САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-механический факультет

Кафедра системного программирования

Представление и обработка запросов на основе множеств объектов с оценками

Дипломная работа

Ярыгиной Анны Сергеевны

Научный руководитель	……	д.ф.-м.н., проф. Новиков Б.А.
	/ подпись /	
Рецензент	……	ст. преп. Луцив Д.В.
	/ подпись /	
“Допустить к защите”	……	д.ф.-м.н., проф. Терехов А.Н.
заведующий кафедрой,	/ подпись /	

Санкт-Петербург

2011

SAINT-PETERSBURG STATE UNIVERSITY

Mathematics & Mechanics Faculty

Software Engineering Chair

Complex query presentation and processing techniques based on sets of objects with scores

by

Iarygina Anna

Supervisor	……	Professor B.A. Novikov
Reviewer	……	Senior lecturer D.V. Luciv
"Approved by"	……	Professor A.N. Terekhov
Head of department,		

Saint-Petersburg

2011

СОДЕРЖАНИЕ

Введение

Быстрый рост объемов хранимой и обрабатываемой информации увеличивает значение качества ее обработки. Важными аспектами качества являются качество поиска информации и возможность анализа объектов сложной природы. Сложность хранимых объектов и взаимосвязи между ними ведут к усложнению задачи поиска и необходимости улучшать качество результата.

Задача поиска и обработки сложных объектов

Сложные объекты представляют собой объекты, которые могут иметь разнообразную структуру; включать в себя атрибуты самых разнообразных типов: численные, строковые, перечисления. Понятие сложного объекта в этой работе согласуется с понятием, используемым в большинстве объектных моделей. Тем не менее, важно отметить, что сложные объекты могут также неявно содержать в себе разные характеристики, то есть включать в себя не только явно выписанные, но и вычисляемые, то есть производные атрибуты.

Примерами таких сложных объектов могут быть изображения, тексты, графы, записи. Многообразие таких объектов порождает задачу осмысленного поиска и обработки объектов сложной природы.

Высококачественные поисковые системы основаны на понятии подобия, и, в частности, при поиске и анализе сложных объектов, как правило, используются подходы, основанные на понятии семантической близости или подобия между объектами. Объекты считаются похожими, если в некотором смысле похожи характеристики их описывающие.

Для отдельных классов объектов, таких как тексты и объекты простых типов, меры подобия и методы поиска достаточно проработаны. Тем не менее, задача построения метрик и функций подобия для сложных объектов по-прежнему интересна и важна.

В представленной работе обсуждаются методы описания сложных видов поиска на основе подобия, позволяющие единообразно формировать поисковые запросы и алгоритмы построения результата.

Среди работ исследовательского сообщества, связанных с поиском и обработкой сложных объектов, доминируют подходы, базирующиеся на использовании природы конкретных сложных объектов. Можно предположить, что выявление общих механизмов работы со сложными объектами позволит упростить и улучшить алгоритмы поиска.

Сложные поисковые запросы

За последние годы не только увеличилась средняя длина пользовательского текстового запроса, но увеличилось количество систем, в которых пользователь имеет возможность задавать некоторую структуру запроса, активно вовлечен в процесс поиска.

Задача поиска сложных объектов является не единственным примером сложных видов поиска. Поиск все чаще представляет собой не разовую операцию, а сложный, многоэтапный, часто интерактивный, процесс. Потребность в адекватном описании и отображении таких процессов поиска самых разнообразных объектов лежит в основе этого класса задач. В центре внимания в этом случае оказывается сам процесс поиска, а не структура объектов.

В связи с этим возникает задача отражения различных схем поиска в терминах сложного запроса внутри поисковой системы. Решение этой задачи должно представлять собой способ выражать высокоуровневые потребности пользователя в терминах формальных операций, на внутреннем языке поисковой системы.

Цели работы

Целью работы является построение формального средства, механизма описания сложных видов поиска на основе подобия. Предлагаемый инструмент должен позволить на логическом уровне единообразно формировать поисковые запросы для сложных видов поиска. Такое средство должно представлять собой однородную систему способов и операций, предназначенную для формирования и декомпозиции поисковых запросов, не зависящую от конкретной задачи поиска.

Мотивировка

Все рассматриваемые нами задачи тем или иным образом являются конкретизациями задач поиска или анализа объектов. Задачи можно разделить на два класса:

- связанные с адекватным отображением и описанием природы процесса поиска,
- направленные на поиск сложных объектов.

Такое деление будет достаточно условным и некоторые из наиболее сложных и интересных задач можно отнести к обоим классам. Мы рассмотрим несколько примеров и покажем, как представленные задачи выражаются в терминах задачи построения сложных запросов.

Природа процесса поиска

Одним из интересных примеров сложных видов поиска является задача уточняющего поиска. В этом случае поисковая система стремится повысить качество результата за счет уточнения запроса.

Уточнение может быть выражено в запросе явным и неявным образом. Задачу использования профиля персонализации при поиске можно рассматривать как задачу неявного уточнения. Профиль пользователя, написанный им самим или полученный на основе анализа его (пользователя) работы, уточняется некоторым запросом. Пользователь получает в качестве результата набор объектов релевантных не только запросу, но и всему профилю пользователя. Таким образом, необходимо построить сложный запрос, в ответ на который поисковая система вернет объекты соответствующие и профилю и запросу.

Другим источником неявного уточнения является сценарий, когда на основе обратной связи (feedback) от пользователя достраивается уточнение конкретного запроса пользователя. Такой вид уточнения активно используется в информационном поиске, особенно в задачах поиска изображений и фактов.

Знание о том, какие из предложенных объектов релевантны запросу с точки зрения пользователя, позволяет не только выделить наиболее важные для пользователя признаки, но и расширить первоначальный запрос новыми объектами, помеченными как релевантные. В этой ситуации, потребность пользователя, состоящая в том, чтобы найти объекты похожие на запрос и на примеры релевантных этому запросу объектов, может быть выражена с помощью сложного запроса.

Построение сложных запросов помогает также в случае использования отрицательной обратной связи от пользователя. Необходимо формировать сложный запрос таким образом, чтобы поисковая система находила объекты релевантные запросу, но не похожие на примеры «неправильных», то есть отклоненных пользователем, результатов.

Все приведенные выше примеры являются частными случаями общей задачи синтеза сложного запроса на основе простых. Для того чтобы выражать и учитывать как четко, так и неявно отраженные в запросе потребности пользователя в уточнении и их соответствие объектам результата, в приведенных схемах поиска необходимо строить некоторый обобщенный запрос на основе простых запросов. Таким образом, конечный, уточненный запрос пользователя представляет собой некоторую сложную структуру, состоящую из простых подзапросов.

Многие существующие решения тесно привязаны к внутренним свойствам синтезируемых запросов. Для более универсального решения описанного класса задач необходимо средство, не зависящее от природы поисковых объектов, в терминах которого можно выражать новый класс синтезированных запросов, адекватно описывающий потребности пользователя.

Поиск сложных объектов

Поиск и обработка сложных объектов также является одной из актуальных задач. В этом классе задач в фокусе рассмотрения находится структура объектов, а не процесс поиска.

При анализе сложного объекта пользователь, как правило, оценивает все его признаки и атрибуты в совокупности, но при построении поисковой системы встает задача точного описания этого интуитивного процесса синтеза нескольких критериев поиска.

При работе со сложными объектами возникает необходимость сопоставления мер разной природы. Такая потребность возникает в ситуации, когда имеется один неделимый сложный объект, например изображение, и мы можем вычислить меру подобия этого объекта по отношению к другим объектам разными способами, например, основываясь на цвете или текстуре изображения.

Когда поисковая система знает несколько разных оценок объекта, возникает вопрос о сопоставлении и комбинировании оценок разной природы. Такая же задача возникает в ситуации, когда внутри сложного объекта можно явным образом выделить несколько атрибутов разных типов.

Важно сказать, что методы поиска простых объектов достаточно проработаны, предложен целый ряд различных метрик и мер близости между объектами. Но в тоже время задача построения мер близости для сложных, комбинированных объектов требует дополнительного исследования.

Вместо того чтобы строить частные решения задач поиска сложных объектов, кажется целесообразным предложить средство, не зависящее от природы обрабатываемых сложных объектов, которое позволит удобно и адекватно выражать процесс сопоставления и синтеза метрик и оценок разной природы. Эта идея поддерживается в современных работах посвященных синтезу данных [18,19,31].

Задача поиска изображений

Поскольку изображения составляют один из важных классов информационных ресурсов, в рамках этой работы мы будем использовать задачу поиска изображений в качестве примера обработки объектов.

Изображения являются интересным примером сложных объектов. Изображения могут рассматриваться как сложные объекты, поскольку при поиске часто используются разнообразные признаки и характеристики изображений: цвет или текстура, размер или яркость, текстовая аннотация.

Когда пользователь самостоятельно выбирает изображения из коллекции, он видит содержание изображений, семантические образы, осмысленные объекты. При автоматическом поиске изображений по содержанию поисковая система, как правило, не может выделить семантически значимые для пользователя образы и работает с низкоуровневыми элементами, например, пикселями. В литературе это несоответствие обозначается как семантический разрыв [14].

Из-за существования семантического разрыва возможности поиска изображений по содержанию ограничены. Тем не менее, поиск изображений только по текстовой аннотации сильно ограничивает набор рассматриваемых коллекций, поскольку построение качественных описаний достаточно трудоемкий процесс.

При поиске изображений по содержанию необходимо найти изображения в коллекции на основе характеристик, вычисленных по составляющим объекты низкоуровневым элементам, например, пикселям.

В качестве запроса, как правило, используются изображения-образцы. При решении задачи поиска изображений по образцу, как сужения задачи поиска изображений по содержанию, предполагается, что пользователь хочет найти в коллекции изображения похожие на изображение-образец.

Результаты работы

В рамках работы предложен подход к построению сложных запросов, основанных на понятии подобия. Разработанная модель построения и выполнения сложных запросов основана на понятии множества объектов с оценками.

Предложенный инструмент определен на уровне работы с множеством поисковых образов внутри системы. Между всеми объектами в этом множестве и объектом-запросом определено некоторое отношение подобия, которое определяет значения оценок объектов

в множестве. Разработанная система операций работает с этим множеством объектов, но не зависит от конкретного способа вычисления меры подобия и оценок.

В рамках работы предложен набор операций над множеством поисковых объектов с оценками, вычисленными на основе подобия, который позволяет описывать высокоуровневые сценарии, выраженные в терминах поискового запроса или необходимые при обработке сложных объектов.

На основе предложенной формальной системы операций представлено несколько спецификаций процедур калибровки и синтеза оценок объектов. В работе также обсуждаются методы настройки описанных операций под специфические, конкретные задачи поиска.

В качестве примера работы со сложными объектами рассмотрена задача поиска изображений по образцу. Использовались подходы, основанные на комбинировании различных признаков объектов, с целью более разностороннего рассмотрения свойств объектов при поиске и потенциального улучшения качества поиска. Результаты, полученные с помощью построенной модели, были проанализированы и сопоставлены с широко известными в информационном поиске подходами к комбинированию нескольких признаков при поиске изображений по содержанию.

Эксперименты продемонстрировали применимость представленной модели к задаче поиска изображений по содержанию.

1 Достижения научного сообщества в данной области

1.1 Работы по поисковым моделям

Разработке поисковых моделей посвящено много работ [13]. Основные модели поиска появились достаточно давно, но задача построения адекватной поисковой модели попрежнему остается актуальной.

В работе [10] рассматривается вероятностная модель поиска, согласно которой основным принципом построения результата поиска является вероятность того, что документ релевантен потребности пользователя. Эта работа также интересна для нас тем, что в ней обсуждается абстрактная модель поиска и ее основные понятия: поисковый запрос, документ, релевантность документа запросу, релевантность документа потребности пользователя. Авторы разграничивают такие понятия как потребность пользователя, поисковый запрос и его представление в поисковой системе; документ и образ документа в поисковой системе.

В [11] затрагивается и объясняется вопрос о важности ранжирования в поисковых системах, а также принцип «The probability ranking principle». Основной целью работы является обсуждение неоднозначного понимания вероятности в разных поисковых моделях. Различные взгляды на использование вероятности в информационном поиске представлены в работах [22,23]. Наша модель позволяет сохранять вероятностную интерпретацию оценок объектов при обработке сложных запросов.

В работе [5] обсуждается возможность использования идей булевой и векторной модели при поиске пассажей (отрывков в тексте). Эта работа интересна с точки зрения нашего направления исследований тем, что авторы обращают внимание на использование операторов внутри поискового запроса. Например, для них интересны работы по булевой модели, в которых используются не только операторы AND и OR, но и NEAR.

1.2 Методы ранжирования результатов

Одной из первых работ по использованию вероятностных моделей при ранжировании результатов была статья [29]. В работе рассматриваются основные проблемы ранжирования результатов: изменчивость информационных потребностей пользователей и противоречивость целевых функций, то есть необходимость нахождения баланса между точностью и полнотой ответа.

В работе [16] обсуждается вероятностный принцип ранжирования (probability ranking principle), предложенный Robertson в 1977 году, и рассматриваются его недостатки с точки зрения поиска мультимедиа данных и интерактивного поиска. В этой работе авторы обсуждают особенности интерактивного информационного поиска, например, такие как необходимость анализа всего процесса взаимодействия с пользователем. Авторы предлагают использовать разную стоимость для разных операций и активностей пользователя. На основе проведенного анализа задачи предложена формальная модель расширения probability ranking principle на случай интерактивного поиска.

Авторы [35] предлагают вероятностный принцип ранжирования мультимедиа документов, позволяющий учитывать время на пересылку документа по сети и его изучение пользователем.

Работа [27] содержит сравнение методов ранжирования, основанных на использовании истории запросов пользователей. В предложенных подходах функция ранжирования также основана на вероятности того, что некоторое слово встречается в контексте, то есть в совокупности запросов пользователя.

Алгебра для построения функций ранжирования результатов предложена в [3]. В работе рассматриваются недостатки глобального ранжирования документов в поисковом результате и преимущества локального построения рангов документов. Авторы предлагают алгебру не только на множестве документов, а на множестве групп документов. На основе предложенной алгебры рассмотрены методы построения рангов документов на основе локального контекста, где под контекстом понимается группа документов, которой принадлежит ранжируемый документ. Для слияния и агрегирования локально построенных рангов предлагается подход основанный на операциях, представленной в этой работе алгебры. Операции алгебры построены на основе использования структуры связей между документами внутри одной локальной группы и взвешенного суммирования рангов, построенных внутри локальных групп.

