ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 49–57

ОСОБЕННОСТИ СЕМАНТИЧЕСКОГО ПОИСКА ИНФОРМАЦИОННЫХ ОБЪЕКТОВ НА ОСНОВЕ ТЕХНОЛОГИИ БАЗ ЗНАНИЙ

М. М. Шарнин[1], И. П. Кузнецов[2]

Аннотация: Рассматривается система семантического поиска информации в больших массивах документов на естественном языке (ЕЯ). Поиск основан на использовании лингвистического процессора, обеспечивающего автоматическое выделение из текстов информационных объектов (именованных сущностей), их признаков, связей и участие в действиях. В результате формируются структуры знаний. Аналогичным образом формируется структура запроса. Поиск, называемый семантическим, обеспечивается за счет сопоставления таких структур, где учитываются связи объектов, а также их участие в событиях, действиях.

Ключевые слова: семантический поиск; семантико-ориентированный лингвистический процессор; извлечение знаний из текстов; база знаний

1 Введение

Одной из актуальных задач в области информационных технологий является поиск информации в больших массивах документов — текстов на *естественном языке*. Для многих профессиональных пользователей поиск определяется их задачами. Например, задачи следователей-аналитиков (из области «Криминалистика») непосредственно связаны с поиском фигурантов, их адресов, деяний, связей между фигурантами, поиском по приметам, поиском похожих фигурантов и происшествий и многим другим. Для поиска используются документы криминальной полиции, имеющие вид текстов на ЕЯ: сводки происшествий и др.

Другой пример — задачи кадровых агентств, где документами являются резюме людей, желающих получить работу. Такие резюме часто пишутся в свободной форме — в виде текстов на ЕЯ. В резюме даются анкетные данные, места учебы и работы с указанием периодов и организаций или учебных заведений и т. д. Задачи кадровых агентств — поиск лиц по запросам клиентов, которые часто задаются на ЕЯ.

Следует отметить, что профессиональных пользователей интересует определенного сорта информация, которая зависит от предметной области. В приведенных выше примерах это лица, где они работают (организации), кем (профессии), чем занимаются (служебные обязанности) или в каких событиях участвовали (деяния лиц) и т. д. Подобную информацию будем называть *информационными объектами* или просто *объектами* (другое название — *именованные сущности*).

Поиск информационных объектов — это самостоятельная задача. Типовые поисковые машины (Google, Яндекс и др.) ищут ресурсы, содержащие слова запроса. Они не учитывают семантическую составляющую — наличие объектов, их связи.

Для поиска объектов требуется предварительная формализация текстов на ЕЯ — выделение не только объектов, но и всего, что с ними связано. Возникают структуры знаний.

В данной статье рассматривается поиск, основанный на сопоставлении таких структур. *Поиск* осуществляется не на уровне слов, а на уровне структур знаний, и поэтому является *семантическим*.

В настоящее время проблема семантического поиска приобретает все большую актуальность. Следует отметить семантическую поисковую систему AskNet (http://asknet.ru), которая «автоматически выбирает смысловые ответы на запросы пользователя», систему Hakia (http://hakia.com), основанную на хранилище семантической информации и технологии ранжирования найденных текстов по смыслу, а также системы Wolfram Alpha, Powerset и др. Во многих из них рассматриваются смысловые связи между терминами, для представления которых разрабатываются специальные формализмы.

Цель данной статьи — описание технологии поиска информационных объектов на основе струк-

[1]Институт проблем информатики Российской академии наук, keywen1@mail.ru
[2]Институт проблем информатики Российской академии наук, igor-kuz@mtu-net.ru

тур знаний, для извлечения которых используется семантико-ориентированный лингвистический процессор [1–3]. Такой поиск не является универсальным (как в системах Яндекс, Google). Его организация требует настройки лингвистического процессора на выделение объектов в определенной предметной области. Набор таких объектов ограничен. Соответственно, система настраивается давать точные ответы на определенный круг запросов.

В основе семантических поисков лежит технология баз знаний (БЗ), разработанная в рамках проектов ИПИ РАН [4]. Она включает в себя формализацию текстов (извлечение структур знаний), формализмы представления и хранения знаний (выделенных «смысловых элементов») в БЗ, а также методы сопоставления запроса и имеющейся информации на уровне структур знаний.

Для организации соответствующего технологического комплекса требуются формализмы, которые должны обладать определенными свойствами: быть как можно более простыми (в синтаксическом плане), обладать высокими изобразительными возможностями для представления знаний и обеспечивать в широких пределах логико-лингвистическую обработку [1]. Данными свойствами обладает язык расширенных семантических сетей (РСС) и продукционный язык их обработки — ДЕКЛ. На этой основе разработан инструментальный комплекс, ориентированный на обработку структур знаний [5, 6]. Комплекс использован для построения класса семантико-ориентированных лингвистических процессоров, преобразующих тексты на ЕЯ в формализм РСС, организации на этой основе БЗ и для разработки множества прикладных программ, обеспечивающих идентификацию объектов, выявление имплицитной информации, преобразование представлений, семантический поиск, принятие экспертных решений и др. Все эти задачи дополняют одна другую и решаются на одном уровне — структур знаний [4].

