ФГАОУ ВПО «УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ ПЕРВОГО ПРЕЗИДЕНТА РОССИИ Б.Н. ЕЛЬЦИНА»

На правах рукописи

Белобородов Александр Владимирович

МЕТОДЫ И МОДЕЛИ АНАЛИЗА БОЛЬШИХ КОЛЛЕКЦИЙ ВЕБ-ДОКУМЕНТОВ МЕДИЦИНСКОЙ ТЕМАТИКИ

Специальность 05.13.18 —
«Математическое моделирование, численные методы и комплексы программ»

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
доктор физико-математических наук, профессор
Волков Михаил Владимирович

Екатеринбург — 2018

Оглавление

Стр.

Стр.

Стр.

Введение

Интернет стал важным источником информации о здоровье для многих людей. В настоящее время в сети доступен огромный объём медицинской информации. Согласно исследованиям, проведенным центром PewResearch в 2013 году, 59% взрослых интернет-пользователей в США искали информацию о состоянии здоровья в сети [1]. В России эта цифра ниже, однако она уже является достаточно большой и продолжает расти: в 2014 году около 21% населения России использовало интернет как источник информации о здоровье, медицине, лекарствах [2].

Тем не менее, выводы интернет-пользователя о состоянии своего здоровья, сделанные на основе веб-данных, могут не соответствовать реальности из-за наличия большого количества неверной информации в открытом доступе или неумения пользователя корректно интерпретировать полученные знания. Медицинская информация, полученная в сети, может послужить сигналом к самодиагностике, самолечению или посещению врача без должных на то оснований, и, как следствие, нанести вред здоровью пользователя. В связи с этим актуальными являются задачи оценки качества медицинской информации в сети, а также развития методов поиска информации о здоровье пользователей и корректной её интерпретации.

Существует много способов доступа к медицинской информации в интернете: универсальные поисковые системы (Яндекс, Google, Bing и т.п.), специализированные и профессиональные поисковые системы (PubMed, Cochrane, Google Scholar), медицинские порталы (Русский медицинский сервер, WebMD), социальные вопросно-ответные сервисы (03.ru, Yahoo Answers, Ответы@Mail.Ru). Диссертационное исследование затрагивает вопросы качества информации в социальных вопросно-ответных сервисах (СВОС) и универсальных поисковых системах.

Цель диссертационной работы — разработка численных методов, моделей и комплексов программ для анализа, оценки и улучшения качества доступа к данным СВОС и веб-страницам о здоровье человека. Поставленная цель достигалась решением следующих **задач**:

– Разработать метод приближенной оценки качества данных вопросно-ответного сервиса. Реализовать соответствующий комплекс программ, провести ручную оценку с привлечением медицинских специалистов.
– Исследовать проблему качества данных CBOC через оценивание пользователей-авторов. Разработать метод оценки компетентности пользователей медицинских разделов CBOC.
– Разработать метод персонализации поиска по коллекции веб-страниц, посвящённых вопросам здоровья человека.

Методология и методы исследования. В диссертационном исследовании использовались методы информационного поиска, тематическое моделирование, лексический анализ текстовых данных, статистические численные методы. Для улучшения качества автоматической обработки текстов разработан модуль исправления ошибок и опечаток. Кроме того, для проверки, интерпретации и дополнения результатов автоматических методов применены методы экспертной оценки.

Основные положения, выносимые на защиту:

1. Разработан метод приближенной оценки качества медицинских разделов CBOC. Разработан алгоритм проверки адекватности предложенной модели качества на основе данных ручной экспертной оценки.
2. Разработан эффективный вычислительный метод и соответствующий алгоритм оценки компетентности пользователя CBOC.
3. Предложен метод моделирования пользователя поисковой системы на основе данных его медицинской карты. Реализован соответствующий алгоритм персонализации поиска по коллекции веб-страниц медицинской тематики.
4. Реализованы соответствующие комплексы проблемно-ориентированных программ. В частности, реализован эффективный алгоритм исправления орфографических ошибок и опечаток.

Научная новизна. В диссертационном исследовании представлена методика экспертной оценки качества медицинских разделов CBOC врачами, разработан комплекс программ для ее практической реализации. Разработан новый метод автоматической оценки компетентности пользователя CBOC в медицинских темах. В качестве одной из составляющих метода предложена модель тематического фокуса пользователя. Кроме того, разработан новый метод пер-

сонализации поиска медицинской информации расширением запроса данными медицинской карты пациента.

Практическая значимость. Результаты диссертационного исследования могут быть использованы для повышения качества и удобства использования вопросно-ответного сервиса, для повышения качества автоматического вопросно-ответного поиска. Предложенный алгоритм оценки компетентности пользователя CBOC в медицинских темах может быть использован для вычисления рейтинга пользователя или маршрутизации нового вопроса конкретному пользователю – специалисту по теме вопроса. Модуль исправления ошибок и опечаток является адаптивным, то есть может быть применён к текстовой коллекции любого вида после полуавтоматического обучения.

Апробация работы. Основные результаты диссертационного исследования докладывались на следующих конференциях и научных семинарах:

- SIGIR'2016 MedIR Workshop: семинар по поиску медицинской информации (Пиза, Италия, 17 – 21 июля 2016 г)
- ISMW-FRUCT'2016: конференция по обработке информации в вебе и социальных медиа (Санкт-Петербург, 2 – 3 сентября 2016 г)
- FDIA'2015: симпозиум по перспективным направлениям в информационном поиске (Салоники, Греция, 31 августа – 4 сентября 2015 г)
- CLEF'2014: конференция по оценке информационного поиска (Шеффилд, Великобритания, 15 – 18 сентября 2014 г)
- ECIR'2013: Европейская конференция по информационному поиску (Москва, 24 – 27 марта 2013 г)
- Молодежная школа-конференция "Современные проблемы математики"(Екатеринбург, 27 января – 02 февраля 2013 г)

Результаты диссертационной работы обсуждались на регулярном семинаре кафедры алгебры и дискретной математики ИМКН УрФУ, на Всероссийской школе-конференции по информационному поиску RuSSIR. Результаты получены частично в рамках проекта «Анализ данных и моделирование пользователей тематических социальных медиа», поддержанного грантом РФФИ №14-07-00589А.

Публикации. Основные результаты диссертационного исследования изложены в 6 печатных работах, 3 из которых проиндексированы в базе Scopus [3—5]; 2 опубликованы в сборниках трудов конференций [6; 7]; 1 – в сборнике тезисов конференции [8]. Кроме того, получено свидетельство о государственной

регистрации одного из программных комплексов, разработанных в рамках диссертации [9; 10].

Личный вклад. Автором диссертационной работы самостоятельно разработаны методы и модели, выносимые на защиту: методика полуавтоматической оценки качества вопросов и ответов о здоровье человека, модель сосредоточенности пользователя СВОС на определённой тематике, численный метод оценки компетентности пользователя, метод персонализации поиска по коллекции медицинских веб-страниц. Автору также принадлежат разработанные в рамках диссертации программные комплексы: экспериментальные реализации предложенных методов, веб-сервис для полуавтоматической оценки, модуль исправления орфографических ошибок и опечаток.

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения и трёх приложений. Полный объём диссертации составляет 107 страниц, включая 16 рисунков и 19 таблиц. Список литературы содержит 91 наименование.

Глава 1. Введение в предметную область. Текстовые данные медицинской тематики

Диссертационное исследование активно использует некоторые теоретические концепции из области информационного поиска и смежных научных дисциплин. В подразделе 1.1 приводятся определения основных понятий предметной области диссертации, таких, как информационный поиск, информационная потребность, релевантность. Также описаны необходимые понятия из области оценки качества информационного поиска, приведены стандартные меры, которые использовались при оценивании качества разработанных методов. Кроме того, в подразделе 1.3 рассматривается область вероятностного тематического моделирования, и даётся определение метода, используемого в главах 2, 3.

Подраздел 1.4 описывает данные, на которых проводились эксперименты в рамках диссертационной работы:

- русскоязычная коллекция вопросов с ответами, предоставленная сервисом Ответы@Mail.Ru, на которой тестировались методы, описанные в главах 2, 3;
- англоязычный набор данных международной инициативы CLEF eHealth'14, использованный для тестирования метода, предложенного в главе 4.

Здесь же обсуждаются некоторые типичные проблемы, возникающие при автоматической обработке данных социальных вопросно-ответных сервисов, и предлагается их решение. В частности, подраздел 1.5 описывает модуль автоматического исправления орфографических ошибок и опечаток в текстовых данных, используемый для коррекции слов в экспериментах с вопросами и ответами.

1.1 Некоторые понятия теории информационного поиска

При анализе текстовых данных (например, веб-страниц), исследователи активно пользуются понятиями теории информационного поиска. Приводимые ниже концепции подробно описаны в источнике [11].

Определение 1.1.1. *Информационный поиск (Information Retrieval, IR) — процесс поиска в большой коллекции (хранящейся как правило в памяти ЭВМ) произвольного неструктурированного материала (обычно документа), удовлетворяющего информационную потребность пользователя. Под неструктурированными данными обычно понимают данные, не имеющие ясной, семантически очевидной структуры, легко реализуемой программно.*

Основной задачей информационного поиска является разработка систем, выполняющих поиск по произвольному запросу. Цель такой системы — найти документы, которые являются наиболее *релевантными* по отношению к произвольной *информационной потребности* пользователя, сообщаемой системе при помощи однократных, обычно текстовых, запросов.

Определение 1.1.2. *Информационная потребность (Information Need) — это тема, о которой пользователь в конкретный момент времени хочет узнать больше при помощи поисковой системы. Её следует отличать от запроса — текстового выражения информационной потребности, сообщаемого системе.*

Определение 1.1.3. *Документ называется релевантным, если, с точки зрения пользователя, он содержит ценную информацию, удовлетворяющую его информационную потребность.*

В частности, технологии систем поиска в сети Интернет (далее интернет или веб) во многом основаны на приведённых выше концепциях.

В настоящее время интернет получает все более широкое распространение. Его используют в качестве среды для общения, развлечений, поиска и получения информации. Объём данных в вебе со временем растёт экспоненциально. Существует исследовательское направление *Big Data Analysis* (дословно «анализ больших данных»), в рамках которого разрабатываются методы, в которых извлечение полезного (то есть нетривиального, ранее неизвестного) знания производится в основном за счёт большого объёма данных.

Универсальной точкой входа в интернет-сервисы обычно является поисковая система общего назначения, например Яндекс, Google или Bing. Для получения специализированной информации могут использоваться поисковые системы, спроектированные специально для извлечения информации конкретного вида. Примерами могут служить TinEye — веб-сервис поиска по изображениям, поисковая система для научных публикаций Google Scholar, каталоги ме-

дицинской информации PubMed, Cochrane. Кроме того существуют различные системы, где информация как создаётся, так и потребляется рядовыми пользователями интернета.

Определение 1.1.4. *Контентом, генерируемым пользователями (User Generated Content, UGC) называется любая текстовая или визуальная информация произвольного формата, создаваемая в вебе не экспертами, а рядовыми пользователями, обычно внутри социальных сервисов.*

Сервисы, позволяющие любому (зарегистрированному или нет) пользователю свободно создавать и потреблять UGC, а также подразумевающие общение, например, социальные сети, блоги или форумы, принято называть социальными. Такие сервисы обычно состоят из множества страниц, каждая из которых посвящена определённому вопросу или узкой теме. Формат форума предполагает обсуждение произвольной темы произвольным числом пользователей в виде последовательных комментариев. Единая логически связанная беседой последовательность комментариев форума называется *тредом*. Блог — это подвид форума, в котором равноправное общение между пользователями трансформируется в отношения вида «один автор — много читателей».

Определение 1.1.5. *Социальный вопросно-ответный сервис (CBOC, Community Question Answering, CQA) — это сервис, предлагающий в качестве треда использовать формат ответов на один вопрос, поставленный в первом сообщении. При этом тред, в котором ответ на вопрос найден и верифицирован по правилам CBOC, закрывается.*

Таким образом, CBOC отличается от форума тем, что в треде не предполагается дальнейшее общение после получения правильного ответа на вопрос. Обычно создатели CBOC организуют внутри сервиса различные рейтинговые системы для стимулирования пользовательской активности, например, введение рейтинга/уровня пользователя, голосование за лучший ответ, геймификация ответов на вопросы. Ниже приводится несколько примеров наиболее крупных и известных CBOC:

- `StackOverflow` — англоязычная система, посвящённая тематике языков и технологий программирования;
- `StackExchange`, поддерживающий более широкий спектр технических тем;

– `Yahoo! Answers` — крупнейший международный сервис, поддерживающий множество языков (включая русский) и любую тематику.

Кроме того существуют крупные аналоги Yahoo! Answers на корейском, китайском и русском языках: `Naver`, `Baidu Zhidao` и `Ответы@Mail.Ru` соответственно. В данной работе для экспериментов использовались вопросы, ответы, а также обезличенные данные пользователей сервиса Ответы@Mail.Ru.

Разработка любых новых методов обычно подразумевает оценку качества их работы. В области IR тестирование новых подходов чаще всего проводится путём их сравнения с существующими методами. Для этого научным сообществом разработаны стандартные меры оценки качества методов информационного поиска. Ниже приводятся определения тех мер, что использовались для тестирования методов, представленных в данной работе.

1.2 Стандартные меры оценки качества методов IR

Оценка качества метода в информационном поиске и анализе текстовых данных чаще всего предполагает сравнение результатов работы этого метода с существующими аналогами на тестовой коллекции документов, которую можно описать формально как тройку (D, Q, R), где D — множество документов, на которых тестируется метод, Q — множество информационных потребностей, R, в свою очередь, — множество троек (q, d, r), в которых документу d, выданному по запросу, выражающему информационную потребность q, ставится в соответствие оценка релевантности r. В литературе тестовые коллекции часто называют *золотым стандартом* (gold standard, ground truth).

Определение 1.2.1. *Асессмент (assessment) — это процесс создания множества оценок релевантности R для тестовой коллекции (D, Q, R).*

Асессмент тестовой коллекции обычно проводится вручную специально приглашёнными для этой цели людьми — *асессорами*. В зависимости от целей тестируемой системы информационного поиска в качестве асессоров могут выступать как эксперты в определённой области, так и рядовые пользователи интернета. В любом случае оценка соответствия конкретного документа тестовой информационной потребности носит субъективный характер. Поэтому, для

получения надёжной и объективной общей оценки системы необходимо иметь достаточно большое число оценённых информационных потребностей. Если в маленьких тестовых коллекциях можно оценить все возможные пары «тестовая информационная потребность — документ», то для больших коллекций такая стратегия будет слишком затратной. Поэтому во втором случае применяют так называемый *метод общего котла* (pooling), при котором оценивается релевантность $D_p \subset D$ — смеси из k первых документов, возвращаемых несколькими (тестируемыми или просто наиболее популярными на момент оценки) поисковыми системами [11].

Для измерения уровня субъективности оценок некоторое число пар q, d оценивается несколькими асессорами. Затем по множеству эквивалентных оценок вычисляется уровень согласия асессоров. В литературе наиболее распространённым показателем согласия называется обычно *каппа-статистика Коэна.*

Определение 1.2.2. *Каппа-статистика Коэна (Kohen's Kappa statistics,* K*) — мера согласованности категориальных оценок, которая делает поправку на случайное совпадение оценок и выражается формулой*

$$\mathrm{K} = \frac{P(A) - P(E)}{1 - P(E)} \tag{1.1}$$

где $P(A)$ — доля совпадений, $P(E)$ — ожидаемая доля случайных совпадений.

В разных источниках называются разные границы приемлемости каппа-статистики, но в общем случае считается, что точные пороговые значения зависят от предназначения данных. В данной работе по аналогии с [12] согласованность считается высокой при $\mathrm{K} > 0{,}75$ и неприемлемой при $\mathrm{K} < 0{,}4$.

В области информационного поиска результат работы метода или системы часто является или может быть представлен в виде множества документов, выдаваемых по запросу. Такое множество документов называется *выдачей.* Оценка производительности метода на тестовых данных производится объединением оценок релевантности в один числовой показатель, который легко сравнить с показателями производительности других методов или систем. Базовыми метриками качества IR считаются *точность* и *полнота.* Чтобы дать им определение, рассмотрим основную задачу IR — поиск релевантных документов в коллекции по запросу — и следующую таблицу сопряжённых признаков:

Таблица 1 — Классификация документов после выполнения запроса

	Релевантные	Нерелевантные
Найденные	Истинно положительные (tp)	Ложно положительные (fp)
Ненайденные	Ложно отрицательные (fn)	Истинно отрицательные (tn)

Определение 1.2.3. *Точность (Precision, P) — это доля релевантных документов среди всех найденных (формула 1.2).*

$$P = \frac{tp}{tp + fp} \tag{1.2}$$

Определение 1.2.4. *Полнота (Recall, R) — это доля найденных релевантных документов среди всех релевантных (формула 1.3).*

$$R = \frac{tp}{tp + fn} \tag{1.3}$$

Использование только одной из оценок точности или полноты не вполне характеризует тестируемый метод по причине крайней несимметричности данных: в коллекциях достаточно большого размера подавляющее большинство документов по запросу являются нерелевантными. Например, система, выдающая всё множество документов коллекции по любому запросу, обеспечивает абсолютную полноту поиска ($R = 1$), являясь при этом непригодной для использования. Поэтому чаще всего для получения адекватной оценки точность и полноту комбинируют с помощью взвешенного среднего гармонического — *F-меры*:

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1 - \alpha)\frac{1}{R}} \tag{1.4}$$

Дополнительным преимуществом такой оценки является то, что в зависимости от целей тестируемой системы, вклад точности и полноты в общую оценку можно варьировать с помощью параметра α.

В диссертационном исследовании точность, полнота и F-мера использовались при разработке метода оценки компетентности, описанного в подразделе 3.4. Качество методов оценивалось с помощью других показателей, которые, благодаря хорошим дискриминирующим свойствам и устойчивости, являются

де-факто стандартом оценивания в современных исследованиях — *макроусредненной средней точности* и *нормированной дисконтированной совокупной выгоды*.

Пусть Q — множество информационных потребностей тестовой коллекции (D, Q, R), $\{d_1, d_2, ..., d_{m_j}\}$ — множество всех документов, релевантных информационной потребности $q_j \in Q$. Рассмотрим упорядоченное множество документов R_{jk}, выдаваемых поисковой системой по запросу q_j, вплоть до k-го релевантного документа.

Определение 1.2.5. *Средняя точность (Average Precision, AP) — это среднее арифметическое точности P множеств R_{jk} для всех натуральных k меньше либо равных m_j.*

Определение 1.2.6. *Макроусредненная средняя точность (Mean Average Precision, MAP) — это AP, усредненная по множеству информационных потребностей Q:*

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \tag{1.5}$$

Все показатели качества, описанные выше, определены в предположении, что показатель релевантности является индикаторной функцией, значение которой равно 1, если документ релевантный, и 0 иначе. Однако часто разработка новых методов требует оценивать степень релевантности документа по некоторой шкале (например, от 1 до 5). Чтобы учесть шкалу релевантности в оценке, современные исследования часто используют *нормированную дисконтированную совокупную выгоду*.

Пусть $rel(j)$ — функция релевантности j-го документа в выдаче.

Определение 1.2.7. *Дисконтированной совокупной выгодой на уровне k (Discounted Cumulative Gain at k, DCG@k) называется сумма взвешенных значений функции релевантности поисковой выдачи вплоть до k-го документа со штрафом за низкое положение в выдаче релевантных документов:*

$$DCG@k = \sum_{j=1}^{k} \frac{2^{rel(j)} - 1}{\log(j+1)} \tag{1.6}$$

Определение 1.2.8. *Нормированная дисконтированная совокупная выгода на уровне k (Normalized Discounted Cumulative Gain at k, NDCG@k) — это DCG@k, нормализованная по значению DCG идеально ранжированной выдачи n_q, и усредненная по всем информационным потребностям Q*

$$NDCG@k = \frac{1}{|Q|} \sum_{q \in Q} n_q DCG@k \tag{1.7}$$

К задачам информационного поиска относят и такие, как классификация и кластеризация, в частности, тематическая кластеризация документов. В её основе лежит математический аппарат теории построения вероятностных моделей. Далее приводится формальное описание *вероятностного тематического моделирования коллекции документов.*

1.3 Тематическое моделирование

Тематические модели — это семейство вероятностных генеративных моделей, используемых для определения тематики документов на основе их содержимого. В общем случае под темой понимается вероятностное распределение над «словами» документа. В качестве иллюстрации можно привести естественный пример текстовых документов (новости на тему *выборы в России*). Другой случай — тематическое моделирование изображений, где под «словами» понимаются небольшие фрагменты, изображающие различные визуальные элементы, которые встречаются на изображениях. Темами здесь могут быть, например, *полоски*, *лица людей* или *текстура дерева.* В тематических моделях обычно предполагается, что каждый документ коллекции содержит в себе смесь различных тем, представленных с определённой вероятностью. Одной из первых тематических моделей считается *модель вероятностного латентно-семантического анализа (Probabilistic Latent Semantic Analysis, PLSA)*, предложенная Т. Хоффманом в 1999 году [13].

В главе 3 диссертации описывается модель тематического фокуса пользователя, которая требует для своей реализации знания о тематическом распределении каждого документа. Для извлечения тематических распределений документов использовалась вероятностная модель порождения коллекции тек-

стов, предложенная Д. Блеем в 2003 г. — *тематическая модель латентного размещения Дирихле (Latent Dirichlet Allocation, LDA)* [14], основанная, в свою очередь, на PLSA.

Формально любая вероятностная модель определяется следующим образом. Пусть W — множество всех терминов коллекции документов D. Предполагается, что существует конечное множество тем T и употребление каждого термина $w \in W$ в каждом документе $d \in D$ связано с некоторой темой $t \in T$, заранее неизвестной. В контексте тематического моделирования коллекция документов рассматривается как множество троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d,w,t)$, заданного на конечном множестве $D \times W \times T$. Тогда можно определить вероятностную тематическую модель порождения данных как

$$P(w|d) = \sum_{t \in T} P(t|d)P(w|t) \tag{1.8}$$

Здесь термины $w \in W$ и документы $d \in D$ являются наблюдаемыми переменными, а темы $t \in T$ - латентной переменной.

Определение 1.3.1. *Построить тематическую модель коллекции документов D — значит найти распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех документов $d \in D$.*

В литературе распределения $p(w|t)$ и $p(t|d)$ обычно представлены в виде матриц Φ и Θ соответственно:

$$\Phi = (\varphi_{wt})_{W \times T}; \varphi_{wt} = p(w|t) \tag{1.9}$$

$$\Theta = (\theta_{td})_{T \times D}; \theta_{td} = p(t|d) \tag{1.10}$$

Определение 1.3.2. *Тематическая модель латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) — это тематическая модель (1.8) при дополнительном предположении, что векторы документов $\theta_d = (\theta_{td}) \in \mathbb{R}^{|T|}$ и векторы тем $\varphi_t = (\varphi_{wt}) \in \mathbb{R}^{|W|}$ порождаются распределениями Дирихле с параметрами $\alpha \in \mathbb{R}^{|T|}$ (уравнение (1.11)) и $\beta \in \mathbb{R}^{|W|}$ (уравнение (1.12)) соответственно.*

$$Dir(\theta_d;\alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \alpha_t > 0, \alpha_0 = \sum_t \alpha_t, \theta_{td} > 0, \sum_t \theta_{td} = 1 \quad (1.11)$$

$$Dir(\varphi_t;\beta) = \frac{\Gamma(\beta_0)}{\prod\limits_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1}, \beta_w > 0, \beta_0 = \sum_w \beta_w, \varphi_{wt} > 0, \sum_w \varphi_{wt} = 1 \quad (1.12)$$

Подробнее тематическое моделирование описано в литературе [13—15]. Источник [15] даёт подробное введение в данную предметную область, описывает различные модели и их свойства. Конкретные модели PLSA и LDA вводятся авторами в работах [13] и [14] соответственно.

