

Московский государственный университет им. М.В. Ломоносова
Ростовский государственный университет
Институт вычислительной математики РАН

Труды Всероссийской научной конференции

**Научный сервис в сети ИНТЕРНЕТ:
технологии параллельного
программирования**



Московский государственный университет им. М.В.Ломоносова
Ростовский государственный университет
Институт вычислительной математики РАН

Труды Всероссийской научной конференции

**Научный сервис в сети ИНТЕРНЕТ:
технологии параллельного программирования**

г. Новороссийск, 18-23 сентября 2006 г.

Всероссийская научная конференция

Научный сервис в сети Интернет:
технологии параллельного программирования

НАПРАВЛЕНИЕ 1:

**ТЕХНОЛОГИИ ПАРАЛЛЕЛЬНОГО
ПРОГРАММИРОВАНИЯ**

К РЕАЛИЗАЦИИ ИДЕАЛЬНОГО ИНТНОАКТИВНОГО ОПРЕДЕЛИТЕЛЯ БИОЛОГИЧЕСКИХ ОБЪЕКТОВ В ИНТЕРНЕТЕ

А.Л. Лобанов, А.Г. Кирейчук, И.С. Смирнов, О.Н. Граничин, А.Т. Вахитов, М.Б. Дианов

Первые опыты применения ЭВМ для диагностики таксономической принадлежности биологических объектов были осуществлены в середине 60-х годов, когда ЭВМ еще были для биологов экзотической техникой. На заре компьютерной биологической идентификации отставания отечественных исследований от зарубежных как по качеству, так и по количеству диагностических систем практически не было. Позднее за рубежом, благодаря развитию компьютерной индустрии, в этой области произошел всплеск активности американских, английских и австралийских биологов, который привел к созданию сразу нескольких машинных систем, и таким образом у зарубежных коллег появился количественный перевес. Определенной вехой этого можно считать выход в 1975 году сборника "Biological Identification With Computers".

К началу 80-х годов сложилось представление о специфических функциях компьютерных диагностических систем. Их полный набор включал: 1. Предварительную обработку диагностической информации о таксонах. 2. Накопление, хранение и анализ информации на машинных носителях. 3. Диалоговый диагноз с оптимизацией со стороны ЭВМ. 4. Автоматическое составление оптимизированных текстовых определителей. 5. Анализ параметров готовых определителей.

Первая в СССР действующая полная машинная диагностическая система "Диагностика-1" (т.е. выполняющая все перечисленные выше основные специфические функции компьютерных диагностических программ) была создана А.Л.Лобановым в 1974 году на ЭВМ "Наири-2". Затем эта система расширялась, модернизировалась в соответствии с растущими возможностями доступных биологам ЭВМ, - сначала ЕС "Ряд-1", затем БЭСМ-6, СМ-4, СМ-1420 и, наконец, IBM PC. Начиная с пятой версии системы ("Диагностика-5") программы разрабатывались только для IBM PC, они использовали базы данных формата DBF и состояли из модулей, написанных на языке Фортран-88 и на внутреннем языке СУБД FoxPro. С 1992 г. к работе над диагностическими компьютерными системами подключился М.Б.Дианов. Усовершенствованная версия системы "Диагностика-5" получила новое название "ViKey5", а входящая в нее специализированная диалоговая

программа была названа "PicKey". Последний вариант комплексной системы "BiKey7" (1996-1998 гг., языки программирования FoxPro и Fortran) был создан совместно [5].

Разработка последней версии системы (Bikey8b/PicKey8b) закончена в 2005 г. Все блоки системы программировались на языке C++ для Windows и использовали базы данных формата DBF. Программа PicKey8b является одной из лучших в мире в своем классе (интерактивные диагностические программы без специальной ориентации на использование в сети Интернет). С использованием перечисленных программ были созданы определители, включающие до 400 таксонов самых разнообразных организмов (древесные растения, медузы, жуки, офиуры, циклопы, нематоды и др.). О программе PicKey можно узнать на специальном веб-сайте: <http://www.zin.ru/projects/pickey>. Этой программе и созданным с ее помощью определителям посвящены десятки публикаций. Ссылки на них можно получить на упомянутом выше сайте PicKey и на веб-странице: http://www.zin.ru/Animalia/Coleoptera/rus/all_ref.htm.

