doi:10.11887/j.cn.201902017 http://journal.nudt.edu.cn

遥感数据用户需求融合处理技术*

郑忠刚[1],付 琨[2],徐崇彦[1],巫震宇[1],周长飞[1]
(1. 北京遥感信息研究所,北京 100192;2. 中国科学院 电子学研究所,北京 100190)

摘 要:遥感数据是国家的基础性和战略性资源,在经济建设、国防建设、抢险救灾、生态环境保护等方面得到了广泛的应用,发挥着越来越重要的作用,各行各业对遥感数据的需求也越来越多。因此,如何提高对地观测资源的利用率,提高服务响应速度成为迫切需要解决的问题。采用自然语言处理技术,提出了一种用户需求融合处理方法,该方法可以有效地融合归并相同或者相似的用户需求,实现一图多用,引入需求预测和需求融合技术以提高需求融合效率,从而提高对地观测资源的利用率,达到事半功倍的效果。

关键词:遥感数据;需求融合;自然语言处理;聚类;语义转换;需求预测;需求挖掘
中图分类号:TP311 **文献标志码**:A **文章编号**:1001-2486(2019)02-115-09

Remote sensing data user request merging technology

ZHENG Zhonggang[1], *FU Kun*[2], *XU Chongyan*[1], *WU Zhenyu*[1], *ZHOU Changfei*[1]
(1. Beijing Institute of Remote Sensing Information, Beijing 100192, China;
2. Institute of Electrics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Remote sensing data, as national fundamental and strategic resources, plays an important role in economy, state security, environment protection and ecology. Since remote sensing data have been widely used in various industries, there is a large amount of data requests from different users, demanding a lot of valuable EOS (earth observation resources). On one hand, users' data requests increase constantly. On the other hand, EOS resources are always limited. Our objective is to ease the contradiction between a large amount of users' requests and limited EOS resources. The key idea is to merge identical or similar users' requests in order to reduce the total number of requests, and then to use requirement forecasting and requirement mining technology to improve requirement fusion efficiency. It is high likely that different users may share identical or similar data requests, since users may show concerns about the same area of the earth over the same time range.

Keywords: remote sensing data; request merging; natural language processing; clustering; semantic transformation; requirement forecasting; requirement mining

自1959年由Mark Ⅱ Reentry Vehicle人造卫星上发回第一张地球相片开始,半个多世纪来的遥感技术发展异常迅速,尤其是近十多年来,其发展速度明显加快,当今的遥感随着太空技术、计算机和地球科学的发展,已经产生了质的飞跃,从航空遥感发展到以卫星为主的航天遥感,目前的遥感应用领域越来越广泛,遥感技术应用于农业可以进行农作物识别分类与面积估算;应用于抢险救灾可以对森林火险进行预警,对火点发生地精确定位;应用于国民经济建设领域,可以实现目标观测和遥感救灾。随着遥感影像应用效果的展现,目前,各行各业对遥感影像数据的需求也越来越多,虽然因应用目的不同,各用户的遥感数据需求存在多样性,但不同用户之间也会存在相似或相同需求,特别是在发生热点事件和自然灾害事件时,各个参与单位会同时申请热点地区和受灾区域的遥感影像数据,这些需求往往会有相同或相似的需求。如何融合来自不同用户的需求,实现最大效率地利用卫星观测资源、地面接收资源、地面数据传输资源、降低卫星对地观测系统的任务负荷,需要开展需求融合归并技术研究,将相同或相似的遥感用户需求进行归并处理。

另外,用户在需求遥感影像时大多要受限于专业门户或软件的要求,需要填写专业的制式表单,表单中包含与平台、传感器等相关的专业参数,这对用户的遥感影像专业知识要求相对较高,普通用户更习惯于使用自然语言表达遥感影像需求。为此,首先需要将自然语言描述的用户需求

* **收稿日期**:2018-02-28
基金项目:国家自然科学基金资助项目(41801349)
作者简介:郑忠刚(1976—),男,辽宁沈阳人,博士研究生,E-mail:yitingxu856@139.com;
付琨(通信作者),男,研究员,博士,博士生导师,E-mail:fukun@mail.ie.ac.cn