Тематическое моделирование используется при решении широкого спектра задач:

- отслеживание тематических трендов в социальных медиа и научных публикациях [16];
- кластеризация, классификация, аннотирование документов [17] и изображений [18];
- разработка рекомендательных систем [19].

1.4 Текстовые данные Ответы@Mail.Ru и CLEF eHealth 2014

Для улучшения качества поиска и доступа к текстовым данным необходимо в первую очередь уметь оценивать качество этих данных, чему и посвящалась первая часть диссертационной работы (главы 2, 3). Эксперименты проводились на данных, предоставленных сервисом Otvety@Mail.Ru для исследовательских целей. В подразделе 1.4.1 приводится описание устройства сервиса и особенностей предоставленных данных.

Кроме того, в рамках исследования решалась задача улучшения качества поиска по веб-страницам медицинской тематики (глава 4). Для тестирования предложенного подхода использовалась тестовая коллекция веб-страниц медицинской тематики, опубликованная в рамках кампании по оценке методов информационного поиска CLEF eHealth'14 в целях развития этой области исследований. Данная тестовая коллекция описывается в подразделе 1.4.2.

Кроме того, в данном разделе нельзя обойти вопрос подготовки данных к исследованию, которая обычно предшествует основным исследовательским экспериментам: нормализации слов, исправления опечаток и т.п. В подразделах 1.4.3 и 1.5 даются основные определения и предлагается модуль исправления опечаток.

1.4.1 Ответы@Mail.Ru

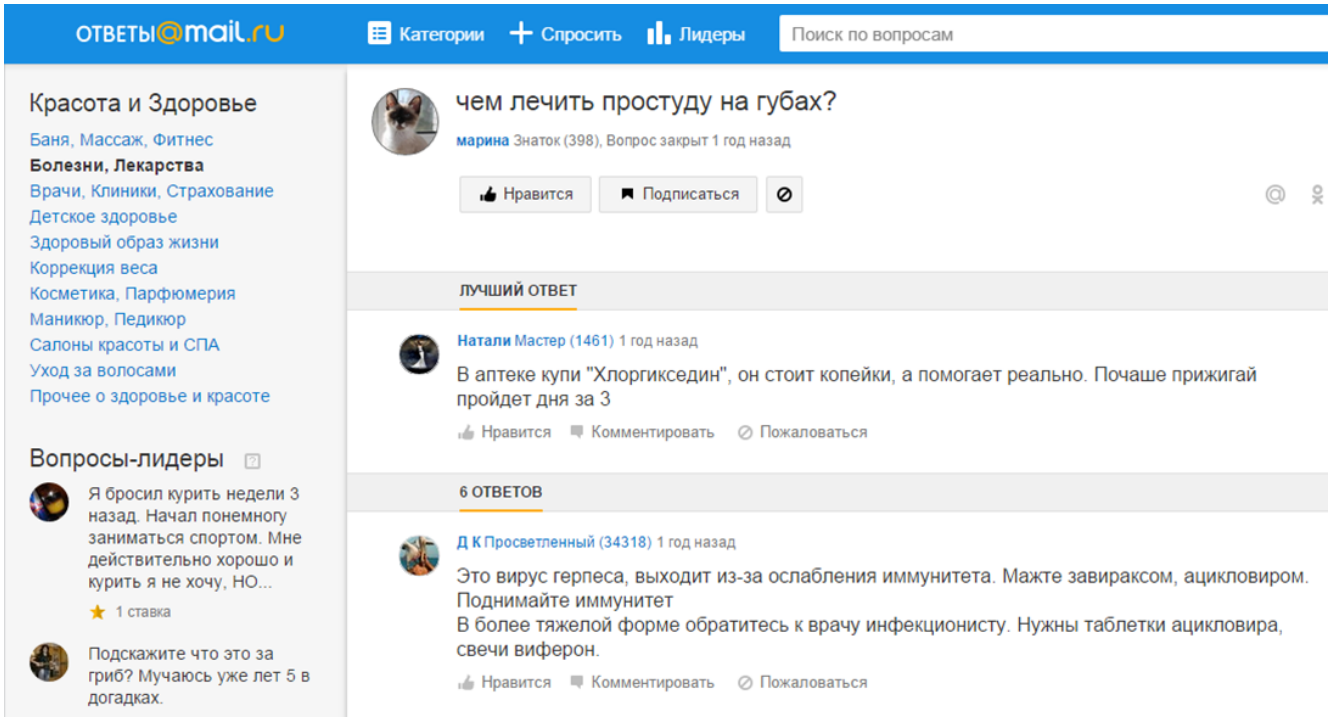
В настоящее время данные СВОС медицинской тематики являются предметом активного интереса исследователей, так как наряду с вопросами и ответами предоставляют для исследований мета-информацию (рейтинг пользователя, дата создания, оценка вопроса/ответа и т.п.), которая позволяет достичь более глубокого понимания данных, восстановить пользовательский и ситуативный контекст вопроса и ответа. Например, рейтинг пользователя применялся в качестве целевой функции для подбора параметров одного из предложенных в диссертации методов.

В силу устройства СВОС, ответы на вопросы часто персонализированы, то есть адресованы конкретно автору вопроса. Восстановленный контекст пользователя позволяет извлекать его персональные характеристики, которые могут быть полезны в различных приложениях, например, в задачах кластеризации, классификации или коллаборативной фильтрации.

Ответы@Mail.Ru – это русскоязычный социальный вопросно-ответный сервис с возможностью свободной регистрации, который был запущен в 2006 году и на момент получения данных (август 2012 года) собрал почти 80 миллионов вопросов и более 400 миллионов ответов [20]. Сайт имеет двухуровневую систему тематических разделов: около 30 разделов верхнего уровня, которые содержат суммарно около 200 подразделов. Пользователю, задающему вопрос, предлагается выбрать подходящий раздел из выпадающего списка. Например, рис. 1.1 демонстрирует страницу сервиса, на которой задан вопрос «Чем лечить простуду на губах?» в тематическом подразделе *Болезни, лекарства* и дано 7 ответов.

Набор предоставленных данных содержит около 11 миллионов вопросов и соответствующих ответов (в среднем 4,85 ответа на вопрос), заданных в 2012

Рисунок 1.1 — Страница СВОС Ответы@Mail.Ru



году. Диссертационное исследование фокусировалось на медицинских разделах сервиса:

– болезни, лекарства;
– врачи, клиники, страхование;
– детское здоровье;
– отвечает врач.

За 2012 год 225 427 уникальных пользователей медицинских разделов задали 227 828 вопросов и ответили на них. В табл. 2 приводятся некоторые статистические показатели медицинских разделов СВОС в сравнении с данными сервиса в целом. В частности, можно отметить, что пользователь медицинских разделов СВОС даёт в среднем значительно больше ответов (248 против 53), причём средняя длина ответа также больше, чем в сервисе в целом (23 слова против 15). Кроме того, ответ, который по результатам голосования сообщества СВОС становится лучшим, в медицинских разделах даётся быстрее (122 минуты против 218). Данные показатели могут говорить о том, что качество медицинских разделов СВОС Ответы@Mail.Ru выше, чем качество данных сервиса в целом. Это является одной из причин, по которым эксперименты с данными СВОС фокусируются в дальнейшем только на медицинских разделах.

В рамках диссертационного исследования данные СВОС Ответы@Mail.Ru прошли следующие этапы предварительной обработки:

Таблица 2 — Сравнение статистических показателей всех данных СВОС Ответы@Mail.Ru и его медицинских разделов

Показатель	Все разделы	Медицинские
Число вопросов	11 170 398,00	227 828,00
Средний размер вопроса (в словах)	9,90	10,00
Среднее число ответов на вопрос	4,85	4,13
Средний размер ответа (в словах)	17,15	22,99
Среднее время получения «лучшего» ответа (в минутах)	217,53	121,61
Число пользователей	2 690 358,00	225 427,00
Число отвечавших пользователей	1 022 090,00	127 602,00
Среднее число ответов пользователя	53,02	248,27
Средний рейтинг пользователя (в процентах)	10,73	10,94

– все словоформы приведены к нижнему регистру;
– исправлены орфографические ошибки и опечатки;
– отброшено 100 наиболее часто встречаемых слов;
– отброшены слова с частотой 1;
– отброшены непробельные небуквенные символы (например, знаки препинания);
– проведена лемматизация или стемминг слов.

На начальном этапе исследования был доступен меньший набор данных — 128 370 вопросов с ответами (в среднем 5 ответов на вопрос) в медицинских разделах СВОС. С ним, в частности, проводились эксперименты, описанные в главе 2. Все слова данного набора были лемматизированы. Набор данных, описанный в табл. 2, имеет значительно больший объём и использовался при разработке моделей и методов, описанных в главе 3. При экспериментах с данным набором использовался стемминг как более производительный по сравнению с лемматизацией метод нормализации слов.

1.4.2 CLEF eHealth 2014

В задаче персонализации поиска использовалась англоязычная тестовая коллекция, опубликованная в рамках кампании по оценке поиска медицинской информации CLEF eHealth в 2014 году [21]. Коллекция содержит около 1 млн веб-страниц о здоровье человека и 50 тестовых *топиков*.

Определение 1.4.1. *Топик — это формальное описание некоторой информационной потребности пользователя. Имеет структурированное представление (часто в формате XML). Обычно включает в себя запрос пользователя, подробное описание запроса, описание гипотетической ситуации, в которой запрос мог быть задан.*

Все топики составлены на основе медицинских карт пациентов, при этом, медицинские карты также прилагаются. Кроме того, для каждого из 50 топиков предоставляются оценки релевантности части веб-страниц в корпусе (около 1000 оценок на одну информационную потребность), полученные методом ручной экспертной оценки.

Коллекция документов собрана методом обхода веб-страниц и покрывает широкий спектр медицинских тем. Среди документов присутствуют как профессиональные статьи, предназначенные для медицинских специалистов, так и страницы, ориентированные на широкую публику. При создании множества тестовых информационных потребностей использовались обезличенные медицинские карты реальных пациентов из тестовой коллекции `MIMIC II`, описанной подробно в [22]. Медицинская карта пациента — это частично структурированный отчет о его истории болезни, который содержит в том числе информацию о состоянии его здоровья на тот момент, когда карта была выдана пациенту на руки. Листинг 1.1 демонстрирует пример медицинской карты пациента.

Топик в тестовой коллекции CLEF eHealth 2014 представляет собой структуру данных на языке XML, состоящую из нескольких полей (листинг 1.2): запрос (`title`), расширенное описание запроса (`desc`), описание ожидаемых результатов (`narr`), описание гипотетической ситуации, в которой мог бы быть задан подобный запрос (`scenario`), дополнительная информация о пациенте, извлеченная из его медицинской карты (`profile`).

Листинг 1.1 Пример медицинской карты пациента

```
Admission Date:    [**2014-03-28**]
Discharge Date:    [**2014-04-08**]
Date of Birth:     [**1930-09-21**]
Sex:               F
Service:           CARDIOTHORACIC
Allergies:
  Patient recorded as having No Known Allergies to Drugs

Attending:         [**Attending Info 565**]
Chief Complaint:   Chest pain
Major Surgical or Invasive Procedure:
  Coronary artery bypass graft 4.
History of Present Illness:
  83 year-old woman, patient of
  Dr. [**First Name (STitle) 5804**] [**Name (STitle) 2275**],
  with increased SOB with activity, left shoulder
  blade/back pain at rest, + MIBI, referred for cardiac
  cath. This pleasant 83 year-old patient notes becoming
  SOB when walking up hills or inclines about one year ago.
  This SOB has progressively worsened and she is now SOB
  when walking [**01-19**] city block (flat surface).
  [...]
Past Medical History:
  arthritis; carpal tunnel; shingles right arm 2000;
  needs right knee replacement; left knee replacement
  in [**2010**]; thyroidectomy 1978; cholecystectomy
  [**1981**]; hysterectomy 2001; h/o LGIB 2000-2001
  after taking baby ASA; 81 QOD
  [...]
```

Структура, показанная в листинге 1.2, удовлетворяет формату TREC — принятому в научном сообществе IR стандарту построения тестовых коллекций. TREC (Text Retrieval Evaluation Conference) [23] — наиболее известная кампания по созданию тестовых коллекций и оценке методов информационного поиска. Кампании TREC организуются ежегодно, начиная с 1992 года, американским национальным институтом по стандартам и технологиям U.S. NIST[1].

[1] https://www.nist.gov/

Листинг 1.2 Пример информационной потребности тестовой коллекции CLEF eHealth 2014 в формате TREC

```
<query>
  <title>
    thrombocytopenia treatment corticosteroids length
  </title>
  <desc>
    How long should be the corticosteroids treatment to cure
       thrombocytopenia?
  </desc>
  <narr>
    Documents should contain information about
    treatments of thrombocytopenia, and especially
    corticosteroids. It should describe the treatment,
    its duration and how the disease is cured using it.
    <scenario>
      The patient has a short-term disease, or has been
         hospitalised after an accident (little or no knowledge
         of the disorder, short-term treatment)
    </scenario>
    <profile>
      Professional female
    </profile>
  </narr>
</query>
```

1.4.3 Предварительная обработка данных

В задачах информационного поиска и анализа данных часто нет необходимости учитывать чрезвычайно редкие или частые слова. Самыми частыми словами в текстах на естественном языке обычно являются служебные слова (частицы, предлоги). Кроме того, в методах, использующих статистические подходы к анализу текстов, необходимо объединять одни и те же слова в разных формах или склонениях. Поэтому, при проведении экспериментов с текстовыми данными, проводят их предварительную обработку.

Определение 1.4.2. *Лемматизация — это процесс приведения каждого слова в документе к его нормальной форме [11].*

Таблица 3 — Нормальные формы некоторых частей речи русского языка

Часть речи	Нормальная форма
Имя существительное	именительный падеж, ед. число
Имя прилагательное	именительный падеж, мужской род, ед. число
Глагол	инфинитив
Причастие	соответствующий глагол в инфинитиве

Нормальные формы для некоторых частей речи русского языка приведены в табл. 3. В диссертационной работе лемматизация проводилась с помощью инструмента MyStem [24] — морфологического анализатора слов русского языка. В случаях, когда анализируемая текстовая коллекция содержит настолько много документов, что проведение лемматизации требует значительных вычислительных ресурсов, её часто заменяют *стеммингом* [25—27].

Определение 1.4.3. *Стеммингом слова называется процесс отсечения его изменяемой части. Стемминг коллекции документов — это отсечение изменяемых частей слов во всех документах коллекции.*

Стемминг, являясь более простой (и поэтому производительной) технологией по сравнению с лемматизацией, имеет более низкое качество нормализации слов для языков с богатой морфологией, в частности, для русского языка. В экспериментах с русскоязычными данными большого объёма, предполагающих предварительную обработку с помощью стемминга, использовался алгоритм Snowball [26] — адаптация алгоритма Портера [25] для большого количества европейских языков, в том числе русского. Англоязычные данные, использовавшиеся в разделе 4, обрабатывались с помощью реализации алгоритма Портера, встроенной в поисковую систему Terrier [28].

В задачах анализа и обработки коллекции текстовых документов обычно не учитывают самые частые слова, так как они встречаются практически в каждом документе, а также служебные части речи, так как они вносят незначительный семантический вклад в документ. Такие слова принято называть *стоп-словами*. Служебные слова в любом естественном тексте также находятся среди наиболее частых слов, поэтому в качестве предобработки документов достаточно отбросить только самые частые слова. Иногда цели решаемой задачи или специфика данных требуют отбросить только некоторые определённые

Рисунок 1.2 — Пример текста с ошибками в одном из ответов СВОС (орфография и пунктуация сохранены)

скорее всего ты ел чтото очень холодное типо мороженово и ел ты его большими кусками и поэтому у тя просто восполение, я не помню у меня тоже такое было темпиратура под 40 градусов была то озноб то жар

слова. Тогда составляется специальный словарь, которым руководствуется система предобработки документов при отбрасывании стоп-слов.

Наряду с частыми словами отбрасываются и наиболее редкие слова — те, что встречаются в коллекции документов единицы раз и бесполезны при работе, например, статистических методов. В источнике [11] показано, что частота терминов в коллекциях текстов на естественном языке убывает согласно степенному закону (закон Ципфа), то есть, достаточно быстро. Эксперименты на данных СВОС Ответы@Mail.Ru показывают, что 61% терминов встречается в коллекции всего 1 раз. Отбрасывание таких слов существенно сокращает объём словаря коллекции.

1.5 Модуль исправления орфографических ошибок и опечаток

Текстовые данные, сгенерированные рядовыми пользователями сети Интернет (UGC), часто содержат большое число опечаток, ошибок, некорректных словоупотреблений и других эффектов, которые усложняют анализ и обработку текстовых данных (например, рис. 1.2). Более того, в вопросах и ответах медицинской тематики подразумевается активное упоминание сложных профессиональных медицинских терминов (симптомов, заболеваний, наименований лекарственных препаратов, названий терапевтических процедур), что влечёт увеличение числа ошибок и опечаток в тексте. Всё это снижает качество автоматической обработки документов. Например, слово *темпиратура*, упомянутое в ответе на рис. 1.2, не будет восприниматься анализатором текста как *температура*. Методы, разработанные в рамках диссертационного исследования, большей частью основаны на словарях медицинских терминов. Чтобы минимизировать число ошибок распознавания таких терминов в данных, был разработан модуль исправления орфографических ошибок и опечаток.

Модуль принимает на вход словарь W_{ref} — множество пар $(w_i, p(w_i)), i \in [1; N]$, где w_i — одно из слов, на которые будут исправлены слова с опечатками, а $p(w_i)$ — вероятность встретить w_i в коллекции документов (формула 1.13). Назовём слова из W_{ref} *эталонными*.

$$W_{ref} = \{(w_1, p(w_1)), (w_2, p(w_2)), ..., (w_N, p(w_N))\} \tag{1.13}$$

Для эффективного поиска подходящего эталонного слова модуль хранит в памяти словарь W_{ref} в виде инвертированного индекса *триграмм* — 3-буквенных подпоследовательностей слов, учитывающих начало и конец слова с помощью дополнительных символов \$ и _ соответственно. Определим инвертированный индекс формально. Пусть W — это множество терминов коллекции документов D, некоторый термин $w \in W$ находится в документе $d \in D$ на позиции i. Назовём *словопозицией термина* w пару (d, i) (d здесь — идентификатор документа).

Определение 1.5.1. *Инвертированный индекс — это отображение* $W \to L$, *где каждый термин* $w \in W$ *переходит в список* $l \in L$ *всех его словопозиций в документах* D.

Инвертированный индекс, который реализован в модуле — это индекс триграмм всех терминов словаря W_{ref}, то есть $D = W_{ref}$, а W — это множество возможных триграмм. Обозначим его $FuzzyIndex$. Понятие словопозиции редуцировано до идентификатора документа, в качестве которого используется термин из W_{ref}. Пример $FuzzyIndex$ для словаря $W_{ref} = \{abcd, bcd, bcde\}$ приведён на рис. 1.3.

Рисунок 1.3 — Пример инвертированного индекса триграмм

$$\begin{aligned}
\$ab &\to [abcd] \\
abc &\to [abcd] \\
\$bc &\to [bcd, bcde] \\
bcd &\to [abcd, bcd, bcde] \\
cd_ &\to [abcd, bcd] \\
cde &\to [bcde] \\
de_ &\to [bcde]
\end{aligned}$$

После инициализации модуля словарем корректных слов W_{ref} и построения инвертированного индекса триграмм $FuzzyIndex$, последующий поиск кор-

ректного термина для слова с ошибкой или опечаткой происходит в 2 этапа. На первом этапе производится поиск слов-кандидатов для слова с опечаткой по алгоритму, представленному на рис. 1.4.

Метод $CalcEditDistance$, вызываемый в алгоритме на рис. 1.4, подсчитывает *расстояние Левенштейна* [29] между двумя словами, которое определяется следующим образом. Назовём *исправлением* в слове одну из следующих операций:

– добавление символа;
– удаление символа;
– замена одного символа на другой.

Определение 1.5.2. *Расстояние Левенштейна (расстояние редактирования) между словами w_1 и w_2 — это минимальное количество исправлений, требуемых для преобразования слова w_1 в слово w_2.*

На втором этапе необходимо проверить слова-кандидаты на пригодность и выбрать в итоге искомое слово. Алгоритм на рис. 1.4 возвращает множество слов с наименьшим расстоянием редактирования до слова с опечаткой. В некоторых случаях (например, когда в словаре W_{ref} нет подходящего слова) это расстояние может оказаться настолько большим, что кандидат на исправление значительно отличается от слова с опечаткой. Тогда модуль не должен возвращать слово, которое он считает корректным. Для реализации этой логики вводится адаптивное пороговое значение T_{edit}, которое является функцией от длины n_w слова w с опечаткой (формула 1.14).

Рисунок 1.4 — Алгоритм поиска корректного слова, реализованный в модуле исправления орфографических ошибок и опечаток

Вход: $mispelledWord$
Выход: $candidates$
1: $mw3grams := Get3gramsFor(mispelledWord)$
2: $words := \{\}$
3: **for** $g \in mw3grams$ **do**
4: $\quad indexWords := FuzzyIndex.Get(g)$
5: $\quad words := words \cup indexWords$
6: **end for**
7: $candidates := \arg\min_{w \in words} CalcEditDistance(w, mispelledWord)$

$$T_{edit}(n_w) = \begin{cases} 2, & n_w > 6 \\ 1, & 4 \leqslant n_w \leqslant 6 \\ 0, & n_w < 4 \end{cases} \tag{1.14}$$

Если наименьшее расстояние редактирования, возвращаемое алгоритмом, меньше T_{edit}, то список кандидатов принимается, иначе отвергается.

Наконец, из списка слов кандидатов необходимо выбрать слово, наиболее подходящее для исправления слова с опечаткой. Для этой цели вводится функция $correct : W \to W_{fixed}$, основанная на предположении, что наиболее подходящим кандидатом является слово, с большей вероятностью встречающееся в коллекции (формула 1.15).

$$correct(w) = \arg \max_{w \in candidates} p(w) \tag{1.15}$$

Из формулы 1.14 следует, что модуль исправления орфографических ошибок и опечаток не применяется к словам длины меньше 4. Это связано с тем, что для коротких слов алгоритм на рис. 1.4 возвращает слишком много кандидатов на исправление. В этом случае для определения верного кандидата недостаточно функции *correct* — нужны более точные методы, учитывающие контекст исправляемого слова и корпуса, в котором производится исправление. Разработка подобных методов является отдельной ветвью исследований автоматической коррекции текстов и остаётся за рамками данной диссертационной работы.

Тестирование модуля на данных СВОС Ответы@Mail.Ru

Словарь W_{ref} для исправления ошибок и опечаток в медицинских разделах СВОС Ответы@Mail.Ru сформирован в первую очередь с помощью национального корпуса русского языка (НКРЯ) [30]. За основу взят частотный словарь НКРЯ, содержащий словоформы частоты не менее 3 [31].

