ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ПОИСКА, ОТБОРА И КЛАССИФИКАЦИИ БОЛЬШИХ МАССИВОВ ИНФОРМАЦИИ

Кулюлина Н.Л.

г. Москва, 10 класс

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

Во время выполнения исследовательской работы по химии встала проблема поиска информации в сети Интернет. Информация эта относится к предметной области – видео-опыты по школьной неорганической химии. Было ясно, что предстоит анализ большого массива малосвязанной и слабоструктурированной информации. В связи с этим понадобилась алгоритмизация механизмов поиска, отбора и классификации информации.

Цель работы:

Разработка системы взаимосвязанных алгоритмов поиска, отбора и классификации больших массивов малосвязанной и слабоструктурированной информации в сети Интернет. По сути система взаимосвязанных алгоритмов является информационной технологией.

Суть данной информационной технологии в том, что большая часть действий пользователя максимально формализованы и выполняются по стандартным схемам с минимальными затратами интеллектуальных ресурсов. Основные решения принимаются пользователем на завершающем этапе работы.

Требуется решить следующие задачи:

1. Разработка плана поиска информации.
2. Выбор программных средств (браузер, поисковая машина).
3. Разработка алгоритма отбора и классификации информации среди результатов поиска.

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

Этапы работы. Подготовительный этап

Разработка и последующее использование информационной технологии состоит из следующих этапов:

1) Подготовительный этап.

2) Выделение и классификация информационных объектов (2 этап).

Подготовительный этап состоит из следующих частей:

- Разработка плана поиска информации.
- Выбор программных средств.
- Выбор поискового запроса.
- Фиксация результатов поиска.

Разработка плана поиска информации. Изначальный план поиска: последовательно просмотреть несколько сотен ссылок – результатов работы поисковой машины, интересующие по тематике ссылки сохранять в закладках браузера. Поисковики периодически обновляют результаты поиска по одному и тому же поисковому запросу. Чтобы гарантированно не просматривать одни и те же ссылки несколько раз, требуется зафиксировать текущее состояние результатов поиска. Для этого нужно сохранить результаты поиска в виде html-файла на жестком диске компьютера, и просматривать результаты поиска уже не из интернета, а из этого файла.

В дальнейшем потребуется классификация найденной информации. Отсюда следуют требования к браузеру и поисковой машине. Браузер должен быть удобен для работы с папками и закладками. Поисковик должен показывать как можно больше (50+) результатов на одной странице для удобства сохранения в виде html-файлов.

Браузеры выбирались из наиболее популярных в России и в мире. В расчет также взяты браузеры, рекомендуемые на интернет-сайтах с соответствующей тематикой. Самыми популярными браузерами в России на октябрь 2016 года являются Google Chrome, Яндекс.Браузер и Mozilla Firefox. Статистика представлена на рис. 1.

Самыми популярными браузерами в мире на декабрь 2016 года являются Google Chrome, Mozilla Firefox и Internet Explorer. Статистика представлена на рис. 2.

Сайты с подобной тематикой [3, 4] также рекомендуют сравнительно новые браузеры российского производства Амиго и Orbitum. В результате кандидатами служили следующие браузер: Google Chrome, Яндекс.Браузер, Mozilla Firefox, Microsoft Edge (вместо устаревшего Internet Explorer), Амиго, Orbitum.

В результате исследования данных браузеров выяснилось, что Google Chrome, Яндекс.Браузер, Амиго и Orbitum созданы на одной и той же платформе Chromium. Из-за этого средства работы с закладками у них аналогичные. Результаты анализа представлены в таблице. По итогам анализа в качестве браузера был выбран Mozilla Firefox.