Работа [3] является достаточно близкой к нашей работе по своим целям, поскольку в ней предлагается система операций, которая позволяет работать с ранжированием документов независимо от их природы. Тем не менее, наша модель будет основана на понятии подобия, а не на использовании связей между документами.

В работе [32] предложен метод построения ранжированных результатов для задачи поиска в структурированных данных. Метод основан на использовании вложенности элементов в структурированном документе.

1.3 Языки запросов

В запросе пользователь выражает свою информационную потребность, поэтому точность выразительных средств языка запросов и его простота оказывают влияние на работу любой поисковой системы.

Поскольку достаточно большой объем информации и данных хранится в структурированном виде, возникает потребность в создании языков запросов, позволяющих учитывать эту структуру внутри пользовательского запроса.

Одной из самых первых работ по проектированию структурированного языка запросов близкого к естественному английскому языку была статья [7].

Авторы [40] объясняют необходимость учета структуры документа в пользовательском запросе. В работе представлено достаточно точное описание недостатков существующих структурированных языков запросов. Авторы предложили язык запросов, позволяющий отражать несколько типов намерений пользователя: поиск по атрибуту, поиск по термину (текстовый поиск), взаимодействие с пользователем (например, использование фильтров уже после отображения результатов поиска), навигация по сети (спуск по иерархии элементов) и поиск по имени.

В работе [39] представлено усовершенствование языка SSQL для поиска документов, в которых структурное содержание появляется после текстового, когда пользователь вручную размечает некоторые атрибуты документа.

Работа [25] предлагает комбинировать гибкость поиска по ключевым словам с выразительностью и точностью поиска по структуре документа.

Авторы статьи [34] позволяют искать документы, файлы одновременно по содержанию и структуре каталогов, для чего используют язык запросов похожий на Xpath и набор правил, расширяющих пользовательский запрос.

Автор [15] и связанной с ней серии работ обсуждает комбинирование поиска в реляционной и мультимедиа базах данных. В работе введено понятие атомарного запроса. Атомарный запрос в мультимедиа базе данных как правило основан на понятии подобия между объектами, в отличие от атомарного запроса в реляционной модели. Ответ на ато-

марный запрос авторы предлагают выражать в терминах нечеткого множетва и определяют логические операции, связывающие подзапросы, через соответствующие теоретико-множественные операции. Основное внимание работы уделено алгоритму эффективного поиска k лучших объектов в рамках описанной модели. Также обсуждается возможность использования весов для изменения влияния отдельных подзапросв на основе предпочтений пользователя.

В работе [9] предлагается алгебра, позволяющай работать с основными задачами обработки запросов на основе подобия. Предложенная аглгебра позволяет рассматривать как нечеткие аттрибуты, так и нечеткие отношения. По-существу предлагается расширение реляционной алгебры, в котором определены операции выборки, проекции, объединения, соединения, разности, булевой разности, пересечения, выбора лучших и сужения. Авторы определяют основные важные свойства операций и две возможных реализации логических операций на базе теории нечетких множеств. На примерах показано, как строятся формальные выражения в рамках представленной алгебры по пользовательским запросам. Предпочтения пользователя задаются с помощью весов соответствующих предикатов.

1.4 Системы синтеза

В этом разделе мы будем рассматривать работы, посвященные изучению задачи комбинирования результатов независимых поисковых систем.

В работе [21] обсуждаются разновидности задачи синтеза результатов поиска. Синтезом данных называется слияние результатов нескольких поисковых систем, работающих с общим набором объектов. Задача синтеза данных несколько отличается от задачи синтеза коллекций, в которой комбинируются результаты поиска в различных или пересекающихся коллекциях [33,37].

Почти во всех рассмотренных нами статьях под результатом поиска понимается набор документов с оценками (scores) или рангами (ranks). Список документов с рангами представляет собой список документов, упорядоченный согласно их релевантности запросу. Результат поиска, представленный в виде множества документов с оценками, является для нас более интересным, так как оценки, помимо упорядоченности документов в результате, позволяют отражать и степень их релевантности запросу. Оценки призваны передавать знание о том, в какой степени один документ более или менее релевантен запросу по сравнению с другими.

Авторы [8] сравнивают качество слияния ранжированных списков методами CombMIN, CombMAX, CombSUM, CombANZ, CombMNZ. Метод CombMNZ показал наиболее хорошие результаты по точности синтезированного результата на рассмотренных экспериментальных данных. Также в статье предложено сравнение этих методов с использованием при слиянии не оценок, а рангов документов.

В работе [4] представлено два алгоритма синтеза результатов поиска для решения задачи мета-поиска: алгоритм, основанный на процедуре «демократического» голосования, и алгоритм на базе интерференции Байеса. Представленные в этой статье алгоритмы сравниваются с другими методами синтеза. Авторы предлагают метод оценки верхней границы эффективности методов синтеза, что позволяет в некоторой степени предсказывать качество результата синтеза. В своей работе мы анализировали свойства операций синтеза и калибровки, что позволило нам делать выводы об их эффективности при решении разных задач поиска.

В работах [20,21] обсуждаются ключевые интуиции в области слияния нескольких поисковых результатов, такие как «Эффект хора» (Chorus effect), «Эффект снятия сливок» (Skimming effect) и «Эффект темной лошадки» (Dark horse effect). Основным достижением работ является разработка способа слияния результатов нескольких поисковых систем на основе одних и тех же поисковых запросов и общей коллекции документов. На базе тестового набора запросов оценивается вероятность того, что документ на некоторой позиции в результате будет релевантен запросу. Такая процедура позволяет оценить шкалу оценок каждой отдельной поисковой системы. На следующем этапе тренировочной фазы вычисляется оценка каждого документа как сумма вероятностей релевантности документа на некоторой позиции во всех поисковых системах. Эксперименты с данными TREC-3 и TREC-5 показали, что предложенный метод слияния показывает лучшие результаты по сравнению с CombMNZ.

Следующая работа [12] интересна тем, что в ней предлагаются эффективные методы синтеза для разных типов поиска в видео коллекции. В рамках работы предложены методы построения комбинированной оценки на основе:

- произведения отдельных оценок (CombJointPr),
- суммы нормализованных оценок среди нескольких лучших результатов (CombSumScore),
- суммы нормализованных оценок среди нескольких лучших результатов (CombSumRank),

- взвешенного среднего нормализованных оценок среди нескольких лучших результатов (CombSumWtScore),
- взвешенного среднего нормализованных рангов среди нескольких лучших результатов (CombSumWtRank) и других.

Эксперименты показали, что CombSumScore хорошо работает при слиянии результатов поиска по одному изобразительному признаку на разных коллекциях данных. Метод CombSumWtScore дает наиболее хорошие результаты при комбнировании результатов поиска по текстовой аннотации и изобразительным признакам. Алгоритмы CombSumScore и CombSumRank получают адекватные результаты при синтезе списков, построенных на основе разных изобразительных признаков для общего запроса. Авторы показали, как сильно влияет характер конкретной задачи поиска на выбор наилучшего метода синтеза поисковых результатов.

Два сценария использования методов синтеза применительно к задаче поиска изображений было рассмотрено в работе [2]: использование синтеза для реализации поиска по частично аннотированной базе с текстовым запросом и синтез методов поиска по цветовым и текстурным характеристикам. Авторы разработали список требований к функции синтеза, подходящей для решения задачи поиска изображений. На основе детального анализа существующих методов синтеза было показано, что они не удовлетворяют всем сформулированным требованиям. В частности, они не позволяют учитывать степень доверия к тому или иному источнику информации, синтезируемому множеству. В [2] была предложена функция WTGF (взвешенное среднее с гравитационной функцией), удовлетворяющая данным требованиям. На основе экспериментов авторы продемонстрировали, что синтеза WTGF для ряда задач поиска изображений выигрывает у CombMNZ.

В работе [28] рассмотрены методы синтеза списков результатов поиска, которые позволяют учитывать иерархическую структуру документа. Статья посвящена способам вычисления оценки для составного элемента с потомками разной природы. На этапе синтеза комбинируются оценки элементов, составляющих документ из поисковой коллекции.

При синтезе списков элементов с оценками, полученными из разных источников необходимо в первую очередь решить задачу нормализации оценок. В работах [8,12,38] нормализация происходит по степени отклонения оценки элемента от минимальной оценки в данном списке. Несколько альтернативных методов нормализации оценок обсуждаются в [38].

Классификация методов синтеза данных по областям применения, целям комбинирования, архитектурам систем слияния и используемому математическому аппарату представлена в [31]. Теория вероятности, нечеткие множества, нейронные сети перечислены среди используемых в литературе математических основ предлагаемых методов синтеза данных. Авторы считают, что следующим шагом в развитии этой области будет извлечение знаний, связанных с построением новых методов синтеза.

1.5 Нечеткие множества

Понятие нечеткого множества часто применяется при синтезе данных и коллекций. Например, в работах [24, 36] обсуждается использование нечетких множеств для решения задачи синтеза коллекций, то есть мета-поиска.

В работе [26] представлена модель поиска на основе теории нечетких множеств. Авторы предлагают теоретическое обоснование согласованности модели нечетких множеств и модели информационного поиска. В основе предложенного авторами подхода лежит описание терма с помощью нечеткого множества, где нечеткое множество строится по инвертированной частоте использования слова в документах.

В работе [17] при решении задачи поиска текстовых документов предлагается использовать нечеткие множества. Каждый терм в тексте представляется в виде нечеткого множества. Оценка документа строится га основе пересечения или объединения нечетких множеств, построенных по словам в тексте. Авторы обсуждают идею использования различных весов при построении нечетких множеств и их автоматического или ручного подбора на основе обратной связи от пользователя. Аналогичная идея представления термов в документе в виде нечетких множеств рассмотрена в работе [42].

2 Основные понятия

2.1 Подобие и расстояние

Поисковая система, как правило, не работает непосредственно с объектом, а строит на его основе поисковый образ объекта, или дескриптор. Образ конструируется таким образом, чтобы он достаточно точно описывал собой соответствующий объект. Основной идеей такого преобразования является предположение, что поисковые образы отражают природу соответствующих объектов. Поисковая система работает с поисковыми образами, но нам на уровне абстрактной модели интересны объекты и их релевантность запросу.

В информационном поиске образы объектов строятся на основе различных признаков объектов и, как правило, представляют собой n-мерные вектора. Например, в системах поиска изображений по содержанию такие вектора могут строиться на основе различных цветовых или текстурных признаков, а также их комбинаций. Способ построения поискового образа во многом определяет качество всей дальнейшей работы с объектами, но эти вопросы остаются вне рамок нашего исследования.

В дальнейшем для построения запросов поисковая система будет оценивать релевантность только на основе поисковых образов, но на уровне абстрактной модели мы будем говорить об объектах, по которым они построены. Нас будет интересовать задача комбинирования запросов, но не поисковых образов.

После того, как внутри поисковой системы объекты заменяются своими поисковыми образами, задача оценки релевантности объекта запросу сводится к оценке релевантности поисковых образов запросу. Как мы уже упоминали во введении большинство подходов к решению задачи поиска основано на использовании понятия подобия.

Определение.

Мерой подобия на множестве S называется функция $sim: SxS \rightarrow [0,1] \ :\forall a, b \in S$ $sim(a,a) = 1,\ sim(a,b) = sim(b,a)$. Совокупность функций подобия, определенных на множестве S, обозначим через $Similarities(S)$.

Функция подобия может возникать из разных источников и определяться на множествах объектов самой разнообразной природы. В литературе рассматривается целый ряд мер подобия в пространствах поисковых образов, например, косинусная мера и меры на основе метрик векторных пространств. Как правило, конкретная мера подобия также как и способ построения поисковых образов объектов выбирается индивидуально для раз-

ных коллекций данных. С точки зрения поисковой системы, образы релевантных друг другу объектов будут близки в терминах меры подобия, определенной для описания конкретного набора данных. Например, при поиске изображений по образцу на основе поисковых образов, построенных по цветовым гистограммам, используется мера подобия на основе Евклидовой метрики.

Таким образом, в предположении о том, что релевантные объекты близки в терминах меры подобия, задача об оценке степени релевантности объектов друг другу сводится к вычислению меры подобия.

Близость между объектами может быть выражена не только в терминах меры подобия, но и с помощью функции расстояния между объектами.

Определение.

Функцией расстояния или псевдометрикой на множестве S называется функция $dist: SxS \rightarrow R$: $\forall a, b \in S\ dist(a,b) \geq 0,\ dist(a,b) = dist(b,a),\ dist(a,a) = 0$. Совокупность функций расстояния, определенных на множестве S, обозначим через $Distances(S)$.

Многие алгоритмы и модели поиска также опираются на неравенство треугольника ($dist(a,b) \leq dist(a,c) + dist(c,b)$), которому удовлетворяют далеко не все используемые на практике функции расстояния. Наша абстрактная модель будет основана на свойствах из определения псевдометрики.

Выбор функций расстояния для разнообразных объектов также широк: геометрические, статистические, строковые функции расстояния. Заметим, что на множестве поисковых образов может быть определено несколько функций расстояния или мер подобия, определяющих разные пространства поиска.

Важно заметить, что если на множестве определена мера подобия, то по ней можно построить некоторую функцию расстояния, и наоборот.

Определение.

Операцией построения меры подобия по функции расстояния называется отображение $similaring: Distances(S) \rightarrow Similarities(S)$: если $dist(a,b) < dist(c,d)$, то $sim(a,b) > sim(c,d)$, где $sim = similaring(dist)$.

Операцией построения функции расстояния по мере подобия называется отображение $distancing: Similarities(S) \to Distances(S)$: если $sim(a,b) < sim(c,d)$, то $dist(a,b) > dist(c,d)$, где $dist = distancing(sim)$.

Отображение между множествами функций расстояния и функций подобия может быть реализовано по-разному. Определяемые нами операции будут строить синтетические функции расстояния и подобия, как правило, не совпадающие с естественными.

Определение.

Операция $similaring$ по функции расстояния $dist \in Distances(S)$ строит функцию подобия $sim \in Similarities(S)$ следующим образом: $\forall a, b \in S\ sim(a,b) = \frac{1}{1+dist(a,b)}$.

Для восстановления функции расстояния по функции подобия достаточно просто строится обратная операция.

Определение.

Операция $distancing$ по функции подобия $sim \in Similarities(S)$ строит функцию расстояния $dist \in Distances(S)$ следующим образом: $\forall a, b \in S\ dist(a,b) = \frac{1}{sim(a,b)} - 1$.

