Московский Государственный Университет Путей сообщения

**(МИИТ)**

# Кафедра АСУ

## Курсовой проект

**«Разработка программных средств конвертирования HTML текстов в семантическую сеть»**

**Выполнила:**

Студентка 5-го курса группы УИС-511 Болотова Е.А.

**Проверил:** Саркисян Р.Е.

### Москва 2001

**Содержание:**

#### Что такое семантическая сеть …………………………..3

Основные сведения о языке HTML …………………….3

1.Введение ………………………………………….3

2.Сруктура HTML-документа ……………………..4

Что такое фреймы ……………………………………….7

Возможности представления знаний на базе

языка HTML ……………………………………………...8

TextAnalyst 2.0 – персональная система

автоматического анализа текста ………………………..14

Принцип работы HTML-конвертора …………………...17

Список использованных источников …………………..21

**Что такое семантическая сеть**

Семантическая сеть – структура для представления знаний в виде ориентированного графа, в котором вершины – это понятия, а дуги - отношения. Термин семантическая означает "смысловая", а сама семантика – это наука, устанавливающая отношения между символами и объектами, которые они обозначают, что есть наука, определяющая смысл знаков.

Самые первые семантические сети были разработаны в качестве языка-посредника для систем машинного перевода. Однако последние версии семантических сетей стали более мощными и гибкими и составляют конкуренцию логическому программированию, фреймовым системам и другим языкам представления.

На сегодняшний день существует множество вариантов семантических сетей. Их терминология и структура различаются, но существуют сходства, присущие всем семантическим сетям:

1. Узлы семантических сетей представляют собой концепты предметов, событий, состояний
2. различные узлы одного концепта относятся к различным значениям, если они не помечены как относящиеся к одному концепту
3. дуги семантических сетей создают отношения между узлами-концептами (пометки над дугами указывают на тип отношения)
4. некоторые отношения между концептами представляют собой лингвистические падежи, такие как агент, объект, реципиент и инструмент (другие означают временные, пространственные, логические отношения и отношения между отдельными предложениями
5. концепты организованы по уровням в соответствии со степенью обобщенности так, как, например, сущность, живое существо, животное, плотоядное.

Несмотря на некоторые различия, сети удобны для чтения и обработки компьютером, а также достаточно мощны, чтобы представить семантику естественного языка.

Наиболее часто в семантических сетях используются следующие отношения:

* связь типа "часть-целое" ("класс-подкласс", "множество-подмножество" и т.п.)
* функциональные связи, определяемые обычно глаголами "производит", "принадлежит" и т.п.)
* количественные ("больше", "меньше", "равно" и т.п.)
* пространственные ("близко от", "далеко от" и т.п.)
* временные ("раньше", "позже" и т.п.)
* логические связи ("и", "или" и т.п.)
* лингвистические связи и т.д.

**Основные сведения о языке HTML**

1. **Введение**

Все стандартные броузеры для сети Интернет используют способы представления текстов, основанные на языке HTML. HTML (Hyper Text Markup Language) – это язык разметки гипертекста. Этот язык «понимают» все компьютеры, он довольно прост, но при этом имеет достаточные выразительные средства для удобного описания разных типов документов. Язык позволяет хранить текст в «чистом» виде (не кодируя его), что делает возможным просмотр HTML скриптов с помощью обычных текстовых редакторов. Этот язык предоставляет авторам Интернет - публикаций средства:

* представления документов, включающих заголовки, тексты, таблицы, списки, «картинки» и т.п. элементы;
* осуществления навигации по отдельным документам и множеству документов путем использования гиперссылок;
* конструирования диалоговых форм для взаимодействия с удаленными сервисами, доступными в сети;
* включения в документы вычисляемых форм (spread-sheets), видео и звука, равно как и разнообразных приложений.

Первая версия языка HTML была разработана Т. Бернерс-Ли из Европейского Центра ядерных исследований (CERN). В дальнейшем язык претерпел существенные изменения. К середине 90-х годов произошла стандартизация его версий, которая стала курироваться международными организациями. В настоящее время наиболее развитой является версия языка HTML 4.0, в которой представлены новые возможности аппаратуры и требования производителей программного обеспечения броузеров, а также пожелания Интернет – авторов.

**2. Структура HTML-документа**

В HTML – файле находится символьная информация. Часть ее – это данные, составляющие содержимое документа, а другая часть – HTML – теги, языковые конструкции, используемые для разметки документа и управляющие его отображением. Для выделения тегов в тексте HTML – документа эти конструкции берутся в угловые скобки. Обычно теги используются парами: открывающий и закрывающий тег.