<< Сен 16 | **октябрь 2016 г.** | Ноя 16 >>

отчет: количество посетителей с разными браузерами — по дням | по неделям | по месяцам

значения: среднесуточные	октябрь 2016 г.		сентябрь 2016 г.		в среднем за 3 месяца	
Google Chrome	16,898,848	42.8%	16,468,305	42.9%	16,230,321	42.8%
Яндекс.Браузер	7,596,265	19.2%	7,205,371	18.8%	7,155,581	18.9%
Firefox	5,047,062	12.8%	4,961,581	12.9%	4,880,404	12.9%
Opera (Blink)	3,436,341	8.7%	3,360,208	8.8%	3,323,414	8.8%
Explorer 11	1,949,928	4.9%	1,921,871	5.0%	1,897,533	5.0%
Хром (Mail.ru)	1,038,831	2.6%	985,787	2.6%	980,590	2.6%
Microsoft Edge	870,553	2.2%	804,088	2.1%	798,710	2.1%
Opera 12	554,991	1.4%	569,144	1.5%	562,168	1.5%

Рис. 1. Статистика популярности браузеров в России [1]

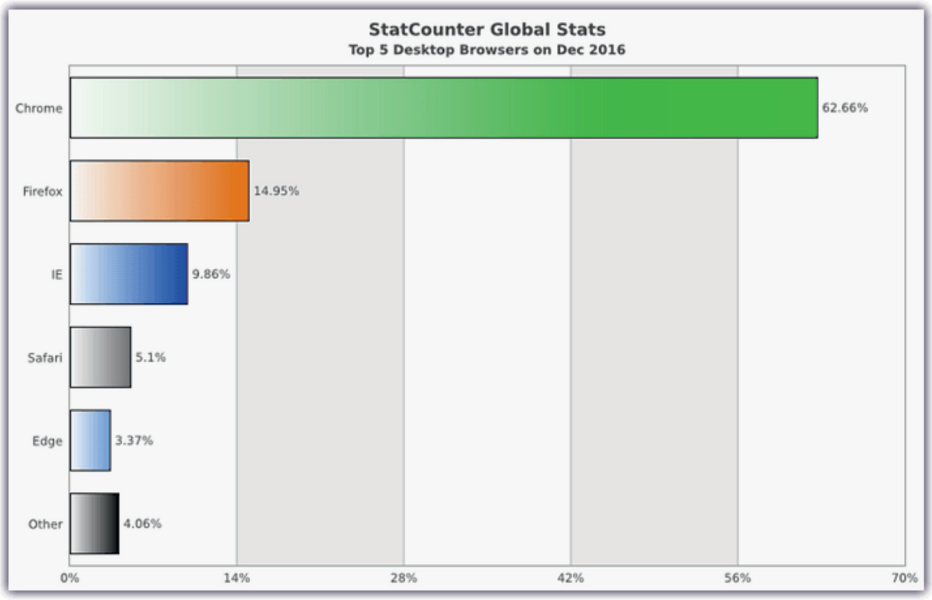

Рис. 2. Статистика популярности браузеров в мире [2]

Анализ функциональности браузеров

Браузер	Есть + / Нет -				Возможность поиска по меткам	Возможность представления закладок на боковой панели
	Имя закладки	Описание закладки	Метки	Папки		
Платформа Chromium	+	-	-	+	-	Нет
Mozilla Firefox	+	+	+	+	+	Есть, панель удобная
Microsoft Edge	+	-	-	-	-	Есть, панель неудобная

Поисковые машины выбирались из наиболее используемых в России. Такими поисковиками являются Google.com и Yandex.ru. Статистика представлена на рис. 3.

С точки зрения удобства сохранения информации нам подходят оба поисковика: Яндекс отображает максимум 50 результатов на странице, Google – 100. Было решено проверить качество поиска Google и Яндекса. Оценив результаты по нескольким поисковым запросам, нами был сделан вывод, что качество поиска Google выше. На этом основании в качестве поисковика был выбран Google.

Выбрав поисковик и браузер, мы начинаем осуществлять поиск Google по выбранному поисковому запросу «Химические опыты видео» с помощью браузера Mozilla Firefox. Результаты сохраняются в виде нескольких html-страниц.

