Адаптивные информационные технологии автоматической классификации документов по их важности и критичности

И.А. Сергеев
Институт математики, информационных технологий и физики
Удмуртский государственный университет
Ижевск, Россия
e-mail: prizrak5097@yandex.ru

Аннотация[1]

В данном исследовании будут рассмотрены различные методы и средства классификации электронных документов по их важности и критичности. Обзор в данной области позволит определить, какие способы анализа электронной неструктурированной информации существуют и посредством чего можно определить и измерить важность электронного документа. Учитывая, что развитие информационных технологий способствует быстрому росту информации, а неструктурированной информации всегда будет больше, чем структурированной, данная тема является достаточно актуальной на текущий момент. Также будут рассмотрены основные проблемы, которые влечёт за собой неконтролируемая генерация большого количества неструктурированных документов.

1. Введение

С древнейших времён информация была одним из важнейших ресурсов, доступных человеку. Люди, не задумываясь, каждую секунду взаимодействуют с информацией во всех сферах жизни. На протяжении всей жизни человек осуществляет структуризацию и извлечение информации из окружающего нас мира. В древние времена полученная информация обеспечивала нам получение средств, необходимых для жизни - таких как еда, оружие, лекарства и одежда, более того, полученная информация передавалась в виде опыта последующим поколениям, что обеспечивало выживание человеческого общества.

Информация непрерывно совершенствовалась, обновлялась и изменялась. Всё новые и новые открытия позволяли человеческому обществу перейти к чему-то новому, более сложному [1].

Со временем информация стала приобретать всё новые и новые формы, она становилась сложнее и многостороннее. Это проявлялось в появлении новых сфер для исследования и развития. А так как информации становилось всё больше и больше, человечеству понадобились новые способы хранения и передачи информации - глиняные таблички, свитки, наскальная живопись, памятники - в древние времена это были одни из немногих способов хранения информации и, разумеется, как информации становилось много, так и способы хранения и передачи совершенствовались и становились более эффективными.

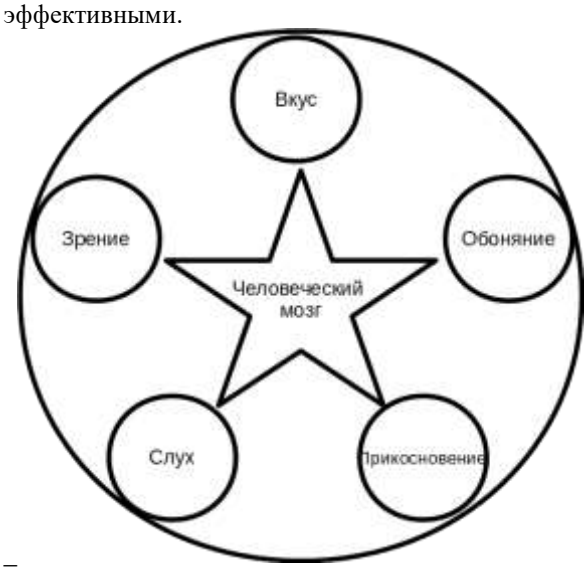

Рис. 1. - Способы восприятия информации

Способы восприятия информации, указанные на рисунке №1, за последние тысячелетия не

Труды Седьмой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-30 мая, Уфа-Ставрополь-Ханты-Мансийск, Россия, 2019

поменялись, как и сама информация. Какой бы облик информация ни принимала, она всегда будет доступна нам лишь посредством пяти органов восприятия, что, в свою очередь, наложило определённое ограничение на способы хранения, передачи, обработки и отображения [2].

