地图检索意图识别计算示例

1.示例数据介绍

本示例涉及空间范围、主题、制图方法与地理要素 4 个维度，各维度本体及概念取值数量如表 1 所示。示例样本集合的预定义意图为{{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}，{空间范围: 北美洲, 主题: 地质, 制图方法: 范围法, 地理要素: 岩石}}，如图 1 所示。相关反馈样本集合如表 2 所示，其中样本编号 1-10 为正样本、11-20 为负样本，其中 14 为反馈噪声，用于模拟用户相关反馈中的漏选情况。

表 1 各意图维度本体

维度	空间范围	主题	制图方法	地理要素
本体库	GeoNames（局部）	自定义	自定义	SWEET
版本	2020.02.14	--	--	3.5.0
3	635	11	8	6699

表 2 相关反馈样本示例

正样本					负样本				
编号	空间范围	主题	制图方法	地理要素	编号	空间范围	主题	制图方法	地理要素
1	加拿大	地质	范围法	岩石	11	巴西	地质	范围法	沉积岩
2	美国	地质	范围法	变质岩	12	智利	地质	范围法	混合岩
3	北美洲	地质	范围法	沉积岩	13	缅因州	水	分级统计图	隔水层
4	内华达州	地质	范围法	斜长石	14	美国	地质	范围法	沉积岩
5	墨西哥	地质	范围法	橄榄岩	15	佛罗里达州	水	范围法	防洪堤
6	加利福尼亚州	水，地质	点状符号法，范围法	泉水，沉积岩	16	印第安纳州	生物多样性	分级统计图	鸟类
7	犹他州	水	质底法	湖泊	17	南美洲	水	分级统计图	降水
8	内华达州	水	范围法	湿地	18	加拿大	水	线状符号法	河流
9	佛罗里达州	水	范围法	水库	19	北美洲	地质	范围法	矿物
10	美国	水	线状符号法	河流	20	巴西	地质	分级统计图	土壤

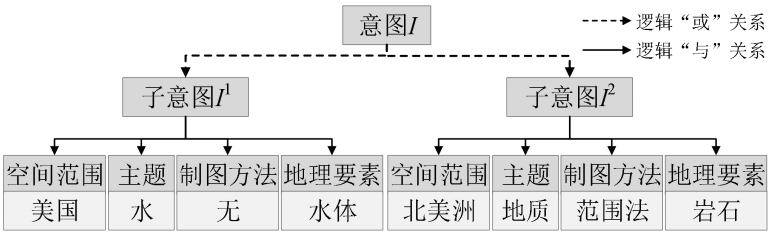

图 1 示例样本集合预定义意图形式化表达

由上述示例数据可得本文方法部分变量取值为：样本集合$S = S^+ \cup S^-$，$S^+ = \{1, 2, \ldots, 10\}$，$S^- = \{11, 12, \ldots, 20\}$；维度数量$d = 4$，维度$D_1$到$D_4$分别为空间范围、主题、制图方法与地理要素，对应的取值集合大小$|C_1|$至$|C_4|$分别为 635、11、8 与 6699。

2.意图识别计算过程

1）参数设置：随机合并数量为 50，子意图覆盖的正样本比例阈值为 0.3。

2）计算样本增强系数α：

①根据正整数通用编码计算公式，$L_N(|S^+| + 1) = log(2.865064) + log(10 + 1) + log(log(10 + 1)) + log(log(log(10 + 1))) \approx 7.6089$ bits；

②依据公式 4 可得不基于意图编码的情况下反馈样本总编码长度$L_{avg}(S|\emptyset) = -(|S^+|/|S|) * log(|S^+|/|S|) - (|S^-|/|S|) * log(|S^-|/|S|) = -0.5 * log(0.5) - 0.5 * log(0.5) = 1$ bit；

③取ε=1，依据公式 5 可得样本增强系数$\alpha = (L_N(|S^+|+1) + |S^+|log\left(\prod_{i=1}^{d}|C_i|\right))/|S|L_{avg}(S|\emptyset) + 1 = (7.6089 + 10 * log(635 * 11 * 8 * 6699))/(20 * 1) + 1 \approx 15.6203$。

