{9}$$

C_{ij} 表示连接两个相邻概念 i 和 j 的路径权重, 用 $i \rightarrow j$ 之间关系的强弱程度进行赋值。L_i 表示概念 i 在概念树中的深度。$a \rightarrow x_1 \rightarrow x_2 \rightarrow \ldots \rightarrow b$ 为从概念 a 到概念 b 的最短路径。对式(9)的计算结果进行归一化, 作为最后的语义相似度值。可以用概念树中两个概念间可能的最大距离 MD 进行归一化, 也可以用对数表达式进行归一化。那么两个主题关键词 a 和 b 的概念相似度 $h(a, b)$的计算公式为

$$h(a,b) = \frac{1}{1 + \ln(1 + \mathrm{TD}(a,b))} \quad \text{或}$$

$$h(a,b) = 1 - \frac{\mathrm{TD}(a,b)}{\mathrm{MD}}。 \tag{10}$$

在实际应用中, 需要对 NT, BT 和 RT 等关系进行加权, Tudhope等[17] 的做法是$w_{\mathrm{NT}} = w_{\mathrm{BT}} \leqslant w_{\mathrm{RT}}$, 也就是通过 RT 关系进行连接的两个概念, 要比通过 NT 或 BT 关系连接的两个概念在语义上离得远些。具体数值的确定过程都是凭经验, 例如可以设定 $w_{\mathrm{NT}} = w_{\mathrm{BT}} = 0.5$, $w_{\mathrm{RT}} = 0.8$ 等。但这种关系权重设置方法并没有考虑概念相似度的有向性问题。由于 $a \rightarrow b$ 与 $b \rightarrow a$ 的语义相似度不同, 因此对 BT 和 NT 关系设置不同的权重, 并令其满足 $w_{\mathrm{NT}} < w_{\mathrm{BT}} < w_{\mathrm{RT}}$。

例如, 某文档中包含“有色金属”和“煤矿”, 查询词为“铁矿”, 要计算主题相关性, 则可表达为

铁矿⊕(有色金属∨煤矿) =

max(铁矿⊕有色金属, 铁矿⊕煤矿)。

根据煤矿领域语言词典的本体图，我们设 BT 和 NT 两种关系的权重分别为 0.6 和 0.4，RT 关系设为 1.0，本题中两个概念之间最远加权距离为 5.0，则分别计算相似度为

$$\text{铁矿}\oplus\text{有色金属}=1-\frac{\frac{0.4}{2}+\frac{0.4}{1}+\frac{0.6}{2}}{5.0}=0.82,$$

$$\text{铁矿}\oplus\text{煤矿}=1-\frac{\frac{0.4}{2}+\frac{0.4}{1}+\frac{0.6}{2}+\frac{0.6}{3}}{5.0}=0.78。$$

利用模糊集理论进行结果合并，铁矿⊕(有色金属∨煤矿)=0.82。上述计算实例表明，本文提出的主题相似度度量方法合理有效。

2.2.2　地理相似度

两个信息单元的地理相似度，是从语义的角度判断其地理信息内容之间的相关性。由于信息单元的地理信息是关于原子地理空间命题的语句，因此地理相似度是两个信息单元中所包含的原子地理空间命题的函数。

一般情况下，文档中的地理空间命题都是简单地名的形式，有时可能包含空间关系谓词的简单组合。检索查询请求的情况与此相同。因此，地理相似度的计算可以归结为两个地名之间语义相似度的计算。

地名的语义包括概念特征和位置特征。对于概念相似度，采用以整体–部分关系进行组织的地名库为基础，计算两个地名对应的概念在本体树上的层次距离[18]，即

$$\mathrm{CS}(p,o)=1-\left(\sum_{x\in\{p.\mathrm{PartOf}-o.\mathrm{PartOf}\}}\frac{\alpha}{L_x}+\sum_{y\in\{o.\mathrm{PartOf}-p.\mathrm{PartOf}\}}\frac{\beta}{L_y}+\sum_{z\in\{p,o\}}\frac{\gamma}{L_z}\right),\quad(11)$$