2. Актуальность

Развитие сферы информационных технологий заставило людей по-новому взглянуть на само значение информации. За последнее столетие человеческое общество активно переходит от машинного производства к информационным технологиям, что, в свою очередь, влияет не только на способы производства, но и на все сферы человеческого жизни. Более того, развитие сферы информационных технологий достаточно сильно повлияло на отношение к самому понятию информации и позволило вывести на совершенно новый уровень воздействие человека на информацию, и это касается абсолютно всех операций воздействия на информацию, в частности, на операции, указанные на рисунке №2

Рис. *1*. - Операции, возможные над информацией

Использование информации стало намного эффективнее, более того, информационные технологии проникли во все сферы нашей жизни, мы видим их абсолютно везде. С каждым годом они становятся всё совершеннее и совершеннее, что напрямую связано с выработкой информации в мире. Появление информационных технологий связано как раз с ростом информации в мире. А так как операции, возможные над информацией, стали проще и совершеннее, то вполне неудивительно, что за последние несколько лет было выработано огромное количество информации. Только в 2002 г. человечеством было произведено информации $18 \cdot 10^{18}$ байт (18 Эксабайт). За пять предыдущих лет человечеством было произведено информации больше, чем за всю предшествующую историю. Объём информации в мире возрастает ежегодно на 30 %, это можно увидеть на рисунке №3 [3,4]:

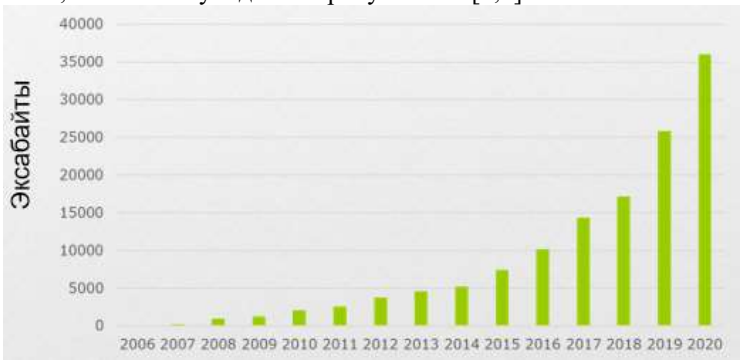

Рис. 3. - Изменение объёма информации за последние годы

В современном мире информация играет очень большую роль в жизни людей. Сейчас это один из главных ресурсов, наряду с такими ресурсами, как финансовые, человеческие или даже материальные. Производство и потребление информации является важным процессом для организации основы эффективного функционирования и развития различных сфер общественной жизни, в частности, экономики. Сейчас ни одно предприятие не может представить свою жизнь без применения информационных технологий в работе. Применение бумажной документации с каждым годом теряет актуальность, всё больше людей предпочитают использовать информацию, представленную в электронном виде, что вполне очевидно, так как это более эффективно для работы [5].

3. Основные проблемы

Тем не менее, во всём этом, казалось бы, полезном развитии информации есть и достаточно ощутимые минусы. Возможность более простой генерации информации в результате даёт нам огромное количество неиспользуемой информации, которая забивает устройства для хранения информации и, например, попросту усложняет поиск документа на компьютере. Особо остро данная проблема стоит на предприятиях. Предприятие может обладать значительными объёмами для хранения информации, и неважно, хранится всё в облачном хранилище или на сервере. При неправильном использовании электронной документации возможны такие ошибки, как:

1. Создание большого количества дубликатов документов.
2. Потеря информации.
3. Сложность идентификации необходимой информации.

Это достаточно неприятные ошибки, которые могут понизить уровень эффективности работы предприятия. Так как обучение и контроль пользователей достаточно проблематичен, возникает необходимость управления информацией пользователей, а для этого необходимо, чтобы

система определяла статус важности того или иного документа, или даже определяла, какую информацию содержит файл. В случае с бумажной документацией для этого существуют архивы, где специальный человек (архивариус) следит за порядком, обеспечивает сохранность документации и может по запросу предоставить нужную информацию. Нечто подобное требуется от идеальной информационной системы, она должна уметь различать документацию в том же значении, в каком это делает человек. Если система сможет определять, какой документ важный, а какой нет, это значительно облегчит работу с документами, не говоря уж о том, что это позволит не человеку, а системе разбираться со всеми проблемами, которые связаны с файлами. Передача работы с документами с человека на машину уже несколько десятилетий является мечтой многих компаний. Примерно подобная ситуация была во времена, когда человечество переходило от ручного производства к машинному [6].

