# СЕМАНТИЧЕСКИЙ СПОСОБ ПОИСКА ИНФОРМАЦИОННЫХ АНОМАЛИЙ ЧЕРЕЗ WEB

С. В. Перминов[1], С. В. Афанасьев[2]

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

[1]<sv_perminov@mail.ru>, [2]<serg@sqlab.nw.ru>

УДК 681.3

*Перминов С. В., Афанасьев С. В.* **Семантический способ поиска информационных аномалий через web** // Труды СПИИРАН. Вып. 3, т. 1. — СПб.: Наука, 2006.

**Аннотация.** *В статье описывается подход к семантическому поиску информационных аномалий. Используются технологии семантического web, предложенные World Wide Web Consortium (W3C), а именно языки для описания семантических данных RDF и OWL, а также язык формальных поисковых запросов к семантическим данным SPARQL. Реализован прототип системы, где в качестве примера описаны семантические данные и запросы для поиска класса информационных аномалий. Приведен обзор существующих поисковых систем.* — Библ. 8 назв.

UDC 681.3

*Perminov S. V., Afanasyev S. V.* **Semantic Search for Information Anomalies through Web Technique** // SPIIRAS Proceedings. Issue 3, vol. 1. — SPb.: Nauka, 2006.

**Abstract.** *Semantic search for information anomalies technique is described. It uses semantic web technologies, recommended by World Wide Web Consortium (W3C) — RDF and OWL languages for semantic data description, SPARQL language for making search queries in semantic data. Prototype of system is implemented; it contains semantic data and queries for searching class of information anomalies. Current search systems are reviewed.* — Bibl. 8 items.

## 1. Введение

Под аномалией мы понимаем отклонение от нормы. Поиск аномалии сводится к нахождению совокупности данных, свидетельствующих о том, что существует отклонение от нормы.

Рассмотрим применимость существующих поисковых технологий в web для обнаружения информационных аномалий. Поиск информации в глобальной сети в том виде, в котором он существует сейчас, не является совершенным. Можно выделить, по крайней мере, три распространенные проблемы поисковых машин:

1) нерелевантность поиска [1];

2) большое разнообразие способов представления и организации данных в глобальной сети web;

3) неполнота поиска [1].

Нерелевантность можно определить как несоответствие результата поиска ожиданиям пользователя. При поиске информационных аномалий это может определяться непредставительной совокупностью данных, выбранной пользователем для описания информационной аномалии, что тесно связано с проблемой отсутствия единообразия данных в web.

Исторические корни последней проблемы связаны с особенностями развития HTML как языка web-документов. Будучи изначально предназначен для описания структуры документа, с развитием WWW в него было введено много элементов представления (например, тэги «b» для полужирного начертания шрифта и «i» для курсива), возникла практика использования элементов языка



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

не по своему прямому назначению (например, создание HTML-таблиц для формирования графической разметки страницы) [2]. Многообразие, к которому это привело, стало причиной трудноразрешимых проблем формального анализа семантической структуры web-документов, поэтому большинство современных поисковых систем в web ограничивают свои задачи поиском совпадений текста по ключевым словам, не акцентируя внимания на структуре.

Неполнота поиска информационных аномалий связана с возможностью различным образом представить одно и то же отклонение от нормы, другими словами возможностью описать семантику аномалии множеством синтаксических конструкций.

В синтаксических системах наиболее распространен поиск по ключевым словам. К менее распространенным видам можно отнести поиск по рубрикатору (например, в системах Yahoo и Rambler), по естественно-языковому запросу (в энциклопедии Britannica) и по образцу текста (например, в Infoseek). Другой способ поиска связан с использованием онтологии, вариант реализации предложен в проекте TAP Стэндфордского университета. У каждого из этих способов есть свои преимущества и недостатки, однако объединяет их синтаксическое поисковое ядро, что не позволяет назвать поиск, осуществляемый в этих системах, смысловым.

