МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра систем информатики

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Ахмадеева Ирина Равильевна

Разработка средств автоматизации построения таксономического ядра онтологий на основе корпуса текстов

Направление подготовки 230100.62 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Руководитель

Загорулько Ю.А.

к.т.н., зав.лаб. ИСИ СО РАН
(уч.степень, уч.звание)

……..
(подпись, дата)

Автор

Ахмадеева И.Р.

ФИТ, гр. 9201

..……..
(подпись, дата)

Новосибирск, 2013г.

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Кафедра систем информатики
(название кафедры)

УТВЕРЖДАЮ

Зав. Кафедрой: Лаврентьев М.М.
(фамилия, И., О.)

……
(подпись, дата)

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ БАКАЛАВРА

Студенту *(ке)* Ахмадеевой Ирине Равильевне
(фамилия, имя, отчество)

Направление подготовки 230100.62 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Тема: Разработка средств автоматизации построения таксономического ядра онтологий на основе корпуса текстов
(полное название темы выпускной квалификационной работы бакалавра)

Исходные данные (или цель работы): разработать и реализовать методы и средства автоматизации построения таксономического ядра онтологии на основе корпуса текстов.

Структурные части работы: провести обзор предметной области, исследовать существующие методы, используемые для автоматизации построения онтологии, выбрать методы наиболее пригодные для автоматизации построения таксономического ядра онтологии на основе корпуса текстов, спроектировать и реализовать программное решение.

Содержание

ВВЕДЕНИЕ

У понятия «онтология» есть несколько значений. В информатику этот термин пришел из философии. Онтология – раздел философии, изучающий фундаментальные принципы бытия. В философии онтология – это учение о бытии, о сущем, о его формах и фундаментальных принципах, о наиболее общих определениях и категориях бытия.

В информатике и искусственном интеллекте онтология – это точная спецификация концептуализации [2].

Под «концептуализацией» понимается абстрактная модель предметной области, описывающая систему понятий предметной области. Это объекты, понятия и другие сущности, которые предполагаются существующими в некоторой предметной области, а также отношения, которые определены между ними. Концептуализация является абстрактной, упрощенной точкой зрения на мир, которая представлена для некоторых целей. «Спецификация» подразумевает описание системы понятий в явном виде.

Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную область знаний. Преимуществом онтологий в качестве способа представления знаний является их формальная структура [5], которая упрощает их компьютерную обработку.

Онтологии могут быть использованы при решении различных задач, как правило, в случаях совместной работы в одной предметной области, для возможности накопления и повторного использования знаний, для создания моделей. Это позволяет специалистам в одной предметной области лучше манипулировать знаниями, обнаруживать новые закономерности, обмениваться опытом.

Часто онтологии используются в тех случаях, где требуется обработка данных, учитывающая их семантику. Например, для повышения эффективности поиска в сети Интернет [4]. Интеллектуальные системы на основе онтологий в последнее время получили широкое распространение.

Построение онтологии предполагает определение классов объектов и описание их отношений с помощью одного из формальных языков. Процесс построения онтологии обычно начинается с создания базовой онтологической структуры, таксономического ядра онтологии, представляющего основные понятия предметной области и родовидовые связи между ними. Необходимым условием для успешного построения онтологии является наличие специалиста, разбирающегося в данной предметной области. Традиционно к созданию онтологии привлекаются эксперты, хорошо знающие данную предметную область. Ручная работа экспертов отнимает много времени и сил, поэтому автоматизация процесса построения онтологий является весьма актуальной задачей. Одним из наиболее

продуктивных подходов на пути к автоматизации построения онтологий является использование методов автоматического извлечения знаний из корпусов текстов, что не требует привлечения дорогостоящих специалистов.

Информацию о терминах и отношениях между ними содержат многие типы источников: базы данных, тексты с разметкой, словари, неструктурированные тексты. Наиболее перспективным представляется подход, основанный на извлечении терминов и отношений между ними из корпуса неструктурированных текстов. Это обусловлено тем, что неструктурированные тексты – наиболее популярный формат. Именно в таком виде хранится большая часть информации.

В последнее время проблеме построения онтологий было посвящено большое число исследований. Однако, несмотря на это, проблема автоматического построения онтологий на текущий момент не имеет удовлетворительного решения.

1 Постановка задачи

Онтология представляет собой систему, состоящую из множества сущностей, связанных отношениями, их свойств и утверждений (аксиом и правил), которые позволяют ограничить смысловые значения сущностей в рамках данной предметной области (ПрО).

Основой любой онтологии являются классы, описывающие понятия моделируемой предметной области, поэтому построение онтологии начинается с составления списка понятий ПрО – классов. Далее определяются свойства понятий и отношения между ними. Наиболее важным отношением между понятиями является отношение обобщения, так как именно на его основе понятия предметной области могут быть выстроены в таксономическую структуру. В связи с этим при построении онтологии отношение обобщения выделяется в первую очередь. В рамках данной дипломной работы рассматривается базовая часть процесса построения онтологии предметной области – построение ее таксономического ядра.

Целью данной работы является разработка методов и средств автоматизации построения таксономического ядра онтологии на основе корпуса текстов.

Для достижения данной цели были поставлены следующие задачи:

1 Анализ существующих методов автоматического построения онтологии.
2 Выбор методов, подходящих для данной цели.
3 Разработка алгоритма автоматического построения таксономического ядра онтологии на основе корпуса текстов.
4 Реализация данного алгоритма.
5 Создание графического интерфейса пользователя.