Так как национальный корпус составлен по большей части из текстов художественных и публицистических произведений, его лексика частично не соответствует контенту, генерируемому пользователями медицинских разделов СВОС. Здесь наблюдается две проблемы:

Таблица 4 — Статистика исправлений орфографических ошибок и опечаток в медицинских разделах CBOC Ответы@Mail.Ru

Множество слов	Мощность
W_{ref}	908 608
W_{CQA}	505 257
$W_{CQA} \setminus W_{ref}$	270 510
W_{fixed}	95 960

1. Медицинские разделы CBOC содержат употребления специальных медицинских терминов — симптомов, заболеваний, лекарств и т.п.
2. UGC-контент большей частью формируется с помощью неформального общения пользователей, что влечёт за собой использование специального сленга.

Как сленг CBOC, так и специальные медицинские термины с большой вероятностью отсутствуют в корпусе НКРЯ, поэтому, для формирования релевантного словаря необходимо дополнить его соответствующими источниками. В частности, для решения проблемы 1 W_{ref} дополнялся лексикой из справочника фельдшера [32], государственного реестра лекарственных средств России [33] и международной классификации болезней 10-го издания [34].

Проблема 2 предполагает обогащение словаря W_{ref} лексикой, которая может считаться некорректной с художественной или публицистической точки зрения, однако для целей автоматической обработки текстов и распознавания слов необходимо учитывать данное лексическое смещение. В этом случае использовался частотный словарь той же коллекции вопросов и ответов, на которой и проводились эксперименты по исправлению ошибок и опечаток: в словарь W_{ref} добавлялись словоформы частотой не менее 10 (это 16% полного словаря коллекции), в предположении, что слова, достаточно часто встречаемые в тексте, вероятно, не содержат ошибок или опечаток. Константа 10 получена путём ручного просмотра срезов словаря на разных частотах на предмет корректности абсолютного большинства слов определённой частоты. Итоговое множество W_{ref} содержит 908 608 пар $(w, p(w))$, из которых 882 385 пар взяты из корпуса НКРЯ, а 20 052 и 6 171 пара — из медицинских справочников и данных CBOC соответственно.

Так как словарь W_{ref} составлен на основе нескольких источников информации, существует проблема сравнимости показателя $p(w)$ для слов из различных источников. Данная проблема решена в два этапа:

1. Назначение приоритета каждому источнику. Медицинские справочники считаются более приоритетным источником, чем НКРЯ, который, в свою очередь, имеет больший приоритет, чем корпус СВОС.
2. Модификация собственной частоты слова в источнике с учётом приоритета.

На практике это реализовано следующим образом. В качестве $p(w)$ слова w из словаря СВОС используется его собственная частота встречаемости в корпусе. Собственные частоты слов источника НКРЯ параллельно сдвигаются так, чтобы минимальная частота слова НКРЯ была больше, чем максимальная частота слова из СВОС. Таким образом, частота любого слова НКРЯ больше частоты любого слова СВОС. Аналогичная операция производится с собственными частотами медицинских словарей относительно новых частот словаря НКРЯ.

Табл. 4 демонстрирует размеры некоторых словарей при работе модуля на данных вопросов и ответов Ответы@Mail.Ru, чей словарь обозначен как W_{CQA}

Таблица 5 — Некоторые типы опечаток/орфографических ошибок пользователей СВОС Ответы@Mail.Ru с примерами исправлений

Тип опечатки/ошибки	Оригинальное слово	Исправленная версия
Орфографические ошибки	восполение симтомы гоноккок ожёга	воспаление симптомы гонококк ожога
«Склеивание» соседних слов	вбрюшной кмышце личногои	брюшной мышце личного
Замена русских букв на латинские аналоги	kогда mожно начинаюt	когда можно начинают
Другие типы	lтироксин анальгин2 регенерациb	тироксин анальгин регенерация

и состоит из списка всех словоформ, встреченных в корпусе хотя бы однажды. Из таблицы видно, что больше половины (270 510) слов корпуса не входит в словарь W_{ref}, выполняющий роль множества «корректных» слов. В частности, показано, что почти одна пятая часть словаря W_{CQA} подверглась исправлению ошибки или опечатки (множество исправленных слов обозначено как W_{fixed} : $W_{fixed} \subset W_{CQA} \setminus W_{ref}$).

В рамках диссертационного исследования не решалась задача классификации типов ошибок и замера их распространённости в данных. Для наглядной демонстрации работы модуля наиболее частые исправления вручную разделены на 4 группы, которые показаны в табл. 5 вместе с соответствующими примерами. В частности, можно отметить примеры исправления ошибок в медицинских терминах (*воспаление, гонококк, анальгин*), что свидетельствует о пользе применения модуля в решении специфических задач, описанных далее, в главах 2 и 3. Модуль исправления орфографических ошибок и опечаток зарегистрирован в государственном реестре программ для ЭВМ [9], копия свидетельства регистрации приведена в приложении В. Исходный код модуля опубликован в источнике [10] и доступен для использования под лицензией MIT [35].

1.6 Выводы

В главе 1 дано введение в предметную область работы: даны определения основным понятиям, описаны некоторые концепции, методы теории информационного поиска и анализа текстовых данных, на которые опирается диссертационное исследование.

Кроме того, описаны основные принципы работы социального вопросно-ответного сервиса Ответы@Mail.Ru и дорожки по оценке методов информационного поиска по медицинским данным CLEF eHealth'14 — источников данных для экспериментов. Приводится описание структурных особенностей документов, преимущества и недостатки данных подобного типа.

Дано описание типичных проблем, возникающих при решении задач предварительной обработки текстовых данных, а также общепринятых способов их решения: лемматизации, стемминга, отбрасывания стоп-слов. Эксперименты показали, что существенная часть слов требует орфографической коррекции. В

подразделе 1.5 приведено описание алгоритма исправления орфографических ошибок и опечаток в текстовых данных произвольной стилистики. На основе алгоритма разработан программный модуль автоматического исправления ошибок и опечаток. С помощью модуля исправлено 18,9% слов в медицинских разделах СВОС Ответы@Mail.Ru — данных для экспериментов, описываемых в следующих разделах. Табл. 5 демонстрирует наиболее частые типы исправлений с примерами.

Глава 2. Метод автоматической оценки качества данных СВОС

Социальные вопросно-ответные сервисы аккумулируют со временем огромные массивы вопросов и ответов, однако на фоне роста количества пользователей исследователи отмечают снижение общего уровня качества данных СВОС [36]. В настоящей главе предложен метод автоматической оценки качества контента, генерируемого пользователями медицинских разделов СВОС. Для верификации подхода разработана методика ручной экспертной оценки качества данных.

2.1 Постановка задачи

Современные исследования качества данных СВОС медицинской тематики достаточно малочисленны и основаны по большей части на ручной обработке небольших объёмов данных. Среди основных причин этому называются относительная молодость СВОС, сложность проверки медицинских утверждений, сложность привлечения врачей к оценке данных, политика доступа к данным [3; 37—39]. В рамках диссертационного исследования была поставлена задача разработать метод полуавтоматической оценки качества данных большого объёма и провести оценку коллекции вопросов и ответов медицинских разделов сервиса Ответы@Mail.Ru.

Согласно исследованиям [36] рост объёма данных СВОС неминуемо влечёт за собой ухудшение качества данных в среднем. Отчасти это связано с тем, что в общей массе контента уменьшается доля *фактологических* вопросов — таких, ответами на которые служат конкретные факты, например: *Какова высота горы Килиманджаро?*. Наряду с этим, увеличивается доля нетривиальных вопросов, подразумевающих сложные ответы, дискуссии, сбор мнений или опыта. С другой стороны, значительный рост числа пользователей уменьшает средний уровень профессионализма в сообществе, из-за чего качество ответов на фактологические вопросы также падает. Данная проблема порождает потребность в методах оценки качества вопросов и ответов с целью их дальнейшей фильтрации, ранжирования или переиспользования в смежных приложениях.

Рисунок 2.1 — Пример пары «вопрос–ответ», удовлетворяющей паттернам $DisMed_1$, $DisMed_2$

Вопрос:	Посоветуйте хорошие капли от [*насморка*]?
Ответ:	Попробуйте [*Санорин*] или [*Назол*] Адванс

Цель исследования, описываемого в настоящей главе, состоит в том, чтобы научиться оценивать качество некоторого подмножества данных СВОС (с набором ограничений), и, далее, попытаться обобщить полученную оценку на всё множество вопросов и ответов медицинской тематики.

Пусть Q — множество вопросов. Определим для $q \in Q$ предикат Dis: $Dis(q) = TRUE$, если q удовлетворяет выражению *«Как вылечить заболевание d?»*, где $d \in D$ — заболевание из множества D, определённого заранее. Определим предикат Med_k: $Med_k(a) = TRUE$, если в ответе a упоминается $j \geqslant k$ различных препаратов из множества лекарств M. Тогда для пары $(q, a), q \in Q, a \in A$, где a — ответ на вопрос q, можно определить паттерн $DisMed_k$ как булеву функцию $Q \times A \to \{0,1\}$:

$$DisMed_k(q,a) = Dis(q) \& Med_k(a) \tag{2.1}$$

Пара (q,a) удовлетворяет паттерну $DisMed_k$, если $DisMed_k(q,a) = TRUE$. Например, рис. 2.1 демонстрирует пару, которая удовлетворяет паттернам $DisMed_1$ и $DisMed_2$, но не удовлетворяет паттернам $DisMed_3$, $DisMed_4$ и т.д.

Требуется разработать метод автоматической оценки пар, удовлетворяющих паттерну $DisMed_k$: $\{(q,a) | DisMed_k(q,a) = TRUE\}$, разработать методику ручной оценки таких пар для валидации автоматического метода, реализовать соответствующее программное обеспечение, оценить качество автоматического метода с помощью ручного асессмента.

2.2 Обзор литературы

Важным этапом в разработке любого метода оценки качества контента является определение и фиксация понятия качества. Разные источники описывают данный термин по-разному. Например, авторы [40] воспринимают качество

ответа как нечто внешнее по отношению к авторам вопроса и ответа — «объективное» знание. Работы [41; 42] напротив, исходят из того, как воспринимается качество автором вопроса, то есть, насколько ответ субъективно удовлетворяет его информационную потребность. Так как диссертационное исследование рассматривало вопросы переиспользуемости знаний из ответов СВОС другими пользователями, что требует рассматривать качество в объективном ключе, данная работа во многом опирается на результаты, описанные в источнике [40].

Некоторые исследования, например [42; 43], рассматривают качество данных только как качество ответов, другие — как качество вопросов [44—46], однако в данной работе подобно [40] рассматривалось качество и вопросов, и ответов. Необходимо отметить, что в автоматической части метода оценки, описанного в подразделе 2.5, оценивается только качество ответа на вопрос, а качество вопроса оценивается в ручной части метода (подраздел 2.6).

Еще одним аспектом качества вопросов и ответов считается цель вопроса. Авторы [44; 45] подразделяют вопросы на следующие типы:

– поиск информации;
– общение;
– развлечение (юмор).

Данная работа сфокусирована прежде всего на оценке качества данных медицинской тематики, поэтому в исследовании рассматриваются вопросы, целью которых является поиск информации.

В области анализа качества СВОС общей тематики описано несколько методов, автоматически оценивающих качество данных. Некоторые источники предлагают оценивать качество контента «на лету», сразу по поступлении вопроса или ответа; другие работы исследуют архивные данные, которые содержат наряду с контентом мета-информацию: пользовательские рейтинги, комментарии, статистику просмотров и т.п. Модели, описанные в литературе, широко используют методы машинного обучения с большим набором признаков. Большинство работ упоминает следующие группы признаков:

– текстовые признаки, отражающие грамотность речи, опечатки, визуальное оформление контента, читаемость и т.п.;
– пользовательские признаки, такие как рейтинг, активность, достижения, уровень экспертизы в теме вопроса, взаимодействие с другими пользователями и т.п.;
– различные статистики, например, количество просмотров и кликов.

Признаки, перечисленные выше, не являются зависимыми от тематики, однако в литературе отмечается, что данные различных тематик отличаются в вопросах поведения пользователей, используемых тезаурусах, и т.д. По этой причине все больше исследований фокусируется на конкретной теме, например, авторы [47; 48] анализировали данные сервиса Stackoverflow (вопросы и ответы на тему программирования и информационных технологий) с учётом специфики предметной области и тематических словарей.

Тема анализа качества медицинского контента CBOC мало представлена в современных исследованиях. Жанг в работе [49] описывает лингвистические, темпоральные и пользовательские (когнитивные) аспекты качества CBOC, мотивацию отвечающих пользователей на 270 вопросах, сэмплированных из медицинских разделов сервиса Yahoo! Answers. Предварительные эксперименты по ручной оценке качества 10 вопросов медицинской тематики описаны в [37]. Трём группам асессоров — пользователям, задававшим вопросы в CBOC, сотрудниками медицинской библиотеки и медсёстрам — предлагалось ответить на вопросы анкеты, затрагивающей разные аспекты качества данных. В дальнейшем авторы увеличили выборку до 400 вопросов, касающихся всех видов заболеваний и симптомов [38; 39]. Это исследование можно считать достаточно масштабным, если учесть тот факт, что качество данных оценивалось вручную.

В работе [50] предложен полуавтоматический метод оценки качества вопросов и ответов Yahoo! Answers про вирус гриппа H1N1 с помощью тематического моделирования. Авторы описали наиболее важные темы вопросов, типы ресурсов, на которые ссылались пользователи, и медицинские концепции, упомянутые в данных. В диссертационной работе использовались похожие методы, например, опросы пользователей и медицинских специалистов, тематическое моделирование, словари медицинских концепций. Упор, однако, делался на возможность автоматической обработки больших массивов данных. Для этого проведен обзор литературы на смежную тему — анализ сообщений социального сервиса Twitter, которые имеют гораздо более лояльную политику доступа к данным.

Данные Twitter показывают большой потенциал решения задач анализа медицинских данных. Работы [51—53] исследуют симптоматическое лечение, использование медицинских препаратов, поведенческие факторы риска заболеваний, географическую локализацию вспышек заболеваемости и т.п. В настоящем диссертационном исследовании похожие идеи применялись к данным

CBOC: аналогично [53] проводилась предварительная обработка данных; такие методы, как тематическое моделирование, использование доменных словарей применялись аналогично работе [52], но для решения задач, связанных с качеством вопросов и ответов.

В литературе описаны также первые попытки построить интеллектуальные системы на основе знаний, извлекаемых из данных CBOC. Например, авторы [54] предлагают экспериментальную диалоговую систему медицинской тематики на данных Yahoo! Answers, однако не затрагивают в работе вопросы качества данных.

2.3 Предварительная оценка качества вопросов и ответов

Современные исследования сервисов Yahoo! Answers [50] и Twitter [52] формулируют общую гипотезу о том, что крупные социальные сообщества чутко реагируют на внешние изменения, и это можно отследить по данным при помощи статистических методов. С целью проверки этого предположения в рамках диссертации поставлено несколько экспериментов с данными сервиса Ответы@Mail.Ru [3]. Кроме того, проведена предварительная оценка качества вопросов и ответов.

Эксперименты проводились на уменьшенной коллекции — 95 002 вопроса с ответами из категории *Болезни, лекарства* за период с апреля 2011 по март 2012 года. Из 133 163 уникальных пользователей-авторов вопросов и ответов 74 760 (56,1%) имеют публичный профиль, что позволяет восстановить для них некоторую информацию, например возраст, пол или регион проживания.

Один из экспериментов посвящён сопоставлению по времени всплесков вопросов определённой тематики некоторым событиям реального мира. Для восстановления тем в данных CBOC применялось тематическое моделирование. Модель строилась с использованием метода латентного размещения Дирихле (LDA), описанного в пункте 1.3 — реализация *GibbsLDA++* с параметрами $T = 100; \alpha = 0{,}5; \beta = 0{,}1$ как наиболее устойчивыми согласно результатам работы [55], где T — число тем, α, β — параметры распределения Дирихле. В качестве документа рассматривалась конкатенация вопроса и ответов на него. Ручной осмотр тем показал, что большинство из них (71 тема) являются зна-

чимыми. Из 100 тем вручную было отброшено 29, содержащих в основном служебные слова, числа и т.п. Примеры некоторых значимых тем приведены в табл. 6.

Таблица 6 — Примеры медицинских топиков, полученных с помощью LDA

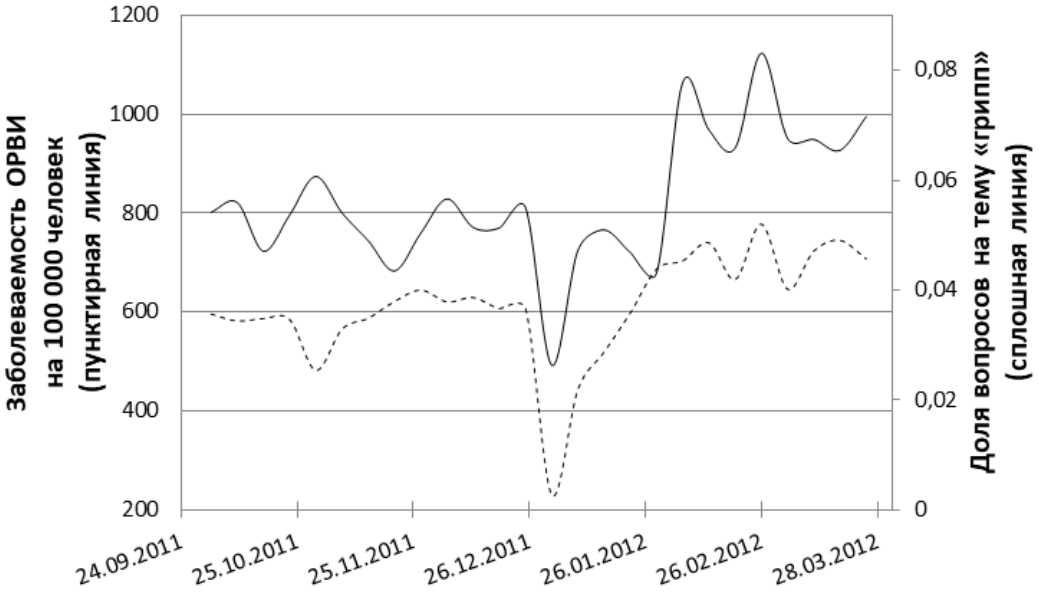
Тема 1	Тема 2	Тема 3	Тема 4	Тема 5
грипп	нос	кашель	рак	печень
37	течь	лёгкое	опухоль	желчь
5	капля	бронхит	клетка	диета
38	насморк	пневмония	стадия	мочевой
грипп	ЛОР	астма	случай	поджелудочный
простуда	промывать	сухой	опасный	орган
повышаться	дышать	мокрота	онкология	УЗИ
тело	сопля	дышать	положение	панкреатит
организм	слизистый	сироп	родинка	острый
высокий	пазуха	дыхание	равномерный	хронический

Примечание — темы представляют собой вероятностные распределения слов и не имеют однозначных названий. Каждая колонка показывает 10 самых вероятных слов в соответствующей теме.

Для сопоставления была выбрана доступная статистика заболеваемости острой респираторной вирусной инфекцией (ОРВИ) в России и тема 1 (табл. 6). Заболеваемость ОРВИ в России за 2011–2012 гг. опубликована в ежемесячном бюллетене EuroFlu всемирной организации здравоохранения [56]. Рис. 2.2 демонстрирует сопоставление данных о заболеваемости ОРВИ и динамики вопросов на тему, связанную с терминами «грипп», «орви», «простуда» (коэффициент корреляции Пирсона $r_{flu} \approx 0{,}45$).

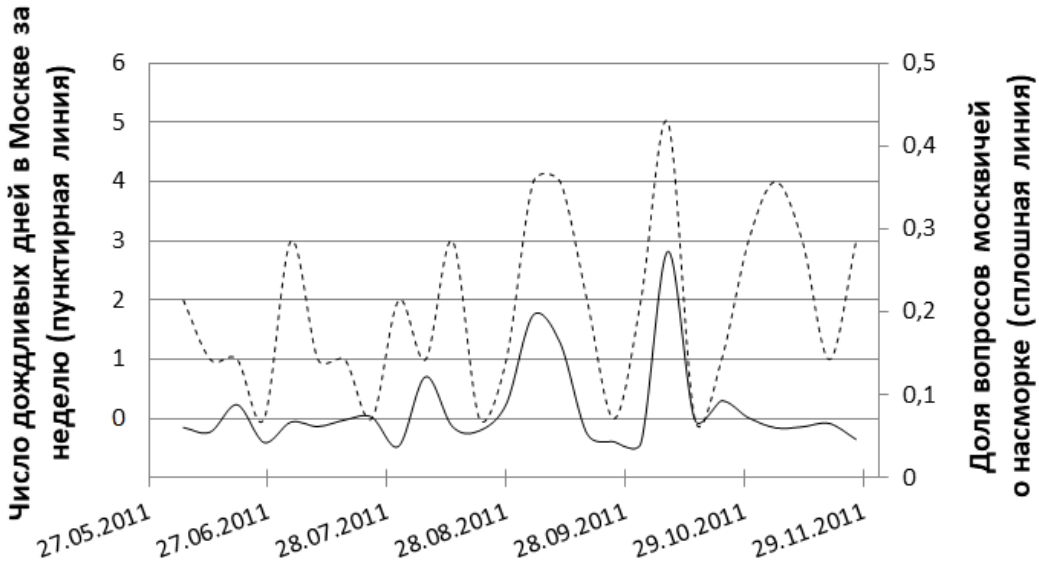
Кроме того, с сайта метеорологической геоинформационной системы *gismeteo.ru* методом итеративного обхода страниц собрана информация об осадках в Москве и Московской области, по которой посчитано число дождливых дней в неделю за тот же период, в который задавались вопросы СВОС. Данные регионы выбраны как самые частотные, указанные в качестве места проживания в открытых профилях пользователей СВОС. Количество дождливых дней в неделю сопоставлялось с динамикой вопросов на тему насморка (табл. 6, те-

Рисунок 2.2 — Динамика вопросов на тему «грипп, орви» на фоне данных заболеваемости ОРВИ в России



ма 2), заданных 3004 пользователями из Москвы и Московской области. Сопоставление всплесков темы о насморке и доступных статистических данных о реальном мире показано на 2.3 (коэффициент корреляции Пирсона $r_{rain} \approx 0{,}55$).

Рисунок 2.3 — Динамика вопросов москвичей о насморке в дождливую погоду



Положительные значения корреляции свидетельствует о наличии частичной прямой линейной зависимости между всплесками тем и событиями реального мира. Недостаточно сильные показатели зависимости могут быть обусловлены несколькими причинами:

– малое число доступных публичных профилей пользователей;

– неполное соответствие темы 1 заболеванию ОРВИ (слово *простуда* указывает на вопросы о простудных заболеваниях);
– наличие сторонних факторов, влияющих на итоговое распределение вопросов (например, влияние дождливой погоды на всплеск насморка может отсутствовать, если температура воздуха высокая).

Наличие в данных структуры «вопрос-ответы» позволяет определять разные типы информационных потребностей пользователей. Авторы работы [57] выделяют «диагностические» медицинские вопросы и подразделяют их на 2 класса по типу информационной потребности:

1. *evidence-directed* — определение заболевания по признакам и симптомам.
2. *hypothesis-directed* — подтверждение конкретного диагноза и поиск информации о лечении.