Для того, чтобы осуществить попытку смыслового поиска в существующих web-документах, необходим предварительный семантический анализ их содержимого, его основная цель — улучшить структурирование информации в документах [3]. Семантический анализ развивает идею распознавания образов, другими словами чрезвычайно сжатых представлений документов, в которых не учитывается ни их конкретное содержание, ни лексика. Можно сказать, что семантический поиск — это поиск по ключевым понятиям, а семантическое представление документа — это множество присутствующих в нем понятий или семантических категорий.

При переходе к семантическому описанию документов происходит сжатие и обобщение информации, что приводит к новому знанию. В то же время семантическое представление информации невозможно без определения закономерностей совместного употребления слов, где значения каждого слова определяется контекстом его использования, т.е. множеством других слов. Проблема вычислимости здесь связана с тем, что значение каждого из слов этого множества в свою очередь определяется собственным контекстом, состоящем из слов, в число которых может попасть и исходное.

Один из вариантов решения этой проблемы — построение математической модели языка. Тогда любое слово в русском языке можно рассматривать как имя функции, где последняя в качестве результата возвращает его семантику [4]. При этом конкретное значение слово получит только после подстановки аргументов (это не обязательно контекст слова), а его смысл будет вычислен по мере выполнения функции. Предложение в данном случае — это законченная суперпозиция функций, а смысл предложения вычисляется при построении и выполнении этой суперпозиции. Такой подход к семантическому анализу позволяет в том числе построить онтологии, описывающие предложения. Кроме того, онтологии, как и тезаурус можно использовать еще на этапе семантического анализа. В этом случае наиболее эффективно их совместное использование, где онтология описывает комплекс понятий и отношений предметной области, а тезаурус формирует подобную систему понятий и отношений, но в рамках лингвистических знаний по предметной области [5].



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Описанные выше методики можно обозначить, как семантический анализ неструктурированной информации, к которой на сегодняшний день можно отнести и множество документов на языке HTML. Эти исследования, в основном, делают попытку решения трудных для формализации задач.

В статье предлагается иной подход к семантическому поиску, который предполагает наличие как предварительно формализованных семантических данных для поиска, так и множества формальных поисковых запросов. При этом рассматривается только поиск информационных аномалий, специализированная область применения дает возможность относительно объективно оценить эффективность предложенного способа поиска. Этот подход имеет общие черты с технологией «Интегрированная система информационных ресурсов» [6], в отличие от нее требованием предлагаемого подхода является предопределение формальных поисковых запросов.

## 2. Семантический способ поиска

Рассмотрим вначале элементы, необходимые для построения универсальной системы семантического поиска, а затем опишем совокупность данных и поискового запроса, при помощи которых можно осуществлять семантический поиск класса информационных аномалий. Все вместе это составляет способ семантического поиска требуемых к обнаружению информационных аномалий в требуемых информационных системах при условии, что составлены соответствующие данные и поисковые запросы.

Рассмотрим общую схему работы поисковой системы, не зависящую от ее типа (см. рис. 1). Серым цветом обозначены элементы, без которых система поиска не может считаться законченной. Важное отличие между системами синтаксического и семантического поиска заключается в формате или форматах данных, с которыми они работают. Если первые осуществляют поисковые запросы к неструктурированным или слабоструктурированным данным, то последние в конечном итоге ищут информацию в семантически структурированной информации (даже если она была получена путем анализа неструктурированных данных). Рассмотрим структуру семантических данных, пригодную для универсальной системы семантического поиска.

### 2.1. Структура семантических данных

Структуру семантических данных можно представить как сеть бинарных схем, где бинарная схема — это сеть бинарных отношений. Граф, где узлы отображают понятия, а дуга — отношение между ними, является формальным обозначением бинарного отношения. Бинарные отношения объединяются в сеть при появлении понятий, которые участвуют в нескольких отношениях.

