В.Б. Мордухович (Минск, МГЛУ)

СПОСОБЫ ФОРМИРОВАНИЯ ГИПЕРТЕКСТА

Термином **гипертекст** первоначально обозначалась технология работы с текстовыми данными, позволяющая устанавливать ассоциативные связи между отдельными фрагментами текста и благодаря этому допускающая не только последовательную работу с текстом, как при обычном чтении, но и произвольный доступ к его компонентам. В настоящее время гипертекст можно определить как не линейно организованный текст. Нелинейность гипертекста означает, что после предъявления каждого фрагмента текст как бы ветвится — для дальнейшего чтения читателю предлагается на выбор несколько возможных продолжений. Элементами данной структуры являются обычные линейные тексты или их фрагменты.

Просмотр гипертекста осуществляется путем следования от узла к узлу по выбираемым связям — гиперссылкам, которые пользователь видит во время ознакомления с содержанием узла. Гипертекст отличается специфическим способом обращения. Информация для гипертекста организована и представлена специальным образом: пользователь постоянно видит перед собой на экране только фрагмент текста. Однако с помощью курсора «мыши» он может перемещаться по выделенным словам-ссылкам, переходить по нажатию клавиш к подробной информации о выделенном слове-ссылке, просматривать оглавление всего гипертекстового документа, возвращаться к ранее просмотренным фрагментам.

Проблема разработки гипертекстовых систем на данный момент очень важна. Налицо внедрение гипертекстовых технологий в программное обеспечение, начиная с операционных систем и прикладных программ и заканчивая глобальной сетью Интернет, где гипертекст является основным способом представления информации. Обычно гипертекст создается вручную. При построении различных справочных систем, электронных книг, каталогов, верстке Web-сайтов и т.п. автор сам расставляет ссылки, руководствуясь определенными целями и задачами, и, тем самым, определенным образом структурирует предложенную в тексте информацию. Однако процесс ручного выделения гипертекстовых переходов весьма трудоемок и занимает много времени. Поэтому важным шагом в создании гипертекстовых систем является автоматизация этого процесса.

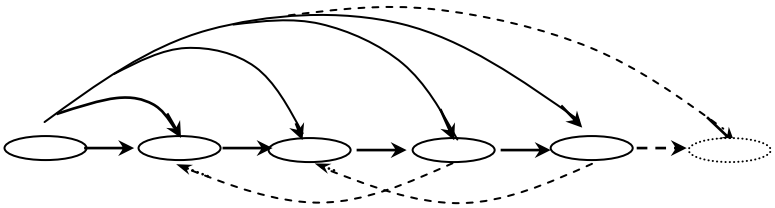
Существуют два основных способа формирования гипертекста, которые непосредственным образом связаны со способом организации связей между их фрагментами: 1) структурирование линейного текста и 2) расширение линейного текста [1].

Процесс структурирования текста состоит из следующих трех основных шагов [2, с. 105]:

Шаг 1. Деление текста на отдельные главы / темы:

Шаг 2. Представление некоторого основного пути чтения гипертекста и расстановка ссылок, ведущих читателя от темы к теме по этому основному пути:

Шаг 3. Выделение в тексте слов-ссылок, по которым пользователь может перейти от основного пути чтения текста к другим возможным путям чтения:



Основной трудностью в процессе автоматического структурирования текста является формализация процесса выделения его семантических блоков/узлов и определения отношений между ними. Это предполагает детальный семантический анализ текста. Как известно, одним из основных параметров, характеризующих план содержания текста, является его семантическая связность. Р.К. Потапова отмечает, что «в качестве критерия семантической связи между предложениями принимается повторение одних и тех же существительных при условии замены личных местоимений их антецедентами. Лексическое значение слова — наименьший элемент семантической структуры текста. Слово выступает в качестве основного компонента всех более крупных единиц текста: синтагм (предикативных и непредикативных), предложений, абзацев, параграфов. Каждая единица представляет собой определенного рода сочетание единиц низшего уровня. Значение каждой из этих единиц может рассматриваться как элемент семантической структуры текста» [3, с. 268]. Таким образом, семантическая структура текста включает несколько уровней, соответствующих указанным элементам. Поэтому адекватное описание этой структуры должно отразить ее многоуровневость. «При анализе семантической структуры текста тип семантической структуры высшего уровня (текста в целом) определяется семантическими связями между абзацами, а тип семантической структуры среднего уровня (абзаца) — семантическими связями между предложениями. Для определения типа семантической структуры текста по семантической связи строится поабзацная семантическая сеть. Критерий семантической связи между абзацами — наличие семантической связи между предложениями, входящими в эти абзацы. Форма поабзацной семантической сети позволяет определить тип семантической структуры текста так же, как форма пофразной сети позволяет определить тип семантической структуры абзаца (или всего короткого текста)» [3, с. 269–270].