2 Методы построения онтологий предметной области

Для автоматического построения онтологии используются различные подходы: статистические методы и методы, основанные на подходах из области искусственного интеллекта. Также многие авторы высказываются о возможности применения анализа формальных понятий для построения формальных онтологий [8, 9].

2.1 Построение онтологии с использованием статистических методов

Наиболее часто для извлечения знаний в различных системах используются статистические методы [3]. Статистические методы основываются на частоте встречаемости слов в тексте на естественном языке. С их помощью выделяются классы и отношения между ними, которые формируют онтологию [6].

При выделении классов учитывается, что:

- Имя класса содержит хотя бы одно существительное.
- Общеупотребительные слова по сравнению с терминами обладают большей частотой встречаемости, приблизительно равной в различных предметных областях.
- Количество информации (основанное на частоте встречаемости) термина из нескольких слов больше, чем количество информации отдельных слов, входящих в его состав.

На первом этапе в каждой коллекции документов выделяются имена существительные и определяется их частота встречаемости. На втором этапе выделяют термины, состоящие из одного слова. Далее, сравниваются частоты встречаемости различных существительных в рамках одной коллекции, также проводится оценка пересечения различных коллекций по используемым существительным.

На следующем этапе на основе взаимной информации выделяются термины, состоящие из нескольких слов. Для случая двухсложных терминов взаимная информация определяется по формуле:

$$mi(x, y) = \frac{P(x,y)}{P(x)P(y)},$$

где *x* и *y* представляют собой отдельные слова термина;

P(x) – частота встречаемости слова *x*;

P(x, y) – частота совместной встречаемости *x* и *y*.

С точки зрения теории вероятности коэффициент взаимной информации является способом проверить независимость появления двух слов в тексте. Взаимная информация будет принимать максимальное значение, когда *x* и *y* будут встречаться только совместно.

Таким образом, чем больше *mi(x, y)*, тем более *x* и *y* зависимы друг от друга и вероятность того, что они вместе являются термином, выше.

Выделенные описанным выше образом термины будут представлять собой классы будущей онтологии.

Для выделения отношения «общее-частное» используется количественный подход к информации. Очевидно, что термин, находящийся на более низком уровне иерархии, обладает большим количеством информации, чем обобщающий термин. Определение того, связаны ли два различных термина с разным количеством информации отношением «общее-частное», выполняется двумя способами.

Первый способ основывается на предположении, что частные термины содержат в своем составе слова из более общих терминов.

Второй способ основывается на понятии «контекста слова». Контекст – это условия употребления слова, позволяющие уточнить его значение. Под контекстом термина понимается некоторое множество слов, которые встречаются одновременно с данным.

В случае если у терминов нет общих слов, но совпадает контекст, и при этом они обладают разным количеством информации, имеет смысл говорить о наличии между ними отношения «общее-частное».

Если контекст слов совпадает, но количество информации у терминов приблизительно равное, то вероятнее всего между терминами существует отношение синонимии.

Следует заметить, что предложенный подход позволяет выделить только базовые отношения, необходимые для построения таксономии.

2.2 Методы, основанные на системах продукций

Данный подход относится к группе методов автоматического построения онтологий, в основе которых лежат подходы из области искусственного интеллекта.

Предложение на естественном языке представляет собой некоторое утверждение. Ситуационный подход акцентирует внимание на том, что для пользующегося языком человека значение слова реализуется через включение его в некоторую более объемную единицу – пропозицию. А это значит, что анализ языковых ситуаций в научном тексте лучше всего выполнять с помощью продукционных правил, ядром которых и будет являться пропозиция [7]. Продукционное правило (продукция) – это пара «условие-действие». Условная часть правила – это шаблон, который определяет, когда это правило может быть применено. Часть действия определяет соответствующий шаг в решении задачи.

Кроме того, применение продукционных правил позволяет обеспечить следующие преимущества:

- простоту и высокое быстродействие;
- модульность (каждое правило описывает небольшой, относительно независимый блок знаний);
- удобство модификации (старые правила можно изменять и заменять на новые достаточно независимо от других правил);
- прозрачность (использование правил облегчает реализацию способности системы к объяснению принятых решений и полученных результатов);
- возможность постепенного наращивания (добавление правил в базу знаний происходит независимо от других правил).

Однако создание продукций может быть достаточно трудоемким процессом. В некоторых работах предлагается автоматизировать этот процесс. Например, в работе [7] предлагается использовать генетический алгоритм для построения продукционных правил.

2.3 Анализ формальных понятий для построения онтологии

При использовании анализа формальных понятий для построения онтологии можно говорить только о построении скелета онтологии – решетки формальных понятий, т.е. выводе множества понятий предметной области и выявлении заданного на этом множестве отношения «общее-частное».

Анализ формальных понятий (АФП) (англ. Formal Concept Analysis, FCA) – ветвь прикладной алгебраической теории решёток. Традиционно АФП относят к области концептуальных структур в искусственном интеллекте.

Часто анализ формальных понятий рассматривается как один из методов анализа данных. С помощью этого метода могут быть визуализированы объектно-признаковые зависимости. Это достигается построением диаграммы решётки формальных понятий. Основная математическая идея анализа формальных понятий – возможность построения полной решётки по любому бинарному отношению и формализация описания понятия в виде пары <объём, содержание>.

