

# 非结构化文本数据的 GIS 描述性查询方法

蒲海霞\*, 李佳田, 李 锐, 何育枫, 王 华

(昆明理工大学 国土资源工程学院, 昆明 650093)

(\*通信作者电子邮箱 pugongying928@126.com)

**摘 要:**针对传统地理信息系统(GIS)结构化或半结构化属性查询方法对查询语句输入的精度及查询范围的限制,提出了以哈尔滨工业大学《同义词词林》扩展版本文本相关度计算为核心的非结构化文本数据 GIS 描述性查询方法。基本过程是根据描述性查询语句计算其与地理要素所关联的文本的相关度,进而以相关度值得出概括性查询结果。对比实验结果表明,描述性查询方法不但支持查询语句输入的多样化,而且能够有效地得出与输入的描述性查询相关联的地理要素。

**关键词:**同义词词林;描述性查询;非结构化数据;地理信息系统

**中图分类号:**TP311.13;TP3-05 **文献标志码:**A

## Descriptive query method based on unstructured text data in GIS

PU Hai-xia\*, LI Jia-tian, LI Rui, HE Yu-feng, WANG Hua

(Faculty of Land Resource Engineering, Kunming University of Science and Technology, Kunming Yunnan 650093, China)

**Abstract:** Since the traditional Geographic Information System (GIS)'s structured or semi-structured attribute query poses sort of limitation on input accuracy and scope of query sentences, a GIS descriptive query method for text-related non-structured text data was suggested based on the expanded version of a dictionary of English synonyms, TongYiCi CiLin. compiled by Harbin Institute of Technology. Basic process is to calculate correlation between descriptive query sentence and text connected to the geographic element, then getting general query results according to it. The comparison experiment shows that descriptive query method not only supports diversity of input query sentence, but also effectively gets geographic element related to input descriptive query.

**Key words:** TongYiCi CiLin; descriptive query; unstructured data; Geographic Information System (GIS)

## 0 引言

地理信息系统(Geographic Information System, GIS)查询方法在信息查询与检索等领域有着广泛的应用,已有的 GIS 查询功能包括空间查询和属性查询,其空间要素的属性数据呈结构化形式,表现为 XML 存储<sup>[1-2]</sup>,或存放在数据库表结构中<sup>[3]</sup>。空间要素的属性数据结构化形式约束了 GIS 属性查询方式为基于字符串的精确或模糊匹配,其对查询输入字段的精度要求比较高,且查询的范围相对狭小,仅限于查询结构化属性字段所涉及的内容,而富含广泛信息的数据以非结构化文本的形式存在,对文本数据的忽略限制了 GIS 查询范围,不能满足查询多样化的需求。

本文提出了一种针对非结构化文本数据的 GIS 描述性查询方法。以非结构化文本数据作为各要素点的属性数据,以描述性查询语句为输入,通过哈尔滨工业大学《同义词词林》扩展版(以下简称《词林》)文本相关度计算方法计算查询语句与要素所关联的文本之间的相关度,根据相关度值过渡到各个要素点,进而得到概括性查询结果。描述性查询的最主要特点是:利用了存在大量信息的非结构化文本数据,并且在应用上对查询的输入无要求,大幅度增强了 GIS 查询的广泛

度。

## 1 GIS 描述性查询方法

GIS 描述性查询方法是以描述性查询语句为输入,以要素点的大小来显示概括性的查询结果。该方法重点需计算描述性语句与要素所关联的文本的相关度,根据相关度值得到查询结果,相关度是描述性查询语句与文本语义上的相关程度。

由于描述性查询语句多样化,本文的相关度计算包括以下几种关系:词语与文本、句子与文本、段落与文本、文本与文本(文本由多个段落组成)。但其核心是词语与词语、句子与句子之间的相关度计算。

词语相关度以《词林》为词源<sup>[4-5]</sup>展开计算,《词林》按照树状的层次结构把所有收录的词条组织到一起,自上到下词义刻画越来越明显。并且对词语进行了编码,编码方式举例如表 1 所示。大类用一个大写字母表示,中类用一个小写字母表示,小类用十进制两位数表示,词群用一个大写字母表示,原子词群用十进制两位数表示。第 8 位的标记有 3 种,分别是“=”、“#”、“@”,末尾的“=”代表相等同义;“#”代表不等同类,属于相关词语;末尾的“@”代表自我封闭、独立,没