Выделение и классификация информационных объектов (2 этап)

Второй этап состоит из нескольких частей – последовательных просмотров информации, результат предыдущего просмотра является исходным для последующего просмотра. Структура второго этапа представлена на рис. 4.

Первые два просмотра – быстрые и максимально формализованные, выполняются с минимальными затратами интеллектуальных ресурсов пользователя. Они не касаются предметного содержания информации.

На разных этапах работы информация представляется в виде гиперссылок, источников (веб-страниц и сайтов) и 3-х типов информационных объектов. Простой информационный объект – это объект, который нельзя поделить на более мелкие доступными техническими средствами, причем без нарушения интересов правообладателей. Составной информационный объект – объект, не являющиеся простыми и обладающие смысловой предметной цельностью. Отложенный информационный объект – это объект, который требует дополнительных сложных технических процедур и/или урегулирования с правообладателями. Формирование информационных объектов является конечной целью данной информационной технологии.

<< Янв 17 **февраль 2017 г.**

отчет: переходы из поисковых систем — по дням | по неделям | по месяцам

значения: **суммарные** / среднесуточные	февраль 2017 г.		январь 2017 г.		в среднем за 3 месяца	
☑ Яндекс	1,452,190,510	48.8%	3,034,031,104	48.3%	2,479,882,170	48.4%
☑ Google	1,351,904,706	45.4%	2,830,540,800	45.1%	2,307,072,150	45.1%
☑ Search.Mail.ru	148,849,174	5.0%	363,620,922	5.8%	289,328,926	5.7%
☑ Rambler	12,455,174	0.4%	23,366,543	0.4%	21,121,213	0.4%
☑ Bing	7,716,886	0.3%	15,836,104	0.3%	13,133,343	0.3%
☐ Yahoo	2,182,662	0.1%	4,743,718	0.1%	3,968,803	0.1%

Рис. 3. Статистика использования поисковиков в России [5]

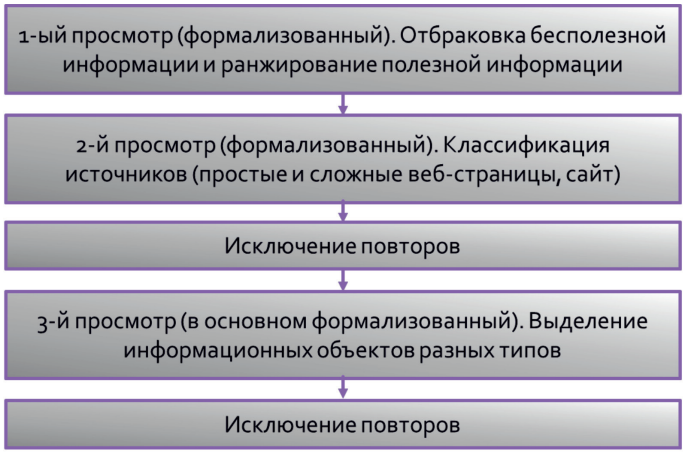

Рис. 4. Структура второго этапа

Отбраковка бесполезной и ранжирование полезной информации (1 просмотр)

Выполняется первый просмотр. При первом просмотре результат работы поисковика – «куча» гиперссылок – классифицируется на полезные и бесполезные гиперссылки (к последним относится, например, реклама). Полезные гиперссылки сохраняются в закладках браузера. Сохранение идет в три папки в зависимости от актуальности и полезности информации:

1. «Основное» – полезная информация, соответствующая школьной программе
2. «Дополнительное» – полезная информация, близкая к школьной программе
3. «Эффектное» – полезная, но не имеющая отношения к школьной программе информация, в т.ч. эффектные опыты – демонстрации, шоу и т.п.

Алгоритм выполнения первого просмотра представлен на рис. 5.

