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. М.И. Петровский
Д.В. Царев

Аннотация

Целью данной работы является исследование и разработка системы сбора и многотемной классификации веб-страниц, с которыми работали пользователи, находящиеся внутри одной сети. Разрабатываемая система сбора должна удовлетворять требованиям масштабируемости (линейного роста расхода ресурсов при увеличении числа подключений) и защищености (пользователь не должен иметь возможности фальцифицировать собранные данные). А модуль классификации данных должен обеспечивать многотемную классификацию на основе машинного обучения с возможностью добавления и удаления тематик.

В ходе работы был проиведен обзор существующих средств многотемной классификации и средств реализации системы сбора. Также была реализова система сбора и многотемной классификации, удовлетворяющая поставленным выше требованиям, проведено тестирование, показавшее приемлемую масштабируемость.

1 Введение

1.1 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.

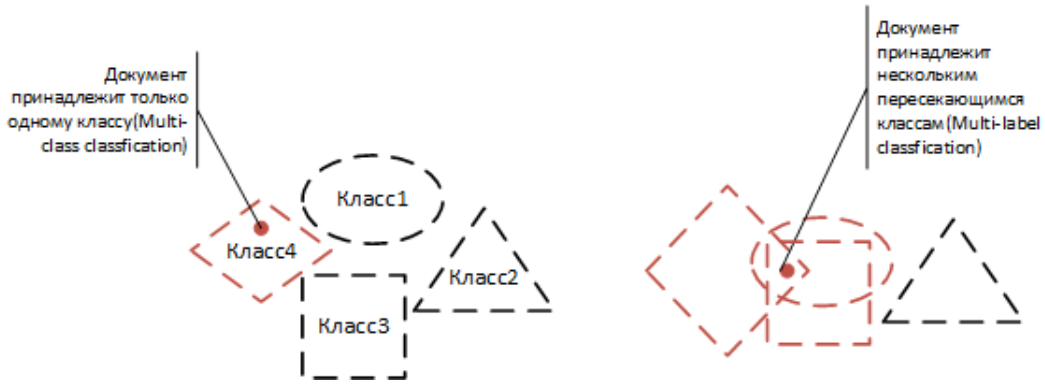

Рисунок 1 – Многотемная и многоклассовая классификация

1.2 Области применения классификации веб-данных

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

– **Анализ работы сотрудников организации.** Определение тематик документов, с которыми работает пользователь. На основе полученной информации применение политик безопасности

– **Фильтрации контента, для обнаружения информации, относящейся к определенному делу (eDiscovery).** Документы и электронная почта могут фильтроваться в зависимость от присвоенных им классов, чтобы гарантировать, что только материалы с запрошенной бизнес-информацией в рамках расследуемого дела были найдены и сохранены.
– **Определения конфиденциальности документа, для предотвращения утечек информации** По различным экспертным оценкам в настоящее время наибольшие риски для информационной безопасности представляют не внешние, а внутренние угрозы [1] Поэтому существует необходимость в средствах определения конфиденциальности документа

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.

1.3 Основные методы классификации данных в корпоративных системах

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

– **Классификация на основе обучающей выборки** (IBM Content Classification [2], Symantec eDiscovery [3]
– **Классификация на основе заданных правил**. 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.

– **Определение категорий в неизвестных документах**. В задаче кластеризации нет предопределенного набора классов. Исходное множество документов разбивается на подмножество таким образом, чтобы документы в различных подмножествах существенно отличались.

В DLP системах выделяют следующие технологии классификации данных:

– **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 VML [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.

1.4 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

Рисунок 2 – Многотемная природа документа

1.5 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.

2 Постановка задачи

Разработать архитектуру и реализовать прототип системы сбора и многотемной классификации текстовых веб-данных пользователя в соответствии с требованиями:

- 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

3 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

3.1 Расширение 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

Таблица 1 – Средства расширения для браузера

	BHO	Crossrider	Kynext
Кроссплатформеность	-(только IE)	+	+
Бесплатность	+	+	+
Поддерживаемые языки	C++,C#	JavaScript	Kynext Rules Lang
Возможность записи в локальную ФС	+	-	-

- **Kynext** позволяет перехватывать содержимое веб – страницы. Поддерживает IE, Firefox, Safari, and Chrome, но для написания необходимо использовать проприетарный язык Kynetx Rules Language. 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 и демо – видео для некоторых задач. Доступ к файловой системе пользователя только с его разрешения.