**收稿日期:**2012-03-14;**修回日期:**2012-05-18。 **基金项目:**国家自然科学基金资助项目(40901197, 41161061);云南省自然科学基金资助项目(2008D032M);云南省教育厅重点基金资助项目(2001Z006)。

**作者简介:**蒲海霞(1987-),女,甘肃天水人,硕士研究生,主要研究方向:空间性文本的建模与计算; 李佳田(1975-),男,黑龙江佳木斯人,副教授,主要研究方向:Voronoi 模型与方法; 李锐(1987-),女,云南曲靖人,硕士研究生,主要研究方向:地标提取中的 Voronoi 方法; 何育枫(1988-),男,江西新余人,硕士研究生,主要研究方向:数字图像处理、ArcGIS; 王华(1988-),女,陕西西安人,硕士研究生,主要研究方向:Voronoi 邻近查询方法。

有与其相同或相关的词语。例如:Cb02A01 = 东南西北 四方,Ba01A03@ 万物,Ba01B10# 导体 半导体 超导体。

表1 词语编码表

编码位	符号举例	符号性质	级别
1	C	大类	第1级
2	b	中类	第2级
3	0	小类	第3级
4	2	词群	第4级
5	A	词群	第4级
6	0	原子词群	第5级
7	1	原子词群	第5级
8	= \# \@		

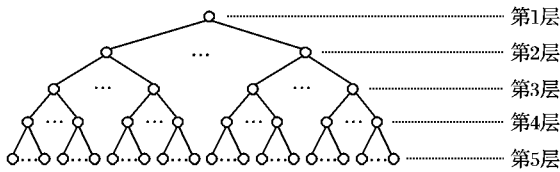


图1 词语树结构组织形式

### 1.1 词语语义相关度计算

- 1) 如果  $a = b$ , 则词语  $a, b$  的相关度  $\text{sim}(a, b) = 1$ 。
- 2) 如果词语  $a \neq b$  且  $a, b$  的前7位编码相同, 则当第8位都为“=”时,  $\text{sim}(a, b) = 0.95$ ; 当第8位都为“#”时,  $\text{sim}(a, b) = 0.85$ 。
- 3) 如果  $a \neq b$  且  $a, b$  的第1位编码不同, 则  $a, b$  不在同一棵树结构上, 其相关度最小,  $\text{sim}(a, b) = 0.1$ 。
- 4) 如果  $a, b$  及编码不满足上述情况, 根据词语编码可以定位  $a, b$  在树中出现的情况 (如图2所示)。

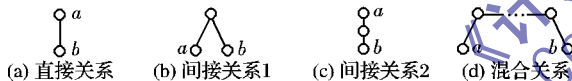


图2 词语之间的关系

图2(a)中词语  $a, b$  为父子关系, 词语  $b$  是对  $a$  进一步的释义, 相关度表现为纵向直接相关。图2(b)~(c)中词语  $a, b$  与第三方是直接关系: 图(b)中相关度表现为横向相关, 词语  $a$  和  $b$  的综合释义对第三方有较大的贡献; 图(c)中相关度表现为间接纵向相关, 词语  $b$  是对  $a$  的二次解释。图2(d)可通过多次间接关系和多次直接关系得出两个词语之间的联系, 其相关度表现为纵横相关。

①词语  $a, b$  为图2(a)或图2(c)中的情况, 以其中一个词语为源头, 派生或进一步对源词进行解释说明, 用树结构中词语之间的距离来模拟相关度值:

$$\text{sim}(a, b) = \text{sim}_v(a, b) = \frac{\alpha \times (d_a + d_b)}{(dis_v(a, b) + \alpha) \times \max(|d_a - d_b|, 1)} \quad (1)$$