В диссертационном исследовании по аналогии с [57] проведен эксперимент по извлечению пар «вопрос–ответ», в которых вопрос принадлежит классу *hypothesis-directed*. Извлекались пары $(q, a), q \in Q, a \in A_q$, соответствующие паттерну $DisMed_1$ (формула 2.1), который, предположительно, хорошо аппроксимирует *hypothesis-directed*-вопросы. Согласно [57], данные такого типа полезны, например, для задач построения ассистирующих диагностических диалоговых систем.

Для извлечения вопросов по шаблону $DisMed_1$ составлены словари заболеваний и лекарственных средств. Словарь заболеваний составлен на основе справочника фельдшера [32] и содержит 1 049 наименований. Для составления словаря медицинских препаратов (размером 4 120) использовался государственный реестр лекарственных средств России (ГРЛС) [33]. Подробно компиляция словаря медикаментов описана в подразделе 2.4.

Так как названия заболеваний и медицинских препаратов являются профессиональными медицинскими терминами, а пользователи СВОС не являются в общем случае медицинскими профессионалами, вопросы и ответы предварительно были обработаны модулем исправления орфографических ошибок и опечаток (пункт 1.5), обученном на составленных словарях.

Из 95 002 вопросов по крайней мере 15 415 (16,2%) содержат хотя бы одну пару, удовлетворяющую шаблону $DisMed_1$. На основе извлеченных пар составлена статистика совместной встречаемости заболеваний и лекарственных

средств (табл. 7), где заболевание упомянуто в вопросе, а лекарство — в соответствующем ответе.

Таблица 7 — Примеры популярных заболеваний в вопросах и лекарственных средств, наиболее часто упоминаемых в соответствующих ответах

Заболевание	Q_{dis}	Лекарство	QA_{med}
Ангина	874	Йод	0,16
		Ромашка	0,15
		Фурацилин	0,13
		Люголь	0,12
		Шалфей	0,10
Молочница	704	Флюкостат*	0,30
		Кандид**	0,20
		Клотримазол**	0,18
		Флюконазол*	0,17
		Дифлюкан*	0,14
Герпес	621	Ацикловир***	0,53
		Зовиракс***	0,26
		Сера	0,17
		Фенистил*	0,07
		Валтрекс	0,06

Примечание:

1. Q_{dis} — число вопросов, упоминающих заболевание.

2. QA_{med} — доля вопросов, в ответах на которые встретилось лекарство.

3. Наименования с равным числом звёздочек используют одинаковые действующие вещества.

Табл. 7 демонстрирует, какие лекарства чаще всего рекомендуют принимать пользователи СВОС при лечении указанных заболеваний. Существует реестр лекарственных средств и товаров аптечного ассортимента [58], в котором для каждого лекарства перечислены заболевания, для которых данное лекар-

ство обычно рекомендуется к применению. Ручная проверка наиболее часто употребляемых пар заболеваний и лекарств с помощью данного ресурса показала высокую степень качества ответов на вопросы, удовлетворяющие шаблону $DisMed_1$. Подраздел 2.5 описывает основной блок экспериментов по автоматизации оценки качества данных СВОС.

2.4 Тематические словари заболеваний и лекарственных средств

Метод оценки качества данных СВОС медицинской тематики, предлагаемый в рамках данного исследования, основывается на словарях заболеваний и лекарственных средств.

Для составления словаря заболеваний была проведена консультация с медицинским специалистом, по результатам которой сформулированы следующие критерии выбора болезней:

- частая встречаемость заболевания/симптома в имеющихся данных СВОС;
- относительная простота диагностики и лечения (с точки зрения врача);
- наличие частых случаев самостоятельного лечения заболевания/симптома (по данным врача);
- относительная простота ручной верификации врачом предлагаемого в ответах лечения.

Итоговый список симптомов и заболеваний, удовлетворяющих всем вышеперечисленным критериям, содержит 13 элементов: *ОРВИ/грипп, насморк, ангина, диарея, отит, аллергия, гастрит, стоматит, кандидоз, герпес, геморрой, дисбактериоз, изжога.*

Словарь заболеваний дополнен синонимами нозологической группы с помощью справочника международной классификации болезней десятого пересмотра (МКБ-10) [34], проверен и скорректирован медицинским специалистом. Для повышения точности извлечения заболеваний из текстов СВОС словарь дополнен также и общеупотребительными наименованиями. Например, для насморка добавлен синоним *сопли.*

Некоторые из перечисленных выше заболеваний имеют слишком подробную внутреннюю классификацию в МКБ-10, тонкости которой становятся

неразличимы при общении пользователей СВОС, которые не являются в общем случае медицинскими профессионалами. Поэтому, для упрощения анализа такие группы классов были объединены в одном заболевании. Например, группа *ОРВИ/грипп* объединяет следующие классы МКБ-10:

- J06. Острые инфекции верхних дыхательных путей множественной и неуточнённой локализации.
- J10. Грипп, вызванный идентифицированным вирусом гриппа.
- J11. Грипп, вирус не идентифицирован.

На основе данных ГРЛС [33] составлен тематический словарь лекарственных средств. Особенность метода извлечения лекарства из текста подразумевает, что наименование препарата должно состоять из одного однозначно идентифицирующего термина.

Предварительная обработка словаря проведена вручную. В первую очередь удалены биологически активные добавки, витаминные комплексы, детское питание, априори не являющиеся средствами лечения выбранных заболеваний. Для каждого наименования препарата выполнены следующие шаги:

1. удалить фармацевтическую форму (*капли, крем, мазь, таблетки, порошок, микстура* и т.п.);
2. удалить модификатор лекарственного средства (например, наименования *Аспирин Кардио, 1000 Аспирин, Аспирин Йорк, Аспирин Экспресс* преобразуются в термин *Аспирин*);
3. выбрать из всех оставшихся в наименовании терминов один, наиболее точно идентифицирующий лекарственное средство (например, наименование *Доктор Мом* трансформируется в *Мом*).

После выполнения вышеперечисленных шагов все термины приведены к нижнему регистру, затем из словаря удалены дубли. В итоге, из 11 926 уникальных наименований отобрано 4 120. Нужно отметить, что лекарства, не имеющие в названии однозначно идентифицирующего термина, не попали в финальный словарь (например, препарат *морская соль* состоит из двух общеупотребительных слов, которые по отдельности не имеют отношения к лекарственному средству).

На основе словарей заболеваний, медикаментов и энциклопедии лекарств и товаров аптечного ассортимента [58] составлена таблица лекарственных средств, рекомендуемых врачами к применению при соответствующем заболевании. Определим ее формально как индикаторную функцию $IsCorrect$.

Пусть D — словарь заболеваний, M — словарь медикаментов, DM — множество всех возможных пар $\{(d,m)|d \in D, m \in M\}$. Определим DM_{true} как множество пар $(d,m) \in DM$, для которых справедливо следующее условие: $(d,m) \in DM_{true}$, если лекарственное средство m рекомендуется к применению при заболевании d согласно источнику [58]. Определим индикаторную функцию $IsCorrect : DM \to \{0,1\}$ следующим образом:

$$IsCorrect(d,m) = \begin{cases} 1, & (d,m) \in DM_{true} \\ 0, & (d,m) \notin DM_{true} \end{cases} \tag{2.2}$$

В итоге выбрано в среднем 126 лекарств на одно заболевание. Табл. 8 демонстрирует распределение числа медикаментов для выбранных заболеваний.

Таблица 8 — Количество уникальных наименований лекарственных средств, рекомендуемых к применению при соответствующем заболевании или симптоме

Заболевание	Количество лекарств
ОРВИ/Грипп	294
Насморк	260
Ангина	167
Диарея	155
Отит	149
Аллергия	118
Гастрит	106
Стоматит	90
Кандидоз	84
Герпес	64
Геморрой	57
Дисбактериоз	53
Изжога	46

2.5 Метод автоматической оценки качества данных СВОС

Для решения задачи, описанной в подразделе 2.1, предлагается метод автоматической оценки больших объёмов вопросов и ответов, удовлетворяющих шаблону $DisMed_k$. Метод основан на простой модели качества пар вопросов и ответов. Кроме того, при реализации метода используется таблица соответствия лекарств болезням, составленная на основе данных энциклопедии [58].

2.5.1 Модель качества пары «вопрос–ответ»

Задача, решаемая в данном разделе, ограничивает рассматриваемое множество вопросов и ответов до пар, удовлетворяющих паттерну $DisMed_k$ (формула 2.1), то есть вопросов о лечении заболеваний и ответов с упоминаниями лекарственных средств. Качественной парой в данном контексте будет считаться та, в которой упомянутые в ответе лекарства действительно рекомендованы к применению при упомянутом в вопросе заболевании. Другими словами, заболевание входит в нозологическую классификацию лекарственного средства. Определим модель качества формально.

Для каждой пары (q, a) из множества $\{(q, a)|DisMed_k(q, a) = TRUE\}$ извлекается заболевание $d \in D$, упоминаемое в вопросе q, и множество лекарств M_a, упоминаемых в ответе a. Тогда на основе функции (2.2) вводится следующая модель качества пары «вопрос–ответ» (q, a):

$$Quality(q, a) = \begin{cases} 1, & \sum_{m \in M_a} IsCorrect(d,m) > \frac{|M_a|}{2} \\ 0, & \sum_{m \in M_a} IsCorrect(d,m) \leqslant \frac{|M_a|}{2} \end{cases} \tag{2.3}$$

То есть, пара «вопрос–ответ» считается качественной, если корректных пар «болезнь–лекарство» больше, чем некорректных.

При таком подходе затрагивается только один из аспектов комплексной оценки качества вопроса и ответа, которую мог бы сделать медицинский специалист, поэтому нельзя говорить о качестве пары только на основании модели (2.3). Преимущество же автоматической оценки заключается в том, что она позволяет быстро оценить данные такого объёма, для которого ручная оцен-

Рисунок 2.4 — Пример ложно положительного результата извлечения пар с помощью паттерна $DisMed_2$

Вопрос:	У меня выскочил [*герпес*] на губе. Стоит ли идти в школу?
Ответ:	Хаха.. 90% населения имеют герпес и ходят в школу и на работу! Возьмите [*Кагоцел*] и [*Ацикловир*] в аптеке. Здоровья!

ка уже слишком дорога, и на основе большого числа «слабых» оценок сделать общий вывод о качестве всей коллекции. Кроме того автоматическая оценка позволяет сравнивать между собой по качеству разные наборы больших данных.

В контексте метода автоматической оценки качества остаётся невыясненным вопрос выбора параметра k в паттерне $DisMed_k$ (формула (2.1)). Для выбора k из коллекции данных СВОС были сэмплированы 1000 вопросов с ответами, удовлетворяющих паттерну $DisMed_1$. Ручная проверка пар «вопрос–ответ» показала, что среди пар, удовлетворяющих $DisMed_1$ только 53% пар соответствует выражению *«Как вылечить заболевание d?»*, которое и является основным свойством, моделируемым функцией (2.1) (см. подраздел 2.1). Так как критерий $DisMed_1$ показал слишком низкую для целей исследования точность, он был усилен до $DisMed_2$. При таком способе отбора уже 79% пар соответствует нужному выражению, однако $DisMed_2$ захватывает существенно меньше данных, чем $DisMed_1$ (255 вопросов из 1000 или 25,5%). Дальнейшее увеличение k сильно уменьшает число извлекаемых данных, не увеличивая точность значительно, поэтому, для проведения экспериментов выбрано $k = 2$, как обеспечивающее удовлетворительную точность моделирования, несмотря на то, что среди извлекаемых пар можно найти и ложно положительные (например, рис. 2.4).

2.5.2 Теоретическая оценка вычислительной сложности алгоритма

Алгоритм автоматической оценки качества пары «вопрос-ответ» медицинской тематики представлен на рис. 2.5. Покажем, что сложность обработки всей коллекции вопросов и ответов с помощью данного метода линейно зависит от числа ответов.

Рисунок 2.5 — Алгоритм вычисления функции $Quality(q, a)$

Вход: q, a, D, M, k

Выход: $Quality$

```
 1: Quality := 0
 2: if ∃!d ∈ D : q.Contains(d)
 3:   if ∃M_a ⊂ M : ∀m ∈ M_a a.Contains(m) & |M_a| ⩾ k
 4:     estimation := 0
 5:     for m ∈ M_a do
 6:       estimation := estimation + IsCorrect(d, m)
 7:     end for
 8:     if estimation > |M_a| / 2
 9:       Quality := 1
10:     end if
11:   end if
12: end if
```

Для обеспечения требуемой эффективности коллекция вопросов и ответов предварительно обрабатывается: проводится лемматизация или стемминг полного словаря коллекции и каждый документ преобразуется в хеш-таблицу слов (что является технической реализацией концепции «мешка слов», широко применяемой в области информационного поиска). Тогда функция $Contains$ представляет собой простую проверку элемента на принадлежность множеству. Реализация, основанная на хеш-таблицах, имеет среднюю сложность $O(1)$.

Технически функция $IsCorrect$ (формула (2.2)) также представляет собой проверку на принадлежность пары (d, m) таблице лекарственных средств, рекомендуемых к применению при определённых заболеваниях (пункт 2.4), что реализуется на основе хеш-таблиц. Таким образом, сложность функции $IsCorrect$ составляет $O(1)$ в среднем.

В силу единственности d (строка 2 рис. 2.5) алгоритм имеет только один проход по множеству D. Реализация проверок в строках 3 и 6 также подразумевает единственный проход по словарю M, так как все операции проверки зависят только от текущего элемента $m \in M$. Это значит, что на одной итерации можно выполнить проверку на принадлежность m множеству M_a, и, в

случае успеха, вычислить значение функции $IsCorrect$. Таким образом, общая вычислительная сложность функции $Quality$ равна $O(|D| + |M|)$ в среднем.

Множества D и M вычисляются один раз до начала обработки коллекции документов, поэтому величину $|D| + |M|$ можно считать константной. Для автоматической оценки качества всего множества вопросов (Q) и ответов (A) необходимо выполнить функцию $Quality$ для всех пар коллекции, число которых очевидно равно числу ответов. В итоге, вычислительная сложность будет возрастать линейно в зависимости от числа ответов: $O(c|A|)$, где константа $c \propto |D| + |M|$.

2.5.3 Результаты автоматической оценки данных Ответы@Mail.Ru

Согласно постановке задачи, описанной в подразделе 2.1, предложен метод (2.3) автоматической оценки больших объёмов данных. Этот подраздел описывает эксперимент по оцениванию вопросов и ответов СВОС Ответы@Mail.Ru, удовлетворяющих формуле (2.1).

В эксперименте использовались словари заболеваний и медикаментов, описанные в подразделе 2.4. Данные СВОС прошли предварительную обработку по алгоритму, описанному в пункте 1.4.1. Всего из 95 002 вопросов выделено 8 285 пар «вопрос–ответ», удовлетворяющих паттерну $DisMed_2$, для 13 болезней, на которых фокусировалось исследование.

Табл. 9 демонстрирует распределение оценок метода по заболеваниям (2.3) для 8 285 пар «вопрос–ответ», удовлетворяющих паттерну $DisMed_2$ (строки упорядочены по убыванию числа извлеченных пар с упоминанием определенного заболевания).

Доля положительных оценок метода варьируется от 0,69 у пар, упоминающих насморк и отит, до 0,91 у пар с упоминанием диареи. Предположительно более высокую оценку получают простые заболевания, которые требуют чаще всего только симптоматического лечения (диарея, герпес на губах) или заболевания, для которых существует меньше рекомендуемых при лечении лекарственных средств (см. табл. 8).

В среднем по заболеваниям метод дает положительную оценку в 80% случаев, что может свидетельствовать об общей адекватности и удовлетворитель-

Таблица 9 — Результаты автоматической оценки данных СВОС Ответы@Mail.Ru методом *Quality* (формула (2.3))

Заболевание	#	0		1	
		#	%	#	%
Насморк	1653	509	0,31	1144	0,69
Аллергия	926	211	0,23	715	0,77
Герпес	920	94	0,10	826	0,90
Ангина	889	143	0,16	746	0,84
Диарея	841	74	0,09	767	0,91
Кандидоз	771	95	0,12	676	0,88
ОРВИ/Грипп	513	88	0,17	425	0,83
Изжога	440	130	0,30	310	0,70
Дисбактериоз	420	63	0,15	357	0,85
Стоматит	359	68	0,19	291	0,81
Гастрит	269	76	0,28	193	0,72
Отит	147	45	0,31	102	0,69
Геморрой	137	22	0,16	115	0,84
Всего:	8285	1618	0,20	6667	0,80

Примечание:

1. # — число пар «вопрос–ответ».
2. % — доля пар с данной оценкой (0; 1).

ном качестве медицинских вопросов СВОС Ответы@Mail.Ru типа «Чем лечить заболевание d?». В следующих подразделах описываются эксперименты по верификации автоматического метода средствами ручного экспертного асессмента и экстраполяции оценки качества на все доступное множество медицинских вопросов и ответов.

2.6 Методология ручной экспертной оценки качества медицинских вопросов и ответов

Авторы современных систем обработки информации уделяют особое внимание оценке качества данных. Чаще всего к оценке привлекается большое число асессоров – людей, вручную оценивающих поисковую выдачу. Огромный объём обрабатываемой информации исключает возможность полной проверки качества имеющихся в системе данных, поэтому асессоры дают оценку некоторому подмножеству документов, на основании которой и делается вывод обо всей коллекции.

В диссертационной работе похожий подход применён к данным медицинской тематики: качество вопросов и ответов о здоровье человека оценивалось асессорами с высшим медицинским образованием. Для ручной оценки использовалось 977 пар, удовлетворяющих паттерну $DisMed_2$ (из 255 вопросов). Для того, чтобы оценить, как показатель качества переносится на все множество вопросов и ответов о здоровье человека, для асессмента было сэмплировано еще 500 пар из медицинских разделов СВОС без каких-либо дополнительных ограничений. Итоговое множество пар, оцениваемых асессорами, имело мощность 1477.

Для проведения асессмента были разосланы приглашения аспирантам и преподавателям Уральской государственной медицинской академии. В итоге, в процедуре ручной оценки приняли участие 7 человек: 6 аспирантов и 1 преподаватель — кандидат медицинских наук.

Инструмент ручной оценки (рис. 2.6) представляет из себя онлайн вебсервис, который спроектирован так, чтобы минимизировать порог вхождения и длину инструктажа асессоров (для экономии их времени). Отдельной страницей сервиса была опубликована подробная инструкция с примерами, доступ к которой был свободен на протяжении всего процесса оценки. Кроме того, каждый оценивающий пользователь идентифицировался адресом электронной почты и мог прервать и продолжить процесс в любое удобное время. Кроме того, от асессора не требовалось сделать какой либо обязательный минимум работы — каждый волен был выбрать комфортное для себя число оцениваемых пар. Это привело к тому, что финальное распределение оценок между асессорами неравномерно: 406 : 267 : 197 : 102 : 58 : 50 : 11.

Среди оцениваемых пар присутствовали группы ответов на один вопрос. Чтобы ответы на один и тот же вопрос не выдавались подряд, влияя на оценку текущей пары, очередь, выдаваемая асессорам, была перемешана случайным образом. В случае сомнений по поводу оценки какой либо пары, асессор имел возможность вернуться к оцениваемой паре позже и изменить свое решение.

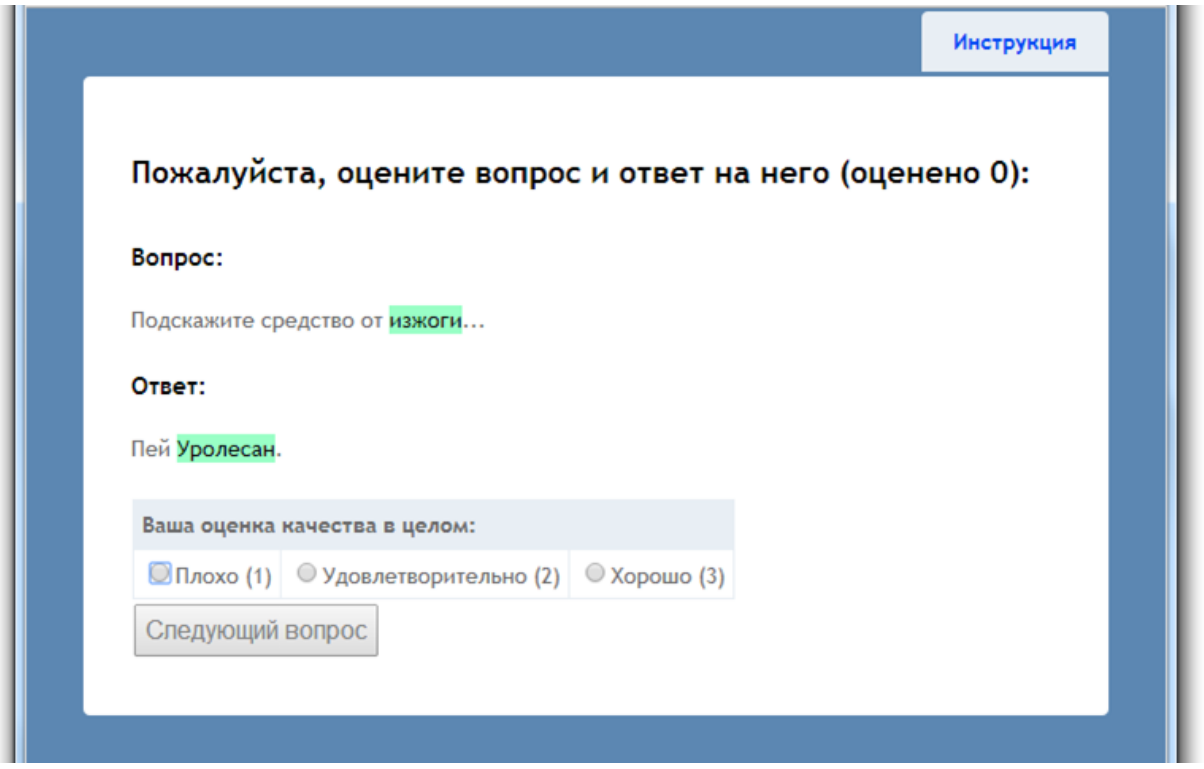
Асессору выдавалась пара «вопрос–ответ», причем, употребления заболевания в вопросе и медикаментов в ответе были выделены цветом. Требовалось оценить пару по шкале оценок, представленных в табл. 10.

Таблица 10 — Шкала оценок качества пар «вопрос-ответ»

Оценка	Описание
0	вопрос или ответ низкого качества (например, ответ может причинить вред)
1	вопрос высокого качества; ответ потенциально полезен (например, ответ может быть неполон)
2	вопрос и ответ высокого качества (точный, полный)

Левая часть табл. 11 (подраздел 2.7) демонстрирует распределение оценок асессоров по заболеваниям. Кроме того, в последней строке таблицы приведены

Рисунок 2.6 — Интерфейс инструмента ручной оценки



результаты оценки вопросов, сэмплированных из медицинских разделов СВОС случайно (без требования соответствия паттерну $DisMed_2$). Распределение ручных оценок по множеству пар «болезнь-лекарство» не отличается статистически значимо от распределения по случайной выборке из 500 пар (на основании критерия χ^2 Пирсона, $p - value = 0{,}05$). Это позволяет обобщить результаты ручной оценки подмножества на все множество вопросов и ответов медицинской тематики.

