ИНФОРМАТИКА

УДК 004.056:004.4'4

А. А. ДЖУСУПОВ, Е. С. ПШЕНИН, Т. В. ГОЛУБЕВА

СТРУКТУРИРОВАНИЕ ПРОСТРАНСТВА ПРИЗНАКОВ ИНФОРМАЦИОННЫХ УГРОЗ ДЛЯ СИНТАКТИКО-СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ В ПСЕВДОЯЗЫКОВОМ КОМПИЛЯТОРЕ

(Представлена академиком НАН РК А. А. Женсыкбаевым)

Введение. В работе предложен псевдоязыковый подход к структурированию пространства признаков информационных воздействий и соответствующих им противодействий. Данный подход основан на создании языка систем информационной защиты (ЯСИЗ) [1] и построения соответствующего компилятора на основе синтактико-семантического анализа текстов по проектированию систем безопасности с целью получения объектного проекта представляющего систему безопасности в виде набора сервисов.

Структурирование пространства признаков в процедурах распознавания информационных угроз. Интеграция вычислительных систем даже в пределах одной организации повышает производительность и эффективность обработки информации, но наряду с достоинствами возникают проблемы, связанные с различными видами воздействий на неё.

Воздействие информационной угрозы – это непосредственный вред, вызываемый угрозой, а методы, средства и действия, применяемые для предотвращения воздействия информационной угрозы, называются противодействием информационной угрозе.

Реальные воздействия информационных угроз, а так же противодействие им отличаются друг от друга какими либо свойствами, но в то же время – многие из них обладают и некоторой общностью, что позволяет их классифицировать по характерным признакам, для определения угрозы и выбора оптимального метода защиты от нее.

Рассмотрим классификацию воздействий и противодействий при информационных угрозах, предложенную Е. С. Пшениным [2].

Итак, информационные воздействия и противодействия им можно разбить на классы, виды, типы, категории, действия, средства, субъекты и объекты.

Классификация субъектов и объектов является единой как для воздействий информационных угроз, так и для противодействий им.

На основе классификации возникает задача распознавания информационных угроз, которую можно рассматривать как частный случай проблемы распознавания образов.

Под образом будем понимать наименование области в пространстве признаков информационных воздействий, в которой отображается множество объектов или явлений.

Рассмотрим классификацию методов защиты информации в зависимости от области их применения, наличия ограничений и недостатков по аналогии с методами применения в распознавании образов у Луценко [3].

Методы распознавания информационных воздействий классифицируем на интенсиональные методы (основанные на операциях с признаками для слабо- и среднезащищенных систем) и экстенсиональные методы (основанные на операциях с объектами для сильнозащищенных систем).

Интенсиональные методы включают в себя 4 направления, это:

– методы, основанные на оценках плотностей распределения значений признаков. Они заимствованы из классической теории статистических решений[4], в которой объекты исследования рассматриваются как реализации многомерной случайной величины, распределенной в пространстве признаков по какому-либо закону [5].

Эта группа методов использует ту или иную интерпретацию формулы условных вероятностей Т. Байеса [6] и применяется для задач с известным распределением (как правило нормальным), что предполагает необходимость набора большой статистики. Недостатками этих методов являются необходимость перебора всей обучающей выборки при распознавании информационных воздействий, а так же высокая чувствительность к

репрезентативности обучающей выборки и артефактам.

– методы, основанные на предположениях о классе решающих функций. В данной группе методов считается известным общий вид уравнения разделяющей поверхности и задан функционал качества разбиения [7].

Самыми распространенными являются представления решающих функций в виде линейных и обобщенных нелинейных полиномов, что позволяет говорить об аналогии этих методов с частными реализациями регрессионного анализа. Функционал качества решающего правила обычно связывают с ошибкой классификации. Классы должны быть хорошо разделяемыми, система признаков – ортонормированной.

Ограничение применения данных методов заключается в том, что должен быть заранее известен вид решающей функции, а так же в невозможности учета новых знаний о корреляциях между признаками.

– логические методы распознавания образов базируются на аппарате булевой алгебры логики и позволяют оперировать информацией, заключенной не только в отдельных признаках, но и в сочетаниях значений признаков [8]. Областью применения данных методов являются задачи небольшой размерности пространства признаков. К недостаткам можно отнести высокую вычислительную трудоемкость, так как при отборе логических решающих правил (конъюнкций) необходим полный перебор.

– лингвистические (структурные) методы распознавания образов основаны на использовании специальных грамматик (т.е. правил построения объектов из «атомарных» элементов), порождающих языки, с помощью которых может описываться совокупность свойств распознаваемых информационных воздействий.