转变为格式化遥感影像需求。为了解决上述问题,本文提出了一种基于自然语言处理的需求融合方法。

1 基于自然语言处理的需求融合方法

其基本原理如图1所示,包括关键信息抽取、需求转义和融合归并处理环节,涉及信息抽取知识库、需求转义知识库和融合知识库。

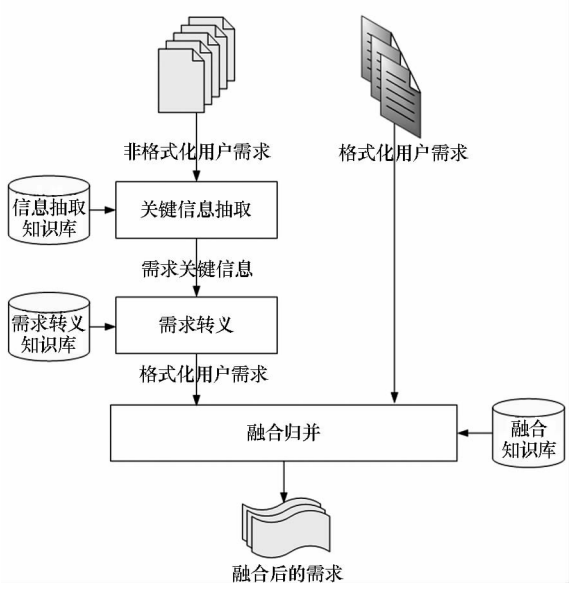

图1 基于自然语言处理的需求融合

Fig. 1 User request fusion based on natural language processing

在关键信息抽取环节,利用自然语言处理技术基于规则的信息抽取方法,从用户需求文本中抽取出时间范围、地域范围、任务、传感器类型要求等参数信息。

在需求转义环节,在用户需求关键信息抽取的基础上,对抽取结果进行规范,使其满足标准化和精确化的要求,达到可与卫星传感器性能指标相匹配的目的。

在融合归并环节,设计了用户需求相似度度量方法,利用聚类技术对批量用户需求进行聚类分析,找出其中潜在的相同或者相似的需求[1-3]。

1.1 关键信息抽取

用户需求中的关键信息抽取步骤主要解决从用户提出需求的文本中抽取遥感影像关键信息元素的问题,抽取的信息包括:时间范围、地域范围、任务、影像参数(空间分辨率、传感器类型和波段)。实际上,关键信息抽取实现的是用户需求的浅层语义分析,主要利用抽取规则实现关键信息的识别和抽取[4]。

基于自然语言处理(Natural Language Processing, NLP)的方法是早期的信息抽取方法,一般效率较低,现已较少使用。

每个用户的需求描述通常遵循某种习惯模式,且具有一定规律性,这种模式和规律性使得采用基于规则的方法进行关键信息抽取成为可能,因此,在本项目中采用基于规则的方法进行关键信息抽取,主要是针对不同关键信息文本片段内部组成的特征规律建立抽取规则,实现关键信息的识别和抽取。

下面就几种典型的关键信息要素如"时间""地理名称""经纬度"等的表现方式进行具体的展开说明。

1)时间。时间关键信息文本片段内部会出现"年、月、日、时、分、秒"等单位,通过对遥感影像用户需求的分析,常见的表现方式是:

方式1:年份数字+"年"+月份数字+"月";

方式2:年份数字+"年"+月份数字+"月"+日数字+"日";

方式3:年份数字+"年"+月份数字+"月"+日数字+"日"+小时数字+"时";

方式4:年份数字+"年"+月份数字+"月"+日数字+"日"+小时数字+"时"+分钟数字+"分";

方式5:年份数字+"年"+月份数字+"月"+日数字+"日"+小时数字+"时"+分钟数字+"分"+秒数字+"秒"。

另外,在用户需求中也会出现相对时间概念,例如,"11月20日",对于这种情况,需要根据上下文线索,确定具体指的是哪一年。

2)地理名称。代表国家地区的地理名称以及地物名称,例如用户需求"天坛5 m全色影像"中的天坛,"澳大利亚大堡礁10 m多光谱影像"中的"澳大利亚大堡礁"。另外,用户需求描述中还会出现一些地理名称和目标名称的缩写形式,例如,"日北海道3 m全色影像"中的"日",它的常见表现形式是:

方式1:国家地区名称或者目标名称;

方式2:作为国家的缩写出现时,例如代表"日本"的"日",其前面不会出现数字,其后会紧随所代表国家的地理名称或者所属目标的名称。

另外,在用户需求中也会出现相对位置概念,例如,"海南岛东北海域""海南岛以北海域",对于这种情况,需要依据领域知识经验,以海南岛为中心推定一个合理的位置。

3)经纬度。经纬度关键信息文本片段内部

格式主要有两种,一种是如“东经 120 度,北纬 23 度”,另一种是“120°E23°N”。上述两种格式特征可以表示为:

方式 1:“东经|西经” + 经度数字 + “度,” + “北纬|南纬” + 纬度数字;

方式 2:经度数字 + “E|W” + 纬度数字 + “N|S”。

另外,通过建立映射关系,地理名称与经纬度之间可以实现相互转换。

4)任务类型。任务类型通常是一些业务术语,例如,“水下地形探测”“农作物估产”“水污染监测”“水资源调查”“冬小麦估产”等。其常见表现形式是:“2016 年 7 月中上旬华北冬小麦估产”中的“冬小麦估产”,出现业务术语词汇的上下文中通常没有“任务类型”这样的引导词。

5)影像参数。遥感影像需求中的影像参数包括:分辨率、传感器类型、幅宽,通过对遥感影像用户需求的分析,上述参数常见的表现方式是:

①分辨率。

方式 1:“分辨率:” + 数字 + “ - ” + 数字 + “m”,例如,“分辨率:1 - 10 m”;

方式 2:“分辨率:” + 数字 + “ - ” + 数字 + “米”,例如,“分辨率:1 - 10 米”;

方式 3:“分辨率:” + 数字 + “m”,例如,“分辨率:10 m”;

方式 4:“分辨率:” + 数字 + “米”,例如,“分辨率:10 米”。

②传感器类型。

方式 1:“传感器类型:” + 传感器类型名称,例如,“传感器类型:多光谱”;

方式 2:传感器类型名称,例如,“美国关岛 10 m多光谱影像”中的“多光谱”。

③波段。

方式 1:“波段包含” + 波段名称 + “(” + 数字 + “ - ” + 数字 + “nm)”,例如,“波段包含近红外(400 - 760 nm)”;

方式 2:“波段包含” + 波段名称 + “和”波段名称,例如,“波段包含可见光和多光谱”;

方式 3:“波段:” + 波段名称 + “、” + 波段名称,例如,“波段:可见光、红外”;

方式 4:“波段含有” + 数字 + “ - ” + 数字 + “nm”,例如,“波段含有 2000 - 3500 nm”。

④幅宽。

方式 1:“幅宽不低于” + 数字 + “km”,例如,“幅宽不低于 200 km”;

方式 2:“幅宽不低于” + 数字 + “公里”,例如,“幅宽不低于 200 公里”;

方式 3:“幅宽” + “十里级/百里级/千里级” + “的影像”,例如,“幅宽百里级的影像”;

方式 4:“十里级/百里级/千里级”的幅宽,例如,“百里级的幅宽”;

方式 5:“百公里级”的幅宽,例如,“幅宽 200 公里以上”。

为了使抽取规则可被计算机理解和执行,需要对信息抽取规则前提条件中的特征谓词逻辑(特征词信息和命名实体信息)进行格式化表达,为此采用正则表达式技术实现规则前提条件的格式化表达。

以时间关键信息为例,相关的抽取规则示例如下:

1)时间信息实体抽取规则 1。

正则表达式:(\\d){4}(-)(\\d){2}(-)(\\d){2};

示例: 抽取形如“2013 - 10 - 29”的时间信息实体。

2)时间信息实体抽取规则 2。

正则表达式:

(\\d){4}(.)(\\d){2}(.)(\\d){2}(-)(\\d){1,2}(:)(\\d){1,2};

示例:抽取形如“2013. 10. 29 - 20:50”的时间信息实体。

3)时间信息实体抽取规则 3。

正则表达式:

(\\d){4}(-)(\\d){2}(-)(\\d){2}(\\d){1,2}(:)(\\d){1,2} (:)(\\d){1,2};

示例:抽取形如“2013 - 10 - 29 20:50:12”的时间信息实体。

4)时间信息实体抽取规则 4。

正则表达式:

(\\d){4}(年)(\\d){2}(月)(\\d){2}(日) (\\d) {1,2}(时)(\\d){1,2}(分)(\\d){1,2}(秒);

示例:抽取形如“2013 年 10 月 29 日 20 时 50 分 12 秒”的时间信息实体。

5)时间信息实体抽取规则 5。

正则表达式:

(\\d) {1,2}(时)(\\d){1,2}(分)(\\d){1,2}(秒);

示例:抽取形如“20 时 50 分 12 秒”的时间实体。

从上述分析可以看出,用户对于需求中各种关键信息描述方式是多种多样的,所对应的抽取

规则业务也是多种多样的,为了有效地组织和管理关键信息的抽取规则,采用了知识本体的方法,形成了信息抽取知识库[5]。

一个用户遥感需求通常包含 4 项概念要素:时间范围、地域范围、任务、影像参数,如图 2 所示。

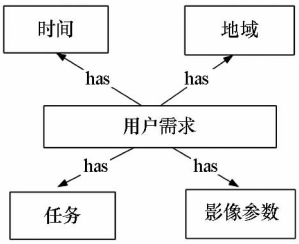

图 2 用户需求概念组成

Fig. 2 Concept of user request

本体包含五个基本的建模元语(modeling primitives),这些元语是:类/概念(classes/concepts)、关系(relations)、公理(axioms)、函数(functions)和实例(instances),每一个概念由关系、函数、公理和实例来界定,如图 3 所示。

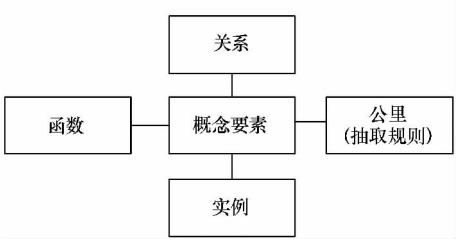

图 3 概念要素的本体描述

Fig. 3 Description of ontology concepts

图 4 是“时间”概念的本体描述,“时间”与“用户需求”之间存在隶属关系,“时间”是“用户需求”的概念要素之一;“公理”是从用户需求文本中抽取时间信息的具体信息抽取规则,正则表达式的形式为“(\\d){2,4}+(年)(\\d){1,2}(月)(\\d){1,2}(日)”;“时间”的实例是各种时间的具体表达形式,例如“09 年 8 月 10 日”“2009 年 8 月 10 日”等;“函数”是时间转换函数,其作用是将非标准的时间转变为标准的时间,例如将“09”转变为正式格式“2009”。

本体知识库的建立涉及两个方面:一方面,通过对用户需求历史数据进行观察分析,识别出其中的概念术语以及相互关系,确定本体知识库架构;另一方面,通过对用户需求历史数据中时间范围、地域范围、任务、影像参数等关键信息出现的上下文特征、关键信息自身特征,关键信息在整个文档中位置特征的分析,形成各个关键信息(概念)的抽取规则,抽取规则作为每个概念的公理存在,实现本体概念与抽取规则的结合和统一管理[6-7],如图 5 所示。

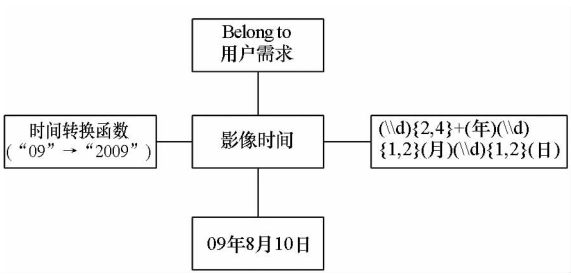

图 4 概念要素的本体描述实例

Fig. 4 Description of instances of concepts

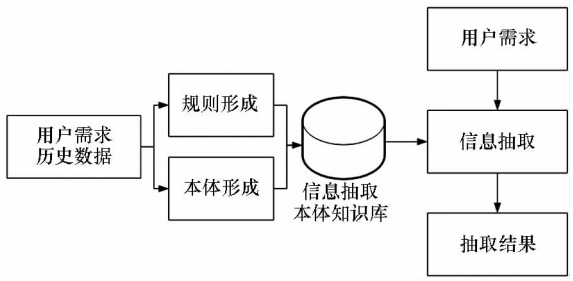

图 5 本体知识库的构建与应用

Fig. 5 Construction and application of ontology

通过对用户需求文本的分析,识别出各种关键信息的触发词、上下文约束条件、区位特征、句子特征、句内特征,基于这些知识构建由特征词汇构成的用户需求解析规则,给定一个用户需求文本,利用特征词汇形成的模式结构,结合前述四种要素的抽取模式,对需求文本进行解析,确定分别包含时间、地域、任务和传感器参数的文本子串,以及各个文本子串中包含的具体时间信息、地域信息、任务信息、影像参数信息。

例如,用户需求“2016 年 4 月下旬安徽省小麦纹枯病监测,采用高光谱影像,空间分辨率优于 5 m”,可利用下面的模式进行解析,【时间】+“对”【地域】+“进行”+【任务】+“采用”+【影响类型】+【空间分辨率】,解析出的时间信息、地域信息、任务信息、影像参数信息如表 1 所示。

表 1 关键信息抽取示例

Tab. 1 Example of key information extraction

关键信息项	关键信息
时间	2016 年 4 月下旬
地域	安徽省
任务	小麦纹枯病监测
影像参数(空间分辨率)	空间分辨率优于 5 m
影像参数(传感器类型)	高光谱影像

1.2　需求转义

需求转义是在用户需求关键信息抽取的基础上,对抽取结果进行规范,使其满足标准化和精确化的要求,实际上,需求转义实现的是用户需求的深层语义分析。

1)时间信息转义。将识别出来的各种格式的时间转变为标准格式。

2)地域信息转义。将识别出来的地域范围转变为由一系列经纬度值定义的多边形。

3)任务信息转义。将任务描述转变为具体的影像参数,任务名称转义基于需求转义知识库,知识库中包含任务与影像参数之间的映射关系,反映的是完成某种任务用户所需的影像参数,适用于各用户的需求转义知识库示例如表 2 所示。

表 2　需求转义知识库示例

Tab. 2　Example of request semantic translation

用户	任务	传感器类型
农业	农业病虫害监测	近红外波段
	农作物估产	多光谱影像、全色
	农作物长势	多光谱影像、全色
	农作物的叶面指数	中分辨率成像光谱仪
	春小麦面积监测	近红外、短波红外、可见光
海洋	海洋测深、水透明度、海流、油膜(泄漏)、海底类型、大气能见度、潮汐、生物体发光、海滩特征、水下危险事件、大气水汽总量、浅海水下地形	高光谱
	海表面温度	红外
	海平面平均高度、大地水准面、有效波高、海面风速、表层流、海面风场、海面温度、海面风速	微波高度计
交通	区域交通压力评价	全色、SAR
能源	油气田勘探	高光谱
环境	大气污染	微波扫描辐射计