Отметим, что структуры знаний в виде РСС автоматически отображаются на языке XML [7] и могут быть использованы для построения прикладных программ на различных языках программирования. Подобный подход при соответствующей технологической доработке может быть основой крайне перспективного направления информатики — «Семантического Интернета».

2 Особенности обработки в базе знаний

Технология семантического поиска объектов на уровне структур знаний была разработана при построении систем «Аналитик» и «Криминал». Последняя была создана для ГУВД г. Москвы [2, 4, 6]. Эти системы ориентированы на работу с текстами на ЕЯ в определенной предметной области. В частности, система «Криминал» ориентирована на работу с большими потоками документов в области криминальной полиции: сводками происшествий, справками по уголовным делам, обвинительными заключениями, записными книжками фигурантов и др. Тексты автоматически формализуется с помощью семантико-ориентированного лингвистического процессора. При этом выделяются информационные объекты (фигуранты, их приметы, адреса, телефоны, даты, оружие, автотранспорт со всеми атрибутами и др.), а также связи между ними и разного рода деяниями, событиями. Участие объектов в одном действии считается одним из видов связи. Более того, сами действия — это тоже информационные объекты, которые связываются с временем, местом, а также причинно-следственными и другими отношениями. В результате возникают сложные структуры. На основе каждого документа формируется семантическая сеть (РСС), называемая *содержательным портретом документа* [8, 9]. Такие портреты образует *базу предметных знаний*, которая запоминается, а сами портреты связываются с соответствующими текстами.

Семантические поиски идут на уровне структур БЗ и включают в себя логический анализ признаков, связей. Например, поиск ответа на запрос в свободной форме (т. е. на ЕЯ) обеспечивается путем сопоставления содержательного портрета, построенного на основе запроса, и содержимого БЗ, т. е. сводится к поиску соответствующей структуры в БЗ. При этом широко используются онтологии, представленные в виде РСС, а также дополнительная информация, которая характеризует поисковый объект или ситуацию, но которая дается в тексте в неявной форме — как имплицитная информация, которую нужно восстанавливать [10].

В данной статье в качестве примера использования технологии БЗ рассматриваются задачи поиска похожих происшествий и лиц (фигурантов). При поиске похожих происшествий учитываются все действия и объекты, составляющие данное происшествие. При поиске похожего фигуранта учитывается только то, что связано с фигурантом. Эти задачи относятся к наиболее важным в области криминальной полиции. Они необходимы для идентификации лиц, установления их связей, порождения и проверки различных гипотез, планирования следственных действий. В данной статье рассматриваются методики и алгоритмы решения этих задач на структурном уровне, т. е. на основе различных видов связей с учетом особенностей

описываемых объектов, событий, происшествий. Ориентация сделана на использование семантических связей, а также методов логического анализа и нечеткого вывода. Отметим, что подобные методики использованы для семантического поиска других информационных объектов.

Задача поиска похожих происшествий и фигурантов решалась в рамках логико-аналитической системы «Криминал» с учетом ее задач и особенностей [4, 6].

В системе «Криминал» онтологии представлены в виде РСС и образуют *онтологическую базу*, которая находится в отдельном файле и объединяется с БЗ в процессе поиска. Онтологическая база определяет семантическое пространство терминов и признаков — с учетом их смысловой близости, синонимии и взаимоотрицания. За счет этого расширяется пространство поиска, повышается точность и надежность результатов, обеспечивается достаточная свобода использования слов и терминов в запросах и заданиях системе.

Все документы и полученные на их основе структуры знаний (содержательные портреты) помещаются в собственную базу данных, ориентированную на большие потоки информации и обеспечивающую их быстрый выбор — за счет индексных файлов (базы данных служат для хранения документов и структур знаний). Эти структуры по мере необходимости подкачиваются в оперативную память и вместе с онтологической базой образуют *оперативную базу знаний* (ОБЗ), где и осуществляется поиск. При этом допускается наличие множества баз данных (со своими БЗ) на различных компьютерах, связанных в сеть. Таким образом обеспечивается работа распределенных БЗ.

3 Содержательные портреты документов

Сеть (РСС), представляющая объекты и связи документа, образует его содержательный портрет, где все слова представлены в канонической форме. Такие портреты служат основой для семантического поиска.