Синтаксические анализаторы, которые представляют полное описание информационного воздействия в виде дерева грамматического разбора, устанавливают его синтаксическую правильность, а именно, может ли фиксированная грамматика, описывающая некоторый класс, породить имеющееся описание информационного воздействия. В противном случае, информационное воздействие либо отклоняется, либо подвергается анализу с помощью других грамматик, описывающих другие классы воздействий.

Данные методы подходят для задач небольшой размерности пространства признаков. Недостаток этой группы методов заключается в том, что задача восстановления (определения) грамматики по некоторому множеству высказываний (описаний воздействий), является трудно формализуемой.

В экстенсиональных методах каждому изучаемому объекту информационных отношений в большей или меньшей мере придается самостоятельное диагностическое значение, при этом роль каждого из них может меняться в самых широких пределах: от главной до весьма косвенного участия в процессе классификации.

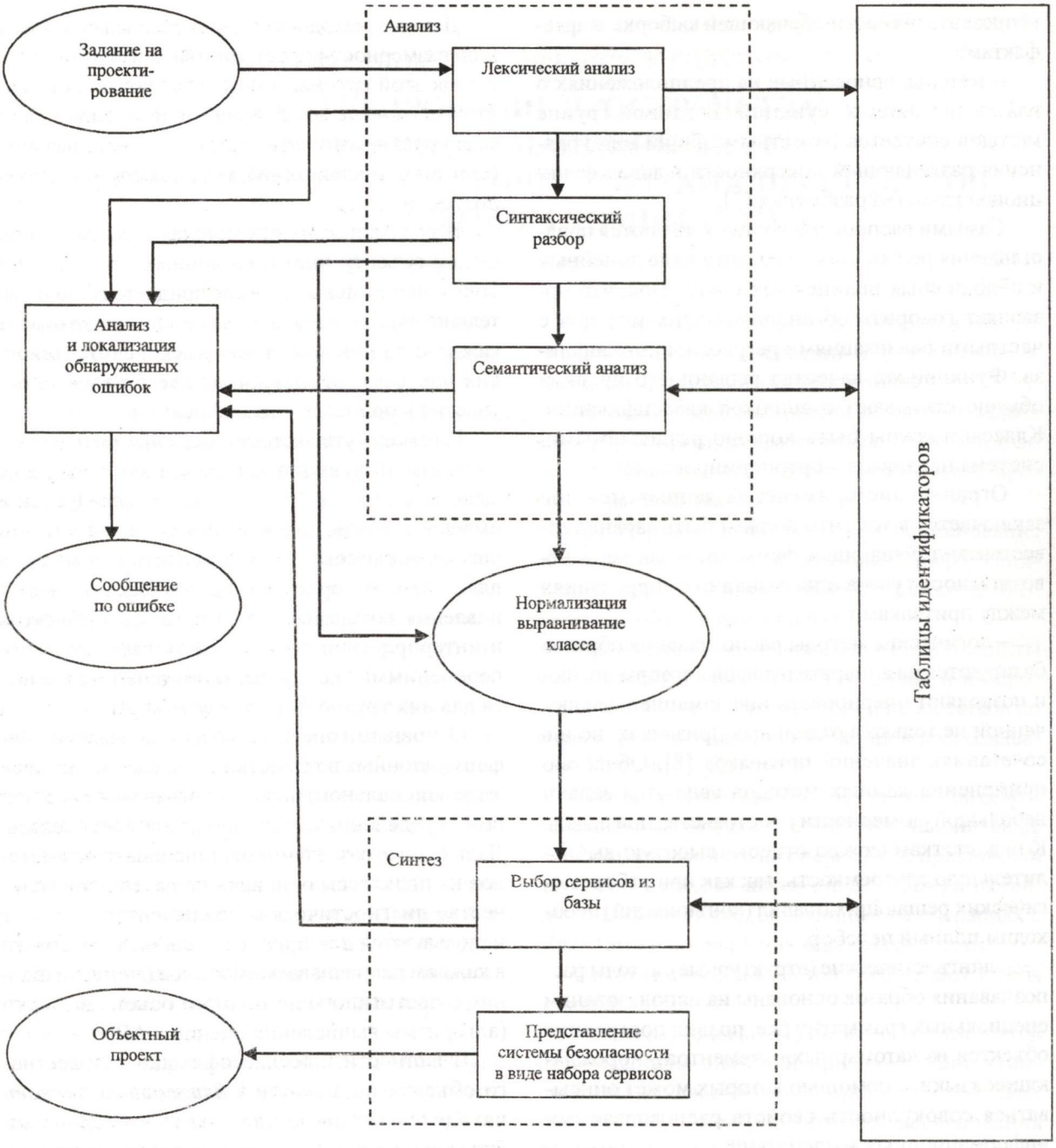
По своей сути экстенсиональные методы рассматривают функции нападения как целостные феномены, каждый из которых индивидуален и имеет особенную диагностическую ценность, что определяет высокую эффективность этих методов для «чистого» прогноза. Однако задача восстановления закономерностей поведения объектов и интерпретации связей между варьируемыми переменными (т.е. функция объяснения) является для них трудно формализуемой.[4]

Основными операциями в распознавании информационных воздействий с помощью методов экстенсионального направления являются операции определения сходства и различия объектов. Дальнейшее разделение экстенсиональных методов на подклассы основано на различии в количестве диагностических прецедентов, которые используются для процесса решения: от одного в каждом распознаваемом классе (метод сравнения с прототипом) до полного объема выборки (алгоритмы вычисления оценок – АВО).