2.2 Запросы и результаты

При решении любой задачи поиска рассматривается множество объектов и запрос. В литературе и промышленности рассматривается два основных класса поисковых систем:

- системы, возвращающие точный ответ на запрос пользователя,
- системы, возвращающие оценки объектов с точки зрения их релевантности запросу.

В наших исследованиях нас будет интересовать задача второго класса поисковых систем, состоящая в том, чтобы оценить степень релевантности каждого объекта запросу.

Классы задач тесно связаны между собой. Когда мы говорим о задаче информационного поиска как об оценке релевантности объектов, то при построении результата важны объекты с высокими оценками. В связи с этим поисковые системы пытаются найти объекты с высокими оценками и не вычислять оценки для всех объектов в коллекции с целью повышения эффективности вычислений. Результат, который получает пользова-

тель, содержит объекты с лучшими оценками и с этой точки зрения оказывается похожим на точный.

Поисковая система не может существовать без запросов, которые выражают потребность пользователя. В рамках работы будет построена модель, описывающая обе парадигмы выполнения запросов и позволяющая комбинировать различные типы подзапросов внутри одного запроса. Центральным понятием нашей модели будет понятие Q-множества. Q-множество является абстракцией над языком запросов, природой объектов в пространстве поиска, методом выполнения запросов, и инкапсулирует (включает в себя) запрос и результат его выполнения.

Определение.

Q-множеством называется множество S вместе с определенной на нем функцией оценки $score: S \to [0,1]$. Класс Q-множеств, определенных на множестве S, обозначим через $Scores(S)$.

Оценкой объекта $s \in S$ в Q-множестве q называется значение функции оценки $score_q(s)$.

Введенное нами определение Q-множества позволяет абстрагироваться от конкретного способа вычисления ответа на запрос. Представления пользователя о запросе и ответе на него возникают одновременно, поскольку человек, формулируя запрос, изначально вкладывает в него некоторое представление о том, как должен выглядеть результат, то есть набор релевантных запросу объектов. Внутри поисковой системы ответ на запрос также заранее предопределен моделью поисковой системы.

Любой запрос, независимо от его природы, структуры, подразумевает, что объектам из множества будут присвоены некоторые оценки, описывающие степень соответствия, релевантности объекта запросу. Пользовательский запрос может представлять собой текст, изображение, некоторый объект с множеством атрибутов, но природа запроса с точки зрения поисковой системы всегда может быть описана некоторой функцией оценки. Важно подчеркнуть, что в рамках работы будет построен набор операций, основанный на таком представлении любого запроса в виде Q-множества независимо от природы объектов и метода построения их оценок.

В этой работе мы используем оценки, все значения которых лежат в отрезке [0,1].

Введенное выше понятие Q-множества по существу совпадает с понятием нечеткого множества, предложенного Заде ([41]). Напомним, что в теории нечетких множеств

рассматриваются множества, на которых определена функция принадлежности элемента множеству. Для таким образом определенных множеств (нечетких множеств) определяются операции на множествах: объединение и пересечение.

Операции и построения теории нечетких множеств будут использованы в нашей модели. Этот набор операций будет расширен дополнительными операциями, поскольку обычные теоретико-множественные операции не обладают многими важными свойствами, которые будут обсуждаться позднее.

Нечеткие множества допускают вероятностную интерпретацию функции принадлежности. Поэтому значения предложенной нами функции оценки могут быть интерпретированы как вероятности того, что объект релевантен запросу.

Определенное таким образом абстрактное понятие Q-множества может включать в себя самые разнообразные типы запросов, например запросы, выраженные с помощью точных языков запросов, или запросы в вероятностных базах данных.

Основным объектом нашей абстракции является Q-множество и соответствующая ему функция оценки и в рамках работы мы будем строить систему операций на множестве Q-множеств, с помощью которой можно будет описывать некоторые задачи поиска. Существующие системы поиска используют понятия расстояния или подобия, и возникает задача сопоставления этих понятий с понятием Q-множества.

Источником первичных Q-множеств могут служить обычные поисковые запросы на основе подобия. Для любого объекта-запроса q ∈ S можно построить Q-множество по любой функции подобия, определенной на множестве S.

Определение.

Функция scoring по объекту $q \in S$, и мере подобия $sim \in Similarities(S)$ строит Q-множество с функцией оценки, определенной следующим образом: $\forall a \in S\ score(a) = sim(q, a)$.

При поиске изображений по образцу из коллекции выбираются изображения, поисковые образы которых ближе к поисковому образу изображения-запроса с точки зрения определенной меры подобия.

Предыдущее определение показывает, как Q-множество можно описать некоторой точкой в пространстве и функцией подобия. Синтетические объект-запрос и функция по-

добия также могут быть сконструированы по Q-множеству. При решении практических задач поиска все происходит в обратном порядке, то есть сначала строятся функция подобия и объект-запрос и уже по ним восстанавливается Q-множество, функция оценки.

Основным элементом в нашей модели является Q-множество, то есть функция оценки, которая каждому объекту из множества сопоставляет оценку. Заметим, что в рамках нашей модели природа объектов в множестве не имеет значения и никак не учитывается. В действительности происхождение объекта и его свойства могут быть учтены при вычислении функции расстояния или подобия (на уровне поисковых образов), на основе которых строится запрос, а соответственно и в оценке, присвоенной объекту.

В рамках представленной работы нас будут интересовать методы работы с Q-множествами и способы синтеза новых Q-множеств, запросов на основе имеющихся.

2.3 Свойства синтеза

В области информационного поиска проработано огромное количество моделей поиска на основе разных алгоритмов и методов представления информации. Каждый конкретный подход превосходит по качеству результата другие, но только в рамках отдельной задачи поиска. В связи с этим возникла задача синтеза нескольких результатов поиска с целью повышения производительности всей системы в целом. В литературе представлен достаточно широкий спектр работ, проверяющих гипотезу о том, что слияние нескольких результатов поиска позволяет повысить качество работы поисковой системы.

Задача синтеза, как правило, рассматривается в одном из двух сценариев ([21]):

- Результат нескольких поисковых систем синтезируется в один (мета-поиск и синтез коллекций);
- Результат выполнения нескольких запросов в одном поисковом пространстве комбинируется в конечный результат. В качестве примера могут служить функции подобия на основе цвета и текстуры при поиске изображений по содержанию.

Последнюю задачу можно также решать с помощью построения сложных дескрипторов и функций подобия. Тем не менее, задача синтеза сильно отличается от подходов, основанных на комбинировании разных признаков объектов на уровне построения поисковых образов. В этой работе нас интересуют именно методы синтеза, поскольку:

- Использование разных признаков внутри одного дескриптора требует сложного процесса нормализации для того, чтобы существующие меры подобия смогли адекватно описывать семантическую близость между объектами;
- Построение дескрипторов объектов на основе нескольких признаков, как правило, сильно увеличивает размерность вектора поискового образа, что сильно увеличивает вычислительную сложность поисковых алгоритмов.

Таким образом, подход к комбинированию различных признаков объектов с помощью решения задачи синтеза оказывается более гибким.

При решении задачи синтеза необходимо учитывать ряд интуитивных представлений пользователя о синтезе результатов поиска. Соответствующие свойства методов синтеза в литературе принято называть эффектами ([20]). Первым свойством является «Эффект хора». Основная идея этого принципа заключается в том, что хорошие объекты по двум признакам должны быть в итоге лучше, чем хорошие только по одному. Например, если изображение близко к изображению-запросу и по текстуре и по цвету, то оно более релевантно запросу, чем то, которое очень похоже на изображение-запрос по цвету, но совсем не похоже по текстуре. Вторым свойством является «Эффект снятия сливок», который состоит в том, что хорошие объекты в смысле хотя бы одного из синтезируемых множеств должны быть хорошими и в результате.

В терминах определенного выше понятия Q-множества задача синтеза двух Q-множеств заключается в построении новой функции оценки на основе первоначальных. При решении задачи синтеза не просто строится новое Q-множество, содержащее объекты из сливаемых множеств. Необходимо корректно строить новую функцию оценки таким образом, чтобы содержащаяся в исходных оценках информация не была утрачена, и любая полученная оценка влияла на Q-множество, полученное в результате слияния.

В разных задачах поиска возникают те или иные информационные потребности и, следовательно, не существует универсального метода синтеза двух запросов. В зависимости от конкретной задачи алгоритмы и формулы, реализующие операцию синтеза должны по-разному учитывать «Эффект хора» и «Эффект снятия сливок». На уровне абстрактной модели операция синтеза устроена одинаково, но методы ее реализации могут быть разными.

С нашей точки зрения, согласно «Эффекту хора», объекты, получившие достаточно хорошую оценку в обоих Q-множествах, должны получить еще более хорошую оценку

после их слияния. Такой способ учета «Эффекта хора» при синтезе Q-множеств позволит сравнивать оценки объектов до и после слияния. Таким образом, тот факт, что объект получил оценку в обоих множествах, а не только в одном, должен усилить синтезированную оценку этого объекта, то есть его синтезированная оценка будет превосходить первоначальные.

Предложенный метод решения задачи синтеза должен учитывать также «Эффект снятия сливок», чтобы объекты с высокими оценками из каждого синтезируемого Q-множества получали в результате более высокие оценки, чем объекты с более низкими оценками. То есть высокие синтезированные оценки для объектов получаются при синтезе таких Q-множеств, что хотя бы в одном из них объект получил высокую оценку.

2.4 Операции модели

Важно отметить, что все предложенные операции на множестве Q-множеств не выводят нас за рамки этого класса. То есть после применения операции мы вновь получаем Q-множество, к которому можно будет применять весь набор операций. Наш класс запросов замкнут относительно предложенного набора операций.

При решении задачи поиска полезно иметь возможность строить синтезированные запросы, которые учитывают первоначальные функции оценок. Представленная ниже операция модели определяет решения задачи синтеза в терминах понятия Q-множества.

Определение.

Синтезом двух Q-множеств называется функция $fusion : Scores(S) \times Scores(S) \rightarrow Scores(S)$: $\forall a, b \in Scores(S)$ $fusion(a, b) = fusion(b, a)$, сохраняет упорядоченность (если $a(x) < a(y)$, $b(x) < b(y)$, то $fusion(a, b)(x) < fusion(a, b)(y)$).

Функция синтеза вместе с необходимыми свойствами, представленными в определении, может обладать дополнительными свойствами: удовлетворять «Эффекту хора»; удовлетворять «Эффекту снятия сливок». Качество функции синтеза зависит от требований к системе, которые определяются конкретной задачей поиска, и от того в какой степени конкретная реализация операции синтеза удовлетворяет этим свойствам. Важно подчеркнуть, что в зависимости от конкретной задачи поиска важность дополнительных свойств функции синтеза может быть различна.

При таком решении задачи синтеза Q-множеств возникает задача сопоставления оценок разной природы. Например, если при поиске изображений по содержанию, синте-

зируются Q-множества на основе цвета и текстуры. Функции расстояния на соответствующих пространствах поисковых образов будут разные, и поэтому, как правило, получаемые оценки будут несопоставимы. Поэтому на уровне нашей абстрактной модели мы будем говорить об операциях на Q-множествах и способах построения оценок, которые были бы в некотором смысле сопоставимы между собой.

Мы считаем функции оценки сопоставимыми, если соответствующие оценки важных, значимых объектов отличаются незначительно. Как правило, значимыми объектами при поиске являются объекты с наибольшими оценками в Q-множестве.

Для решения задачи приведения функций оценки к сопоставимым рассматривается задача калибровки оценок Q-множества. В рамках этой работы мы рассмотрим две операции, которые можно будет использовать при калибровке Q-множеств: нормализация и усиление.

Под нормализацией будем понимать пропорциональное изменение длины диапазона значений оценок в Q-множестве или значений соответствующей функции расстояния.

Для сопоставимости оценок объектов необходимо в первую очередь привести оценки к одному множеству значений, но определение функции оценки уже содержит в себе это ограничение.

Важно отметить, что сопоставимость оценок в первую очередь важна на правой части диапазона значений функции оценки, поскольку именно объекты с оценками из этого диапазона в первую очередь формируют конечный набор объектов, который увидит пользователь.

Результатом развития этой идеи стала операция нормализации, которая позволяет приводить разнообразные оценки к сопоставимым. На самом деле, операция нормализации строит новое Q-множество, в котором оценки также описывают множество объектов, но лучше сопоставимы с оценками другого Q-множества.

Определение.

Операцией нормализации называется монотонная неубывающая функция $norm: Scores(S) \to Scores(S)$: не изменяет соотношения между оценками внутри Q-множества.

Введенная операция нормализации работает независимо от того, каким образом было получено первоначальное Q-множество, поскольку на абстрактном уровне все Q-множества устроены одинаково.

Чтобы после применения операции нормализации оценки объектов сохранили свои свойства, то есть новые построенные оценки сохранили информацию об объектах, операция нормализации должна монотонно не убывать.

Также как и операция синтеза, функция нормализации может обладать набором дополнительных свойств, которые влияют на качество нормализации в зависимости от конкретных требований задачи поиска. Операция нормализации может обладать тем свойством, что результирующая функция оценки сильнее зависит от хороших оценок в первоначальном Q-множестве, чем от плохих. Это свойство обосновано тем, что при построении результата объекты с более высокими оценками оказываются более значимыми.

К сожалению, для сопоставления и сравнения Q-множеств с функциями оценки разной природы не достаточно общего диапазона значений оценок. Например, в случае, когда все оценки первого Q-множества почти равны 1, а оценки другого Q-множества близки к 0, функции оценки не сопоставимы. Поэтому для реализации сопоставимости оценок объектов на нашем уровне абстракции необходимо ввести еще одно ограничение на операцию нормализации. При нормализации выбросы, то есть слишком большие и слишком маленькие оценки, не должны сильно влиять на результирующие нормализованные оценки. Таким образом, вторым дополнительным свойством операции нормализации является то, что она не должна быть чувствительна к выбросам.

При решении задачи синтеза Q-множеств также возникает необходимость учитывать разное качество оценок. Задача состоит в том, чтобы изменять величину оценок того или иного Q-множества в зависимости от их важности, например, усиливать оценки одного и ослаблять оценки другого. Методы решения задачи усиления должны усиливать высокие оценки в Q-множестве и ослаблять низкие. В рамках этого требования была построена операция усиления и ослабления Q-множеств, то есть функций оценки. Эта пара операций позволяет изменять относительное влияние оценок с учетом их важности.