Пример 1. Типовой документ (с номером 221) из сводок происшествий: *1.05.98 г. в 7.10 Фирсова Владимира Николаевича 1953 г.р. прож.: ул. Глаголева 25-1-273, работает АОЗТ «ХДУ», зам. директора, о том, что 1-05-98 г. неизвестные от д. 22 кор. 3 по ул. Тухачевского, похитили а/м ГАЗ 31029, черная, 1995 г/в, дв. 402-0019476. . .*

Его содержательный портрет имеет вид:

ДОК_(221,‘TEXT_98.TXT’,‘S_CRI.NL’)
ДАТА_(#1.5.1998,1998,МАЙ,∼1,7.1/4+)
FIO(ФИРСОВ,ВЛАДИМИР,НИКОЛАЕВИЧ,1953/5+)
АДР_(УЛ.,ГЛАГОЛЕВА,25,1,273/6+)
ПРОЖ.(5-,6-/7+)
ОРГ_(АОЗТ,ХДУ/8+)
РАБ_(5-,8-,ЗАМ.,ДИРЕКТОР/9+)
FIO(“ ”,“ ”,“ ”,НЕСКОЛЬКО/10+) НЕИЗВЕСТНЫЙ(10-)
АВТО_(ГАЗ,31029,ЧЕРНЫЙ,1995,Г/В,ДВ.,402-0019476/11+)
УГНАТЬ(10-,11-/12+)
ДАТА_(#1.5.1998,1998,МАЙ,∼1/14+)
КОГДА(12-,14-)
АДР_(УЛ.,ТУХАЧЕВСКОГО,ДОМ,22,КОРП.,3/15+)
ГДЕ(12-,15-)
ПРЕДЛ_(221,4-,5-,6-,8-,9-,О,ТОМ,12-,14-,15-)

Первый фрагмент ДОК_(221,‘TEXT_98.TXT’, ‘S_CRI.NL’) указывает на то, что содержательный портрет построен на основе документа 221 из файла ‘TEXT_98.TXT’. При этом были использованы лингвистические знания ‘S_CRI.NL’. Второй фрагмент представляет дату. Третий фрагмент представляет фигуранта — *Фирсова Владимира Николаевича*. Ему сопоставлен внутренний код 5+, с помощью которого представлено, где он проживает — ПРОЖ.(5-,6-/7+), где «5-» — код адреса. Здесь же представлены и другие объекты — организация (ОРГ_), место работы (РАБ_), автотранспорт (АВТО_) и др. Фрагмент УГНАТЬ(10-,11-/12+) представляет действие неизвестного лица (с кодом 10+), который *похитил* (= *угнал*) автомашину (с кодом 11+). Последний фрагмент ПРЕДЛ_(221,. . .) содержит коды других фрагментов и представляет порядок расположения соответствующей информации в тексте документа.

Такие портреты (в виде РСС) запоминаются в БЗ. Поиск сводится к сопоставлению таких портретов — запроса и содержимого БЗ [4, 6]. При поиске похожих фигурантов и происшествий важную роль играют не только объекты, но и действия типа УГНАТЬ и др. Помимо этого используется дополнительная информация, представленная в виде аналитических фрагментов (см. разд. 4).

4 Оценка документа по ключевым позициям

При организации семантических поисков важную роль играют признаки, задающие общий характер происшествий (*способы проникновения*, *совершения преступления* и др.), особенности фигуранта (их приметы) или особенности любого другого информационного объекта. Они могут в явном виде не присутствовать в тексте и требуют специального логического анализа для их выявления. С этой целью в процессе ввода документов с их формализацией производится оценка документа по ключевым

позициям. Она необходима для быстрого и качественного поиска, а также для выдачи информации в сжатом виде и объяснения результатов.

Оценка документа по ключевым позициям осуществляется на уровне структур знаний с помощью специальной программы постлингвистической обработки, реализующей идеологию семантических фильтров [3, 8, 9]. Оценка заключается в выделении особенностей описанного в документе происшествия (или особенностей какого-либо информационного объекта) и его соотнесении с соответствующими ветвями типовых классификаторов, находящихся в онтологической базе. Такое соотнесение осуществляется автоматически — на основе анализа содержательного портрета документа.

В результате строятся так называемые *аналитические фрагменты*, которые представляют в сжатом виде наиболее значимую информацию об объекте или происшествии и которые дополняют содержательный портрет документа. Они играют важную роль при поиске и аналитической обработке.

Рассмотрим примеры работы программы постлингвистической обработки на документах из области криминальной полиции.

Пример 2. Формирование по тексту описания словесного портрета фигуранта.

Основные классы онтологической базы, характеризующие фигурантов (лиц): *пол*, *возраст*, *рост*, *особые приметы*, *индивидуальные особенности*, *телосложение*, *тип лица*, *волосы*, *глаза*, *лоб*, *брови*, *нос*, *рот*, *губы*, *зубы*, *подбородок*, *уши*, *одежда*.