4. Анализ собранной информации

Информацию, полученную в ходе анализа файла, можно разделить на три основных блока, которые мы можем увидеть в представленной ниже таблице №1:

Таблица *1* - Блоки информации, полученной в ходе сборки информации о файле

Файл		
Внешняя информация	**Содержимое файла**	**Сведения о файле**
Информация о местонахождении файла	Формулы	Формат
В какой подсети находится рабочая станция.	Текст	Имя
Кому принадлежит рабочая станция и кто создатель файла	Изображения	Размер
Связь с другими документами	Мультимедиа	Прочие метаданные о файле

После сбора информации по каждому из блоков мы получаем информацию, которая будет участвовать в анализе, а именно:

1. **Внешняя информация**

Информация в данном блоке прежде всего позволяет узнать, кем является человек, который создал файл и которому принадлежит рабочая станция, род занятий, профессию уровень доступа и прочую информацию подобного типа. Информация о месторасположении рабочей станции и подсети, в которой находится рабочая станция, позволит уточнить, например, откуда поступает информация, а, следовательно, вывести путь информационного потока, место скопления файлов и т.д. При наличии связи файла с другими файлами можно вывести структуру их взаимодействия.

2. **Содержимое файла**

В этом блоке информация является наиболее важной, так как именно по анализу содержимого можно определить, насколько важен файл, даже просто краткий анализ наполнения файла позволяет вывести крайне полезную информацию о нём. Например, если в информации множество формул и часто упоминается слово «деньги», можно сделать вывод, что документ является бухгалтерским или финансовым. Посредством даже такой простой информации можно многое узнать.

3. **Сведения о файле**

Как таковых сведений о файле или мета информации достаточно большое количество, но в таблице №1 указаны лишь самые основные. Посредством их анализа мы можем определить, о чём данный файл, его основное назначение и объём потребляемой информации, уже на основе этих данных можно многое узнать [7].

При наличии достаточно количества информации по каждому из блоков мы можем определить степень важности файла и, более того, полученная для анализа информация сама по себе является полезными статистическими данными. Важно учесть, что вся собранная информация должна быть записана в созданную для этого базу. Для определения степени важности и полезности информации по каждому из блоков ниже представлен рисунок №4:

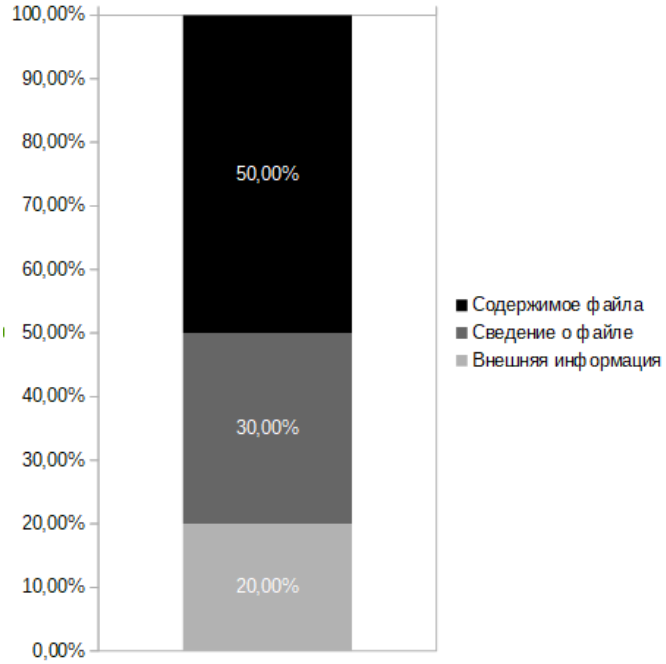

Рис. 4. Разделение собранной информации по важности

Согласно информации, представленной на рисунке №4, основным и наиболее важным блоком является блок, содержащий информацию о содержимом файла, за ним по важности следует блок сведений о файле, а после - блок внешней информации. Каждый из этих блоков является важной составляющей необходимых для анализа данных. Также стоит отметить, что чем более важен полученный блок, тем больше усилий затрачивается на его обработку в целом. После получения данных о файле мы можем их проанализировать и на основе этой информации сделать вывод, нужен файл или нет. Также полученную информацию можно использовать для статистики [8].

Данные пользователей, как и любая другая информация, не являются однородными, они разделяются на структурированную, неструктурированную и слабоструктурированную информацию.

1. **Неструктурированная информация** – это информация, которая не имеет какой-то заданной структуры или даже простейших связей между данными.
2. **Структурированная информация** – это информация с конкретно заданной связью между заданными элементами информации и метаданными.
3. **Слабоструктурированная или полу-структурированная информация** является неструктурированной информацией с простейшей внутренней структурой.

Для большей наглядности того, что представляют из себя указанные типы данных, ниже представлена таблица №2:

Таблица 2 – Типы данных и их представление

Тип данных	Вид информации
Неструктурированная информация	Книги
	Документы
	Веб-сайты
	Диалоги
	Мультимедиа
	Изображения
Слабоструктурированная информация	Заказы
	Платёжные документы
	Счета
	Xml-документы
Структурированная информация	Таблицы
	Базы данных
	Объекты и связи

В таблице №2 представлена не вся разделённая на три типа информация, а лишь часть, для того, чтобы более объективно понимать, чем же отличаются вышеуказанные типы данных. Основным и самым важным отличием является наличие связи между элементами. Структурированная информация отличается от информации на основе файлов тем, что она хранится в формате, который сохраняет не только сами данные, но и различные сведения об этих данных и об их структуре. Слабоструктурированную информацию обычно достаточно редко упоминают, тем не менее эта форма данных содержит теги и другие маркеры для отделения семантических элементов и для обеспечения иерархической структуры записей и полей в наборе данных [9].

В начале статьи уже упоминалось о непрекращаемом росте информации в мире, по последним данным количество неструктурированной информации в мире намного больше, чем структурированной. Особо явно это можно проследить по разнице в типах данных предприятий, указанных на рисунке №5

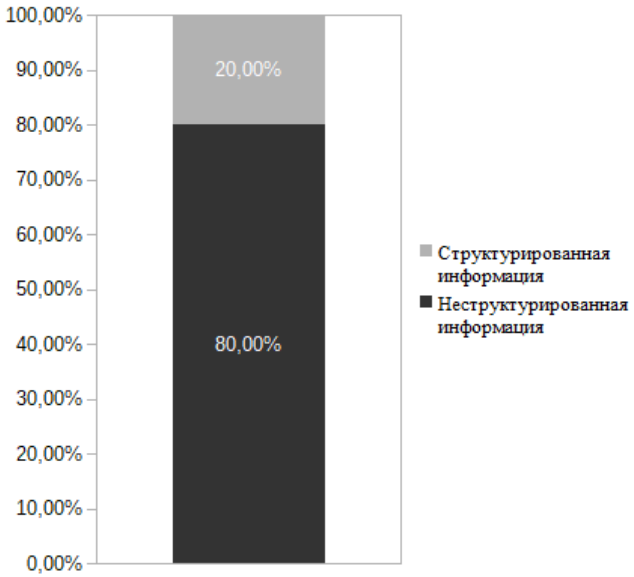

Рис. 5. - Разница в типах данных информации предприятия

Такая существенная разница в типах данных заключается в способах генерации информации. Под определение неструктурированной информации попадает намного большее количество различной информации, как это было указано в таблице №2. Генерация неструктурированной информации осуществляется намного проще, а взаимодействуют с ней часто. И это не обязательно могут быть документы, это могут быть абсолютно любые данные. Например, на рабочей станции среди неструктурированной информации структуры хранения информации нет, пользователю приходится создавать структуру по собственному видению. Старая информацию либо дубликаты пользователем не удаляются, а наоборот, множатся. Страх потери или повреждения информации толкает их на дополнительную генерацию файлов. А так как пользовательской информации меньше не становится, а резервные копии данных на пользовательских рабочих станция не делают,

следовательно, риск потери информации возрастает многократно.

Структурированная информация генерируется несколько медленней, так как управлением информации помимо пользователя занимается система, более того, в 90% случаев структурированная информация сохраняется в виде резервной копии. В таблице №2 было указано, в каком виде представлена структурированная информация. Обычно под структурированной информацией на предприятии понимают различные ресурсы, взаимодействующие с базой данных. Это могут быть различные СЭДО, справочная система, например, КонсультантПлюс, то есть практически любые программные продукты, которые взаимодействуют с базой данных. Структурированная информация может генерировать неструктурированную, а неструктурированная может быть добавлена в базу данных [10].

5. Методы и средства для определения важности документа

Определение важности документа в структурированной информации осуществляется просто. Каждый файл, каждая таблица и связь имеют точно выведенную структуру, а раз есть структура, значит, можно проследить практически всю информацию о файлах и данных в целом. Но неструктурированная информация структуры не имеет и в лучше случае может быть просто разбита пользователем по папкам. Простейшая схема алгоритма на определение степени важности неструктурированного файла представлена на рисунке №6

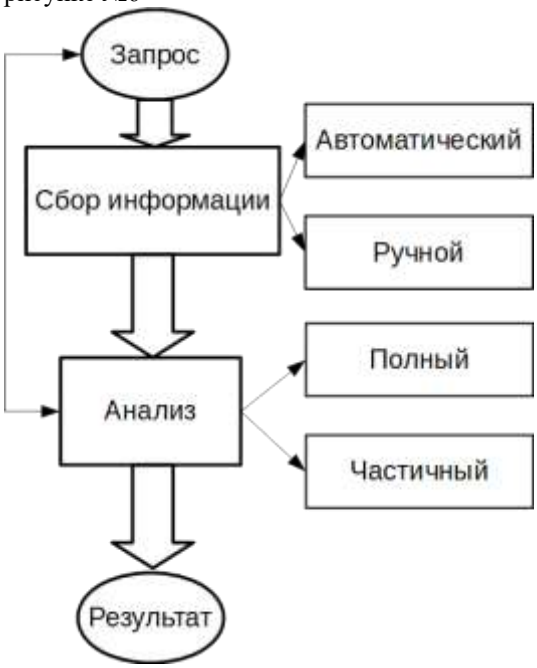

Рис. 6. - Простейший алгоритм действий по анализу информации

Подаётся запрос, затем осуществляется сбор и запись информации о файле в базу данных. Сбор может быть, как полностью автоматическим, то есть всё будет делать только сама система, так и ручным – информацию о файле будет вносить сам пользователь. Далее идёт анализ содержимого базы данных, как указано на рисунке №6, он может быть полным, то есть идёт обработка всей информации, и частичным, где для анализа доступна не вся информация. Такое может быть, если осуществляется анализ видеофайла либо изображения в плохом качестве, то есть внешняя информация и сведения о файле у нас есть, а вот содержимое файла мы получить не можем. Определение важности неструктурированных файлов осуществляется посредством использования комплекса моделей, методов и средств, например, таких как:

1. Искусственные нейронные сети
2. Машинное обучение
3. Большие данные
4. Интеллектуальный анализ данных

Совокупность всех этих средств позволяет осуществлять анализ документов с последующим выведением нужного нам файла [11,12,13].

6. Примеры аналогичных программ

Вообще, как таковой программы, специализирующейся на определении важности и критичности файла не существует, но существуют различные программы для обработки документов. Несколько примеров представлены в таблице №3:

Таблица 3 – Варианты программных продуктов

Наименования ПО	Функции	Разработчик
ABBYY Intelligent Search	Корпоративный поиск по всем источникам данных	ABBYY
word2vec	программа, позволяющая построить векторные представления слов на заданных массивах текстовой информации	Google
Apache OpenNLP	обработки текстов на основе методов машинного обучения	The Apache Software Foundation
WordSmith Tools	программный комплекс для исследования поведения слов в текстах	School of English, University of Liverpool

Большинство программ с наиболее эффективным интеллектуальным анализом данных являются платными и отличаются способами анализа и сбора информации [14]. Большинство программ такого типа делятся по отраслям производства и видам деятельности. А учитывая, что такие программы чаще всего предназначены для крупных компаний, следовательно, разработкой занимаются внутренние отделы предприятий, и в результате чего каждая компания работает в своей программе. Программы, представленные для общего пользования, являются малой частью тех программ, доступ к которым имеют внутри предприятий. Это является залогом безопасности информации, а также позволяет руководству самому корректировать, в какую сторону будет идти разработка.

7. Заключение

Как можно заметить, в мире активно разрабатываются технологии для осуществления работы информационной системы с документами. Возможность определения степени важности документа системой является достаточно важным шагом в области развития средств интеллектуального анализа файлов.

Список используемых источников

1. Белов, В.М. Теория информации. Курс лекций: Учебное пособие для вузов. / В.М. Белов, С.Н. Новиков, О.И. Солонская. - М.: РиС, 2016. - 143 с
2. Киселев, А.Г. Теория и практика массовой информации: общество-СМИ-власть: Учебник / А.Г. Киселев. - М.: ЮНИТИ, 2014. - 431 с.
3. Lyman P., Varian H.R. How much information. Release of the University of California. Oct.27, 2003.
4. Чернавский, Д.С. Синергетика и информация: Динамическая теория информации / Д.С. Чернавский; Предисл. и послесл. Г.Г. Малинецкий. - М.: ЛИБРОКОМ, 2013. - 304 с.
6. А. Л. Ноогенез и теория интеллекта. Краснодар: СовКуб, 2005. — 356 с.
7. Кулаичев, А.П. Методы и средства комплексного анализа данных: Учебное пособие / А.П. Еремин Горяинова, Е.Р. Прикладные методы анализа статистических данных: Учебное пособие / Е.Р. Горяинова, А.Р. Панков, Е.Н. Платонов. — М.: ИД ГУ ВШЭ, 2012. — 310 с
5. Малюк, А.А. Теория защиты информации. / А.А. Малюк. - М.: Горячая линия -Телеком, 2015. - 184 с.Y.-C. Wang, M. Joshi, W. W. Cohen, and C. P. Rosé, “Recovering Implicit Thread Structure in Newsgroup Style Conversations.,” in ICWSM, 2008.
8. Holzinger, Andreas. Combining HCI, Natural Language Processing, and Knowledge Discovery – Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field // Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data / Andreas Holzinger,Christof Stocker, Bernhard Ofner … [и др.]. — Springer, 2013. — P. 13–24.
9. Christopher Cook Diet for a Dead Planet: Big Business and the Coming Food Crisis; **Высшая школа** - Москва, **2010**. - 336 с.
10. Bill Schmarzo Big Data: Understanding How Data Powers Big Business; Wiley - M., 2013. - 240 с
11. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман. — М.: ДМК, 2016. — 498 с.
12. Сравнение документов посредством ScanDifFinder SDK: технологии от ABBYY. Оригинал статьи: *https://www.kp.ru/guide/obrabotka-dokumentov.html*