其中，CS(p, o)是地名 p 和 o 的概念相似度，L_x, L_y, L_z 分别代表 x, y, z 三个地名在本体树上的层数，p.partOf 和 o.PartOf 分别代表地名 p 和 o 在本体树上所有祖先地名的集合，α, β, γ 是调和系数，一般令 $\alpha=\beta=1.0$，而当 p 和 o 是兄弟关系时，令 $\gamma=1.0$，否则 $\gamma=0$。

位置相似度的计算基于“拓扑空间关系为主，度量关系精化”[19]的原则。当地名 p 和 o 对应的空间范围尺度较大(如省级及以上级别)时，可以直接采用拓扑关系语义相近度[20]进行评价，无须通过距离进行精化。当空间范围尺度较小(如市级及以下级别)时，采用综合考虑拓扑和度量的计算方法[21]。

第一步，计算拓扑相似度:

$$\mathrm{Inclusion}(p,o)=\begin{cases}\dfrac{\mathrm{NumDescendants}(o)+1}{\mathrm{NumDescendants}(p)+1}, & o\subseteq p,\\ 0, & \text{其他},\end{cases}\quad(12)$$

其中 NumDescendants(o)和 NumDescendants(p)分别为地名 o 和 p 在本体树上子地名的数量。

第二步，计算度量相似度:

$$\mathrm{Proximity}(p,o)=\frac{1}{1+\mathrm{Distance}(p,o)/\mathrm{diagnoal}(p)},\quad(13)$$

其中，Distance(p, o)为地名 p 与 o 之间的欧几里得距离，diagnoal(p)为地名 p 的 MBR 的对角线长度。

第三步，判断是否为兄弟关系。当 p 和 o 具有相同的祖先时，令 Sibling(p, o) = 1；否则令 Sibling(p, o) = 0。

第四步，合并上述 3 个数值，得到位置相似度 LS(p, o):

$$\mathrm{LS}(p,o)=b\times\{\mathrm{Inside}(p,o)+\mathrm{Proximity}(p,o)\}+(1-b)\times\mathrm{Siblings}(p,o),\quad(14)$$

其中，b 是介于 0 与 1 之间的调和参数。

最后，将概念相似度 CS 与位置相似度 LS 合成，得到地理相似度 g(p, o):

$$g(p,o)=w_g\mathrm{LS}(p,o)+w_h\mathrm{CS}(p,o),\quad(15)$$

其中，w_g 和 w_h 是介于 0 和 1 之间的调和系数，一般令 w_g 和 w_h 分别为 0.6 和 0.4[22]。实际上，上述各式中调和系数的取值均与查询及文档地理范围的尺度有关。

2.2.3　主题和地理相似度的组合

将主题和地理的相似度组合形成一个综合的评价结果，则文档 d 的一个信息单元 u_i 与子查询 q_j 的相似度 Ru(u_i, q_j)可以表示为

$$\mathrm{Ru}(u_i,q_j)=l(\mathrm{Ru}_g(u_i,q_j),\mathrm{Ru}_t(u_i,q_j)),\quad(16)$$

其中 Ru_g 为地理相似度，Ru_t 为主题相似度，l 为合成函数。合成函数可以采用几何平均值、算术平均值、乘积、加权算术平均值、欧氏距离等方法。

2.3　文档与子查询的相似度计算

对于文档 d 与子查询 q_j 之间相似度评价 Rq(d, q_j)的计算，可以将其建模为一个不确定性推理的过

程, 合成文档所有信息单元与子查询的相似度。文档中的每个信息单元 u_i 都可以看成是证明文档 d 与 q_j 是否相关的一个证据, 证据之间是独立出现的。这样, u_i 与 q_j 之间的相似程度可以视为证据来证明“文档 d 与子查询 q_j 相关”的可信度。因此, 采用 Dempster-Shafer(D-S)证据理论[23], 合并所有证据的可信度, 最终导出文档与子查询的相似度。