Текст на входе:

. . . На вид 45 лет, рост 170–175 см, полного т/сл., одет в рыжую лохматую шапку, зеленый пуховик, черные брюки, зимние ботинки коричневого цвета. . .

На выходе — аналитический фрагмент, представляющий в формализованном виде следующую информацию:

ВОЗРАСТ: 45,
РОСТ: 170, 175,
ТЕЛОСЛОЖЕНИЕ: ПОЛНЫЙ,
ОДЕЖДА: ШАПКА (РЫЖИЙ, ЛОХМАТЫЙ),
ПУХОВИК (ЗЕЛЕНЫЙ), БРЮКИ (ЧЕРНЫЙ),
БОТИНОК (ЗИМНИЙ, КОРИЧНЕВЫЙ).

Каждое слово с двоеточием представляет класс. Далее следуют подклассы. Слова в скобках поясняют или уточняют эти подклассы.

Подобное формализованное описание играет роль реферата. Оно строится по аналитическому фрагменту автоматически с помощью обратного лингвистического процессора. В данном случае программа постлингвистической обработки осуществляет автоматическое построение словесного портрета по тексту описания с его формализацией.

Пример 3. Выявление из текста описания основных характеристик происшествия.

Основные классы онтологической базы, характеризующие криминальные происшествия: *предварительные действия*, *способ проникновения*, *способ совершения преступления*, *преступные действия*, *предлог*, *организация*, *оружие*, *транспортные средства*, *ценные бумаги*, *драгоценные изделия*, *ценные изделия*.

Текст на входе:

. . . Найдена а/м ВАЗ 2109 темно-вишневого цвета г. н. К 939 ЕМ 70, в которой на передних сиденьях находятся два трупа мужчин кавказской национальности на вид 30–35 лет. Исследование показало, что смерть данных лиц наступила от огнестрельного ранения. На месте преступления были найдены и изъяты стреляные 2 гильзы от пистолета ТТ и 5 стреляных гильз, пуля от ПМ.

На выходе — аналитический фрагмент, представляющий в формализованном виде следующую информацию:

Преступные действия: РАНЕНИЕ (ОГНЕСТРЕЛЬНЫЙ),
ЛИЧНОСТЬ: НАЦИОНАЛЬНОСТЬ (Лицо кавказской национальности),
ОРУЖИЕ: ПИСТОЛЕТ (ТТ,ПМ),
АВТОМАШИНА: ВАЗ.

Подобное описание (как и в предыдущем примере) строится автоматически и играет роль сжатого описания или реферата.

5 Этапы поиска

Поиск похожих происшествий и фигурантов (как и других информационных объектов) осуществляется по запросам и заключается в анализе содержательных портретов документов на предмет их совпадения с содержательным портретом запроса [4, 6, 7]. Анализ осуществляется на уровне структур знаний, находящихся в оперативной БЗ. Вначале выделяются объекты запроса (например, фигуранта), их признаки — *приметы*, *деяния*, а также связанные с ними *адреса*, *телефоны*, *машины* и др. Анализ сводится к проверке наличия в документах объектов (фигурантов) с аналогичными признаками и связями. При этом используются следующие признаки и связи:

- первичные признаки (значимые слова запроса в каноническом виде);
- вторичные признаки (близкие по смыслу слова, уточняющие слова и др.), порожденные первичными признаками за счет информации онтологической базы;

– аналитические признаки (*способ проникновения*, *способ совершения преступления* и др.), взятые из аналитических фрагментов;
– свойства объектов (например, для фигурантов — *неизвестный*, *потерпевший*, *безработный*, для действий и событий — *время*, *место*);
– участие объектов в действиях.

Отметим, что в качестве запроса может быть взят любой документ или словесный портрет фигуранта. Тогда вначале будет сформирован содержательный портрет, в котором будут все объекты запроса. Далее пользователь может выбрать любой из них. Тогда система на содержательном уровне будет искать похожие объекты. Если пользователь выберет документ, то будет инициирован поиск похожих происшествий. Поиск является нечетким, так как не требуется полного совпадения слов-признаков. Находится только то, что является общим и объединяет запрашиваемый и найденный объекты. Это важно, так как точный поиск часто не дает результата.

