Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра автоматизации систем вычислительных комплексов

Колосов Алексей Михайлович

Информационно-ориентированные сети: система именования текстовых объектов

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:
д.ф.-м.н., чл.-корр. РАН, профессор
Р.Л.Смелянский

Москва, 2019

Аннотация

Работа посвящена теме информационно-ориентированных сетей(ИОС). В таких сетях доступ к информационному объекту(ИО) осуществляется по имени, определяемому системой именования, а не по адресу.

В работе в качестве ИО рассматриваются научные тексты. Вводятся требования к системе именования. Проводится обзор возможных свойств системы именования, определяется какие свойства в большей степени удовлетворяют требованиям системы именования: уникальность имени; возможность определения ИО по имени, известному частично; положительная корреляция разностей между именами и ИО.

В результате обзора сделан вывод о том, что предлагаемая система именования применяется для поиска по четко сформулированным запросам; для каждого ИО происходит построение машинного имени; для построения имени ИО доступна информация о других ИО; при наступлении коллизии имён в ходе изменения набора ИО построение имени происходит заново.

Для организации системы именования текстовых объектов, обладающей данными свойствами, предлагается использовать тематическое моделирование. В работе описана предлагаемая система именования и проведено экспериментальное исследование, в котором использован корпус выпускных и курсовых работ студентов Лаборатории Вычислительных Комплексов. Целью экспериментального исследования является оценка степени соответствия требованиям предлагаемой системы именования. Результаты экспериментов показывают, что тематическое моделирование может быть использовано для организации системы именования текстовых объектов, при условии достаточных размера выборки и качества тематической модели.

Также в работе рассмотрены влияние системы именования на построение ИОС и вопросы о существовании единой для всех типов ИО системе именования и едином имени, являющемся одновременно и машинным, и пользовательским.

Оглавление

Основные понятия

ИО — информационный объект

Тип ИО — идин из типов графический, аудио, видео и текстовый ИО

ИОС — информационно-ориентированная сеть

Задача ИОС — одна из задач Именование, Поиск, Передача, Хранение

Служба ИОС — одна из служб Публикации, Подписки, Анонсирования

Уровень функционирования ИОС — один из уровней Представления данных и Доступа к данным

Введение

В настоящее время объём данных, передаваемых по сети Интернет, постоянно увеличивается. В то же время пропускная способность каналов связи также растёт. Однако становится больше и конечных устройств. В результате всё более актуальны вопросы, что предпочтительнее: переносить вычисления к данным или перемещать данные к вычислениям; обрабатывать данные на стороне клиента или удалённо; передавать уже подготовленные к использованию данные или данные, которые ещё нужно преобразовывать.

Ответить на вопрос про перенос вычислений и перемещение данных позволяет кэширование, благодаря которому данные остаются там, где над ними проводятся вычисления; бессерверные вычисления и удалённый вызов процедур позволяют ответить на вопрос про сторону обработки данных, когда вызов функции производится на стороне клиента, но вычисления происходят удалённо; наконец, graphics streaming позволяет не передавать по сети видеопоток или проводить рендеринг полностью на клиенте: вычислительная часть также происходит удалённо, но часть, связанная с отображением графики, выполняется на стороне клиента, таким образом по сети передаются только инструкции, состоящие из функций графического API, и данные, содержащие предвычисленные аргументы.

Graphics streaming является примером использования кэширования, бессерверных вычислений и подготовленных к использованию данных. Отдельно стоит отметить, что использование в graphics streaming кэширования позволяет снизить объём передаваемых по сети данных приблизительно на 80%. [1]

Таким образом, перемещение данных к вычислениям является приоритетной задачей, тем более, что в некоторых случаях возможно предсказать распределение востребованности данных в сети и переместить данные к вычислениям до начала вычислений.

В случае перемещения данных к вычислениям данные больше не зависят от устройств, которые когда-либо их содержали, поскольку любые данные могут быть востребованы в любой точке сети и в любой момент времени могут содержаться на любом устройстве. Это означает, что в сети больше нет понятия адреса устройства, за которым данные закреплены на постоянной основе.

Подход, когда обращение к данным производится по их имени, не используя адрес устройства, их содержащего, называется информационно-ориентированным, а сети, построенные с его помощью — информационно-ориентированными сетями(ИОС). В работе предложено построение ИОС, состоящее из системы именования информационных объектов(ИО), системы поиска ИО по имени, системы передачи и хранения ИО.

Работа посвящена разработке системы именования текстовых объектов, поскольку система именования существенно влияет на построение ИОС, так как способ построения имени определяет функционирование ИОС — подробнее см. раздел 6. Важно отметить, что система именования различается у разных типов ИО — подробнее см. раздел 7. Таким образом, разработка системы именования текстовых объектов является актуальной.

Цель работы

Целью работы является разработка системы именования текстовых объектов, позволяющей по текстовому объекту определить его имя.

Система именования должна удовлетворять требованиям уникальности имени; возможности восстановления имени, известного частично; положительной корреляции разностей между именами и ИО. Выбор данных требований должен быть обоснован.

1 Построение информационно-ориентированных сетей

Построение ИОС проводится по модели «Поставщик-Потребитель»(«Publishing-Subscribing»)[2]. Можно предложить построение ИОС, состоящее из решения задач Именование, Поиск, Передача, Хранение:

1. В задаче Именование требуется сопоставить каждому ИО уникальное имя. Для этого необходимо определить систему именования — задать класс ИО, способ его отображения на множество имен, а также определить понятие имени ИО, по которому ИО идентифицируются в ИОС.

2. В задаче Поиск требуется представить механизм определения имени ИО по заданному имени, при этом имя может быть известно не полностью. Для этого необходимо определить систему разыменования.

3. В задаче Передача определяется способ доступа к запрашиваемому ИО. Для этого необходимо определить жизненный цикл запроса ИО, состоящий из формирования запроса, отправление запроса, получение запроса, реализация запроса.

4. В задаче Хранение определяется место ИО. Для этого необходимо определить жизненный цикл ИО, состоящий из первичного размещения ИО, перемещения ИО, обнаружения ИО, выдачи ИО по запросу.

Рисунок 1: Соответствие задач и уровней функционирования ИОС.

Данные задачи соответствуют двум логическим уровням функционирования ИОС (см. Рис. 1): уровню представления данных и уровню доступа к данным. Так, задачи Именование и Поиск относятся к уровню представления данных, а задачи Передача и Хранение — к уровню доступа к данным.

На уровне доступа к данным решаются задачи, связанные с передачей и хранением данных. В частности, на данном уровне решаются вопросы безопасности, множественной доставки ИО, а также перемещения ИО в места, где они наиболее востребованы, и кэширования ИО.

Информационные объекты «существуют» на уровне представления данных. На уровне доступа к данным они «погружаются» в контейнеры, носители данных, которые обладают своим набором атрибутов — атрибутами носителя, например, сведениями о дате создания; пользователе, добавившем объект в систему; правах доступа к объекту; размере объекта. Отдельно стоит отметить, что идентификатор Поставщика это один из атрибутов носителя, также как и идентификатор Потребителя, который используется для Подписки на ИО.

1.1 Функционирование ИОС

В общем случае ИОС функционирует следующим образом. В модели «Поставщик-Потребитель» пользователь ИОС может совершать два действия — Подписываться на ИО и Публиковать ИО. Тот, кто подписывается на ИО, называется Потребителем, а тот, кто публикует ИО — Поставщиком. Потребитель не может изменять содержимое ИО без разрешения Поставщика. Публикация контента производится через службу Публикации, подписка — через службу Подписки. Подписка бывает одноразовая и постоянная. В первом случае Потребитель получает ИО один раз, во втором — каждый раз, когда ИО обновляется Поставщиком. Подписка на ИО происходит по имени ИО, поэтому если содержимое ИО меняется таким образом, что это приводит к изменению имени ИО, все Потребители, у которых есть постоянная подписка на данный ИО, получают обновление имени ИО.

Для того, чтобы опубликовать ИО, Поставщик выбирает и передает службе Публикации файл-источник. Данная служба формирует имя ИО и загружает его в ИОС. После этого Поставщик получает имя ИО и затем может анонсировать полученное имя Потребителям через службу Анонсирования.