Типичный HTML – документ имеет следующую структуру:

< ! DOCTYPE HTML PUBLIC “-//W3C//DTD HTML 4.0//EN”

“http://www.w3.org/TR/REC-html40/strict.dtd”>

<HTML>

<HEAD>

<TITLE> Наименование документа </TITLE>

<META> name=keywords content=”Ключевые слова”>

</HEAD>

<BODY>

Тело документа

</BODY>

</HTML>

В приведенном фрагменте комментарий <!DOCTYPE …> фиксирует текущее состояние спецификации версии языка HTML. Документ должен открываться тегом <HTML>, закрываться тегом </HTML> и состоять из двух частей: заголовка и тела документа.

Заголовок находится между тегами <HEAD> и </HEAD> и содержит служебную информацию. В представленном выше фрагменте HTML – документа показаны два компонента заголовка:

* <TITLE> Заглавие </TITLE> - строка символов, которая отображается в заголовке окна броузера.
* <META> - дополнительная информация об HTML – документе (в нашем примере этот тег с помощью параметров name и content фиксирует значение первого атрибута как keywords, а второго – как ключевые слова «Представление знаний» и «Мультиагентные системы»). Этот тег ориентирован на аннотирование Интернет – документов и значительно облегчает задачу индексирования их, например, с помощью сетевых роботов.

Кроме этих компонентов заголовок может содержать еще и другие теги:

* <BASE> - базовый адрес, используемый при обработке относительных URL,
* <LINK> - используется для связи HTML-документа с другими источниками данных.

Собственно содержание документа находится в теле между тегами <BODY> и </BODY>. Иногда вместо этих тегов можно встретить тег <FRAMESET>, который определяет специальный тип документа – Web-страницу с кадрами или фреймами.

Как правило, тело HTML-документа состоит из последовательности структурных единиц, базисными из которых являются:

* заголовки разного уровня (текст, заключенный между тегами <Hi> и </Hi>), где i может меняться от 1 до 6;
* параграфы (текст, заключенный между тегами <P> и </P>).

Это минимальные средства форматирования документа. Естественно, что в HTML эти средства значительно богаче (всевозможные выравнивания, табуляция, несколько видов списков, таблицы и т.д.).

Наиболее важными базовыми конструкциями языка HTML являются якоря. Синтаксически эти конструкции представлены тегами <A> и </A> с атрибутами name и href. Якорем является конструкция вида:

<A name=«Метка»>Текст</A>

Эта конструкция обеспечивает уникальное в пределах документа имя начала определенного фрагмента HTML-текста. При этом текст, заключенный между тегами <A> и </A>, как правило, задает семантически значимое наименование заголовка.

Для ссылок на помеченные таким образом части Интернет - документа используются конструкции следующего вида:

1. <A href=«#Метка»>Текст</A> - Задает локальную ссылку на часть документа, начинающуюся с указанной метки
2. <A href=«URL»>Текст</A> - Задает глобальную ссылку на документ в сети, однозначно идентифицируемый с помощью URL (Unified Resource Locator) – Интернет – адрес: имя домена, уточненное названием протокола и собственное имя документа, включая путь к нему в пределах данного домена.

Важными конструкциями HTML также являются таблицы с богатыми возможностями многоуровневых заголовков и формы, с помощью которых в язык введены средства обеспечения диалога с читателями Интернет - документов. Базисными конструкциями форм являются редактируемые текстовые поля, элементы выбора, различные кнопки и т.д.

Пример создания таблицы:

<TABLE>

<TR>

<TD> столбец1, строка1 </TD><TD> столбец2, строка1 </TD>

</TR>

<TR>

<TD> столбец1, строка2 </TD><TD> столбец2, строка2 </TD>

</TR>

Простейший пример создания формы:

<FORM METHOD=”POST” …>

<P> Вы можете ввести в поле одну строку

<INPUT NAME=”entry”>

</P> Для обработки результатов ввода нажмите кнопку

<INPUT TYPE=”submit” VALUE=”Принять запрос”>

</P>

</FORM>

Еще одна важная конструкция – это фреймы (frames). С их помощью можно разделить документ на части и представлять их в отдельных, неперекрывающихся областях экрана. Такое представление информации характерно для многооконных приложений.

Кроме этого имеются и другие конструкции разметки Интернет – документов. Среди этих конструкций надо отметить более четкое разделение между структурой документа и его представлением за счет использования таблиц стилей, скриптов, поддерживающих, в частности, создание динамических страниц, новый механизм интеграции текстовых и графических ссылок, исполняемый на стороне клиента и стандартизацию механизма подключения к HTML-документам базисных медиаобъектов и приложений.