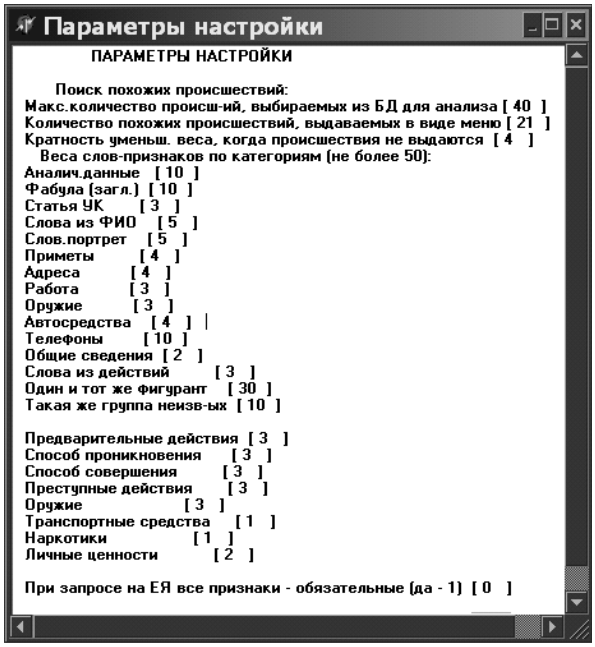
Этапы поиска похожих происшествий и фигурантов

Первый этап. Выделение значимых слов-признаков из содержательного портрета запроса с присвоением им весов. Они образуют первичные признаки. Под значимыми понимаются слова, которые не являются предлогами, союзами, понятиями широкого объема, вспомогательными глаголами и др. Напомним, что значимые слова блоком морфологического анализа приводятся в каноническую форму [2].

Если стоит задача поиска похожих происшествий, то используются все значимые слова запроса, дополненные аналитическими признаками. Если решается задача поиска похожих фигурантов, то из запроса выделяются только те части, которые относятся к указанному фигуранту: его ФИО, а также *приметы*, *адрес*, *деяния* и т. д. Только из этих частей берутся слова, которые в дальнейшем будут играть роль первичных признаков.

Выделенным словам-признакам присваиваются веса в зависимости от уникальности слова и вида информации (куда отнесено слово — к приметам, адресам и др.). Наибольшие веса присваиваются аналитическим признакам, относящимся к характеру происшествия и к фигурантам, описанным в запросе.

Отметим, что веса фактически отражают степень значимости той или иной информации при анализе степени сходства. В системе «Криминал» имеются специальные настроечные фреймы, дающие возможность пользователю изменять веса аналитических данных (*способ проникновения*, *оружие* и т. п.) и категорий (*приметы*, *адреса* и т. п.). Соответственно, будут меняться веса аналитических и других признаков. Таким способом акцентируется внимание на определенных моментах. Пример настроечного фрейма, задающего веса при поиске похожих происшествий (веса получены экспериментальным путем):



Аналогичный вид (с другими категориями и весами) имеет настроечный фрейм, определяющий веса слов-признаков при поиске похожих фигурантов.

Например, если дать высокий вес классу «*способ проникновения*» то система будет присваивать высокие веса происшествиям (документам) со способом проникновения, описанным в запросе. Если дать высокий вес адресам, то большой вес будет присваиваться документам с такими же улицами, номерами домов и квартир, как в запросе.

Второй этап. Дополнение набора признаков за счет онтологической базы. Это необходимо для расширения пространства поиска. Нужно учесть различные способы и средства описания того, что есть в запросе. На базе имеющихся признаков запроса порождаются вторичные признаки:

– близкие по смыслу термины (на основе фрагментов NEAR);
– поясняющие термины (на основе фрагментов SUB);
– противоречивые признаки (на основе фрагментов OR_OR).

Примеры фрагментов, взятых из онтологической базы:

NEAR(АТЛЕТИЧЕСКИЙ,МОГУЧИЙ,МОЩНЫЙ,
БОГАТЫРСКИЙ,СПОРТИВНЫЙ,К РЕПКИЙ)
SUB(ДОКУМЕНТ,ПАСПОРТ)
SUB(ДОКУМЕНТ,УДОСТОВЕРЕНИЕ)
SUB(ДОКУМЕНТ,“Водительские права”)
SUB(ДОКУМЕНТ,“Воинский билет”)
SUB(ДОКУМЕНТ,МЕТРИКА)
SUB(ДОКУМЕНТ,ПРОПУСК)
OR_OR(МОЛОДОЙ,“средний возраст”, ПОЖИЛОЙ)
NEAR(ПОЖИЛОЙ,СТАРЫЙ)
. . .

Поясним роль этих фрагментов на примерах. Если в запросе встретился признак БОГАТЫРСКИЙ (относящийся к фигуранту), то за счет фрагмента NEAR будут сформированы вторичные признаки: АТЛЕТИЧЕСКИЙ, МОГУЧИЙ, МОЩНЫЙ, СПОРТИВНЫЙ, КРЕПКИЙ, которые будут принимать участие при поиске.

Если в запросе встретился термин ДОКУМЕНТ, то за счет фрагментов типа SUB будут сформированы способы его расшифровки: это может быть ПАСПОРТ, УДОСТОВЕРЕНИЕ, «Водительские права», «Воинский билет», МЕТРИКА, ПРОПУСК. Они будут также учитываться при поиске.