3）初始化意图（即子意图集合）$R = \emptyset$，剩余正样本集合$S_r^+ = \{1, 2, \ldots, 10\}$，剩余负样本集合$S_r^- = \{11, 12, \ldots, 20\}$，$R$对应样本总编码长度$L(S, R) = L(R) + \alpha|S|L_{avg}(S|\emptyset) = log(2.865064) + log(1) + \alpha|S|L_{avg}(S|\emptyset) = 1.5186 + 15.6203 * 20 * 1 \approx 313.9246$ bits。

4）第 1 次迭代搜索

①样本随机合并。从S_r^+与R的并集中随机选取两个元素，根据 2.2 节合并生成候选子意图。假设被选取元素为样本 6 与样本 10，合并时根据两样本的标签集合及各标签间的从属关系生成候选子意图在各维度的取值。具体合并过程如下：

a. 分别计算样本各维度标签对应概念的信息量。以地理要素维度为例，样本 6 与样本 10 包含标签“泉水”、“沉积岩”与“河流”，其中“泉水”对应概念（记为c）的信息量计算如下：

$$IC(c) = f_{depth}(c) * (1 - f_{leaves}(c)) + f_{hypernyms}(c)$$
$$= \frac{log(depth(c))}{log(max\_depth)} * (1 - \frac{log(leaves(c) + 1)}{log(max\_leaves + 1)}) + \frac{log(hyper(c) + 1)}{log(max\_nodes)}$$
$$= \frac{log(6)}{log(14)} * (1 - \frac{log(0\ +\ 1)}{log(5446\ +\ 1)})\ +\ \frac{log(5\ +\ 1)}{log(6699)} \approx 0.8823$$

式中$depth(c)$为概念c在 SWEET 本体中的深度，值为 6；max_depth为 SWEET 本体的最大深度，值为 14；$leaves(c)$为从属于概念c的叶子概念数量，值为 0；max_leaves为所有叶子概念数量，值为 5446；$hyper(c)$为概念c的所有上位概念数量，值为 5；max_nodes为 SWEET 本体概念总数，值为 6699。按照上式，“沉积岩”与“河流”对应概念的信息量分别为 0.8451 与 0.8823。

b. 计算各维度可合并标签对的语义相似度。以地理要素维度为例，可合并的标签对包括（泉水，河流）与（沉积岩，河流）。依据公式 6，标签对（泉水，河流）的语义相似度为$Sim(\text{泉水}, \text{河流}) = 2IC(LCA(\text{泉水}, \text{河流}))/(IC(\text{泉水}) + IC(\text{河流})) = 2 * 0.5644/(0.8823 + 0.8823) \approx 0.6397$，其中$LCA(\text{泉水}, \text{河流})$为“泉水”与“河流”的最低公共祖先概念“水体”（如图 2 所示），信息量为 0.5644。标签对（沉积岩，河流）的最低公共祖先概念为“Thing”，信息量为 0，故语义相似度为$Sim(\text{泉水}, \text{河流}) = 0$。

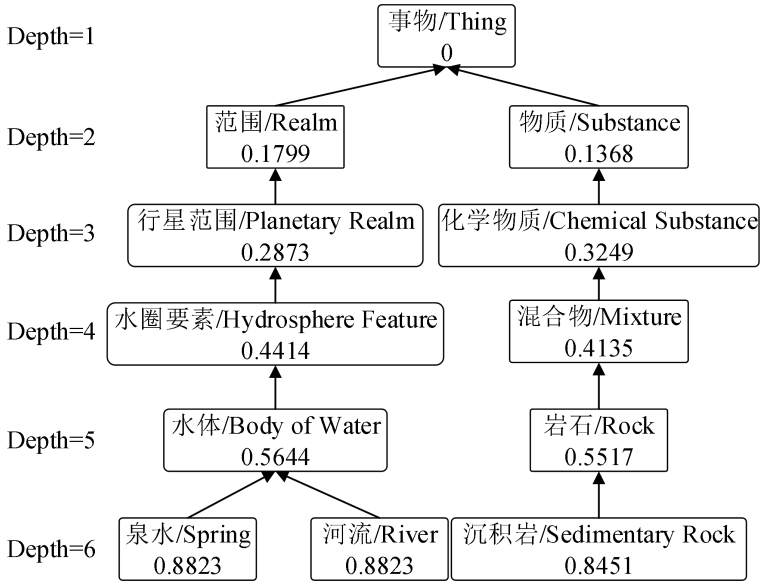

图 2 样本 6 与 8 地理要素维度标签对应概念及其上位概念的信息量

c. 使用具有最大语义相似度标签对的最低公共祖先概念作为候选子意图在各维度的取值，得到样本合并结果。各维度标签对相似度如表 3 所示，故样本 6 与样本 10 合并结果为{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}。

