

Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра автоматизации систем вычислительных комплексов

Треско Константин Игоревич

**Исследование и разработка средств многотемной классификации веб-страниц**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Научные руководители:**

к.ф-м.н Петровский Михаил Игоревич

Царев Дмитрий Владимирович

Москва, 2015

Оглавление

[Введение 3](#_Toc407053867)

[Системы управления корпоративным контентом (ECM) 4](#_Toc407053868)

[Методы классификации данных 5](#_Toc407053869)

[Сбор информации 6](#_Toc407053870)

[Системы предотвращения утечки данных 6](#_Toc407053871)

[Выводы из обзора систем 8](#_Toc407053872)

[Постановка задачи 9](#_Toc407053873)

[Обзор 9](#_Toc407053874)

[Обзор средств написания расширения для браузера 9](#_Toc407053875)

[Модуль многотемной классификации 11](#_Toc407053876)

[Исследование и построение решения 11](#_Toc407053877)

[Агент мониторинга 12](#_Toc407053878)

[Агент консолидации 13](#_Toc407053879)

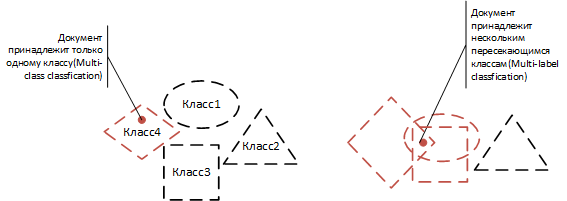
[Описание практической части 13](#_Toc407053880)

[Список используемых источников 15](#_Toc407053881)

# Введение

## Задача классификации

Настоящая работа посвящена исследованию и разработке программных средств сбора и многотемной классификации текстовых данных веб-страниц. Задача классификации многотемных документов (multi-label classification), заключается в определении принадлежности документа к одному или нескольким классам (из предопределённого набора классов) на основании анализа совокупности признаков, характеризующих данный документ. В отличие от традиционной задачи классификации, классы могут пересекаться или быть вложенными, то есть документ может принадлежать нескольким классам. Классы, к которым принадлежит документ называются релевантными.



1. Многотемная и многоклассовая классификация

На сегодняшний день существует ряд задач, для решения которых требуются системы сбора и классификации контента, частным случаем которого является веб-контент:

* Предотвращения утечек конфиденциальной информации
* Анализ работы пользователей
* Эффективная навигация по информационным ресурсам

Далее будут рассмотрены классы индустриальных систем, функционал которых включает в себя классификацию текстовых данных (в том числе веб-данных), с которыми работают пользователи.

* Системы управления корпоративным контентом (англ. Enterprise Content Management, ECM) — программные решения для управления информационными ресурсами предприятия предоставляют программные средства сбора, анализа, управления, накопления, хранения и доставки документов в масштабах организации. В настоящее время большинство ECM систем включают в себя системы обнаружения данным, связанных с определенным судебным делом (англ. eDiscovery). Средства eDiscovery обеспечивают процесс, с помощью которого организации находят, получают, сохраняют и анализируют документы, связанные с делом.
* Системы предотвращения утечки данных (англ. Data Loss Prevention, DLP) — программные решения для предотвращения утечек конфиденциальной информации и минимизации других рисков, связанных с внутренними угрозами.

## Системы управления корпоративным контентом (ECM)

С учетом постоянно растущего объема данных возникает необходимость в средствах, помогающим обеспечить эффективный доступ и управление информационными ресурсами. Системы управления корпоративным контентом предоставляют средства сбора, анализа, управления, накопления, хранения и доставки информации всем пользователям организации. В данном контексте информация (контент) является слабоструктурированной – это могут быть офисные текстовые и табличные электронные документы, документы в формате PDF и так далее.

Gartner (исследовательская и консалтинговая компания, специализирующаяся на рынках информационных технологий) выделяет следующие компоненты:

* управление документами (англ. Document management) — экспорт, импорт, контроль версий, безопасность и службы библиотек для деловых документов;
* управление образами документов (англ. Image-processing applications) — захват, преобразование и управление бумажными документами;
* управление потоками работ (англ. Content workflow) — поддержка бизнес-процессов, передача контента по маршрутам, назначение рабочих задач и состояний, создание журналов аудита;
* управление записями (англ. Records management, в соответствии с последним переводом стандарта IEEE 15489 — ГОСТ Р ИСО 15489-1-2007, «управление документами») — долгосрочное архивирование, автоматизация политик хранения и соответствия нормам регулирующих органов, обеспечение соответствия законодательным и отраслевым нормам;
* управление веб-контентом (англ. Web content management) — набор программных средств для управление веб-содержимым организации;
* социальный контент (англ. Social content) — реализует функционал для совместного использования документов, взаимодействия сотрудников, поддержки проектных команд;
* расширяющие компоненты (англ. Extended components) — компоненты, реализующие следующие функции: поиск и аналитика контента, электронное обнаружение или раскрытие информации (англ. electronic discovery, eDiscovery), архивирование данных (в том числе электронной почты, данных из других информационных источников) и т.п.