Для того, чтобы подписаться на ИО, необходимо в службе Подписки указать имя ИО. Одному имени соответствует один ИО. В ответ на запрос, служба Подписки возвращает ИО, имя которого соответствует заданному имени. С каждым ИО связан список допустимых действий, которые Потребитель может совершать с данным ИО: единоразовое получение ИО, подписка на обновления ИО, единоразовое получение ИО и подписка на обновления ИО.

1.2 Применение ИОС

Основная область применения ИОС — построение сетевых поисковых систем, в которых поиск осуществляется самой сетью, то есть для обнаружения данных не требуется использование дополнительных систем. В общем случае поиск может одновременно проводиться по ИО всех типов, чего не могут традиционные поисковые системы, в которых поисковая выдача индивидуальна для каждого типа ИО. Стоит отметить, что традиционная сеть это тоже своего рода поисковая система, которая по адресу находит устройство, а если его нет, то сообщает об этом.

Также ИОС могут быть использованы для организации рекомендательных систем; систем доставки контента(Content Delivery System, CDN); систем без обратной связи, например, стриминговых систем; систем с обратной связью, например, систем облачного гейминга.

2 Система именования ИОС

Система именования задаёт отображение множества информационных объектов на множество имен и операцию, реализующую это отображение.

2.1 Понятие информационного объекта

Каждый Информационный объект(ИО) представлен в виде данных, описывающих некоторый факт или явление. Данными могут быть текст, аудио, видео или графический объект. В общем случае любой ИО может быть представлен как бинарный объект. Каждый ИО содержит информацию — интерпретацию данных.

2.2 Понятие имени

Имя позволяет идентифицировать ИО. Чаще всего имя определяется содержанием ИО, однако может и не зависеть от него, например, являюсь порядковым номером ИО в системе. В случае когда имя определяется содержанием ИО, оно является признаковым описанием ИО. Задача имени — отразить содержание ИО.

2.3 Операция системы именования

Система именования представлена следующей операцией:
1. def put(IO) return name — по ИО определить его имя

Подробнее данная операция рассмотрена в разделе Предлагаемая система именования.

2.4 Операции системы разыменования

С системой именования связана система разыменования, которая необходима для решения задачи Поиск. Система разыменования задаёт отображение множества имён на множество информационных объектов и операции, реализующие это отображение.

Система разыменования представлена следующими операциями:
1. def get(name) return IO — по имени определить ИО
2. def find(partly_name) return {name} — по частично известному имени определить множество соответствующих имён

3. Обзор свойств системы именования

Обзор организован следующим образом. Определены и обоснованы требования к системе именования. Рассмотрены свойства, определяющие систему именования. Свойства сгруппированы в пары. В результате обзора определены те свойства, которые в большей степени соответствуют требованиям системы именования. Данными свойствами должна обладать предлагаемая система именования.

Рассмотрены следующие группы свойств системы именования:
1) Тип поиска: поиск по четко сформулированным запросам, разведочный поиск.
2) Тип имени: машинное, пользовательское.
3) Доступная для построения имени информация: локальная, глобальная.
4) Переименование при коллизии: наращивание имени, построение имени заново.

Также рассмотрено дополнительное свойство: сложность поиска. Однако поскольку данное свойство является количественным, а не качественным, для него не проводится обзор соответствия требованиям.

3.1 Требования системы именования

К системе именования предъявляются следующие требования:

1) Уникальность имени(далее Требование уникальности)

   Имя каждого ИО должно быть уникально в том смысле, что каждому имени соответствует ровно один ИО. При этом одному ИО может соответствовать несколько имен. Данное требование необходимо поскольку если нескольким ИО соответствуют одинаковые имена, эти объекты должны совпадать, в противном случае не будет возможности их различить без получения полностью.

2) Возможность восстановления имени, известного частично(далее Требование восстановления по части)

   В случае когда имя известно не полностью должна быть возможность определить множество имён ИО, соответствующих имени, известному частично. Данное требование необходимо поскольку чаще всего информационная потребность пользователя не является исчерпывающий. В этом заключается существенное отличие использования системы именования в решении задач Поиск и Хранение — в задаче Хранение доступ осуществляется только по полностью известным именам.

3) Положительная корреляция разностей между именами и ИО(далее Требование положительной корреляции)

   Данное требование означает, что из разности между именами должна следовать разность между ИО. Обратное утверждение следует из данного, поскольку если ИО отличаются незначительно, имена не могут отличаться значительно. Разность определяется по-разному для машинного и пользовательского имён. Кроме того разность вычисляется не между именами и ИО, а между их моделями. Имя является моделью по построению.

Стоит отметить, что данные требования могут быть применены к системе именования ИО любого типа, не только текстовых объектов. Различаются только модели ИО.

Далее рассмотрены группы свойств системы именования и для входящих в них свойств определена степень соответствия требованиям. Для каждого свойства рассматриваются все требования. Для свойства Доступная для построения имени информация рассматривается дополнительное требование: независимость сложности построения имени от количества ИО. Данное требование является дополнительным, поскольку необходимость его выполнения существенно сужает пространство возможных систем именования. Кроме того, данное требование является качественным, в отличие от остальных требований, которые являются количественными и могут быть использованы как мера качества предлагаемого решения. Выполняемость данного требования также рассматривается в главе Предлагаемое решение.

3.2 Тип поиска

Система именования может быть использована для поиска по четко сформулированному запросу(known-item search), когда запросом является имя или его часть и разведочному поиску(exploratory search), когда запросом является ИО или его часть. [3]

3.2.1 Поиск по четко сформулированным запросам

Примерами поиска по четко сформулированному запросу является поиск по ключевым словам и полнотекстовый поиск. Результатом поиска является точное соответствие.

Соответствие требованиям:

1) Требование уникальности — выполнимо

   Для данного свойства возможно выполнить требованием уникальности, поскольку для точного соответствия может быть организовано однозначное отображения множества имён на множество ИО.

2) Требование восстановления по части — выполнимо

   Для данного свойства возможно выполнить требование восстановления по части, поскольку могут быть использованы шаблоны, в которых Пользователь указывает известные ему атрибуты, а оставшиеся поля оставляет пустыми.

3) Требование положительной корреляции — выполнимо частично

   Для данного свойства возможно лишь частично выполнить требование положительной корреляции, поскольку понятие разности может быть не определено для имен, например, в случае использования категориальных атрибутов.

3.2.2 Разведочный поиск

Пользователь может обладать информационной потребностью в систематизации некоторой области знаний. В этом случае используется разведочный поиск, позволяющий производить изучение новых предметных областей. [4]

В данном типе поиска запросом может быть текст произвольной длины или даже подборка текстов. Результатом поиска может являться «карта» предметной области, позволяющая эффективно организовать обучение.

Соответствие требованиям:

1) Требование уникальности — не выполнимо

   Для данного свойства нет возможности выполнить требованием уникальности, поскольку в данном типе поиска нет понятия имени в явном виде.

2) Требование восстановления по части — выполнимо частично

   Для данного свойства возможно лишь частично выполнить требование восстановления по части, поскольку из-за отсутствия в данном типе поиска понятия имени в явном виде возможен только вариант восстановления ИО по его части.

3) Требование положительной корреляции — выполнимо

   Для данного свойства возможно выполнить требование положительной корреляции, поскольку неявные имена могут удовлетворять данному требованию, так как для них может быть задана необходимая метрика, определяющая понятие разности.

Вывод

В Таблице 1 приведены результаты сравнительного анализа соответствия требованиям свойств из группы Тип поиска.

Из Таблицы 1 делается вывод о том, что Поиск по четко сформулированным запросам в большей степени удовлетворяет требованиям системы именования.

	Поиск по четко сформулированным запросам	Разведочный поиск
Уникальность имени	+	-
Восстановление по части	+	-+
Положительная корреляция	-+	+

Таблица 1. Сравнительный анализа соответствия требованиям свойств группы Тип поиска.