Для оценки меры согласованности решений, принимаемых асессорами, некоторые пары выдавались на оценку двум различным асессорам. В качестве меры согласованности использовалась каппа-статистика Коэна (см. пункт 1.2). Из 100 пар оценки асессоров совпали в 21% случаев, $\kappa = 0{,}51$, что показывает приемлемый, но достаточно низкий уровень согласованности.

По завершении эксперимента ручной оценки были проведены беседы с наиболее активными асессорами для получения более детальной картины результатов. В целом, медицинские специалисты считают интернет ненадёжным источником информации о здоровье человека. В определённом смысле интернет-ресурсы воспринимаются врачами как некий конкурирующий сервис — известно много случаев, когда пациент приходит на приём к специалисту, уже имея на руках «диагноз», полученный после знакомства с материалами в сети. Возможно, вследствие этого факта медицинские специалисты склонны давать ответам СВОС вида *«Обратитесь к врачу»* в среднем более высокие оценки. Кроме того, врачи часто воспринимают вопрос абстрактно, предполагая больше, чем написано в вопросе, и могут иметь индивидуальное мнение о применимости и эффективности той или иной терапии. В целом, специалисты склонны занижать общую оценку, руководствуясь принципом *primum non nocere* («прежде всего — не навреди»). В то же время, возможны и обратные ситуации: врач может дать ненулевую оценку ответу, который в общем не является корректным лечением заболевания, указанного в вопросе, но «не навредит» — например, совет придерживаться здорового образа жизни.

Таким образом, можно заключить, что область медицины и здоровья человека является достаточно специфической темой, и довольно сложно провести ручную оценку с высоким уровнем согласованности асессоров и минимальными инвестициями в обучение и тренировку медицинских специалистов. Даже в простых оценочных сценариях требуется большая подготовительная работа,

подробный инструктаж и совместное оценивание нескольких случаев с последующим обсуждением оценок для повышения согласованности.

2.7 Сравнение автоматической и ручной оценки

В данном подразделе приводится сравнение результатов ручной и автоматической оценки. Мнение медицинского специалиста совпало с автоматической оценкой в 60% случаев, что показывает достаточно низкий уровень согласованности ручной и автоматической оценки. Это могло быть обусловлено несколькими факторами. Во-первых, в автоматической оценке использовался только один критерий качества (соответствие лекарства заболеванию), тогда как медицинский специалист оценивает данные комплексно, используя одновременно множество критериев, на основании образования и личного опыта. Во-вторых, причиной мог послужить и низкий уровень согласия самих асессоров.

Правая часть табл. 11 содержит оценки качества, полученные автоматическим методом $Quality(q, a)$ для тех же пар, на которых проводилась ручная оценка. Колонка $M \cap A$ показывает количество и долю совпадений показателей ручной и автоматической оценки по каждому заболеванию (ручные оценки приводились к автоматическим по следующему правилу: $0 \to 0; \{1,2\} \to 1$). Учитывая это правило, можно заключить, что из 977 пар, оцененных вручную, 63% содержат вопрос хорошего качества и потенциально полезный или хороший ответ согласно мнению врачей и 83% содержат «корректные» ответы согласно автоматической оценке. Таким образом, оценка автоматического метода оказывается достаточно завышенной относительно оценки медицинских специалистов.

2.8 Анализ случаев несогласия методов

Сравнение результатов автоматической и ручной оценки выявил несколько типичных случаев несогласия методов.

Таблица 11 — Сравнение ручной и автоматической оценки

Заболевание	#	Ручная (M)			Авто (A)		$M \cap A$	
		0	1	2	0	1	#	%
Насморк	182	0,37	0,38	0,24	0,33	0,67	112	0,62
Аллергия	149	0,54	0,26	0,19	0,16	0,84	56	0,38
Кандидоз	148	0,32	0,45	0,23	0,12	0,88	96	0,65
Герпес	136	0,36	0,42	0,22	0,11	0,89	88	0,65
Ангина	107	0,21	0,47	0,32	0,14	0,86	79	0,74
Диарея	70	0,39	0,40	0,21	0,10	0,90	44	0,63
Изжога	53	0,32	0,38	0,30	0,09	0,91	33	0,62
Стоматит	48	0,38	0,50	0,13	0,15	0,85	31	0,65
Дисбактериоз	28	0,57	0,21	0,21	0,07	0,93	14	0,50
ОРВИ/Грипп	24	0,29	0,46	0,25	0,08	0,92	17	0,71
Отит	13	0,23	0,31	0,46	0,38	0,62	9	0,69
Гастрит	12	0,42	0,33	0,25	0,25	0,75	6	0,50
Геморрой	7	0,29	0,43	0,29	0,43	0,57	4	0,57
Итого	**977**	**0,37**	**0,39**	**0,24**	**0,17**	**0,83**	**589**	**0,60**
Случайные	500	0,41	0,40	0,19				

Смена диагноза. Пользователь в вопросе описывает свои симптомы и предполагает некоторый диагноз; отвечающий пользователь ставит другой диагноз и предлагает лечение от вновь предложенного заболевания (рис. 2.7). Если диагноз, предложенный в ответе, верен, то медицинский специалист правильно поставит паре хорошую оценку, тогда как автоматический метод оценит пару на 0, имея в виду диагноз, упомянутый в вопросе.

Рисунок 2.7 — Пример смены диагноза в ответе

Вопрос:	У меня очень странная [*аллергия*]!!! Маленькие волдыри по всему телу, очень зудят! Помогите!
Ответ:	<...> Если волдыри вылазят в области царапин, тогда у вас кожная инфекция. <...>

Субъективная оценка лекарств в ответе. Упоминание лекарства не обязательно означает рекомендацию к применению при заболевании, указанном в вопросе. Довольно частой ситуацией является реклама (спам) лекарств нетрадиционной медицины, сопровождаемая критикой лекарственных препаратов, обычно назначаемых врачом. Пользователи могут сомневаться в эффективности лекарства или даже предостерегать от его использования (рис. 2.8). Подобным случаям врачи дают отрицательную оценку, в то время, как автоматический метод делает ложно положительное срабатывание.

Множество лекарств в ответе. Функция *Quality* (формула (2.3)) даёт бинарную оценку на основе большинства корректных или некорректных пар заболеваний и лекарств (формула (2.2)). Такая оценка очевидно не может всесторонне отразить качество ответа на вопрос. Пользователи часто предлагают несколько препаратов, которые помогут, по их мнению, комплексно справиться с заболеванием. Например, в случае герпеса может быть рекомендован крем против герпеса на губе и иммуномодуляторы; при отите — антибиотики наряду с пробиотиками для борьбы с дисбактериозом как возможным побочным эффектом лечения антибиотиками.

2.9 Выводы

Диссертационное исследование демонстрирует, что методы, эксплуатирующие в своей работе доменную специфику (то есть особенности тематики дан-

Рисунок 2.8 — Примеры субъективной оценки и предостережения от использования лекарства

Вопрос:	Как избавиться от [*герпеса*]?
Ответ:	[*Ацикловир*] вообще не помогает.
Вопрос:	Что может помочь ребенку при [*диарее*]? (2 с половиной года) Неделю назад пропили курс антибиотиков
Ответ:	<...> Вам не хватило этой химии? Все эти лекарства [*Линекс*], [*Хилак*] форте и т.п. сделаны химическим путем из неорганических соединений.

ных), значительно углубляют понимание данных, позволяя достичь большей производительности по сравнению с методами анализа текстов без привязки к конкретной теме.

В настоящей главе предложен метод *Quality* автоматической оценки качества данных медицинских разделов СВОС (формула (2.3)). Автоматический метод верифицировался с помощью ручной оценки подмножества вопросов и ответов медицинскими специалистами. Реализован комплекс программ: реализация автоматического метода, программное обеспечение для ручного асессмента. В ходе эксперимента врачами оценено порядка 1500 пар «вопрос-ответ». Полученный набор данных опубликован в свободном доступе для исследовательских целей[1]. Он позволяет производить верификацию и сравнение автоматических методов оценки качества вопросов и ответов.

Результат оценки медицинских разделов СВОС свидетельствует о достаточно хорошем качестве вопросов и ответов. Данные в целом пригодны для переиспользования (возможно, с использованием простых методов фильтрации очевидно некачественного контента). В качестве дальнейших направлений исследования рассматриваются:

- анализ тональности текстов для определения полярности оценки медицинских препаратов (позитивная/нейтральная/негативная);
- более точный выбор данных для автоматической оценки: например, переход от уровня вопросов к уровню предложений — классификация их на описания болезней и описания лекарств по аналогии с [59].

Существует несколько сценариев, при которых полезно повторное использование данных СВОС: поиск по ответам, автоматический ответ на вновь задаваемый вопрос, поиск похожих вопросов. Помимо непосредственной задачи автоматической оценки качества контента, предложенный подход и его модификации могут быть применены, например, в исследованиях об употреблении лекарственных средств или в методах выявления заблуждений пользователей о лечении тех или иных заболеваний.

[1] http://kansas.ru/cqa/data2/

Глава 3. Модель компетентности пользователя медицинских разделов СВОС

Качество контента, генерируемого произвольным пользователем СВОС (UGC), может быть оценено как через анализ самого контента, так и путём ответа на вопрос, является ли автор экспертом в предметной области. Данная глава описывает задачу автоматического вычисления *степени компетентности* (то есть квалификации в определённой сфере) пользователя медицинских разделов СВОС.

3.1 Опросы активных пользователей социальных онлайн-сервисов медицинской тематики

В качестве предварительных экспериментов в рамках диссертационной работы проведено два онлайн-опроса, основной задачей которых было выяснить отношение общества к распространению медицинской информации в сети, приобретающему со временем всё больший размах. Кроме того, автора интересовало, что мотивирует людей отвечать на вопросы о здоровье человека в интернете.

Первый опрос проводился среди врачей — пользователей профессионального сообщества «Доктор на работе» (вопросы анкеты приведены в приложении А). Сообщество является закрытым — для регистрации требуется предоставить диплом о высшем медицинском образовании.

Среди прочих врачам был задан вопрос о том, знают ли они, что такое социальные вопросно-ответные сервисы, и если да, то являются ли их пользователями. Большинство медицинских специалистов (78,8%) осведомлены о существовании вопросно-ответных сервисов. Более того, 28% респондентов отвечает на медицинские вопросы онлайн, мотивируя это желанием помочь людям и поделиться опытом.

Кроме того, респондентам задавался вопрос об их отношении к широкому распространению личного опыта и информации о здоровье рядовыми пользователями социальных онлайн-сервисов. Хотя значительная часть врачей относит-

ся к этой идее скорее положительно (41,2% из 85 участников), почти половина респондентов (48,2%) не одобряет её вследствие следующих причин:

– неспособность пользователя-непрофессионала в общем случае корректно интерпретировать ответ, полученный в сети;
– популярность методов «народного лечения», часто не имеющих отношения к доказательной медицине.

Результаты опроса врачей в целом подтверждают выводы главы 2 об общем скептицизме профессионального медицинского сообщества в отношении информации, свободно доступной на сайтах социальных сервисов медицинской тематики. Среди основных аргументов приводится также и тот факт, что потенциальный вред открытых данных такого рода может превалировать над возможной пользой.

Второй опрос проводился среди активных пользователей медицинских разделов вопросно-ответного сервиса Ответы@Mail.Ru (вопросы анкеты приведены в приложении Б). Основными мотивирующими факторами ответов на вопросы респонденты называют желание поделиться знаниями и опытом (аналогично результатам первого опроса); сочувствие болеющим людям; желание донести до широкой аудитории корректную медицинскую информацию.

По результатам опроса неожиданно высокая доля пользователей имеет медицинское образование (48% из 172 участников). Кроме того, среди врачей–участников первого опроса почти треть отвечает на медицинские вопросы хотя бы изредка. С учётом этих результатов имеет смысл искать и выделять среди пользователей СВОС профессионалов или тех, кто похож по своим ответам на медицинского специалиста. Поэтому, в рамках исследования была поставлена задача разработать метод автоматической оценки профессионализма пользователя медицинских разделов СВОС.

3.2 Постановка задачи

Рассмотрим пользователей социального вопросно-ответного сервиса. Назовём конкатенацию всех ответов i–го пользователя квазидокументом «пользователь» и обозначим как A_i. Будем считать пользователя *активным*, если число различных терминов в A_i больше некоторого порогового значения T_{active}.

Требуется разработать модель оценки компетентности активного пользователя i как функцию его квазидокумента:

$$expertise_i = f(A_i), |\{w|w \in A_i\}| \geqslant T_{active} \tag{3.1}$$

3.3 Обзор литературы

В отличие от главы 2, данный раздел рассматривает качество CBOC под другим углом: фокус сдвигается с контента на пользователей, генерирующих контент.

Результаты опроса о мотивации пользователей CBOC, полученные в рамках диссертационного исследования, согласуются с результатами аналогичного опроса, проведённого Рабаном и Харпером в работе [60]. Эта информация дополняется в источниках [61; 62], где авторы исследовали проблему, почему некоторые пользователи *не отвечают* на вопросы. Опрос 135 активных пользователей сервиса Yahoo! Answers выявил такие причины, как сомнение в том, что ответ будет правильно интерпретирован, и уверенность в том, что на подобные вопросы ранее дано достаточно ответов.

Сайленс и др. в работе [63] исследуют вопросы отношения пользователей к контенту CBOC и фокусируются на медицинских разделах. В работе описан парадокс: по результатам опроса, проведённого авторами [63], пользователи, высоко оценивая опыт авторов ответов, в то же время выказывают сомнения относительно надёжности и применимости информации, найденной в CBOC.

Поиск экспертов среди пользователей CBOC — задача, которая исследуется достаточно активно. Гуо и др. в [64] пытаются использовать потенциал социальных вопросно-ответных сообществ, определяя конкретных пользователей, способных ответить на вновь задаваемые вопросы. Работы [65; 66] решают задачу поиска экспертов с помощью Гауссовских моделей классификации и других вероятностных методов, учитывающих связи между пользователями и похожесть или непохожесть тем вопросов.

Задача определения степени компетентности пользователя медицинских разделов CBOC, описываемая в настоящей главе, может также быть рассмотрена в виде более общей проблемы поиска экспертов. Подобные задачи реша-

ются в современных исследованиях для многих профессий. Например, авторы работ [67; 68] пытаются автоматически выделять среди пользователей социального сервиса Twitter учёных и журналистов соответственно. Методы, описываемые в статьях, используют открытые профили пользователей, связи между ними (чтение, ретвиты и упоминания), и обучающее множество пользователей, для которых достоверно известна нужная профессия. Эти данные характерны для Twitter, но не для СВОС, поэтому описываемые методы не могут быть применены в задаче поиска медицинских профессионалов в вопросно-ответном сервисе. Кроме того для данных, на которых решалась задача, отсутствует обучающее множество пользователей-врачей, отвечающих на вопросы. Поэтому в качестве решения задачи предложен метод, в основе которого лежит две составляющих, которые используют имеющиеся данные: тематический фокус пользователя и его лексикон.

3.4 Метод оценки компетентности пользователя СВОС

3.4.1 Модель тематического фокуса пользователя СВОС

Решение задачи построения модели тематического фокуса пользователя предполагает восстановление множества тем, которые затрагивает данный пользователь в своих ответах.

Практически все СВОС имеют подразделение вопросов на тематические категории, из которых пользователь может выбрать подходящую, когда задаёт вопрос. Несмотря на разнообразие, одна тематическая категория может охватывать тему слишком широко. Например, раздел 2-го уровня СВОС Ответы@Mail.Ru *Болезни, лекарства* включает в себя вопросы обо всех заболеваниях и медикаментозных препаратах, включая и «народные» методы лечения. Очевидно, что настолько грубое тематическое обобщение не позволяет понять картину интересов конкретного пользователя. Кажется естественным, что отдельные вопросы затрагивают более узкий набор тем. Например, в рамках раздела *Болезни, лекарства* может быть задан вопрос на тему гриппа.

Таким образом, в задаче построения модели тематического фокуса возникает потребность восстановить тематическое распределение ответов пользователя, имеющее большую степень гранулярности по сравнению с тематическими разделами СВОС. В диссертационном исследовании такое распределение восстанавливалось с помощью тематических моделей.

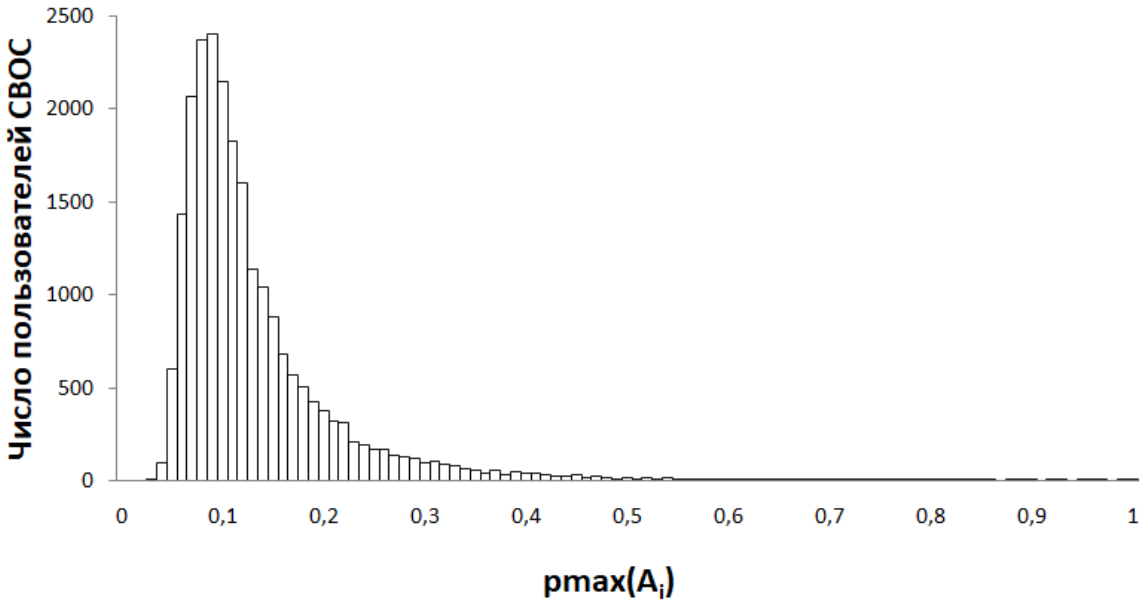
Пусть W — множество всех терминов коллекции квазидокументов $\{A_1,A_2,\ldots,A_N\}$. Предполагается, что существует конечное множество тем T, и употребление каждого термина $w \in W$ в каждом квазидокументе A_i связано с некоторой заранее неизвестной темой $t \in T$. Тогда можно определить вероятностную тематическую модель порождения данных $P(w|A_i) = \sum_{t\in T} P(t|A_i)P(w|t)$, как описано в пункте 1.3.

Взяв за основу тематическую модель порождения квазидокументов, построим функцию, оценивающую тематический фокус i–го пользователя как максимальную вероятность того, что слова, генерирующие квазидокумент A_i, взяты из некоторой темы $t \in T$:

$$pmax(A_i) = max_{t\in T} P(t|A_i) \tag{3.2}$$

Рис. 3.1 показывает распределение значений $pmax$ среди активных пользователей медицинских разделов СВОС Ответы@Mail.Ru. Можно отметить, что для 95% квазидокументов $pmax(A_i) < 0{,}3$. При уменьшении показателя максимальной вероятности распределение тем у пользователя стремится к равномер-

Рисунок 3.1 — Распределение показателя $pmax$ среди активных пользователей медицинских разделов СВОС



ному. Очевидно, что пользователя, чьё тематическое распределение достаточно равномерно (то есть, максимум вероятности не отличается значимо от среднего), нельзя считать сосредоточенным на одной теме. Поэтому, для выделения искомых пользователей вводится пороговое значение T_{focus}. Предполагается, что пользователь достаточно сосредоточен на одной теме, если значение функции $pmax$ превышает T_{focus}. Тогда итоговая модель тематического фокуса пользователя выражается формулой 3.3.

$$focus(A_i) = \begin{cases} pmax(A_i), & pmax(A_i) \geqslant T_{focus} \\ 0, & pmax(A_i) < T_{focus} \end{cases} \tag{3.3}$$

3.4.2 Примеры тем, экстремальных по числу пользователей и среднему рейтингу

Для определения тем, привлекающих высококвалифицированных пользователей, были выявлены самые частые темы, на которых сосредоточено внимание пользователей вообще, и темы, на которых концентрируются пользователи, имеющие самые высокие оценки согласно рейтинговой системе СВОС[1].

Таблица 12 — Топ-3 тем по числу пользователей N, сфокусированных на теме и среднему рейтингу R сосредоточенных на теме пользователей

Тема (5 самых вероятных слов)	N	R
столовая [ложка], сок [растения], трава, стакан, лист	124	15,22
продукты, диета, овощи, мясо, рыба, фрукты	41	18,36
головной, мозг, нарушение, расстройство, нервный	32	15,70
беременность, гинеколог, менструация, тест, таблетки	12	27,07
малыш, месяц, кормление, молоко, педиатр	2	23,67
диабет, железа, сахар, гормон, норма	9	22,90

Таблица 12 демонстрирует, что на темах «беременность», «педиатрия», «гормональные заболевания» сосредоточены пользователи с наиболее высоким

[1] Детали реализации эксперимента и описание рейтинговой системы СВОС приведены в подразделе 3.5

средним рейтингом. Следовательно, среди пользователей, отвечающих на вопросы данных тематик, с большей долей вероятности можно встретить экспертов.

3.4.3 Оценка разнообразия медицинского лексикона

Человек, имеющий медицинское образование, склонен употреблять в ответах профессиональные медицинские термины чаще, чем их употребляет рядовой пользователь. Наряду с моделированием тематического фокуса пользователя, при оценке степени компетентности предлагается оценивать долю медицинских терминов в его лексиконе.

Для анализа лексикона пользователей был сформирован словарь медицинских терминов W_{med}. Схема алгоритма автоматического формирования словаря представлена на рис. 3.2. Источниками медицинской лексики послужили 10-я редакция международной классификации болезней [34], государственный реестр лекарственных средств России [33] и справочник фельдшера [32].

Данные источники предназначены для широкого круга лиц и содержат определения медицинских концепций, описания клинических случаев и разъяснения способов лечения, которые не могут быть выражены только медицинскими терминами, и содержат также стоп-слова и общую лексику. Например, фрагмент «... продуктов, в которых *тифопаратифозная инфекция* сохраняется...» содержит только 2 медицинских термина (выделены курсивом). Остальные слова не нужны для целей эксперимента и должны быть отброшены.

Для очистки W_{med} от общеупотребительных слов применялась фильтрация словарём общей лексики. Из начального множества вычиталось множество общеупотребительных слов. В качестве словаря общей лексики был взят набор наиболее частотных словоформ национального корпуса русского языка [31], усечённый по частоте 3.

Для эффективной работы с корпусом, все вхождения словаря были нормализованы методом стемминга. В качестве реализации использовалась версия алгоритма Snowball для русского языка (подробнее стемминг описан в пункте 1.4.3).