一般地, 对于 GIR 的检索模型, 需要证明的命题有两个: T=文档与查询相关, F=文档与查询不相关。基于 D-S 证据理论, 识别框 $\Theta=\{T, F\}$。Θ 是由互不相容的基本命题组成的完备集合, 表示对某一问题的所有可能答案, 但其中只有一个答案是正确的。Θ 的幂集记做 2^{Θ}, 则有

$$2^{\Theta}=\{\phi,\{T\},\{f\},\{T,F\}\}。\tag{17}$$

对于 2^{Θ} 中给定的命题(或子集) A, 基础信任分配函数 $m(A)$ 表示证据对命题 A 的支持程度, 信任度函数 Bel(A) 表示对命题 A 的信任程度, 似然函数 Pl(A)表示对命题 A 非假的信任程度, 也即对 A 似乎可能成立的不确定性度量。区间[Bel(A), Pl(A)]表示所有提交的证据给出的 A 为真的可信度的波动范围, 且有

$$\mathrm{Bel}(A)=\sum\nolimits_{B|B\subseteq A}m(B),\tag{18}$$

$$\mathrm{Pl}(A)=\sum\nolimits_{B|B\cap A\neq\phi}m(B),\tag{19}$$

其中 m 函数满足

$$\begin{cases}m:2^{\Theta}\to[0,\ 1],\\ 0\le m(x)\le 1,\forall x\in 2^{\Theta},\\ \sum_{x\in 2^{\Theta}}m(x)=1,\\ m(\phi)=0。\end{cases}\tag{20}$$

假设一个信息单元 u_i 与子查询 q_j 之间的相似度为 α, 即 Ru(u_i, q_j)=α (基于 2.2 节的方法计算), 如果把该信息单元视为一个证据, 那么它证明命题集合$\{T\}$成立的可信度为 α。没有证据直接证明$\{F\}$命题组, 那么根据合成理论, 最好的方法是将 $1-\alpha$ 赋给全集$\{T, F\}$, 即

$$\begin{cases}m(\{T\})=\alpha,\\ m(\{F\})=0,\\ m(\{T,F\})=1-\alpha,\\ m(\{\phi\})=0。\end{cases}\tag{21}$$

那么, Rq(d, q_j)可以由 Ru(u_i, q_j)迭代导出。对于一个完整的文档, 依次加入文档中的信息单元作为证据, 计算每个证据的基础信任分配, 然后使用 D-S 的证据合成方法合成多个证据的确信度, 即

$$m_{i+1}(C)=\frac{\sum_{A\cap B=C}m_i(A)m^{i+1}(B)}{\sum_{A\cap B\neq\phi}m_i(A)m^{i+1}(B)},\tag{22}$$

其中, m_i 表示合成了 i 个证据后的 m 函数值, m^{i+1} 表示第 i+1 个证据出现后给出的基础信任分配。如果出现交集为空的元素, 依据 D-S 证据合成算法, 需要扣除交集为空集的计算结果, 然后对每个 m 函数值重新进行归一化。这样, 每个命题集合的 m 函数值迭代更新(表 1)。

表 1 D-S 证据合成算法的计算过程

Table 1 Computation of the evidence combination algorithm

m 函数	$m^{i+1}(\{T\})$	$m^{i+1}(\{T,F\})$
$m_i(\{T\})$	$m_i(\{T\})\cdot m^{i+1}(\{T\})$	$m_i(\{T\})\cdot m^{i+1}(\{T,F\})$
$m_i(\{T,\mathrm{F}\})$	$m_i(\{T,F\})\cdot m^{i+1}(\{T\})$	$m_i(\{T,F\})\cdot m^{i+1}(\{T,F\})$

在所有的证据全部加入后, 得到最终每个命题集合的 m, Bel 和 Pl 函数值。对命题 T 和 F 支持的可信度 $p(T)$和 $p(F)$将分别满足:

$$\begin{cases}\mathrm{Bel}(\{T\})\le p(T)\le\mathrm{Pl}(\{T\}),\\ \mathrm{Bel}(\{F\})\le p(F)\le\mathrm{Pl}(\{F\})。\end{cases}\tag{23}$$

Bel($\{T\}$), Pl($\{T\}$), Bel($\{F\}$)和 Pl($\{F\}$)是用于度量文档与子查询相似度的基本指标。可以直接使用 Bel($\{T\}$)值作为相似度, 或更进一步地, 将相似度定义为

$$\mathrm{Rq}(d,q_j)=\frac{\mathrm{Bel}(\{T\})}{\mathrm{Bel}(\{F\})}\quad 或$$

$$\mathrm{Rq}(d,q_j)=1-\frac{1}{\ln\left(\mathrm{e}+\dfrac{\mathrm{Bel}(\{T\})}{\mathrm{Bel}(\{F\})}\right)}。\tag{24}$$

