Министерство Образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК

Трофимов
Алексей Дмитриевич

Нечеткий поиск данных в хеш-таблице

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1–40 80 02 « Системный анализ, управление и обработка информации »

_____

Научный руководитель

Гуринович Алевтина Борисовна
канд. физ.-мат. наук., доцент

_____

Минск, 2014

КРАТКОЕ ВВЕДЕНИЕ

В настоящее время хеширование, как способ обеспечения защиты и шифрования электронной информации, встречается практически повсеместно.

С хешированием мы сталкиваемся едва ли не на каждом шагу: при работе с браузером (список Web-ссылок), текстовым редактором и переводчиком (словарь), языками скриптов (Perl, Python, PHP и др.), компилятором (таблица символов). Заглядывая в адресную книгу, энциклопедию, алфавитный указатель, мы даже не задумываемся, что упорядочение по алфавиту является не чем иным, как хешированием.

Хеширование есть разбиение множества ключей (однозначно характеризующих элементы хранения и представленных, как правило, в виде текстовых строк или чисел) на непересекающиеся подмножества (наборы элементов), обладающие определенным свойством. Это свойство описывается функцией хеширования, или хеш-функцией, и называется хеш-адресом.

Алгоритмы нечеткого поиска (также известного как поиск по сходству или fuzzy string search) являются основой систем проверки орфографии и полноценных поисковых систем вроде Google или Yandex. Например, такие алгоритмы используются для функций наподобие «Возможно вы имели в виду …» в тех же поисковых системах.

Методы искажения:

1. Вставлен неправильный символ(опечатка)
2. Пропущен символ
3. Вставлен лишний символ
4. Транспозиция(перестановка двух соседних букв)

Основное ограничение в искажении накладывается на глубину искажения, т.е. количество допустимых ошибок для каждого из методов искажения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель данной магистерской диссертации – разработать алгоритм нечеткого поиска данных в хеш-таблице.

Для реализации поставленной цели были определены следующие задачи:

1. Изучение и анализ учебной, научной и периодической литературы по алгоритмам хеширования и хеш-таблицам.
2. Изучение и анализ учебной, научной и периодической литературы по алгоритмам нечеткого поиска.
3. Разработка алгоритма нечеткого поиска с учетом ограничений;
4. Программная реализация Веб-сервиса с нечетким поиском данных в хеш-таблице.

Дипломная работа состоит из введения, трех разделов, заключения, библиографического списка, приложений.

В первой главе рассматриваются типы хеш-таблиц, а также принципы наборных и фонетических алгоритмов нечеткого поиска, и производится постановка задачи для работы над диссертацией.

Во второй главе раскрываются теоретические основы алгоритмов нечеткого поиска. Производится детальное сравнение и анализ фонетических и нефонетических(наборных) алгоритмов нечеткого поиска, и видов хеширования. Методологическую основу составляют труды таких специалистов, как Левенштейн В.И., Дамерау К.К., Ниделман А., Вугш К.

В третьей главе проводится анализ предмета работы алгоритма, разрабатывается наборная часть алгоритма нечеткого поиска с применением алгоритма сигнатур. На основе алгоритма русского Metaphone создаются новые правила его работы. Разрабатывается Веб-приложение, с реализацией разработанного алгоритма нечеткого поиска данных по таблице, содержащей рассчитанные хеши входных данных и на основе полученных метрик производится корректировка исходного алгоритма.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из введения, трех разделов, заключения, библиографического списка, приложений.

Во введении магистерской диссертации обоснована актуальность темы по разработке алгоритма нечеткого поиска данных в хеш-таблице.

В первой главе рассматриваются типы хеш-таблиц, а также принципы наборных и фонетических алгоритмов нечеткого поиска, и производится постановка задачи для работы над диссертацией.

Во второй главе раскрываются теоретические основы алгоритмов нечеткого поиска. Производится детальное сравнение и анализ фонетических и нефонетических(наборных) алгоритмов нечеткого поиска, и видов хеширования. Методологическую основу составляют труды таких специалистов, как Левенштейн В.И., Дамерау К.К., Ниделман А., Вугш К.

В третьей главе проводится анализ предмета работы алгоритма, разрабатывается наборная часть алгоритма нечеткого поиска с применением алгоритма сигнатур. На основе алгоритма русского Metaphone создаются новые правила его работы. Разрабатывается Веб-приложение, с реализацией разработанного алгоритма нечеткого поиска данных по таблице, содержащей рассчитанные хеши входных данных и на основе полученных метрик производится корректировка исходного алгоритма.

В заключении приводятся результаты магистерской диссертации и строятся выводы с недостатками и достоинствами разработанного алгоритма.

ЗАКЛЮЧЕНИЕ

Результатом поэтапной реализации данного проекта стал разработанный алгоритм нечеткого поиска данных в хеш-таблице и программное средство, реализующее эту методику и позволяющее производить поиск входного текста с ошибками по хешированным данным. Учитываются ошибки как фонетического характера, так и наборного.

В процессе изучения предметной области были изучены принципы хеширования и хранения данных в хеш-таблицах и их практическое применение, алгоритмы нечеткого поиска, а также фонетические алгоритмы.

Был проведен сравнительный анализ различных алгоритмов поиска входных данных с ошибками, также был проведен анализ затрат ресурсов для различных алгоритмов.