По данным Gartner лидером на рынке является продукт компании IBM, поэтому в силу схожести классов ECM систем, а так же наличие наиболее полной документации у продукта компании IBM, далее основная часть обзора будет производить на примере ECM системы компании IBM, но будут включены в рассмотрение также иные лидеры.

### Методы классификации данных

Классификация данных поддерживается в большинстве ECM системах. Существуют три основных подхода к классификации данных

1. Классификация на основе обучающей выборки (IBM Content Classification[2],Symantec eDiscovery[1])
2. Классификация на основе заданных правил (пример: IBM Content Classification — Decision plans [3]). Принадлежность документа к тому или иному классу определяется на основе эвристических правил (сигнатур). Правила могут формироваться на основе контекста – имя отправителя, директория создания и т.п., а также на основе контента – шаблоны текста, ключевые слова и т.п.
3. Определение категорий в неизвестных документах (IBM Content Classification — Taxonomy Proposer [3]). В задаче кластеризации нет предопределенного набора классов. Исходное множество документов разбивается на подмножество таким образом, чтобы документы в различных подмножествах существенно отличались

### Сбор информации

Для сбора данных в ECM системах существуют отдельные компоненты, в терминологии IBM называемые Искатели[6]. Искатели установлены на каждом из источников информации и передают информации либо непрерывно, либо по заданному расписанию. Для создания выборки анализируемых документов задается набор и конфигурация сбора

## Системы предотвращения утечки данных

Системы предотвращение утечки данных (Data Loss Prevention, DLP) - программные решения для предотвращения утечек конфиденциальной информации и минимизации других рисков, связанных с внутренними угрозами.

Согласно Gartner DLP системы включают:

* набор технологий классификации контентной информации (например, данные, содержащейся в файлах, электронных письмах, хранилищах данных), которая может находится в трёх состояниях: хранимые данные (англ. Data-at-Rest), используемые/обрабатываемые данные (англ. Data-in-Use), передаваемые данные (англ. Data-in-Motion);
* контролирующие техники для применения политик безопасности к контентным данным: ведение журнала событий, уведомление, перемещение (карантин), шифрование, маркеровка (англ. tagging) .и др.

Далее обзор будет проведен на основе лидеров по данным Gartner.



2. Данные Gartner 2013

Основными технологиями определения конфиденциальности документа являются:

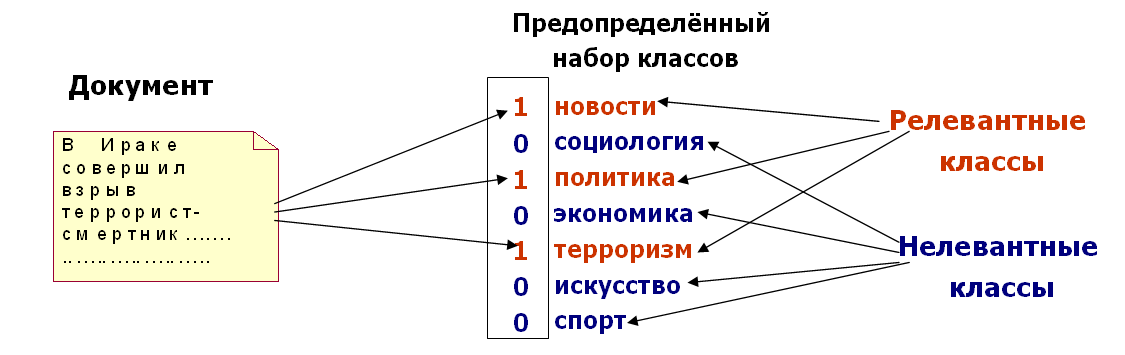
* Цифровые отпечатки(англ. digital fingerprint) — технология предназначена для защиты больших по объему документов, содержание которых не изменяется или меняется незначительно. Детектор цифровых отпечатков позволяет автоматически обнаруживать в анализируемом тексте цитаты из документов-образцов, содержащих конфиденциальную информацию.
* Анализ шаблонов – технология предназначена для детектирования алфавитно-цифровых объектов по шаблону данных (маске) и позволяет наиболее эффективно выявлять факты пересылки персональных данных или финансовой информации. Кроме того, данная технология может использоваться как вспомогательный метод для обнаружения фактов несанкционированной пересылки внутренних документов, содержащих формализованные данные, образованные по определенному шаблону (например, договоров или счетов в случае детектирования банковских реквизитов, кодов классификаторов и т.д.).
* Машинное обучение (InfoWatch[4], Symantec[5]) . Примерная схема работы классификатора такова: на вход подается обучающий набор документов, состоящий как из конфиденциальных так и не конфиденциальных документов. По этим наборам производится обучение и строится статистическая модель. Далее полученная модель используется для классификации неизвестным документов. При обнаружении конфиденциального документа производятся действия, предписанные политиками безопасности.