В частности, при классификации неизвестного объекта по *методу k-ближайших соседей* находится заданное число (***k***) геометрически ближайших к нему в пространстве признаков других объектов с уже известной принадлежностью к распознаваемым классам. Дальнейшее решение принимается, например, с помощью простого подсчета голосов.

Рассмотрим более подробно экстенсиональные методы:

– метод сравнения с прототипом применяется для задач небольшой размерности пространства признаков. К недостаткам относится высокая зависимость результатов классификации от меры расстояния (метрики).



Общая схема работы компилятора

– метод *k*-ближайших соседей применим для задач небольшой размерности по количеству классов и признаков. Недостатками являются: высокая зависимость результатов классификации от меры расстояния (метрики), необходимость полного перебора обучающей выборки при распознавании, а так же вычислительная трудоемкость.

– алгоритмы вычисления оценок (голосования) АВО подходят для задач небольшой размерности по количеству классов и признаков. Данный метод обладает высокой технической сложностью. К его недостаткам относятся зависимость результатов классификации от меры расстояния (метрики) и необходимость полного перебора обучающей выборки при распознавании.

– коллективы решающих правил применимы для задач небольшой размерности по количеству классов и признаков. Они обладают очень высокой технической сложностью метода, а также теоретическими проблемами, как при определении областей компетенции частных методов, так и в самих частных методах.

Так как различные алгоритмы распознавания проявляют себя по-разному на одной и той же выборке объектов, то закономерно встает вопрос о синтетическом решающем правиле, адаптивно использующем сильные стороны этих алгоритмов [5]. В коллективах решающих правил применяется двухуровневая схема распознавания. На первом уровне работают частные алгоритмы распознавания, результаты которых объединяются на втором уровне в блоке синтеза. Наиболее распространенные способы такого объединения основаны на выделении «областей компетентности».

Синтактико-семантический анализ текстов в псевдоязыковом компиляторе. Псевдоязыковый компилятор – это программа, которая будет переводить задание на проектирование, поступающее на вход компилятора, в эквивалентный ей выходной объектный проект на основе соответствующего представления системы безопасности в виде набора сервисов.

Процесс компиляции состоит из двух основных этапов – синтеза и анализа (см. рисунок) [9].

На этапе анализа выполняется распознавание текста задания на проектирование, создаются и заполняются таблицы идентификаторов. Результатом работы данного этапа служит некое внутреннее представление понятное компилятору.

На основании внутреннего представления задания на проектирование и информации, содержащейся в таблице или таблицах идентификаторов, **на этапе синтеза** создается текст объектного проекта.

Кроме того, в составе компилятора обязательной является часть, ответственная за анализ и исправление ошибок, которая при наличии ошибки в тексте задания на проектирование должна максимально полно информировать пользователя о типе ошибки и месте ее возникновения (лучше если компилятор предложит пользователю вариант исправления ошибки).

Эти этапы, в свою очередь, состоят из более мелких этапов, называемых фазами компиляции.

С точки зрения теории формальных языков, компилятор выполняет две основные функции:

1) распознавание языка исходной программы. То есть псевдоязыковый компилятор получает на вход цепочку символов входного языка (генерируемую автором задания на проектирование), проверяет ее принадлежность языку, а так же выявляет правила, по которым эта цепочка была построена;

2) компилятор является генератором для результирующего объектного проекта. На выходе компилятора, мы получаем цепочку символов выходного языка, построенную по определенным правилам, соответствующим представлению системы безопасности в виде набора сервисов.

Рассмотрим подробно этап анализа псевдоязыкового компилятора, который будет состоять из 3-х частей:

Лексический анализатор или сканер – это часть компилятора, которая читает посимвольно задание на проектирование на исходном языке и строит из литер слова (лексемы) исходного языка. На вход лексического анализатора поступает исходный текст задания на проектирование, а выходная информация передается далее для обработки компилятором на этапе синтаксического разбора. Теоретически, лексический анализатор не является обязательной частью компилятора, но, в случае реализации псевдоязыкового компилятора, он позволит избавить сложный по структуре синтаксический анализатор от решения примитивных задач.