Например, если при поиске изображения по образцу заранее известно, что для пользователя текстура важнее цвета изображения, то встает задача усиления оценок, построенных на основе текстуры. В этом примере для решения задачи требуется знание о том, как построены оценки изображения, хотя как уже говорилось ранее, вводимые нами

операции не должны зависеть от природы оценок. Автоматическое усиление оценок высокого качества возможно в ситуации, когда имеется некоторый тренировочный набор данных, близкий по своим свойствам к базовой коллекции, на основе которого можно проанализировать разные виды функций оценки.

Для извлечения информации о важности Q-множеств можно использовать relevance feedback (обратную связь) от пользователя. Во многих работах обсуждаются способы автоматического подбора весов оценок при синтезе запросов на основе relevance feedback. Вводимая нами операция позволит изменять значимость оценок более естественных образом.

Определение.

Операцией усиления называется отображение $strengthen: Scores(S) \times [0,1] \rightarrow Scores(S)$: $\forall a, b \in S, \forall score \in Scores(S), \forall level \in [0,1]$ выполняются следующие свойства:

если $score(a) \leq score(b)$, то

$$strengthen(score(a), level) \leq strengthen(score(b), level);$$

$strengthen(score(a), level) \geq score(a)$, если $|\{y: score(y) \geq score\ (a)\}| \leq level * |S|$;

$strengthen(score(a), level) \leq score(a)$, если $|\{y: score(y) \geq score\ (a)\}| > level * |S|$.

Определение.

Операцией ослабления называется отображение $weaken: Scores(S) \times [0,1] \rightarrow Scores(S)$: $\forall a, b \in S, \forall score \in Scores(S), \forall level \in [0,1]$ выполняются следующие свойства:

если $score(a) \leq score(b)$, то $weaken(score(a), level) \leq weaken(score(b), level)$;

$weaken(score(a), level) \leq score(a)$, если $|\{y: score(y) \geq score\ (a)\}| \leq level * |S|$;

$weaken(score(a), level) \geq score(a)$, если $|\{y: score(y) \geq score\ (a)\}| > level * |S|$.

Следующими операциями, которые мы ввели на уровне нашей модели, являются операции усиления и ослабления. Операция ослабления является обратной к операции усиления. Операции усиления или ослабления монотонно изменяют оценки объектов в множестве. Операция усиления, как и операция ослабления, применяется к Q-множеству и

возвращает новое Q-множество. Операция усиления сокращает маленькие оценки и усиливает большие оценки внутри одного Q-множества.

Таким образом, важными свойствами операций калибровки являются:

- Чувствительность к выбросам (оценка отдельного объекта не должна сильно влиять на результат калибровки);
- Скос (высокие оценки внутри Q-множества сильнее влияют на результат калибровки);
- Эффективность (разница в качестве Q-множеств, являющихся аргументами операции синтеза, принимается во внимание при калибровке).

3 Реализация модели

В этом разделе будет представлена конкретизация операций нашей модели и некоторые конкретные формулы и алгоритмы, которые выражают наши операции и отвечают необходимым свойствам, а также порождают новые. В результате будут представлены некоторые интуитивно правильные формулы, описывающие наши операции, но их применимость при решении реальных задач будет обсуждаться и проверяться чуть позже.

3.1 Унарные операции

3.1.1 Нормализация

В рамках этой работы будет рассмотрено несколько подходов к реализации операции нормализации.

В качестве первого метода нормализации оценок мы рассмотрели стандартную процедуру, используемую в литературе при решении задачи синтеза для разных коллекций данных. Как правило, эта операция используется для приведения диапазонов значений оценок к сопоставимым, что не является необходимым в нашей модели, так как все оценки принадлежат отрезку от 0 до 1 по определению. Тем не менее, важным преимуществом этой операции является наличие в нормализованном Q-множестве объектов с максимальной, то есть равной 1, и минимальной оценкой.

Определение.

Операция $norm - maxmin$ по запросу $score \in Scores(S)$ строит запрос $score' \in Scores(S)$: $\forall e \in S\ score'(e) = \frac{score(e) - min(score(x))}{max(score(x)) - min(score(x))}$.

Во втором подходе Q-множества нормализуются с помощью приведения среднего значения расстояний между всеми объектами в множестве к фиксированному числу. Преимущество этой реализации операции нормализации состоит в том, что такой метод нормализации оценок не чувствителен к выбросам.

Определение.

Операция $norm - avg$ по запросу $score \in Scores(S)$ строит запрос $score' \in Scores(S)$: $\forall e \in S\ score'(e) = similaring(\frac{distancing(score(e))}{avg(distancing(score(x)))})$.

Прежде, чем обсуждать остальные рассматриваемые нами методы нормализации, важно отметить, что при построении результата в первую очередь рассматриваются объекты с большими оценками в Q-множестве. Используемые методы приведения функций

оценки к сопоставимым могут приводить к близким распределения оценок не на всем множестве значений, а только на той его части, куда попадают объекты с наибольшими оценками. В следующих реализациях операции нормализации мы будем автоматически подбирать значения параметров так, чтобы учесть это наблюдение.

Третий метод нормализации оценок позволяет точнее калибровать Q-множества с целью приведения распределений значений соответствующих функций оценок к сопоставимым. Для нормализации оценок соответствующие расстояния на уровне, отделяющем «важные» значения расстояний, приводятся к сопоставимым. Попробуем формализовать это утверждение.

Определение.

Операция $norm - dist(A)$ по запросу $score \in Scores(S)$ строит запрос $score' \in Scores(S)$: $\forall e \in S\ score'(e) = scoring(distancing(score(e)) * A)$, где A - некоторый вещественный параметр.

Определение.

Параметр $p \in [0,1]$ определяет процент объектов с наибольшими оценками и наименьшими расстояниями, который считается «важным» для пользователя при поиске. Процедура $normalize - dist_p$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S\ \ score(e) = norm(A)(score_1(e))$, где $A = \frac{distancing(score_2(a))}{distancing(score_1(b))}$, где a и b: $|\{y: score_2(y) \geq score_2(a)\}| = n * |S|$ и $|\{y: score_1\ (y) \geq score_1\ (b)\}| = n * |S|$.

Преимущество этого метода калибровки Q-множеств состоит в том, что при построении результата в первую очередь рассматриваются объекты с большими оценками.

Четвертый метод нормализации оценок по своей идее близок к третьему, но к сопоставимым приводятся непосредственно оценки, а не соответствующие им расстояния. Этот подход иногда бывает технически более предпочтителен, если система работает только с оценками, так как не возникает необходимости строить соответствующие расстояния.

Определение.

Операция $norm - score(B)$ по функции оценки $score \in Scores(S)$ строит функцию оценки $score' \in Scores(S)$: $\forall e \in S\ score'(e) = score(e)\ * B$, где B некоторый вещественный параметр.

Определение.

Параметр $p \in [0,1]$ определяет процент объектов с наибольшими оценками и наименьшими расстояниями, который считается «важным» для пользователя при поиске. Процедура $normalize - score_p$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S\ score(e) = norm(B)(score_1(e))$,где $B = score_2(a)/score_1(b)$, где a и b: $|\{y: score_2(y) \geq score_2(a)\}| = n * |S|$ и $|\{y: score_1(y) \geq score_1(b)\}| = n * |S|$. Заметим, что $B = \frac{score_1(b)+1}{A+score(b)+1}$, где A – коэффициент из предыдущего определения процедуры калибровки.

3.1.2 Усиление и ослабление

Операция ослабления вычисляется следующим образом: построение нового Q-множества, в множество элементов которого попадут все объекты из ослабляемого множества, а функция оценки которого будет определена с помощью возведения в степень функции расстояния ослабляемого Q-множества. Оценки объектов, которые попали в некоторый процент лучших, то есть высокие, уменьшатся, а остальные оценки - увеличатся.

Определение.

Операция ослабления Q-множества $weaken(n)$ по функции оценки $score \in Scores(S)$ и параметру $level \in [0,1]$ строит Q-множество с функцией оценки $score' \in Scores(S)$: $\forall e \in S\ score'(e) = similaring((distancing(score(e))/M)^n)$, где $n < 1$- параметр процедуры, $M: |\{y: distancing(score(y)) \leq M\}| = level * |S|$.

Операция усиления строится аналогично операции ослабления.

Определение.

Операция усиления Q-множества $strengthen(n)$ по функции оценки $score \in Scores(S)$ и параметру $level \in [0,1]$ строит Q-множество с функцией оценки $score' \in Scores(S)$: $\forall e \in S\ score'(e) = similaring((distancing(score(e))/M)^n)$,), где $n > 1$- параметр процедуры, $M: |\{y: distancing(score(y)) \leq M\}| = level * |S|$.

Другими словами, операция $weaken(n)$ определяется как $strengthen(1/n)$.

Важно отметить, что действие операций усиления и ослабления сильно зависит от распределения значений расстояний относительно параметра $level$. Например, после нормализации оценок с помощью операции $norm - avg$, операция усиления с параметром

$level$: $M = 1$ усилит оценки, соответствующие расстояния которых выше среднего внутри заданного Q-множества.

3.1.3 Дополнение

Определение.

Операция построения дополнения к Q-множеству $complement$ по функции оценки $score \in Scores(S)$ строит новую функцию оценки $score' \in Scores(S)$: $\forall e \in S$ $score'(e) = 1 - score(e)$.

Операция complement строит Q-множество, у которого функция оценки является дополнением к первоначальной. Эта операция описывает построение такого запроса, ответом на который являются те объекты, которые нерелевантны первоначальному запросу.

В рамках этой работы операция построения дополнения к Q-множеству будет использоваться в качестве вспомогательной для других операций.

3.1.4 Дискретизация

Еще одной операцией на классе Q-множеств является операция дискретизации. Результатом этой операции является точное Q-множество, то есть функция оценки, которая принимает значения на множестве $\{0,1\}$.

Определение.

Операция дискретизации строит по функции оценки $score \in Scores(S)$ и параметру $threshold \in [0,1]$ новую функцию оценки $score' \in Scores(S)$, $\forall e \in S$

$$score'(e) = \begin{cases} 1, \textit{если } \text{score(e)} \geq \text{threshold} \\ 0, \textit{иначе} \end{cases}$$

Операция дискретизации преобразует Q-множество к форме похожей на представление точного запроса и его результата в виде Q-множества. Эта операция может быть использована, чтобы исключить объекты с низкими оценками из дальнейшего рассмотрения внутри Q-множества. Чтобы сохранить оценки объектов остающихся в Q-множестве, необходимо пересечь результат дискретизации с первоначальным Q-множеством с помощью операции пересечения нечетких множеств. Операцию дискретизации можно использовать на разных этапах работы с Q-множествами, а не только на этапе получения конечного списка результатов.

3.2 Бинарные операции

Все ниже определенные бинарные операции обобщаются на операции над несколькими Q-множествами. Рассмотренные операции являются примерами реализации операции синтеза ($fusion$).

Как уже упоминалось выше, понятие Q-множества по существу совпадает с понятием нечеткого множества. Некоторые из определяемых ниже операций унаследованы из теории нечетких множеств.

3.2.1 Объединение

Определение.

Операция объединения $union$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S\ score(e) = max(score_1(e), score_2(e))$.

Объединение Q-множеств определено через объединение соответствующих нечетких множеств. Заметим, что в литературе ([9]) на такое определение операции объединения нечетких множеств ссылаются как на стандартное (fuzzy standard).

Утверждение.

Операция $union$ коммутативна, ассоциативна, сохраняет упорядоченность, не удовлетворяет «Эффекту хора», не учитывает низкие оценки внутри синтезируемых Q-множеств и удовлетворяет «Эффекту снятия сливок».

Поведение операции объединения можно проиллюстрировать диаграммой, представленной на рисунке 1. По осям наложена равномерная сетка, представляющая оценки двух синтезируемых Q-множеств, таким образом, точки плоскости представляют собой комбинацию двух оценок. На графике показаны изолинии, содержащие точки, имеющие одинаковое значение синтезированной оценки. Все последующие реализации операции fusion будут также проанализированы на основе аналогичных диаграмм. По диаграмме 1 видно, что одинаково высокую синтезированную оценку может получить объект, у которого высокая одна из оценок или обе оценки высокие.

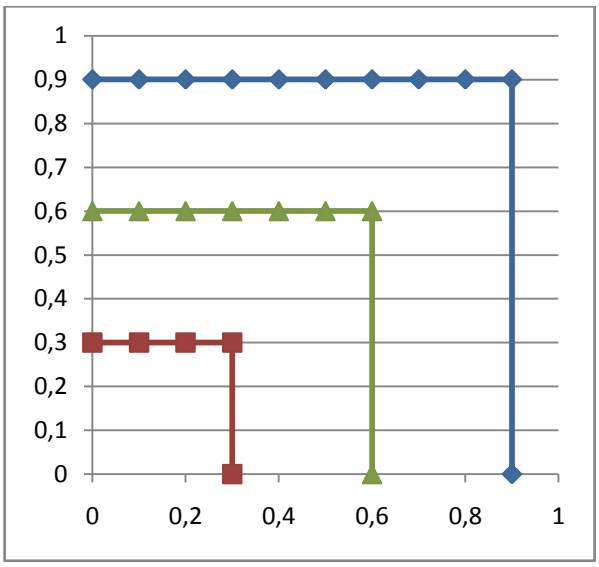

Рисунок 1 Изолинии, иллюстрирующие поведение операции объединения.

3.2.2 Пересечение

Определение.

Операция пересечения $intersect$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S\ score(e) = min(score_1(e), score_2(e))$.

Утверждение.

Операция $union$ коммутативна, ассоциативна, сохраняет упорядоченность, не удовлетворяет «Эффекту хора», не учитывает низкие оценки внутри синтезируемых Q-множеств и не удовлетворяет «Эффекту снятия сливок».

Следующая диаграмма на рисунке 2 аналогично предыдущей показывает процесс синтеза и демонстрирует характер поведения операции пересечения. Операция intersect по сравнению с операцией объединения ведет себя несколько иначе. Высокую синтезированную оценку может получить только объект, у которого обе исходные оценки высокие.