Если в запросе фигурант был охарактеризован как ПОЖИЛОЙ (это могла сделать и сама система путем анализа возраста), то за счет фрагментов типа OR_OR будут сформированы опровергающие признаки: МОЛОДОЙ, «средний возраст», которые используются при оценке степени сходства со знаком минус.

Вторичным признакам также присваиваются веса — в зависимости от веса признака, который их породил.

При наличии в запросе фигурантов производится анализ их ФИО. Отметим, что полные имена и отчества при построении содержательного портрета уже преобразуются к единому виду — в каноническую форму. На данном этапе на их основе порождаются инициалы, которые тоже играют роль признаков. И наоборот, по инициалам порождаются возможные имена и отчества. Это позволяет при поиске более полно охватить возможные случаи написания ФИО.

Третий этап. Быстрый поиск (по индексным файлам) в базах данных содержательных портретов документов с указанными признаками. Поиск осуществляется на основе выделенных из запроса слов-признаков и заключается в подсчете взвешенной суммы весов совпавших признаков. В качестве результата выдаются номера найденных документов — в порядке взвешенных сумм.

При наличии в запросе фигурантов с ФИО производится дополнительный поиск документов, при котором обязательными признаками делаются пары: полные фамилия и имя или полные имя и отчество. Словом, находятся документы, где есть и то, и другое. Это позволяет избежать потерь при поиске и идентификации фигурантов.

Четвертый этап. Подкачка из базы данных в оперативную память семантических сетей — содержательных портретов документов с наибольшими взвешенными суммами. В результате (вместе с онтологической базой) образуется ОБЗ, представляющая собой большую семантическую сеть, доступную для быстрого выполнения сложных операций сравнения и логического анализа.

Пятый этап. Детальный анализ на совпадение слов-признаков, связанных с объектами (или выбранным объектом) запроса и объектами, находящимися в ОБЗ. При этом сравнение идет по категориям: ФИО фигуранта из запроса сравнивается с ФИО фигурантов из ОБЗ, связанные с ними приметы сравниваются с приметами, свойства — со свойствами, действия — с действиями и т. д. В результате находятся похожие объекты. Подсчитывается их вес, который определяет степень сходства с объектами (или объектом) запроса. Выбираются объекты с наибольшими весами. При этом учитываются следующие факторы:

- веса совпавших порожденных аналитических признаков, определяющих характер объекта или особенности фигуранта;
- веса совпавших слов-признаков (в том числе вторичных);
- соотнесенность признаков к той или иной категории;
- связь признаков, заданная в онтологической базе (близкие по смыслу или поясняющие);
- сильное совпадение по какой-либо категории признаков (например, совпадает большинство примет);
- наличие противоречивых признаков.

Каждое совпадение дополняет общий вес выбранного объекта — к нему добавляется вес совпавшего признака. При наличии противоречивых признаков их веса вычитаются.

При анализе чисел и интервалов на их совпадение (например, ВОЗРАСТ, РОСТ, номер дома, квартиры, год и др.) рассматриваются различные варианты:

- равенство чисел;
- число входит в интервал;
- пересечение интервалов;
- близость числовых значений.

В зависимости от варианта совпадения и от категории (приметы, адрес, время и др.) к общему весу документа добавляется определенная величина.

При поиске похожих происшествий в первую очередь учитывается сходство аналитических признаков и криминальных действий, а уже затем объектов, участвующих в этих действиях. Для этого категориям присваиваются соответствующие веса. Общий вес выбранного (из ОБЗ) происшествия подсчитывается как сумма весов входящих в него признаков и объектов (действие — это тоже объект).

При поиске фигурантов различается два случая.

Первый случай — когда в запросе заданы ФИО фигуранта. Тогда в ОБЗ производится поиск фигурантов с аналогичными ФИО. При этом учитываются случаи совпадения инициалов с полными именами или отчествами (такое совпадение дает меньший вес). Подсчитывается общая степень совпадения — в зависимости от совпавших признаков.

Множество найденных фигурантов с высокими весами является основой для дальнейшего анализа. К ним добавляются веса, полученные от совпадения свойств, а также от совпадения связанных с фигурантами примет, адресов, телефонов и др. В результате находятся фигуранты с высокими весами, отражающими степень сходства с лицом, описанным в запросе.

Если в ОБЗ не найдено фигурантов с ФИО, заданными в запросе, то ищутся фигуранты, у которых может быть другое имя или отчество. Возникают противоречивые признаки, которые уменьшают вес анализируемого фигуранта. При этом акцент смещается на сравнение связей.