Синтаксический разбор – это основная часть анализатора компилятора, которая выполняет главную функцию – распознавание текста входного задания на проектирование. Она должна будет выполнять выделение синтаксических конструкций в тексте задания на проектирование, прошедшего лексический анализ. На этой же фазе компиляции будет проверяться синтаксическая правильность задания на проектирование.

Семантический анализ – это часть компилятора, ответственная за проверку правильности исходного текста с точки зрения семантики входного задания на проектирование. Кроме того, семантический анализатор должен выполнять преобразование текста, требуемое семантикой входного языка на основе классификации воздействий при информационных угрозах [2] (например, добавление функций неявного преобразования типов воздействий) и требуемого класса безопасности.

В задачу синтаксического анализатора входит: найти и выделить основные синтаксические конструкции в тексте задания на проектирование, установить тип и проверить правильность каждой синтаксической конструкции в виде, удобном для дальнейшей нормализации и выравнивания класса, а так же для осуществляемых на этапе синтеза выбора сервисов из базы и представления системы безопасности в виде набора сервисов с целью получения результирующего объектного проекта.

В основе синтаксического анализатора будет лежать распознаватель текста задания на проектирование на основе грамматики входного языка, основанной на совокупности баз реляционного типа, состоящих из записей, ключевые элементы которых обозначим через представление Дьюи. Синтаксические конструкции ЯСИЗ должны быть описаны с помощью КС-грамматик(контекстно-свободные грамматики), на основе синтаксиса которых должен быть построен КС-язык.

Выходом лексического анализатора является таблица лексем (или цепочка). Она образует вход синтаксического анализатора и устанавливает, удовлетворяет ли она структурным условиям, явно сформулированным в определении синтаксиса задачи проектирования систем информационной безопасности [1].

Грамматика ЯСИЗ должна содержать правила 2-х типов: первые, определяющие синтаксические конструкции языка, которые легко поддаются формальному описанию. Вторые, определяющие семантические ограничения языка, должны излагаться в неформальной форме. Поэтому должны в начале формально излагаться правила построения синтаксических конструкций, а потом на естественном языке даваться описание семантических правил для пользователя, а для компилятора семантические ограничения необходимо излагать в виде алгоритмов проверки правильности задания на проектирование (семантические ограничения на исходный текст). Такой проверкой в псевдоязыковом компиляторе должен заниматься семантический анализатор – специально для этого разрабатываемая часть компилятора.

Для определения грамматики псевдоязыка допустимо использовать формальное описание грамматики, построенное на основе системы правил, которые классифицируем как контекстно-свободную грамматику.

Итак, основу синтаксического анализатора псевдоязыкового компилятора, составит распознаватель, построенный на основе КС-грамматик. Главную роль в том, как функционирует синтаксический анализатор и какой алгоритм лежит в его основе, играют принципы построения распознавателей КС-языков, без них невозможно выполнить эффективный синтаксический разбор предложений входного языка.

Таким образом, псевдоязыковый компилятор ЯСИЗ на основе методов распознавания информационных угроз, в соответствии с классификацией информационных воздействий должен производить выбор оптимальных противодействий в зависимости от необходимого уровня защищенности информационной системы.

ЛИТЕРАТУРА

1. *Голубева Т.В.* Языковый подход в проектировании систем защиты информации: Труды Республиканской научной конференции «Молодые ученые – будущее науки». Ч. 2. Алматы: КазНТУ, 2004. С. 94-99.

2. *Пшенин Е.С.* Теоретические основы защиты информации: Учебное пособие. Алматы: Каз НТУ, 2000. 125 с. ISBN 9965-487-36-7.

3. *Луценко Е.В.* Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (на примере универсальной автоматизированной системы распознавания образов «ЭЙДОС-5.1»). Краснодар: КЮИ МВД РФ, 1996. 280 с.

4. *Кендалл М.Дж., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.

5. *Горелик А.Л., Скрипкин В.А.* Методы распознавания. Учебное пособие для вузов. М.: Высш. школа, 1977. 222 с.

6. *Боровков А.А.* Теория вероятностей. М.: УРСС, 2003. 472 с.

7. *Аркадьев А.Г., Браверман Э.М.* Обучение машины классификации объектов. М.: Наука, 1971.

8. *Горелик А.Л., Гуревич И.Б., Скрипкин В.А.* Современное состояние проблемы распознавания. М.: Радио и связь, 1985. 160 с.

9. *Грис Д.* Конструирование компиляторов для цифровых вычислительных машин. М.: Мир, 1975. 544 с.

Резюме

Ақпараттық қауіпсіздік жүйесін жобалаудағы жалғантілдік қадамның авторлармен өңделуі қарастырылған.

Summary

Pseudo language approach to the problem of designing the information protect systems, projected by the authors, is considered in this article.

КазНТУ им. К.И. Сатпаева
г.Алматы

Поступила 19.04.07г.