|  |
| --- |
| Секция: Программирование |
| ГБОУ "Лицей № 1533 (информационных технологий)", г. Москва |
| 119296, Москва, г. Москва, Ломоносовский проспект 16 |
| тел.: (495) 133-20-00; E-mail: [info@lit.msu.ru](mailto:info@lit.msu.ru) |
| Кластеризация результатов поиска |
| Гайдамашко Даниил Олегович |
| тел.: +7 985-258-27-38; E-mail: [dan-gai@yandex.ru](mailto:dan-gai@yandex.ru) |
| Класс: 11 |
| 121351, Москва, ул. Кунцевская 4 корп.1, кв. 20 |
| Карпенко Максим Дмитриевич |
| тел.: +7 925-618-10-63; E-mail: [maxim020998@gmail.com](mailto:maxim020998@gmail.com) |
| Класс: 11 |
| 119571, Москва, Академика Анохина 9, кв. 25 |
| Научный руководитель: Завриев Николай Константинович, преподаватель ГБОУ ЛИТ 1533 |

Проблема большинства интернет-поисковиков заключается в том, что поисковая выдача представляет собой набор ссылок на статьи, описывающие разные объекты, что существенно усложняет поиск необходимой информации.

Проблема решается с помощью кластеризации результатов поиска – разделения результатов поиска на смысловые группы с возможностью для пользователя искать в той группе, которая ему нужна. При этом алгоритм кластеризации, который направлен на выявление семантической сущности текста, имеет потенциал использования и в других интернет-сервисах.

Для решения проблемы кластеризации текстов был выбран алгоритм, получивший название алгоритма латентно-семантического анализа [3], позволяющий выявить зависимости между документами и терминами, в них встречающихся. Результатом выполнения алгоритма является множество вершин, которые соответствуют документам и тегам, относящимся к этим документам. Вершины имеют координаты в некоем многомерном пространстве, таким образом о семантической близости документов и тегов можно судить, оценивая расстояния между ними.

В нашем проекте в процессе выполнения алгоритма вершины разделяются по принципу тег/документ, затем происходит кластеризация множества вершин-тегов при помощи алгоритма dbscan [2], после чего вершины-документы присоединяются к ближайшим кластерам. Распределение документов по смысловым группам происходит на основе полученных кластеров.

Таким образом, разработанный проект состоит из двух частей. Первая, основная часть (далее *кластерный поисковик*) – непосредственно веб-инструмент, выполняющий кластеризацию результатов интернет-поиска. Вторая – это графический исследовательский инструмент, реализующий и демонстрирующий кластеризацию введенных текстов, а также позволяющий настраивать параметры алгоритма и улучшать тем самым качество кластеризации. Именно на нём тестировались и отрабатывались алгоритмы, использованные в кластерном поисковике.

К функциям кластерного поисковика относятся:

* Получение поискового запроса пользователя и отправление его в существующий веб-поисковик;
* Получение страниц результатов поиска, извлечение из нее необходимой информации и сохранение в базе данных;
* Обработка полученных данных, выполнение латентно-семантического анализа и алгоритма выделения кластеров;
* Формирование кластеров и их вывод на веб-страницу в виде папок, описанных набором тегов и содержащих набор поисковых результатов, относящихся к данному кластеру.

К функциям графического инструмента относятся:

* Получение на вход некоторое количество текстов, введенных пользователем;
* Выполнение латентно-семантического анализа и формирование кластеров;
* Графическое отображение результатов группировки текстов, где каждый кластер представляет собой графически выделенное множество тегов и текстов.

Программная логика веб-кластеризатора реализована на языке C# в среде Microsoft Visual Studio с использованием фреймворка ASP.NET MVC 4, а для написания интерфейса сайта использовались языки HTML, CSS и JavaScript.

Для выполнения автоматических запросов к поисковику «Яндекс» и дальнейшей публикации обработанной поисковой выдачи на веб-сайте используется сервис «Яндекс.XML».

Графический инструмент также был реализован на языке C# в среде разработки Microsoft Visual Studio с помощью .NET Framework 4.5.

В качестве дальнейших разработок планируется дальнейшая оптимизация и улучшение качества работы алгоритма кластеризации, испытания альтернативных методов кластеризации. В более долгосрочной перспективе планируется учитывать синонимичность слов и усовершенствовать методы обработки текстов (см. алгоритм латентно-семантического анализа). Что касается развития кластерного поисковика, то помимо улучшения качества его работы планируется добавить возможность работы с поисковой выдачей других существующих поисковых сервисов, наиболее приоритетным из которых является Google.

Использованные материалы:

* Фридл Дж. Регулярные выражения, 3-е издание. СПб.: Символ-Плюс, 2008 г.
* Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011 г.
* Edunov «Латентно-семантический анализ» [Электронный ресурс]:  
  <http://habrahabr.ru/post/110078/>
* andreycha «Обзор алгоритмов кластеризации данных» [Электронный ресурс]:  
  <http://habrahabr.ru/post/101338/>
* «SVD-разложение прямоугольной матрицы» [Электронный ресурс]:  
  <http://alglib.sources.ru/matrixops/general/svd.php>
* Онлайн-книг «Изучаем ASP.NET MVC 4»  
  <http://metanit.com/sharp/mvc/>
* «htmlbook.ru» [Электронный ресурс]:  
  <http://htmlbook.ru/>
* «Учебник HTML» [Электронный ресурс]:  
  <http://webremeslo.ru/html/glava0.html>
* «Учебник CSS» [Электронный ресурс]:  
  <http://webremeslo.ru/css/glava0.html>