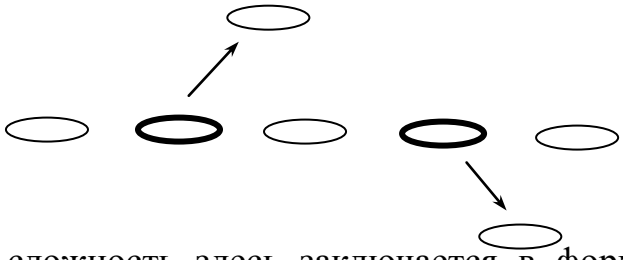
Описанный подход анализа и представления плана содержания связного текста может использоваться при автоматическом индексировании и реферировании текстов. Он может также применяться при автоматическом

анализе плана содержания текстов во время создания и модификации гипертекстовых систем и, следовательно, для автоматической генерации семантических связей между текстами / фрагментами гипертекста.

Технология построения гипертекста путем расширения линейного текста выглядит следующим образом:

Шаг 1. В тексте нужно выделить те слова-ссылки, которые нуждаются в толковании, комментарии, пояснении, либо могут служить отправным пунктом для перехода к другому документу:

Шаг 2. Нужно связать эти слова-ссылки с поясняющими их элементами, документами, находящимися за пределами данного текста:



Основная сложность здесь заключается в формализации процесса выделения узловых моментов текста для перехода к другому документу. Это может быть слово, словосочетание, обращение к первоисточнику и т.д. Прежде всего, следует отметить, что ссылки могут быть двух типов: 1) поясняющие ссылки и 2) ссылки на другие документы. Поясняющие ссылки дополняют, поясняют, раскрывают значение выделенного слова, могут служить комментарием. Такие слова выделить несложно. Обычно они задаются заранее в виде списка терминов и аббревиатур определенной предметной области. Например, в технических текстах по машиностроению встретилось слово *gear*. Кликнув по этой ссылке, можно перейти к следующему пояснению: *set of toothed wheels working together in a machine, esp. such a set to connect the engine of a motor-vehicle with the road wheels*. Что касается ссылок на другие документы, то принцип их выделения будет значительно отличаться от описанного выше метода. Его можно применять при работе с текстами практически любой предметной области: с нормативными, техническими, военными, экономическими документами и т.д. Рассмотрим принцип работы такого автоматического анализатора на примере текстов англоязычных юридических документов. В них потенциальные связи выделены определенными языковыми конструкциями, как правило, содержащими следующие атрибуты: «название документа», «вид документа», «номер документа», «дата принятия, регистрации или подписания документа». Например: *according to the amendment №4 to the Constitution of the United States, passes State Standard 12.307-94 , acting on the basis of the Regu-lation* и т.д. Анализатор выделяет варианты атрибутов документа и их местоположение в документе, откуда должен осуществляться переход. Далее происходит поиск новых документов по выделенным атрибутам и среди них отбираются наиболее вероятные варианты.

В рамках данного исследования предлагается несколько иная формальная модель системы автоматического выделения гипертекстовых переходов в англоязычных текстах. Она опирается на важные в семантическом плане единицы текста. На их основе и создаются поясняющие ссылки.

ЛИТЕРАТУРА

1. Bernstein, Mark. Patterns of hypertext // Pro-ceedings of Hypertext [Electronic resource]. — Mode of access: http://www/easgate.com/patterns/print.html. — Date of access: 14.01.2012.

2. Ованесбеков, Л.Г. Технология построения гипертекстов: дис. …канд. физико-мат. наук: 05.13.11. — М., 1993.

3. Потапова, Р.К. Новые информационные технологии в лингвистике. — М., 2002.