Эти определители охватывали лишь небольшое число таксонов и работали на отдельных персональных компьютерах. Для более совершенного механизма определения необходимо было создать алгоритм, позволяющий производить диагностику значительно большего числа предполагаемых объектов. С появлением и широким распространением Интернета появилась идея написания программы, которая бы давала возможность дистанционного определения [6].

В ходе подготовки к реализации задуманного, при поддержке РФФИ (грант N 02-07-90105), в течение 2002-2004 гг. был создан электронный Атлас жуков России и сопредельных стран, который уже размещен на сайте Зоологического института Российской академии наук: <http://www.zin.ru/Animalia/Coleoptera/index.htm>. В 2005 г. была получена поддержка РФФИ специально на разработку программного обеспечения к многоходовым политомическим определителям с использованием сети Интернет (грант N 05-07-90179-в). Работа над этим проектом стала возможной благодаря участию специалистов в области сетевых компьютерных систем на основе баз данных - О.Н.Граничина и А.Т.Вахитова. Так родился проект WebKey-X: <http://www.zin.ru/projects/webkey-x/index.html>. Интернет-определитель реализован с использованием технологии Java 2 Enterprise Edition. Приложение использует сервер Tomcat. Для используемых WebKey-X баз данных был разработан единый формат (набор сущностей, атрибутов, связей, таблиц и полей, им соответствующих). Базы данных могут работать под управлением любой SQL-совместимой СУБД. Само приложение состоит из нескольких компонент, реализующих следующие функции: web-интерфейс, управление пользовательской сессией, доступ к базе данных, алгоритм сортировки. Набор данных для конкретного ключа, работающего внутри WebKey-X, включает в себя удовлетворяющую установленному формату базу данных и набор изображений для иллюстрации признаков, состояний и таксонов.

Предусматривается механизм локализации приложения, средствами J2EE. Рассматривается возможность реализации части функций WebKey-X также на основе web-служб, а также распределенную работу с приложением. В целом, Webkey-X в настоящее время должен рассматриваться как базовый инструмент, который может использоваться в более сложных системах биологической идентификации. Для этого мы придерживаемся наибольшей открытости интерфейсов определителя и его модульности. Последнее выражается в легкости замены сортирующего алгоритма (планируется реализация нескольких и обеспечение возможности выбора) и в легкости конфигурации новой базы данных.

Проведенный в 2005 г. обзор существующих диагностических систем, показал наличие довольно развитых программных продуктов, обеспечивающих определение различных групп животных и растений [6]. На основе анализа компьютерных систем была создана таблица, идея и часть содержания которой заимствованы у M. Dallwitz'a: <http://delta-intkey.com/www/comparison.htm>. Таблица сильно модифицирована: в нее добавлены новые характеристики (которые выделены в раздел "New features") и ряд отсутствовавших в ней программ (в том числе наши разработки - PICKEY и WebKey-X, а также программа Д.Дмитриева - 3I): <http://www.zin.ru/projects/webkey-x/index.html>. Таблица активно дополняется оценками для новых характеристик и для добавленных программ. В таблицу включены следующие программы: WebKey-X, PICKEY, 3I (by D.Dmitriev), Flora Search, Intkey, LucID, MEKA, PollyClave, IdentifyIt, Visual Key, NaviKey, XID, DAP, DAWI.