Создадим по одному бинарному отношению для двух файлов. Под файлом мы будем понимать логический информационный блок, который хранится на носителе информации. Таким образом, если мы будем рассматривать файлы, как конкретные экземпляры обобщенного понятия «файл», у нас появляется возможность объединить бинарные отношения в сеть (см. рис. 2). Данную сеть (схему) можно формально описать на любом языке, в котором есть поддержка бинарных отношений. Понятие «экземпляр» и отношение «тип», которые в ней использованы, являются универсальными по своей природе, для



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

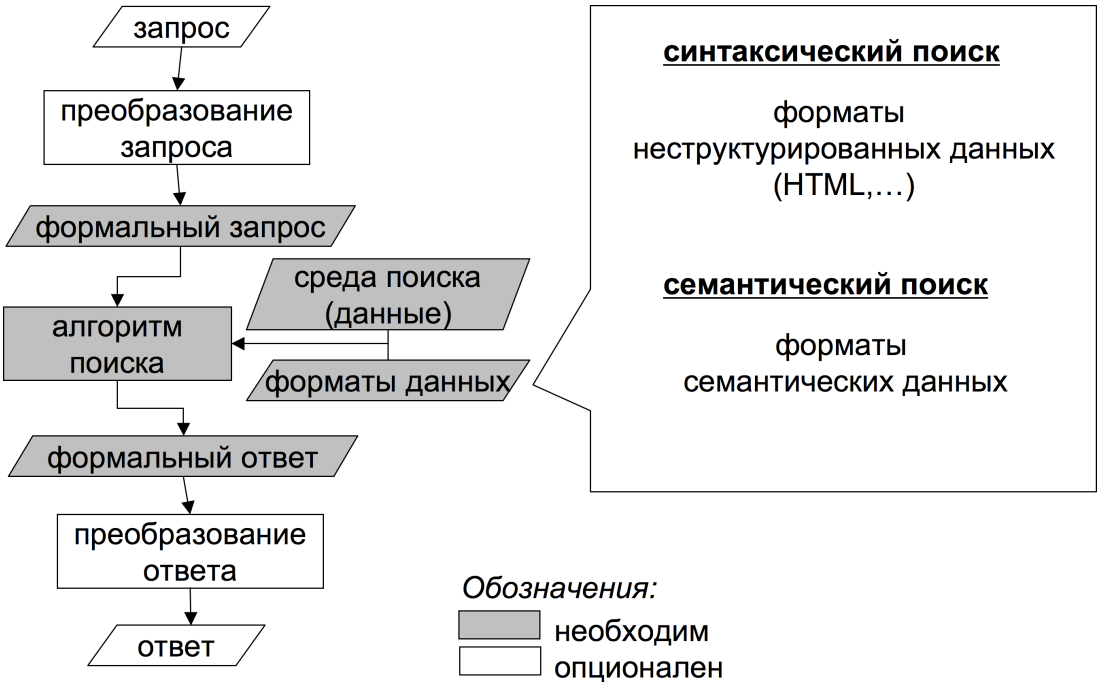


Рис. 1. Универсальная схема работы поисковой системы.

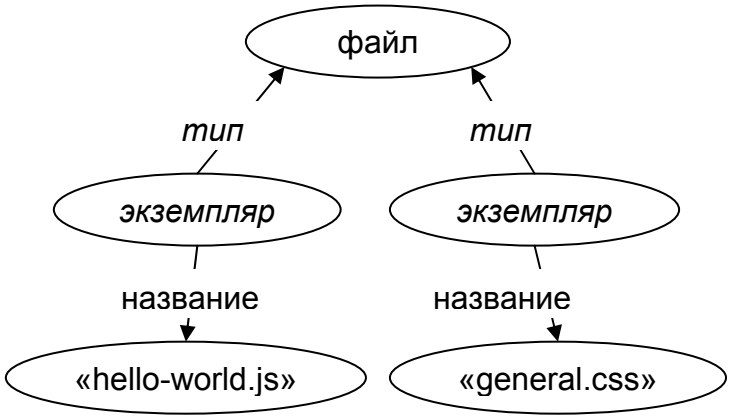


Рис. 2. Бинарная схема 1.