3.3 Тип имени

Имя может быть машинным в том смысле, что для человека оно неинтерпретируемо, но позволяет вычислительно эффективно проводить операцию именования и поиска по имени, возможно известном частично. Также имя может быть пользовательским в том смысле, что пользователь по имени может предположить содержание ИО, то есть имя является интерпретируемым для него.

Имя в явном виде из предыдущего раздела это пользовательское имя. Неявное имя — машинное.

Изначально вопросом именования является вопрос о том, может ли существовать «единое» имя, являющееся одновременно и машинным, и пользовательским. Из истории развития вычислительной техники таких примеров неизвестно. Поэтому скорее всего такого имени существовать не может. Таким образом, предполагается необходимость построения двух имен.

3.3.1 Машинное имя

Примером машинного имени может быть вектор чисел. Машинное имя в некотором смысле интерпретируемо для систем, когда может рассматриваться как координаты.

Соответствие требованиям:

1) Требование уникальности — выполнимо

   Для данного свойства возможно выполнить требованием уникальности, поскольку, вектора действительных чисел могут быть попарноразличны.

2) Требование восстановления по части — выполнимо частично

   Для данного свойства возможно лишь частично выполнить требование восстановления по части, поскольку понятие части может быть не определено, например, в случае когда имя это одно число.

3) Требование положительной корреляции — выполнимо

   Для данного свойства возможно выполнить требование положительной корреляции, поскольку для машинных имён могут быть заданы метрики, удовлетворяющие данному требованию, например, для векторов это может быть косинусная метрика.

3.3.2 Пользовательское имя

Примером пользовательского имени может быть аннотация ИО ключевыми словами. По данному имени пользователь имеет возможность предположить содержание ИО.

Соответствие требованиям:

1) Требование уникальности — не выполнимо

   Для данного свойства нет возможности выполнить требованием уникальности, поскольку у близких по смыслу объектов могут полностью совпадать части, по которым происходит построение имени.

2) Требование восстановления по части — выполнимо

   Для данного свойства возможно выполнить требование восстановления по части, поскольку Пользователь может не указать ряд значений, однако указанные значения образуют множество, для которого возможно определить надмножества, в которое указанное множество входит целиком.

3) Требование положительной корреляции — выполнимо

   Для данного свойства возможно выполнить требование положительной корреляции, поскольку в случае, когда имя сохраняет информацию об исходном объекте, похожим с точки зрения пользователя именам соответствуют похожие с точки зрения пользователя объекты.

Вывод

В Таблице 2 приведены результаты сравнительного анализа соответствия требованиям свойств группы Тип имени.

Из Таблицы 2 делается вывод о том, что Машинное имя в большей степени удовлетворяет требованиям системы именования.

	Машинное имя	Пользовательское имя
Уникальность имени	+	-
Восстановление по части	-+	+
Положительная корреляция	+	+

Таблица 2. Сравнительный анализа соответствия требованиям свойств группы Тип имени.

3.4 Доступная для построения имени информация

Для построения имени может быть доступна информация только о том ИО, для которого производится построение имени, локальная информация. Также может быть доступна информация о других, в общем случае всех ИО, включая их имена — глобальная информация.

Кроме того для данного свойства рассматривается дополнительное требование: независимость сложности построения имени от количества ИО.

3.4.1 Локальная информация

Примером локальной информации может быть множество слов, встречающихся в тексте, который именуется, и их частотности.

Соответствие требованиям:

1) Требование уникальности — не выполнимо

   Для данного свойства нет возможности выполнить требование уникальности, поскольку в случае использования только локальной информации невозможно гарантировать отсутствие коллизий, поскольку количество ИО не ограничено.

2) Требование восстановления по части — выполнимо

   Для данного свойства возможно выполнить требование восстановления по части, поскольку локальная информация индивидуальна и может быть использована для идентификации объекта в случае, если доступна только часть сведений об объекте.

3) Требование положительной корреляции — не выполнимо

   Для данного свойства не возможности выполнить требование положительной корреляции, поскольку наличие только функции расстояния не позволит вычислительно эффективно находить ближайший к заданному объект — для этого потребуется полный перебор. Для эффективного решения этой задачи объекты должны обладать координатами, для определения которых необходима глобальная информация.

4) Требование независимости от количества ИО — выполнимо

   Для данного свойства возможно выполнить требование независимости от количества ИО, поскольку локальная информация не зависит от количества ИО.

3.4.2 Глобальная информация

Примером глобальной информации может быть множество слов, встречающихся в коллекции текстов и их частотности.

Соответствие требованиям:

1) Требование уникальности — выполнимо

   Для данного свойства возможно выполнить требованием уникальности, поскольку коллизии могут разрешаться на этапе построения. При это возникает вопрос, о том какой из объектов переименовывать при коллизии. Данный вопрос рассматривается в следующей секции.

2) Требование восстановления по части — выполнимо частично

Для данного свойства возможно лишь частично выполнить требование восстановления по части, поскольку для частично известной информации может существовать множество вариантов восстановление, поэтому определение всех таких вариантов является вычислительно трудоёмкой задачей.

3) Требование положительной корреляции — выполнимо

   Для данного свойства возможно выполнить требование положительной корреляции, поскольку для глобальной информации возможно задание координат объектов, которое позволяет выполнить данное требование.

4) Требование независимости от количества ИО — выполнимо частично

   Для данного свойства возможно лишь частично выполнить требование независимости от количества ИО, поскольку для выполнения данного требования необходимо ограничить количество информации, которое используется для построения имен новых объектов и не зависит от их количества.

Вывод

В Таблице 3 приведены результаты сравнительного анализа соответствия требованиям свойств группы Доступная для построения имени информация.

Из Таблицы 3 делается вывод о том, что Глобальная информация в большей степени удовлетворяет требованиям системы именования.

	Локальная информация	Глобальная информация
Уникальность имени	-	+
Восстановление по части	+	+-
Положительная корреляция	-	+

Таблица 3. Сравнительный анализа соответствия требованиям свойств группы Доступная для построения имени информация.

3.5 Переименование при коллизии

В случае изменения набора ИО, имя нового ИО может совпадать с именем ИО, уже содержащегося в наборе, в этом случае происходит коллизия и необходимо определить, какой ИО переименовать для того, что выполнять требование уникальности. Первым вариантом является расширение имени выбранного ИО, его наращивание, с тем, чтобы имя вновь стало уникальным. Другим вариантом является построение имени выбранного ИО заново.

Стоит отметить, что если при расширении множества ИО, коллизий не возникает, количество объектов может быть произвольным. Поэтому важен именно вопрос о переименовании объектов в случае коллизии. Также важен вопрос о том, какой из объектов переименовывать.

3.5.1 Наращивание имени

Наращивание имени может быть произведено разными способами, как за счёт удлинения, так и за счёт дополнительных итераций в процессе построения имени.

Соответствие требованиям:

1) Требование уникальности — выполнимо

   Для данного свойства возможно выполнить требованием уникальности, поскольку в случае отсутствия ограничений на наращивание, имя может наращиваться до тех пор, пока не станет уникальным.

2) Требование восстановления по части — выполнимо

   Для данного свойства возможно выполнить требование восстановления по части, поскольку имея инкрементальную структуру имя может быть представлено в виде уровней, то есть не быть монолитным. В этом случае для него определимо понятие части.

3) Требование положительной корреляции — выполнимо частично

   Для данного свойства возможно лишь частично выполнить требование положительной корреляции, поскольку при наращивании имен у объектов могут начать различаться размерности, что будет значить их несравнимость в смысле разности.

3.5.2 Построение имени заново

Соответствие требованиям:

1) Требование уникальности — выполнимо

   Для данного свойства возможно выполнить требованием уникальности, поскольку, поскольку новое имя может отличаться от всех остальных имён.

2) Требование восстановления по части — выполнимо частично

   Для данного свойства возможно лишь частично выполнить требование восстановления по части, поскольку имя может быть монолитным, например, в случае, когда учитывает связи внутри объекта, поэтому для такого имени понятие части не определено.

3) Требование положительной корреляции — выполнимо

   Для данного свойства возможно выполнить требование положительной корреляции, поскольку имена могут строиться таким образом, чтобы обладать одинаковой размерностью и являться координатами.