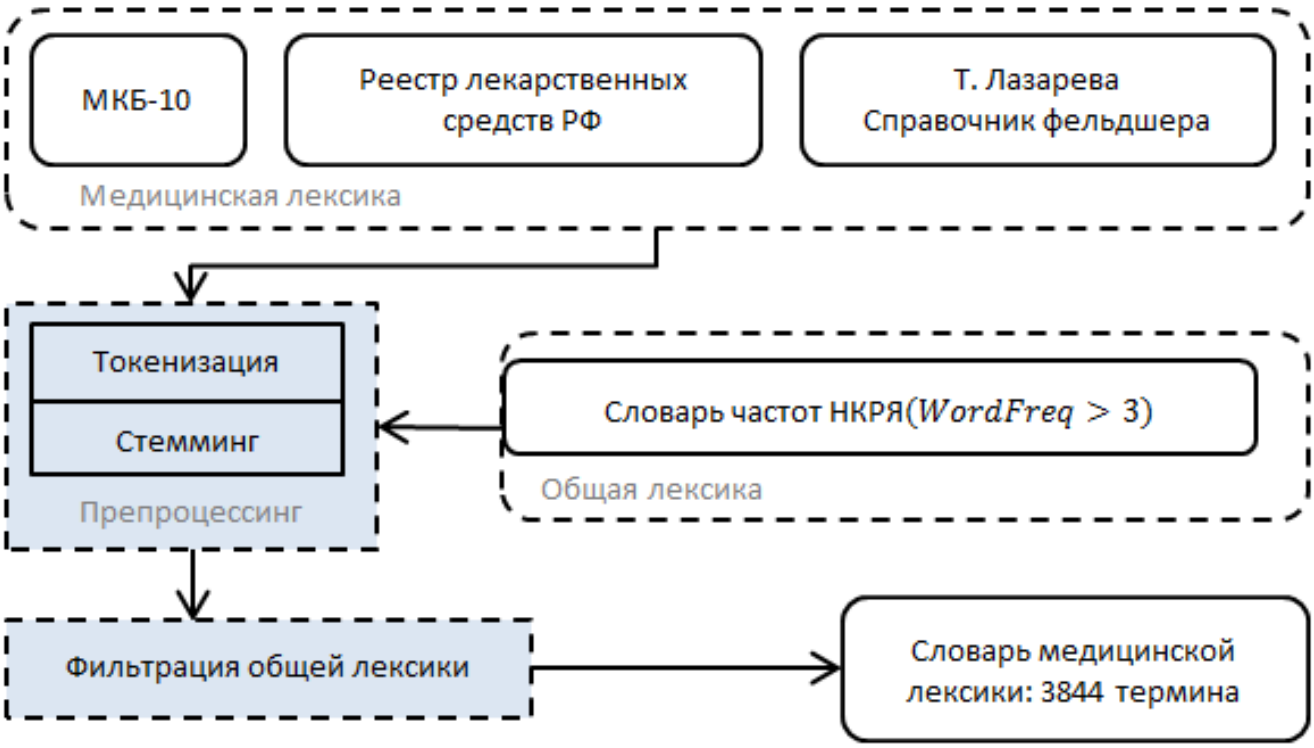
Пусть W_i – множество различных терминов пользователя i, хотя бы раз встречающихся в его ответах, W_{med} – словарь медицинских терминов, универсальный для всех пользователей. Для выявления наиболее эффективного способа оценки разнообразия медицинского лексикона пользователя СВОС сравнивались следующие показатели:

– доля медицинских терминов в лексиконе пользователя;
– разнообразие медицинского лексикона — используемая доля словаря W_{med};
– мера Жаккара для множеств W_i и W_{med}: $J(W_i,W_{med})$ (формула (3.4));
– абсолютное значение количества медицинских терминов в лексиконе пользователя (уникальных или с повторениями).

$$J(P,Q) = \frac{|P \cap Q|}{|P \cup Q|} \tag{3.4}$$

Кроме того, проводились эксперименты с объединением нескольких показателей в различные комбинации. Например, согласно источнику [11] для получения более надёжной оценки имеет смысл объединить долю медицинских терминов в лексиконе пользователя с показателем полноты его медицинского лексикона с помощью F-меры по определению, данному в разделе 1.2.

Рисунок 3.2 — Схема построения словаря медицинской лексики



Из всех способов анализа лексикона, приведённых выше, только доля медицинских терминов в лексиконе пользователя показывает приемлемый результат. Полнота медицинского лексикона пользователя зависит от объёма его общего лексикона. Чем больше пользователь отвечает на вопросы, тем более разнообразным является его общий лексикон. Следовательно, полнота медицинского лексикона также находится на высоком уровне. Такой способ оценки плохо работает для неактивных пользователей, редко отвечающих на вопросы. Стандартной мере Жаккара (формула (3.4)) присущ тот же недостаток. Кроме того, для пользователей, имеющих достаточно большой лексикон, множеством W_{med} в $J(W_i,W_{med})$ можно пренебречь — показатель достаточно точно аппроксимируется долей медицинских терминов. Абсолютные значения показывают плохую производительность, так как не позволяют сравнить двух пользователей со значимо различающимися объёмами используемых лексиконов.

В силу причин, перечисленных выше, в качестве оценки разнообразия медицинского лексикона пользователя i предлагается использовать долю медицинских терминов в общем лексиконе этого пользователя:

$$medlex(A_i) = \frac{|W_i \cap W_{med}|}{|W_i|} \tag{3.5}$$

Тогда оценивать компетентность пользователя предлагается по формуле:

$$expertise_i = focus(A_i)medlex(A_i) \tag{3.6}$$

Таким образом, оценка компетентности пользователя прямо зависит от двух факторов: разнообразие его медицинского лексикона и степень сосредоточенности на одной теме.

Таблица 13 — Топ-3 тем по среднему показателя *expertise* (формула (3.6))

Тема (5 самых вероятных слов)	N	R	$expertise$
недостаточный, мозг, кислород, кроветворение, сердце	4	13,32	0,16
нос, тепло, раствор, ингаляция, молоко	6	10,83	0,14
почка, массаж, паразит, желчный, кишечник	3	7,36	0,13

Табл. 13 показывает 3 темы с наиболее высокой средней степенью компетентности пользователей, сосредоточенных на данных темах.

3.4.4 Теоретическая оценка вычислительной сложности метода

Наиболее сложной с вычислительной точки зрения частью метода является обучение тематической модели для последующего вычисления формулы (3.3). Вычислительная сложность вариационных алгоритмов, которые обычно применяются при обучении тематических моделей, в общем случае оценивается как $O(T|D||V|)$ [69], где D — коллекция текстовых документов, V — словарь коллекции и T — число тем. Примером может служить классический EM-алгоритм [14; 15]. Несмотря на то, что в последнее время разработаны различные модификации, позволяющие оптимизировать алгоритмы обучения как по времени, так и по используемой алгоритмом памяти [70; 71], данный процесс имеет недостаточно высокую производительность и потому проводится один раз перед выполнением основного цикла вычисления общей формулы (3.6).

С учётом рассуждений, приведённых выше, предлагается следующий алгоритм (рис. 3.3) оценки компетентности i-го пользователя на основе квазидокумента A_i его ответов, словаря W_{med} медицинских терминов, матрицы $\Theta : |U| \times |T| \to \mathbb{R}$ распределений тем для каждого пользователя ($|U|$ — общее число пользователей, $|T|$ — число тем, $\theta_{i,j} = p(t_j|A_i)$ — доля j-й темы в ответах i-го пользователя) и порогового значения T_{focus}.

Покажем, что сложность оценки компетентности множества U всех пользователей вопросно-ответной системы в среднем равна $O(\sum_{i=1}^{|U|} |A_i|)$.

Выполнение алгоритма на рис. 3.3 предполагает проверку на принадлежность слов множествам W_i и W_{med} в цикле по всем словам в ответах пользователя. При реализации множеств с помощью хеш-таблиц, проверка на принадлежность имеет сложность $O(1)$ в среднем, поэтому время обработки одного пользователя линейно зависит от числа слов в его ответах. Подсчёт переменной $Focus$ предполагает вычисление функции максимума, которая имеет линейную сложность относительно размера входного массива чисел. В нашем случае вычисление $Focus$ линейно зависит от числа тем. Следственно, вычислительную сложность алгоритма на рис. 3.3 можно оценить как $O(|T| + |A_i|)$.

На практике число тем не превышает нескольких тысяч (подбор подходящего числа тем не является предметом исследования в рамках диссертации, поэтому в экспериментах использовалось значение по умолчанию: $|T| = 100$), поэтому его можно считать константой. Более того, переменную $Focus$ необхо-

Рисунок 3.3 — Алгоритм оценки компетентности i-го пользователя

Вход: $A_i, W_{med}, \Theta, T_{focus}$

Выход: $Expertise$

1: $Focus := \max\limits_{j \in [1,|T|]} \theta_{i,j}$
2: **if** $Focus < T_{focus}$
3: $\quad Focus := 0$
4: **endif**
5: $userMedWordsCount := 0$
6: $W_i := \{\}$
7: **for** $word \in A_i$ **do**
8: $\quad$ **if** $word \notin W_i$
9: $\quad\quad W_i.Add(word)$
10: $\quad\quad$ **if** $word \in W_{med}$
11: $\quad\quad\quad userMedWordsCount := userMedWordsCount + 1$
12: $\quad\quad$ **endif**
13: $\quad$ **endif**
14: **endfor**
15: $Medlex := \dfrac{userMedWordsCount}{|W_i|}$
16: $Expertise := Focus\ Medlex$

димо вычислять только один раз для каждого пользователя, а число пользователей в реальных вопросно-ответных системах обычно на несколько порядков меньше, чем количество слов в их ответах. Таким образом, $O(|T| + |A_i|) = O(|A_i|)$, тогда суммарную сложность метода оценки компетентности для всех пользователей СВОС можно оценить как $O(c|U| + \sum_{i=1}^{|U|} |A_i|) = O(\sum_{i=1}^{|U|} |A_i|)$.

Вычисление формулы 3.6 для нового $(|U| + 1)$-го пользователя системы требует предсказать для его ответов распределение тем, что имеет сложность $O(|A_{(|U|+1)}|)$, а затем выполнить алгоритм на рис. 3.3. Поэтому, сложность данной операции можно оценить как $O(c|A_{(|U|+1)}|)$

3.5 Оценка качества метода

Для оценки модели *expertise* ранжирование пользователей данным методом сравнивалось с ранжированием согласно стандартному рейтингу пользователей внутри CBOC. Основной вопрос при оценке качества модели формулировался следующим образом:

Насколько выше в среднем метод оценки профессиональной принадлежности ранжирует медицинских специалистов среди всех пользователей CBOC по сравнению с ранжированием согласно стандартному рейтингу сервиса?

Стандартный рейтинг пользователя CBOC Ответы@Mail.Ru, называемый внутри сервиса *КПД* (коэффициент полезного действия), обозначается в данной работе как *rating* и для i—го пользователя определён CBOC по формуле 3.7.

$$rating_i = \frac{best_i}{n_i} \cdot 100\% \tag{3.7}$$

Здесь n_i — общее число ответов i—го пользователя, $best_i$ — число его ответов, выбранных лучшими ответами на соответствующий вопрос с помощью голосования сообщества сервиса[2]. Это, фактически, единственный способ оценки уровня экспертизы пользователя, предоставленный сервисом на момент получения данных.

Таким образом, задача оценки предлагаемого в диссертационной работе метода — показать, превосходит ли он значимо по качеству ранжирования стандартный рейтинг CBOC. Для сравнения качества ранжирования использовался показатель *макроусредненной средней точности (MAP)*, наиболее широко используемый в современных исследованиях (полное определение MAP дано в разделе 1.2).

Для выполнения задачи оценки необходимо иметь множество пользователей, для которых достоверно известна их профессиональная принадлежность к медицинской области — тех, которые при вычислении MAP будут считаться релевантными. Насколько известно автору, в настоящий момент не существует стандартных тестовых коллекций для оценки методов решения задачи, рассматриваемой в данном разделе. По этой причине в рамках диссертационного иссле-

[2]https://help.mail.ru/otvety-help/others/kpd

дования по имеющимся данным СВОС было составлено собственное тестовое множество пользователей, с большой вероятностью являющихся медицинскими профессионалами.

3.5.1 Извлечение тестового множества медицинских специалистов

На начальном этапе исследования [7] в тестовое множество попадали пользователи, указавшие на свою врачебную специальность в одном из опросов, описанных в разделе 3.1 (83 человека). Такой размер множества не является достаточным для получения надёжной оценки качества метода. Проблема осложняется тем, что информация об образовании и профессии не является обязательной при заполнении регистрационного профиля пользователя СВОС. Несмотря на то, что некоторая часть пользователей указывает образование или профессию, эта информация отсутствует в открытом доступе. Поэтому, в рамках диссертационной работы решено было собрать подобную информацию о пользователях СВОС по косвенным признакам и сформировать тестовое множество самостоятельно.

Каждый пользователь при ответе на вопрос в рамках СВОС имеет возможность указать источник своего ответа. В процессе исследования было замечено, что многие пользователи оставляют в качестве источника информации выражения с указанием на врачебную специальность, например «высшее медицинское образование», «опыт работы врачом», «я сам педиатр».

Для извлечения пользователей, оставляющих в источниках ответов подобные выражения, на основе источника [72] был составлен словарь W_{doc} терминов, указывающих на профессию врача. Пользователи извлекались в несколько этапов:

1. Извлечение всех полей «Источник ответа», содержащих термины словаря W_{doc}.
2. Отсечение нерелевантных источников с помощью ручной фильтрации (убирались, например, выражения «недавно ходил к врачу», «знакомый педиатр подсказал»).
3. Извлечение ответов пользователей, указывавших выбранные источники.

4. Составление для каждого пользователя i соответствующего квазидокумента A_i.
5. Отсечение неактивных пользователей, то есть тех, для которых выполняется условие $|\{w|w \in A_i\}| < T_{active}$.

Кроме того, извлекались пользователи, указавшие на свою специальность в опросах, проведённых в рамках исследования. Таким образом, из пользователей, хотя бы раз указавших на свою врачебную специальность в опросах или источниках ответов, было составлено тестовое множество из 762 пользователей, которые с большой вероятностью являются медицинскими специалистами. Обозначим данное множество как $U_{doctors}$.

3.5.2 Численный эксперимент

При реализации модели тематического фокуса пользователя (формула 3.3) в качестве базовой тематической модели в диссертационном исследовании использовалось латентное размещение Дирихле с параметрами по умолчанию $\alpha = 0{,}5$ и $\beta = 0{,}1$; вывод апостериорных распределений осуществлялся методом сэмплирования Гиббса, как дающий достаточно точный результат при меньшей вычислительной сложности по сравнению с другими методами вероятностного вывода, например, EM-алгоритмом.

Эксперименты проводились на данных СВОС Ответы@Mail.Ru (подробное описание см. в подразделе 1.4.1). Так как методы, разработанные в рамках диссертационного исследования, опираются по большей части на анализ текстов, генерируемых пользователями, для экспериментов из всего множества пользователей медицинских разделов СВОС (225 427 уникальных авторов) были отобраны достаточно активные участники — те, чьи ответы содержат по крайней мере 50 различных нормализованных слов (23 259 авторов). Обозначим множество таких пользователей как U. Из-за большого объёма анализируемых данных в качестве метода нормализации слов вместо лемматизации применялся стемминг (подраздел 1.4.3).

В экспериментах качество ранжирования методами *rating* и *expertise* сравнивалось для разных пороговых значений $T_{focus} \in (0; 1)$ для определения оптимального порогового значения. Для каждого T_{focus} из множества U извле-

калось подмножество $U_{T_{focus}}$ тех, для кого верно неравенство $pmax(A_i) \geqslant T_{focus}$. Так как $T_{focus} > 0$ и по формуле (3.3) $focus(A_i) = pmax(A_i)$, то $focus(A_i) > 0$. Для остальных пользователей ($U \setminus U_{T_{focus}} = \{A_i | pmax(A_i) < T_{focus}\}$) значение $focus$ по формуле (3.3) обращается в 0. Поэтому, согласно формуле (3.6), значение функции $expertise$ для $U \setminus U_{T_{focus}}$ обращается в 0, то есть сравнение методов ранжирования для этого множества не имеет смысла. Поэтому эксперименты для каждого T_{focus} проводились на соответствующем множестве $U_{T_{focus}}$.

Для получения надёжной оценки для каждого T_{focus} эксперимент по сравнению ранжирования запускался 100 раз на случайных подмножествах тестовых пользователей $U_{randdocs} \subset U_{doctors} \cap U_{T_{focus}}$, которые сэмплировались с соблюдением условия (3.8).

$$|U_{randdocs}| \approx \frac{1}{2}|U_{doctors} \cap U_{T_{focus}}| \tag{3.8}$$

По результатам каждого запуска эксперимента считалась средняя точность AP ранжирования множества $U_{T_{focus}}$, в котором релевантными считались пользователи из $U_{randdocs}$. Затем AP усреднялась по 100 запускам, давая в результате итоговую оценку MAP для данного T_{focus}.

Из того, что MAP по определению — это усреднённое значение AP для множества запусков эксперимента (см. раздел 1.2), следует, что для проверки статистической значимости отличий в показателях MAP для каждого значения T_{focus} достаточно провести тест на значимость различия средних для выборок AP. Так как выборки AP распределены нормально (проверено с помощью критерия Шапиро-Уилка, уровень значимости 0,05), то в качестве теста на статистическую значимость различия средних использовался критерий Стьюдента.

Табл. 14 показывает значения MAP методов $rating$ и $expertise$ для $T_{focus} \in [0{,}05; 0{,}2]$, перебираемого с шагом 0,01. Таблица демонстрирует, что, при $T_{focus} \geqslant 0{,}11$ ранжирование методом $expertise$ статистически значимо превосходит стандартное ранжирование $rating$ ($p-value < 0{,}001$). Для $T_{focus} > 0{,}2$ число пользователей $U_{T_{focus}}$, как и $U_{randdocs}$, становится слишком мало. Полнота результата в этом случае не удовлетворяет поставленной задаче, поэтому соответствующие строки не присутствуют в табл. 14. При переборе порогового значения от меньшего к большему меняется два показателя: уменьшается число врачей в выборке $U_{T_{focus}}$ (полнота) и увеличивается разница $MAP_{expertise} - MAP_{rating}$ в пользу метода $expertise$ (точность). Это значит, что выбор конкретного значения T_{focus} обуславливается соблюдением баланса меж-

ду этими двумя величинами. В зависимости от конкретного приложения можно увеличивать точность метода и уменьшать полноту или наоборот.

Таблица 14 — Сравнение показателей MAP ранжирования пользователей методами *rating* и *expertise*

T_{focus}	$\lvert U_{T_{focus}}\rvert$	$\lvert U_{randdocs}\rvert$	MAP_{rating}	$MAP_{expertise}$
0,05	22983	304	0,0186	0,0158
0,06	22024	299	0,0189	0,0175
0,07	20252	276	0,0191	0,0179
0,08	17950	243	0,0198	0,0180
0,09	15500	209	0,0207	0,0191
0,10	13253	179	0,0204	0,0213
0,11	11198	152	0,0205	**0,0250**
0,12	9494	128	0,0166	**0,0263**
0,13	8137	109	0,0157	**0,0283**
0,14	7014	91	0,0154	**0,0285**
0,15	6034	79	0,0164	**0,0316**
0,16	5252	70	0,0170	**0,0319**
0,17	4633	64	0,0169	**0,0340**
0,18	4108	58	0,0167	**0,0352**
0,19	3636	53	0,0172	**0,0361**
0,20	3241	49	0,0180	**0,0368**

Примечание — жирным выделены показатели метода *expertise*, статистически значимо превосходящие аналогичные показатели *rating*, $p - value < 0{,}001$.

3.6 Выводы

В настоящей главе исследована проблема оценки качества вопросов и ответов медицинской тематики с точки зрения оценивания квалификации пользователей, отвечающих на вопросы в разделах CBOC о здоровье человека. В качестве решения данной задачи предложен и реализован эффективный чис-

ленный метод оценки компетентности *expertise* (формула (3.6)). В основу метода легла модель сосредоточенности пользователя на определённой теме, разработанная также в рамках диссертационного исследования (формула (3.3)). Эффективность метода с точки зрения вычислительной сложности показана в пункте 3.4.4.

На данных СВОС Ответы@Mail.Ru проведён вычислительный эксперимент, результаты которого демонстрируют значимо более высокий уровень качества ранжирования пользователей с помощью метода *expertise* по сравнению со стандартной рейтинговой системой сервиса.

В качестве недостатка эксперимента можно отметить отсутствие надёжной репрезентативной выборки пользователей, имеющих высокую медицинскую квалификацию, достаточного размера для подбора параметров метода и тестирования. Несмотря на то, что выборка, полученная способом, описанным в пункте 3.5.1, имеет размер, достаточный для демонстрации подхода, предложенного в данной главе, требуется дополнительная работа по её дополнению и уточнению для более надёжного подбора параметров и пороговых значений.

К недостаткам метода *expertise* можно отнести невозможность (в силу его устройства) ранжирования пользователей с тематическим фокусом ниже порогового значения: $\{A_i | pmax(A_i) < T_{focus}\}$. При этом, среди таких пользователей, вероятно, остаётся значимое количество квалифицированных медицинских специалистов, которые могут быть не сосредоточены на одной теме или иметь принципиально отличающиеся паттерны поведения. В дальнейшей работе планируется доработать модель *expertise* так, чтобы учесть и этих пользователей. Будущие исследования предполагают смещение акцента метода на анализ медицинского лексикона и добавление новых факторов, для определения которых необходимо дополнительное изучение поведенческих моделей пользователя, и выделение среди них тех, что отличают и подчёркивают высокую квалификацию пользователей с тематическим фокусом ниже порогового значения.

Глава 4. Персонализация поиска по медицинским веб-страницам с помощью моделирования пользователя

В четвёртой главе рассмотрена задача моделирования пользователя поисковой системы с помощью данных о здоровье, взятых из его медицинской карты. Основная цель исследования состояла в повышении качества информационного поиска по англоязычной коллекции веб-страниц медицинской тематики путём его персонализации.

4.1 Постановка задачи

С развитием интернета всё более популярным становится онлайн-поиск информации о здоровье человека. Исследования (например, [73; 74]) показывают, что поисковые системы общего назначения, такие как Google, Yahoo! или Microsoft Bing, не всегда обеспечивают удовлетворительное качество обработки подобных запросов. В рамках диссертационной работы исследовался вопрос о том, может ли *персонализация* запросов улучшить качество информационного поиска данных о здоровье человека.

Определение 4.1.1. *Под персонализацией в настоящей работе понимается эксплуатация данных, ассоциированных с пользователем произвольным образом, с целью уточнения, дополнения и улучшения понимания его информационной потребности.*

Выписываясь из больницы, пациент получает на руки медицинскую карту, содержащую данные о состоянии его здоровья. Обычно медицинские карты заполняются врачами профессиональным языком и предназначены для чтения сотрудниками медицинских учреждений и другими врачами. Пациент без специальных знаний может неправильно воспринимать и интерпретировать данные медицинской карты о своём здоровье. С другой стороны, информация такого типа может быть успешно обработана поисковой системой для предоставления результатов поиска с учётом персональных особенностей пациента: пола, возраста, истории болезней и т.п.