##### Что такое фреймы

Термин фрейм был предложен в 70-е годы для обозначения структуры знаний для восприятия пространственных сцен. Фрейм – это абстрактный образ для представления некоего стереотипа восприятия. Фреймом также называется и формализованная модель для отображения образа.

Различают:

* фреймы-образцы (прототипы) – хранятся в базе данных
* фреймы-экземпляры – создаются для отображения реальных фактических ситуаций на основе поступающих данных
* фреймы-структуры – используются для обозначения объектов и понятий
* фреймы-роли
* фреймы-сценарии
* фреймы-ситуации

Обычно структура фрейма представляется как список свойств:

(ИМЯ ФРЕЙМА:

(имя 1-го слота: значение 1-го слота),

(имя 2-го слота: значение 2-го слота),

--------------------------

(имя N-го слота: значение N-го слота)

В качестве значения слота может выступать имя другого фрейма, так образуется сеть фреймов.

Существует несколько способов получения слотом значений во фрейме-экземпляре:

* по умолчанию от фрейма-образца (значение default)
* через наследование свойств от фрейма, указанного в слоте АКО (a kind of)
* по формуле, указанной в слоте
* через присоединенную процедуру
* явно из диалога с пользователем
* из базы данных.

**Возможности представления знаний на базе языка HTML**

Рассмотрим, каким образом HTML-документ может быть представлен в виде семантической сети. Нам необходимо выделить те конструкции языка, которые могут быть полезными для решения этой задачи.

Прежде всего, к числу таких конструкций относятся теги типа <TITLE>, <META…> и <A…>. Первый тег важен для фиксации семантики всего HTML – документа, так как текст, заключенный между тегами <TITLE> и </TITLE> чаще всего отражает его назначение или содержание.

Теги типа <META…> вводят имена атрибутов и их значения с помощью параметров name=”…” и content=”…”, а ссылки и якоря фиксируют отношения между частями одного документа или между отдельными документами.

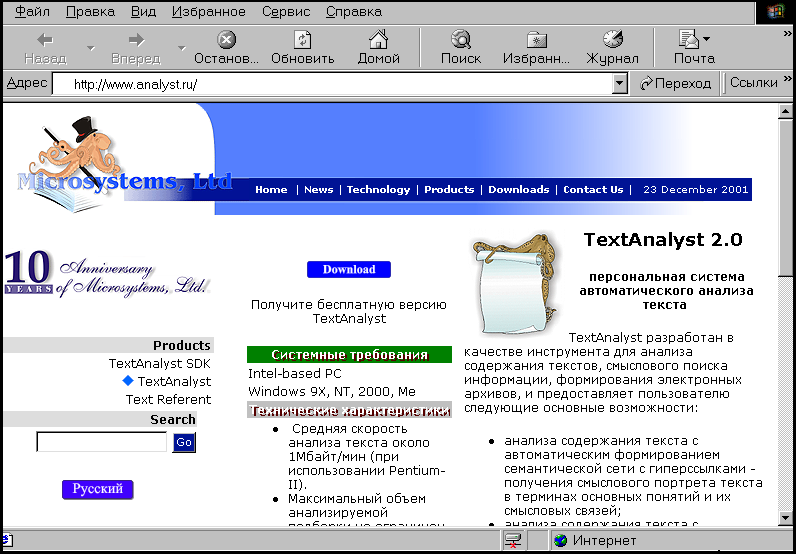
Теги типа <META…> явно вводят семантику значений атрибутов, одинаково интерпретируемых броузерами за счет ключевых слов, которые могут быть значениями параметра name.

Теги типа <A…> фиксируют лишь факт наличия отношения между ссылкой и ее якорем. В некоторых случаях этому отношению можно «приписать» имя SeeAlso (смотри также), в других случаях – ConsistOf, PartOf или иное подходящее имя, но семантика данной конструкции имплицитна, а встроенная интерпретация ее связана лишь с переходом по ссылке и визуализацией начала соответствующего фрагмента документа или загрузкой нового документа для просмотра.

Другими полезными конструкциями являются заголовки разделов и подразделов (тексты между тегами <Hi> и </Hi>), списки, таблицы и другие элементы языка.

Но в целом, выделение значимых для семантической интерпретации конструкций является экспертной задачей, решаемой каждый раз автором соответствующей Интернет - публикации по-своему. Но существуют определенные стереотипы. Например, на страницах Интернет – магазинов каталоги товаров в большинстве случаев представляются таблицами или списками, либо «зашиты» в чувствительные для щелчка мыши графические объекты. Это характерно и для индексов на сайтах машин поиска.