其中:  $d_a, d_b$  分别为词语  $a, b$  在树中的深度,  $dis_v(a, b)$  为  $a, b$  的纵向连通距离,  $\alpha$  为可调节参数, 使得  $\text{sim}(a, b) \leq 1$ ,  $\text{sim}_v(a, b)$  为纵向相关度。

② $a, b$  为图1(b)中的情况, 词语  $a, b$  在同一层, 第三方蕴含很多不同的语义, 语义之间的相关度大小用词语所在行的节点顺序值来确定, 即语义分歧度量:

$$\text{sim}(a, b) = \text{sim}_h(a, b) = \frac{\beta \times n \times d}{\max(\frac{n_a + n_b}{2}, \frac{n}{2})} \quad (2)$$

其中:  $\text{sim}_h(a, b)$  表示横向相关度,  $\beta$  为可调节参数,  $n_a, n_b$  为  $a, b$  所在行的节点顺序值,  $n$  是  $a, b$  所在行的节点数。

③ $a, b$  为图1(d)中的情况, 词语  $a, b$  可以同行, 也可以不同行:

若  $a, b$  为混合关系, 则  $a, b$  可以回溯到共同的根节点  $w_k$ ,  $k$  为  $w_k$  所在的深度,  $w_k$  经过多次直接关系和多次间接关系得到  $a, b$ , 设  $a$  的深度为  $d_a$ ,  $b$  的深度为  $d_b$ , 从  $a$  回溯到  $w_k$  经过的节点集合为  $p_a = \{a_k, \dots, a_{da}\}$ ; 从  $b$  回溯到  $w_k$  经过的节点集合为  $p_b = \{b_k, \dots, b_{db}\}$ , 集合  $p_a$  和  $p_b$  是对根节点  $w_k$  的多次展开, 在一定程度上  $a, b$  既不是派生关系, 也不具备横向相关关系,  $a, b$  之间隐含有介于派生和横向相关之间的纵横关系, 即夹角关系, 根节点与集合中的节点构成了相关度权重向量  $S_a, S_b$ :

$$S_a = \left( \frac{n_{a_k}}{n_k} \text{sim}_v(w_k, a_k), \dots, \frac{n_{a_{da}}}{n_{da}} \text{sim}_v(w_k, a_{da}) \right) \quad (3)$$

$$S_b = \left( \frac{n_{b_k}}{n_k} \text{sim}_v(w_k, b_k), \dots, \frac{n_{b_{db}}}{n_{db}} \text{sim}_v(w_k, b_{db}) \right) \quad (4)$$

其中:  $\text{sim}_v(w_k, a_k)$  表示  $w_k, a_k$  的纵向相关度,  $n_i$  表示第  $i$  层的节点总数,  $n_{a_i}$  表示  $a_i$  从左到右的节点次序。最后使用两个向量的余弦来计算  $\text{sim}(a, b)$ 。若两个权重向量  $S_a, S_b$  越接近<sup>[6-8]</sup>, 余弦值越大, 词  $a, b$  的相关度越大:

$$\text{sim}(a, b) = \cos(S_a, S_b) = \frac{S_a \cdot S_b}{|S_a| \cdot |S_b|} \quad (5)$$

### 1.2 句子语义相关度计算

计算句子  $l_1$  和  $l_2$  的相关度, 为了提高计算结果的准确性, 故把短句定为目标句, 长句定为比较句, 短句中的词语为目标词。本文除了计算句子分词后词语之间的相关度外, 还考虑了目标词左右词语在比较句中的偏移量。词语的相互衔接组成句子, 偏移量体现比较句中的词语相对于目标句中词语的衔接程度, 即偏移量越小, 衔接程度越大, 句子相关度越大。

对句子  $l_1$  和  $l_2$  进行分词, 并且进行词性标注且去除虚词。词性标注为: 名称(N), 动词(V), 形容词(A), 数词(M), 量词(Q), 代词(R)。对  $l_1$  和  $l_2$  进行相应的操作后变为  $T_1$  和  $T_2$ ,  $T_1(l_1)$  和  $T_2(l_2)$  为分词及去除虚词后句子中词语个数:

$$T_1 = (N(\dots w_{1i} \dots), V(\dots w_{1j} \dots), A(\dots w_{1k} \dots), M(\dots w_{1l} \dots), Q(\dots w_{1e} \dots), R(\dots w_{1q} \dots)) \quad (6)$$

$$T_2 = (N(\dots w_{2i} \dots), V(\dots w_{2j} \dots), A(\dots w_{2k} \dots), M(\dots w_{2l} \dots), Q(\dots w_{2e} \dots), R(\dots w_{2q} \dots)) \quad (7)$$