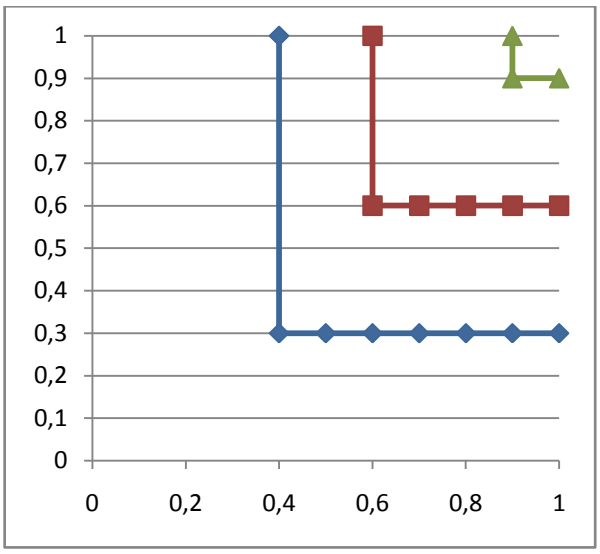

Рисунок 2 Изолинии, иллюстрирующие поведение операции пересечения.

Недостатком операций объединения и пересечения является то, что в синтезированном Q-множестве учитывается только одна из первоначальных оценок: наибольшая или наименьшая.

3.2.3 Супер-пересечение

Определение.

Операция супер-пересечения $super - intersect$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S \quad score(e) = score_1(e) * score_2(e)$.

Для реализации операции супер-пересечения мы использовали алгебраическое определения операции пересечения нечетких множеств (fuzzy algebraic) ([9]).

При таком методе синтеза двух Q-множеств учитываются все оценки синтезируемых множеств, но получаемая результирующая оценка получается достаточно маленькой и для дальней работы с Q-множеством, скорее всего, потребуется использование операции усиления.

Утверждение.

Операция $super - intersect$ коммутативна, ассоциативна, сохраняет упорядоченность, не удовлетворяет «Эффекту хора» и не удовлетворяет «Эффекту снятия сливок».

Важно отметить, что операция $super - intersect$ удовлетворяет «Эффекту хора», то есть учитывает обе синтезируемые оценки, но комбинированная оценка получается ниже исходных, что не соответствует нашему определению этого свойства.

Изолинии, изображенные на рисунке 3, показывают, что предложенная операция отражает интуитивное понимание синтеза двух Q-множеств. При использовании супер-пересечения на результирующую оценку влияют обе первоначальные оценки, что более естественно по сравнению с обычным пересечением. В терминах пользователя при поиске будет учитываться и первый запрос и второй. Например, при поиске изображений по текстуре и цвету, пользователю нужно вернуть в качестве результата только изображения похожие на запрос одновременно по текстуре и цвету.

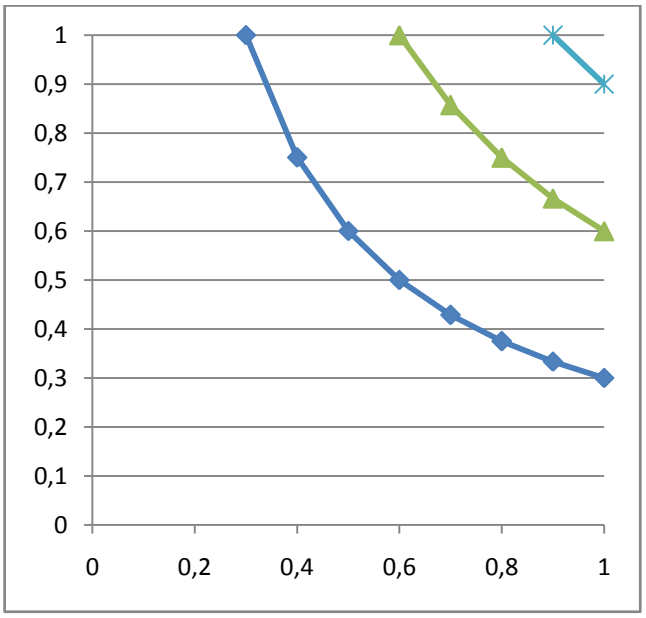

Рисунок 3 Изолинии, иллюстрирующие поведение операции супер-пересечения.

3.2.4 Супер-объединение

Определение.

Операция супер-объединения $super - union$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S \quad score(e) = 1 - (1 - score1e*(1-score2e)$.

Утверждение.

Операция $super - union$ коммутативна, ассоциативна, сохраняет упорядоченность, удовлетворяет «Эффекту хора» и «Эффекту снятия сливок».

Операция супер-объединения несколько иначе учитывает синтезируемые Q-множества (рисунок 4). Эта операция синтеза учитывает хотя бы одну из первоначальных оценок. Например, при поиске изображений по текстуре и цвету, в коллекции может не быть изо-

бражений похожих на запрос и по цвету и по текстуре одновременно, но пользователю можно вернуть в качестве результата изображения похожие по одному из признаков.

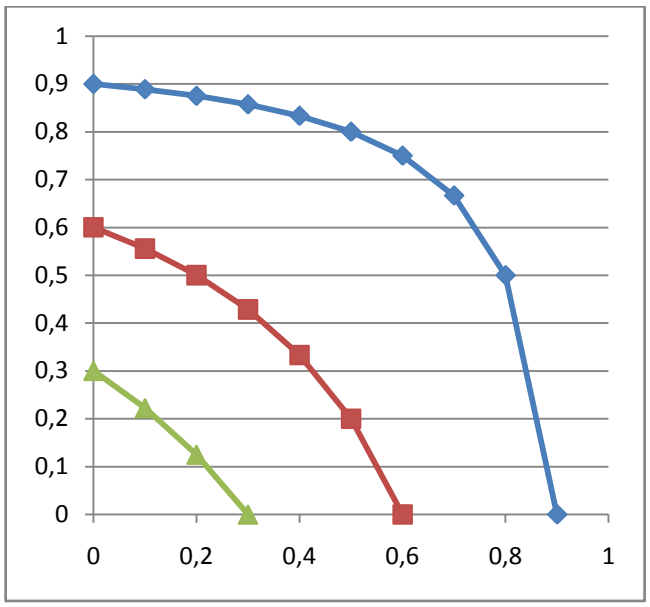

Рисунок 4 Изолинии, иллюстрирующие поведение операции супер-объединения.

Использование операций супер-пересечения и супер-объединения позволяет сохранять первоначальную вероятностную интерпретацию оценок, если синтезируемые Q-множества независимы.

3.2.5 CombMNZ

В наших экспериментах мы будем сопоставлять действие описанных операций синтеза (объединения, пересечения, супер-объединения и супер-пересечения) с решением задачи синтеза, которое считается базой для сравнения – операцией $CombMNZ$.

Определение.

Операция $CombMNZ$ по Q-множествам $score_1, score_2 \in Scores(S)$ строит Q-множество $score \in Scores(S)$: $\forall e \in S\ score(e) = \big(score_1(e) + score_2(e)\big) * R$, где R – количество Q-множеств, для которых объект e имеет ненулевую оценку.

Утверждение.

Операция $CombMNZ$ коммутативна, сохраняет упорядоченность, удовлетворяет «Эффекту хора» и «Эффекту снятия сливок».

Важно отметить, что значения синтезированных оценок не принадлежат диапазону [0,1], что делает результирующие оценки не сопоставимыми с первоначальными. Для построения Q-множества требуется последующая нормализация.

Операция $CombMNZ$ достаточно хорошо учитывает «Эффект хора», особенно в случае обобщения операции на синтез нескольких функций оценки. Тем не менее представленная операция проработана в первую очередь для синтеза ответов на запросы, в которые не вкдлючены объекты с низкими оценками, и на ее поведение сильно влияет порог дискретизации. Важно отметить, что в случае решения задачи синтеза, когда множества объектов, по которым построены Q-множества совпадают, операция $CombMNZ$ превращается в операцию CombSUM, рассмотренную в работе [8].

В отличие от операции супер-объединения при таком методе синтеза «Эффект хора» влияет на результирующее Q-множество сильнее, чем «Эффект снятия сливок».

Операция $CombMNZ$ может быть более естественно выражена, через комбинацию операций супер-объединения и супер-пересечения. Супер-объединение Q-множеств, построенных как супер-объединение и супер-пересечение первоначальных функций оценки, будет синтезировать оценки согласованно с операцией $CombMNZ$. Преимуществом наших операций является то, что в результате их применения строится новая функция оценки с множеством значений на отрезке [0,1]. Тем не менее в последующих экспериментах мы будем использовать первоначальное определение операции $CombMNZ$ в качестве базы для сравнения.

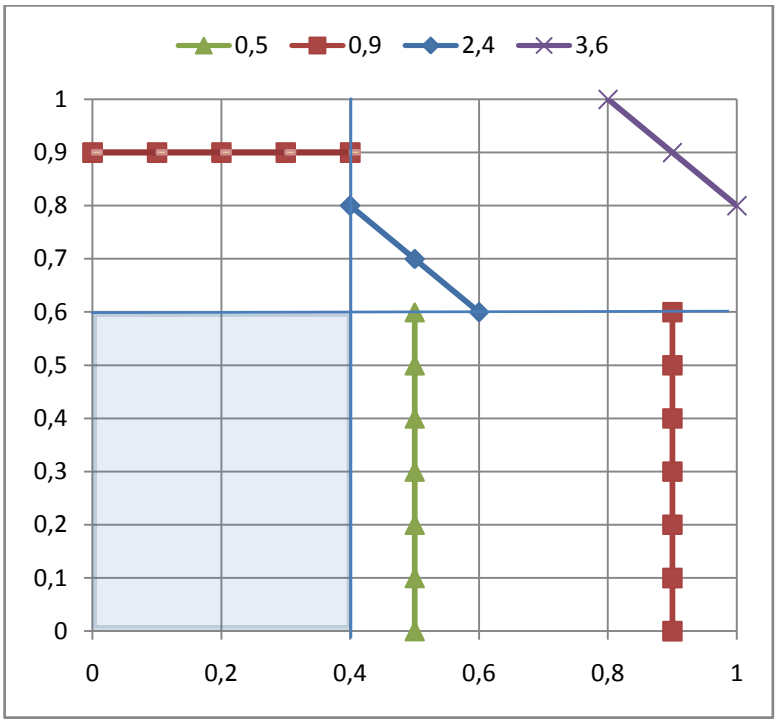

Рисунок 5 Изолинии, иллюстрирующие поведение операции CombMNZ.

Чтобы продемонстрировать природу операции $CombMNZ$, на рисунке 5 проиллюстрирован процесс синтеза двух Q-множеств после дискретизации по порогам 0.4 и 0.6.

После того как мы определили несколько реализаций операции синтеза двух Q-множеств важно показать на модельном примере, как конкретный метод синтеза влияет на результирующие оценки. На следующей диаграмме (рисунок 6) показано как изменяется значение синтезированной оценки, если одна оценка зафиксирована и равна 0.7, а вторая изменяется равномерно от 0 до 1 с шагом 0.1.

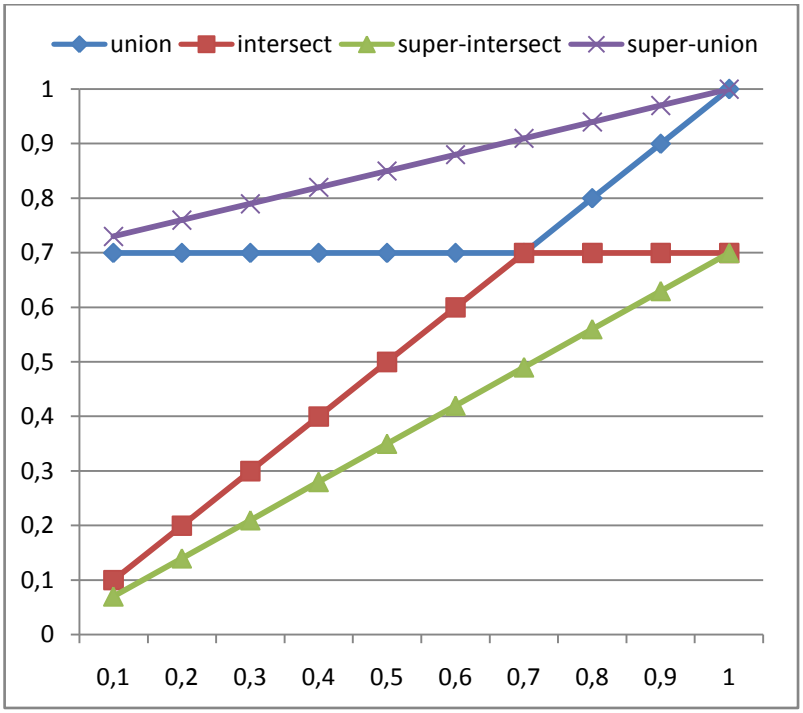

Рисунок 6 Сопоставление различных реализаций операции $fusion$

4 Эксперименты

В этом разделе представленная модель будет применена к задаче поиска изображений по образцу. В ходе экспериментов на тестовой коллекции изображений будет проанализирована адекватность описанной модели.

4.1 Экспериментальное окружение

4.1.1 Набор данных

Все эксперименты проводятся коллекции, которая содержит 1087 изображений ([1]). Коллекция строилась следующим образом. Сначала из коллекции Corel Photo Set было выбрано 101 изображение-запрос. Изображения-запросы были распределены двумя экспертами в 16 тематических групп (Clouds, Sunrise, Skyscraper, Field, Sunflowers, People, Forest, Trees, City, Cats, Winter wood, Coastal sunrise, Coastal, Bears, Lakes, Winter wood - 2). В таким образом построенной коллекции изображения находящиеся в одной тематической группе считаются релевантными друг другу. После этого коллекция была пополнена случайными изображениями, которые считаются нерелевантными к изображениям-запросам.

4.1.2 Используемые цветовые и текстурные признаки

В экспериментах рассматривалось несколько пространств поисковых образов. Каждое изображение в коллекции представлено дескриптором в трех пространствах признаков:

- Цветовые моменты (color moments) – цветовые метрики и признаки, основанные на моментах распределения цвета ([30]);
- Цветовые гистограммы (color histograms) – цветовые гистограммы с учетом информации о пространстве ([1]);
- Текстура (texture) – свертка изображения с помощью ICA фильтров использовалась в качестве текстурных признаков и дивергенция Kullback-Leibler – в качестве текстурной метрики ([6]).

Для каждой пары изображений были вычислены расстояния в каждом пространстве признаков, по которым были построены соответствующие Q-множества.

4.2 Постановка экспериментов

4.2.1 Измеряемые характеристики

Поскольку наши эксперименты будут направлены на проверку гипотезы о том, что предложенная модель не только упрощает задачу поиска изображений по образцу, но и повышает качество результата, мы будем анализировать метрики, используемые для оценки качества поиска.