Вывод

В Таблице 4 приведены результаты сравнительного анализа соответствия требованиям свойств группы Переименование при коллизии.

Из Таблицы 4 делается вывод о том, что Построение имени заново в большей степени удовлетворяет требованиям системы именования.

	Дополнение имени	Построение имени заново
Уникальность имени	+	+
Восстановление по части	+	+-
Положительная корреляция	-+	+

Таблица 4. Сравнительный анализа соответствия требованиям свойства группы Переименование при коллизии.

3.6 Сложность поиска

Данное свойство является дополнительным. Сложность поиска является количественным свойством и зависит от того, на каком пространстве реализуется поисковый запрос. Поисковый запрос может реализовываться как на пространстве имен, так и на пространстве ИО. Сложность поиска определяется тем как устроена связь между пространством имен и пространством ИО.

Значение данного свойства рассматривается в главе Предлагаемая система именования.

3.7 Выводы

В результате обзора делается вывод о том, что следующие свойства систем именования в большей степени удовлетворяют требования системы именования:
1) Тип поиска: поиск по четко сформулированным запросам.
2) Тип имени: машинное.
3) Доступная для построения имени информация: глобальная.
4) Переименование при коллизии: построение имени заново.

Предлагаемая система именования должна обладать данными свойствами.

4. Предлагаемая система именования

В работе решается задача о существовании системы именования, обладающей свойствами, определёнными в обзоре. Поскольку такая система именования существует, проведение обзора систем именования не требуется. В качестве меры качества предлагаемой системы именования рассматривается степень соответствия требованиям, определяемая с помощью экспериментов.

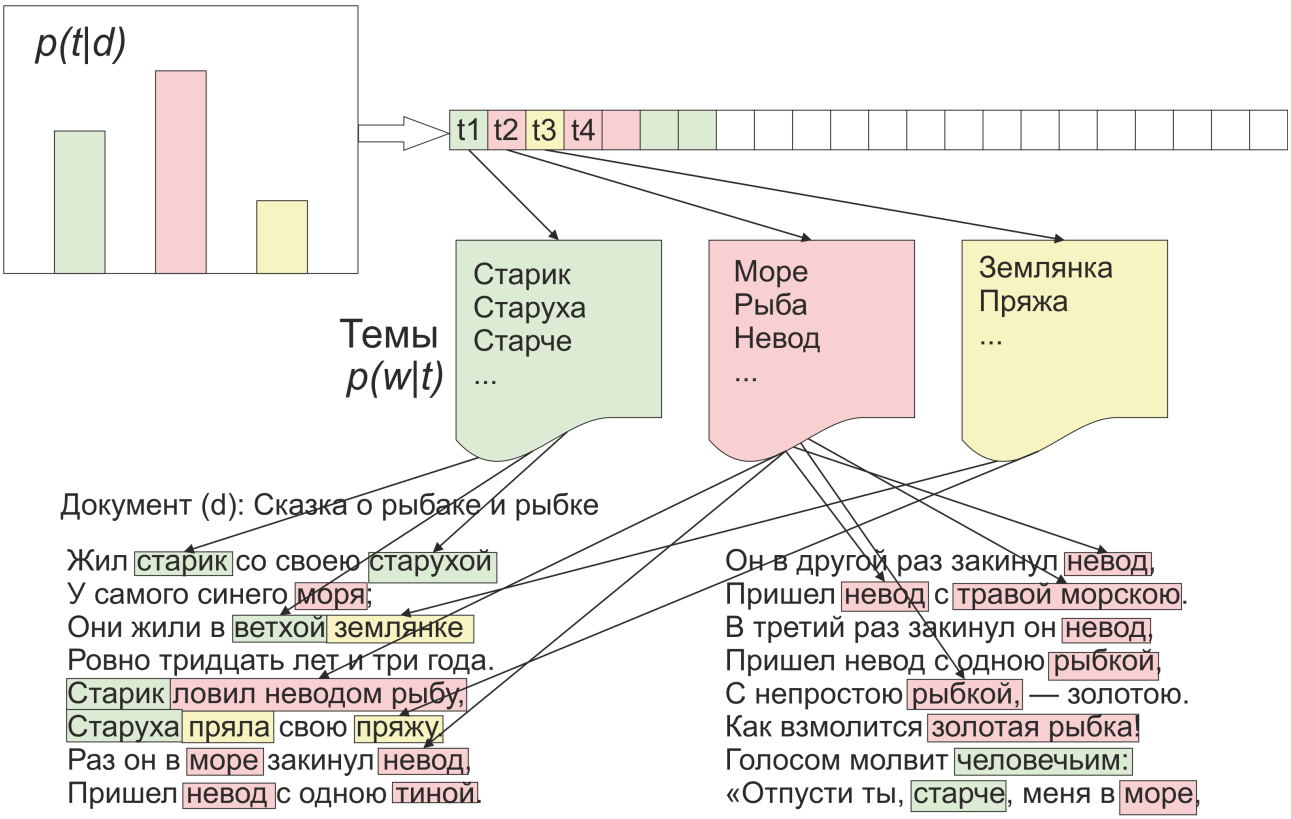

Рисунок 2. Тематическая модель[5].

Для организации системы именования предлагается использовать вероятностное тематическое моделирование[6]. Этот подход позволяет сопоставить документу вектор вероятностей соответствия документа множеству тем(см. Рис. 2). Темы представляют собой распределения характеризующих их слов. Формально решается задача матричного разложения. В исходной матрице в строках находятся слова из словаря коллекции, в столбцах номера документов. Данная матрица раскладывается на две: Фи и Тетта. В матрице Фи в строках находятся слова из словаря коллекции, в столбцах темы, в ячейке вероятность принадлежности слова к соответствующей теме(см. Таблица 5). В матрице Тетта в строках находятся темы, в столбцах номера документов, в ячейке вероятность принадлежности темы к соответствующему документу(см. Таблица 6).

Стоит отметить, что темы являются скрытыми переменными, то есть не могут быть предъявлены в явном виде. Однако их можно достаточно точно определять по семантическим ядрам, т.е. множествам слов, имеющих высокую вероятность принадлежности к данной теме и низкую ко всем остальным.

Также стоит отметить, что тематическая модель осуществляет так называемую «мягкую» кластеризацию, когда один документ может относиться сразу к нескольким темам с соответствующими вероятностями[6]. Но также возможен и случай «жесткой» кластеризации, когда документ относится с вероятностью единица к одной теме и ноль ко всем остальным.

	topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8
(@default_class, тестирование_эффект)	0.000023	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
(@default_class, ошибка_пользователь)	0.000000	0.000000	0.000000	0.000024	0.000000	0.000019	0.000000	0.000000	0.000000
(@default_class, цвет_нахождение)	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.447888e-05	0.000000

Таблица 5. Пример матрицы Фи.

	200	201	202	203	204
topic_0	0.000000	0.0	0.0	0.0	0.0
topic_1	0.000000	0.0	0.0	0.0	0.0
topic_2	0.000000	1.0	0.0	1.0	0.0
topic_3	0.000000	0.0	0.0	0.0	0.0
topic_4	0.000000	0.0	0.0	0.0	1.0
topic_5	0.000000	0.0	0.0	0.0	0.0
topic_6	0.042908	0.0	0.0	0.0	0.0
topic_7	0.000000	0.0	0.0	0.0	0.0
topic_8	0.957092	0.0	1.0	0.0	0.0

Таблица 6. Пример матрицы Тетта.

4.1 Модель информационного объекта

Для представления текста используется модель мешка слов. Данная модель не учитывает порядок слов в тексте, но учитывает количество вхождений слова в текст. [6]

4.2 Имя информационного объекта

Именем документа является вектор вероятностей принадлежности документа к темам. Число тем определяется заранее. Длина вектора равна количеству тем. Каждая компонента вектора

содержит вероятность принадлежности документа к соответствующей данной компоненте теме.

4.3 Соответствие требованиям

В данном разделе описывается возможность соответствия предлагаемой системы именования требованиям системы именования. Оценка степени соответствия данным требованиям проводится проводится в разделе Экспериментальное исследование.