若  $T_1(l_1) < T_2(l_2)$ , 则  $l_1$  是目标句,  $l_2$  是比较句, 先取出  $w_{11}$  及其词性  $X$ , 得:

$$\text{sim}(w_{11}, w_{2a}) = \max(\text{sim}(w_{11}, X(\dots w_{2a} \dots))) \quad (8)$$

式(8)的目的是: 在  $l_2$  中寻找与  $l_1$  中的第1个词语  $w_{11}$  词性相同且相关度最大的词  $w_{2a}$ , 即最大相关组合, 若没有相同词性的词, 则依次查找相关度最大的词  $w_{2a}$ , 并记录词语  $w_{11}$  和  $w_{2a}$  的映射序列对 (在各自所属的句子中的词语顺序值), 依此类推, 直到  $T_1(l_1) = \emptyset$ , 最后, 得到词语最大相关组合集  $R$  和映射序列对集合  $B$ :

$$R = \{\text{sim}(w_{11}, w_{2a}), \text{sim}(w_{12}, w_{2b}), \dots, \text{sim}(w_{1m}, w_{2n})\} \quad (9)$$

$$B = \{(1, a), (2, b), \dots, (m, n)\} \quad (10)$$

偏移量  $\Delta d(w_{1i})$  是目标句子中任意一个词语  $w_{1i}$  所对应的左右词语及本身在比较句中的映射序列值之间的偏差, 如

图3所示。

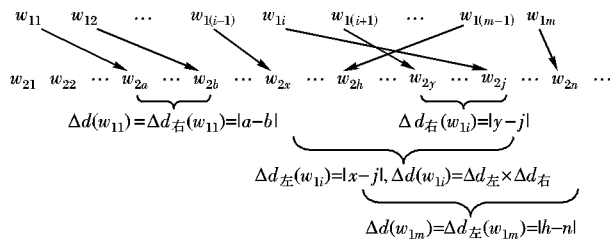


图3 偏移量解析

$$\Delta d(w_{1i}) = \begin{cases} |a-b|, & i=1, \text{目标词 } w_{11} \text{ 具备右偏移量} \\ |x-j| \times |y-j|, & 1 < i < T_1(l_1), \text{目标词 } w_{1i} \text{ 具备左右偏移量} \\ |h-n|, & i=m=T_1(l_1), \text{目标词 } w_{1m} \text{ 具备左偏移量} \end{cases} \quad (11)$$

$$\text{sim}(l_1, l_2) = \varphi \sum_{\substack{v=1, i=1 \\ y \leq T_2(l_2)}}^{T_1(l_1)} \text{sim}(w_{1v}, w_{2y}) \frac{1}{\Delta d(w_{1i})} \quad (12)$$

其中:偏移量  $\Delta d(w_{1i})$  与相关度值成反比,  $\varphi$  为可调节参数。

### 1.3 段落语义相关度计算

由句子的语义相关度计算扩展到段落的语义相关度计算。有段落  $t_1$  和  $t_2$ , 组成段落的句子集合分别为:  $t_1 = \{p_{11}, p_{12}, \dots, p_{1m}\}$ ,  $t_2 = \{p_{21}, p_{22}, \dots, p_{2n}\}$ , 句子的任意组合为:

$$\begin{bmatrix} p_{11}p_{21} & p_{11}p_{22} & \cdots & p_{11}p_{2n} \\ p_{12}p_{21} & p_{12}p_{22} & \cdots & p_{12}p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{1m}p_{21} & p_{1m}p_{22} & \cdots & p_{1m}p_{2n} \end{bmatrix} \quad (13)$$