На основе полученных данных шла разработка структуры типового интерактивного определителя (компьютерного ключа) и создание пилотных вариантов различных определителей. Принципиальные свойства интерактивного определителя или компьютерного ключа могут быть условно разделены на 2 группы - структурные и динамические. Структурные свойства - особенности структуры базы данных ключа. Динамические свойства - специфические особенности шага идентификации, представляющего собой итеративно повторяющийся цикл диагноза (под шагом, более точно, понимается рабочий элемент диагноза, обычно включающий выбор признака, ввод в компьютер информации о состоянии признака и получение ответа компьютера с текущим набором таксонов обладающих данным состоянием признака.). Структурные свойства ключа: а) важнейшее свойство - число входов для начала каждого шага диагноза. Существуют одновходовые (пользователю предлагается один признак) и многоходовые (предлагается несколько признаков для выбора наиболее надежного и удобного) ключи; б) число состояний признака - существуют дихотомические (каждый признак имеет только 2 состояния) и политомические (признак включает 3 и более состояний) ключи; в) способность оперировать изображениями признаков и состояний признаков; существуют ключи, управляемые образами (изображения признаков и их состояний - экранные кнопки управления диагнозом) и ключи, управляемые словесными формулировками (используются альтернативные текстовые формулировки). Изображения дают возможность мгновенно понять признак и верно выбрать его состояние; текстовая формулировка требует времени для понимания, сравнения, выбора соответствующего признака и состояния; г)

способность оперировать количественными признаками; возможность фильтровать набор таксонов по диапазону значений количественного признака может существенно сократить путь диагноза [4].

В дальнейшем была предпринята попытка представить себе идеальный компьютерный определитель или "опредетель-идеал", к которому нужно стремиться при разработке подобных систем. На основе сравнения всех свойств компьютерных определителей, созданных к настоящему времени как за рубежом, так и в СНГ, можно смоделировать оптимальный идентификационный ключ. Это интерактивный, многоходовый, политомический, управляемый изображениями ключ со следующими динамическими свойствами: ранжированием признаков на каждом шаге в зависимости от их диагностической ценности; возможностью видеть названия и изображения таксонов текущего набора, а также значения вероятности их идентичности определяемому объекту; возможностью возврата на один или несколько шагов диагноза для коррекции ошибок определяющего; возможностью выбрать несколько признаков в каждом шаге и отметить "невозможные" состояния признаков; активным оперированием как качественными, так и количественными признаками с возможностью использования для последних математических функций (диапазон минимум- максимум, среднее, формулы дискриминантного анализа). В этом направлении и ведутся поисковые работы, так как ни одна из существующих систем пока не обладает полным набором таких свойств [1, 2, 3, 6 и т.д.].

Работа поддержана грантом РФФИ N 05-07-90179-в, а также частично - грантом РФФИ 05-07-90354-в, программой "Биоразнообразие" и проектом "Антарктика" (ФЦП "Мировой океан").

ЛИТЕРАТУРА:

1. Лобанов А.Л. Проблемы создания единой системы диагностической информации в биологии // Единая система и информационно-поисковые языки. Тезисы докладов Всесоюзной научной конференции. Юрмала, 6-8 сентября 1977 г. 1977. С. 84-87.
2. Лобанов А.Л. Принципы построения определителей насекомых с использованием электронных вычислительных машин. - Автореферат диссертации на соискание ученой степени канд. биол. наук. Л.: ЗИН АН СССР, 1983. С. 1-19.
3. Лобанов А.Л., Дианов М.Б. Компьютерная диагностическая система VIKEY и возможности ее использования в защите растений // Защита растений в условиях реформирования агропромышленного комплекса: экономика, эффективность, экологичность. Всеросс. съезд по защите растений. Тезисы докл. 1995. С. 548-549.
4. Лобанов А.Л., Рысс А.Ю. Компьютерные идентификационные системы в зоологии и ботанике: современное состояние и перспективы // Информационно-поисковые системы в зоологии и ботанике. Труды Зоологического института РАН. 1999. Т. 278. Р. 17-19.
5. Лобанов А.Л., Смирнов И.С. Место и роль информационных технологий в исследованиях Зоологического института РАН // Фундаментальные зоологические исследования: Теория и методы. (По материалам Международной конференции «Юбилейные чтения, посвященные 170-летию Зоологического института РАН», проходившей 23-25 октября 2002 г.), М.-СПб.: Товарищество научных изданий КМК. 2004: 283-318 (резюме на англ. яз.).
6. Лобанов А.Л., Кирейчук А.Г., Смирнов И.С., Дианов М.Б., Граничин О.Н. Интернет и интерактивные определители биологических объектов // Научный сервис в сети ИНТЕРНЕТ: Труды Всероссийской научной конференции (19-24 сентября 2005 г., г. Новороссийск). - М.: Изд-во МГУ, 2005. с. 132-134.