Требование уникальности

Имя является вектором действительных чисел. Поэтому при проведении мягкой кластеризации и достаточно высокой размерности вектора, то есть значительном количестве тем, коллизий быть не должно. Однако в случае жесткой кластеризации и не значительном количестве тем коллизии возможны. Влияние жесткой кластеризации на количество коллизий в зависимости от количества тем рассматривается в экспериментальном исследовании. В качестве модельного примера можно привести случай, когда все документы подвергнуты жесткой кластеризации, документов N, размерность пространства < log_2(N). В этом случае коллизии неизбежны.

Таким образом, степень соответствия предлагаемой системы именования данному требованию зависит от параметров тематической модели.

Требование восстановления по частично известному имени

В случае тематического моделирования Пользователю может быть известна часть тем имени объекта, либо известна часть объекта, если темы неизвестны.

В первом случае Пользователь ставит единицу в позициях имени, соответствующих известным ему темам. Соответствие позиций с единицами и тем определяется вручную, по значимым словам тем. После этого в матрице Тетта по полученному шаблону производится поиск всех соответствующих данному шаблону имён.

Во втором случае для документа строится имя и затем по функции расстояния определяется ближайшее имя в матрице Тетта. В качестве функции расстояния в тематическом моделировании принято использовать косинусную метрику[4]. Формально это величина схожести двух объектов, однако по формуле 1 - cos(d1, d2) может быть получено расстояние между объектами.

Важным является тот факт, что выполняется дополнительное требование о том, что сложность построения имени для новых объектов не зависит от количества объектов. Сложность построения нового имени определяется размерами матриц Фи и Тетта, которые фиксированы и не зависят от количества документов после обучения модели. Выполнение дополнительного свойства важно для случая, когда известна часть документа, поскольку и когда документ, часть которого известна, присутствует в коллекции и когда отсутствует, необходимо производить построение имени, так как частично известный объект является новым.

Стоит отметить, что для случая когда известна часть документа, обнаружение ближайшего по имени документа приводит к обнаружению близкого по смыслу документа только когда выполнено требование положительной корреляции разностей. Для случая, когда задаётся шаблон имени, данное требования также важно, поскольку если оно выполнено, то документам будут присвоены соответствующие их содержанию темы.

Требование положительной корреляции разностей

В случае если выделены семантические ядра тем, данное требование выполнено по построению.

Таким образом, степень соответствия предлагаемой системы именования данному требованию зависит от качества тематической модели.

4.4 Свойства

В данном разделе определяются свойства предложенной системы именования и их соответствие свойствам, определённым в обзоре.

Тип поиска

С помощью предложенной системы именования возможно осуществление поиска по четко сформулированным запросам. В этом случае Пользователю известна часть тем имени объекта. Все объекта, имена которых соответствуют полученному шаблону, являются результатами поискового запроса. Результаты поиска точно соответствуют поисковому запросу.

Также предложенная система именования может быть использована и для разведочного поиска. В этом случае Пользователю известна часть объекта. Все, объекты, имена которых похожи в заданном смысле на имя частично известного объекта, являются результатами поискового запроса. Поисковая выдача является рекомендациями.

Таким образом предложенная система именования может использоваться как для поиска по чётко сформулированным запросам, так и для разведочного поиска.

Доступная для построения имени информация

Для построения имени используется глобальная, но ограниченная информация — используется информация о всей обучающей выборки. Для новых объектов информация о других новых объектах не учитывается.

С другой стороны имя строится на основе локальной информации об объекте.

Таким образом, предложенная система именования использует для построения имени как глобальную, так и локальную информация.

Тип имени

Предложенная система именования производит автоматическое построение машинного имени — вектора вероятностей принадлежности документа к темам.

Пользовательское имя может быть получено путем ручного аннотирования тем. В случае появления решения, автоматизирующего решение данной задачи, будет производиться единовременное построение обоих имен.

Таким образом, предложенная система именования позволяет производить последовательное построение машинного и пользовательского имени. При это машинное имя строится автоматически, а пользовательское вручную.

Переименование при коллизии

В случае коллизии на обучающей выборке возможно как проведение обучение модели заново, так и проведение дообучения. Однако после обучения модели, для новых объектов, в случая появления коллизии, возможен только вариант обучения модели заново. Строго говоря, дообучить модель возможно, но в этом случае потребуется перестраивать имена всех ранее обработанных новых объектов.

При коллизии производится переименование не одного объекта, а всех. Поэтому вопроса о том, какой из объектов, имена которых совпадают, переименовывать, не возникает.

Таким образом, предложенная система именования позволяет производить заново построение имен объектов. Также возможно и проведение наращивания имени, однако не за счёт удлинения, а за счёт дополнительных итераций в обучении модели.

Дополнительное свойство: сложность поиска

Отдельно стоит отметить, что в предложенной системе именования длина имени определяется количеством тем, которое заметно меньше чем количество ключевых слов в документе и гораздо меньше чем размер словаря документа.

Важно отметить, что сложность поиска определяется не количеством позиций в имени, а размером словаря. Поэтому проведение поиска в пространстве тем требуется существенно меньше операций по сравнению с поиском в пространстве имён.

Таким образом, предложенная система именования позволяет эффективно организовать поисковые операции.

4.5 Реализация

Для реализации предложенной системы именования используется библиотека тематического моделирования bigartm[7]. Программная реализация, включающая программы для подготовки данных, обучения модели и проведения экспериментов, находится в открытом доступе[8].

5 Экспериментальное исследование

5.1 Цель и методика

Целью исследования является получение экспериментального подтверждения соответствия требованиям предлагаемой системы именования.

Для проведения исследования используется следующая методика. Проводится выделение интерпретируемых тем из коллекции документов. Затем для полученных векторов принадлежности документов к темам проводится серия экспериментов для определения уникальности имён, возможности восстановления имени по своей части, а также расчёт минимального и среднего расстояния между документами по косинусной метрике для определения соблюдения требования положительной корреляции.

5.2 Описание экспериментальных данных

В качестве экспериментальных данных используется текстовая коллекция архива лаборатории ЛВК кафедры АСВК за период с 2004 по 2016 годы — курсовые и дипломные работы, рецензии и отзывы

Характеристики коллекции:
1) всего 2074 файлов, из них .doc, .docx, .pdf 626, из них в мешок слов попали 559 документов — остальные не удалось преобразовать
2) в самом мешке 499470 ненулевых счётчиков
3) в словаре 120969 слов

Коллекция представлена в виде мешка слов в формате vowpal wabbit[9]. Мешок слов это структура данных, не учитывающая последовательность слов, но учитывающая их частотность. В этой модели объекты считаются совпадающими, если совпадают их мешки слов. Таким образом, тексты, в которых выполнена перестановка слов, имеют совпадающие мешки слов и считаются одинаковыми.

Мешок слов коллекции получен путем объединения мешков слов документов. Для каждого документа коллекции проведены следующие действия:
1) определение кодировки документа и представление его в виде последовательности символов
2) выделены токены — последовательности символов между пробельными символами
3) среди токенов выделены слова — последовательности букв
4) оставлены только слова, не входящие в список стоп-слов
5) проведена лемматизация — то есть приведение слов к начальной форме
6) получены юниграмы и биграмы лемм с учетом их частотности в документе
7) убраны юниграмы и биграмы с частотностью единица

Последние два действия выполнялись, поскольку были проведены следующие построения мешков слов документов:
1) юниграмы без счётчиков, равных единице
2) юниграмы со счётчиками, равными единице
3) юниграмы и биграмы со счётчиками, равными единице

4) юниграмы и биграмы без счётчиков, равных единице

В итоге было выбрано последнее построение, однако из словаря были убраны юниграмы — таким образом мешок слов эквивалентен мешку слов с биграмами без счётчиков, равных единице. Данное построение выбрано, поскольку на нём были получены наиболее интерпретируемые результаты.

