ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН

ҚазҰТЗУ ХАБАРШЫСЫ

ВЕСТНИК КазНИТУ

VESTNIK KazNRTU

№3 (127)

АЛМАТЫ 2018 ИЮНЬ

Учредитель:

Казахский национальный исследовательский технический университет
имени К.И. Сатпаева

Регистрация:

Министерство культуры, информации и общественного согласия
Республики Казахстан № 951 – Ж “25” 11. 1999 г.

Основан в августе 1994 г. Выходит 6 раз в год

Адрес редакции:

г. Алматы, ул. Сатпаева, 22,
каб. 616, тел. 292-63-46
Nina. Fedorovna. 52 @ mail.ru

[13] Петрова Е.В. Взгляд в будущее – обзор новинок, представленных на выставке ConExpo-Con / Agg. // Строительная техника и технологии. – 2014. –№3 (103). – С. 30-48.

[14] Клушанцев Б.В., Косарев А.И., Муйземнек Ю.А. Дробилки. Конструкция, расчет, особенности эксплуатации. – М.: Машиностроение, 1990. – С. 320.

[15] Предварительный патент РК №19801, В02С 4/30, бюл. №8, 15.08.2008г.

Гурьянов Г.А., Ким В.А., Васильева О.Ю.

Қатты бөлшектерді бұзу процесінің механика-математикалық моделін әзірлеу және кедір-бұдырлы біліктердің геометриялық параметрлерін анықтау

Аңдатпа: Мақалада кедір-бұдырлы біліктердің конструкциясын біліктердің геометриялық профилі пішінін өзгерту арқылы жетілдіру, сондай-ақ кесек материалды жаңа профилді біліктің қармау процесінің компьютерлік моделі нәтежиесі көрсетілген.

Түйінді сөздер: кесек материал, қармау бұрышы, ұнтақтау процесі, білікті уатқыш, компьютерлік модель.

Guryanov G.A., Kim V.A., Vasileva O.Yu.

Development of the mechanical-mathematical model of the process of the destruction of solid particles and definition of geometric parameters of convex-concave rollers

Summary: The article considers a variant of improving the design of a roller crusher by changing the geometric shape of the profile of rolls, as well as the results of computer simulation of the process of gripping a piece of material in a roller crusher with a new roll profile.

Keywords: lump material, gripping angle, grinding process, roller crusher, computer model.

M. Mansurova, M. Kaipoldayev
(Al-Farabi Kazakh National University, Almaty, Kazakhstan
E-mail: mansurova.madina@gmail.com, m.kaipoldaev@gmail.com)

DEVELOPMENT OF A MODULE FOR DETECTING DUPLICATE TEXTS

Abstract: This paper is devoted to solve the problem of eliminating duplicating articles in the news web-portal. A model of the system for detecting and eliminating such articles is described. Also, two algorithms for comparing texts are given and an analysis of their effectiveness is made. The main task was to develop a module for eliminating duplicate articles for the news web-portal.

Keywords: w-shingling algorithm, Hirschberg's algorithm, duplication of texts, text processing.

М.Е. Мансурова, М.Е.Қайполдаев
(Казахский Национальный Университет им. Аль-Фараби, Алматы, Республика Казахстан.
E-mail: mansurova.madina@gmail.com, m.kaipoldaev@gmail.com)

РАЗРАБОТКА МОДУЛЯ ОБНАРУЖЕНИЯ ДУБЛИРУЮЩИХСЯ ТЕКСТОВ

Аннотация: Данная работа посвящена решению задачи устранения дублирующихся статей в новостном портале. Описана модель системы обнаружения и устранения таких статей. Также приведены два алгоритма сравнения текстов и сделан анализ их эффективности. Основной задачей являлась разработка модуля устранения дублирующихся статей для новостного портала.

Ключевые слова: алгоритм Шинглов, алгоритм Хиршберга, дублирование текстов, обработка текста.

1. Введение

Проблема повторяющегося содержимого текста очень становится очень актуальной в современных реалиях. Если содержимое различных текстов описывает одно и то же событие или объект, то это может привести к различным негативным последствиям. К примеру в поисковых системах необходимо выводить данные содержащие различный контент, так-как пользователю не желательно просматривать множество схожих сайтов. Также любой документ претендующий на уникальность должен пройти процедуру выявления плагиата. В таких ситуациях необходима разработка систем и алгоритмов, которые выявляют сходства содержимого текстов и устраняет дублирование документов.