Точность (precision) наиболее полно отражает изменения качества поиска. В нашей коллекции количество релевантных объектов для разных изображений-запросов разное, поэтому мы будем использовать R-precision для оценки наших результатов. Эта метрика показывает, сколько релевантных изображений оказывается среди R изображений с наибольшими оценками, где R – количество изображений релевантных заданному запросу.

4.2.2 Цели экспериментов

В ходе экспериментов мы проанализируем поведение разных модификаций нашей схемы поиска изображений по образцу на основе нескольких признаков.

Поведение предложенной реализаций операции $fusion$ мы будем сравнивать с алгоритмом синтеза CombMNZ (база для сравнения), который считается наиболее подходящим для комбинирования результатов поиска изображений по нескольким признакам и активно используется в исследованиях [2]. Сравнение результатов поиска на основе наших методов синтеза и CombMNZ покажет, насколько точно представленный набор операций описывает процесс слияния нескольких результатов поиска.

4.2.3 Программная реализация

Экспериментальное окружение было реализовано на основе СУБД Oracle версии 10.2. Такой выбор программной реализации был обусловлен тем, что в данном сервере управления базами данных реализованы высокопроизводительные и эффективные методы соединения и агрегирования данных.

Исходные данные коллекции изображений были представлены в формате CSV, что позволило их автоматически загружать в хранилище.

На рисунке 7 представлена диаграмма, описывающая основные таблицы схемы базы данных, использованной в экспериментах.

Специфика задачи требовала использования операций агрегирования во всех проводимых экспериментах, что требовало полного просмотра большинства таблиц.

В реализованной схеме были использованы индексы только по ключам таблиц, поскольку извлечение данных вне соединения не требовалось. Использование вспомогательных представлений и таблиц позволило ускорить проведение экспериментов и упростить анализ результатов. Например, информация о нормализованных значениях расстояний сохранялась в отдельной вспомогательной таблице перед проведением экспериментов с различными реализациями операции fusion.

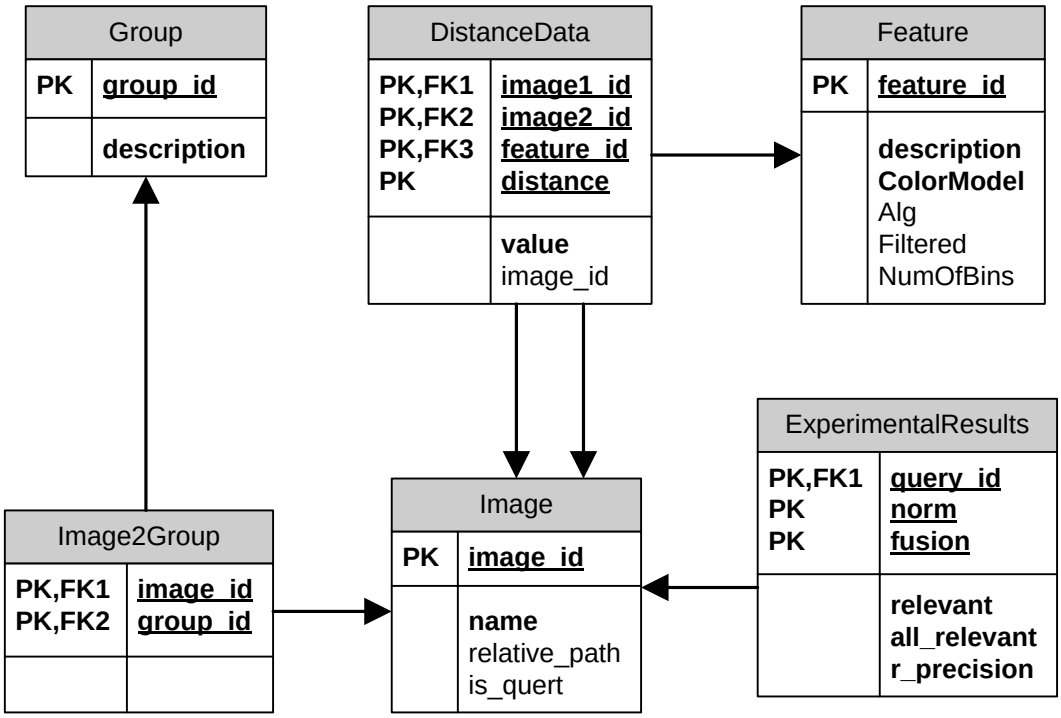

Рисунок 7 Основные таблицы базы данных, использованной в экспериментах

Программная часть работы включала в себя подготовку и проведение около 400 измерений (то есть построения одной гистограммы или вычисления отдельной метрики), каждое из которых занимало от 15 секунд до 5 минут. В конечный результат вошли эксперименты построения 90 квадратных гистограмм, 30 плоских гистограмм распределения оценок в Q-множествах, измерения 90 значений R-precision результирующих Q-множеств после применения различных операций калибровки и синтеза. Также было проведено множество экспериментов, которое помогло нам в построении модели и определении важных свойств операций.

В рамках работы проводились эксперименты (около 15 измерений), в которых метрикой качества поиска являлись точность, полнота и сбалансированная F-мера. Эксперименты показали, что действительное изменение качества поиска при использовании различных методов синтеза отражает именно R-precision. Именно эта метрика и была использована в окончательных экспериментах.

Большая серия экспериментов (около 18 измерений) была проведена с использованием метода нормировки $norm - maxmin$ и различных реализаций операции синтеза. Сначала предполагалось использование различных порогов дискретизации с целью увеличения точности синтезированного результата.

Анализ гистограмм распределений оценок в Q-множествах по диапазону значений позволил сделать выводы о недостаточности калибровки оценок с помощью операции $norm-maxmin$ и подтолкнул нас к обсуждению основных свойств операций нормализации и усиления функций оценки. Было построено около 15 гистограмм распределения оценок, как в отдельных Q-множествах, так и агрегированных, то есть для всех Q-множеств, построенных по одному признаку.

Нами была проведена серия экспериментов, проверяющая идею о том, что можно подобрать значения параметров операции усиления так чтобы минимизировать норму разности гистограмм распределений оценок внутри Q-множеств. Эксперименты (около 30 измерений) показали, что такой метод приведения распределений оценок к сопоставимым не дает требуемых результатов.

Анализ квадратных гистограмм, демонстрирующих распределение оценок одновременно внутри двух синтезируемых Q-множеств, позволил нам объяснить различное влияние реализаций операции синтеза на точность результирующего множества.

4.3 Анализ экспериментов

4.3.1 Анализ характера функций расстояния

Как обсуждалось ранее, оценки изображений, построенные на основе разных признаков или разными методами, могут очень сильно отличаться по диапазону значений и характеру распределения. Поскольку в нашем распоряжении имелась коллекция изображений с расстояниями между ними, посчитанными для разных функций расстояния, мы начали с анализа диапазона значений функций расстояния. Это позволило нам убедиться в том, что используемые нами функции расстояния действительно по-разному описывают всю коллекцию изображений. Таблица 1 показывает разброс значений функций расстояния.

Все описанные нами алгоритмы, в том числе и база для сравнения, CombMNZ, работают не со значениями функции расстояния, а построенными по ним значениями функции подобия и с оценками. Следующая гистограмма (рисунок 8) показывает распределение значений функций подобия.

Таблица 1 Максимальное и минимальное значение различных функций расстояния, вычисленные на основе тестовой коллекции изображений

Distance function	min distance	max distance
colour moments	0,017905196	6,7895699
colour histograms	0,21907941	199,49219
texture	0,002919512	13,466594

Гистограмма на рисунке 8 показывает, что хотя формально значения функции подобия лежат в отрезке от 0 до1, распределения этих значений сильно отличаются. Именно для решения задачи сопоставимости таких и других оценок разной природы в нашей модели и алгоритме CombMNZ предлагаются различные методы калибровки функций оценки.

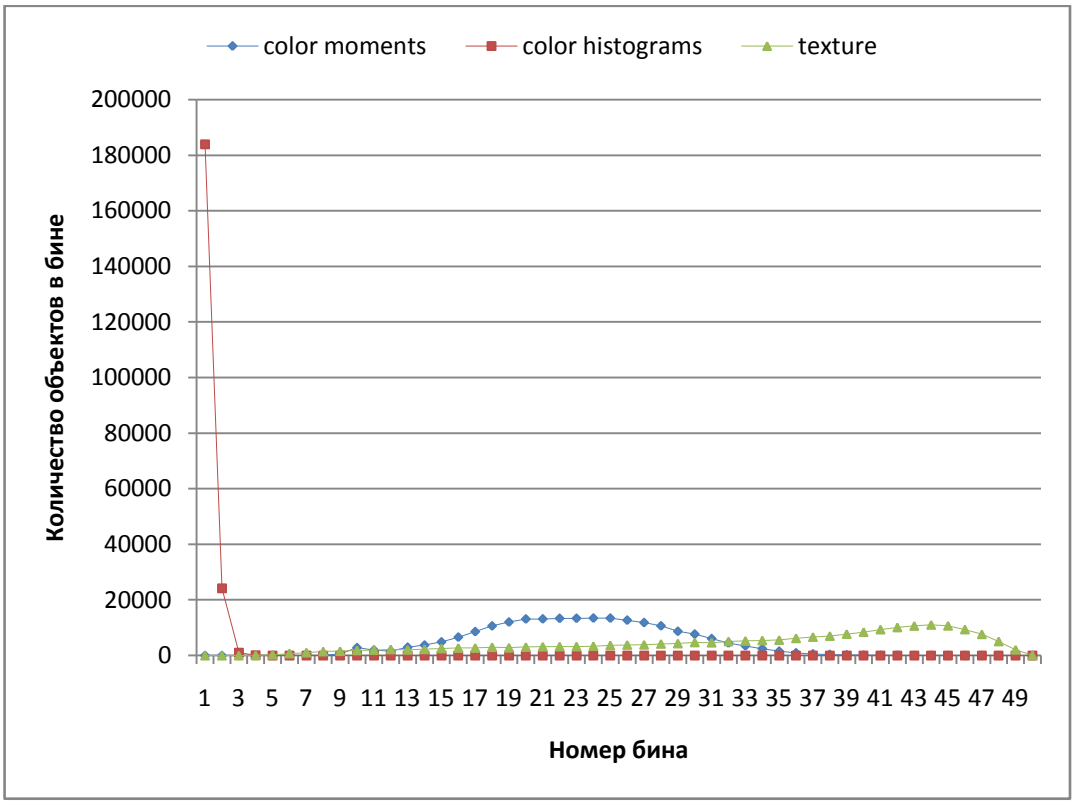

Рисунок 8 Распределение значений функций оценки

4.3.2 Методы приведения распределений оценок к сопоставимым

Напомним, что операция $norm - maxmin$ направлена лишь на приведение диапазонов значений оценок к одному, а не их распределений. Гистограмма на рисунке 9 показывает распределение оценок после нормализации с помощью операции $norm - maxmin$. Важно отметить, что распределение оценок в Q-множествах до нормализации и после применения операции $norm - maxmin$ отличаются незначительно.

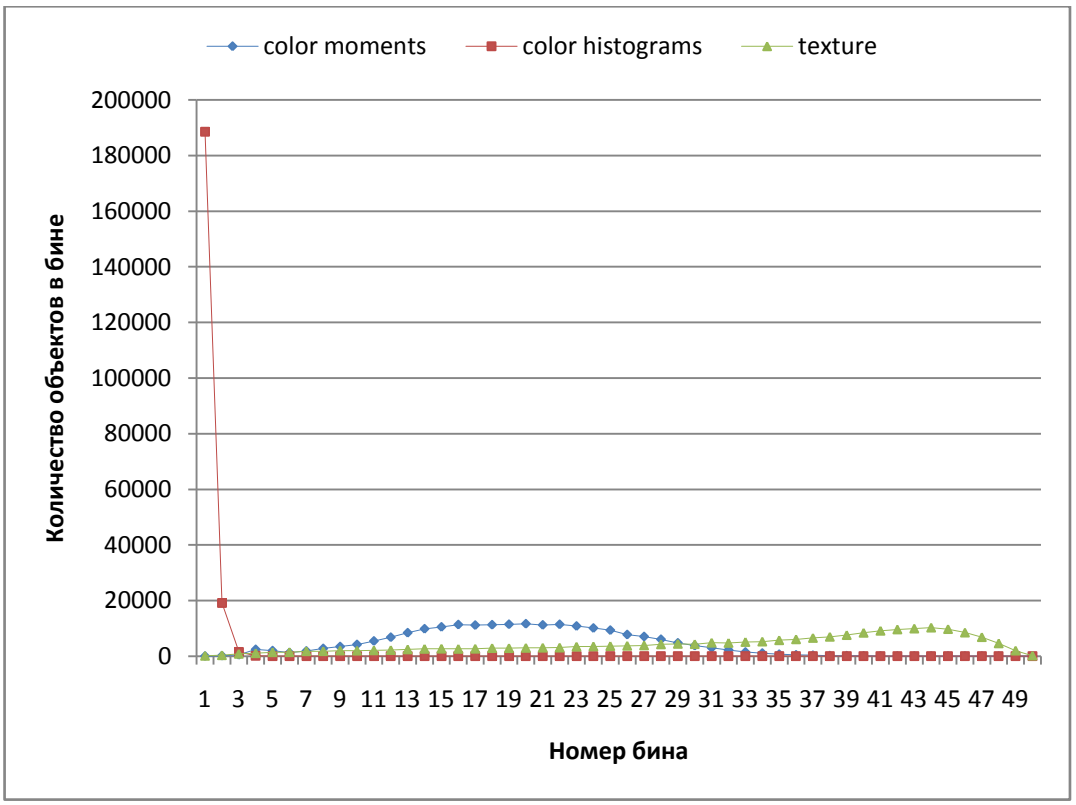

Рисунок 9 Распределение значений функций оценки после применения операции $norm - maxmin$

На рисунке 10 представлена гистограмма, демонстрирующая распределение оценок в Q-множествах после нормализации с помощью операции $norm - avg$.