3.2 Методы многотемной классификации

Можно выделить три основных подхода к многотемной классификации

- «Оптимизационный» подход
- Подход на основе декомпозиции в набор независимых бинарных проблем
- Подход на основе ранжирования

3.2.1 Методы, основанные на "оптимизационном"подходе

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 данным алгоритмам можно отнести:

- Основанные на AdaBoost алгоритме (AdaBoost.MH [?],ADTBoost.MH [?]) - минимизируется функция Hamming Loss)
- Multi-Label-kNN [?] максимизируются апостериорные вероятности принадлежности классам

Основной недостаток алгоритмов, основанных на оптимизационном подходе, заключается в том, что для решения задачи оптимизации необходимо хранение всего тренировочного набора, таким образом, нет возможности добавления и удаления тематик или дообучения

3.2.2 Методы, основанные на декомпозиции в набор независимых бинарных проблем

Кратко суть методов, основанных на декомпозиции в набор независимых бинарных проблем, можно описать так:

- Пусть дано N классов. Для каждого из N классов строится бинарный классификатор
- Далее, основываясь на решающей функции для каждого из классификаторов, можно определить релевантность класса

Основным поводом для критики методов этой группы является то, что строятся независимые классификаторы, которые не учитывают корреляции между классами, что существенно для многотемной классификации. Еще одним недостатком является высокая вычислительная сложность - что количество бинарных подзадач равно числу классов, а каждая подзадача обучается на всём тренировочном наборе.

3.2.3 Методы, основанные на подходе ранжирования с последующим отсечением нерелевантных классов

Работу алгоритмов можно разделить на два этапа

- Первый этап состоит в обучении алгоритма ранжирования, который упорядочивает все классы по степени их релевантности для заданного классифицируемого объекта
- Второй этап заключается в построении функции многотемной классификации, отделяющей релевантные классы от нерелевантных

4 Исследование и построение решения

4.1 Разбиение на подзадачи

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

- Агента мониторинга
- Агента консолидации
- Модуля многотемной классификации

4.1.1 Агент мониторинга

Агент мониторинга состоит из нескольких компонент:

- Расширение для браузера
- Модуль передачи данных агенту консолидации

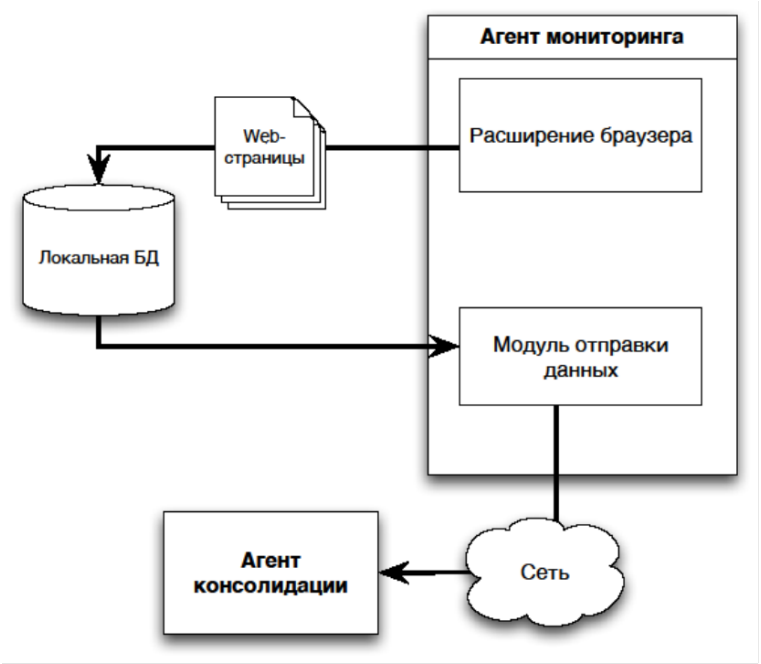

Рисунок 3 – Архитектура агента мониторинга

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 расписанию, 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.

4.1.2 Агент консолидации

Агент консолидации сохраняет данные, полученные от всех агентов мониторинга, и сохраняет их в единую базу данных.Также в базу данных заносится дополнительная информация о страницах:

- Логин пользователя, посещавшего веб-страницу
- Имя компьютера, на котором находится пользователь
- Дата посещения
- ID веб-страницы

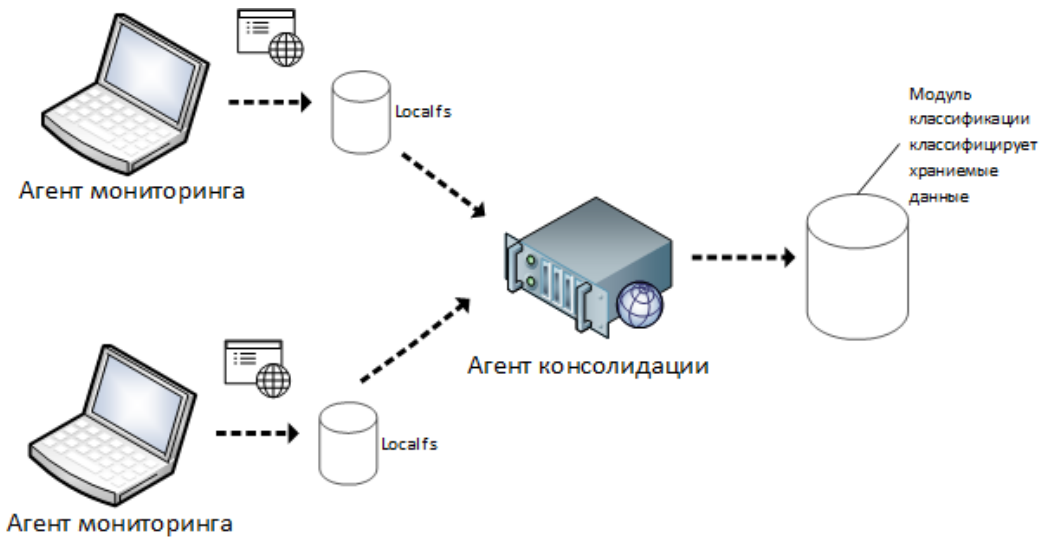

Рисунок 4 – Архитектура агента сбора

4.1.3 Модуль классификации

В ходе обзора было принято решение использовать модуль классификации, реализованный в лаборатории Технологий Программирования. Он предоставляет следующие сценарии работы:

- **Обучение.**Построение модели классификации на основе совокупности заранее рубрицированных гипертекстовых документов
- **Классификация.**Применение построенной модели к новому классифицируемому документу
- **Дообучение.**Модификация модели классификации на основе дообучения на новых документах с релевантными для них тематиками
- **Удаление темы.**Удаление тематики классификации из модели без необходимости последующего обучения "с нуля"

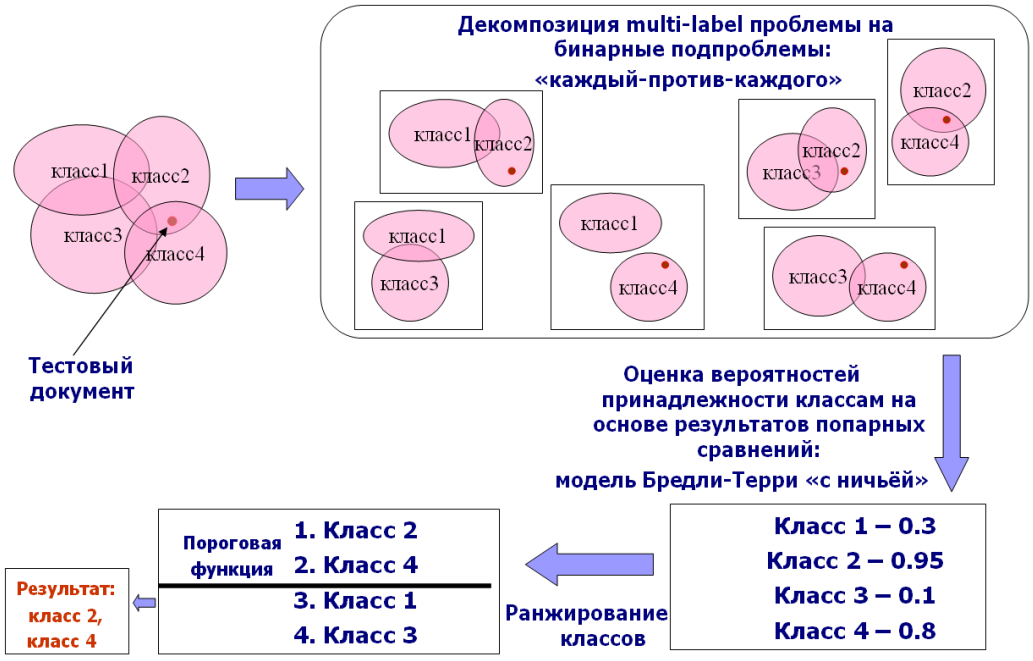

Рисунок 5 – Работа модуля классификации

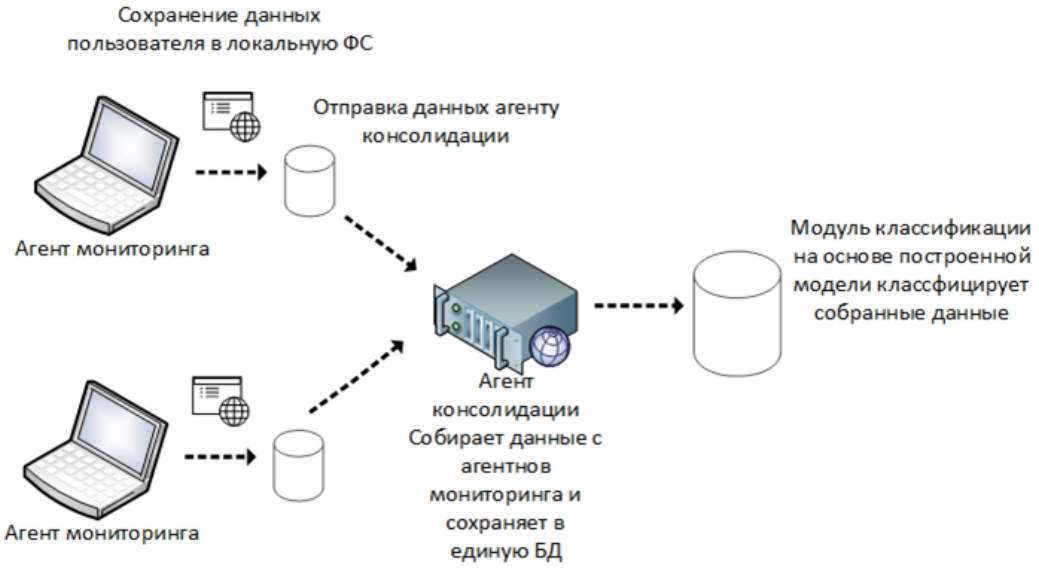

Рисунок 6 – Архитектура

4.1.4 Архитектура предложенного решения

Список литературы

1. *Аналитический Центр InfoWatch, Безопасность информации в корпоративных информационных системах. Внутренние угрозы* (http://www.infowatch.ru/analytics/reports/4609)
2. *Component overview (Content Classification 8.8.0)*
3. *Предиктивное обучение* http://www.symantec.com/ru/ru/predictive-coding/
4. *Infowatch БКФ* http://www.infowatch.ru/technologies
5. *Symantec Machine Learning* http://eval.symantec.com/mktginfo/enterprise/