устранения неопределенности в их интерпретации они могут быть встроены в формальный язык.

Если описать данную бинарную схему на формальном языке, она может быть доступна в среде web по некоторому URL-адресу, в данном случае ее можно назвать семантическим web-ресурсом. Допустим, что так оно и есть, а также то, что по другому URL-адресу доступна бинарная схема, изображенная на рис. 3.

Поставим задачу каким-либо образом связать информацию, представленную в двух схемах. Это возможно, если разные ресурсы описывают схожие сущности, даже если при этом сущности обозначаются другими понятиями, которые участвуют в других бинарных отношениях. В построенных схемах иллюстрацией к этому являются понятия «файл» и «ресурс данных», которые в данном случае представляют одну и ту же сущность. При объединении двух схем возникнет конфликт имен. Дело в том, что в первой схеме слово «название» используется для обозначения отношения, связывающего экземпляр понятия «файл» со значением его названия, во второй тем же словом обозначено от-



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ношение экземпляра схожего понятия «ресурс данных» с описанием класса его содержимого. Для того, чтобы устранить конфликт этих имен, необходимо заменить их на уникальные идентификаторы. Уникальность должна соблюдаться в рамках рассматриваемого множества схем, в нашем примере их две. В рамках сети web потребуется глобальная уникальность идентификаторов, здесь возможно использование стандарта URI (или его подмножества URL) для их написания.

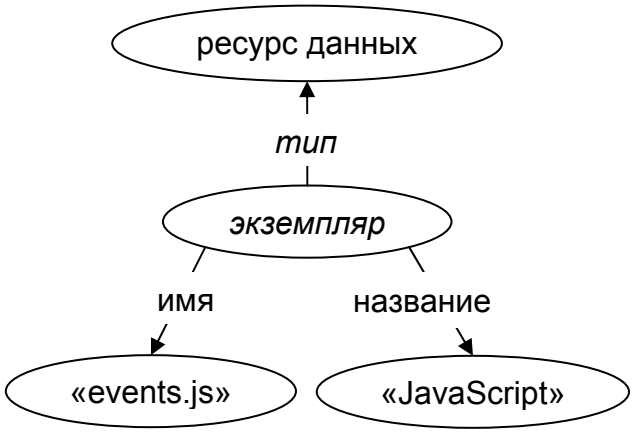


Рис. 3. Бинарная схема 2.

Для того чтобы связать две бинарные схемы, необходимо создать хотя бы одно бинарное отношение между понятиями (но не значениями) из первой схемы и понятиями из второй схемы. Создадим отношение эквивалентности между понятием, обозначенным словом «файл» из первой схемы и понятием, обозначенным словом «ресурс данных» из второй схемы. Таким образом, мы получили сеть бинарных схем, для которой понятия «файл» и «ресурс данных», описанные на разных web-ресурсах, обозначают одну и ту же сущность. Отношение эквивалентности универсально по своей природе, оно может быть встроено в формальный язык. Обычно язык, поддерживающий такие отношения, называется языком описания онтологий. Примерами подобных отношений являются эквивалентность, противоположность, надкласс, подкласс.

Одно из распространенных определений онтологии — «точная спецификация концептуализации». Другими словами, онтология — это множество объектов (терминов) и отношений между ними. Структурная иерархия и объектно-ориентированная классификация — это тоже онтологии, но достаточно простые, обладающие ограниченными возможностями. Например, в объектно-ориентированном подходе существует понятие дочернего и родительского класса, однако нет понятий класса-синонима (отношение эквивалентности) и класса-антонима (отношение противоположности).