Значительной проблемой анализа данных на дублирование является то, что с развитием интернет технологий количество веб-документов очень сильно возросло, что в свою очередь привело к увеличению числа схожих документов. Данный фактор привел к необходимости разработки более эффективных методов выявления дублирования.

2 Постановка задачи

На данный момент в лаборатории НИИ ММ КазНУ им. аль-Фараби разработан веб-портал который занимается сбором и анализом информаций о чрезвычайных ситуациях на территории Казахстана. Данный сайт собирает и публикует новостные данные с официальных источников МВД РК. Однако, вследствии того что многие новостные сайты публикуют схожие новости, возникает проблема дублирующихся новостей.

Для обеспечения оригинальности публикующихся новостей необходимо разработать и добавить модуль фильтрации новых новостей и оставлять лишь один из многих дублирующихся новостей (рис. 1). Предварительно все собранные за небольшой промежуток времени новости будут храниться в кэш-памяти системы. Затем модуль выявления дублирующихся записей будет обрабатывать данные кэша и удалять записи которые удовлетворяют условиям дублирования статей. После чего на сайте публикуются лишь оригинальные новости. Эти же новости отправляются на хранение в базу данных новостей в качестве архива. На портале будет публиковаться лишь самая первая по дате публикации новость из всех одинаковых записей.

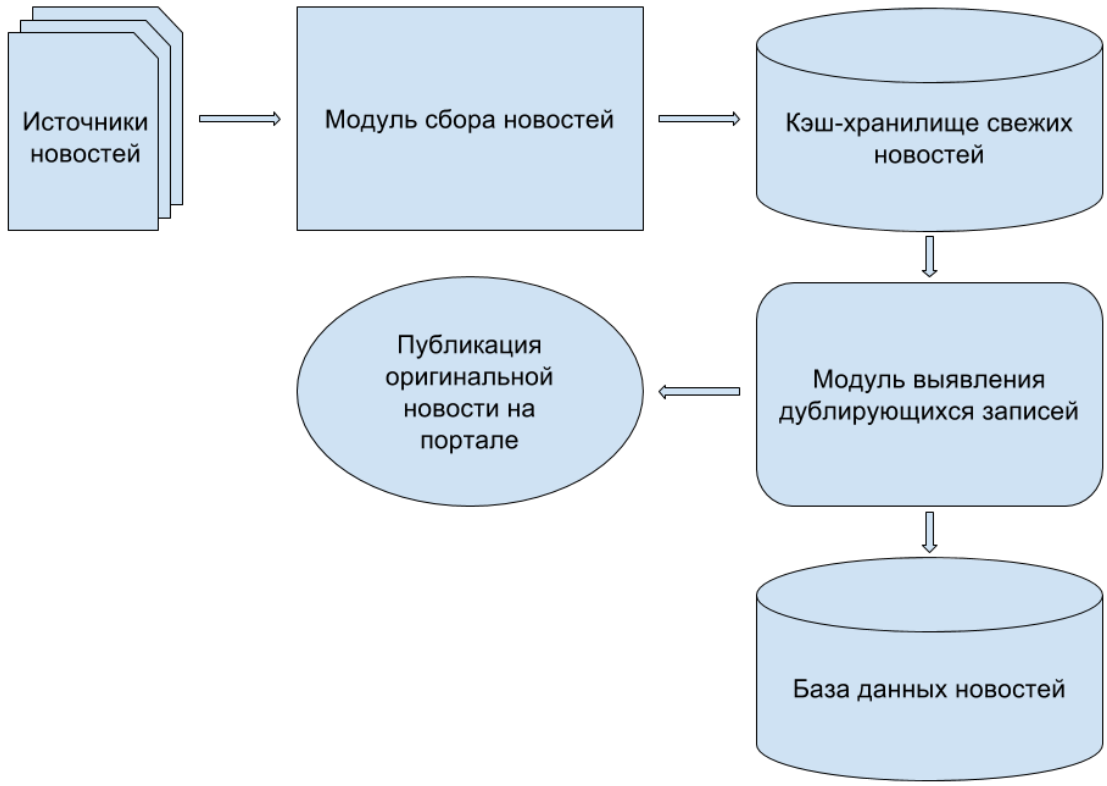

Рис. 1. Модель системы сбора и публикации новостей

3 Существующие решения

Для определения схожести текста вычисляются различные метрики. Одним из часто используемых мер схожести является расстояние Левиншейна[1]. Данное расстояние определяется как наименьшее количество операции удаления и вставки символа с помощью которых из одного текста можно получить второе. Имеются также модификации данного алгоритма которые учитывают вес операции. К примеру, операции удаления и вставки символа имеют вес 1, а операция замены символа имеют вес 2. Данная метрика называется “расстояние редактирования”.