Анализ формальных понятий в качестве исходных данных о предметной области использует соответствие «объект-свойство» (в котором объектам ставятся в соответствие характерные для него свойства), называемое формальным контекстом, формирование которого в большинстве случаев выходит за рамки работ, касающихся построения онтологий на основе анализа формальных понятий.

Алгоритмы порождения всех формальных понятий предметной области из ее формального контекста широко известны [1]. Как правило, они сразу ориентированы на упорядочение понятий в соответствии с обычным пониманием отношения обобщения и строят решетку формальных понятий предметной области. В общем случае для полученной таким образом онтологии характерны следующие свойства:

- Множественное наследование свойств.
- Существование понятий с нулевым содержанием (такое понятие, что у объектов, входящих в него, отсутствуют общие свойства).
- Описание реальных объектов предметной области «листовыми понятиями», остальные же будут абстрактными обобщениями.

2.4 Обзор существующих решений

Рассмотренные методы автоматического построения онтологий дают разработчикам широкий выбор средств автоматизации построения онтологии, однако, следует отметить, что данные методы не лишены недостатков.

Методы, основанные на продукционном подходе, достаточно удобны в применении и хорошо понимаются экспертами. Системы на их основе способны к объяснению принятых решений. Однако среди их недостатков можно выделить недостаточную семантическую связность между правилами. Также недостатком является сложность в формировании правил. К тому же они не всегда эффективны и нестандартные шаблоны построения предложений могут привести к ложным срабатываниям.

Статистический подход также является достаточно универсальным и не зависит от языка, на котором представлены тексты. Он может одинаково успешно применяться и для русского, и для английского языка. Однако и этот подход имеет свои недостатки, главным из которых является то, что эти методы рассматривают анализируемые тексты как упорядоченный набор слов, не учитывая их связность.

Анализ формальных понятий позволяет выявлять таксономические отношения между понятиями, строя формальные понятия, однако исходными данными для этого метода является формальный контекст, построение которого осуществляется экспертом.

Существует много подходов к построению онтологии предметной области. У каждого подхода есть свои достоинства и недостатки. При применении сразу нескольких подходов можно воспользоваться достоинствами каждого из них.

3 Алгоритм построения онтологии

Для автоматизации построения ядра онтологии предлагается использовать анализ формальных понятий. Так как этот метод в качестве исходных данных использует таблицу «объект-свойство», предлагается предварительно строить такую таблицу с использованием статистических методов и методов, основанных на системах продукций. Рисунок 1 показывает обобщенный алгоритм построения онтологии.

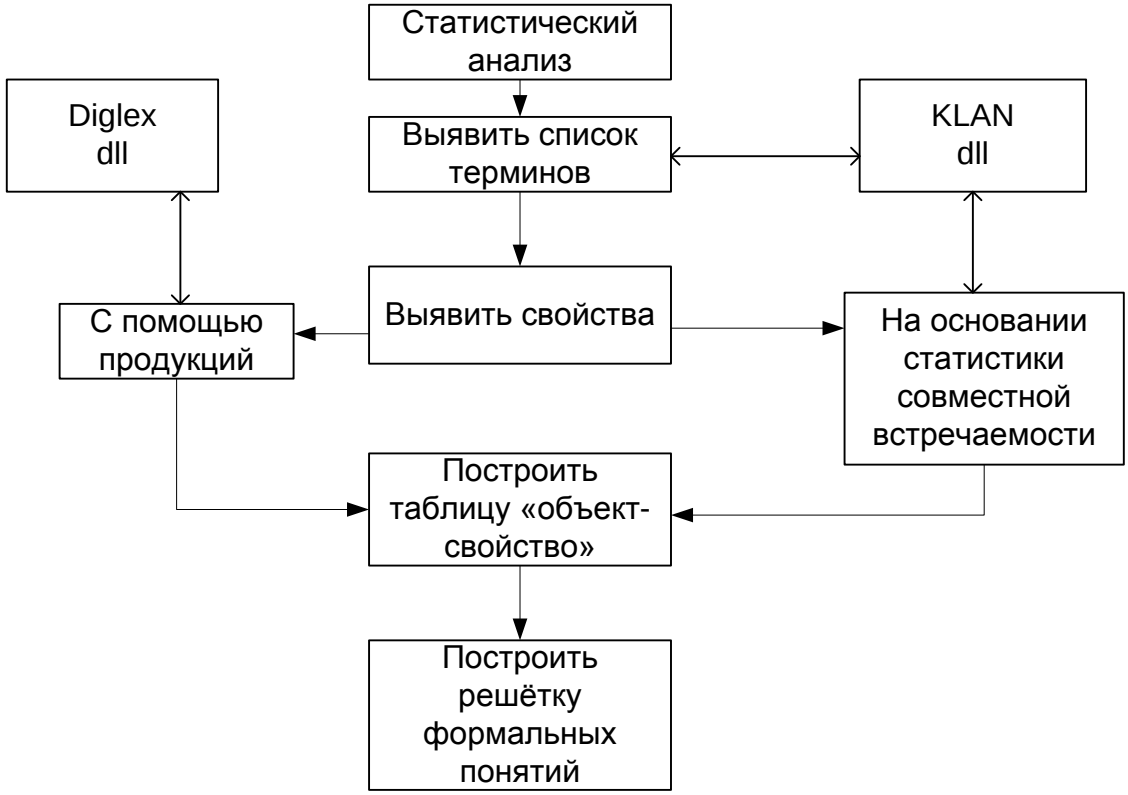


Рисунок 1. Обобщенный алгоритм построения онтологии

В данном алгоритме можно выделить два этапа: на первом этапе на основе корпуса текстов строится таблица «объект-свойство», на втором этапе данная таблица преобразуется в решетку формальных понятий, являющуюся таксономическим ядром онтологии.