В противовес предыдущему определению одним из формальных является представление онтологии в виде четверки:

$$O = (E, D, R, P)$$

где $E$ — множество сущностей (состоит из терминов, объектов, классов, отношений, функций);

$D$ — множество определений сущностей;

$R$ — множество отношений между сущностями;

$P$ — множество правил использования сущностей [5].



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

## 2.2. Форматы семантических данных

Мы используем 2 связанных между собой формата, которые удовлетворяют описанной структуре семантических данных и предназначены для использования в web. Они описаны в пирамиде семантического web (см. рис. 4), предложенной организацией World Wide Web Consortium (сокр. W3C). Формат описания web-ресурсов Resource Description Framework (сокр. RDF) третьего уровня пирамиды позволяет описать множество бинарных схем. Формат описания онтологий Web Ontology Language (сокр. OWL) четвертого уровня позволяет связать различные бинарные схемы в единую сеть [7]. Таким образом, формальная структура данных для разрабатываемой системы определяется форматами RDF и OWL, их совместное использование позволяет построить распределенную модель данных, где блоки RDF-данных умеют узкую информационную специализацию, а OWL-блоки объединяют и систематизируют множество данных RDF. Такая схема имеет общие черты с моделью распределенных поисковых систем [8].

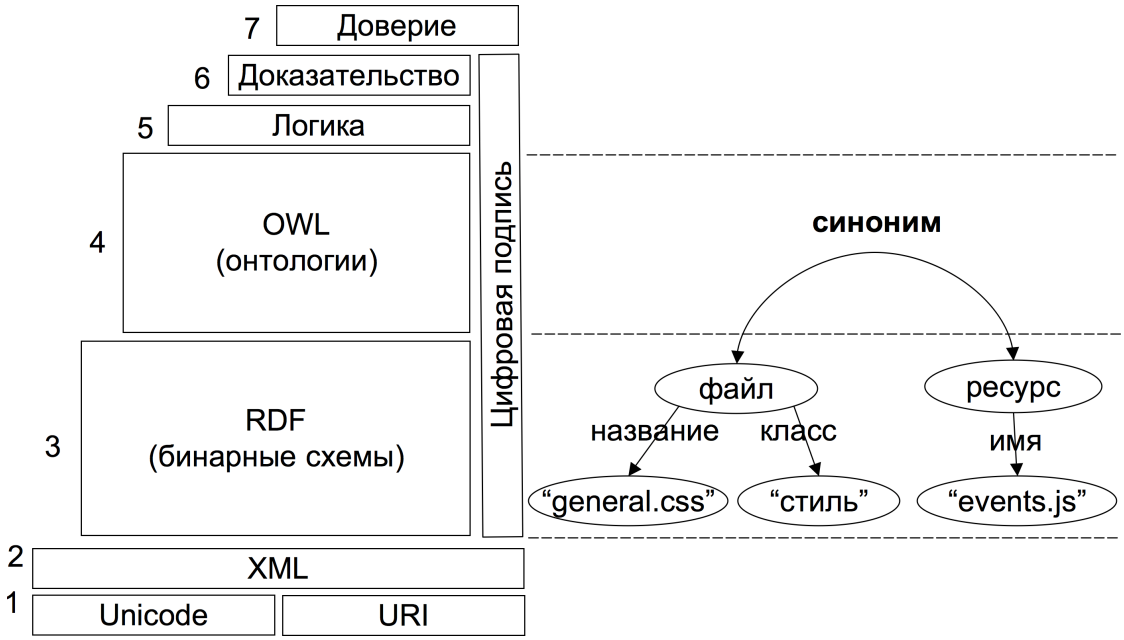


Рис. 4. Форматы семантических данных.