Рассмотрим в качестве примера страницу официального сайта компании Microsystems LTD, расположенную в сети по адресу http://www.analyst.ru. На этой странице располагается по информация по программе TextAnalyst 2.0. Экранная форма этой страницы показана на рисунке.



Фрагмент соответствующего HTML – текста представлен ниже:

<html>

<head>

---------------------------------

<meta name="KEYWORDS" content="Microsystems, TextAnalyst, text mining, knowledge discovery, textmining, e-commerce, classification, semantic analysis, neuro networks, natural linguistc, text processing, Микросистемы, анализ текстов, база знаний, документооборот, классификация, семантический анализ, нейронные сети, натуральные языки, текст процессор">

<meta name="GENERATOR" content="Microsoft FrontPage 4.0">

<meta name="ProgId" content="FrontPage.Editor.Document">

<title>Microsystems, Ltd</title>

<link rel="stylesheet" type="text/css" href="style.css">

</head>

-----------------------------------------------

<body topmargin="0" leftmargin="0">

<table border="0" cellspacing="0" cellpadding="0">

<tr>

<td valign="top" width="239"><a href="/index.php?lang=eng"><img

border="0" src="/images/top\_logo.gif"></a></td>

<td valign="top" align="left">

<table border="0" cellspacing="0" cellpadding="0">

<tr><td width="100%"><img border="0" src="/images/top\_up.gif">

</td></tr>

<tr> <td width="100%">

<table border="0" cellspacing="0" cellpadding="0" width="100%"

bgcolor="#001395" height="23">

<tr><td width="100%">

-------------------------------------

</table>

</td></tr>

</table>

<!-- end menu -->

----------------------------------------------

<!-- start menu here -->

<table border="0" cellspacing="0" cellpadding="0" width="100%">

<tr> <td width="241" valign="top" align="left">

<table border="0" width="100%" cellspacing="0" cellpadding="0">

<tr><td width="100%">

<p>&nbsp;</p>

<p align="center">

<img border="0" src="/images/10thyear\_s.gif" width="210" height="52">

</p>

</td></tr>

<tr><td width="100%" valign="top" align="left">

<!-- left menu-->

<table border="0" width="218">

<tr><td width="210" bgcolor="#DDDDDD" valign="top" align="left">

<p align="right"><b>Products</b></td>

</tr>

<tr onmouseout="this.style.backgroundColor='transparent'"

onmouseover="this.style.backgroundColor='#6B8ADE'">

<td width="210" align="right">

<a href="/index.php?lang=eng&dir=content/products/&id=body&left=

content/products/menu.txt " target="\_self">

<span style="color: #000000; text-decoration: none">

TextAnalyst SDK</span></a>

</td></tr>

<tr><td width="210" align="right">

<p align="right"><img border="0" src="/images/bd14580\_.gif" width="12"

height="12">

TextAnalyst</p>

</td></tr>

<tr onmouseout="this.style.backgroundColor='transparent'"

onmouseover="this.style.backgroundColor='#6B8ADE'">

<td width="210" align="right">

<a href="/index.php?lang=eng&dir=content/products/&id=

tref&left=content/products/menu.txt

" target="\_self">

<span style="color: #000000; text-decoration: none">

Text Referent</span></a>

</td></tr>

<!-- end left menu-->

----------------------------------------

</table>

<!-- end here -->

</td>

<!-- free space -->

<td valign="top" align="left">

&nbsp;&nbsp;&nbsp;&nbsp;

</td>

<!-- end free space -->

<td valign="top" align="left" width=100% >

<!-- content started here -->

<table border="0" cellspacing="6" cellpadding="0">

<tr> <td width="100%" valign="top" align="left">

<head>

<meta name="DESCRIPTION" content="TextAnalyst - personal text mining system">

<meta name="KEYWORDS" content="TextAnalyst, personal, text mining">

<title>TextAnalyst</title>

</head>

<div align="left">

<table border="0" align="left" cellspacing="4" cellpadding="3">

<tr> <td valign="top" align="center" width="250">

<p align="center">&nbsp;</p>

<p align="center"><a href="cgi-bin/stat/loadfile.pl?file=ta\_rus">

<img border="0" src="images/downloads.gif"></a></p>

<p align="center">Получите бесплатную версию

TextAnalyst</p>

<table border="0">

<tr><td width="100%" bgcolor="#008000">

<p align="center" class="menu">

<font color="#FFFFFF">Системные требования</font></td> </tr>

<tr><td width="100%">Intel-based PC</td> </tr>

<tr> <td width="100%">Windows 9X, NT, 2000, Me</td> </tr>

<tr> <td width="100%" bgcolor="#C0C0C0">

<p align="center" class="menu">

<font color="#FFFFFF">Технические характеристики</font></td>/tr>

<tr><td width="100%" valign="top" align="left">

<ul>

<li>&nbsp;Средняя скорость анализа текста около 1Мбайт/мин (при использовании Pentium-II).</li>