表 3 样本合并示例（样本 6 与 10）

	空间范围	主题	制图方法	地理要素
所有标签对语义相似度	Sim(内华达, 美国) ≈ 0.6	Sim(地质, 水) = 0 Sim(水, 水) = 1	Sim(线状符号法, 点状符号法) = 0 Sim(范围法, 点状符号法) = 0	Sim(泉水, 河流) ≈ 0.6397 Sim(沉积岩, 河流) = 0
最相似标签对	(内华达, 美国)	(水, 水)	(范围法, 点状符号法)	(泉水, 河流)
子意图取值	美国	水	无	水体

②将上步所得候选子意图与R合并得到候选意图$candR$ = {{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}}，并计算$candR$对应的样本总编码长度$L(S, candR)$。$candR$仅包含一个子意图I^1，即$m = 1$。该子意图覆盖的正负样本集合为$S_1 = S_1^+ \cup S_1^-$，$S_1^+ = \{6, 7, 8, 9, 10\}$，$S_1^- = \emptyset$。剩余正负样本集合分别为$S_r^+ = \{1, 2, 3, 4, 5\}$，$S_r^- = \{11, 12, 13, \ldots, 20\}$。$candR$对应总编码长度计算过程为：

a. 依据公式 2，意图编码长度 $L(candR) = L_N(m + 1) + log(\prod_{i=1}^{d} |C_i|) = log(2.865064) + log(2) + log(log(2)) + log(635 * 11 * 8 * 6699) \approx 30.9983$ bits；

b. 依据公式 4，给定意图$candR$后，S_1与S_r中样本平均编码长度分别为$L_{avg}(S_1|I^1) = -(|S_1^+|/|S_1|) * log(|S_1^+|/|S_1|) - (|S_1^-|/|S_1|) * log(|S_1^-|/|S_1|) = -(5/5) * log(5/5) - (0/5) * log(0/5) = 0$，$L_{avg}(S_r|candR) = (-(5/15) * log(5/15) - (10/15) * log(10/15)) \approx 0.9183$ bits；

c. 由于单子意图不存在样本重复覆盖问题，故无需去除重复样本；

d. 依据公式 3，给定意图后的样本编码长度 $L(S|candR) = \alpha|S_1|L_{avg}(S_1|I^1) + log(\alpha|S_1|) + \alpha|S_r|L_{avg}(S_r|candR) = 0 + log(15.6203 * 5) + 15.6203 * 15 * 0.9183 = 221.4491$ bits；

e. 依据修正后的公式 1，总编码长度 $L(S, candR) = L(candR) + L(S|candR) = 30.9983 + 221.4491 = 252.4474$ bits。

③将步骤①与步骤②按照参数随机合并数量重复 50 次，得到 50 个候选意图及其对应样本总编码长度。去除重复项后，样本总编码长度最短的 3 个候选意图如表 4 所示。由于总编码长度最小值$L_{min} = 252.4474$ bits $< L = 313.9246$ bits，因此更新意图R = {{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}}，$L = L_{min} = 252.4474$，$S_r^+ = \{1, 2, 3, 4, 5\}$，$S_r^- = \{11, 12, 13, \ldots, 20\}$，并继续迭代搜索。

表 4 第 1 次迭代候选意图及对应样本总编码长度（升序排序，前 3）

候选意图序号	空间范围	主题	制图方法	地理要素	总编码长度/bits
1	美国	水	无	水体	252.4474
2	北美洲	无	无	无	277.4919
3	北美洲	地质	范围法	岩石	283.2971