Спецификации стандарта RDF состоят из нескольких частей и предназначены для семантического описания любого ресурса в сети. Конкретные описания сайтов осуществляются с помощью ранее представленных бинарных схем, которые проецируются в язык RDF как логические тройки, каждая из которых состоит из субъекта, предиката и объекта. Набор используемых в описании субъектов, предикатов и объектов может быть описан в языке RDFS, другой части стандарта RDF. RDFS также позволяет описать простую онтологию на базе этого набора данных и встроенных в язык отношений: класс, подкласс, свойство, подсвойство и др.

Однако при построении сложных онтологий предпочтительно использования языка OWL, который является синтаксическим и семантическим расширением языка RDFS, что связано со следующими ограничениями RDFS:

1. Невозможно ввести непересекающиеся классы. Другими словами, если мы скажем, что класс Мужчина и класс Женщина являются подклассами класса



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Человек, мы не можем сказать, что классы Мужчина и Женщина не пересекаются.

2. Невозможно описать ограничение на некоторое условие. К примеру, если мы утверждаем, что корова — животное, а также то, что животные едят мясо, мы не можем сказать, что корова не ест мясо;

3. Невозможно описывать классы как логическую комбинацию других классов. Мы не можем сказать, что класс Человек является логической суммой классов Мужчина и Женщина;

4. Нельзя описать ограничения на значения свойств классов. Например, нельзя сказать, что у человека может быть только двое родителей, или, что учебный курс должен вести по крайней мере один лектор;

5. Нельзя описать свойство свойства, такое как транзитивность, уникальность или, например, свойство, обратное другому свойству.

## 3. Реализация

Мы берем за основу схему работы поисковой системы, изображенную на рис. 1. Данные среды поиска описываются на языках RDF и OWL, поисковые запросы — на формальном языке запросов SPARQL, который наследует конструкции SQL — SELECT, FROM, WHERE, процедура поиска осуществляется при помощи процессора SPARQL (см. рис. 5). Таким образом, совокупность собранных вместе элементов позволяет искать информацию в предварительно подготовленных, формализованных на языках RDF и OWL семантических данных. В глобальной сети web есть незначительный объем таких данных, мы надеемся, что он будет увеличиваться.

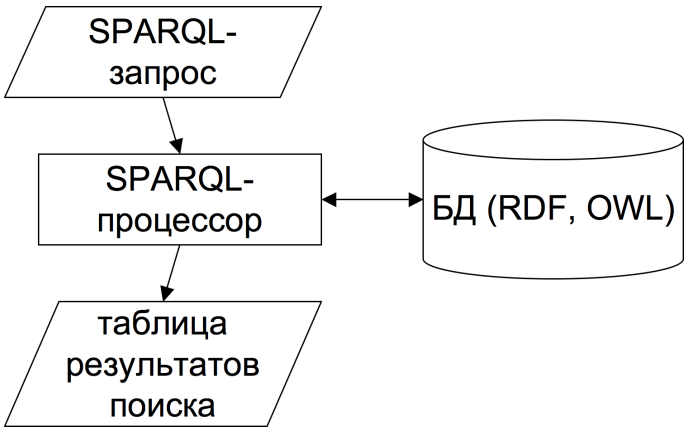


Рис. 5. Схема работы системы семантического поиска.

Опишем семантические данные для поиска класса аномалий. Как это ни парадоксально, семантические данные описывают не сами аномалии, а информационную систему, в которой они потенциально могут присутствовать. В прототипе семантические данные описывают состояние информационной системы типа сайт в текущий момент времени (см. рис. 6).

Здесь описаны данные, преобразование и представление системы, состоящие из файлов. Для каждого из файлов описана оригинальная и текущая контрольная сумма. Информационная аномалия, которую требуется найти — искажение файлов. Можно описать признак этой аномалии в подготовленных семантических данных — несовпадение оригинальной и текущей контрольной



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

суммы. Для формального поиска этой аномалии можно описать запрос на языке SPARQL, что и было сделано. Так как семантическая сеть описывает текущий момент времени, ее необходимо обновлять. Этой цели служит небольшая программа, которая с периодичностью в 60 секунд сканирует список и содержимое файлов, составляющих информационную систему. Для тестирования прототипа был реализован контролируемый источник информационных аномалий, а именно программа, которая скрывает данные в файлах XML методом стеганографии и тем самым искажает их.