Полученный мешок слов для одного из документов:
|text рецензент:2 дипломный:7 работа:8 кафедра:2 вычислительный:2 факультет:2 математика:2 кибернетика:2 мгу:2 имя:2 ломоносов:2 алтухов:5 исследование:4 метод:8 сеть:10 основной:3 цель:3 это:2 разработка:2 средство:3 оценка:4 задержка:15 пакет:6 пкс:5 коммутатор:5 уровень:2 достаточно:4 высокий:3 точность:4 нагрузка:3 сбор:2 информация:4 первый:3 еще:3 задача:4 связь:2 анализ:7 практически:2 год:2 openflow:3 являться:4 построение:2 использование:4 многий:2 новый:4 получать:4 возможность:7 один:3 сторона:2 например:2 пересылка:2 контроллер:4 правило:4 алгоритм:2 показывать:3 раздел:7 полностью:2 предлагать:3 число:4 результат:3 реализовать:2 позволять:2 значение:3 особенность:2 измерение:3 интервал:2 виток:3 цикл:6 проходить:2 последний:2 отмечать:3 использовать:2 счетчик:2 аппарат:2 n:2 использоваться:3 замечание:3 ряд:2 термин:2 пояснение:2 этот:2 достоинство:2 измерять:2 заданный:2 математический:2 дипломный_работа:7 мгу_имя:2 имя_ломоносов:2 основной_цель:2 цель_дипломный:2 разработка_метод:2 метод_средство:2 средство_оценка:2 оценка_задержка:2 нагрузка_сеть:2 сбор_информация:2 сеть_еще:2 задача_анализ:2 анализ_задержка:4 задержка_один:2 задержка_пкс:2 коммутатор_правило:2 значение_задержка:2 предлагать_метод:2 число_виток:2 виток_цикл:3 цикл_проходить:2 проходить_пакет:2 работа_алтухов:2 измерять_цикл:2

Далее на полученном мешке слов коллекции проводилось тематическое моделирование. Количество тем выбрано равным девяти, поскольку в результате моделирования для данного числа тем были получены наиболее разделимые по семантическим ядрам темы.

Поскольку выбран подход к тематическому моделированию, использующий аддитивную регуляризацию, для улучшения интерпретируемости получаемых результатов использовались регуляризаторы. Были использованы следующие регуляризаторы:
1) регуляризатор декорреляции матрицы Фи
2) регуляризатор сглаживания/разреживания матрицы Фи
3) регуляризатор сглаживания/разреживания матрицы Тетта

В итоге наилучшие результаты были получены после применения регуляризатора разреживания матрицы Тетта со значением Тау=-225.

Итоговые показатели разреженности матриц:
1) матрица Фи — 0.84
2) матрица Тетта — 0.88

Стоит отметить, что итоговые распределения слов по темам были получены полуавтоматически, поскольку после обучения и вывода слов с максимальной вероятностью принадлежности к данной теме, из словаря вручную убирались слова, которые не могли принадлежать семантическому ядру темы, что давало косвенную обратную связь на модель

после обучения модели заново. Семантическое ядро темы это слова, которые значимы для данной темы и не значимы для всех остальных тем. Формально это слова, вероятности которых в данной теме превышают заданное верхнее пороговое значение, а в других темах не превышают заданное нижнее пороговое значение.

В результате тематического моделирования были получены следующие распределения слов по темам(см. Таблица 7) — соответствие тем и направлений лаборатории определяется вручную.

В Лаборатории Вычислительных Комплексов изучаются следующие направления:
1) Системы программирования
2) Генетические алгоритмы
3) Безопасность
4) Программно-конфигурируемые сети
5) Системы реального времени
6) Планирование вычислений

Отдельно стоит отметить, что если получить интерпретируемые темы не удаётся, проведение дальнейших экспериментов не имеет смысла, поскольку в этом случае имена не характеризуют объекты и нет возможности проверить соответствие системы именования требованию положительной корреляции.

topic_0 среда_выполнение - 0.005 такой_образ - 0.004 логический_процесс - 0.004 модельный_время - 0.003 база_данные - 0.002 блок_обработка - 0.002 html_http - 0.002 представлять_себя - 0.002 домашний_страница - 0.002 описание_модель - 0.002	**topic_3** время_выполнение - 0.006 порядок_устройство - 0.004 среда_разработка - 0.003 f_f - 0.003 выполнение_программа - 0.003 реальный_время - 0.003 устройство_кольцо - 0.002 кольцо_арбитраж - 0.002 среда_диана - 0.002 такой_образ - 0.002	**topic_6** система_аксиома - 0.011 нештатный_поведение - 0.01 генетический_алгоритм - 0.006 эталонный_траектория - 0.006 временный_ряд - 0.005 элементарный_условие - 0.004 целевой_функция - 0.003 участок_нештатный - 0.003 разметка_эталонный - 0.003 алгоритм_построение - 0.003
topic_1 обучать_выборка - 0.004 обнаружение_атака - 0.003 передача_данные - 0.003 w_w - 0.003 такой_образ - 0.002 передача_сообщение - 0.002 директивный_интервал - 0.002 граф_зависимость - 0.002 данный_работа - 0.002 условный_оператор - 0.002	**topic_4** построение_расписание - 0.01 нейронный_сеть - 0.006 рабочий_интервал - 0.006 целевой_функция - 0.006 решение_задача - 0.005 работа_алгоритм - 0.005 задача_построение - 0.004 имитация_отжиг - 0.004 такой_образ - 0.003 директивный_интервал - 0.003	**topic_7** базовый_блок - 0.003 такой_образ - 0.003 уязвимость_уязвимость - 0.003 поток_управление - 0.003 время_выполнение - 0.002 поток_данные - 0.002 входной_данные - 0.002 данный_работа - 0.002 совместимость_требование - 0.002 виртуальный_регистр - 0.002
topic_2 библиографический_ссылка - 0.005 такой_образ - 0.003 извлечение_метаинформация - 0.003 научный_статья - 0.003 база_данные - 0.003 виртуальный_сеть - 0.002 экспериментальный_исследование - 0.001 name_val - 0.001 представлять_себя - 0.001 исходный_файл - 0.001	**topic_5** вычислительный_система - 0.008 программный_компонент - 0.005 механизм_отказоустойчивость - 0.003 генетический_алгоритм - 0.003 целевой_функция - 0.003 аппаратный_компонент - 0.003 внесение_неисправность - 0.003 решение_задача - 0.003 рвс_рв - 0.003 данный_работа - 0.003	**topic_8** инструментальный_окно - 0.005 обнаружение_атака - 0.005 нормальный_поведение - 0.004 система_обнаружение - 0.004 такой_образ - 0.004 html_http - 0.003 данный_работа - 0.002 метод_обнаружение - 0.002 текстовый_редактор - 0.002 a_b - 0.002

Таблица 7. Распределения слов по темам.

5.3 Описание экспериментов

Эксперимент 1

Данный эксперимент посвящен определению уникальности имён. Для данного эксперимента проводится расчёт числа жёстко кластеризуемых векторов и коллизий в зависимости от количества тем при разных типах моделях(юниграмная и биграмная, биграмная, юниграмная) без счётчиков, равных единице.

Учитывая жесткую кластеризацию, предположительно вызываемую биграммной моделью, ожидается, что доля уникальных имён будет расти с увеличением числа тем.

Так для 559 объектов и 9 тем при условии жесткой кластеризации, коллизии быть должны, однако уже при 10 темах коллизий быть не должно.

Предполагается, что на малых размерностях коллизии связаны с жесткой кластеризацией.

Результаты эксперимента 1

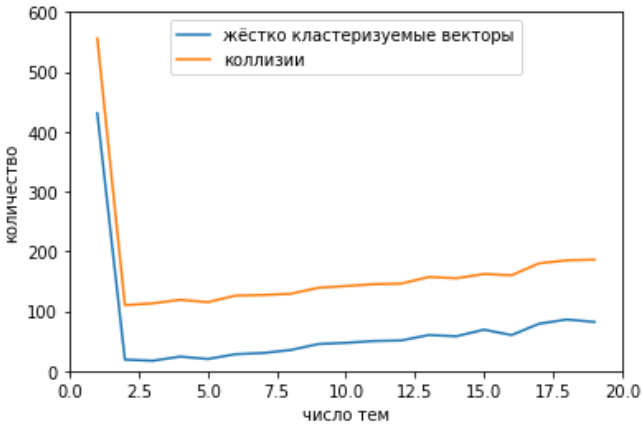

Рисунок 3. Юниграмная и биграмная модель.