<li>Максимальный объем анализируемой подборки не ограничен и зависит от объема ресурсов компьютера и настройки TextAnalyst.</li>

<li>Собственный объем TextAnalyst не превышает 5Мб.</li>

<li>Форматы обрабатываемых файлов:</li>

<li>\*.txt (ANSI, DOS), \*.rtf</li>

<li>Экспорт информации в форматы: \*.txt,

\*.csw (электронные таблицы).</li>

</ul> </td></tr>

</table>

-----------------------------------

<h1 align="center">

<img border="0" src="../../images/octopus\_shaden.gif" align="left" width="99" height="112">TextAnalyst 2.0&nbsp;</h1>

<p align="center"><b>персональная система автоматического анализа текста&nbsp;</b></p>

<p>TextAnalyst разработан в качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов, и предоставляет пользователю следующие основные возможности:&nbsp;</p>

<ul>

<li>анализа содержания текста с автоматическим формированием семантической сети с гиперссылками - получения смыслового портрета текста в терминах основных понятий и их смысловых связей;&nbsp;</li>

<li>анализа содержания текста с автоматическим формированием тематического древа с гиперссылками - выявления семантической структуры текста в виде иерархии тем и подтем;&nbsp;</li>

<li>смыслового поиска с учетом скрытых смысловых связей слов запроса со словами текста;&nbsp;</li>

<li>автоматического реферирования текста - формирования его смыслового портрета в терминах наиболее информативных фраз;&nbsp;</li>

<li>кластеризации информации - анализа распределения материала текстов по тематическим классам;</li>

<li>автоматической индексации текста с преобразованием в гипертекст;&nbsp;</li>

<li>ранжирования всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;&nbsp;</li>

<li>автоматического/автоматизированного формирования полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации;&nbsp;</li>

</ul>

<p align="center"><b>Не пугайтесь обилия возможностей!</b></p>

<p align="center"><i>Работа с TextAnalyst покажется Вам неожиданно простой и приятной, а его аналитические способности сэкономят массу полезного времени...

</i></p></td> </tr>

</table></div></td> </tr>

</table>

----------------------------------

</body>

</html>

Сравнив приведенные экранную форму и HTML-текст, видим, что семантически значимыми элементами данного документа являются:

* ключевые слова, относящиеся к данному документу: Microsystems, TextAnalyst, text mining, knowledge discovery, textmining, e-commerce, classification, semantic analysis, neuro networks, natural linguistc, text processing, Микросистемы, анализ текстов, база знаний, документооборот, классификация, семантический анализ, нейронные сети, натуральные языки, текст процессор (тег <META>);
* все меню организованы в виде таблиц (тег <TABLE>), в ячейках которых (тег <TD>) расположены ссылки (тег <A href …>), с помощью которых можно перейти к другой интересующей информации. Например, можно получить информацию о продуктах данной компании, выбрав их название из левого меню.
* текст описания возможностей программы TextAnalyst организован в виде списка (тег <li>).

Т.о. можно видеть, семантически значимые характеристики документа могут быть разбросаны по разным частям документа или по разным документам. Это сильно затрудняет семантический анализ Интернет – документов.

Решение этой проблемы в настоящее время связано с использованием двух подходов. Первый подход предполагает, что семантическая разметка документа выполняется вручную его автором на основе специальных метатегов, а второй подход связан с автоматическим или полуавтоматическим преобразованием исходного текста в специальное семантическое представление. Целесообразно конвертировать HTML-тест в более удобную форму представления для дальнейшей обработки.