其中:根据以上阐述的句子语义相关度计算方法得出所有的  $p_{1i}p_{2j}$ , 找出最大的相关度组合  $\text{sim}_{\max 1} = \text{sim}(p_{1x}, p_{2y})$ , 然后从矩阵(13)中去除  $x$  行  $y$  列, 依次循环, 直至矩阵为空, 最后形成了句子最大相关度组合集  $L = (\text{sim}_{\max 1}, \text{sim}_{\max 2}, \dots, \text{sim}_{\max k})$ , 则段落相关度为:

$$\text{sim}(t_1, t_2) = \frac{1}{k} \sum_{i=1}^k \text{sim}_{\max i} \quad (14)$$

### 1.4 文本语义相关度计算

文本  $h_1$  与文本  $h_2$  之间的相关度计算是将文本分成若干段落, 将段落自由组合, 形成矩阵, 最后得出段落最大相关度组合集  $G = \{(\text{sim}_{\max 1}, \text{sim}_{\max 2}, \dots, \text{sim}_{\max m})\}$ , 则:

$$\text{sim}(h_1, h_2) = \frac{1}{m} \sum_{j=1}^m \text{sim}_{\max j} \quad (15)$$

### 1.5 相关度算法

定义输入的描述性查询语句 *InputSentence*, 其包括3种形式: 句子 *Sentence*, 段落 *Paragraph*, 文本 *Text*,  $s_1, s_2, \dots, s_m$  代表句子组合,  $p_1, p_2, \dots, p_i$  为段落组合。

1) 如输入为 *Sentence*, 将要素所关联的文本分成  $s_1, s_2, \dots, s_m$ , 计算每个  $\text{sim}(\text{Sentence}, s_k)$ , 其中  $k \in \{1, 2, \dots, m\}$ , 句子与文本的相关度定义为所有相关度值的平均值。

2) 如输入为 *Paragraph*, 将要素所关联的文本分为  $p_1, p_2, \dots, p_i$ , 计算每个  $\text{sim}(\text{Paragraph}, p_k)$ , 其中  $k \in \{1, 2, \dots, i\}$ , 段落与文本的相关度定义为所有相关度值的平均值。

3) 如输入为 *Text*, 按照文本相关度计算方法, 得出文本与文本之间的相关度。

### 1.6 相关度值测试

以上方法的参数设置为:  $\alpha = 0.55, \beta = 0.085, \varphi = 0.25$ , 为了验证相关度值的准确性, 本文对实验相关度方法和人工直觉方法进行了简单对比(见表2~3)。

表2 词语相关度计算

词语1	词语2	相关度	词语1	词语2	相关度
父亲	母亲	0.87	好人	坏人	0.88
男人	女人	0.89	美女	大象	0.13
男人	责任	0.85	开门	苹果	0.11

表3 句子相关度计算

句子1	句子2	相关度
我们今天很开心。	今天是她的生日,我们去了很多地方,玩得很尽兴。	0.87
学生的任务是好好学习。	作为一个学生,我们必须尽自己最大的努力好好学习。	0.91
改革开放是社会主义发展的前提。	我们必须抓紧社会主义建设。	0.21

从表2~3可看出,实验相关度计算和人工直觉测试出的词语相关度、句子相关度基本符合,因此相关度计算方法较准确。

## 2 实验与分析

### 2.1 实验构成

句子相关度计算,对句子分词后词语个数分别为  $n$  和  $m$ , 计算句子相关度值的时间复杂度为  $O(nm)$ ; 段落相关度计算,每个段落分别有  $e$  和  $k$  个句子,每个句子有  $n$  和  $m$  个词,则计算段落相关度值的时间复杂度为  $O(ek + nm)$ 。

### 2.2 分词词典

非结构化文本数据是本文的主要操作对象,要实现非结构化文本数据的GIS描述性查询,必须要对训练文本和输入的描述性语句进行分词和词性标注。本文采用基于词典的分词方法进行中文分词<sup>[9-12]</sup>,以双Hashing结构词典为分词词典,词典的源文件是一个文本文件,这种词典的分词速度快,效率高,时间复杂度为  $O(1)$ 。双Hashing结构词典是嵌套型哈希表,父哈希表的值存储词语首字的Unicode编码,键为子哈希表,子哈希表的值为词语的长度,子哈希表的键为一个动态数组,这个数组用来存储长度等于子哈希表值的词。

