ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ СТРУКТУРИРОВАНИЯ И ИНДЕКСАЦИИ БОЛЬШИХ МАССИВОВ ИНФОРМАЦИИ

Кулюлина Н.

г. Москва, семейное образование, 10 класс

Научный руководитель: Хачатурьян Л.П., ЦОМТМП, г. Москва

В статье «Информационная технология поиска, отбора и классификации больших массивов информации» [1] была опубликована система взаимосвязанных алгоритмов поиска, отбора и классификации больших массивов малосвязанной и слабоструктурированной информации в сети Интернет, по сути являющейся информационной технологией. Суть данной информационной технологии состоит в том, что большая часть действий пользователя при поиске, отборе, классификации больших массивов информации максимально формализованы, выполняются по стандартным схемам «не задумываясь»; основные решения принимаются пользователем на завершающем этапе работы.

В результате применения данной информационной технологии формируется первичная база информационных объектов ([1], рис. 1). Простой информационный объект – это объект, который нельзя поделить на более мелкие доступными техническими средствами, причем без нарушения интересов правообладателей. Составной информационный объект – объект, не являющийся простым и обладающий смысловой предметной цельностью. Отложенный информационный объект – это объект, который требует дополнительных сложных технических процедур и/или урегулирования с правообладателями. Разработка информационной технологии была не завершена в части анализа сложных источников информации – веб-страниц 3-го типа и сайтов (терминология из [1]).

Целью настоящей работы является разработка дополнений, которые следует внести в упомянутую информационную технологию, а также механизмов дальнейшей работы с полученной базой информационных объектов.

Требуется решить следующие задачи:

1. Разработка алгоритмов анализа веб-страниц 3-го типа и сайтов и выделения информационных объектов – простых, составных и отложенных.
2. Разработка алгоритмов структурирования и индексации найденной информации для ее дальнейшего использования.
3. Формулирование требований к плагину для браузера, автоматизирующего ручные процессы.