Основное преимущество машинного обучения заключается в том, что в отличие от других описанных технологий, оно предназначено для работы не со статическими, а с постоянно меняющимися документами.

## Специфика задачи классификации текстовых веб-данных

Веб – страницы имеют следующие особенности

* Содержимое веб-страниц не подлежит шаблонному разбору
* Содержимое имеет многотемную природу, то есть каждая из страниц может одновременно принадлежать к нескольким тематикам



## Выводы

На сегодняшний день существует ряд прикладных задач, требующих классификации текстовых данных. Обзор индустриальных систем показал, что наиболее актуальными технологиями классификации текстовых данных являются

* Метод шаблонов
* Метод цифровых отпечатков
* Метод машинного обучения

Первые две технологии применимы только к статическим данным, в то время как метод машинного обучения позволяет позволять адаптироваться к тому, что содержимое и состав анализируемых данных постоянно меняется.

Для задачи классификации текстовых веб – данных из рассмотренных технологий подходит только машинное обучение с возможностью дообучения, так как веб – данные и набор классов постоянно меняются.

Рассмотренные системы решают задачу многоклассовой классификации, так как документ может принадлежать только к одному из предопределенного набора классу. Так как большая часть веб – данных имеет многотемную природу, то существует актуальность разработки системы сбора и многотемной классификации текстовых веб - данных пользователей. С учетом большого объема анализируемой информации к модулю классификации предъявляется требование масштабируемости.

Так как в рассмотренных системах присутствуют компоненты сбора информации, расположенные на пользовательских машинах, то при разработке системы сбора нужно учитывать то, что деятельность компонент никак не должна сказывать на работе пользователя, то есть к системе предъявляется требование производительности. Также в рассмотренных системах пользователь не имеет доступа к собираемой информации, поэтому система сбора должна удовлетворять требованию защищенности.

# Постановка задачи

Осуществить сбор и многотематическую классификацию веб-данных пользователей с обеспечением следующих требований:

* Модуль сбора должен обеспечивать
  + Масштабируемость (линейный рост расхода ресурсов при росте числа подключений)
  + Производительность (компоненты сбора, установленные на пользовательских машинах, не должны влиять на работу пользователя)
  + Защищенность (пользователь не должен иметь возможности фальсифицировать данные)
  + Функционирование под ОС Windows и браузером IE
* Модуль классификации должен обеспечивать
  + Многотемную классификацию на основе машинного обучения с возможностью дообучения

# 

# Обзор

Проведенный анализ систем ECM и DLP в контексте задач классификации показывает, что каждый из пользователей является источником анализируемой информации, собираемая информация должна храниться к единой базе данных. Для управления потоком данных из различных источников используется мультиагентная система. Под агентом подразумевается автономный процесс, способный реагировать на среду исполнения и вызывать изменения в среде возможно, в кооперации с пользователями или с другими агентами.

Разрабатываемая система должна состоять из:

* Агентов мониторинга, расположенных на каждом из пользовательских компьютеров, собирающих текстовые данные просматриваемых пользователем веб – страниц и отправляющих их агенту консолидации
* Агента консолидации, принимающего данных от всех агентов мониторинга и сохраняющего их в единую базу данных
* Модуля классификации