2.4 文档与查询的相似度计算

将文档与所有子查询的相似度进行合成, 得到最终文档与查询的相似度 $R(d, q)$。查询和子查询都是基于命题逻辑表达, 查询根据其内部的逻辑结构, 分解为若干个形如<theme><spatial relationship><location>的原子的子查询, 且子查询之间通过逻辑连接符{∧, ∨, ¬}相连。

由于文档 d 与子查询 q_j 之间的相似度为 Rq(d, q_j), 则有

$$\mathrm{Rq}(d,\ \neg q_j)=1-\mathrm{Rq}(d,\ q_j)。\tag{25}$$

设查询 q 由两个子查询$\{q_1,\ q_2\}$构成, 子查询通过的逻辑连接符$\uplus$连接, 可以为$\vee$或$\wedge$, 那么文档 d 与 q 查询的相似度 $R(d,\ q)$计算公式为

$$R(d,q)=\begin{cases}\max\{\mathrm{Rq}(d,q_1),\mathrm{Rq}(d,q_2)\},\ \uplus=\vee,\\ \min\{\mathrm{Rq}(d,q_1),\mathrm{Rq}(d,q_2)\},\ \uplus=\wedge。\end{cases}\tag{26}$$

任意复杂的查询都可以先进行转换, 然后按照上述方法求解与文档的相似度。通过上述工作, 最终得到文档 d 与查询 q 的相似度 $R(d,\ q)$, 据此可排序候选文档。

3　实验

实验文档集取自中国矿业网[①], 所有文档都用中文自然语言编写, 内容限定为地质矿产专题领域。在限定领域的条件下, 一个有限且有效的领域本体相对容易制定, 在检索过程中, 该领域概念本体将作为专题知识库使用。信息单元的提取采取人工和机器相结合的方式进行, 先通过人工方式将文档内容分割成内容独立的若干段落, 再用中文分词工具对各个段落进行分词处理, 并提取相关的主题词信息和地名信息。通过上述处理, 整理汇编其中 50 个文档作为测试, 平均每个文档包含 2~3 个信息单元, 每个信息单元包含 1 个主题词和 1 个表示地理位置的地名。

例如, 对于内容为"……我国铬铁矿储量分布于 13 个省(区), 但主要集中在西藏、内蒙古、新疆、甘肃。……我国铝土矿分布于 20 个省(区), 但主要集中在山西、贵州、河南和广西……"的文档, 可以表达为 2 个信息单元, 即

d={({铬铁矿}, {西藏, 内蒙古, 新疆, 甘肃}), ({铝土矿}, {山西, 贵州, 河南, 广西})}。

针对文档涉及的地质矿产领域, 建立一个简化的地质矿产资源类型本体, 如图 1 所示。该领域本体中包含 1 个一级类别, 3 个二级类别, 73 个三级类别, 47 个四级类别, 17 个五级类别。

实验采用的地名库是覆盖县级行政区划的全国行政区地名库, 其中包含省级行政区划 34 个, 地级市地名 360 个, 县级地名 2939 个, 乡镇级地名 20266 个。

实验中采用传统的地理信息检索方法作为与定性检索方法的比较。该对比方法使用基于向量模型计算文档与查询的主题相似度, 采用外包矩形相交比率作为地理相似度, 对两者采用几何平均值方法合成。