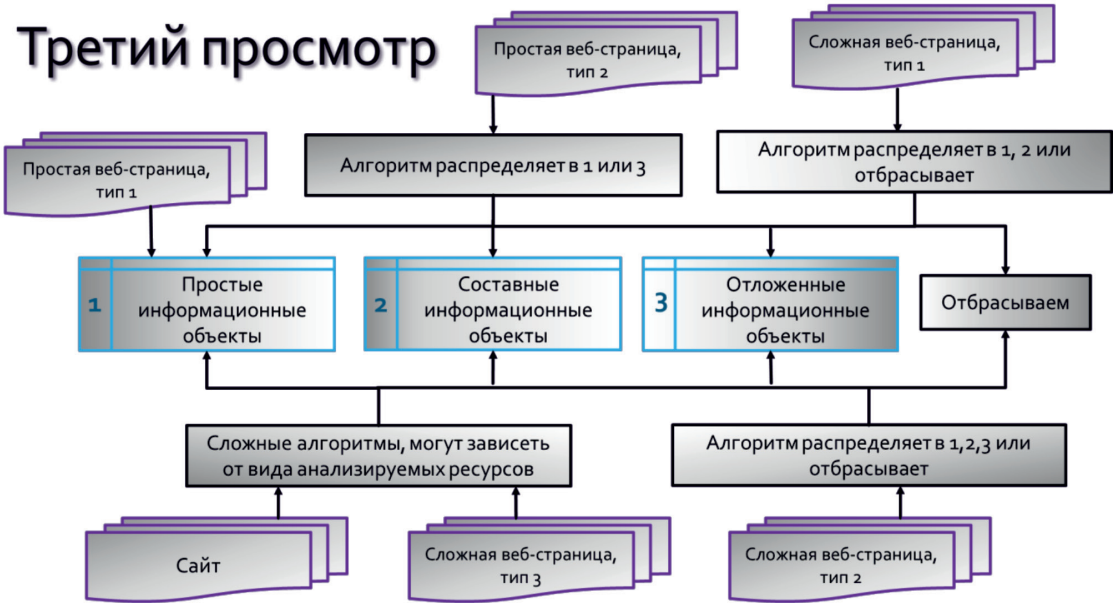

Рис. 1. Алгоритм выполнения третьего просмотра

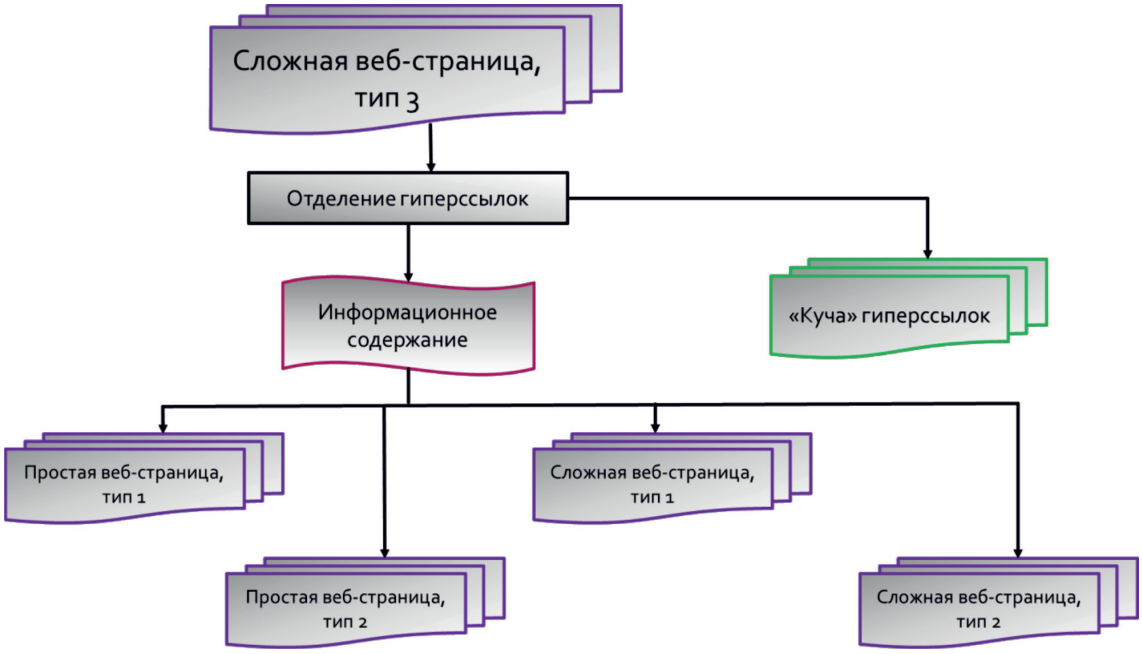

Рис. 2. Алгоритм анализа сложных веб-страниц 3-го типа во время 3-го просмотра

В результате применения данной информационной технологии будет сформирована первичная структурированная и проиндексированная база информационных объектов, готовая к дальнейшему использованию.

Выделение информационных объектов из сложных источников информации

Алгоритмы выделения информационных объектов из сложных веб-страниц третьего типа представлен на рис. 2.

Сложные веб-страницы 3-го типа содержат в себе ссылки на другие информационные объекты по интересующей нас теме. Собственное информационное содержание такой страницы соответствует простой веб-странице 1-го или 2-го типов или сложной веб-странице 1-го или 2-го типов. В моем случае это видео-опыт(ы) и дополнительные к ним материалы [1, рисунок 6].

Первым шагом необходимо отделить гиперссылки, сохранить их в закладках браузера – формируется «куча» гиперссылок. Дальнейшие действия с «кучей» гиперссылок выполняются по стандартным алгоритмам 2-го этапа информационной технологии поиска, отбора и классификации больших массивов информации [1] начиная с 1-го просмотра.

Собственное информационное содержимое веб-страницы, очищенное от гиперссылок, разбирается на информационные объекты по заданным алгоритмам 3-го просмотра 2-го этапа информационной технологии поиска, отбора и классификации больших массивов информации ([1], рис. 1).