Выше приведенные метрики применимы для сравнения строк, однако для сравнения больших текстов они не подходят, так-как эти метрики вычисляются с помощью сравнения символов. Для сравнивания текстов объектом сравнения необходимо брать целые слова. В связи с чем широко распространен алгоритм шинглов разработанный A. Broder et al[2]. В данном методе сравниваемые тексты делятся на множество частей (шинглов), которые попарно сравниваются друг с другом. Дальнейшим развитием данного метода является такой алгоритм, при котором вычисляются 84 контрольных сумм каждого документа. Далее эти 84 шингла образуют 6 супершинглов. Затем каждый документ представляется всеми парными сочетаниями из 6 супершинглов, эти сочетания называются мегашинглами. Два текста будут считаться схожими если у них совпадает не менее одного мегашингла.

Другой подход основанный на лексической характеристике текста создан A. Chowdhury et al. Этот алгоритм основывается на вычислении дактилограммы I-Match. Для этого необходимо составить словарь со средними значениями IDF (*inverse document frequency*). После этого находится пересечение множеств слов из составленного словаря и всех слов текста. Затем, если значение этого пересечения больше заданного порога, то вычисляется дактилограмма I-Match. А два документа можно считать схожими по содержанию если данные дактилограммы идентичны.

Существуют подходы основывающиеся на сходстве фонетического звучания слова. На данный момент большинство фонетических алгоритмов разработаны для английского языка, следовательно не могут корректно применяться для языков не схожих с данным. Наиболее популярными фонетическими алгоритмами являются Soundex, New York State Identification and Intelligence System, Cologne phonetics и т.д.

4 Выбор алгоритма и его реализация

Для решения задачи нахождения схожести текстов нами были выбраны два алгоритма: алгоритм w-shingling и алгоритм Хиршберга[3]. Выбор данных методов обусловлен их эффективностью и относительной простотой реализации.

4.1 Алгоритм w-shingling

W-shingling - это алгоритм оценки сходства двух документов. Суть данного алгоритма заключается в нахождении количества шинглов принадлежащих обоим документам. Затем находится отношение общих шинглов двух текстов к общему количеству шинглов в двух текстах.

Шингл - это последовательность, фиксированной длины, подряд идущих слов в тексте. Например, рассмотрим следующее выражение: “Мороз и солнце; день чудесный!”. Разделим данное предложение на шинглы с длинной 3. Для наглядности не будем учитывать знаки препинания. Получаем: “Мороз и солнце”, “и солнце день”, “солнце день чудесный”. Несложно догадаться, что количество шинглов равно

количество слов в тексте - длина шингла + 1.

Перед применением данного алгоритма необходимо провести предварительную обработку текста, т.е. удалить все знаки препинания, удалить стоп-слова (слова не имеющие смысловой нагрузки), привести слова к стандартной форме (именительный падеж, единственное число и т. п.).

Предобработка текста для применения алгоритма w-shingling

●Удаление всех лишних символов. Для осуществления данной операции была написана функция, которая удаляет все символы кроме A-Z, a-z, А-Я, а-я.

●Удаление стоп-слов. Для этого действия применялась функция которая сравнивает слова текста со стоп-словами, которые хранились в файле где записаны все слова такого типа. При совпадении из текста удалялось данное слово.

●Приведение слов к нормальной форме. Для этого применялся лемматизатор Apache Lucene.

Для нахождения дактилограммы шинглов была применена хэш-функция CRC32. Дактилограмма необходима для упрощения процесса сравнивания шинглов, так как гораздо быстрее сравнивать числовые значения хэш-функций, нежели сравнивать шинглы в виде строк.

При написании программы пришлось совместить деление слов на шинглы и нахождение дактилограммы (т.е. контрольной суммы), так как имелась предстала следующая проблема. У шинглов, с одинаковым набором слов но расположенных в разном порядке, дактилограмма должна быть одинаковой. К примеру шинглы “солнце встает над Алматы” и “над Алматы встает солнце” должны иметь одинаковое значение дактилограммы.

Physico-mathematical sciences