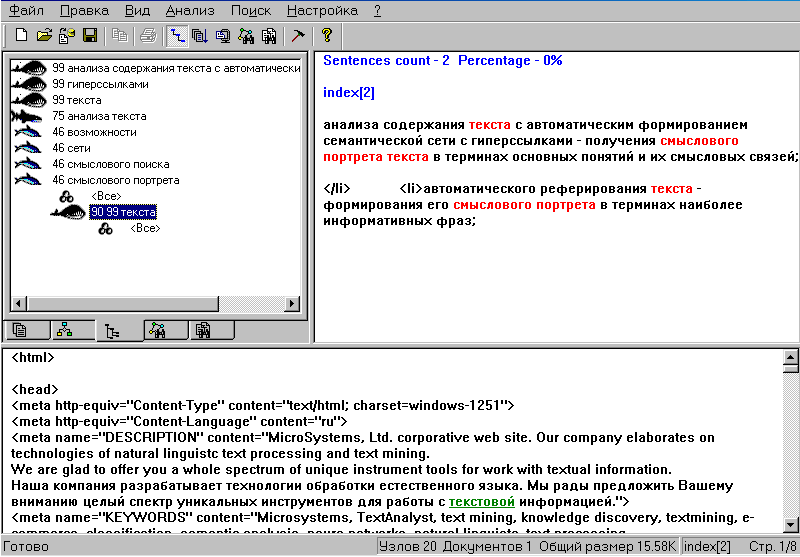
**TextAnalyst 2.0 – персональная система автоматического анализа текста**

TextAnalyst разработан в качестве инструмента для анализа содержания текстов, смыслового поиска информации, формирования электронных архивов, и предоставляет пользователю следующие основные возможности:

* анализ содержания текста с автоматическим формированием семантической сети с гиперссылками - получение смыслового портрета текста в терминах основных понятий и их смысловых связей;
* анализ содержания текста с автоматическим формированием тематического древа с гиперссылками - выявление семантической структуры текста в виде иерархии тем и подтем;
* смысловой поиск с учетом скрытых смысловых связей слов запроса со словами текста;
* автоматическое реферирование текста - формирование его смыслового портрета в терминах наиболее информативных фраз;
* кластеризация информации - анализ распределения материала текстов по тематическим классам;
* автоматическая индексация текста с преобразованием в гипертекст;
* ранжирование всех видов информации о семантике текста по «степени значимости» с возможностью варьирования детальности ее исследования;
* автоматическое/автоматизированное формирование полнотекстовой базы знаний с гипертекстовой структурой и возможностями ассоциативного доступа к информации.

Рассмотрим на нашем примере работу этой программы. После запуска TextAnalyst, необходимо открыть текстовый файл, в котором расположен HTML-документ нашего примера. Программа выполняет анализ предложенного текста и выдает результаты (см. рисунок)

Изучив предложенный текст, TextAnalyst формирует сеть наиболее значимых понятий, содержащихся в данном тексте. В такую включены те термины текста, которые несут основную смысловую нагрузку. Т.о. сеть позволяет отбросить несущественную информацию и представить содержание текста в сжатом виде. Каждое понятие, появляющееся множество раз в различных частях текста, в сети представлено единственным узлом. Различные формы слов для отображения в один узел сети представляются к общей грамматической форме.



Каждый элемент сети характеризуется числовой оценкой – смысловым весом. Связи между понятиями также характеризуются весами. Значение смыслового веса (от 1 до 100) показывает, насколько важную роль играет понятие для смысла всего текста, т.е. как много информации в тексте касается данного понятия. Максимальное значение, равное 100, говорит о том, что понятие является ключевым и представляет важнейшую тему текста. Маленькое, близкое к единице значение показывает, что соответствующая тема лишь вскользь упомянута в тексте и в нем очень мало информации, относящейся к данному понятию. Второе число, стоящее перед смысловым весом, ближе к раскрытому узлу, представляет вес связи от понятия в вершине раскрытого списка к данному. Большое значение веса связи (близкое к 100) указывает на то, что подавляющая часть информации в тексте, касающаяся первого, касается в то же время и второго понятия. Малое (близкое к 1) значение означает, что первое понятие слабо связано со вторым и очень мало информации по первой теме касается в тоже время и второй.

По умолчанию на экране отображаются понятия с весом не менее 5. Вид сети на экране можно настраивать, изменяя количество отображаемых понятий и связей, а также способ их сортировки.

TextAnalyst предоставляет услугу автоматического реферирования. Формируемый реферат содержит список наиболее информативных предложений текста. Это позволяет быстро ознакомиться с содержанием текста. Подробность реферата можно настраивать, изменяя количество формирующих его предложений. Каждое предложение характеризуется относительной степенью значимости во всем тексте.