В рамках исследования рассматривалась информационная потребность пользователя в дополнительных знаниях о своём здоровье после выписки из больницы. В предположении, что поисковой системе известны такие поля медицинской карты, как пол (обозначен в таблице результатов как `gender`), возраст (`age`), основные жалобы на здоровье (`compl`) и история хирургических вмешательств (`proc`), перед диссертантом была поставлена задача разработать метод моделирования пользователя данными его медицинской карты для персонализации его поисковых запросов.

Чаще всего в качестве запроса пользователя фигурирует симптом или название заболевания, например, «желудочно-кишечное кровотечение».

Гипотеза 4.1.1. *В общем случае формулировки симптома или заболевания в запросе недостаточно для точной интерпретации информационной потребности пациента и выдачи, соответственно, полных и точных результатов.*

Помимо основной задачи было предложено проверить гипотезу 4.1.1, поэтому наряду с полями медицинской карты пациента при персонализации использовалось развёрнутое описание информационной потребности пользователя (`desc`), которое предоставлялось вместе с каждым его запросом.

4.2 Обзор литературы

Базовым сценарием персонализации информационного поиска является использование истории прошлых действий пользователя в поисковой системе для улучшения качества результатов его текущего запроса. Например, авторы работы [75] используют предыдущие запросы пользователя, данные посещённых им веб-страниц и просмотренных документов для построения модели интересов пользователя, на основе которой и происходит дальнейшая персонализация поисковой выдачи.

Источник [76] описывает использование социальных связей пользователя для персонализации поиска в социальной сети: модель интересов в этом случае строится на основе поисковых стратегий «близких» пользователей (то есть тех, с кем искомый общается больше всего) или пользователей, ведущих себя в социальной сети похожим образом. В источнике [77] приводится сравне-

ние эффективности персонализации поиска в различных условиях. В частности утверждается, что методы, основанные на моделях интересов пользователей, показывают нестабильное улучшение по сравнению с простым анализом кликов пользователя.

Вообще говоря, пользовательский контекст может быть рассмотрен под разными углами. Например авторы [78] различают контекст устройства (данные, полученные с устройства пользователя), пространственно-временной, а также личностный контекст, под которым понимаются характеристики конкретного пользователя. Личностный контекст объединяет понятия персонального (например [75]) и социального (как в [76]) контекста. Таким образом способы получения, обработки и анализа контекстной информации зависят от типа используемого контекста.

Существующие методы персонализации информационного поиска опираются, в основном, на классические подходы к извлечению релевантной информации. Часто стадия персонализации встраивается в обычный цикл обработки запроса и выдачи результатов. В литературе встречается методы встраивания на этапе обработки запроса (например, модификация запроса), во время ранжирования (модификация функции ранжирования) или после ранжирования (переранжирование поисковой выдачи).

Одной из первых работ по расширению поисковых запросов является [79]. Эксперименты по сравнению различных способов расширения описываются, например, в источнике [80]. В работе [81] используется модификация формулировки запроса для улучшения его соответствия профилю пользователя. Авторы [82] предлагают фреймворк адаптивного расширения запросов пользователя на основе различных стратегий: анализа отдельных слов и словосочетаний запроса, использования совместной встречаемости терминов в документах, привлечения внешних тезаурусов.

Переранжирование результатов поиска используется, например, в работе Сперетты и Гоч [83]. Авторы составляют профиль пользователя на основании истории его запросов и активности на сайте поисковой системы. Авторы [84] фокусируются на поиске по медицинским статьям. В работе производится переранжирование выдачи классической поисковой машины согласно релевантности каждой из найденных статей обобщённому профилю пользователя. В качестве обобщённого профиля используется множество терминов, извлечённых из медицинской карты искомого пациента, а также из медицинских статей, харак-

теризующих его состояние здоровья (например, «высокое кровяное давление», «хроническая сердечная недостаточность»).

В диссертационном исследовании поиск производится по веб-страницам о здоровье человека, а не медицинским статьям, как это сделано в работе [84]. Качество данных в вебе может очень сильно различаться. Предложенный метод не использует обобщение профиля пациента с помощью сторонних текстов — извлекаются только поля его медицинской карты. Потенциально это позволяет значительно увеличить точность поиска по медицинским документам за счёт уменьшения полноты.

4.3 Методы персонализации поиска

Классические алгоритмы поиска моделируют пользователя поисковой системы с помощью данных о его прошлой активности. Простейшим примером служит история прошлых запросов пользователя. В настоящей работе предлагается метод моделирования пользователя поисковой системы, использующий дополнительную обезличенную информацию о состоянии его здоровья в специфичном сценарии поиска по медицинским документам.

Методы, исследуемые в работе, основаны на встраивании этапа персонализации в стандартный сценарий работы поисковой системы. При реализации данного этапа использовались следующие подходы:

– модификация (расширение) исходного запроса пациента;
– переранжирование путём смешивания нескольких поисковых выдач на основе подхода, предложенного Шоу и Фоксом в работе [85].

4.3.1 Расширение поискового запроса

При поиске информации, касающейся вопросов состояния здоровья, пользователь интуитивно исходит из некоторых знаний о своём организме, априори имея в виду, например, свой возраст, пол, предыдущие заболевания и т.п. Классические поисковые системы в общем случае не обладают такими знаниями в

процессе поиска и извлечения релевантных страниц. Это может приводить к выдаче результатов, которые не удовлетворяют информационную потребность пользователя. Основной целью метода персонализации, разработанного в рамках диссертационного исследования, является снабжение поисковой системы априорными знаниями о пациенте, вносящими существенный вклад в поиск релевантных ему страниц. Для этого используется техника расширения поискового запроса.

Определение 4.3.1. *Под расширением поискового запроса понимается его модификация путём добавления дополнительных терминов. Начальная формулировка запроса при этом остаётся неизменной.*

Изначальный запрос расширялся полями медицинской карты пациента. В обычной ситуации медицинские карты пользователей отсутствуют в публичном доступе, поэтому для экспериментов использовались данные, предоставленные международной инициативой по оценке информационного поиска CLEF (дорожка eHealth, созданная для независимой оценки методов поиска по медицинским документам) в 2014 году. Подробно данные описаны в подразделе 1.4.2.

Алгоритм персонализации поиска с помощью расширения начального запроса пользователя полями его медицинской карты представлен на рис. 4.1. Ключевой идеей предлагаемого подхода является добавление к начальному запросу $InitialQuery$ нескольких типов данных контекста пользователя, представленных различными полями его медицинской карты $MedReport$. Значение $fieldValue$ каждого поля $fieldName$ извлекается из медицинской карты и, проходя этап предобработки (строка 5 на рис. 4.1), добавляется к запросу $Query$ с определённым весом $weight$ для соблюдения баланса важности терминов начальной и добавочной частей. Запрос $Query$ в этом случае является изменяемой копией начального запроса пользователя $InitialQuery$.

Оптимальные веса определяются перебором для каждого типа добавляемой информации. При добавлении к запросу большого числа полей подбор весов является достаточно сложной задачей с точки зрения вычислительной мощности. В силу этих ограничений эксперименты проводились с добавлением одного или двух полей, так как тестирование добавления трёх полей требует уже значительно больших вычислительных ресурсов. Подробно детали тестирования описаны в подразделе 4.4.

Рисунок 4.1 — Алгоритм персонализации поиска с помощью расширения запроса пользователя

Вход: $SearchEngine, InitialQuery, MedReport, MedReportFieldNames$
Выход: $SearchResults$

```
1: Query := InitialQuery
2: for fieldName ∈ MedReportFieldNames do
3:     weight := fieldName.GetWeight()
4:     rawFieldValue := MedReport.ExtractField(fieldName)
5:     fieldValue := rawFieldValue.Preprocess()
6:     Query.AddField(fieldValue, weight)
7: end for
8: SearchResults := SearchEngine.Search(Query)
```

4.3.2 Переранжирование

Невозможно встроить всю имеющуюся о пользователе информацию в процесс поиска только путём расширения запроса — это приводит к значительному увеличению длины запроса и соответствующему падению качества поиска. Одним из возможных решений проблемы является смешивание нескольких поисковых выдач в одну и последующее переранжирование результатов.

В работе предложен метод, примешивающий к выдаче по начальному запросу несколько выдач по запросам, расширенным полями медицинской карты пользователя.

Запрос, расширенный одновременно большим числом полей медицинской карты пациента, может получиться достаточно длинным. Исследования показывают, что увеличение длины запроса в среднем улучшает качество поиска [86]. Наряду с этим, при обработке подобных запросов возникает ряд проблем. Средняя длина запроса в поисковых системах сети Интернет составляет по разным оценкам 2–3 слова [87; 88], что заставляет разработчиков современных систем оптимизировать метрики качества поиска для коротких запросов. В случае с длинными запросами поисковые системы могут вести себя по-разному: в процес-

се обработки обрезать запрос, пытаться искать все термины в одном документе (что влечёт за собой падение полноты поиска) и т.п. Кроме того, в поисковых системах часто используется модель «мешка слов» (bag of words), согласно которой в документах и запросе не учитывается порядок слов. Это обстоятельство может стать причиной падения точности поиска в случае извлечения документов по подмножеству слов запроса. Такая ситуация возможна из-за того, что длинный запрос порождает большое число подмножеств, многие из которых не являются релевантными.

В качестве решения вышеописанной проблемы в диссертационной работе предлагается вместо одного запроса, расширенного большим числом полей, делать несколько запросов, расширенных малым числом полей. Полученные поисковые выдачи документов предлагается смешивать и упорядочивать с помощью дополнительной функции ранжирования, в общем случае отличной от той, что используется внутри основной поисковой системы. Тогда в каждом случае поисковая система будет выполнять более конкретную задачу, а в финальной выдаче будет учтён не только основной запрос пользователя, но и персональные данные его организма.

В качестве функции ранжирования при смешивании выдач использовался метод $CombSUM$, предложенный в 1994-м году в работе [85] (уравнение (4.1)). Данный метод нормализует и комбинирует ранжирующие показатели документов, выдаваемые базовой поисковой системой. Для n смешиваемых выдач $L_i : i \in [1, n]$ новое ранжирующее значение документа d такого, что $\exists i : d \in L_i$, вычисляется как взвешенная сумма Min-Max-нормализованных ранжирующих значений $score(d, L_i)$ документа d в выдаче L_i:

$$CombSUM(d) = \sum_{i=1}^{n} \alpha_i score(d, L_i), \sum_{i=1}^{n} \alpha_i = 1 \tag{4.1}$$

Коэффициент α_i в формуле (4.1) показывает относительную важность выдачи L_i при смешивании.

Теоретически предлагаемый подход может использовать сколь угодно большое число данных о пользователе, однако в силу ограничений производительности в экспериментах смешивалось только две выдачи. Таким образом, формула (4.1) может быть представлена в упрощённом виде:

$$CombSUM(d) = \alpha score(d, L_1) + (1 - \alpha) score(d, L_2) \tag{4.2}$$

В экспериментах перебором α в пределах интервала $(0, 1)$ оптимизировалась метрика качества поиска NDCG@10. По подобранному значению α можно судить, какой вклад каждого из используемых полей медицинской карты наиболее выгоден при поиске.

4.4 Эксперименты

Для проверки подхода были проведены эксперименты с тестовыми данными на английском языке, предоставленными инициативой по оценке информационного поиска по медицинским документам CLEF eHealth в 2014 году.

В корпус данных входит коллекция веб-страниц о здоровье человека, медицинские карты некоторых пользователей, тестовые топики — описания их информационных потребностей, построенные на основе предоставленных медицинских карт (используемых в качестве возможной отправной точки при поиске), и ручные оценки релевантности документов коллекции для каждой информационной потребности. Подробно корпус данных описан в подразделе 1.4.2.

Коллекция веб-страниц индексировалась поисковой системой Terrier, широко используемой в последнее время исследователями в области информационного поиска [28]. Для стемминга использовался алгоритм Портера [25], встроенный в Terrier, и являющийся, фактически, стандартом стемминга текстов на английском языке [27]. Подробнее стемминг описан в подразделе 1.4.3.

Оценка методов проводилась с помощью метрик точности (P@10) и нормализованной дисконтированной кумулятивной выгоды (NDCG@10) по 10 первым результатам поиска (подробные определения метрик даны в подразделе 1.2). Веса терминов исходного запроса приравнивались к 1. Веса дополнительных терминов подбирались для максимизации метрик. В качестве отправной точки (*baseline*) использовался исходный запрос.

Требуемая для поиска ранжирующая модель «по умолчанию» подбиралась с помощью предварительного прогона тестовых топиков. В итоге была выбрана языковая модель с Байесовским сглаживанием, использующая априорное распределение Дирихле [89], с параметром $\mu = 2400$. Оптимальное значение параметра, как и сама модель, подбиралось с помощью максимизации показателя NDCG@10 ранжирования результатов тестовых запросов.

4.4.1 Расширение поискового запроса

Медицинские карты пациентов являются полуструктурированными текстовыми документами (листинг 1.1). Частичная структура данного типа документов позволяет реализовать автоматическое извлечение некоторых персональных характеристик состояния здоровья пациента. Для расширения запросов рассматривались следующие поля: пол (`gender`), возраст (`age`), основная жалоба на здоровье (`compl`) и прошлые хирургические вмешательства (`proc`). Данные поля выбраны по причине относительной лёгкости их извлечения из медицинских карт пациентов. Кроме того, их относительно легко *формализовать* — то есть, преобразовать в термины запроса по некоторым формальным правилам, объединяющим группы схожих значений в классы.

Некоторые типы информации не могут быть использованы для расширения запроса без предварительной обработки: пол пациента обычно схематически обозначается одной буквой (в имеющихся данных `F` — `Female`, `M` — `Male`), возраст — числом полных лет на момент заполнения медицинской карты. Подобные обозначения в запросе не доносят по умолчанию до поисковой машины информацию о возрасте и поле, поэтому должны быть формализованы — преобразованы в некоторое множество терминов, которое содержится в поисковом индексе. В данной работе выделялось несколько возрастных групп пациентов, для каждой из которых были подобраны термины, характеризующие именно эту группу (табл.15).

Таблица 15 — Термины для обозначения разных возрастных групп

Возрастной интервал	Термины
0 – 10	infant, baby, child, kid
11 – 20	child, adolescent, teenager
21 – 60	adult, middleaged
больше 60	senior, elderly, elder

Для каждого пола также были подобраны характерные термины (табл. 16). При этом учитывался тот факт, что корпус состоит из веб-страниц о здоровье человека, которые в общем случае не являются специализированными медицинскими документами. Поэтому при расширении запросов

использовались как медицинские, так и общеупотребительные обозначения понятий.

Таблица 16 — Термины для обозначения пола

Пол	Термины
Женский	female, feminine, woman, girl
Мужской	male, masculine, man, boy

Наряду с полями медицинских карт пациентов для расширения запроса использовалось его подробное описание (`desc`) для проверки гипотезы 4.1.1. В экспериментах предполагалось, что данное поле по определению уточняет информационную потребность пользователя, поэтому над ним не проводилось никакой предварительной обработки.

Подбор весов полей медицинской карты пациента, расширяющих запрос, осуществлялся методом перебора с оптимизацией меры релевантности выдачи (P@10 и $NDCG$@10). Под весом поля здесь понимается вес каждого термина, входящего в поле и добавляемого к запросу. Веса перебирались на отрезке $[0, 1]$ с шагом 0,05. Таким образом, расширение запроса *gastrointestinal bleed* 55-летней пациентки с желудочно-кишечным кровотечением может выглядеть как *gastrointestinal bleed adult^0,1 middleaged^0,1 female^0,2 feminine^0,2 woman^0,2 girl^0,2*, где полям `age` и `gender` присвоены веса 0,1 и 0,2 соответственно.

Таблица 17 содержит значения метрик релевантности P@10 и $NDCG$@10, полученные после подбора оптимальных весов. Каждая строка таблицы показывает значение, усреднённое по 50 тестовым запросам, где каждый запрос расширен соответствующим образом. Например, первая строка таблицы содержит усреднённые значения P@10 и $NDCG$@10 результатов поиска по запросам, расширенным полями `age` и `desc` с весами $w_1 = 0{,}20$ и $w_2 = 0{,}15$ соответственно. Для строк, где запрос расширен одним полем, определено только одно значение веса — w_1.

Строка *baseline* показывает результаты поиска по оригинальным запросам. Для экономии места в таблице демонстрируются только те запросы, что превосходят *baseline*.

Табл. 17 показывает, что наибольшее значение метрики достигается при добавлении к запросу возраста (`age`) и пола (`gender`), а также развёрнутого

описания информационной потребности пользователя (`desc`). Пример оригинального запроса и его версии, расширенной полями `age` и `desc`, для которого метрика $NDCG@10$ значимо улучшилась, приведён в табл. 18. Увеличение значений метрик свидетельствует о том, что стандартные поисковые системы при обычном запросе о здоровье человека в среднем могут быть улучшены методами персонализации поиска с помощью дополнительных данных о пациенте.

Таблица 17 — Результаты расширения запросов полями медкарты пациента, превосходящие по обеим метрикам ($P@10$ и $NDCG@10$) качество поиска по исходному запросу (baseline). Результаты упорядочены по убыванию $NDCG@10$.

Комбинации полей	w_1	w_2	$P@10$	w_1	w_2	$NDCG@10$
`age-desc`	0,20	0,15	0,790	0,30	0,30	0,733
`desc-gender`	0,50	0,15	0,790	0,30	0,15	0,730
`age-gender`	0,10	0,20	0,788	0,10	0,20	0,722
`desc`	0,30	—	0,778	0,30	—	0,721
`age`	0,15	—	0,782	0,30	—	0,716
`compl-desc`	0,15	0,40	0,770	0,35	0,60	0,715
`gender`	0,10	—	0,780	0,10	—	0,714
`desc-proc`	0,30	0,20	0,782	0,30	0,20	0,709
`age-proc`	0,20	0,15	0,786	0,25	0,15	0,707
baseline			0,766			0,704
eHealth'14			0,756			0,744

Кроме того, в табл. 17 можно заметить, что 5 из 10 способов расширения запроса, опережающих *baseline*, содержат поле `desc`. Этим, в частности, подтверждается гипотеза 4.1.1: более подробный запрос о заболевании или симптоме в целом улучшает качество возвращаемых результатов. В верхних строках таблицы в основном фигурируют поля `age` и `gender`, из чего следует, что соответствующая информация о пациенте является достаточно важной и значимой в процессе персонализации поиска.

Расширение запросов с помощью `proc` и `compl` показывает более скромные результаты по сравнению с остальными полями, что может объясняться

Таблица 18 — Пример запроса, для которого показатель $NDCG@10$ значительно улучшился после расширения полями `age` и `desc`

Запрос	$NDCG@10$
chronic duodenal ulcer	0,67
chronic duodenal ulcer senior^0.3 elderly^0.3 elder^0.3 How^0.3 common^0.3 is^0.3 it^0.3 that^0.3 the^0.3 ulcer^0.3 starts^0.3 to^0.3 bleed^0.3 again^0.3	0,95

следующим фактом: ручное исследование медицинских карт пациентов показало, что только около 50% карт содержит значимую информацию в графах, соответствующих прошлым хирургическим вмешательствам и основной жалобе на здоровье. Кроме того, подобные поля часто содержат неформализованный текст — например, основная жалоба на здоровье обычно записывается медсестрой при первичном приёме пациента. Поэтому следует проводить более глубокую предобработку текста, например, переходить из пространства терминов в пространство концепций, как это сделано для полей типа `compl` в работе [90].

Результаты экспериментов также сравнивались с лучшим результатом участников дорожки CLEF eHealth'14. Табл. 17 демонстрирует, что подход с расширением запросов превосходит результат участников в точности ($P@10$), но уступает в качестве ранжирования ($NDCG@10$).

4.4.2 Переранжирование

Эксперименты с переранжированием призваны решить проблему использования большого множества полей медицинской карты для персонализации поиска методом расширения запроса.

В данной работе переранжирование проводилось путём смешивания нескольких выдач поисковой системы по запросам, расширенным различными полями медицинской карты пользователя. Под *выдачей* здесь понимается результат работы поисковой системы, ограниченный снизу некоторым лимитом, например, 1000 первых результатов поиска с соответствующими значениями релевантности. Эта техника подробно описана в подразделе 4.3.2.

Основной идеей экспериментов с переранжированием является дальнейшая работа по улучшению наиболее успешного результата расширения запроса. Согласно экспериментам, такой результат демонстрируется расширением запросов полями `age` и `desc`. К выдаче, полученной таким образом, предлагается примешивать выдачи, полученные расширением начального запроса другими полями.

В каждом запуске эксперимента смешивалось две выдачи, при этом выдача `age-desc` всегда участвовала с весом α, который подбирался в интервале $(0; 1)$ с шагом 0,01 для оптимизации метрики релевантности $NDCG@10$ (вторая выдача участвовала, соответственно, с коэффициентом $1 - \alpha$). Табл. 19 демонстрирует пять лучших результатов.

Таблица 19 — Результаты смешивания выдач по запросу, расширенному полями `age` и `desc`, с выдачами по другим расширенным запросам. Запросы `age-desc` участвуют в формуле (4.2) с коэффициентом α

Расширение 2-го запроса	α	$1 - \alpha$	$NDCG@10$
`gender`	0,99	0,01	0,733
`proc`	0,97	0,03	0,733
`gender-proc`	0,99	0,01	0,733
`compl`	0,98	0,02	0,733
`compl-gender`	0,98	0,02	0,732
baseline			0,733

К сожалению, подход с переранжированием не показал улучшения в качестве поиска по сравнению с *baseline*-подходом — расширением запроса полями `age` и `desc`. В табл. 19 можно заметить, что значение коэффициента α для лучших результатов смешивания всегда близко к 1, что означает, в частности, отсутствие значимого влияния на качество поиска выдач, подмешиваемых к «эталонной».

Реализация переранжирования с помощью техники смешивания выдач мотивирована необходимостью привнесения в существующую выдачу, полученную с использованием некоторых аспектов пользовательского контекста, новых документов, найденных с помощью других пользовательских данных. Анализ полученных результатов позволяет сделать вывод о том, что используемый под-

ход в данной форме является слишком «грубым», так как наряду с релевантными документами привносит в финальную выдачу также и большое число нерелевантных позиций. С этой точки зрения метод нуждается в более тонкой настройке. Возможно, следует смешивать только наиболее успешные выдачи.

Ещё одним вариантом улучшения переранжирования является использование более сложного смешивания, использующего *нормализацию со смещением*, предложенного в работе [91]. В любом случае, требуется более глубокий анализ проблемы и дополнительные исследования в этой области.

4.5 Выводы

Данная глава посвящена исследованию вопроса о том, может ли дополнительная контекстная информация о пользователе системы поиска по медицинским документам успешно использоваться для улучшения качества поиска. Эксперименты по персонализации поиска проводились на данных дорожки eHealth'14 инициативы CLEF с использованием техник расширения поискового запроса данными медицинской карты пациента и переранжирования путём смешивания выдач.

Кажется очевидным, что при поиске информации о диагнозе пациента дополнительные данные его медицинской карты имеют достаточно большое значение. Эксперименты показывают, что даже простые подходы, использующие специальный пользовательский контекст, улучшают качество поиска. Требуется более глубокий анализ результатов и дальнейшее исследование темы персонализации для понимания того, как конкретно данные профиля пользователя влияют на качество поиска, почему некоторые аспекты пользовательского контекста дают улучшение, тогда как другие — нет.