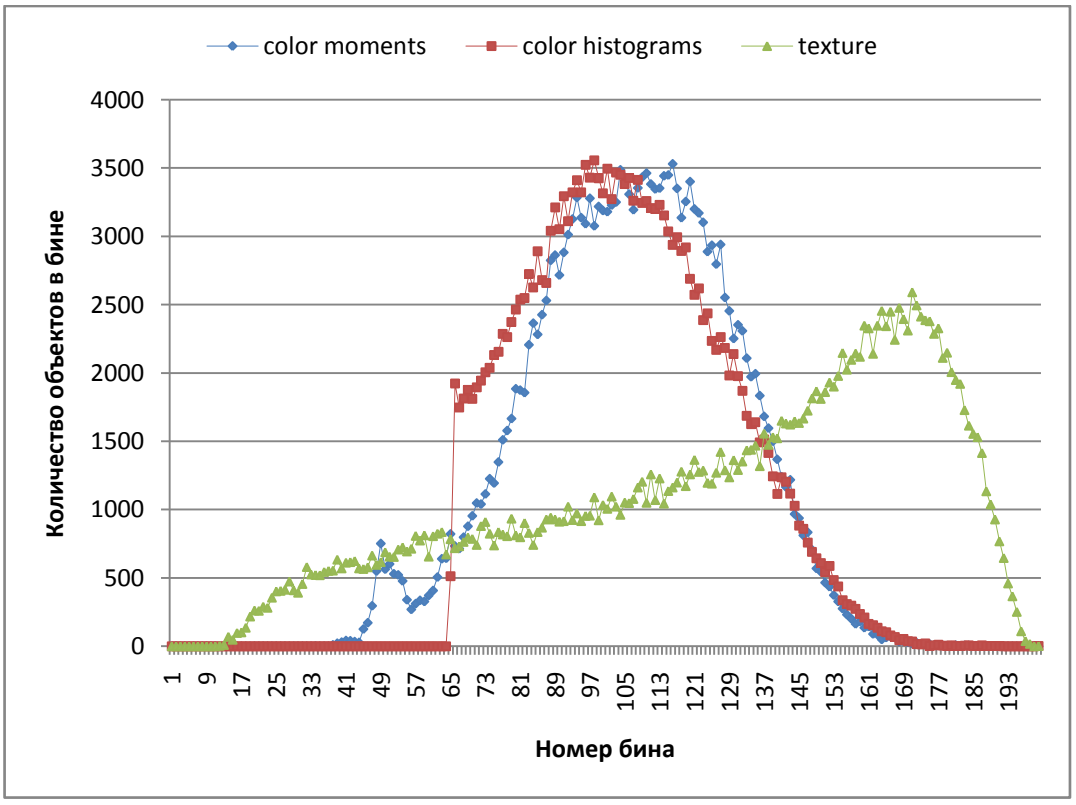

Рисунок 10 Распределение значений функций оценки после применения операции $norm - avg$

Предложенный в нашей модели метод калибровки с помощью операции $normalize - dist_p$ нацелен не только на приведение значений оценок к общему диапазону, но и на некоторое изменение распределений оценок. При калибровке оценок в Q-множествах с помощью операции $normalize - dist_p$ значение параметра p было выбрано равным 0.1. Мы полагаем, что в нашей коллекции изображений 10% составляют ту долю коллекции, которая в первую очередь влияет на результат поиска. Распределение оценок по диапазону значений представлено на рисунке 11.

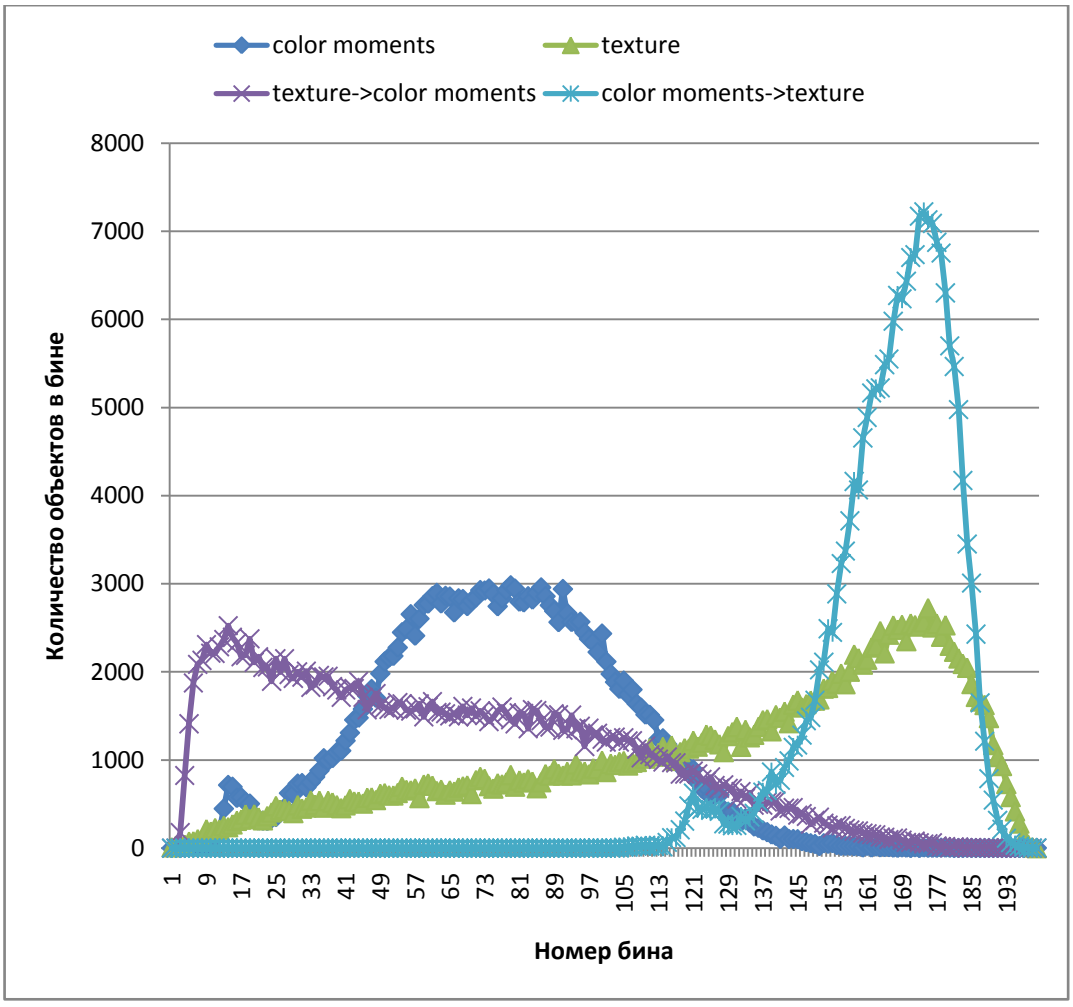

Рисунок 11 Распределение значений функций оценки после применения операции $normalize - dist_{0.1}$

Мы также проанализировали влияние операций усиления и ослабления. В своих экспериментах мы нормализовали Q-множества с помощью операции $norm - avg$, после чего применили операцию усиления. Значение параметра n в операции $strengthen(n)$ было выбрано таким образом, чтобы значения оценок в рассматриваемых Q-множествах на уровне 10% совпадали. В дальнейшем мы будм ссылаться на эту процедуру калибровки как на $norm\&strengthen$. Результат применения такой процедуры калибровки продемонстрирован на гистограмме на рисунке 12.

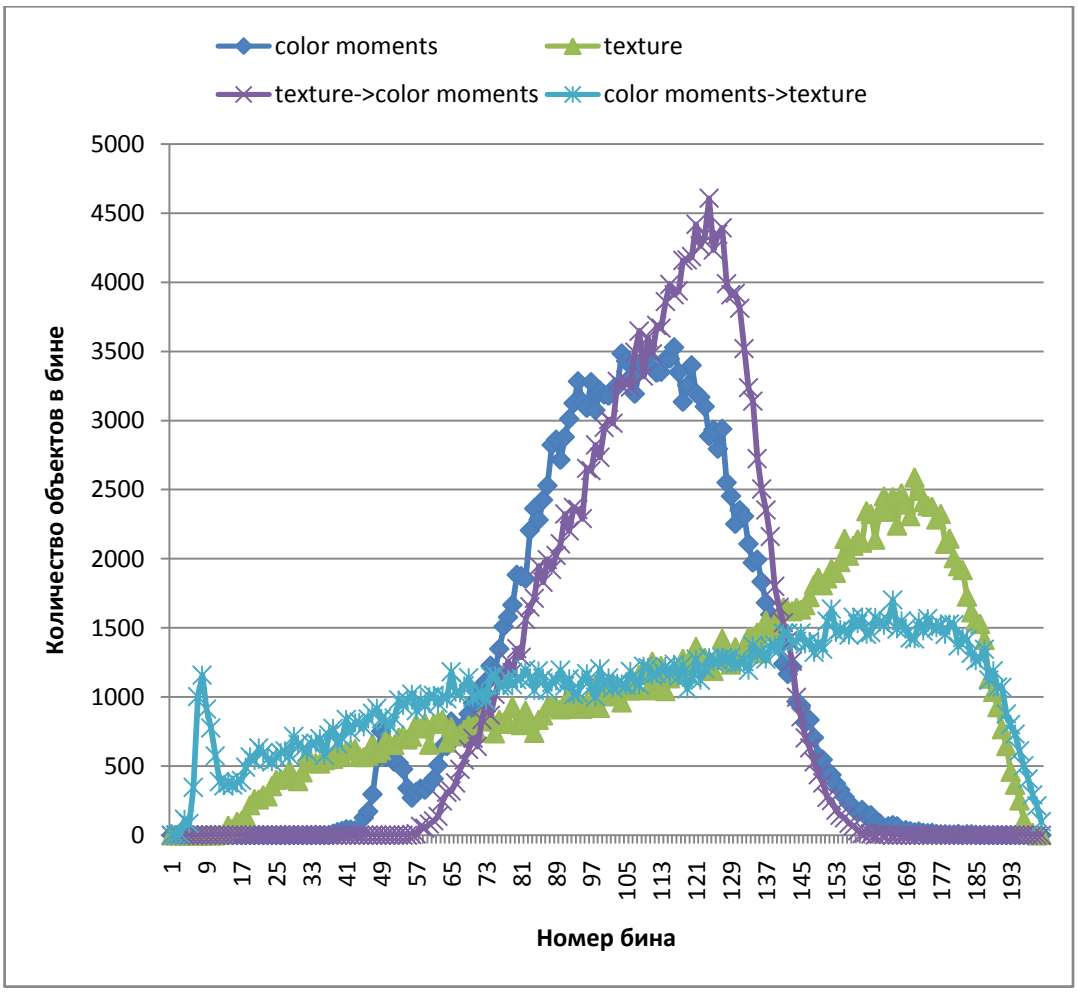

Рисунок 12 Распределение значений функций оценки после применения процедуры $norm\&strengthen$

4.3.3 Анализ характера методов синтеза

Этот раздел будет посвящен анализу представленных методов синтеза на реальных данных. Ранее мы моделировали синтез двух Q-множеств с помощью разных реализаций операции $fusion$. Следующий эксперимент покажет, как разные методы синтеза соотносятся между собой.

Ко всей коллекции изображений последовательно применяли все операции синтеза. После чего выбирались 10% изображений, получивших наибольшие оценки, и среди них выбиралось наименьшее значение оценки, с которым объект попадал в 10% лучших. Для каждого метода синтеза такая наименьшая оценка получается разной, но при использовании этого значения оценки в качестве пороговой можно выбрать 10% «лучших» изображений.

На рисунке 13 представлены изолинии, построенные для всех видов синтеза по определенному пороговому значению синтезированной оценки. Представленные линии «отрезают» правый верхний угол, содержащий изображения с наибольшими синтезированными

оценками. На диаграмме видно засчет каких объектов изменяется ответ на запрос пользователя при использовании разных функций синтеза.

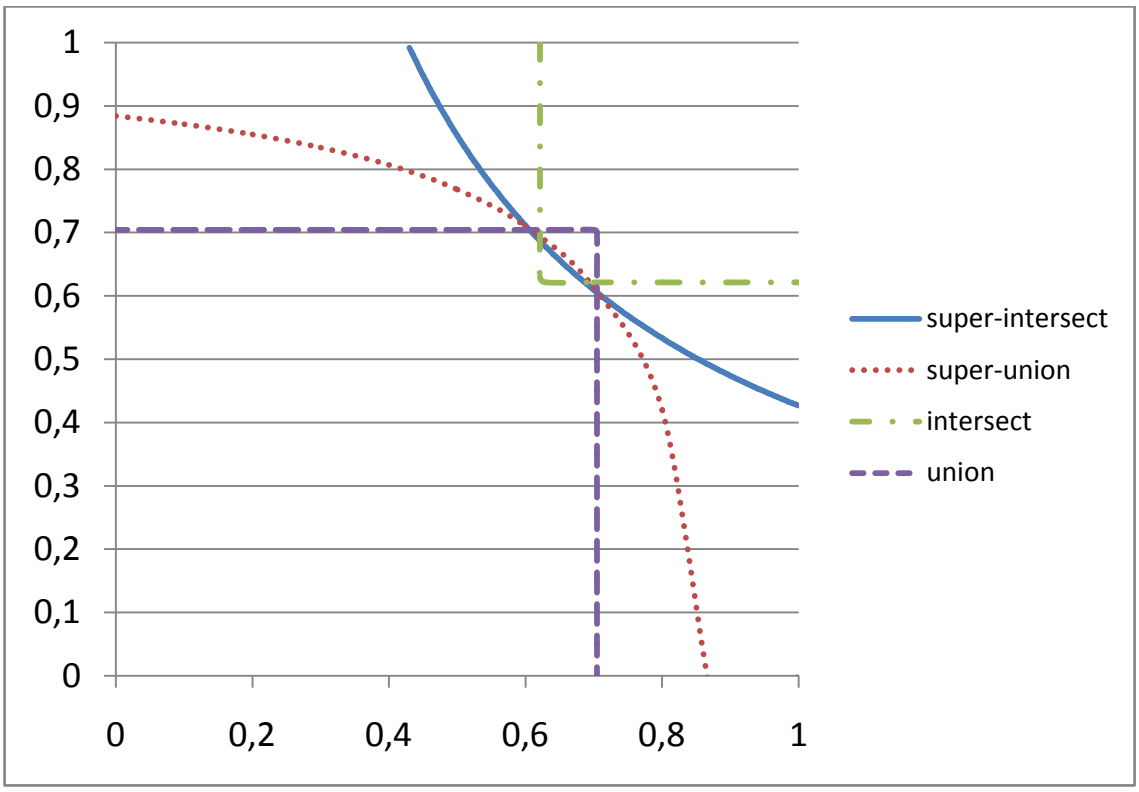

Рисунок 13 Изолинии, иллюстрирующие, поведение различных реализаций операции $fusion$

4.3.4 Анализ качества поиска

Анализ качества поиска изображений по разным признакам следует начать с обсуждения результатов, которые можно получить с использованием всех метрик по отдельности, без синтеза соответствующих Q-множеств. Таблица 2 представляет значение R-precision, полученное при поиске изображений на основе цветовых моментов, цветовых гистограмм и текстуры. Из представленных результатов видно, что текстурные признаки плохо описывают коллекцию изображений с точки зрения поиска. Таким образом, в рамках нашей модели ослабление Q-множества на основе текстуры, скорее всего, приведет к улучшению качества синтезированного результата.