Алгоритм для сайта (собрания веб-страниц) состоит из 2-х этапов:

1. Разбор сайта на перечень веб-страниц разных типов, сохраняемых по-отдельности, с последующим исключением повторов среди веб-страниц.
2. Обработка каждой из веб-страниц в соответствии с ранее описанными алгоритмами. Далее проводится процедура исключения повторов среди полученных информационных объектов.

Содержательный просмотр, структурирование и индексация информации

При этом просмотре состав информационных объектов остается неизменным. В отличие от предыдущих, предметно-независимых этапов, содержание этого этапа предметнозависимо: характер структурирования и индексации информационных объектов зависит от специфики предметной области и решаемой пользователем задачи. В моей предметной области – видео-опыты по школьной неорганической химии за 8-9 классы – главным структурообразующим фактором является тип химической реакции или проводимого процесса в широком смысле.

Структурирование выполняется с помощью папок панели закладок. Структура папок выстроена по типу химической реакции: 4 стандартных типа школьных хими-

ческих реакций (соединение, разложение, обмен, замещение), сложные и многостадийные реакции, прочие процессы. Каждый объект распределяется по папкам в зависимости от того, какой тип реакции представляет собой.

Индексация выполняется с помощью меток (тегов). Они проставляются в специальном окошке в описании закладки браузера. Метки разделены по группам, каждой из групп выделен собственный код. Стандартный вид тега: <Код, 2 знака> – <метка>. Были выделены следующие группы меток:

1. «Тип реакции». Совпадают со структурой папок. Код – ТР.

2. «Исходные вещества». Представляются в виде химической формулы. На каждое вещество – отдельный тег. Пример – H2O, KOH. Код – ИВ.

3. «Продукты реакции». Представляются в виде химической формулы. На каждое вещество – отдельный тег. Код – ПР.

4. «Типы неорганических соединений и веществ», задействованных в реакции как в качестве исходных веществ, так и в качестве продуктов. На каждый тип соединения – отдельный тег (металлы, неметаллы, 3 типа оксидов, основания, кислоты, соли, комплексные соединения и т.п.). Код – ТС.

5. «Тип процесса». Только для типа реакции 6 (прочие процессы). Сюда относятся определение кислотности среды, нагрев, фильтрование и т.п. Код – ТП.

Формулирование требований к плагину для браузера

В данной информационной технологии от пользователя требуется много ручных операций по манипулированию информацией («перетаскивание», сохранение, копирование и т.п.). Подобные операции могут быть автоматизированы с помощью плагина для используемого браузера.

Перечислим такие операции:

- установка набора целевых папок для копирования гиперссылок;
- выбор целевой папки для копирования текущей гиперссылки;
- автоматическое или по выбору исключение повторов;
- редактирование списка меток;
- предъявление контекстно зависимого списка меток;
- выбор из списка меток и сохранение выбранных меток;
- поисковые операции по меткам, в том числе по сложным запросам с несколькими условиями.
- формирование списка гиперссылок с данной страницы
- сохранение списка гиперссылок в целевую папку
- определение типа источника ссылки – внутренняя или внешняя (на определенном сайте)
- предоставление определенной группы операций в зависимости от типа рассматриваемого объекта.

Плагин, соответствующий указанным требованиям, может быть найден среди большого набора имеющихся, доступных на рынке, или заново разработан.

Выводы

Исходная информационная технология поиска, отбора и классификации больших массивов информации доработана, расширена путем дополнения её алгоритмами структурирования и индексации, апробирована, в результате чего сформирована первичная структурированная и проиндексированная база информационных объектов, готовая к дальнейшему использованию. Данная информационная технология позволяет упорядочить, упростить и существенно ускорить:

1. Первичную отбраковку ненужной информации.

2. Сортировку и классификацию полезной информации.

3. Формирование структурированной и проиндексированной базы информационных объектов.

а также:

4. Значительно уменьшить вероятность ошибочных действий пользователя.

Данная информационная технология применима к поиску, классификации и структурированию больших массивов малосвязанной и слабоструктурированной информации в сети Интернет для любой предметной области.

Список литературы

1. Кулюлина Н.Л. Информационная технология поиска, отбора и классификации больших массивов информации // Старт в науке. – 2017. – № 4-1. – С. 37-42.