在该文档集上, 分别采用定性方法和传统方法

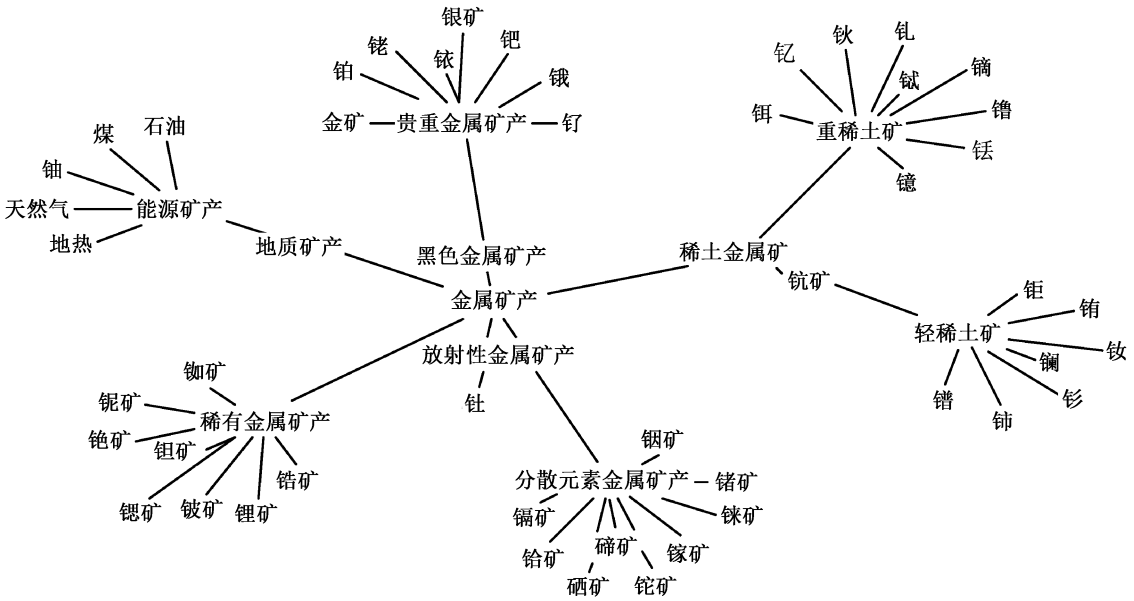

图 1　实验中采用的地质矿产领域本体

Fig. 1　Mineral ontology used in the experiment

① http://www.chinamining.com.cn

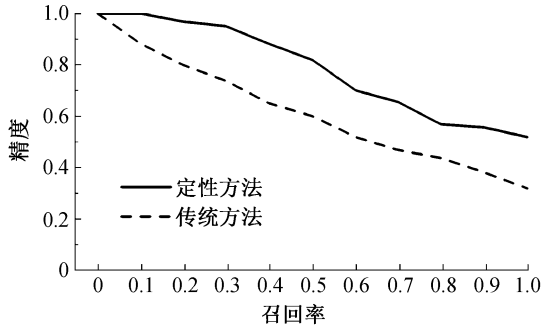

图 2 两种检索方法的 11 点召回率精度对比评估
Fig. 2 Comparison of 11-points recall rates of two GIR methods

进行多次查询，查询过程中采用通用的<theme><location>形式(这里基于简化的形式省略了<spatial relationship>部分)，例如“贵重金属矿产 河北省”、“金属矿产 石家庄市”等。每个查询的相关文档集采用人工方法判定。两种方法的查询结果采用 11 点标准召回率平均精度进行评估，如图 2 所示。实验表明，定性方法表现出更好的检索精度，并可以度量文档中比较细微的相似度差别，较少出现定量方法那样完全不匹配的情况。

在一般情况下，该定性方法可以取得比传统方法更好的结果，特别是命题表达和定性推理规则更适用于处理中大空间尺度的查询。对于小尺度的查询，例如街区尺度，传统的定量方法则可以给出更精确的结果。因此，更好的策略是在一个地理信息检索系统中同时集成定性和定量的方法。当查询的空间尺度较大(城市级别或更大)时，采用定性方法更符合常识性空间认知，检索处理也更容易、更快捷。当查询的空间尺度适中(区县级别)时，组合使用定性和定量的方法，首先使用定性方法进行空间范围的粗过滤，然后使用定量方法精确处理。当查询的空间尺度很小(街区级别)时，可以直接使用定量方法得到精确结果。

4 结论

本文基于命题逻辑和不确定性推理，提出一种地理信息检索的定性模型，为提高现有地理信息检索技术的检索效率提供一种新方法。相比于传统的地理信息检索方法，该定性方法顾及了语义信息，没有对空间进行简化，而是直接采用地理本体或地名库，因而可以更加客观真实地刻画文档中的信息和用户的检索请求。所提方法中基于真实信息内容的推理匹配方法，在大中尺度下得到的检索结果比传统定量方法更接近于使用者的常识性认知。并且，该定性地理信息检索方法也可以同时支持传统的定量方法，可以实现定量和定性检索模型的良好结合。