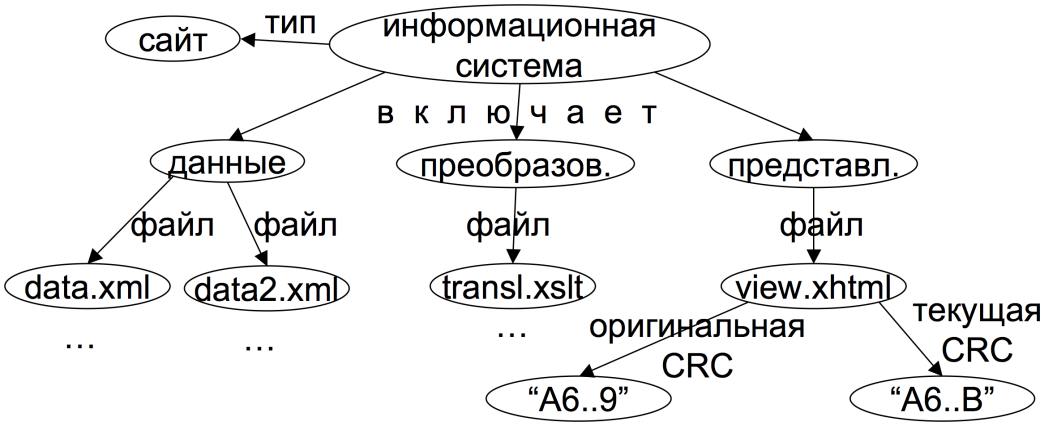


Рис. 6. Семантические данные, описывающие состояние информационной системы.

Web-интерфейс и пример работы прототипа — на рис. 7, 8.

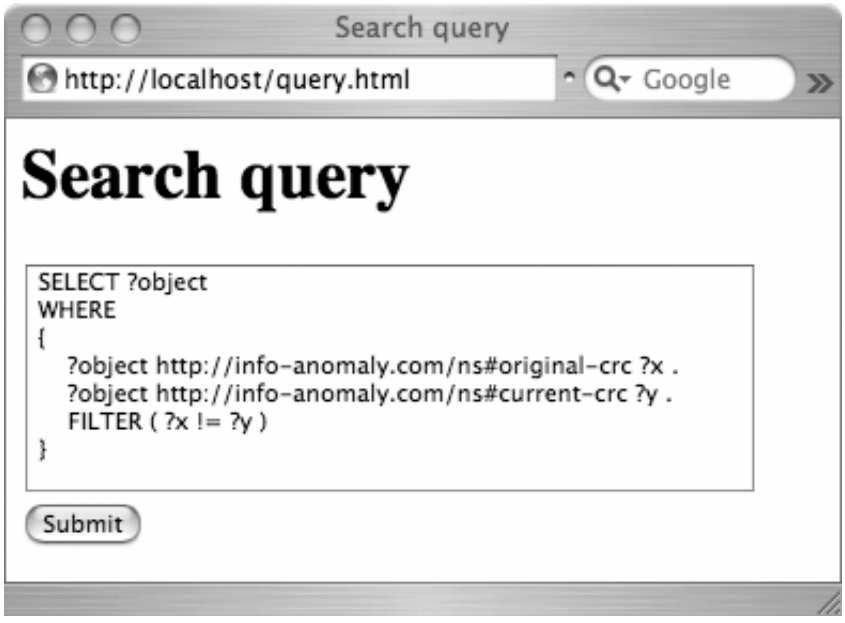


Рис 7. Окно ввода поискового запроса.