3.1 Предварительная обработка текста

Нулевой этап в построении онтологии заключается в предварительной обработке корпуса текстов. Под корпусом текстов здесь понимается совокупность текстов, собранных в соответствии с общей предметной областью.

На этом этапе производится лексическая обработка текста. Основная задача лексического анализа состоит в том, чтобы выделить последовательность лексем из входной последовательности одиночных символов. Каждый символ входной

последовательности либо принадлежит какой-нибудь лексеме, либо является разделительным символом.

Результатом лексического анализа является множество лексем, часть из которых в дальнейшем будет определена как термины (понятия) или их свойства.

3.2 Извлечение терминов

На следующем шаге определяется список терминов. Для этого проводится статистическая обработка текста, основанная на частотных характеристиках текста: частота вхождения каждого слова в текст, частота совместного вхождения нескольких слов. В данном случае отношения между словами не анализируются с лингвистической точки зрения.

С помощью статистического анализа выделяются наиболее часто встречающиеся слова и словосочетания. Существительные и словосочетания вида «существительное + зависимое слово» объявляются терминами.

3.3 Извлечение свойств

На следующем этапе выявляются свойства понятий, полученных на первом этапе. В результате строится таблица «объект-свойство», где по вертикали расположены понятия, по горизонтали все возможные свойства, а на пересечениях указано значение данного свойства у объекта.

Поиск свойств терминов выполняется двумя способами: с помощью продукционных правил и на основе статистики совместной встречаемости.

Для формирования продукций используются следующие шаблоны:

- <термин> обладает свойством | признаком | параметром <свойство>
- <термин> зависит от признаков | свойств | параметров <список свойств>
- <термин> имеет признаки | свойства | параметры <список свойств>
- <свойство> характерно для <термин>
- <термин> обладает <свойство>
- свойством | признаком | параметром <термин> является <признак>
- <термин> характеризуется наличием <свойство>
- <термин> характерно <свойство>

С помощью данных шаблонов можно выявить свойства, характерные для каждого термина. Однако успешное нахождение свойств таким методом сильно зависит от характера текстов данной предметной области.

Поэтому параллельно с методом выделения свойств, основанном на использовании продукционных правил, применяется метод, основанный на статистике. Для этого для

каждого термина рассматриваются все словосочетания, в которые входит данный термин и строится статистика участвующих в них слов. Слова, с которыми данный термин наиболее часто входит в словосочетания, считаются свойствами.

3.4 Построение таблицы «объект-свойство»

Таблица «объект-свойство» представляет собой таблицу, где каждому объекту (термину) ставится в соответствие значение характерного ему свойства. В частном случае значением свойства может быть «обладает» либо «не обладает». Ее построение требует сначала выделения терминов предметной области из корпуса текстов, а затем определения списка свойств для каждого термина.

В результате выполнения данного этапа формируется таблица «объект-свойство», которая является входным параметром для работы следующего этапа – анализа формальных понятий.

3.5 Формирование онтологии на основе таблицы «объект-свойство»

Таблица «объект-свойство» поставляет исходные данные для анализа формальных понятий. Основная математическая идея АФП заключается в возможности построения полной решётки по любому бинарному отношению и формализация описания понятия в виде пары *<объём, содержание>*. (Здесь под *объемом* понимается некоторое множество объектов, а под *содержанием* – общие для них признаки.)

В основе решеток формальных понятий лежит так называемое соответствие Галуа, задаваемое на множестве объектов и признаков и обладающее известным из философского определения понятий свойством уменьшения объёма с ростом содержания.

Суть метода заключается в построении решетки формальных понятий на основе формального контекста. Формальный контекст – это тройка $K = (G, M, I)$, где

- G – множество объектов
- M – множество признаков
- $I \subseteq G \times M$ (отношение обладания признаком)

Пара $<g, m> \in I$ или gIm показывает, что объект g обладает признаком m.

Часто формальный контекст представляют в виде бинарной матрицы. Таким образом, таблица «объект-свойство», построенная на предыдущих этапах, является формальным контекстом, иными словами исходными данными для анализа формальных понятий.

Задача данного этапа заключается в том, что, имея формальный контекст, необходимо выделить в нем формальные понятия, которые и будут являться классами результирующей онтологии, а также построить решетку формальных понятий, для выявления иерархических отношений.

В определении формального понятия используются операторы Галуа:

Пусть $A \subseteq G$ и $B \subseteq M$, тогда:

- $A' = \{m \in M \mid \forall g \in A : gIm\}$, т.е A' – множество признаков, которыми обладают все объекты из множества A.
- $B' = \{g \in G \mid \forall m \in M : gIm\}$, т.е. B' – множество объектов, которые обладают всеми признаками из множества B.

Формальное понятие – это пара (A, B), где $A \subseteq G$, $B \subseteq M$, такие что $A'=B$ и $B'=A$. A называется объемом, а B – содержанием понятия. В матрице формального контекста формальное понятие представляет собой подматрицу, состоящую из единиц.

Для двух формальных понятий (A_1, B_1) и (A_2, B_2) некоторого контекста, (A_1, B_1) называется подпонятием (A_2, B_2) если $A_1 \subseteq A_2$ (эквивалентно $B_2 \subseteq B_1$). В этом *случае* (A_2, B_2) является надпонятием (A_1, B_1), и это обозначают как $(A_1, B_1) \leq (A_2, B_2)$. Множество всех

понятий контекста *(G, M, I)*, упорядоченных по вложению объемов, называется решеткой понятий.

В работе [1] было показано, что множество всех понятий формального контекста образует полную решетку. Множество понятий контекста *K* создает частичный порядок по вложению объемов понятий и всегда имеет наименьшее и наибольшее по вложению понятия. Наибольшее по объему понятие представляет собой понятие, содержащее все объекты и признаки, которыми обладают все объекты (чаще всего ни одного признака). Наименьшее – понятие, содержащее все признаки.

Таким образом, в ходе данного этапа строится решетка формальных понятий для формального контекста, построенного на предыдущем этапе. Далее каждое формальное понятие *(A, B)* рассматривается как отдельный класс онтологии. Для любого *i* формальное понятие (A_i, B_i), такое что $(A_i, B_i) \leq (A, B)$ будет подклассом данного класса. И аналогично для любого i формальное понятие (A_i, B_i), такое что $(A_i, B_i) \geq (A, B)$ будет надклассом данного класса.

В итоге результатом работы алгоритма на данном этапе будет множество классов, связанных отношением наследования, что и будет таксономическим ядром онтологии предметной области.

4 Программная реализация

4.1 Структура программы

Программа для автоматического построения ядра онтологии на основе корпуса текстов состоит из двух модулей, работающих поэтапно.

На первом этапе модуль построения таблицы «объект-свойство» на основе статистики, а также с помощью продукционных правил выделяет понятия и их свойства, а затем строит таблицу «объект-свойство».

На втором этапе модуль построения онтологии на основе таблицы «объект-свойство» строит решетку формальных понятий, являющуюся ядром онтологии.

Данная программа обладает графическим интерфейсом пользователя, который позволяет пользователю программы следить за процессом построения онтологии и при необходимости корректировать полученный результат.

Для написания программы использовался язык C++, разработка велась в среде Borland C++ Builder 6.0. Выбор языка был обусловлен наличием готовых библиотек для анализа текста, разработанных в Институте Систем Информатики, которые используются для решения исходной задачи.

Рисунок 2 Архитектура системы

4.2 Модуль построения таблицы «объект-свойство»

Модуль построения таблицы «объект-свойство» решает две задачи:

- Выделение списка понятий из корпуса текстов на естественном языке;
- Выделение списка свойств для каждого понятия.

Исходными данными для данного модуля является корпус текстов – набор файлов формата TXT в кодировке ANSI.

Термины выделяются на основе статистики. Для статистической обработки используется библиотека KLAN.

Свойства понятий согласно вышеизложенному алгоритму выделяются двумя способами:

- На основе статистики совместной встречаемости (используется библиотека KLAN);
- С использованием продукционных правил (используется библиотека Diglex DSL).

Рисунок 3 представляет пример формального контекста, построенного программой (входной корпус текстов по теме «Млекопитающие»).

	жизнь	домашний	дикий	полезный	форма	морской	вид	царство	низкий	водный
животное	X	X	X	X	X	X	X	X	X	X
тело					X					
организм										
лошадь		X								
птица		X								
мышь		X								
кошка		X	X							
размножение										
яйцо										
лев						X				
собака		X								
область										
вид										
млекопита...										
обезьяна										
человек										
звук										
хвост										
тигр										
зверь			X							

Рисунок 3 Фрагмент формального контекста

4.3 Модуль построения решетки формальных понятий

Модуль построения решетки формальных понятий использует формальный контекст, созданный модулем построения таблицы «объект-свойство». Модуль строит решетку формальных понятий. Так как формальное понятие представляет собой два множества: множество объектов и множество свойств, то имя класса онтологии, в который переходит данное формальное понятие, может дать только эксперт.

Пример построенной решетки формальных понятий изображен на рисунке 4. Из рисунка видно, что при построении решетки генерируется большое количество формальных понятий.