В нашем примере реферат выглядит таким образом:

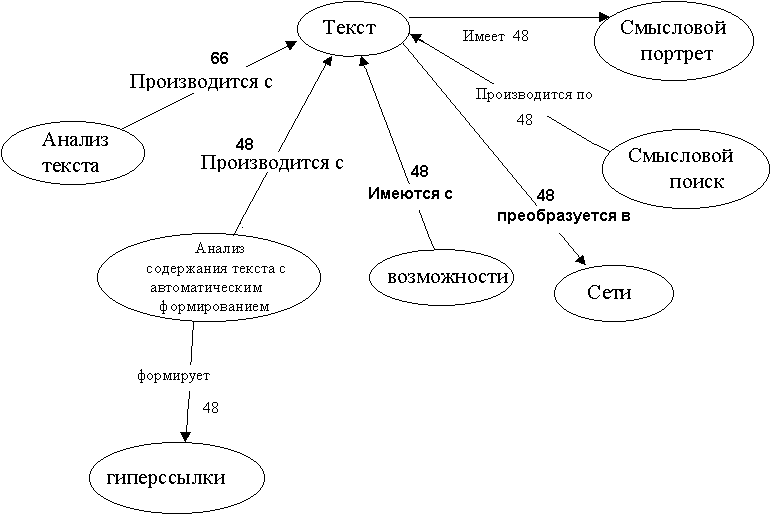
98 анализа содержания текста с автоматическим формированием семантической сети с гиперссылками - получения смыслового портрета текста в терминах основных понятий и их смысловых связей;

98 </li> <li>анализа содержания текста с автоматическим формированием тематического древа с гиперссылками - выявления семантической структуры текста в виде иерархии тем и подтем;

Цифры показывают степень значимости предложений в тексте. Значение веса, близкое к 100, означает, что данное предложение представляет важнейшую информацию, касающуюся главных понятий текста. Эти понятия в реферате выделяются цветом.

По умолчанию на экране отображаются предложения реферата с весами не менее 90.

Для рассматриваемого выше примераHTML-текста описания страницы Analyst.ru фрагменты семантической сети выглядят следующим образом:



**Принцип работы HTML-конвертора**

Рассмотрим в качестве примера следующее подмножество HTML-языка, которое может быть задано следующими определениями:

HTML-text :: = <HTML> HEAD BODY </HTML>

HEAD :: = TITLE{HEAD}| META{HEAD}| LINK{HEAD}…

TITLE :: = <TITLE> строка </TITLE>

META :: = <META name="KEYWORDS" content="строка">

KEYWORDS :: = …

BODY :: = <BODY> HTML-BODY </BODY>

HTML-BODY :: = PARAGRAPH{HTML-BODY} | TABLE {HTML-BODY} | LIST{HTML-BODY} | ANCHOR{HTML-BODY} | …

PARAGRAPH :: = <P> текст </P>

TABLE :: = <TABLE> TABLE-CELLS </TABLE>

###### TABLE-CELLS :: = STROKA{TABLE-CELLS} | …

STROKA :: = <TR> CELL </TR>

CELL :: = <TD> текст </TD>

LIST :: = <UL> LIST-ATOM </UL>

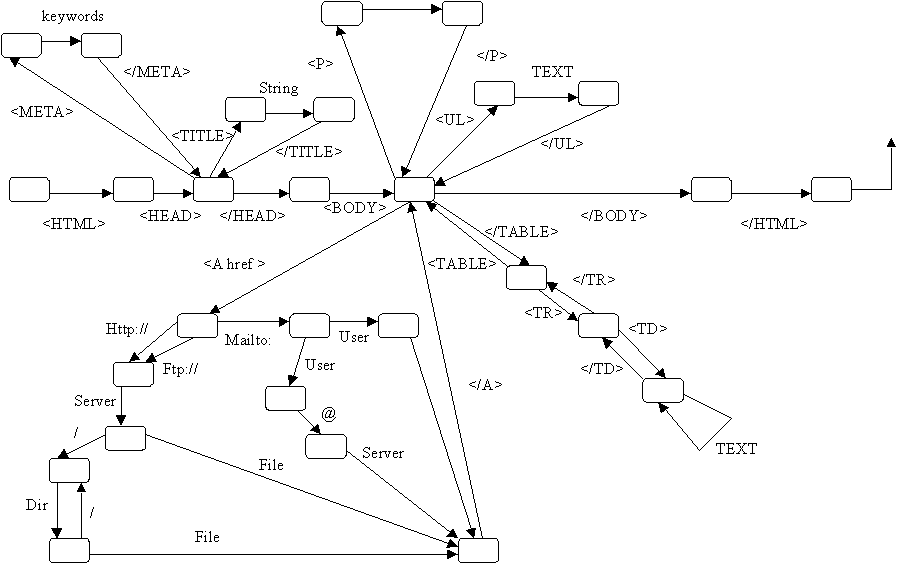
LIST-ATOM :: = …

ANCHOR :: = <A HREF=LINK> TEXT </A>

TEXT :: = …

LINK: = …

Синтаксическая диаграмма, соответствующая этим правилам выглядит следующим образом:



С теоретической точки зрения HTML – это простой язык программирования с контекстно-свободной грамматикой. Для анализа HTML-текстов можно использовать нисходящие распознаватели, реализуемые на базе метода рекурсивного спуска. Рассмотрим продукционно-фреймовый формализм представления знаний и разработку на его основе интеллектуальный HTML-конвертор.