### 2.3 实验数据

本实验选取了两种数据类型:一种是矢量空间数据;另一种是非结构化文本数据。矢量空间数据是采用1:1000昆明市地图作为空间数据底图,以昆明市的50个单位作为主要的研究对象。非结构化文本数据来源于这50个单位的文本语料库(来自于互联网上各个单位的简介及要闻),每个要素所关联的文本字数基本相同,文本中的内容以段落形式存在。表4给出了非结构化文本数据的相关参数。

表4 非结构化文本数据集

参数	值
数据来源	昆明市50个单位的文本语料库
文本大小	32 MB
训练文本总数	5500
每要素的训练文本数	110
每个要素连接的文本类别	地理位置、单位性质、业务范围、业绩、待遇、前景6个方面



2.4 实验查询与查询结果

实现两种不同的 GIS 描述性查询:第一种查询为  $Q_1$ , 输入的描述性查询语句为句子, 需要运用句子语义相关度计算方法, 计算出查询与各个要素所关联的文本的相关度, 如相关度值越大, 则要素的符号显示越大; 否则越小。第二种查询为  $Q_2$ , 输入的描述性查询语句为段落, 需运用段落语义相关度计算方法, 计算出查询与每个要素所关联的文本的相关度, 相关度值的大小决定了要素符号显示的大小。下面就两种查询分别给出一个例子。

表 5 描述性查询语句

查询种类	查询输入类别	查询输入内容
$Q_1$	输入为句子	云南省红塔集团昆明卷烟厂的位置。
$Q_2$	输入为段落	云南省红塔集团有限责任公司是一家 以烟草为主业、多元化、国际化经营的 跨地区、跨行业、跨所有制经营的大型 国际化公司, 拥有“玉溪”、“红塔山”、 “红梅”三大主力品牌, 是中国目前唯一 一家有三个品牌被列入“中国名牌”的 卷烟生产企业。

2.5 实验结果分析

图 5 为本实验的查询结果, 与传统的 GIS 查询模式所得

结果(见图 6)对比较明显。在图 6(a)中, 当输入为  $Q_1$  或者  $Q_2$  时, 由于各个要素的结构化属性数据不能完全与输入相匹配, 因而得不到查询结果; 在图 6(b)中, 当输入与存储的某一属性数据相匹配时, 能得到唯一的查询结果; 在图 5 中, 当分别输入  $Q_1$  和  $Q_2$  这样的描述性语句时, 依次计算出描述性语句与每一个要素所关联的文本的相关度, 根据相关度值, 得出查询结果。在图 5 中, 相对其他的参照要素, 分别有一个较大的点状要素图标, 这就清晰地阐述了输入的描述性查询语句与此点关联的文本的相关程度最大。从实验结果可看出, 非结构化文本数据的 GIS 描述性查询方法不但对输入无要求, 而且可以得到概括性查询结果, 文献[1-3]基于半结构化或结构化数据的 GIS 查询却未能实现。

3 结语

本文提出一种非结构化文本数据的 GIS 描述性查询方法, 该方法以《词林》为词源, 结合《词林》的内部结构和编排方式主要给出了词语相关度计算公式和句子相关度计算公式。