На рис. 7 показано окно, позволяющее ввести запрос на языке SPARQL. В данном случае в запросе описано требование найти все объекты, оригинальная и текущая сумма которых не совпадает. Рис. 8 отображает результат поиска, в данном случае это файлы с названиями «data2.xml» и «view.xml». Так как предварительно в эти файлы были внесены умышленные искажения методом стеганографии, можно говорить, что данные результаты поиска соответствуют действительности.



Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

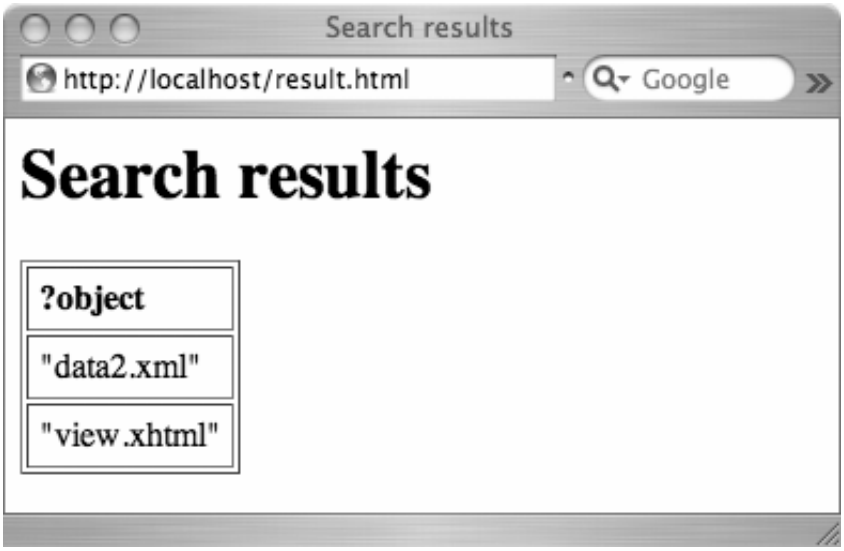


Рис. 8. Окно результата поиска.

## 4. Заключение

Для осуществления эффективного поиска синтаксический метод не всегда эффективен. В настоящей статье предложен способ семантического поиска на примере информационных аномалий. Способ использует технологии RDF, OWL и SPARQL и дополняет идеи, предложенные в концепции семантического web, подходом к поиску информации в предопределенных сетях бинарных отношений.

В прототипе описаны семантические данные для поиска класса информационных аномалий. Использование открытых стандартов и универсальность системы позволяет осуществлять поиск в разработанных сторонними организациями семантических данных, если они доступны (например, через web) и формализованы при помощи языков RDF и OWL.

### Литература


1. *Поляков В. Н.* Интеллектуальная поисковая машина [Электронный ресурс] // <http://www.geocities.com/SiliconValley/Campus/7926/Polyakov/IntelSE.htm> (по состоянию на 21.03.2006).
2. *Meyer E.* Cascading Style Sheets, 2nd edition. O'Reilly, 2004. 528 p.
3. *Шумский С. Я.* Интернет разумный // Открытые системы. 2001. № 3. С. 43–46.
4. *Тузов В. А.* Компьютерная семантика русского языка СПб.: Изд-во С.-Петерб. ун-та, 2004. 400 с.
5. *Нариньяни А. С.* Кентавр по имени ТЕОН: Теазаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Аксаково, 2001. Т. 1. С. 184–188.
6. *Серебряков В. А.* Интегрированная система информационных ресурсов. Архитектура, реализация, приложения. М.: Вычислительный центр РАН им. А. Дородницына, 2004. 240 с.
7. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities // Scientific American. 2001. No. 5. P. 34–43.
8. *В. А. Жигалов и др.* Предел однородности поиска в Интернете // Системная информатика. Вып. 8. Новосибирск: Наука, 2002. С. 29–71.




Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru