**М****ИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ**

**Кафедра дискретной математики и алгоритмики**

Новик Никита Сергеевич

**АВТОМАТИЧЕСКИЙ ПОИСК РЕЛЕВАНТНЫХ ТЕКСТОВЫХ ДОКУМЕНТОВ В МНОГОЯЗЫЧНОЙ ИНТЕРНЕТ-СРЕДЕ**

Магистерская диссертация   
  
специальность 1-31 81 09 «Алгоритмы и системы обработки больших объемов информации»

Допущен к защите Научный руководитель:

*«*\_\_\_*»* \_\_\_\_\_\_\_\_\_\_\_\_\_ 2018 г. *Совпель Игорь Васильевич*

Зав. кафедрой дискретной математикидоктор технических наук, и алгоритмики ФПМИ, профессор, кафедра ИСУ ФПМИ  
доктор технических наук, профессор

Котов Владимир Михайлович

Минск 2017

# Реферат

Дипломная работа, 10 страниц, Х рисунков, Х источников.

**Ключевые слова:** ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, ВЫДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ, ИНТЕРНЕТ, ПОИСКОВЫЕ СЛУЖБЫ, ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР, РЕЛЕВАНТНЫЕ ДОКУМЕНТЫ.

**Объект исследования –** алгоритмы и способы составления образа текстовых документов для их последующего использования в поиске релевантных документов в многоязычной Интернет-среде.

**Цель работы –** разработать алгоритм составления образа документа, разработать систему для автоматического поиска релевантных текстовых документов с помощью некоторой поисковой службы, проанализировать полученные результаты.

**В результате исследования** разработаны и реализованы алгоритмы составления образа текстового документа вместе с автоматической системы поиска релевантных текстовых документов по заданному документу.

**Методы исследования –** анализ литературы, статистический и семантический анализ.

**Область применения результатов –** системы информационного поиска, классификации и анализа текстовых документов.

# ОГЛАВЛЕНИЕ

[Реферат 1](#_Toc497759822)

[ОГЛАВЛЕНИЕ 4](#_Toc497759823)

[ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ, СИМВОЛОВ И ТЕРМИНОВ 5](#_Toc497759824)

[ВВЕДЕНИЕ 6](#_Toc497759825)

[Глава 1. Спутниковая интерферометрия как один из способов дистанционного зондирования 8](#_Toc497759826)

[1.1 Понятие дистанционного зондирования Земли 8](#_Toc497759827)

[1.2 Физика спутниковой интерферометрии 11](#_Toc497759828)

[1.3 Спутниковая и спутниковая радарная интерферометрии 13](#_Toc497759829)

[1.4 Алгоритм построения спутниковой интерферометрии 14](#_Toc497759830)

[1.5 Постановка задачи 16](#_Toc497759831)

[1.6 Выводы 17](#_Toc497759832)

[список использованных источников 18](#_Toc497759833)

# 

# ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ, СИМВОЛОВ И ТЕРМИНОВ

ЕЯ – естественный язык;

ПО – программное обеспечение;

NLP – natural language processing.

# ВВЕДЕНИЕ

В настоящее время существует огромное количество информации, часть которой находится в Интернете или в базах данных. Большая часть данной информации находится в неструктурированном виде, что делает задачу поиска конкретной информации сложной задачей. Существуют готовые решения для поиска информации, которыми люди пользуются каждый день. Примерами таких решений выступают поисковые службы (Google, Yandex, Bing и т.д.). Для поиска определенной информации человеку требуется некоторым образом составить поисковое выражение (ключевые слова для поисковой системы в Интернете, SQL-выражение для поиска в SQL базе). Поиск релевантных документов не является исключением. Однако для поиска человеку необходимо знать суть и содержание этого документа, чтобы он мог составить поисковое выражение. Автоматизирование данного процесса является шагом к увеличению скорости поиска необходимой информации.

Чаще всего для поиска релевантных документов из исходного документа выделяют ключевые слова или ключевые фразы, которые будут в дальнейшем использованы для составления поискового образа. Для работы с текстом используются лингвистические процессоры.

На данный момент уже существуют достаточно эффективные решения задачи поиска релевантных документов в одноязычной информационной среде. Однако наличие в сети Интернет и полнотекстовых базах данных огромного числа текстовых документов, представленных на различных языках, может существенно повысить показания полноты и точности решения указанной задачи. Создание системы поиска релевантных документов в многоязычной среде (возможно, вместе с автоматических переводом документа на язык исходного документа) позволило бы существенно увеличить количество найденных релевантных документов. При этом, конечно, требуется ориентироваться на уже существующие поисковые службы, работающие с Интернетом, которые в свою очередь накладывают определенные ограничения на выбор решения.

Целью данной диссертации является исследование этих вопросов с целью реализации так называемой функциональности cross-language для поиска релевантных документов в сети Интернет, а также разработка системы для автоматического поиска релевантных текстовых документов с помощью некоторой поисковой службы.

# **Глава 1. Понятие поиска релевантных документов**

## 1.1 Понятие информационного поиска

Информационный поиск (information retrieval) — это процесс поиска в источнике информации (как правило, в памяти компьютеров) некоего неструктурированного материала (обычно - документа), удовлетворяющего информационные потребности. В зависимости от степени привлечения к информационному поиску технических средств и участия в нем человека различают ручной, машинный или автоматизированный информационный поиск. Автоматизированный информационный поиск может представлять из себя систему или набор пакетов для более удобного и быстрого поиска.

Под понятием «неструктурированные данные» будем подразумевать данные, которые не имеют ясной, семантически очевидной и легко реализуемой на компьютере структуры. В свою очередь, под понятием «структурированные данные» будем подразумевать данные, чья информация хранится в некотором упорядоченном виде. Из таких данных можно выбрать конкретную информацию исходя из требования информационного поиска. Одним из примеров таких данных является SQL базы данных, извлечение информации из которых выполняется посредством специально сформированного SQL запроса. Стоит отметить, что полностью неструктурированных данных не существует, т.к. любой текст, хранящийся в некотором документе на компьютере, использует некоторую структуру сильно приближенную к естественному языку (ЕЯ).

Рассмотрим основные термины информационного поиска:

1. Запрос - формализованный способ выражения информационных потребностей человеком. Для выражения информационной потребности используется язык поисковых запросов, синтаксис варьируется от системы к системе. Кроме специального языка запросов, современные поисковые системы позволяют вводить запрос на естественном языке.
2. Объект запроса - информационная сущность, которая хранится в базе автоматизированной системы поиска. Несмотря на то, что наиболее распространенным объектом запроса является текстовый документ, не существует никаких принципиальных ограничений. В частности, возможен поиск изображений, музыки и другой мультимедиа информации.
3. Релевантность - степень соответствия между сутью запроса и сутью найденного. Очень близким по смыслу является термин адекватность - оценка степени практической применимости результата.

Основной задачей информационного поиска является найти необходимую человеку информацию, которая наиболее точно выполняет его информационные потребности. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

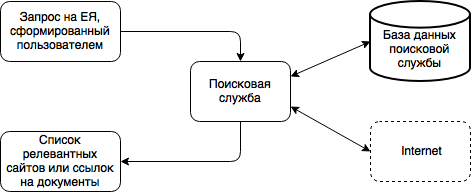
## 1.2 Поиск релевантных документов в Интернет-среде

Типичными средствами поиска любой информации в Интернет-среде являются поисковые службы. Поисковая служба - компьютерная система, предназначенная для поиска информации в сети Интернет. Примерами таких служб являются поисковые службы Google, Yandex, Yahoo, Bing и т.д.

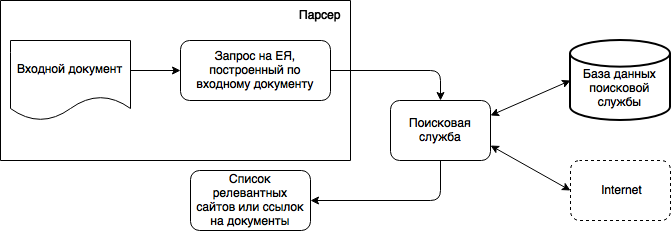
Эти службы подразделяются на несколько категорий в зависимости от способа их функционирования:

* Системы, основанные на поисковых роботах. Такие системы состоят из следующих основных частей: crawler, index и самого поисковика. Crawler является программой, которая сканирует вэб-страницы Интернета в поисках информации, которая поможет в дальнейшем системе подобрать более релевантную информацию для конкретного запроса пользователя, и формирует ее в определенный формат.  Index добавляет эту информацию в базу данных системы, чтобы она в последствии могла быть испольщована для формирования результатов запроса. Сама поисковая служба анализирует запрос на ЕЯ и с помощью NLP и других средств собирает нужные результаты из базы и показывает их пользователю.
* Системы, основанные на каталогах. Данные системы формируются посредством внесения в их каталог списка вэб-страниц вместе с дополнительной информацией о каждой из них (например, заголовок страницы, краткое описание содержимого страницы, список тэгов для этой страницы и т.д.). Когда будет обрабатываться запрос, введенный пользователем, подбор релевантных сайтов будет производится исходя из дополнительной информации.
* Гибридные системы. Системы, которые объединяют в себе работу двух предыдущих режимов.

Исходя из этой информации, работу поисковой службы можно представить следующим образом:

**Рисунок 1.1 – схема работы поисковой службы**

Поскольку разрабатываемая нами система должна принимать на вход документ, то схема выше приобретает следующий вид:



**Рисунок 1.2 – схема работы автоматической системы поиска релевантных документов**

Итак, мы составили примерную схему работы нашей системы.

## 1.3 Оценка качества результатов

Для того чтобы понять, насколько хорошо наша система отработала, введем следующие оценки.

Введем следующие обозначения. Пусть множество представляет из себя множество всех документов, к которым имеет доступ наша поисковая система. Под будем понимать запрос в виде удобном для обработки нашей системе. Тогда под будем понимать множество найденных документов по запросу , а под будем понимать множество релевантных документов для запроса . Перейдем к определению оценок.

Точность (precision). Под точностью понимается отношение количества найденных релевантных документов к количеству найденных документов. Точность характеризует насколько релевантным получается наш результат:

Полнота (recall). Под полнотой понимается отношение количества найденных релевантных документов к количеству релевантных документов. Полнота характеризует насколько много мы нашли релевантных документов из всего количества релевантных документов в нашей системе:

Выпадение (fall-out). Под выпадением понимается отношение количества найденных нерелевантных документов к количеству нерелевантных документов. Данная оценка не является основной для нашего исследования:

F-мера (F-score, мера Ван Ризбергена). F-мера является взвешенным средним гармоническим оценок точности и полноты. Эта оценка нужна для того, чтобы можно было судить о качестве нашего алгоритма исходя из одной оценки, а не двух. Высчитывается она следующим образом:

Можно обратить внимание, что параметр влияет на то, какая оценка для нас важнее. Так, например, выбирая мы отдаем предпочтение только полноте, а выбирая - точности. Если для нас обе оценки одинакова важны, то имеет смысл выбрать . Тогда наша формула примет вид:

## 1.4 Постановка задачи автоматического поиска релевантных документов

Задача автоматического поиска релевантных документов в контексте данной диссертации состоит в получении по исходному документу, состоящему из последовательности слов и знаков препинания, набора релевантных документов, полученных посредством поисковой службы. Язык исходного документа необязательно должен совпадать с языком полученного документа.

Для решения данной необходимо выполнить следующие задачи:

* Разработать алгоритм составления поискового образа по заданному документу;
* Составление по поисковому образу документа запроса для поисковой службы на ЕЯ;
* Выбор наиболее подходящей поисковой службы, а также ее использование в нашей системе.

Целью данной работы является разработка алгоритма составления образа документа, разработка системы для автоматического поиска релевантных текстовых документов с помощью некоторой поисковой службы, анализ полученные результатов.

* Разработать алгоритм автоматического распознавания ключевых слов в текстовом документе;
* Разработать алгоритм составления из ключевых слов запроса на ЕЯ для поисковой службы;
* Осуществить поиск документов с помощью поисковой службы на нескольких языках;
* Разработать ПО системы;
* Протестировать систему и сравнить полученные для нее оценки.

## 1.5 Выводы

В первой главе были получены следующие результаты:

* проанализирована литература по информационному поиску, структуре поисковых служб;
* приведен способы оценки результатов поиска в поисковой системе;
* изучены основные определения и термины;
* сформулирована постановка задачи.

# 

# список использованных источников

1. Manning C. Intriduction to Information Retrieval / Manning C., Raghavan P., Schütze H. // Camridge University Press, 2009 – 544 p.
2. Risvik K. Search engines and web dynamics / Risvik K., Michelsen R. // Computer Networks, 2002 – 289-302 p.
3. Bird S. Natural Language Processing with Python / Bird S., Klein E., Loper E. // O’Reilly, 2009 – 482 p.