Классификация источников информации (2 просмотр)

Выполняется второй просмотр; этот просмотр имеет служебный характер. В нем рассматривается содержимое папки «Основное» (результат первого просмотра). Осуществляется классификация источников по следующим типам:

1. Сайт. Является собранием веб-страниц (имеются гиперссылки, которые ведут на другие сложные веб-объекты, которые могут быть полезны)

2-3. Простая веб-страница. Не является собранием веб-страниц. Содержит материалы из одного информационного интернет-источника. Простая веб-страница 1-го типа содержит в себе один интересующий нас объект (видеоопыт), 2-го типа – более одного.

4-5. Сложная веб-страница 1-го и 2-го типов. Не является собранием веб-страниц. Содержит материалы из разных информационных интернет-источников. Сложная веб-страница 1-го типа содержит в себе один интересующий нас объект (видеоопыт), 2-го типа – более одного.

6. Сложная веб-страница 3-го типа. Не является собранием веб-страниц. Содержит в себе ссылки на другие простые веб-объекты. Собственное информационное содержание соответствует пунктам 2, 3, 4 или 5.

Алгоритм выполнения второго просмотра представлен на рис. 6. В дальнейшем к каждому из типов источников применятся свой алгоритм дальнейшего анализа при третьем просмотре.

После второго просмотра некоторые источники могут дублироваться. Для исключения повторов выполняется автоматическая процедура.

Выделение информационных объектов разных типов (3 просмотр)

Выполняется третий просмотр. Из каждого типа источников по специальным полностью формальным алгоритмам выделялись простые, составные и отложенные информационные объекты для дальнейшего формирования информационной базы. Ненужная и бесполезная информация отбрасывалась. Алгоритм выполнения третьего просмотра представлен на рис. 7.