Был разработан уникальный алгоритм нечеткого поиска данных по хешированным данным из хеш-таблиц, создан веб-сервис, реализующий данный алгоритм.

Достоинствами данного алгоритма является относительная простота его реализации, масштабируемость и универсальность применения без привязки к виду хеш-функции, используемой для шифрования данных в хеш-таблице. Разработанный веб-сервис позволяет получить результат нечеткого поиска для любого объема данных. Другим плюсом разработанного алгоритма является непривязанность его к наборным средствам: алгоритм будет отлично работать как с реальной клавиатурой, так и с виртуальной. А также алгоритм поддерживает набор слов для нечеткого поиска с помощью голоса: такие алгоритмы не будут иметь наборные ошибки, но большой долей вероятности могут содержать фонетические ошибки.

Также достоинством разработанного алгоритма можно считать то, что он не привязан к конкретному веб-сервису, для которого он был разработан, а также его независимость от языка программирования, на котором алгоритм должен быть реализован.

К недостаткам данного алгоритма можно отнести его эффективную работу для небольшой глубины его нечеткости, привязку к языку исходных

данных, и, соответственно, наличия слов-исключений для которых алгоритм будет неэффективен, а также не реализованную поддержку не-альфанумерических символов, логической «звездочки» в теле запроса и поиска по части слова.

За время работы над магистерской были реализованы все сформулированные задачи и требования к алгоритму и программному продукту, а поставленная цель была достигнута. Разработку и развитие полученного алгоритма и веб-сервиса можно продолжать и дальше. Разработанный веб-сервис можно дополнять новой функциональностью, увеличивая глубину и степень искажения одновременно улучшая алгоритм поиска.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А Трофимов, А.Д. Модели криптостойких хеш-таблиц / А.Д. Трофимов, А. Б. Гуринович // Информационные технологии и системы 2013 (ИТС 2013): материалы международной научной конференции, БГУИР, Минск, Беларусь, 23 октября 2013 г. = Information Technologies and Systems 2013 (ITS 2013): Proceeding of The International Conference, BSUIR, Minsk, 24th October 2013 / редкол.: Л. Ю. Шилин [и др.]. - Минск : БГУИР, 2013. – С. 242-243;

2-А Трофимов, А.Д Хранение данных в хеш-таблицах / А.Д. Трофимов // Информационные технологии и управление: Материалы 50-й научной конференции аспирантов, магистрантов и студентов. – Минск : БГУИР, 2014- С. 38.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 W3C Working Group Note 11 February 2004. Web Services Glossary

2 James Bach. Exploratory Hashing Explained, April 2003

3 Hash Tables Glossary. Editor: Eng Van Veenendaal. December 2007

4 Garry Pinkies, Shawn Brown. Cryptography for Enterprises. Marks Technologies, 2012

5 Cem Kaner. An Introduction to Distorted Search. June, 2008

6 Karl Semitone, Rona Atkins. Sound algorithms. March, 2003

7 Andre Yee. Working with Hash Tables: What You Need to Know. 2005

8 Richard Monson-Haefel. Hash Functions. Addison, 2009

9 Dave Kelly. Search Engine Basics, 2005

10 Syed Simnani. Cryptography – What the future holds. Wirpo Technologies, 2008

11 Elena Petrovskaya. Rigth Way to Hash, December 20010

12 Iienne Burstein. Search Algorithms. Springer, 2002

13 Mark Swissman, J.K. Klarks. Hashing Algorithms. Windsberg, 2002

14 Sameer Tyagi. Patterns and Strategies for Data search in Web Services. September 2004

15 Debu Panda, Reza Rahman, Derek Lane. Hash Tables in Action. Manning, 2007

16 Marry Kelman. Hashing functions. Grender, 2005

17 Harry Manal, Pierre Donald, Robin Panek. Hash Tables and Big Data for Enterprises. Windsberg, 2012

18 Raghu R. Kodali, Jonathan Wetherbee, Peter Zadrozhny. Working with search engines. Apress, 2006

19 W3C Note 15 March 2001. Web Services Data Hashing

20 UML. Классика CS. 2-у изд./Пер. с англ.; Под общей редакцией проф. С.Орлова - СПб.: Питер, 2006. - 736 с.: ил.

21 W3C Recommendation 26 June 2007. Web Services Description Language (WSDL) Version 2.0

22 W3C Recommendation 16 August 2006, edited in place 29 September 2006. MD5 Hash Functions 1.1 (Second Edition)

23 Future of Language Recognition Algorithms. AppLabs, July 2008

24 Akhil Saha, Sven Graupner. Data Base Security in the Enterprise. Springer, 2005

25 Addison Wesley. J2EE Web Services. Addison Wesley – 2013

26 Arthur Griffith. JavaScript, XML and JSON. John Wiley and Sons, Inc., 2002

27 Фергюсон Н., Шнайер Б. Практическая криптография. — Диалектика – 2005 – 418 с.

28 Matsumoto, Makoto. Cryptographic Mersenne Twister and Fubuki Stream/Block Cipher – 2005

29 Хеширование данных [Электронный ресурс] – Электронные данные. – Режим доступа: http://securitylab.ru

30 Методика создания алгоритмов [Электронный ресурс] – Электронные данные. – Режим доступа: http://algoritmy.info