5）第 2 次迭代搜索

参照第一次迭代进行 50 次样本随机合并。假设某次合并过程中被选取的元素为样本 1 与样本 2，则生成的候选意图为$candR = \{I^1, I^2\}$，I^1 = {空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}，I^2 = {空间范围: 北美洲, 主题: 地质, 制图方法: 范围法, 地理要素: 岩石}，$m = 2$。各子意图的样本覆盖情况及剩余样本集合为：$S_1^+ = \{6, 7, 8, 9, 10\}$，$S_1^- = \emptyset$；$S_2^+ = \{1, 2, 3, 4, 5, 6\}$，$S_2^- = \{14\}$；$S_r^+ = \emptyset$，$S_r^- = \{11, 12, 13, 15, \ldots, 20\}$。$candR$对应总编码长度计算过程为：

a. 依据公式 2，意图编码长度 $L(candR) = L_N(m + 1) + log(\prod_{i=1}^{d} |C_i|) = log(2.865064) + log(3) + log(log(3)) + log(635 * 11 * 8 * 6699) * 2 = 60.7275$ bits；

b. 依据公式 4，给定意图$candR$后，S_1、S_2与S_r中样本的平均编码长度分别为$L_{avg}(S_1|I^1)=0$ bits，$L_{avg}(S_2|I^2)\approx 0.5917$ bits，$L_{avg}(S_r|candR)=0$ bits；

c. 由于I^1与I^2同时覆盖样本 6，故需进行函数修正以消除重复覆盖。由于$L_{avg}(S_1|I^1)<L_{avg}(S_2|I^2)$，则将样本 6 分配给$I^1$，修正后$S_1$不变，$S_2^+=S_2^+-S_{2-exclude}=\{1,2,3,4,5,6\}-\{6\}=\{1,2,3,4,5\}$，$S_2^-=\{14\}$；

d. 依据公式 3，给定意图后的样本编码长度$L(S|candR)=\alpha|S_1|L_{avg}(S_1|I^1)+log(\alpha|S_1|)+\alpha|S_2|L_{avg}(S_2|I^2)+log(\alpha|S_2|)+\alpha|S_r|L_{avg}(S_r|candR)=0+log(15.6203*5)+15.6203*6*0.5917+log(15.6203*6)+0\approx 68.2928$ bits；

e. 依据修正后的公式1，总编码长度$L(S,candR)=60.7275+68.2928=129.0203$ bits。

去除重复项后，样本总编码长度最短的 3 个候选意图如表 5 所示。由于总编码长度最小值$L_{min}=129.0203\text{ bits}<L=252.4474\text{ bits}$，因此更新意图$R=$ {{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}, {空间范围: 北美洲, 主题: 地质, 制图方法: 范围法, 地理要素: 岩石}}，$L=L_{min}=129.2427$，$S_r^+=\emptyset$，$S_r^-=\{11,12,13,15,\ldots,20\}$，并继续迭代搜索。

表 5 第 2 次迭代候选意图及对应样本总编码长度（升序排序，前 3）

候选	子意图 1				子意图 2				总编码
意图序号	空间范围	主题	制图方法	地理要素	空间范围	主题	制图方法	地理要素	长度/bits
1	美国	水	无	水体	北美洲	地质	范围法	岩石	129.0203
2	美国	水	无	水体	北美洲	地质	范围法	火成岩	230.2383
3	美国	水	无	水体	美国	地质	范围法	火成岩	263.0667

6）第 3 次迭代搜索

由于$S_r^+=\emptyset$，样本随机合并过程中选取的两个元素只能为R中的两个子意图，合并后可生成候选子意图$subI=$ {空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}。由于$subI$覆盖R中两个子意图，故将$subI$加入R时需去除R中已有子意图，得到候选意图$candR=$ {{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}}。又由于此时$L(S,candR)\approx 277.5009\text{ bits}>129.0203\text{ bits}$，故停止迭代搜索。

7）错误子意图过滤

R中两个子意图覆盖的正样本比例分别为$C^1=|S_1^+|/|S^+|=5/10=0.5$，$C^2=|S_2^+|/|S^+|=5/10=0.5$，均大于参数正样本覆盖占比阈值 0.3，因此均被保留，即最优意图为$R=$ {{空间范围: 美国, 主题: 水, 制图方法: 无, 地理要素: 水体}，{空间范围: 北美洲, 主题: 地质, 制图方法: 范围法, 地理要素: 岩石}}。