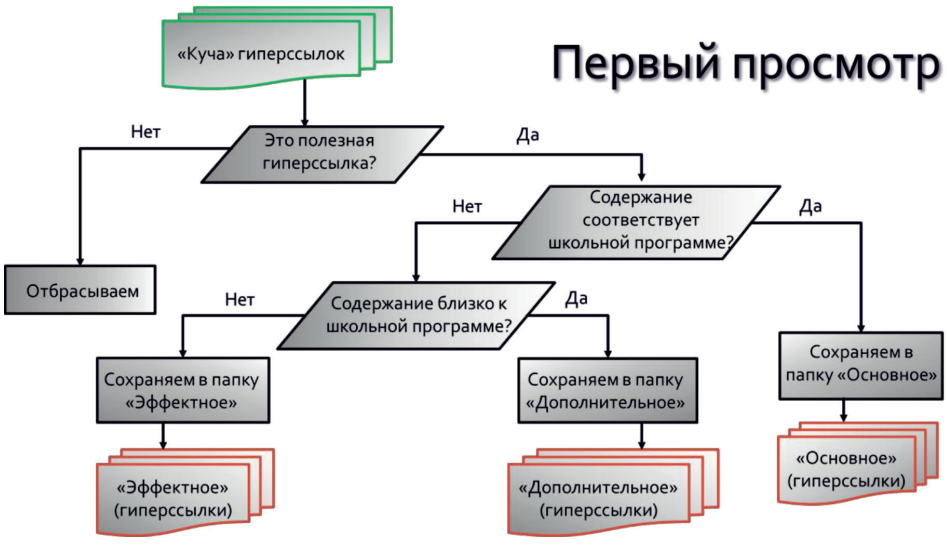

Рис. 5. Алгоритм выполнения первого просмотра

Рис. 6. Алгоритм выполнения второго просмотра

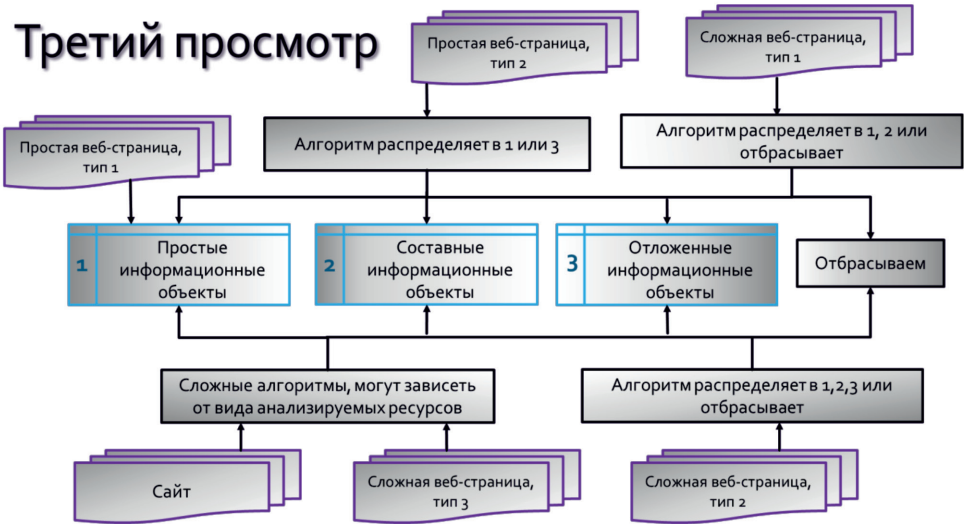

Рис. 7. Алгоритм выполнения третьего просмотра

Алгоритмы выделения информационных объектов из сайтов и сложных веб-страниц третьего типа являются более сложными, могут зависеть от конкретного вида анализируемых ресурсов, включают в себя процедуры обхода дерева, использование стека и другие.

Таким образом нами создается первичная база классифицированных информационных объектов, пригодная для дальнейших структурирования, индексации и использования.

Практические результаты работы

С помощью данной информационной технологии проанализирован большой массив текстовой и видео-информации из сети Интернет по тематике лабораторных и практических работ по неорганической химии за курсы 8-9 классов (с сохранением интересной сопутствующей информации).

Полностью завершены подготовительный этап (объем «кучи» гиперссылок – около 400), 1-й просмотр (объем папки «Основное» – 123 гиперссылки) и 2-й просмотр. Продолжается 3-й просмотр, на данный момент проанализировано 53 простых и сложных веб-страницы и сайта, выделено и классифицировано 64 простых и составных информационных объектов.

Применение данной информационной технологии существенно упорядочило

и упростило обработку информации, ускорило работу пользователя и значительно уменьшило вероятность ошибочных действий. Таким образом, информационная технология показала высокую эффективность в поиске, отборе и классификации больших массивов информации в сети Интернет.

Выводы

Разработана и апробирована информационная технология – система взаимосвязанных алгоритмов, позволяющая упорядочить, упростить и существенно ускорить:

1. Первичную отбраковку ненужной информации.
2. Сортировку и классификацию полезной информации.
3. Формирование первичной базы классифицированных информационных объектов.
4. Дальнейшую работу с полученными информационными объектами.
5. Значительно уменьшить вероятность ошибочных действий пользователя.

Данная информационная технология применима к поиску, отбору и классификации больших массивов малосвязанной и слабоструктурированной информации в сети Интернет для любой предметной области. Полученная база информационных объектов может быть в дальнейшем структурирована и проиндексирована. Для дальнейшего облегчения рутинной работы целесообразно использовать плагин к браузеру.

Список литературы

1. URL: https://my-chrome.ru/statistika-brauzerov/

2. URL: http://www.itrew.ru/brauzery/statistika-ispolzovaniya-brauzerov-2016.html

3. URL: http://softcatalog.info/ru/obzor/vybiraem-luchshiy-brauzer

4. URL: http://pcpro100.info/luchshie-brauzeryi-2016/

5. URL: http://www.liveinternet.ru/stat/ru/searches.html?period=month;total=yes