Системы персонализированного поиска могут быть полезны для пациентов как дополнительный источник информации о диагнозе и других аспектах течения заболевания (например, о побочных эффектах, наиболее вероятных в случае с конкретным пациентом). Кроме того, подобные системы могут использоваться медицинскими учреждениями для предоставления адаптированных ознакомительных и образовательных материалов, связанных с текущим состоянием здоровья пациента.

Заключение

Основные результаты работы заключаются в следующем.

1. Разработан приближенный метод оценки качества медицинских разделов вопросно-ответных сервисов. Анализ результатов проведённых экспериментов показал, что вопросы и ответы медицинской тематики имеют приемлемое качество для сценариев переиспользования и извлечения знаний.
2. Разработан эффективный вычислительный метод оценки компетентности пользователей социального вопросно-ответного сервиса. Результат численного эксперимента на данных СВОС Ответы@Mail.Ru продемонстрировал статистически значимо более высокий уровень качества ранжирования пользователей с помощью предложенного метода по сравнению со стандартной рейтинговой системой сервиса.
3. Предложен метод моделирования пользователя поисковой системы на основе данных его медицинской карты. Численный эксперимент с персонализацией поиска на основе построенных моделей тестовых пользователей показал, что даже простые подходы, использующие такие данные как пол, возраст, основную жалобу на здоровье, прошлые хирургические вмешательства, улучшают качество поиска по медицинским документам.
4. Для решения подзадачи орфографической коррекции слов при их нормализации для дальнейшего применения методов автоматической обработки текстов реализован и адаптирован для медицинской тематики модуль исправления орфографических ошибок и опечаток.

Методы, использующие тематические особенности текстовых данных, значительно углубляют их понимание, позволяя достичь большего качества по сравнению с методами анализа текстов без привязки к конкретной теме. Важно отметить, что методы, разработанные в рамках данного диссертационного исследования, вообще говоря, применимы к любой тематике, для которой существует собственная терминологическая база. Применение методов к новым данным в большинстве случаев подразумевает формирование подходящих тематических словарей, а также подбор соответствующих параметров и пороговых значений в конкретных приложениях.

Область анализа данных медицинской тематики достаточно молода, при этом имеет существенный потенциал по применению в будущих системах обработки и выдачи информации. Оценка качества текстовых данных медицинской тематики в интернете является важной задачей, так как информация плохого качества может нанести потенциальный вред здоровью пользователя. В рамках развития темы диссертации планируется рассмотреть частные аспекты оценки качества медицинской информации на более детальном уровне. Среди таких аспектов рассматриваются, например, методы оценки доверия отдельным сообщениям о здоровье человека, учёт данных пользователя в оценке качества предоставляемой ему информации, разработка корпуса данных для надёжной верификации методов оценки качества медицинской информации.

В заключение автор выражает благодарность и большую признательность научному руководителю Волкову М.В. за поддержку и руководство, научному консультанту Браславскому П.И. за помощь, обсуждение результатов и определение приоритетов дальнейшей работы. Автор сердечно благодарен своей жене Белобородовой А.Н. за моральную поддержку и помощь в трудные моменты, а также Рожковой Н.Н., Рожкову Д.А., Лизуро Т.Е. и Лизуро О.В. за ценные замечания в ходе подготовки диссертационной работы.

Список литературы

1. *Fox S.*, *Duggan M.* Health Online 2013 [Электронный ресурс]. — 2013. — Режим доступа: http://www.pewinternet.org/2013/01/15/health-online-2013/.
2. Интернет как источник получения потребителями информации о здоровье, медицине, препаратах // Дайджест HealthIndex360. — Synovate Comcon Healthcare, 2015. — Т. 19.
3. *Beloborodov A.*, *Kuznetsov A.*, *Braslavski P.* Characterizing Health-Related Community Question Answering // Proc. of the 35th European Conf. on IR research (ECIR'13): LNCS. – Moscow. Vol. 7814. — 2013. — P. 680–683.
4. *Beloborodov A.*, *Braslavski P.*, *Driker M.* Towards Automatic Evaluation of Health-Related CQA Data // Proc. of the 5th International Conf. of the CLEF Initiative (CLEF'14): LNCS. – Sheffield, UK. Vol. 8685. — 2014. — P. 7–18.
5. *Beloborodov A.*, *Braslavski P.* Does Everybody Lie? Characterizing Answerers in Health-Related CQA // Proc. Of the AINL-ISMW FRUCT Conf. – Saint-Petersburg: ITMO. — 2016. — P. 3–8.
6. *Beloborodov A.*, *Goeuriot L.* Improving Health Consumer Search with Contextual Information // Proc. of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR). — Pisa, Italy, 2016.
7. *Beloborodov A.* Whether a CQA User is a Medical Professional? Work in Progress // Proc. of the 6th Symposium on Future Directions in Information Access (FDIA'15): eWiC Series. — Thessaloniki, Greece, 2015. — P. 71–73.
8. *Белобородов А.*, *Кузнецов А.* Извлечение именованных сущностей из тематического корпуса вопросов и ответов // Современные проблемы математики: тезисы междунар. (44-й всероссийской) конф. — Екатеринбург, 2013. — С. 302—304.
9. *Белобородов А.* Модуль исправления орфографических ошибок и опечаток: а. с. 2016662094 РФ. — 2016. — Бюл. № 11. 869.
10. *Белобородов А.* Модуль исправления орфографических ошибок и опечаток: исходный код [Электронный ресурс]. — 2016. — Режим доступа: https://github.com/bellal89/SpellIt.

11. *Маннинг К. Д.*, *Рагхаван П.*, *Шютце Х.* Введение в информационный поиск. : Пер. с англ. / под ред. П. И. Браславского, Д. А. Клюшина, И. В. Сегаловича. — М. : ООО "И.Д. Вильямс", 2011. — 528 с.

12. *Landis J.*, *Koch G.* The measurement of observer agreement for categorical data // Biometrics. — 1977. — P. 159–174.

13. *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in Information Retrieval. — New York, NY, USA: ACM, 1999. — P. 50–57.

14. *Blei D. M.*, *Ng A. Y.*, *Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 993–1022.

15. *Воронцов К. В.* Вероятностное тематическое моделирование [Электронный ресурс]. — М. : MachineLearning.ru, 2013. — Режим доступа: http://www.machinelearning.ru/ wiki/images/2/22/Voron-2013-ptm.pdf.

16. Textflow: Towards Better Understanding of Evolving Topics in Text / W. Cui [et al.] // IEEE Transactions on Visualization and Computer Graphics. — 2011. — Vol. 17(12). — P. 2412–2421.

17. Statistical Topic Models for Multilabel Document Classification / T. Rubin [et al.] // Machine Learning. — 2012. — Vol. 88. — P. 157–208.

18. *Feng Y.*, *Lapata M.* Topic Models for Image Annotation and Text Illustration // Human Language Technologies Conference. — 2010. — Vol. 17(12). — P. 831–839.

19. *Yeh J.-H.*, *Wu M.-L.* Recommendation Based on Latent Topics and Social Network Analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. — 2010. — Vol. 1. — P. 209–213.

20. Новости проекта Ответы@Mail.Ru [Электронный ресурс]. — М., 2012. — Режим доступа: http://otvet.mail.ru/news/#hbd2012.

21. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval / L. Goeuriot [et al.] // Proceedings of CLEF 2014 online working notes. — 2014.

22. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database / M. Saeed [et al.] // Critical Care Medicine. — 2011. — Vol. 39(5). — P. 952–960.

23. Text REtrieval Conference (TREC) [Электронный ресурс]. — National Institute of Standards, Technology, 2016. — Режим доступа: http://trec.nist.gov/.

24. *Zobnin A., Nosyrev G.* Morphological Analyzer MyStem 3.0 // Труды института русского языка им. В.В. Виноградова. — 2015. — Т. 6(1). — С. 300—307.

25. *Porter M.* An Algorithm for Suffix Stripping // Program. — 1980. — Vol. 14(3). — P. 130–137.

26. *Porter M.* Snowball: A Language for Stemming Algorithms [Электронный ресурс]. — 2001. — Режим доступа: http://snowball.tartarus.org/texts/introduction.h

27. *Willett P.* The Porter Stemming Algorithm: Then and Now // Program. — 2006. — Vol. 40(3). — P. 219–223.

28. From Puppy to Maturity: Experiences in Developing Terrier / C. Macdonald [et al.] // Proceedings of the Workshop in Open Source in Information Retrieval at SIGIR. — 2012. — P. 60–63.

29. *Левенштейн В.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. — 1965. — Т. 163(4). — С. 845—848.

30. Национальный корпус русского языка [Электронный ресурс]. — М., 2013. — Режим доступа: http://www.ruscorpora.ru/index.html.

31. *Ляшевская О.* О частотном словаре Национального корпуса русского языка // Слово и словарь = Vocabulum et vocabularium: сб. науч. тр. по лексикографии. — Гродно, 2007.

32. *Лазарева Г. Ю.* Справочник фельдшера. — М. : Рипол Классик, 2013.

33. Государственный реестр лекарственных средств России [Электронный ресурс]. — М. : Минздрав России, 2012. — Режим доступа: http://grls.rosminzdrav.ru.

34. Международная классификация болезней (МКБ-10) [Электронный ресурс]. — М. : Регистр лекарственных средств России, 2012. — Режим доступа: https://www.rlsnet.ru/mkb_tree.htm.

35. The MIT License [Электронный ресурс]. — Massachusetts Institute of Technology. — Режим доступа: https://opensource.org/licenses/MIT.

36. *Liu Y.*, *Agichtein E.* On the Evolution of the Yahoo! Answers QA Community // Proceedings of SIGIR'08. — 2008. — P. 737–738.

37. *Kim S.*, *Oh S.* Users' Relevance Criteria for Evaluating Answers in a Social Q&A Site // Journal of the Association for Information Science and Technology. — 2009. — Vol. 60(4). — P. 716–727.

38. *Oh S.*, *Worrall A.*, *Yi Y. J.* Quality Evaluation of Health Answers in Yahoo! Answers: A Comparison between Experts and Users // Proceedings of the American Society for Information Science and Technology. — 2011. — Vol. 48(1). — P. 1–3.

39. *Oh S.*, *Yi Y. J.*, *Worrall A.* Quality of Health Answers in Social Q&A // Proceedings of the Association for Information Science and Technology. — 2012. — Vol. 49(1). — P. 1–6.

40. Finding high-quality content in social media / E. Agichtein [et al.] // Proceedings of WSDM'08. — 2008. — P. 183–194.

41. *Agichtein E.*, *Liu Y.*, *Bian J.* Modeling information-seeker satisfaction in community question answering // ACM Trans. Knowl. Discov. Data. — 2009. — Vol. 3(2). — P. 1–27.

42. *Shah C.*, *Pomerantz J.* Evaluating and predicting answer quality in community QA // Proceedings of SIGIR'2010. — 2010. — P. 411–418.

43. Knowledge sharing and Yahoo! Answers: Everyone knows something / L. A. Adamic [et al.] // Proceedings of WWW'08. — 2008. — P. 665–674.

44. *Harper F. M.*, *Moy D.*, *Konstan J. A.* Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites // Proceedings of CHI'09. — 2009. — P. 759–768.

45. *Rodrigues E. M.*, *Milic-Frayling N.* Socializing or knowledge sharing?: Characterizing social intent in community question answering // Proceedings of CIKM'09. — 2009. — P. 1127–1136.

46. Analyzing and predicting question quality in community question answering services / B. Li [et al.] // Proceedings of WWW'12. — 2012. — P. 775–782.

47. *Correa D.*, *Sureka A.* Fit or unfit: Analysis and prediction of 'closed questions' on Stackoverflow // Proceedings of COSN'13. — 2013. — P. 201–212.

48. *Lezina G.*, *Kuznezov A.*, *Braslavski P.* Learning to predict closed questions on Stackoverflow // Kazan. Gos. Univ. Uchen. Zap. Ser. Fiz.-Mat. Nauki. — 2013. — Vol. 155, no. 4. — P. 118–133.

49. *Zhang Y.* Contextualizing Consumer Health Information Searching: an Analysis of Questions in a Social Q&A Community // Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10). — 2010. — P. 210–219.

50. *Kim S.*, *Pinkerton T.*, *Ganesh N.* Assessment of H1N1 questions and answers posted on the web // American Journal of Infection Control. — 2012. — Vol. 40(3). — P. 211–217.

51. *Lampos V.*, *Bie T. D.*, *Cristianini N.* Flu Detector — Tracking Epidemics on Twitter // Lecture Notes in Computer Science. — Springer Berlin Heidelberg, 2010. — Vol. 6323. — P. 599–602.

52. *Paul M.*, *Dredze M.* You Are What You Tweet: Analyzing Twitter for Public Health // Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. — 2011. — P. 265–272.

53. *Bhattacharya S.*, *Tran H.*, *Srinivasan P.* Discovering health beliefs in Twitter // AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text. — 2012.

54. *Wong W.*, *Thangarajah J.*, *Padgham L.* Health conversational system based on contextual matching of community-driven question-answer pairs // Proceedings of CIKM'11. — 2011. — P. 2577–2580.

55. *Phan X.*, *Nguyen C.* GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA). — 2007.

56. Бюллетень EuroFlu [Электронный ресурс]. — Европейский центр профилактики и контроля заболеваний ВОЗ, 2012. — Режим доступа: http://euroflu.org.

57. *Cartright M.-A.*, *White R. W.*, *Horvitz E.* Intentions and attention in exploratory health search // Proceedings of SIGIR'11. — 2011. — P. 65–74.

58. Энциклопедия лекарств и товаров аптечного ассортимента [Электронный ресурс]. — М. : Регистр лекарственных средств России, 2012. — Режим доступа: https://www.rlsnet.ru.

59. Shallow information extraction from medical forum data / P. Sondhi [et al.] // Proceedings of COLING'2010. — 2010. — P. 1158–1166.

60. *Raban D.*, *Harper F.* Motivations for Answering Questions Online // New media and innovative technologies. — 2008. — Vol. 73.

61. *Dearman D.*, *Truong K.* Why Users of Yahoo! Answers Do not Answer Questions // Proceedings CHI'2010 Conference. — 2010. — P. 329–332.

62. *Pelleg D.*, *Yom-Tov E.*, *Maarek Y.* Can You Believe an Anonymous Contributor? On Truthfulness in Yahoo! Answers // PASSAT/SocialCom. — 2012. — P. 411–420.

63. *Sillence E.*, *Hardy C.*, *Briggs P.* Why Don't We Trust Health Websites that Help Us Help Each Other?: An Analysis of Online Peer-to-Peer Healthcare // Proceedings of WebSci'13 Conference. — 2013. — P. 396–404.

64. Tapping on the Potential of Q&A Community by Recommending Answer Providers / J. Guo [et al.] // Proceedings of CIKM'2008. — 2008. — P. 921–930.

65. *Pal A.*, *Konstan J.* Expert Identification in Community Question Answering: Exploring Question Selection Bias // Proceedings of CIKM'2010. — 2010. — P. 1505–1508.

66. Topic-sensitive Probabilistic Model for Expert Finding in Question Answer Communities / G. Zhou [et al.] // Proceedings of CIKM'2012. — 2012. — P. 1662–1666.

67. *Hadgu A.*, *Jäschke R.* Identifying and Analyzing Researchers on Twitter // Proceedings of WebSci'14 Conference. — 2014. — P. 23–32.

68. *Bagdouri M.*, *Oard D.* Profession-based Person Search in Microblogs: Using Seed Sets to Find Journalists // Proceedings of CIKM'2015 Conference. — 2015. — P. 593–602.

69. *Blei D.* The computational complexity of LDA [Электронный ресурс]. — 2008. — Режим доступа: https://lists.cs.princeton.edu/ pipermail/topic-models/2008-April/000211.html.

70. *Воронцов К. В.*, *Потапенко А. А.* Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. — 2013. — Т. 1, № 6. — С. 657—686.

71. *Воронцов К. В.* Вероятностное тематическое моделирование: обзор моделей и аддитивная регуляризация [Электронный ресурс]. — М. : MachineLearning.ru, 2013. — Режим доступа: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf.

72. Об утверждении Единого квалификационного справочника должностей руководителей, специалистов и служащих, раздел «Квалификационные характеристики должностей работников в сфере здравоохранения» [приказ № 541н: принят Минздравсоцразвития РФ 23 июля 2010 г.] — 2016. — Режим доступа: https://www.rosminzdrav.ru/documents/.

73. *White R.*, *Horvitz E.* Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search // ACM Transactions on Information Systems (TOIS). — 2009. — Vol. 27(4). — P. 23.

74. *Zuccon G.*, *Koopman B.*, *Palotti J.* Diagnose This If You Can: On the Effectiveness of Search Engines in Finding Medical Self-diagnosis Information // Proceedings of ECIR'15 Conference. — 2015. — P. 562–567.

75. *Teevan J.*, *Dumais S.*, *Horvitz E.* Personalizing Search via Automated Analysis of Interests and Activities // Proceedings of SIGIR'05 Conference. — 2005. — P. 449–456.

76. Personalized Social Search Based on the User's Social Network / D. Carmel [et al.] // Proceedings of CIKM'09 Conference. — 2009. — P. 1227–1236.

77. *Dou Z.*, *Song R.*, *Wen J.* A Large-Scale Evaluation and Analysis of Personalized Search Strategies // Proceedings of WWW'07 Conference. — 2007. — P. 581–590.

78. *Tamine-Lechani L.*, *Boughanem M.*, *Daoud M.* Evaluation of Contextual Information Retrieval Effectiveness: Overview of Issues and Research // Knowledge Information Systems. — 2010. — Vol. 24. — P. 1–34.

79. *Jones K. S.* Automatic Keyword Classification for Information Retrieval // Butterworth. — London, 1971.

80. *Xu J.*, *Croft W.* Query Expansion Using Local and Global Document Analysis // ACM SIGIR Forum. — New York, 1996. — Vol. 51(2). — P. 168–175.

81. *Sieg A.*, *Mobasher B.*, *Burke R.* Inferring User's Information Context from User Profiles and Concept Hierarchies // Classification, Clustering, and Data Mining Applications. — 2004. — P. 563–573.

82. *Chirita P.*, *Firan C.*, *Nejdl W.* Personalized Query Expansion for the Web // Proceedings of SIGIR'07 Conference. — 2007. — P. 7–14.

83. *Speretta M.*, *Gauch S.* Personalized Search Based on User Search Histories // Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. — 2005. — P. 622–628.

84. Personalizing Retrieval of Journal Articles for Patient Care / S. Teufel [et al.] // Proceedings of the AMIA Symp. American Medical Informatics Association. — 2001. — P. 696–700.

85. *Shaw J.*, *Fox E.* Combination of multiple searches // NIST special publication SP. — 1994. — P. 243.

86. Query Length in Interactive Information Retrieval / N. Belkin [et al.] // Proceedings of SIGIR'2003 Conference. — 2003. — P. 205–212.

87. *Croft W.*, *Cook R.*, *Wilder D.* Providing Government Information on the Internet: Experiences with THOMAS // Digital Libraries Conference. — 1995. — P. 19–24.

88. *Jansen B.*, *Spink A.*, *Saracevic T.* Real Life, Real Users and Real Needs: A Study and Analysis of Users' Queries on the Web // Information Processing and Management. — 2000. — Vol. 36(2). — P. 207–227.

89. *Mackay D.*, *Peto L.* A Hierarchical Dirichlet Language Model // Natural Language Engineering. — 1995. — Vol. 1(3). — P. 289–307.

90. *Travers D.*, *Haas S.* Using Nurses' Natural Language Entries to Build a Concept-oriented Terminology for Patients' Chief Complaints in the Emergency Department // Journal of Biomedical Informatics. — 2003. — Vol. 36(4). — P. 260–270.

91. *Ould-Amer N.*, *Mulhem P.*, *Géry M.* LIG at CLEF 2015 SBS Lab // Working Notes of CLEF'2015 Conference. — 2015.

Список рисунков

Список таблиц

Приложение А

Пункты опроса врачей — пользователей профессионального сообщества «Доктор на работе»

1. Пол.
2. Основная специальность.
3. Работа в настоящее время:
   - амбулаторный приём;
   - стационар;
   - образование и наука;
   - другое.
4. Врачебный стаж.
5. Насколько по вашему мнению достоверна информация о здоровье и медицине в интернете?
   - в основном достоверная;
   - есть и достоверная, и недостоверная информация;
   - в основном недостоверная информация;
   - всё зависит от умения анализировать и интерпретировать информацию.
6. Как вы относитесь к специализированным интернет-сервисам медицинской тематики в интернете?
   - положительно;
   - скорее положительно;
   - безразлично;
   - скорее отрицательно;
   - отрицательно.
7. Как вы относитесь к тому, что пациенты активно ищут медицинскую информацию в интернете?
   - хорошо;
   - скорее хорошо;
   - безразлично;
   - скорее плохо;
   - плохо.

8. Знакомы ли вы с социальными вопросно-ответными сервисами и форумами?
9. Если вы пользуетесь подобными сервисами, то какими?
10. Если вы пользуетесь подобными сервисами, насколько часто вы отвечаете на вопросы о здоровье и медицине?
11. Что побуждает вас отвечать на вопросы о здоровье и медицине в интернете?
    - профессиональный долг;
    - желание поделиться знаниями и опытом;
    - желание помочь людям;
    - желание пообщаться с людьми;
    - чем большему числу людей я отвечу в интернете, тем меньше пациентов придет на приём к врачу.

Приложение Б

Пункты опроса активных пользователей медицинских разделов вопросно-ответного сервиса Ответы@Mail.Ru

1. Пол.
2. Возраст.
3. Город/область.
4. Образование:
   - среднее;
   - среднее специальное;
   - неполное высшее;
   - высшее;
   - учёная степень;
   - другое.
5. Связано ли ваше образование с медициной и здоровьем?
6. Если ваша работа в настоящее время связана с медициной, здоровьем или индустрией красоты, укажите вашу профессию.
7. Связана ли ваша профессия с тематикой вопросов, на которые вы отвечаете?
8. Почему лично вы отвечаете на вопросы других пользователей сервиса Ответы@Mail.Ru?
   - это форма общения с другими людьми;
   - я зарабатываю очки и повышаю свой рейтинг;
   - отвечая на вопросы, я учусь и узнаю много нового;
   - это связано с моей профессиональной деятельностью (например, я использую сервис для рекламы товаров, услуг и т.п.);
   - я хочу поделиться моими знаниями и опытом;
   - я сопереживаю людям, у которых проблемы, и хочу помочь им;
   - я хочу, чтобы люди получали достоверную информацию;
   - я ожидаю, что люди так же помогут мне советом, когда у меня будет необходимость в этом;
   - это просто интересно;
9. Как часто вы посещаете Ответы@Mail.Ru?

- несколько раз в день;
- 1 раз в день;
- несколько раз в неделю;
- 1 раз в неделю;
- реже.

10. Как давно вы пользуетесь сервисом Ответы@Mail.Ru?
    - до полугода;
    - полгода – год;
    - 1 – 2 года;
    - 2 – 3 года;
    - больше 3 лет.
11. Укажите, пожалуйста, ваш рейтинг в сервисе Ответы@Mail.Ru.

Приложение В

Копия свидетельства о государственной регистрации модуля исправления орфографических ошибок и опечаток

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2016662094

Модуль исправления орфографических ошибок и опечаток

Правообладатель: ***Белобородов Александр Владимирович (RU)***

Автор: ***Белобородов Александр Владимирович (RU)***

Заявка № **2016619511**
Дата поступления **09 сентября 2016 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***31 октября 2016 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев