基于 XML 的有害藻华监测数据交换机制及灾害预警聚类方法的研究

姬 鹏，张承慧，李 俊

（山东大学控制科学与工程学院，山东 济南 250061）

摘 要： 有害藻华监测数据具有复杂性、数据来源的多源性、数据平台异构、数据格式不规范等特点。而 XML 具有可扩展性、简单性、开放性、互操作性及支持多国语言等优势。基于此，运用 XML 技术建立了有害藻华监测数据交换机制模型 HABML，有效解决了有害藻华监测数据的管理、传输及交换等问题，为有害藻华的高效预警奠定了基础。同时，将对 FCM 聚类分析算法进行改进，引入到有害藻华灾害预警中，直观地反映了海洋要素样本的分布规律，为有害藻华的灾害预警提供了一种新的思路。

关键词： XML，HABML，有害藻华，聚类分析

中图分类号： P714 **文献标识码：** A **文章编号：** 1003 － 0239（2008）2－ 0064 － 10

1 引言

近年来，伴随着人类大规模开发利用海洋，海洋的污染程度日趋严重，导致有害藻华灾害发生的频率越来越高，规模越来越大，持续时间越来越长，对我国沿海的海洋生态环境产生了严重的影响并导致巨大的经济损失。因此，迫切需要研制有效的有害藻华监测与预报系统，以满足海洋资源开发、利用和社会、经济发展的需求[1～2]。

有害藻华的预测需要有大量、准确、连续和实时的海洋环境要素监测数据做支持，同时需要有一个跨平台的系统的数据交换机制来封装、存储、交换、配置这些海洋数据。考虑到 XML (eXtensible Markup Language) 语言具有可扩展性、简单性、开放性、互操作性及支持多国语言等优势，以及国际上已经投入使用的海洋 XML (MarineXML[3～4]) 在海洋数据交换机制中的成功运用，结合当前国外先进的研究成果，本文分析了与有害藻华相关的的海洋数据资料的特点，建立了基于 XML 的有害藻华监测数据交换机制平台架构 HABML（Harmful Algal Blooms Markup Language)。同时，以 HABML 架构为依托，引入了 FCM 聚类分析算法，对与有害藻华相关的海洋数据进行聚类分析[5]，发现这些数据之间潜在的联系，为建立实时、高效的有害藻华预警系统的提供了可靠保障。

2 有害藻华监测要素需求分析

收稿日期： 2007-09-12

基金项目： 山东省重点科技攻关项目（2004GG2205108）

作者简介： 姬 鹏（1975-），男，博士研究生，研究方向为信号处理、数据挖掘技术，海洋检测技术，控制理论与控制工程。

海洋是一个复杂的大系统，为了跟踪、研究、预测海洋环境发生的变化，需要收集大量的海洋环境要素的数据资料，通常包括水文、气象、化学和生物四大类数据资料，每类数据资料又包含若干类数据。因此，数据资料种类非常繁多。其中和有害藻华相关的监测要素就多达 23 种。

除了以上监测要素以外，还需要对海洋数据监测台站名称、监测台站的位置、海洋数据获取时间、海洋数据的质量等一系列数据进行系统的管理，设计出合理、实用的数据结构和数据交换机制，为有害藻华灾害准确监测和预警提供了必要保障。图 1 描述了有害藻华监测数据管理的整体流程，作为 HABML 设计的依据，HABML 的设计围绕有害藻华监测数据管理工作流程进行。

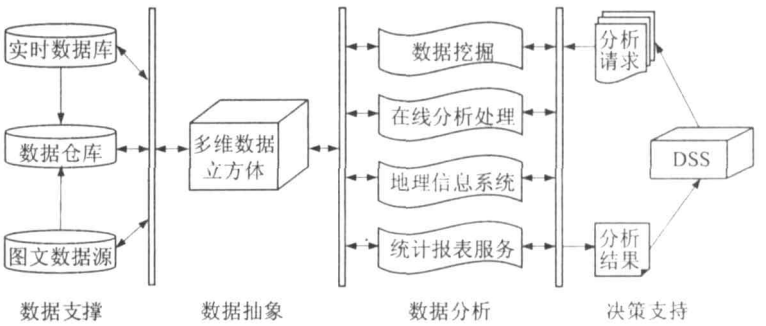

图 1　数据管理工作流程

为了满足有害藻华监测和预警的要求，本文根据“海洋调查规范（GB/T 12763）”、“海洋监测规范(GB/T 17378-1998)”、“海洋有害藻华监测技术导则(HY/T 069-2003)”、“海洋站自动化观测通用技术要求(HY/T 059-2002)”、“海洋调查观测监测档案业务规范”、“工程海冰技术规范”等标准，以图 1 为基础，设计了合理的数据结构模型 HABML，并以 HABML 架构为依据，引入聚类分析算法，为有害藻华的预警提供了有力支持。

3 有害藻华监测数据交换机制设计

有害藻华监测数据交换机制设计，主要包括 HABML 架构设计、HABML 数据架构定义、XML 代码开发三个部分。

3.1 HABML 架构设计

HABML（Harmful Algal Blooms Markup Language）即为有害藻华标记语言。HABML 架构旨在建立有害藻华监测系统内部的监测数据的交换机制，主要实现以下六个目标。

(1) 提供一种对海洋数据进行编码的方法，以方便对种类繁多、性质各异的海洋数据进行存储和移植。

(2) 海洋数据编码易于理解，具有自描述功能。

(3) 结合 XML 技术开发具有可互操作性的海洋数据管理软件，合理地管理海洋数据，有效地为有害藻华预警服务。

(4) 可将 XML 定义的海洋数据以表格、图表、网页等需要的格式进行输出。

(5) 将所有的海洋数据单元结合在一起，建立一个完整的海洋数据系统。

(6) 为有害藻华灾害预警提供数据支持。

要实现以上六个目标，就需要建立一个合理的数据架构用来封装不同种类、不同类型的海洋数据。该架构必须能够准确、详细地对不同海洋数据的空间、时间、内容、质量等属性做出描述，既要有机、合理、层次分明地组织海洋数据，又要便于各种海洋信息的检索、交换、存储。

本文在充分分析了与有害藻华相关的各种海洋数据的属性并充分考虑数据交换架构设计目标的基础上，将全部海洋信息分为两大类：将与监测台站相关的海洋空间、时间信息分为一类；将与有害藻华灾害有关的海洋监测数据实体分为一类，在这两个类下面又根据海洋信息的不同特点进行了详细分类，得到了 HABML 架构。图 2 为 HABML 架构图。

图 2 描述了用来封装有害藻华监测数据的 HABML 架构的整体结构。在 HABML 中，HABData 是整个数据集的父元素(根元素)，它包含两个类 Location 和 Monitor。其中 Location 类反映了海洋监测数据的空间和时间信息；Monitor 类则是海洋监测数据实体，包含相关的监测数据及其相关属性。Location 和 Monitor 两个类的具体结构如下文所述。

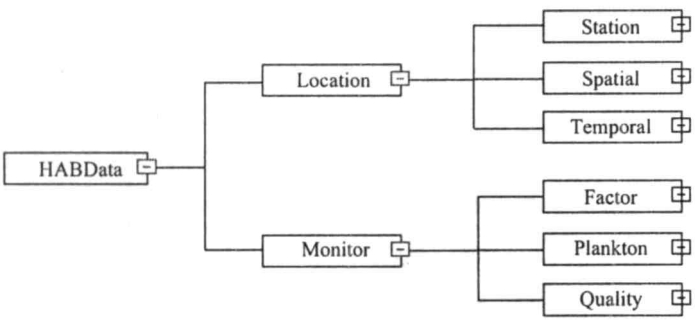

图 2 HABML 架构

3.1.1 Location 架构分析

有害藻华海洋数据主要是从不同台站监测得到的，所以不同的海洋数据具有不同的台站、空间、时间属性。Location 元素主要用来描述与海洋数据和台站相关的特性(见图 2)。

Location 元素包含了 Station、Spatial、Temporal 三个子元素，分别用来表示台站相关信息、海洋要素的空间属性以及海洋要素的时间属性。Station(台站信息)包含 Name (台站名称)、Abbreviation (台站拼音简称)、Code (台站站号)、Region (台站所属区域) 四个子元素；Spatial (海洋要素的空间属性) 包含 Latitude (纬度) 和 Longitude (经度) 两个子

元素，而 Latitude 和 Longitude 又都分别具有各自的子元素；Temporal（海洋要素的时间属性）包含 Date（日期）和 Time（时间）两个子元素。

3.1.2 Monitor 架构分析

Monitor 元素主要用来描述海洋数据各监测要素的数值、单位以及监测要素的获取方法，整体结构（见图 2）。

Monitor 元素包含了 Factor、Plankton、Quality 三个子元素，分别用来表示普通监测要素、浮游生物要素以及海洋要素的质量。Factor（普通监测要素）包含 Name（监测要素名称）、Type（监测要素类型）、Units（监测要素单位）、Value（监测要素值）、Method（监测方法）、Instrument（监测仪器）六个子元素，其中 Method 和 Instrument 又都分别具有各自的子元素；Plankton（浮游生物要素）包含 Type（浮游生物类型）、Name（种名）、Species（种属）、Quantity（数量）、Units（单位）五个子元素；Quality（海洋要素质量）包含 Symbol（数据质量符）和 Principal（数据质量负责人）两个子元素。

3.2 HABML 数据架构定义

在 HABML 架构基础上，需要设计具体的数据定义，以便使 HABML 架构形象化。HABML 相关元素实体的具体定义完全按照 HABML 架构图的结构定义了有害藻华监测数据，直观、详细地表述了不同海洋监测数据间的组织关系及各种海洋监测数据的表示方法，为编写 XML 代码铺平了道路。

3.3 XML 代码开发

XML 具有较成熟的数据交换机制，在许多行业中普遍形成了以 XML 数据文档为核心的集中的星状交换模型[6]（见图 3）。其中每个系统都将其内部的数据转换成符合行业标准的基于 XML 的数据文档，并将其作为系统间数据交换的媒介。数据在数据源被封装到 XML 数据文档中，通过传输介质传输到数据的请求端，再由解析器解读 XML 代码，最后转换到需要的格式进行显示或应用。

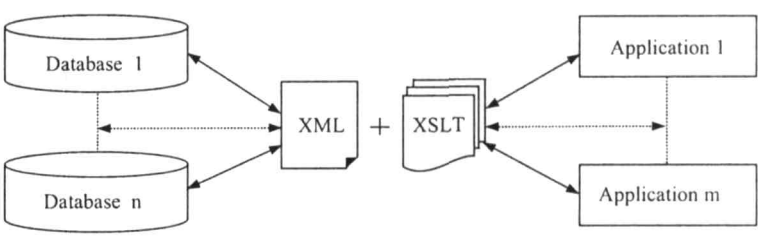

图 3　基于 XML 的三层交换模型

在 XML 三层交换模型、HABML 架构、及 HABML 数据定义的基础上，要在实时海洋数据传输网络系统中实现 XML 数据文档的交换，需要做的就是编写对应于 HABML

的 XML Schema [7~8]，来规范 XML 文档的结构，然后便可按照 XML Schema 规定的格式编写具体的 XML 应用文档。下面给出描述海洋监测数据 XML 结构的 XML Schema 代码片段。

```
<? xml version="1.0" encoding="utf-8" ?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd" elementFormDefault=
 "qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="location">
<xs:complexType>
<xs:sequence>
<xs:element name="station" type="station_brick" />
<xs:element name="spatial" type="spatial_brick" />
<xs:element name="temporal" type="temporal_brick" />
</xs:sequence>
</xs:complexType>
</xs:element>
```

XML Schema 代码编写完成后，便可使用 XML Schema 定义的 XML 文档结构来编写 XML 文档，对海洋数据进行操作，实现海洋数据的存储、移植、发布等功能。下面给出一段基于 XML Schema 定义的描述海洋监测台站属性的 XML 数据文档片段。

```
<station >
<name>ChengShanTou</name>
<abbreviation>CST</abbreviation>
<code>06</code>
<latitude>52.83°</latitude>
<longitude>122.32°</longitude>
</station>
```

4 有害藻华监测数据聚类分析

有害藻华监测数据交换机制设计的一个重要目的就是为有害藻华预警服务。本文在经典的聚类分析算法 FCM (fuzzy c-means) 算法[9~10]的基础上进行了改进，得到了一种更适合有害藻华预警的算法，并将这种算法引入到有害藻华灾害预测中，提高了有害藻华预警的精度。

4.1 经典 FCM 聚类分析算法应用原理

运用经典的 FCM 聚类分析算法预警的原理为：对历史海洋数据样本经进行分簇聚类，得到有害藻华爆发的簇，然后将新海洋数据样本参与聚类，如果大量新海洋数据样本被聚类到有害藻华爆发的簇，则说明有可能将要发生有害藻华灾害。

4.1.1 海洋要素定义

首先，结合 FCM 算法，对海洋要素进行如下定义：

(1) 给定海洋要素数据集 $X=\{x_1,\ x_2,\ \cdots,\ x_n\}\subset R^s$ 为有害藻华监测数据模式空间中 n 个模式的一组有限观测样本集，$x_k=(x_{k1},\ x_{k2},\ \cdots,\ x_{ks})^T\in R^s$ 为观测样本 x_k 的特征矢量，对应特征空间中的一个点，x_{kj} 为特征矢量 x_k 的第 j 维特征上的赋值。$x_{k1},\ x_{k2},\ \cdots,\ x_{ks}$ 为与有害藻华相关 s 个的监测要素。对数据集 X 进行聚类分析就是要产生 X 的 c 划分 ($2\leq c\leq n$)。

(2) 隶属函数 $\mu_{ik}=\mu_{Xi}(x_k)$ 表示样本 x_k 与子集 X_i ($2\leq c\leq n$) 的隶属关系，其中 $\mu_{ik}\in[0,1]$，则 c 划分产生的划分矩阵为 $U=[\mu_{ik}]_{c\times n}$，X 的模糊 c 划空间为：

$$M_{fc}=\{U\in U\in R^{cn}\mid \mu_{ik}\in[0,\ 1],\ \forall i,\ k;\ \sum_{i=1}^{c}\mu_{ik}=1,\ \forall k;\ 0<\sum<\mu_{ik}<n,\ \forall i\} \tag{1}$$

(3) 聚类模式中心集合 P 的定义为，$P=\{p_1,\ p_2,\ \cdots,\ p_c\}$，$p_i(i=1,\ 2,\ \cdots,\ c)$ 表示第 i 类的类中心，即簇心，$p_i\in R^s$，d_{ik} 表示第 i 类中的样本 x_k 与第 i 类样本簇心 P_i 之间的失真度，用两个矢量之间的距离来衡量。FCM 算法的目标函数如下：

$$\begin{cases} J_m(U,\ P)=\sum_{k=1}^{n}\sum_{k=1}^{c}\mu_{ik}^m d_{ik}^2, & m\in[1,\ \infty] \\ s.\ t.\ U\in M_{fc} \end{cases} \tag{2}$$

其中，m 称为加权指数。

(4) 考虑到 d_{ik} 可能为 0，进行了如下定义。对于 $\forall k$ 定义集合 I_k 和 $\bar{I}_k$ 为：

$$I_k=\{i\mid 1\leq i\leq c,\quad d_{ik}=0\} \tag{3}$$

$$\bar{I}_k=\{1,2,\ \cdots,\ c\}-I_k \tag{4}$$

在以上海洋要素定义的基础上，引入 FCM 算法，FCM 算法是基于目标函数的聚类算法，求出使目标函数 J_m 达到最小值的划分 (隶属度) 矩阵 $U=[\mu_{ik}]_{c\times n}$ 与聚类原型 (簇心) $P=\{p_1,\ p_2,\ \cdots,\ p_c\}$。初始时，簇心和隶属度矩阵都是随机产生的，之后不断进行迭代更新。

迭代规则公式如下：

$$\rho_i=\frac{\sum_{k=1}^{n}\mu_{ik}^m x_k}{\sum_{k=1}^{n}\mu_{ik}^m},\ i=1,\ 2,\ \cdots,\ c \tag{5}$$

$$\begin{cases}\mu_{ik}=[\sum_{j=1}^{c}(\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}]^{-1} & \text{当 } I_k=\varphi \\ \mu_{ik}=0,\ \forall i\in\bar{I}_k, \text{ 以及 } \sum_{i\in I_k}\mu_{ik}=1, & \text{当 } I_k\neq\varphi\end{cases} \tag{6}$$

4.1.2 样本聚类方法

给出以上定义之后，便可根据以下步骤进行操作，对海洋数据进行聚类，得到理想的据类结果，分析结果数据，预测有害藻华灾害。

FCM 算法的基本步骤如下：

初始化：给定聚类类别数c，$2\leq c\leq n$，n 是数据样本集 X 中样本个数，设定迭代停止阀值 ε，初始化聚类原型模式 $P^{(0)}$，设置迭代计数器b=0；

步骤 1 ：用式（7）计算或更新划分矩阵$U^{(b)}$；

对于$\forall i,\ k$,如果$\exists d_{ik}^{(b)}>0$则有：

$$\mu_{ik}^{(b)}=\left[\sum_{j=1}^{c}(\frac{d_{ik}^{(b)}}{d_{ik}^{(b)}})^{\frac{2}{m-1}}\right]^{-1} \tag{7}$$

如果$\exists i,\ r$,使得 $d_{ir}^{(b)}=0$ 则有：

$$\mu_{ir}^{b}=1, \text{ 且对} j\neq r,\ \mu_{ij}^{(b)}=0 \tag{8}$$

步骤 2 ：用迭代规则公式(9)更新聚类原型模式矩阵$P^{(b+1)}$；

$$p_i^{(b+1)}=\frac{\sum_{k=1}^{n}(\mu_{ik}^{(b+1)})^m x_k}{\sum_{k=1}^{n}(\mu_{ik}^{(b+1)})^m},\ i=1,2,\ \cdots,\ \mathrm{c} \tag{9}$$

步骤 3 ：如果 $\|P^{(b)}-P^{(b+1)}\|<\varepsilon$，则算法停止并输出划分矩阵$U$和聚类原型$P$，否则$b=b+1$，转向步骤 1。

4.2 改进 FCM 聚类分析算法应用原理

在经典的 FCM 算法中，样本 $x_k=(x_{k1},\ x_{k2},\ \cdots,\ x_{ks})^T\in R^s$ 中不同维上的值 x_{kj}（$1\leq j\leq s$）在整个向量中所占的权重是相同的，$x_{k1},\ x_{k2},\ \cdots,\ x_{ks}$ 分别代表与有害藻华相关的海洋监测要素，这就表示在聚类过程中，不同海洋监测要素对有害藻华灾害预测结果的影响是相同的。但实际的情况不是这样，不同监测要素对有害藻华的影响程度是不同的。所以，直接使用经典的 FCM 算法对海洋监测数据进行聚类会产生误差，为了解决这一问题，本文提出了一种改进的 FCM 聚类分析方法。

首先，分别对样本集中所有特征矢量的各维分量分别进行聚类。从历史数据中取出

包含 m 个样本的样本集 $X=\{x_1, x_2, \cdots, x_m\}\subset Rs$ ($m\geq 100$),这些样本分别取自不同时期,它所对应的时期是否爆发过有害藻华灾害是已知的。对所有样本矢量第 j ($1\leq j\leq s$) 维分量的集合{ $x_{1j}, x_{2j}, \cdots, x_{mj}$}共 m 个元素运用 FCM 算法进行分簇聚类,对于第 j 维分量聚类结束后,有 z_j 个元素 $x_{1j}, x_{2j}, \cdots, x_{zj}$ 被聚类到正确的簇中,那么第 j 维分量的分类正确率为:

$$\eta_j=\frac{z_j}{m} \tag{10}$$

η_j 值越大,第 j 维分量的聚类正确率越高,则第 j 维分量聚类结果与有害藻华灾害的一致性就越强,说明第 j 维海洋监测要素对有害藻华的影响越大。如果 $\eta_j<0.5$, 说明第 j 维要素聚类的错误率高于正确率,即第 j 维监测要素与有害藻华无关或关系甚小,对整体的聚类分析具有增大计算量、影响聚类结果等负面作用,须从特征矢量中剔除第 j 维分量。经过分维聚类、正确率计算和筛选之后,特征向量的维数由 s 维变为 s' 维 ($s'\leq s$)。同时,通过各维分量的正确率推导出第 j 维分量在特征向量各维分量中所占权重比为:

$$\gamma_j=\frac{\eta_j}{\sum_{j=1}^{s'}\eta_j} \tag{11}$$

由此,得到如下新的海洋监测要素的定义:

给定海洋要素数据集 $X'=\{x'_1, x'_2, \cdots, x'_n\}\subset R^{s'}$ 为有害藻华监测数据模式空间中 n 个模式的一组有限观测样本集,$x'_k=(x'_{k1}, x'_{k2}, \cdots, x'_{ks'})^T=(\gamma_1 x_{k1}, \gamma_2 x_{k2}, \cdots, \gamma_{ks'} x_{ks'})^T\in R^{s'}$ 为观测样本 x'_k 的特征矢量,对应特征空间中的一个点,x'_{kj} 为特征矢量 x'_k 的第 j 维特征上的赋值。$x'_{k1}, x'_{k2}, \cdots, x'_{ks'}$ 为与有害藻华相关 s' 个的监测要素。对数据集 X 进行聚类分析就是要产生 X' 的 c 划分 ($2\leq c\leq n$)。

在以上新的海洋监测要素定义的基础上,运用(1)式中的 FCM 算法的公式和聚类方法,便可得到更加准确的聚类分析结果。

4.3 有害藻华聚类分析实验

以我国石城岛和王家岛海域的一组历史海洋数据[11]作为样本,样本中包含未发生有害藻华灾害时和发生有害藻华灾害时的海洋监测数据。将这些样本经过改进后的 FCM 算法分三簇聚类后,加入新的海洋数据样本,再次进行聚类。如果大量新样本聚类到发生有害藻华灾害的簇或在其边缘,说明有可能将要发生有害藻华灾害(见图 4),上面的簇为已经爆发严重的有害藻华灾害的样本簇,中间的簇为爆发状况较轻的有害藻华灾害的样本簇,下面的簇为没有爆发有害藻华灾害的样本簇,方框表示各簇的簇心,三角形表示新样本。该方法形象、直观地反映了海洋监测数据与有害藻华的潜在联系,为有害藻华灾害预警提供了有力支持。

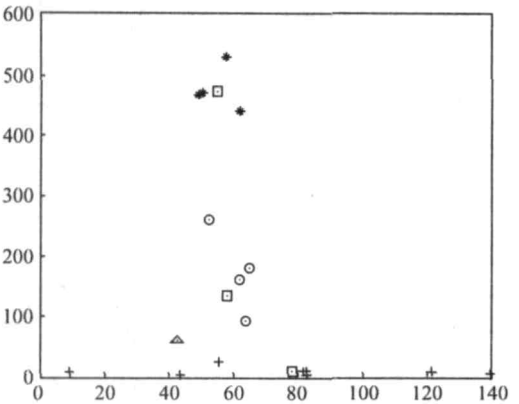

图 4 聚类分析结果

5 结语

针对有害藻华监测数据复杂、多源、跨平台等特点，本文采用具有简单、开放、可扩充、灵活、自描述等特性的 XML 技术，设计了针对有害藻华监测数据交换的 HABML 架构，并投入了使用，运行效果良好，成功地解决了海洋数据封装、存储、交换等问题，并在此基础上将 FCM聚类分析算法引入到有害藻华灾害预警系统中，针对有害藻华的特点对 FCM 算法进行了改进，实验结果证明，改进后的算法能有效、直观的将海洋信息样本进行分类，为有害藻华预警提供了有力支持。

同时，HABML 架构有效地屏蔽了数据来源、数据格式、平台特性等导致数据集成难度大、数据利用率低等根源问题，实现了开放、可靠、高效的数据传输，有极高的应用价值。我国海洋信息业今年来快速发展，XML及聚类分析在海洋信息领域中的应用前景非常广阔，本文为 XML 及聚类分析在海洋领域中的应用提供了有价值的参考，旨在促进 XML 在我国海洋信息领域的应用，促进我国海洋信息业的发展。

参考文献：

[1] 冯士筰，李凤岐，李少菁. 海洋科学导论[M]. 北京：高等教育出版社, 2000.
[2] 王 旭，张占海，吴辉碇. 赤潮的研究和预报[J]. 海洋预报, 2001,18 (1)：65～72.
[3] Anthony W. Isenor. Canadian Investigations of the Keeley Bricks With Application to Profile Data Transfer Using XML [C]. DRDC Atlantic SL 2003～029.
[4] Belinda Ronai, Paul Sliogeris, Matthew de Plater, Krystyna Jankowska. Development and Use of Marine XML within the Australian Oceanographic Data Centre to Encapsulate Marine Data[R]. Australian Oceanographic Data Centre.
[5] Margaret H. Dunham. 数据挖掘教程 [M]. 北京：清华大学出版社, 2005,5：107～110 .
[6] A Vakali, B Catania, A Maddalena. XML data stores: emerging practices[J]. Internet Computing IEEE, 2005,9 (2)：62～69.
[7] J Roy, A Ramanujan. XML schema language：taking XML to the next level[J]. IT Professional, 2001,3(2)：37～40.
[8] S Boucher, R Steinmetz. Embedding XML schema constraints in search-based intersection tests for XPath query optimization[C]. Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on, Aug. 2005，842 ～ 846.
[9] 高新波. 模糊聚类分析及其应用[M]. 西安：西安电子科技大学出版社, 2004.

[10] L Groll, J Jakel. A new convergence proof of fuzzy c-means[J].Fuzzy Systems,IEEE Transactions on, 2005,13(5)：717～720.
[11] 董　婧，刘海映，李培军，等. 海洋岛与王家岛周围海域赤潮生物夜光藻生态初探[J]. 海洋环境科学, 1999,18(4)：48～51.
[12] Dong Qian. A Predictive Model for the Density of Red Tide Alga Noctiluca Scientillans around Shicheng Island and Wangjia Island[J]. Marine Environmental Science, 1999,18 (4)：48～51.

Study on Monitoring Data Exchange Mechanism of HAB and Cluster Method of Disaster Predicting Based on XML

JI Peng, ZHANG Cheng-hui, LI Jun

(School of Control Science and Engineering, Shandong University，Jinan Shandong 250061 China)

Abstract：HAB monitoring data have the features of complexity, multi-driven data sources, different data platforms, nonstandard data formats, etc. XML has the advantages of expansibility, simplicity, openness, interoperability, support for the multinational language, etc. Based on the reasons above, XML is used to build the monitoring data exchange mechanism of HAB. The problems of HAB monitoring data management, transmission and exchange are effectively solved, which lays the foundation for the disaster warning of HAB. Meanwhile, FCM cluster analysis algorithm is improved and introduced into the disaster predicting of HAB, which visually reflects the distribution of samples of marine elements, and this provides a new idea for the disaster predicting.

Keywords：XML；HABML；Harmful Algal Blooms；Cluster Analysis