该定性检索方法仍需进一步完善，还需要深入研究文本分割技术，实现文档中信息单元的正确高效提取，改善定性推理评价模型，并完善地理知识库建设。

参考文献

[1] Larson R R. Geographic information retrieval and spatial browsing // Smith L, Gluck M. Proceedings of the data processing clinic — geographic information systems and libraries: patrons, maps, and spatial information. Urbana-Champaign: University of Illinois, 1996: 81–124

[2] Jones C B, Purves R S. Foreword of GIR’06 // Workship on Geographic Information Retrieval, SIGIR’06. New York, 2006: 40–41

[3] Jones C B, Purves R S. Geographical information retrieval. International Journal of Geographical Information Science, 2008, 22(3): 219–228

[4] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: ACM Press, 1999

[5] Hill L L, Frew J, Zheng Q. Geographic names: the implementation of a gazetteer in a georeferenced digital library. D-Lib Magazine, 1999, 5(1): 1–19

[6] Alani H, Jones C B, Tudhope D. Voronoi-based region approximation for geographical information retrieval with gazetteers. International Journal of Geographical Information Science, 2001, 15(4): 287–306

[7] Frontiera P, Larson R, Radke J. A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographical Information Science, 2008, 22(3): 337–360

[8] Naaman M, Song Y J, Paepcke A, et al. Assigning textual names to sets of geographic coordinates. Computers, Environment and Urban Systems, 2006, 30: 418–435

[9] Liu Y, Yuan Y, Xiao D, et al. A point-set-based approximation for areal objects: a case study of representing localities. Computers, Environment and Urban Systems, 2010, 34(1): 28–39

[10] Beard K, Sharma V. Multidimensional ranking for

data in digital spatial libraries. International Journal on Digital Libraries, 1997, 1(2): 153–160

[11] Larson R R, Frontiera P. Spatial ranking methods for geographic information retrieval (GIR) in digital libraries // Heery R, Lyon L. Lecture notes in computer science 3232. Berlin: Springer, 2004: 45–57

[12] de Sabbata S, Reichenbacher T. A probabilistic model of geographic relevance // Proceedings of the 6th Workshop on Geographic Information Retrieval. Zürich: ACM, 2010: 23–24

[13] 金鑫, 耿海燕, 高勇, 等. 空间方位关系在不同场景下的模糊性探讨. 北京大学学报: 自然科学版, 2009, 45(6): 1025–1032

[14] Cohn A G, Renz J. Qualitative spatial representation and reasoning. Foundations of Artificial Intelligence, 2008, 3: 551–596

[15] 刘瑜, 龚咏喜, 张晶, 等. 地理空间中的空间关系表达和推理. 地理与地理信息科学, 2007, 23(5): 1–7

[16] Zhang Yi, Gao Yong, Xue Lulu, et al. A common sense geographic knowledge base for GIR. Science in China Series E: Technological Sciences, 2008, 51(Suppl 1): 26–37

[17] Tudhope D, Taylor C. Navigation via similarity: automatic linking based on semantic closeness. Information Processing and Management, 1997, 33(2): 233–242

[18] Jones C B, Alani H, Tudhope D. Geographical information retrieval with ontologies of place. Lecture Notes in Computer Science, 2001, 22(3): 322–335

[19] Mark D M. Spatial representation: a cognitive view // Maguire D J, Goodchild M F, Rhind D W, et al. Geographical information systems: principles and applications. New York: John Wiley & Sons, 1999: 81–89

[20] Bruns T, Egenhofer M. Similarity of spatial scenes // Kraat M J, Molenaar M. Seventh Internaltional Symposium on Spatial Data Handling (SDH'96). Delft, 1996(4A): 31–42

[21] Andrade L, Silva M J. Relevance ranking for geographic IR // Workshop on Geographic Information Retrieval, SIGIR'06. Seattle, WA: ACM, 2006: 1–4

[22] Jones C B, Alani H, Tudhope D. Geographical information retrieval with ontologies of place // Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science (COSIT). Morro Bay, CA: Springer Berlin/Heidelberg, 2001: 322–335

[23] Shafer G. A mathematical theory of evidence. Princeton, NJ: Princeton University Press, 1976