1.3　融合归并

需求融合归并是在统一、标准化的时间、地域、影像参数格式的基础上进行的,融合归并的用户需求既包括格式化的需求也包括非格式化的需求。

融合归并是在融合知识库的支持下完成的,融合知识库中包含着多种融合规则。

根据时间、地域、传感器类型、光谱分辨率、空间分辨率、幅宽等方面对用户需求之间的相似度进行分析计算,根据计算结果进行需求融合归并。

需求的融合归并问题实际上是用户需求的聚类过程,经过聚类运算将一批用户需求聚为若干个簇,簇内的用户需求在时间、地域、传感器类型、光谱分辨率、空间分辨率、幅宽等方面相同或者相似。

为了计算需求之间的相似度,需要确定时间、地域、传感器类型、光谱分辨率、空间分辨率、幅宽等方面相似度的量化标准。

为了计算需求之间的相似度,需要对需求在时间 T、地域 A、传感器类型 S、光谱分辨率 V、空间分辨率 P、幅宽 W 等指标上的相似度进行量化处理,实现在统一量纲下的相似度评估,量化处理的示例见表 3,可以根据具体情况改变量化方法。

表 3　需求指标相似度量化标准

Tab. 3　Similarity metrics for quantification

需求指标要求	指标相似情况	量化结果	说明
时间 T	完全相同或者完全包容	10	如果两个需求时间段相同,或者一个需求完全包含在另一个需求的时间段内,则量化结果为 10
	时间段部分相交	$10\times\lambda$,当 $\lambda\geq 50\%$;0,当 $\lambda<50\%$	如果两者相交部分与由两者中最早开始时间及最晚结束时间构成的时间段的比(λ)小于 50%,则相似度为 0
地域 A	完全相同或者完全包容	10	如果两个需求地域范围完全相同,或一个需求的地域范围包含在另一个需求的地域范围内,则量化结果为 10
	地域范围部分相交	$10\times\lambda$,当 $\lambda\geq 50\%$;0,当 $\lambda<50\%$	如果两者相交部分与由两者融合后形成区域的面积之比(λ)小于 50%,则相似度为 0
传感器类型 S	传感器类型相同	10	传感器类型不同,则相似度为 0
	传感器类型不同	0	

表3(续)

需求指标要求	指标相似情况	量化结果	说明
光谱分辨率 V	光谱分辨率相同	10	光谱分辨率不同,且差值大于1时,相似度为0
	光谱分辨率不同	8,当两者数值相差小于等于1时; 0,当两者数值相差大于1时	
空间分辨率 P	空间分辨率相同	10	空间分辨率不同,且差值大于1时,相似度为0
	空间分辨率不同	8,当两者数值相差小于等于1时; 0,当两者数值相差大于1时	
幅宽 W	空间分辨率相同	10	幅宽不同,且差值大于10时,相似度为0
	空间分辨率不同	8,当两者数值相差小于等于10时; 0,当两者数值相差大于1时	

相似度的计算公式为:

$$Similarity = T \times \lambda_1 + A \times \lambda_2 + S \times \lambda_3 + V \times \lambda_4 + P \times \lambda_5 + W \times \lambda_6$$

其中:$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ 和 λ_6 是权重系数;$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 = 1$。

相似度的确定是需求融合归并的关键,确定了相似度的值才能通过聚类算法进一步得出类内用户之间合适的总体内聚度。

进一步采用的聚类算法是基于类内用户需求间相似度值柱状图进行的,相似度值柱状图是类内所有用户需求相似度的简洁表示。算法的基本思想是,保证在聚类运算过程中类内用户需求之间的总体内聚度尽可能高。总体内聚度是通过计算类内两两用户需求之间的相似度并观察相似度值的分布而得出的。把相似度值划归到几个值的区间,每个区间用一个长方柱表示,落在每个区间内相似度值的数量代表该区间在整个分布中的比重,长方柱高度与落入区间内的相似度值数量成正比,包含较大相似度值区间的比重越大,说明该类的内聚度越大。

假定 n 是某一聚类内的用户需求数量,那么就有 $m = n(n-1)/2$ 个需求之间的相似度值。设 $S = \{s_1, s_2, \cdots, s_m\}$ 是 m 个相似度值集合。把这些相似度值划归到 B 个区间,$H = \{h_1, h_2, \cdots, h_B\}$ 表示所有相似度值在 B 个区间内的分布情况。

类内需求间值的内聚度 C 可由下式计算,即

$$C = \frac{\sum_{i=T}^{B} h_i}{\sum_{j=1}^{B} h_j}, \quad T = \lfloor S_T \cdot B \rfloor$$

其中,S_T 为相似度阈值(设定一个阈值,需求间的相似度超过这一阈值则认为它们是相似的),T 为与相似度阈值对应的相似度区间号。

在聚类过程中,要始终把保持较高的类内聚度作为目标,为此,当对一个新需求进行归类时,要对这一需求的加入对该类造成的影响进行评估。如果由于新需求加入导致类的相似度值分布明显变坏,就不能把新需求分至该类。当然也可以采取更为严格的条件,即要求新需求的加入必须使该类的相似度分布有所改善,也就是要好于原来的情况才允许新需求加入。苛刻条件带来的负面影响是,即使新需求与某个类中大部分的需求都相似,此类也会拒绝新需求的加入。在实际应用中可以不采取这样严格的条件,即允许相似度分布有所变坏,但通过设定一些附加条件,防止情况发生大的变化。

如前所述,希望保持每个类的 C 值越大越好,但实际应用中一个新需求的加入可能使相似度值分布略有变坏。所以,一个类的 C 值有可能随新需求的不断加入变得越来越低。为了防止由于 C 值连续降低,最终导致类的内聚度持续变坏的情况发生,为 C 设置一个最低值限制。如果一个新需求的加入会使 C 值小于设置的最低值,那么就拒绝新需求加入该类。另外,即使一个新需求的加入未使 C 值降低到小于设置的最低值,但 C 值的降幅过大(超出某个设定的阈值)也不能把新需求加入该类。这样就可以防止由于一个需求的原因导致一个类的内聚度急剧变坏。

其中:C_{Min} 表示要求的最小类内聚度;C_{New} 表示新文档加入后类的内聚度;C_{Old} 表示新文档加入前

类的内聚度;ε 表示允许的内聚度降低限度。

聚类算法如算法 1 所示。算法 1 的描述,对于每一个新需求,都计算假定新需求加入一个现有类后的 C 值,并将其与原来的 C 值相比较。若新 C 值大于或等于旧 C 值,则把新需求加入该类。如果新 C 值小于旧 C 值,但降低量小于 ε 且新 C 值大于 C 值最小要求值,也可以把新需求加入该类。若不是上述情况,则不能把新需求加入该类。如果一个新需求不能被分给任何一个现有类,则创建一个新类,并把它放在这个新类中。

算法 1　聚类算法

Alg. 1　Clustering algorithm

输入:新需求 d、聚类列表 L
输出:添加新需求 d 后的聚类列表 L

```
1. L← Empty list/ Cluster List
2. for each document d do
3.     for each cluster c in L do
4.         C_Old = C
5.         Simulate adding d to C
6.         C_New = C
7. If (C_New ≥ C_Old) OR ((C_New ≥ C_Min) AND (C_Old - C_New < ε))
8. then
9.         Add d to C
10.    End if
11.    End for
12. If d was not added to any cluster then
13.    Create a new cluster C
14.    Add d to C
15.    Add C to L
16. End if
17. End for
```

与传统常用的 K 最近邻分类(K-Nearest Neighbor, K-NN)算法相比,相似度柱状图方法能够更准确地表示类内需求的相似程度,以及新需求给类的内聚度带来的影响。

1.4　验证与分析

构建遥感数据用户需求融合处理原型系统,原型系统的组成如图 6 所示。

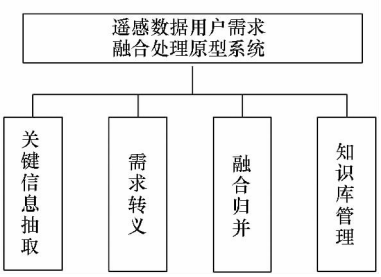

图 6　遥感数据用户需求融合处理原型系统组成

Fig. 6　Composition of remote sensing request fusion system

关键信息抽取模块负责抽取用户需求文本中的时间、地理范围、任务、传感器参数等关键信息;需求转义模块负责将抽取出的关键信息转变为标准化和精确化的指标要求;融合归并模块负责将相同或者相似的用户需求合并;知识库管理模块负责维护管理信息抽取知识库、需求转义知识库和需求融合归并知识库。

利用农业生产、国土资源、防灾减灾领域各 100 份用户需求,共计 300 份需求进行了需求融合试验,试验结果表明,融合归并的正确率大于 90.2%。表 4 是部分用户需求关键信息抽取、需求转义及最终融合归并结果示例[8]。

表 4　用户需求关键信息抽取、需求转义及最终融合归并结果示例

Tab. 4　Examples of request key information extraction, semantic translation and request fusion

序号	需求样例	浅层解析结果	深层解析结果	融合后需求
1	2008 年天津春小麦面积监测	时间:2008 年 地点:天津 对象事件:春小麦面积监测	时间:2008-1-1 0:00:00—2008-12-31 23:59:59 地点:中国天津市,west:116.657 888,east:118.026 289,north:40.194 066,south:38.548 975 对象事件:春小麦面积监测 分辨率:10~20 m 观测时间:6 月 1 日—6 月 30 日 光谱段:近红外、短波红外、可见光	时间:2008-1-1 0:00:00—2008-12-31 23:59:59 地点:中国河北省,west:113.439 867,east:119.802 968,north:42.562 44,south:36.038 584 对象事件:春小麦面积监测 分辨率:10~20 m 观测时间:6 月 1 日—6 月 30 日 光谱段:近红外、短波红外、可见光
2	2008 年北京春小麦面积监测	时间:2008 年 地点:北京 对象事件:春小麦面积监测	时间:2008-1-1 0:00:00—2008-12-31 23:59:59 地点:中国北京市,west:115.404 177,east:117.464 825,north:41.057 009,south:39.417 053 对象事件:春小麦面积监测 分辨率:10~20 m 观测时间:6 月 1 日—6 月 30 日 光谱段:近红外、短波红外、可见光	
3	2008 年河北春小麦面积监测	时间:2008 年 地点:河北 对象事件:春小麦面积监测	时间:2008-1-1 0:00:00—2008-12-31 23:59:59 地点:中国河北省,west:113.439 867,east:119.802 968,north:42.562 44,south:36.038 584 对象事件:春小麦面积监测 分辨率:10~20 m 观测时间:6 月 1 日—6 月 30 日 光谱段:近红外、短波红外、可见光	