Для начала необходимо задать регулярное отображение каждого правила спецификации HTML-конструкций в соответствующий объект базы знаний на уровне фрейма-прототипа. Система таких прототипов даст нам описание языка, а множество фреймов-экземпляров – спецификацию конкретных и синтаксически правильных HTML-текстов. Основные правила такого отображения таковы:

* каждому концепту из левой части BNF-определения ставим в соответствие имя фрейма-прототипа;
* альтернативам из правой части BNF-определения при этом должны соответствовать имена слотов этого фрейма;
* для концептов-нетерминалов соответствующий слот должен иметь тип frame;
* для концептов-терминалов соответствующие слоты будут, как правило, иметь тип numb или string;
* рекурсия в BNF-определениях заменяется итерацией, а соответствующие слоты становятся множественными.

После применения данных правил к BNF-определениям языка HTML получим следующее множество фреймов-прототипов:

[html is\_aprototype, if\_added HTML();

HEADframe, restr\_by head;

BODYframe, restr\_by body ];

[head is\_aprototype, if\_added HEAD();

BODY{frame}, restr\_by one\_of {title, meta, …}];

[title is\_aprototype, if\_added TITLE();

BODYstring ];

[meta is\_aprototype, if\_added META();

BODYstring ];

………………….

[body is\_aprototype, if\_added BODY();

SENT{frame}, restr\_by one\_of {header, paragraph, table, …}];

[paragraph is\_aprototype, if\_added PARAGRAPH();

[LIST is\_aprototype; ATOM{frame}, if\_added LI() ];

BODYframe, restr\_by text];

[table is\_aprototype; if\_added TABLE();

TAB{frame}, restr\_by one\_of {stroka,…};]

[stroka is\_aprototype, if\_added TR();

CELLS{frame}, restr\_by one\_of{cell,…}];

[cell is\_aprototype, if\_added TD();];

……………………

[anchor is\_aprototype;

BODYframe, restr\_by text];

……………………

[link is\_aprototype;

URLframe, restr\_by one\_of {http,ftp,…}];

MAILframe, restr\_by mail];

[url is\_alink; without\_slot MAIL];

[http is\_aurl, if\_added HTTP();

SERVERstring;

DIR{string};

FILEstring];

[ftp is\_aurl, if\_added FTP();

SERVERstring;

DIR{string};

FILEstring];

…………………………

В соответствии с приведенными фреймами-прототипами и синтаксическими диаграммами, можно специфицировать процедурную часть конвертора как систему демонов, присоединенных к фреймам или к их слотам.

Спецификация одного из таких демонов представлена ниже на языке Java:

public class HTML extends FramePrototype {

HEAD head=null;

BODY body-null;

………….

String keyword;

Public void HTML (String name) {

Super (name);

keyword=getToken();

if (keyword.compareTo (“<HTML>”) = =0 {

head = new HEAD (getNewName());

body = new BODY (getNewName());

};

keyword = getToken ();

if (keyword.compareTo (“</HTML>”)= =0) return;

}

…………………

Public void (String nam) {

Super (name);

Keword=getToken();

If (keyword compareTo (“<BODY>”) = = 0 {

paragraph = new PARAGRAPH (getNewName());

header = new HEADER (getNewName());

table = new TABLE (getNewName());

};

keyword = getToken ();

if (keyword compareTo (“</BODY>”) = = 0) return;

}

…………………………..

}

По существу, такой демон является конструктором класса HTML, а запуск конвертора осуществляется с помощью оператора создания нового объекта этого класса:

HTML currPage = new HTML (get\_new\_name());

При этом будут рекурсивно вызываться конструкторы других классов (на верхнем уровне это HEAD, BODY), что, в конечном счете, приведет к построению множества фреймов-экземпляров, представляющих анализируемую HTML-страницу.

Получение полезной в дальнейшем базы знаний предполагает дальнейшую семантическую интерпретацию фреймового представления и построение в конечном счете семантической сети, отражающей смысл исходного Интернет – документа.

**Список использованных источников:**

* 1. Т.А. Гаврилова, В.Ф. Хорошевский «Базы знаний интеллектуальных систем», учебник, Санкт-Петербург, «Питер», 2001
  2. www.citforum.ru
  3. www.bur.oivta.ru
  4. www.analyst.ru