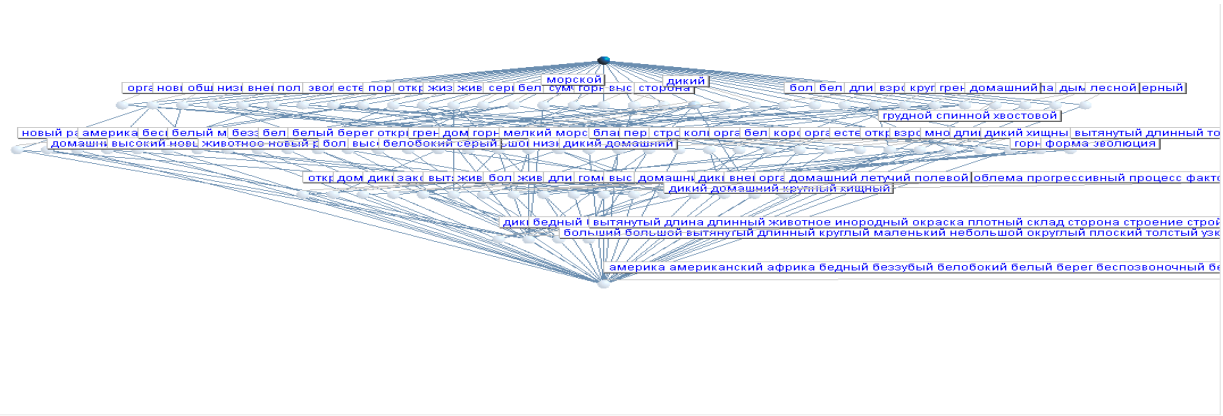


Рисунок 4 Решетка формальных понятий

На рисункахРисунок ***5*** и Рисунок ***6*** представлены фрагменты решетки формальных понятий. Красным цветом обозначены объекты, входящие в данное понятие (объем понятия), а синим – свойства, общие для этих объектов (содержание понятия). На рисунке Рисунок ***5*** выделено два формальных понятия. Понятие с содержанием {дикий, хищный} является подпонятием для понятия с содержанием {дикий}, таким образом в онтологии два класса: «дикие» и «дикие хищники», связанные таксономическим отношением.

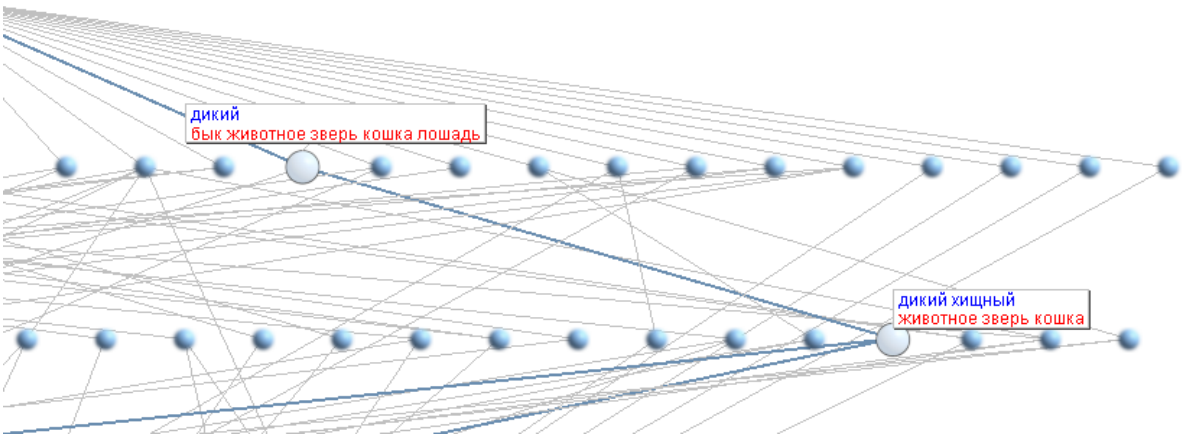


Рисунок 5 Фрагмент решетки формальных понятий

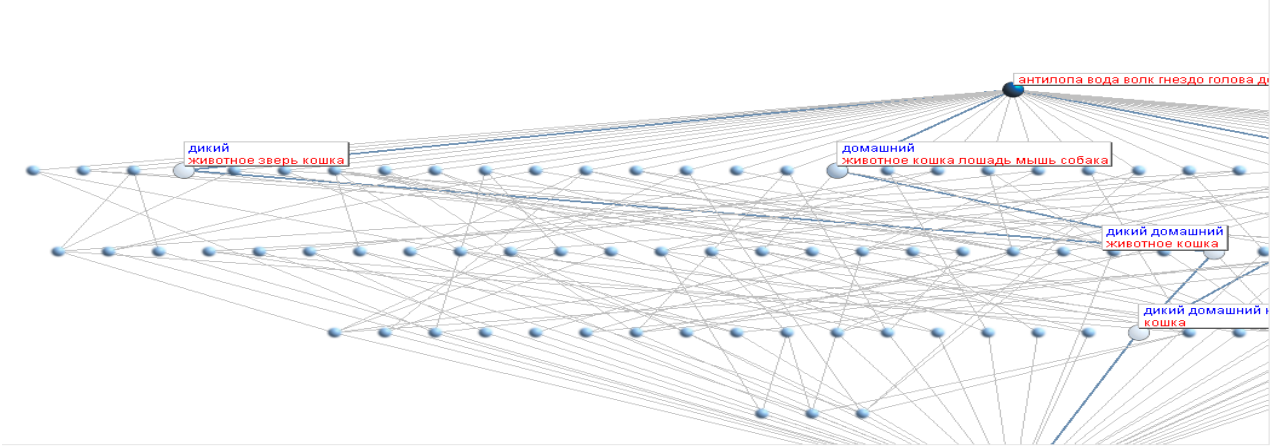


Рисунок 6 Фрагмент решетки формальных понятий

На рисунке Рисунок ***6*** формальное понятие с объемом {кошка} является подпонятием для формального понятия с содержанием {дикий, домашний}. Классы онтологии, которые можно сопоставить данным формальным понятиям, изображены на рисунке Рисунок ***7***.