2 用户历史遥感需求的挖掘分析技术

需求融合完成了提高需求处理效率的第一步,为了进一步增加需求融合的效率,需要对用户长期累积的历史需求及潜在的需求进行挖掘分析,并与现有的用户需求进行融合。

对累积的用户遥感任务需求和用户遥感影像数据需求进行挖掘分析,发现需求的来源、对地观测区域范围、对地观测时间、传感器类型等参数之间的关联规律。依据这些知识规律可以事先自动生成用户需求,实现用户需求的自动智能提交,保证用户的规律性需求能够尽早提交到卫星管控部门。

由于用户需求数量众多,具体内容多种多样,如何从海量用户需求中找出用户的特点和规律是一个难点问题。其难点在于,代表需求参数之间关联关系的规律是隐式的,而不是显式的,需要对数据进行复杂的数据分析才能发现,特别是在海量数据规模条件下解决上述问题就显得尤其复杂。数据挖掘技术能够从海量数据中发现潜在的、有价值的规律和知识,是分析海量数据的有效工具,是解决上述问题的有效技术手段。

根据遥感任务需求的数据特点,设计了挖掘分析算法,算法分为两步进行:第一步,根据遥感任务需求数据库生成一棵 FP-tree;第二步,根据第一步生成的 FP-tree,利用观测任务需求挖掘分析算法生成所有频繁项目集。下面分别介绍如何建立 FP-tree 和如何使用观测任务需求挖掘分析算法[9-11]。

FP-tree 的定义如下:

1)它有一个标记为"null"的根节点,它的子节点为一个项前缀子树(item prefix subtree)的集合,还有一个频繁项(frequent item)组成的头表(header table)。

2)每个项前缀子树的节点有三个域:item - name,count,node_link。item - name 记录了该节点所代表的项的名称;count 记录了所在路径代表的交易(transaction)中达到此节点的交易个数;node_link 指向下一个具有同样的 item - name 域的节点,要是没有这样一个节点,就为 null。

3)频繁项头表的每个表项(entry)由两个域组成:item - name;node_link。node_link 指向 FP-tree 中具有与该表项相同 item - name 域的第一个节点。

根据一个数据库建立一棵 FP-tree 算法的形式化描述如算法 2 所示。

算法 2 FP-tree 算法

Alg. 2 FP-tree algorithm

输入:一个交易数据库 D 和一个最小支持度阈值
输出:FP-tree

1. 扫描数据库 D 一遍。得到频繁项的集合 F 和每个频繁项的支持度。把 F 按支持度递降排序,结果记为 L
2. 创建 FP-tree 的根节点,记为 T,并且标记为'null',然后对 D 中的每个交易 Transaction 做如下操作
3. 根据 L 中的顺序,选出 Transaction 中的频繁项并对其进行排序。把 Transaction 中排好序的频繁项列表记为$[p|P]$,其中 p 是第一个元素,P 是列表的剩余部分。调用函数 insert_tree($[p|P]$,T)

函数 insert_tree($[p|P]$,T)的功能如下:

如果 T 有一个子结点 N,其中 N. item - name = p. item - name,则将 N 的 count 域值增加 1;否则,创建一个新节点 N,使它的 count 为 1,使它的父节点为 T,并且使它的 node_link 和那些具有相同 item - name 域串起来。如果 P 非空,则递归调用 insert_tree(P,N)。

FP-tree 是一个压缩的数据结构,它用较少的空间存储了后面频繁项集挖掘所需要的全部信息。

第二步以第一步产生的 FP-tree 为基础。它会递归调用自己,并且反复调用新产生的 FP-tree。观测需求挖掘算法如算法 3 所示。

算法 3 需求挖掘分析算法

Alg. 3 Demand mining algorithm

输入:一棵在第一步建立的 tree
输出:所有的频繁项集

1. Procedure 观测任务需求挖掘分析算法(tree,α)
2. {
3. if tree 只有一条路径 P
4. then 对 P 中节点的每一个组合(记为 β)做步骤 1
5. 产生频繁项集 β∪α,并且把它的支持度指定为 β 中节点的最小支持度
6. else 对 tree 的头表从表尾到表头的每一个表项(记为 a)做步骤 2 ~ 5
7. 产生频繁项集 β = a∪α,支持度为 a
8. 建立 β 的条件模式库(conditional pattern base)和 β 条件树(conditional FP-tree)tree2
9. if tree2! = ∅
10. then 调用遥感任务需求挖掘分析算法(tree2,β)
11. }
12. 一旦在数据库 D 的事务中找出频繁项集,就可以使用下列公式产生各种关联规则,即
13. Confidence(A => B) = support_count(A ∪ B)/support_count(A)
14. 其中,support_count(A∪B)是包含 A∪B 的事务数,support_count(A)是包含 A 的事务数