Второй случай — когда запрашивается неизвестное лицо (фигурант). Тогда поиск и сравнение идет по связанным с этим лицом приметам, действиям, адресам и другим объектам. В ОБЗ ищутся лица с аналогичными связями.

При поиске похожих происшествий найденные фигуранты с их степенями совпадения запоминаются, а сами степени дополняют вес соответствующего документа. Помимо этого учитываются веса других совпавших признаков. В результате находятся происшествия (документы) с высокими весами, отражающими степень сходства с запросом.

Шестой этап. Выдача похожих происшествий или фигурантов, упорядоченных по степени сходства, в виде списка или меню.

Седьмой этап. Выдача объяснений. Пользователь может выбрать из упомянутого меню любой пункт, соответствующий происшествию или фигуранту. Система на основе совпавших признаков формирует объяснение сходства в виде понятного текста на русском языке.

6 Выдача и объяснение результатов

Как отмечалось ранее, вся обработка в системе «Криминал» осуществляется на уровне семантических сетей в рамках специально созданного для этого инструментария языка ДЕКЛ [5, 6]. Находятся фрагменты семантической сети, представляющие похожие происшествия или фигурантов с совпавшими признаками. При выдаче соответствующего меню и объяснении результатов такие фрагменты преобразуются на понятный пользователю язык — естественный. Это делается с помощью обратного лингвистического процессора.

При формировании меню формируются краткие описания происшествий или фигурантов. При объяснении результатов (когда пользователь выбирает из меню интересующее его происшествие или фигуранта) дается краткое описание выбранного происшествия (фигуранта), указываются совпавшие и противоречивые признаки, а также дается сам текст описания. Этого достаточно, чтобы помочь пользователю самому оценить степень сходства или адекватности запросу.

Пример 4. Проиллюстрируем сказанное на примере выдачи результатов поиска похожих криминальных происшествий и похожих фигурантов.

Текст на входе (взят из документа с номером 63): *. . . На лестничной площадке 3-го этажа двое неизвестных из неустановленного оружия нанесли два сквозных ранения в голову и живот Лихомову Владимиру Ивановичу, 1954 г.р., неработающий, прож.: Тюменская. . . С места происшествия изъято: 1 пуля и 1 гильза калибра 7.62 мм предположительно от пистолета ТТ. . .*

Меню похожих происшествий выглядит следующим образом:

На документ 63 содержательно похожи:

– Док-т 1231 (БЗ-1) УБИЙСТВО 29.3.1996 (вес 142);
– Док-т 4323 (БЗ-1) ОГРАБЛЕНИЕ 20.6.1996 (вес 111);
– Док-т 81 (БЗ-2) УБИЙСТВО 1.7.1995 (вес 92);
– . . .

При выборе пункта 1 данного меню на экран будет выдано объяснение причин сходства документов 63 и 1231:

Похожее происшествие — 1231 из БЗ-1 (вес 142).

В происшествии 1231 встретились те же признаки:

Преступные действия: РАНЕНИЕ, ГОЛОВА.
Оружие: ПИСТОЛЕТ, ТТ.
Фабула: УБИЙСТВО.
Работа: НЕРАБОТАЮЩИЙ.
Действие: ИЗЪЯТЬ ГИЛЬЗА.
Общие сведения: РЕЗУЛЬТАТ, ПРОВЕДЕНИЕ, СОТРУДНИК, ОКАЗАТЬСЯ, КАЛИБР.
<Текст документа 1231 из БЗ-1>.

Пример 5. Меню похожих фигурантов выглядит следующим образом:

На фигуранта ЛИХОМОВ ВЛАДИМИР ИВАНОВИЧ похожи:

– ЛИХОМОВ ВЛАДИМИР ПЕТРОВИЧ 1943, док. 4437 из БЗ-1 (вес 42);
– без ФИО в кол-ве 1, док. 81 из БЗ-1 (вес 35);
– КОВАЛЕВ ВЛАДИМИР ИВАНОВИЧ 1956, док. 24 из БЗ-2 (вес 30);
– . . .

При выборе пункта 1 данного меню на экран будет выдано объяснение причин сходства фигурантов ЛИХОМОВ ВЛАДИМИР ИВАНОВИЧ и ЛИХОМОВ ВЛАДИМИР ПЕТРОВИЧ:

Похожий фигурант (вес 44) — ЛИХОМОВ ВЛАДИМИР ПЕТРОВИЧ 1943.
Особые приметы: ОТМЕТИНА (РАНЕНИЕ),
ТЕЛОСЛОЖЕНИЕ: ТОЛСТЫЙ,
СТАТУС: ПОТЕРПЕВШИЙ (РАНЕНИЕ).