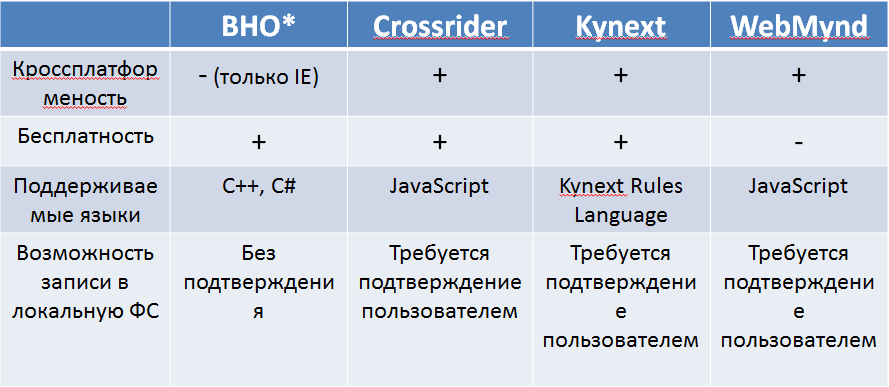
Агент мониторинга состоит из двух компонент:

* Расширение для браузера, сохраняющее html код просматриваемой пользователем веб – страницы в локальную базу данных, расположенную на пользовательском компьютере
* Модуль передачи информации из базы данных агенту консолидации

## Обзор средств написания расширения для браузера

Для решения задачи необходимо расширение для браузера, имеющее возможность сохранить html код просматриваемой пользователем веб – страницы в локальную базу данных. Так как система сбора должна удовлетворять требованиям производительности, то сохранение и обработка данных должны происходить без участия пользователя.

* BHO (Browser Helper Object) - DLL-модуль, разработанный как плагин для Internet Explorer для обеспечения дополнительной функциональности. В BHO API существует возможность получения доступа к DOM текущей страницы. Также существует возможность обработки событий и управлению навигацией. Для задачи перехвата контента данное решение подходит. Минусом является то, что данное решение подходит только для браузера IE. Но BHO имеет возможность записи и чтения данных из файловой системы пользователя, без его участия
* Kynext позволяет перехватывать содержимое веб – страницы. Поддерживает IE, Firefox, Safari, and Chrome, но для написания необходимо использовать проприетарный язык [Kynetx Rules Language](http://wiki.kynetx.com/pages/Overview_of_KNS). Еще одним минусом является то, что работа написанное расширение зависит от работоспособности самого расширения Kynext.
* WebMynd поддерживает IE, Firefox, Safari, and Chrome. Пока доступна бета версия, и не известно, будет ли дальнейшая поддержка продукта. Поддерживает JavaScript API. Не является бесплатным обеспечением.
* Crossrider позволяет быстро создавать кросс-браузерные расширения. Использует один API и поддерживает JavaScript и jQuery, так что разработчик с базовыми знаниями JavaScript может писать и поддерживать свой код. Имеется возможность писать под IE, Firefox, Safari, and Chrome. Бесплатен в использовании. Существует [документация](http://docs.crossrider.com/) и [демо – видео](http://crossrider.com/developer/demo) для некоторых задач. Доступ к файловой системе пользователя только с его разрешения.



3. Обзор средств написания расширения для браузера

Требованию производительности удовлетворяет и защищенности удовлетворяет Browser Helper Object

## Обзор средств многотемной классификации

Можно выделить три основных подхода к многотемной классификации

* «Оптимизационные» подход [4,7,8,9]
* Подход на основе декомпозиции в набор независимых бинарных проблем [5,6]
* Подход на основе ранжирования[6,11,12]

### Оптимизационный подход

К оптимизационным относятся методы классификации, в которых в явном виде задан показатель качества, который необходимо обратить в экстремум (максимум или минимум) по множеству допустимых разбиений.

К алгоритмам этой группы можно отнести

* AdaBoost.MH [7]

Суть заключается в том, что если у нас есть набор эталонных объектов (точки на плоскости), т.е. есть значения и класс, к которому они принадлежат (например, -1 – красная точка, +1 – синяя точка), кроме того имеется множество простых классификаторов (набор вертикальных или горизонтальных прямых, которые разделяют плоскость на две части с наименьшей ошибкой), то мы можем составить один лучший классификатор. При этом в процессе составления или обучения финального классификатора акцент делается на эталоны, которые распознаются «хуже», в этом и заключается адаптивность алгоритма, в процессе обучения он подстраивается под наиболее «сложные» объекты.

Основной недостаток алгоритма заключается в том, что нет возможности дообучение и добавление или удаления тематики

* Multi-Label-kNN
* метод на основе модели смешивания, обученной с помощью метода EM

## Модуль многотемной классификации

Для решения задачи многотемной классификации был выбран модуль, разработанный в Лаборатории Технологий Программирования, так как он удовлетворяет следующим требованиям

* классификация многотемных документов на основе машинного обучения с возможностью дообучения
* возможность удаления существующих и добавления новых тематик классификации
* скорость классификации, удовлетворяющая интерактивному режиму работы пользователей
* высокое качество классификации, не уступающее современному уровню разработок по данной проблеме

# Исследование и построение решения

Для решения задачи сбора и многотемной классификации текстовых данных веб – страниц пользователя был выбран мультиагентный подход[6], состоящий из:

* Агентов мониторинга
* Агента консолидации
* Модуля многотемной классификации

## Агент мониторинга

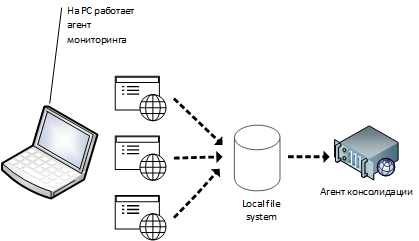
Агент мониторинга состоит из нескольких компонент

* Расширение для браузера
* Модуль передачи данных агенту консолидации

Расширение для браузера, написанное с помощью BHO, считывает html код просматриваемой пользователем веб-страницы и сохраняет ее в локальную базу данных. Сохранение в локальную базу данных осуществляется для контроля нагрузки на агент консолидации и возможности отправления данных по расписаниюрер

, а также на случай потери связи с агентом консолидации.

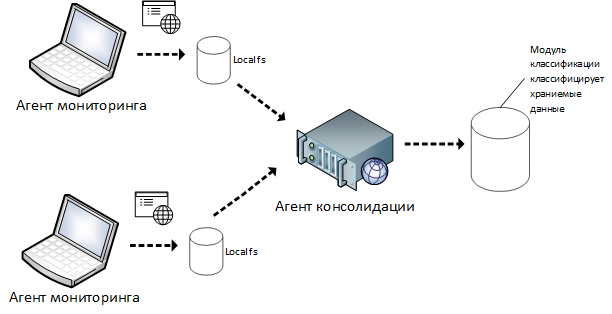
Модуль передачи данных агенту подключается к локальной базе данных и отправляет хранящуюся в ней информацию агенту консолидации, при получении ответа от агента консолидации отправленные данные удаляются из локальной базы данных, чтобы размер локальной базы данных не увеличивался постоянно.



4. Работа агента мониторинга

## Агент консолидации

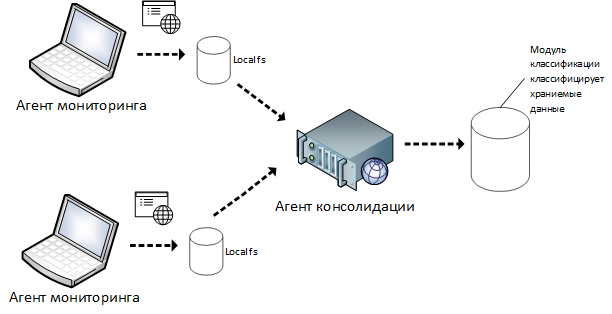
Агент консолидации сохраняет данные, полученные от всех агентов мониторинга, и сохраняет их в единую базу данных.



5. Работа агента консолидации

# Описание практической части

* Расширение для браузера написано с помощью BHO на языке C#. Данные сохраняются в локальную базу данных реализованную с помощью SQLite
* Модуль передачи данных агенту консолидации написан на языке C#, передача данных осуществляется с помощью System.Net.Sockets
* Агент консолидации написан на языке C#



6. Архитектура системы сбора и многотемной классификации

# Список используемых источников

1. Symantec eDiscovery (<http://www.symantec.com/ru/ru/predictive-coding/>)
2. Component overview (Content Classification 8.8.0) (<http://www-01.ibm.com/support/knowledgecenter/SSBRAM_8.8.0/com.ibm.classify.admin.doc/c_AG_about_rme.htm>).
3. Infowatch БКФ (<http://www.infowatch.ru/technologies/linguistic_analysis/create_bkf>)
4. Symantec (<http://eval.symantec.com/mktginfo/enterprise/white_papers/b-dlp_machine_learning.WP_en-us.pdf>)
5. «Распределенные системы. Принципы и парадигмы» [Эндрю Таненбаум](http://www.ozon.ru/person/1218176/), [М. ван Стеен](http://www.ozon.ru/person/1431775/)
6. Component overview (IBM Watson Content Analytics 3.5.0) (<http://www-01.ibm.com/support/knowledgecenter/SS5RWK_3.5.0/com.ibm.discovery.es.nav.doc/iiysaovcomp.htm?lang=en>).