在词语相关度计算中, 充分考虑了《词林》的树状结构特点, 并加以分析度量, 计算较为全面可靠。

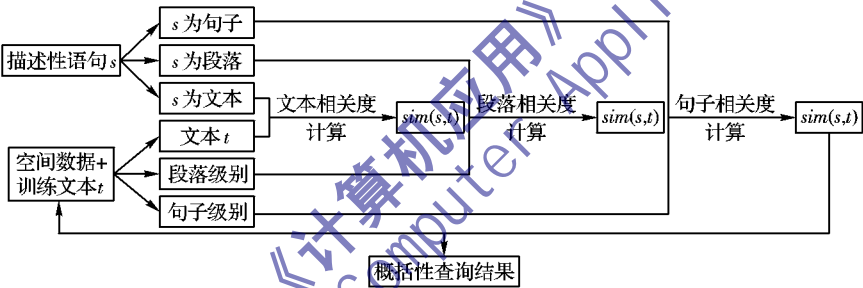
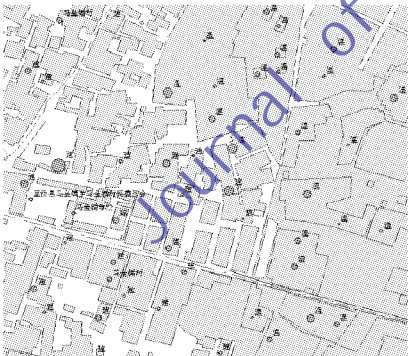


图 4 实验组成



(a) 输入为  $Q_1$



(b) 输入为  $Q_2$

图 5 本文方法



(a) 输入为  $Q_1$  或者  $Q_2$



(b) 输入与存储的属性数据相匹配

图 6 结构化查询

在句子相关度计算中,跳过了句法分析的难度,从词语之间的相互依存性和衔接性出发,不仅考虑到了词语之间的相关度,而且涉及到了词语偏移量,计算更为合理。

相关度计算方法还存在一些问题有待于进一步研究。首先需要更深层次的研究《词林》的组织结构,以及计算方法中的参数值选取,其次是计算效率,相关度计算的一个重要应用领域是信息检索,如果没有很高的效率,信息检索将无法实际应用,这个问题在于算法的数据结构需要进一步优化。

#### 参考文献:

- [1] BADARD T, RICHARD D. Using XML for the exchange of updating information between geographical information systems[J]. Computers, Environment and Urban Systems, 2001, 25(1): 17-31.
- [2] GUSTAVSSON M, SEIJMONSBERGEN A C, KOLSTRUP E. Structure and contents of a new geomorphological GIS database linked to a geomorphological map — with an example from Liden, central Sweden[J]. Geomorphology, 2008, 95(3/4): 335-349.
- [3] GARCIA - CUMBRERAS M A, PEREA - ORTEGA J M, GARCIA - VEGA M, *et al.* Information retrieval with geographical references. Relevant documents filtering vs. query expansion[J]. Information Processing & Management, 2009, 45(5): 605-614.
- [4] 廉站俊, 吕学强, 张玉杰, 等. 基于句子相似度计算的信息抽取[J]. 现代图书情报技术, 2007, 45(6): 38-41.
- [5] CHEN YANMIN, LIU BINGQUAN, WANG XIAOLONG. Automatic text summarization based on textual cohesion[J]. Journal of Electronics (China), 2007, 24(3): 338-346.

- [6] SONG FEI, CROFT W B. A general language model for information retrieval[C]// CIKM '99: Proceedings of the 8th International Conference on Information and Knowledge Management. New York: ACM Press, 1999: 316-321.
- [7] HUANG FENG-LONG, YU MING-SHING. Study on good-turing and a novel smoothing method based on real corpora for language models[C]// Proceedings of 2004 IEEE International Conference on Systems, Man and Cybernetics. Piscataway, NJ: IEEE Press, 2004, 4: 3741-3745.
- [8] 李庆虎, 陈玉健, 孙家广. 一种中文分词词典新机制: 双字哈希机制[J]. 中文信息学报, 2003, 17(4): 13-18.
- [9] CLARIZIA F, de SANTO M, NAPOLETANO P. A probabilistic method for text analysis[C]// Proceedings of 2009 the 9th International Conference on Intelligent Systems Design and Applications. Washington, DC: IEEE Computer Society, 2009: 932-937.
- [10] ZHAI CHENGXIANG, LAFFERTY J. A study of smoothing methods for language models applied to Ad Hoc information retrieval [C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2001: 334-342.
- [11] APAN A A, PETERSON J A. Probing tropical deforestation: the use of GIS and statistical analysis of georeferenced data[J]. Applied Geography, 1998, 18(2): 137-152.
- [12] 孙茂松, 肖明, 邹嘉彦. 基于无指导学习策略的无词表条件下的汉语自动分词[J]. 计算机学报, 2004, 27(6): 736-742.

(上接第2482页)

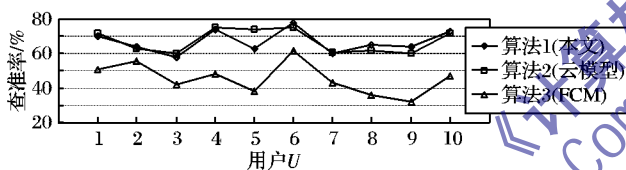


图1 3种算法推荐查准率比较

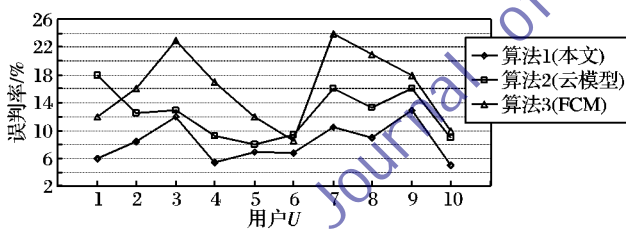


图2 3种算法推荐误判率比较

## 4 结语

针对在日益庞大的Web资源环境下,如何正确地将Web资源进行分类,以方便用户快捷搜索到所需要的信息,并实现Web资源的主动推荐这一问题,提出了一种改进的基于直觉模糊C均值聚类的Web资源推荐方法,并结合电影分类及推荐这一实例进行了阐述说明。理论分析及对比实验表明,相对于传统的用户等级评价或普通模糊隶属度评价方法,直觉模糊数更能客观细腻地描述用户对项目资源的兴趣意愿,在Web资源推荐过程中有较高的推荐质量。但本文方法还存在以下两方面的不足:一是推荐质量的评价问题,实验中所提出的评价方法过于简单,包括查准率和误判率,如某一用户在访问Web资源时,其对某些类的访问数据相同(如观看某几个类中的电影数目都一样),如何确定其感兴趣类;二是直觉模糊初始化聚类中心的获取过于复杂,计算量太大。这些问题,将在下一步的研究中逐步完善。

#### 参考文献:

- [1] ZAN H, HSINCHUN C, DANIEL Z. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering [J]. ACM Transactions on Information Systems, 2004, 22(1): 116-142.
- [2] BARRAGANS-MARTINEZ A B, COSTA-MONTENEGRO E, BUR-GUILLO J C, *et al.* A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition[J]. Information Sciences, 2010, 180(22): 72-78.
- [3] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Analysis of recommendation algorithms for E-commerce[C]// EC '00: Proceedings of the 2nd ACM Conference on Electronic Commerce. New York: ACM Press, 2001: 158-167.
- [4] ATANASSOV K. Intuitionistic fuzzy set[J]. Fuzzy Sets and Systems, 1986, 20(1): 87-96.
- [5] XU Z S. Some similarity measures of intuitionistic fuzzy sets and their applications to multiple attribute decision making[J]. Fuzzy Optimization and Decision Making, 2007, 6(2): 109-121.
- [6] 徐泽水. 直觉模糊信息集成理论及应用[M]. 北京: 科学出版社, 2008: 30-44.
- [7] KHATIBI V, MONTAZER G A. Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition[J]. Artificial Intelligence in Medicine, 2009, 47(1): 43-52.
- [8] 贺正洪, 雷英杰, 王刚. 基于直觉模糊聚类目标识别[J]. 系统工程与电子技术, 2011, 33(6): 1283-1286.
- [9] 吴成茂. 模糊C均值算法在直觉模糊聚类中的应用[J]. 计算机工程与应用, 2009, 45(16): 141-145.
- [10] 徐小来, 雷英杰, 赵学军. 基于模糊熵的直觉模糊聚类[J]. 空军工程大学学报, 2008, 9(2): 80-83.
- [11] 刘守生, 王忠, 张露. 基于遗传算法的直觉模糊C均值聚类算法[J]. 科学导报, 2011, 29(14): 56-59.
- [12] 张光卫, 李德毅, 李鹏, 等. 基于云模型的协同过滤推荐算法[J]. 软件学报, 2007, 18(10): 2403-2411.