3　结论

随着我国在轨卫星数量和质量的逐步提升,遥感数据的应用也日益成熟和扩展,遥感用户需求的数量和来源也越来越广泛。为了提高卫星的应用效益,对遥感用户的需求进行融合处理已成为必然。本文利用自然语言处理技术、需求分析技术和聚类技术对遥感数据用户需求进行分析,在分析的基础上,对相同或者相似的需求进行融合归并,实现了一图多用的目的,提高了对地观测资源的利用率,实验结果表明,该方法能够有效地对自然语言形式的用户需求进行融合归并处理,在卫星任务管控领域具有应用价值。

参考文献(References)

[1] 马满好, 祝江汉, 范志良, 等. 一种对地观测卫星应用任务描述模型[J]. 国防科技大学学报, 2011, 33(2): 89 - 94.
MA Manhao, ZHU Jianghan, FAN Zhiliang, et al. A model of earth observing satellite application task describing [J]. Journal of National University of Defense Technology, 2011, 33(2): 89 - 94. (in Chinese)

[2] 马万权, 张学庆, 崔庆丰, 等. 多用户对地观测需求统筹处理模型研究[J]. 测绘通报, 2014(S1): 141 - 143.
MA Wanquan, ZHANG Xueqing, CUI Qingfeng, et al. Research on multi-user earth observation demand coordination processing model[J]. Bulletin of Surveying and Mapping, 2014(S1): 141 - 143. (in Chinese)

[3] 巫兆聪, 刘培, 巫远. 一种多光谱遥感应用需求综合方法[J]. 应用科学学报, 2017, 35(5): 658 - 666.
WU Zhaocong, LIU Pei, WU Yuan. Synthesis of requirements in applications of multi-spectral remote sensing[J]. Journal of Applied Sciences, 2017, 35(5): 658 - 666. (in Chinese)

[4] 文坤梅. 基于本体知识库推理的语义搜索研究[D]. 武汉: 华中科技大学, 2007.
WEN Kunmei. Research on semantic search based on ontology repository reasoning [D]. Wuhan: Huazhong University of Science and Technology, 2007. (in Chinese)

[5] 刘玉龙. Web 信息抽取规则的设计和实现[D]. 南京: 南京大学, 2013.
LIU Yulong. Design and implementation of Web information extraction rules[D]. Nanjing: Nanjing University, 2013. (in Chinese)

[6] 谭玉玲. 基于正则表达式的数据处理应用 [J]. 武汉理工大学学报(信息与管理工程版), 2010, 32(2): 249 - 252.
TAN Yuling. Application of regular-expression based data processing [J]. Journal of Wuhan University of Technology (Information & Management Engineering), 2010, 32(2): 249 - 252. (in Chinese)

[7] 杨威. 基于正则表达式的 Web 信息抽取系统的研究与实现[D]. 西安: 西安电子科技大学, 2013.
YANG Wei. The research and implementation of Web information extraction system based on the regular expression[D]. Xi′an: Xidian University, 2013. (in Chinese)

[8] 张素香. 信息抽取中关键技术的研究[D]. 北京: 北京邮电大学, 2007.
ZHANG Suxiang. Research on key technologies of the information extraction[D]. Beijing: Beijing University of Posts and Telecommunications, 2007. (in Chinese)

[9] 何国金, 张晓美, 焦伟利, 等. 基于数据挖掘机制的卫星遥感信息智能处理方法研究[J]. 科学技术与工程, 2005, 5(24): 1911 - 1915.
HE Guojin, ZHANG Xiaomei, JIAO Weili, et al. A study on the data mining strategy based intelligent information processing technologies for satellite remote sensing [J]. Science Technology and Engineering, 2005, 5(24): 1911 - 1915. (in Chinese)

[10] 蔡伟杰, 张晓辉, 朱建秋. 关联规则挖掘综述[J]. 计算机工程, 2001, 27(5): 31 - 33.
CAI Weijie, ZHANG Xiaohui, ZHU Jianqiu. Survey of association rule generation [J]. Computer Engineering, 2001, 27(5): 31 - 33. (in Chinese)

[11] 花红娟, 张健, 陈少华. 基于频繁模式树的约束最大频繁项集挖掘算法[J]. 计算机工程, 2011, 37(9): 78 - 80.
HUA Hongjuan, ZHANG Jian, CHEN Shaohua. Mining algorithm for constrained maximum frequent itemsets based on frequent pattern tree [J]. Computer Engineering, 2011, 37(9): 78 - 80. (in Chinese)