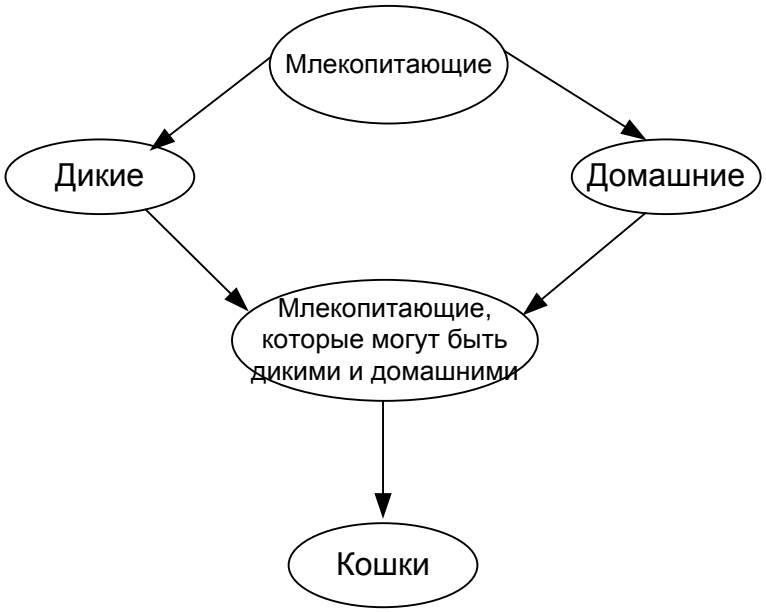


Рисунок 7 Фрагмент онтологии

Таким образом, на данном этапе программа строит решетку формальных понятий, понятия которой соответствуют классам таксономического ядра онтологии.

4.4 Интерфейс пользователя

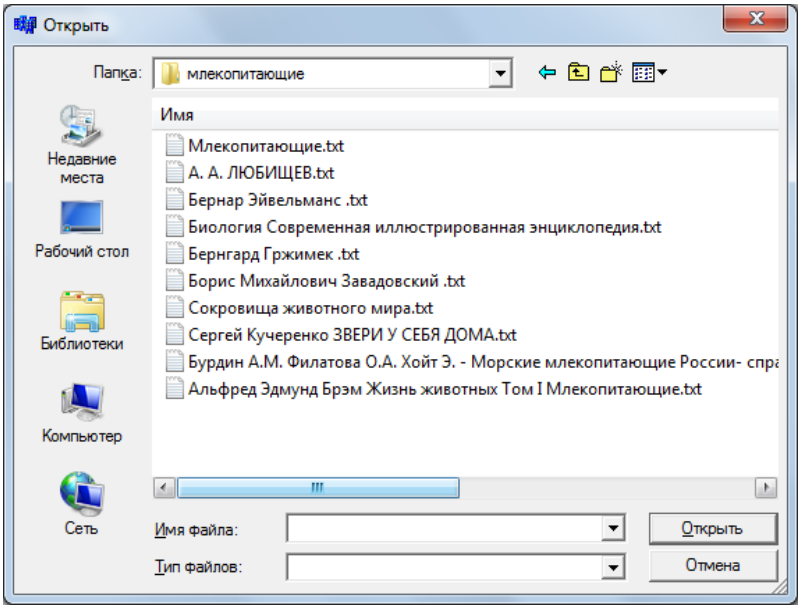
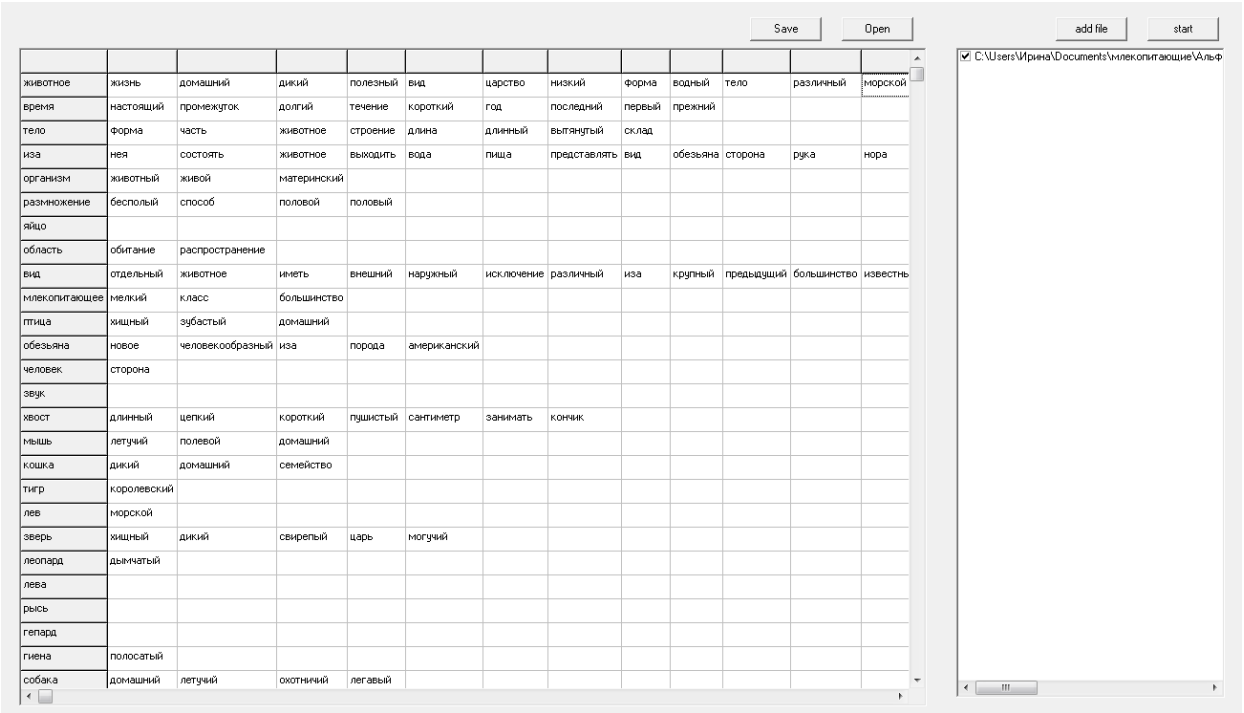


Рисунок 8 Добавление файлов в корпус текстов

Программная система для автоматизации построения онтологии на основе корпуса текстов имеет графический интерфейс пользователя. Пользователь может добавлять/удалять текстовые файлы в корпус текстов. После первого этапа работы системы пользователь может просмотреть результат работы программы в виде таблицы, отображающей объекты и их свойства. При необходимости можно удалять или добавлять новые свойства.