У фигуранта встретились те же признаки:

ТЕЛОСЛОЖЕНИЕ: ТОЛСТЫЙ,
СТАТУС: ПОТЕРПЕВШИЙ РАНЕНИЕ,
Работа: НЕРАБОТАЮЩИЙ.
ФИО: ЛИХОМОВ ВЛАДИМИР.
Не совпадают ФИО: 1943 (*было* 1954).
Не совпадают ФИО: ПЕТРОВИЧ (*было* ИВАНОВИЧ).
Входит в документ с номером 4437 из БЗ-2.
<Текст документа 4437>.

Отметим некоторые важные моменты.

Во-первых, в содержательных портретах представлены объекты и действия. Их сопоставление играет важную роль при поиске похожих происшествий, при классификации лиц как *потерпевших* или *преступников* и во многих других случаях.

Во-вторых, дается оценка документа по ключевым позициям, представляющая в сжатом виде наиболее значимую информацию и играющая роль реферата.

В-третьих, при анализе степени сходства запроса и документа используются признаки типа «*Преступные действия*», «*Угроза оружием*» и др., которые в явном виде могут отсутствовать в тексте и которые выявляются системой в процессе постлингвистической обработки. Соответственно, такие признаки вводятся в объяснения.

В-четвертых, допускается поиск похожих фигурантов без ФИО (поиск неизвестных лиц) по связанной информации, например по словесному портрету.

И последнее: использование привычных человеку классификаторов (они представлены в онтологической базе) делает результат реферирования и объяснения более понятным.

7 Заключение

Особенность предлагаемых в данной статье методик и алгоритмов семантического поиска состоит в следующем:

(1) вся обработка осуществляется на уровне структур знаний, т. е. содержательных портретов документов. Они образуют БЗ, которая вместе с правилами преобразования (продукциями языка ДЕКЛ) образует законченный технологический комплекс, ориентированный на сложные задачи, связанные с логическим выводом, преобразованием представлений, экспертными решениями. В результате обеспечивается анализ высокой степени глубины и сложности;

(2) выделяются и используются разнообразные признаки. Учитывается наличие множества объектов (лиц, телефонов, оружия и т. п. — до 40 типов) и аналитических признаков, характеризующих происшествия и фигурантов. Для расширения пространства поиска используется онтологическая база;

(3) допускается работа с многими БЗ, связанными через сеть или Интернет. Они образуют распределенную БЗ.

Описанные методики и алгоритмы семантического поиска реализованы в рамках систем «Аналитик», «Криминал», «Поток» и апробированы при работе с различными корпусами текстов, среди которых: сообщения СМИ, сводки происшествий, обвинительные заключения, записные книжки фигурантов и др. Эти методики использованы в различных приложениях и показали высокую степень универсальности. В перспективе они могут послужить основой для создания комплекса поисковых программ, составляющих «Семантический Интернет».

Литература

1. *Кузнецов И. П.* Семантические представления. — М.: Наука, 1986. 290 с.
2. *Кузнецов И. П., Кузнецов В. П., Мацкевич А. Г.* Система выявления из документов значимой информации на основе лингвистических знаний в форме семантических сетей // Диалог-2000: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям. — Протвино, 2000. Т. 2.
3. *Kuznetsov I., Matskevich A.* System for extracting semantic information from natural language text // Диалог-2002: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям (Протвино). — М.: Наука, 2002. Т. 2.
4. Лаборатория компьютерной лингвистики ИПИ РАН: Официальный сайт. www.IpiranLogos.com.
5. *Кузнецов И. П., Шарнин М. М.* Продукционный язык программирования ДЕКЛ // Система обработки декларативных структур знаний Деклар-2. — М.: ИПИ РАН, 1988.
6. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе баз знаний. — М.: МТУСИ, 2007. 173 с.
7. *Kuznetsov I. P., Kozerenko E. B.* Linguistic processor Semantix for knowledge extraction from natural texts in Russian and English // ICAI 2008: 2008 Conference (International) on Artificial Intelligence Proceedings. — Las Vegas: CSREA Press, 2008. P. 835–841.
8. *Кузнецов И. П., Мацкевич А. Г.* Особенности организации базы предметных и лингвистических знаний в системе АНАЛИТИК // Диалог-2003: Труды Междунар. конф. по компьютерной лингвистике и интеллектуальным технологиям. — Протвино, 2003. С. 373–378.
9. *Кузнецов И. П., Мацкевич А. Г.* Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка // Диалог-2005: Труды Междунар. конф. по компьютерной лингвистике и интеллектуальным технологиям (Звенигород). — М.: Наука, 2005. С. 303–311.
10. *Kuznetsov I. P.* Identifying role functions of persons on the basis of knowledge structures) // Компьютерная лингвистика и интеллектуальные технологии: По матлам ежегодной Междунар. конф. «Диалог 2011». — М.: РГГУ, 2011. Вып. 10(17). С. 391–402.