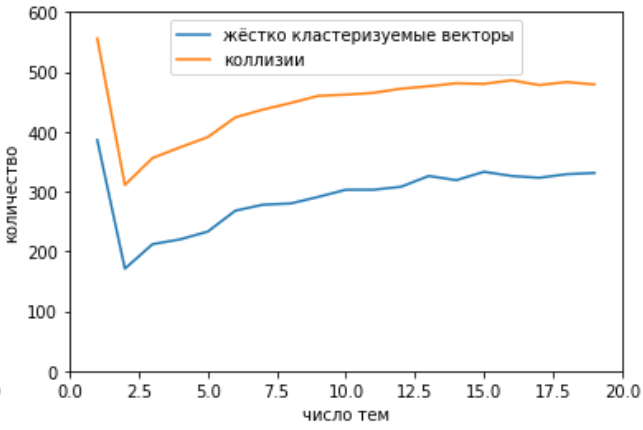

Рисунок 4. Биграмная модель.

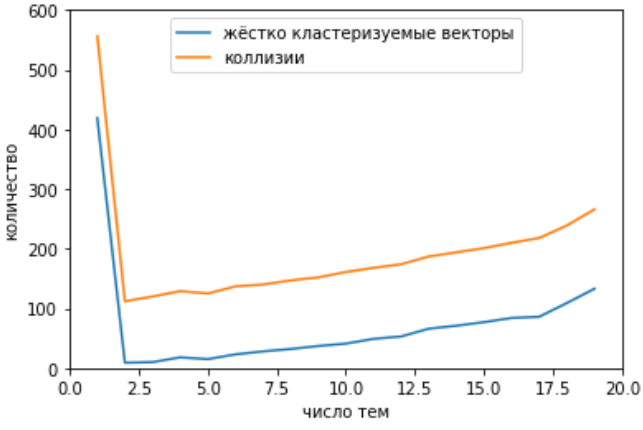

Рисунок 5. Юниграмная модель.

Получена следующая зависимость числа коллизий от количества жёстко кластиризируемых объектов(см. Рис. 3, 4, 5). Результаты эксперимента подтверждают, что число коллизий связано с количеством жёстко кластеризируемых объектов. Зависимость прямая.

Эксперимент 2

Данный эксперимент посвящен определению способности системы именования к восстановлению имени по своей части.

В случае тематического моделирования Пользователю может быть известна часть тем имени объекта, либо известна часть объекта, если темы неизвестны.

Решено не проводить данный эксперимент в связи со значительным числом коллизий, определённым в Эксперименте 1.

Множественные коллизии могут быть связаны как со средним качеством модели, так и с недостаточным объёмом выборки, а именно невысокими показателями счётчиков в мешке слов.

Эксперимент 3

Данный эксперимент посвящен определению наличия положительной корреляции между разностями имён и разностями объектов. В качестве разности между именами используется косинусная метрика. Разности между объектами определяется вручную Пользователем.

Для данного эксперимента проводится расчёт количества единиц в матрице попарных расстояний и среднего расстояния, а также их отношения — в зависимости от количества тем при разных типах моделях(юниграмная и биграмная, биграмная, юниграмная) без счётчиков, равных единице.

Существует проклятье размерности, когда минимальное расстояние стремится к среднему[10] — в этом случае корреляция отсутствует. Таким образом эксперимент заключает в определении отношения минимального расстояния к среднему на множестве имён. Также интересна зависимость данного отношения от количества тем, поскольку близость отношения к единице задаёт верхнюю границу на длину имени, т.е. на количество тем.

Минимальное расстояние равно 0, но в косинусной метрике меряется похожесть. Она вычисляется как 1 - расстояние. Таким образом оценивается доля похожих векторов, похожесть которых равна единице, от общего числа векторов. Также оценивается средняя похожесть, вычисляемая как суммарная попарная похожесть, разделённая на квадрат количества векторов. Кроме того интересно отношение доли единиц в матрице похожести к средней похожести в зависимости от количества тем.

Ожидается, что с ростом размерности доля единиц в матрице похожести уменьшается, как и средняя похожесть.

Увеличение отношения доли единиц в матрице похожести к средней похожести будет говорить о невыполнении требования положительной корреляции.

Результаты эксперимента 3

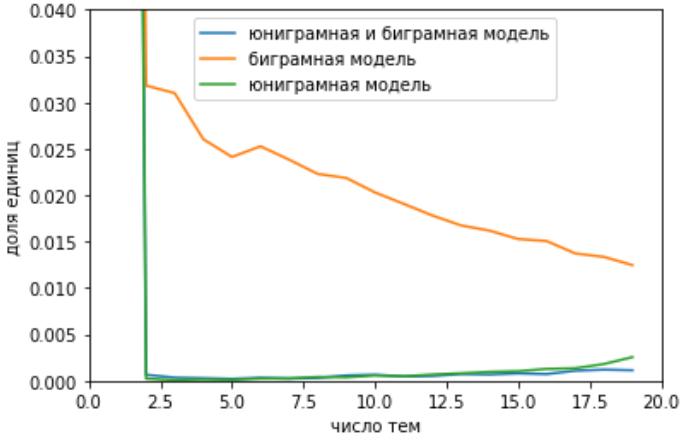

Рисунок 6. Доля единиц в матрице попарной похожести.

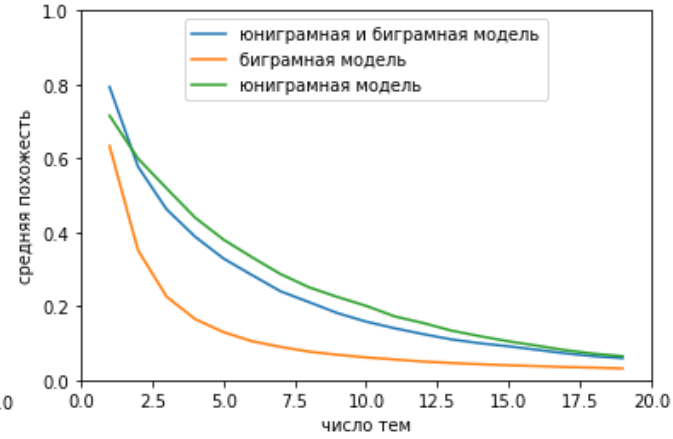

Рисунок 7. Средняя похожесть в матрице попарной похожести.

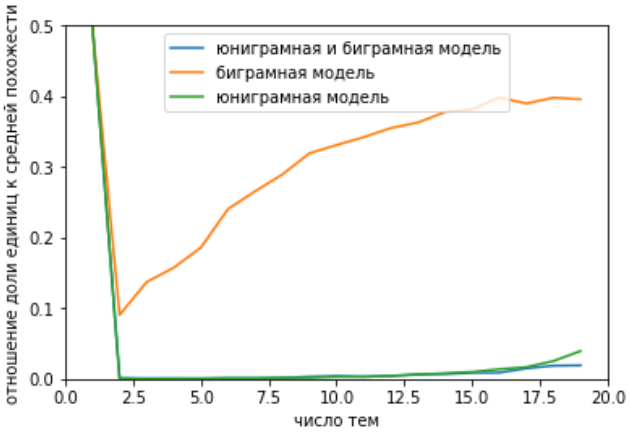

Рисунок 8. Отношения числа единиц к среднему расстоянию в матрице попарной похожести.

Получена следующая зависимость доли единиц в матрице попарной похожести, средней похожести и их отношения от числа тем(см. Рис. 6, 7, 8). Результаты эксперимента подтверждают, что доля единиц и средняя похожесть уменьшаются с увеличением числа тем. Также делается вывод о том, в биграмной модели не выполнено требование положительное корреляции. Возможно именно из-за этого в биграмной модели наблюдается повышенное число коллизий по сравнению с остальными моделями.

5.4 Результаты

Результаты экспериментов показывают, что тематическое моделирование может быть использовано для организации системы именования текстовых объектов, при условии достаточных размера выборки и качества тематической модели.

6. О влиянии системы именования на построение ИОС

В данном разделе рассматривается влияние предложенной системы именования на решение задач построения ИОС.

Предложенная система именования соответствует дополнительному требованию о том, что сложность построения имени для новых объектов не зависит от числа объектов. Сложность построение имени определяется количеством операций над матрицами Фи и Тетта, которые имеют фиксированный размер. Таким образом, предложенная система именования влияет на задачу Именование, позволяя минимизировать число операций построения имени для новых объектов.

Предложенная система именования позволяет существенно уменьшить количество поисковых операций, по сравнению с полнотекстовым поиском и поиском по аннотациям ключевыми словами. Для поиска по имени важно не число позиций в имени, а размер словаря, значения которого могут присваиваться данным позициям. Так, словарь тематик имеет существенно ограниченный размер, по сравнению со словарём документа и словарём ключевых слов документа. Поэтому поиск по нему осуществляется заметно быстрее. Особенно это должно быть заметно на больших коллекциях. Таким образом, предложенная система именования влияет на задачу Поиск, позволяя сократить число поисковых операций.

Наличие мягкой кластеризации в предложенной системе именования позволяет определить как сгруппировать объекты для хранения. Новый объект относится к той группе, тематика которой имеет наибольший вес в его имени. Хранение может быть организовано таким образом, что группы объектов свободно перемещаются по сети, в зависимости от вектора запросов Пользователей. Точнее перемещаются центры групп, это не документы, это центроиды кластеров. Таким образом, предложенная система именования влияет на задачу Хранение, позволяя определить каким образом группировать документы для хранения.

Группы объектов могут свободно перемещаться по сети, поскольку в ИОС отсутствует понятие адреса. Однако для организации подобного перемещения необходимо, что количество правил маршрутизации не зависело от числа ИО в сети, поскольку иначе таблицы правил маршрутизации со временем и увеличением количества ИО будут переполнены. Организовать данную независимость можно за счёт маршрутизации по меткам, когда каждой метке соответствует узел сети и таким образом маршрут является последовательностью меток. В этом случае сложность маршрутизации не зависит от количества имен. Сложность маршрутизации определяется количеством правил в таблице маршрутизации.

Предложенная система именования соответствует требованию положительной корреляции разностей между именами из разностей между объектами. Благодаря этому становится возможной "маршрутизация по смыслу", когда запрос поступает из любой точки сети и идёт по центроидам групп в сторону, где находится имя, наиболее близкое к запросу. Возможна аналогия с маршрутизацией по расстоянию Хэмминга. Для организации такой маршрутизации необходимо, чтобы похожие центроиды находились рядом в пространстве имен. Это условие выполнено поскольку предложенная система именования соответствует требованию положительной корреляции, а имена являются координатами.

Таким образом возникает задача отображения пространства координат центроидов групп на физическую топологию. Центроиды находятся рядом в пространстве имен по построению, в случае если тематическое моделирование произведено успешно. Однако два соседних центроида могут быть отображены на не соседние узлы, например, если соответствующие данным узлам Пользователи запросили соответствующий данным центроидам контент. В этом случае задержка до запрошенных объектов у Пользователей уменьшится, однако количество правил маршрутизации для поддержания целостности отображения пространства центроидов на пространства узлов увеличится. Таким образом возникает задача оптимизации, в которой нужно определить баланс между задержкой до Пользователей и количеством правил маршрутизации на узлах. Важно понимать, что не всегда можно «перестроить» существующее отображение координат центроидов на пространство координат узлов, иногда нужно построить его «с нуля». Также стоит отметить, что количество правил маршрутизации минимально, когда из разницы между именами следует разница между узлами, на которые отображаются данные имена; иными словами, близкие имена отображаются на близкие узлы. Таким образом, система именования влияет на задачу Передача, поскольку "маршрутизация по смыслу" возможна только когда соблюдается требование положительной корреляции, соблюдение которого зависит от успешности выделения тематик.

7. О единой системе именования и едином имени

Предложенная система именования используется для именования текстовых объектов. В то же время требования, которым она соответствует, также относятся и к системам именования других типов объектов, а именно аудио, видео и изображений.

Возникает вопрос о существовании единой для всех типов ИО системе именования, которая с одной стороны соответствует требованиям системы именования, а с другой позволяет получать унифицированные имена для всех типов ИО.

Возможен следующий подход к определению такой системы именования. Данный подход связан с приведением ИО всех типов к канонической форме, по которой затем строится имя. Данная форма называется синтетическим объектом, это пятый тип ИО. Содержательно это трёхмерная реконструкция сцены, описанной в ИО. Такая сцена может состоять из нескольких синтетических объектов, если исходный объект был составным. Однако более интересен случай, когда каждым из элементов данной сцены является не непосредственно синтетический объект, а его собирательный образ, архиобъект. Может быть проведена аналогия с центроидом в кластере. Исходные ИО могут быть подвергнуты различным преобразованиям, а именно сдвиг, поворот, отражение и масштабирование. Поэтому каноническая форма имени, архиобъект, возможно составной, должна быть инвариантна данным преобразованиям. Такой формы может не существовать. В этом случае может быть предложена функция, которая по двум архиобъектам определяет, является ли первый из них результатом преобразований второго.

Также важен вопрос о едином имени. В предлагаемой системе именования происходит последовательное построение машинного и пользовательского имён. Их единовременное построение возможно в случае автоматического именования тем. Однако единое имя, одновременно являющееся и машинным, и пользовательским, с помощью предлагаемой системы именования получено быть не может, поскольку пара <вероятность принадлежности к теме, название темы> не агрегируется. В то же время для изображений единое имя построено быть может. Примером такого имени является phash[11]. Содержательно это с одной стороны сжатое изображение, а с другой последовательность битов, хэш. Для phash выполняется требование уникальности и требование положительной корреляции, однако для выполнения требования восстановления по части необходима модификация структуры данных, например, использование мешка пикселей.

Таким образом, вопросы о существовании единой системы именования и единого имени для каждого типа ИО, включая синтетические объекты, являются развитием темы данной работы.

Заключение

В работе получены следующие результаты:

1) Сформулированы требования к системе именования
2) Предложен подход к формированию имени на основе семантического анализа содержимого информационного объекта
3) Экспериментально показано, что предложенная система именования способна удовлетворять сформулированным требованиям

Список литературы

[1] Xiaofei Liao; Li Lin; Guang Tan; Hai Jin; Xiaobin Yang; Wei Zhang. Bo LiLiveRender: A Cloud Gaming System Based on Compressed Graphics Streaming. 2016.
https://ieeexplore.ieee.org/document/7166339

[2] Anders Eriksson; Borje Ohlman; Karl-Ake Persson. What are the Services of an Information-centric Network, and Who Provides Them? 2012.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.5419&rep=rep1&type=pdf

[3] Еремеев М.А; Янина А.О. Разведочный поиск на основе тематического моделирования. 2019.
https://lomonosov-msu.ru/archive/Lomonosov_2019/data/16103/uid341423_report.pdf

[4] Воронцов К. В. BigARTM: от лего-конструктора тематических моделей к сервисам разведочного поиска. 2019.
http://www.machinelearning.ru/wiki/images/0/0d/Voron-2019-05-11-bigartm.pdf

[5] Serg Karpovich. Тематическая модель. 2017.
https://commons.wikimedia.org/wiki/File:Тематическая_модель.png

[6] Воронцов К. В. Обзор вероятностных тематических моделей. 2019.
http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

[7] Проект BigARTM. Библиотека тематического моделирования. 2019.
http://bigartm.org

[8] Колосов Алексей. Система именования текстовых объектов. 2019.
http://github.com/cheptil/nicn

[9] John Langford. Vowpal Wabbit. 2017.
http://hunch.net/~nyoml/vowpal_wabbit.pdf

[10] Джоэл Грас. Data Science. Наука о данных с нуля. 2017.
Санкт-Петербург. «БХВ-Петербург». Стр. 186-190.

[11] Zauner, Christoph: Implementation and Benchmarking of Perceptual Image Hash Functions. 2010.
https://www.phash.org/docs/pubs/thesis_zauner.pdf