животное	жизнь	домашний	дикий	полезный	вид	царство	низкий	форма	водный	тело	различный	морской
время	настоящий	промежуток	долгий	течение	короткий	год	последний	первый	прежний			
тело	форма	часть	животное	строение	длина	длинный	вытянутый	склад				
иза	нея	состоять	животное	выходить	вода	пища	представлять	вид	обезьяна	сторона	рука	нора
организм	животный	живой	материнский									
размножение	бесполый	способ	половой	половый								
яйцо												
область	обитание	распространение										
вид	отдельный	животное	иметь	внешний	наружный	исключение	различный	иза	крупный	предыдущий	большинство	известны
млекопитающее	мелкий	класс	большинство									
птица	хищный	зубастый	домашний									
обезьяна	новое	человекообразный	иза	порода	американский							
человек	сторона											
звук												
хвост	длинный	цепкий	короткий	пушистый	сантиметр	занимать	кончик					
мышь	летучий	полевой	домашний									
кошка	дикий	домашний	семейство									
тигр	королевский											
лев	морской											
зверь	хищный	дикий	свирепый	царь	могучий							
леопард	дымчатый											
лева												
рысь												
гепард												
гиена	полосатый											
собака	домашний	летучий	охотничий	легавый								

Рисунок 9 Интерфейс пользователя. Просмотр терминов и слов.

Заключение

В ходе данной квалификационной работы были выполнены следующие задачи:

1. Выполнен анализ существующих методов автоматического построения онтологии.
2. Выбраны методы, подходящие для автоматизации построения таксономического ядра онтологии.
3. Разработан алгоритм автоматического построения таксономического ядра онтологии на основе корпуса текстов.
4. Предложенный алгоритм реализован в виде программы с графическим интерфейсом пользователя.

Анализ существующих методов автоматического построения онтологий показал, что, хотя в данном направлении существует достаточно много исследований, на данный момент не существует приемлемого решения проблемы автоматического построения онтологии предметной области путем извлечения знаний из текстов на естественном языке.

Практическим результатом данной работы является программная система для автоматизации построения таксономического ядра онтологии на основе корпуса текстов. Данная программа дает пользователю возможность следить за ходом построения ядра онтологии и предоставляет удобный графический интерфейс.

В дальнейшем планируется исследование возможной оптимизации автоматически построенного формального контекста, а также способов редуцирования решетки формальных понятий с целью уменьшения числа формальных понятий, затрудняющих выделение существенных черт ПрО.

Также дальнейшее развитие данной работы может быть связано с исследованием методов извлечения различных типов отношений и расширением функциональности системы до возможности извлечения таких отношений.

Литература

1. Ganter Bernhard. Formal Concept Analysis: Mathematical Foundations / Bernhard Ganter, Rudolf Wille. — Springer-Verlag New York, 1997.
2. Gruber Thomas R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing // International Journal of Human-Computer Studies. — 1992. — С. 907-928.
3. Амурский, К.А. Проблема извлечения знаний в информационных системах / К.А. Амурский, В.В. Дрождин, Ю.Н. Слесарев // Известия ПГПУ им. В.Г.Белинского. – 2010. – №18 (22) – С. 96-98.
4. Загорулько Ю.А. Применение онтологий для поиска информации в Интернет / Ю.А. Загорулько, О.И. Россеева, Л.И. Гладкова // Труды III-й международной конференции "Проблемы управления и моделирования в сложных системах" – Самара: Самарский Научный Центр РАН, 2001. –С. 503-508..
5. Клещев А.С. Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология» / А.С. Клещев, И.Л Артемьева // Научно-техническая информация, серия 2 «Информационные процессы и системы». – 2001. – № 2. – С. 20-27..
6. Мозжерина Е. С. Автоматическое построение онтологии по коллекции текстовых документов // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции – RCDL 2011 – Воронеж, 2011 – С. 293 – 298.
7. Найханова Л.В. Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования: Монография / Л.В. Найханова –Улан-Удэ: Изд-во БНЦ СО РАН, 2008. – 244 с.
8. Оробинская Е.А. Метод FCA для построения онтологии на основе текстового корпуса / Е.А.Оробинская, Н.В. Шаронова // БИОНИКА ИНТЕЛЛЕКТА. — 2011. — 2 (76). — С. 129–135.
9. Смирнов С.В. Построение онтологий предметных областей со структурными отношениями на основе анализа формальных понятий / С.В. Смирнов // Знания - Онтологии - Теории: Труды Всероссийской конференции с международным участием ЗОНТ-2011. — Новосибирск : Институт математики СО РАН, 2011. — Т. 2. — С. 103-112.