Таблица 2 R-precision Q-множеств на основе цветовых моментов, цветовых гистограмм и текстуры

colour moments	colour histograms	texture
0,48	0,40	0,15

Таблицы 3-5 показывают R-precision при поиске изображений на основе комбинации двух различных признаков.

Результаты, представленные в таблице 3, показывают, что, несмотря на использование процедур калибровки, операция $super - union$ дает неудовлетворительную точность ре-

зультирующего Q-множества при синтезе Q-множеств на основе текстуры и цветовых моментов. Такое поведение операции $super - union$ объясняется тем, что одно из синтезируемых Q-множеств описывается функцией оценки, которая плохо отражает релевантность объектов запросу. На рассмотренных ранее гистограммах видно, что даже после калибровки оценки Q-множества на основе текстуры доминируют по сравнению с оценками на основе цветовых моментов. Только в случае калибровки с помощью процедуры $norm\&strengthen$ операция $super - union$ синтезирует Q-множества согласно представлениям пользователя. Этот факт согласуется с наблюдением, что на рисунке 12 высокие оценки на основе цветовых моментов доминируют. Аналогичные наблюдения объясняют низкую точность синтезированного Q-множества после применения операции $norm - dist(0.1)$ для всех реализаций операции синтеза.

Таблица 3 R-precision результата синтеза Q-множеств на основе текстуры и цветовых моментов

texture and color moments	without norm	norm-avg	norm-maxmin	norm-dist(0.1)	norm&strengthen
CombMNZ	0,45	0,44	0,46	0,38	0,45
super-intersect	0,46	0,45	0,47	0,38	0,45
super-union	0,32	0,33	0,31	0,37	0,46

Таблица 4 демонстрирует R-precision синтезированного Q-множества, полученного из Q-множеств на основе текстуры и цветовых гистограмм. Результаты, представленные в таблицы 4, показывают чувствительность операций $CombMNZ$ и $super - union$ к процедуре предварительной калибровки Q-множеств.

Таблица 4 R-precision результата синтеза Q-множеств на основе текстуры и цветовых гистограмм

texture and color histograms	without norm	norm-avg	norm-maxmin	norm-dist(0.1)	norm&strengthen
CombMNZ	0,20	0,41	0,20	0,37	0,43
super-intersect	0,45	0,41	0,45	0,37	0,43
super-union	0,16	0,35	0,16	0,37	0,44

Результаты синтеза Q-множеств на основе цветовых признаков представлены в таблице 5. Важно отметить, что правильная процедура калибровки исходных Q-множеств позволяет

значительно увеличить точность синтезированного результата. В случае синтеза Q-множеств на основе цветовых моментов и цветовых гистограмм различные реализации операции $fusion$ показывает сопоставимые результаты.

Таблица 5 R-precision результата синтеза Q-множеств на основе цветовых моментов и гистограмм

color moments and histograms	without norm	norm-avg	norm-maxmin	norm-dist(0.1)	norm&strengthen
CombMNZ	0,49	0,53	0,49	0,53	0,53
super-intersect	0,50	0,53	0,50	0,53	0,53
super-union	0,49	0,52	0,50	0,53	0,52

Результаты, представленные в таблицах 3-5, также показали, что качество синтезированного Q-множества в значительной степени зависит от точности первоначальных множеств.

4.3.5 Анализ влияния методов синтеза на качество поиска

В этом разделе мы попробуем показать поведение различных операций синтеза на реальных примерах. На рисунке 14 для Q-множеств на основе текстуры и цветовых моментов до нормализации представлены квадратные гистограммы, показывающие какое количество объектов попадает в бин, соответствующий паре оценок. На рисунках отражено распределение всех объектов, распределение релевантных объектов и отношение количества релевантных объектов к количеству всех объектов в каждом бине соответственно. Гистограммы хорошо показывают неравномерность распределения оценок на основе текстуры. Гистограмма распределения оценок релевантных объектов показывает область, в которой расположены релевантные объекты. Методы синтеза должны присваивать наибольшие оценки объектам из этой области, чтобы результат синтеза содержал наибольшее количество релевантных объектов. Точность результата синтеза зависит от того, какие оценки будут присвоены объектам в зоне, выделенной на третьей диаграмме.

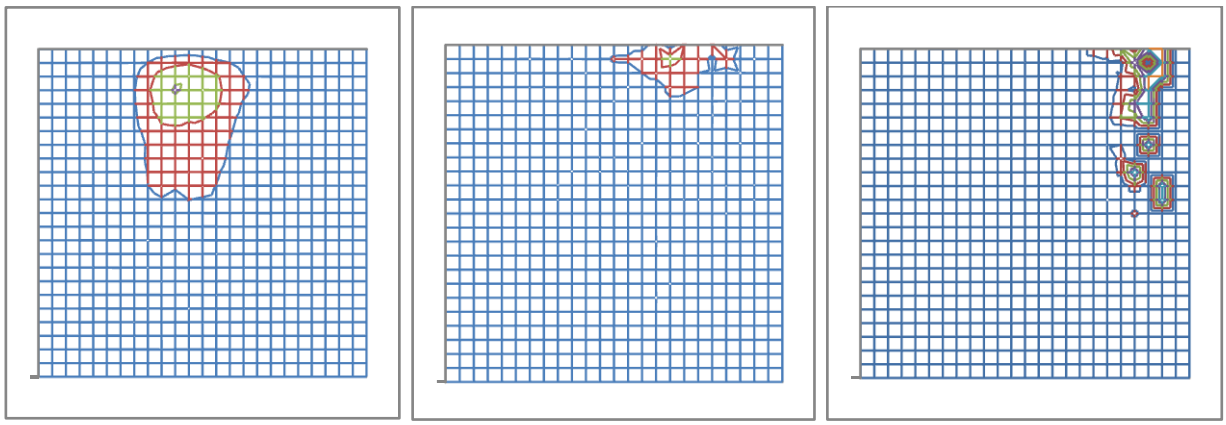

Рисунок 14 Квадратные гистограммы распределения оценок в Q-множествах на основе текстуры и цветовых моментов

На рисунке 15 изображены аналогичные диаграммы для Q-множеств на основе текстуры и цветовых моментов после применения процедуры $norm\&strengthen$. Гистограммы показывают, насколько сопоставимыми становятся распределения значений оценок двух Q-множеств поле калибровки.

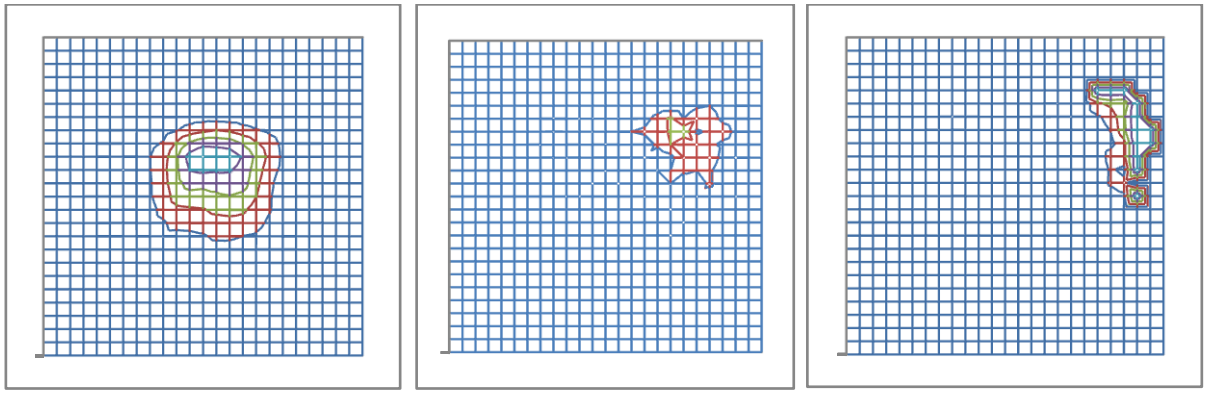

Рисунок 15 Квадратные гистограммы распределения оценок в Q-множествах на основе текстуры и цветовых моментов после применения процедуры $norm\&strengthen$

Заключение

В рамках работы построен инструмент описания сложных видов поиска на основе подобия. Разработанное средство представляет собой однородную систему способов и операций, предназначенную для формирования поисковых запросов, не зависящую от конкретной задачи поиска.

Введено понятие Q-множества, которое является абстракцией над языком запросов, природой объектов в пространстве поиска, методом выполнения запросов, и включает в себя запрос и результат его выполнения. Предложен набор операций над классом Q-множеств, построенных на основе подобия, который позволяет выражать высокоуровневые операции необходимые при обработке сложных объектов и сложных видов поиска. Набор операций представляет собой формальную систему методов калибровки и синтеза Q-множеств и позволяет описывать сложные поисковые запросы и методы их построения.

В рамках работы выявлены и проанализированы общие принципы и механизмы работы с Q-множествами, а также определены свойства предложенных операций и их взаимосвязь с потребностями пользователя и свойствами конкретных задач построения сложных запросов.

В качестве примера работы со сложными объектами рассмотрена задача поиска изображений по образцу на основе нескольких признаков. Для улучшения качества поиска изображений за счет более полного анализа свойств объектов использовались подходы, основанные на комбинировании разных признаков.

На базе предложенного набора операций сформирована схема синтеза результатов поиска изображений по нескольким разнородным признакам. Результаты были проанализированы и сопоставлены с результатами, полученными на основе известных подходов к комбинированию нескольких признаков при поиске изображений по содержанию.

Эксперименты показали сопоставимость результатов. Тем не менее, детальный анализ свойств представленной модели позволяет предсказывать поведение операций, в зависимости от конкретных требований задачи поиска. Таким образом, в рамках работы был предложен подход к единообразному описанию сложных видов поиска на основе подобия.

Список литературы

1. *Н. Васильева.* Построение и комбинирование признаков в задаче поиска изображений по содержанию. Дис. … канд. физ.-мат. наук. / Санкт-Петербургский Государственный Университет. –Санкт-Петербург, 2010.
2. *Н. Васильева, А. Дольник, И. Марков.* Поиск изображений. Синтез различных методов поиска при формировании результатов. // Интернет-Математика 2007: Сборник работ участников конкурса. –Екатеринбург: изд-во Урал. ун-та, 2007. –С. 46-55.
3. *K. Aberer and J. Wu.* A framework for decentralized ranking in web information retrieval. // 5th Asia-Pacific web conference on Web technologies and applications, APWeb'03, 2003, Proceedings. -Berlin, Heidelberg: Springer-Verlag, 2003. –P. 213-226.
4. *J. A. Aslam and M. Montague.* Models for metasearch. // 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, 2001, Proceedings. -New York, NY, USA: ACM, 2001. –P. 276-284.
5. *M. Beigbeder.* Integrating boolean and vector models of information retrieval with passage retrieval. // 4th international symposium on Information and communication technologies, WISICT '05, 2005, Proceedings. –Dublin: Trinity College, 2005. –P. 123-128.
6. *H. Borgne, A. Guerin-Dugue, and A. Antoniadis.* Representation of images for classification with independent features. // Pattern Recognition Letters. –2004. –Vol. 25. –P. 141-154.
7. *Chamberlin, Donald D. and Boyce, Raymond F.* SEQUEL: A structured English query language. // ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control, SIGFIDET '74, Ann Arbor, Michigan, 1974, Proceedings. –New York, NY, USA: ACM, 1974. –P. 249-264.
8. *Chowdhury, O. Frieder, D. A. Grossman, and M. C. McCabe.* Analyses of multiple-evidence combinations for retrieval strategies. // W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, SIGIR, 2001. –ACM, 2001. –P. 394-395.
9. *Pablo Ciaccia, Danilo Montesi, Wilma Penzo, and Alberto Trombetta.* Imprecision and User Preferences in Multimedia Queries: A Generic Algebraic Approach. // K.-D. Schewe, B. Thalheim, editors, FoIKS 2000, Proceedings, volume 1762 of Lecture Notes in Computer Science.–Berlin / Heidelberg: Springer, 2000. –P. 50-71.
10. *F. Crestani, M. Lalmas, C. J. Van Rijsbergen, and I. Campbell.* "Is this document relevant?. . . probably": a survey of probabilistic models in information retrieval. // ACM Comput. Surv. –December 1998. –Vol. 30. –P. 528-552.
11. *F. Crestani and C. J. van Rijsbergen.* A study of probability kinematics in information retrieval. // ACM Trans. Inf. Syst. –July 1998. –Vol. 16. –P. 225-255.

12. *K. Donald and A. Smeaton.* A comparison of score, rank and probability-based fusion methods for video shot retrieval. // W.-K. Leow, M. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. Bakker, editors, Image and Video Retrieval, volume 3568 of Lecture Notes in Computer Science. –Berlin / Heidelberg: Springer, 2005. –P. 61-70.
13. *Hai Dong, F.K. Hussain, Chang.* A survey in traditional information retrieval models. // 2nd IEEE International Conference on Digital Ecosystems and Technologies, 2008. – DEST, Feb. 2008. –P. 397-402.
14. *Peter G. B. Enser and Christine J. Sandom.* Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval. // Erwin M. Bakker and Thomas S. Huang and Michael S. Lew and Nicu Sebe and Xiang Sean Zhou, editors, Image and Video Retrieval, Second International Conference, CIVR 2003, Urbana-Champaign, IL, USA, July 24-25, 2003, Proceedings, volume 2728 of Lecture Notes in Computer Science.–Berlin / Heidelberg: Springer, 2003. –P. 291-299.
15. *Ronald Fagin.* Fuzzy Queries in Multimedia Database Systems. // 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, PODS '98, Seattle, WA, USA, 1998, Proceedings. –ACM, 1998. –P. 1-10.
16. *N. Fuhr.* A probability ranking principle for interactive information retrieval. // Inf. Retr. –2008. –Vol. 11, No.3. –P. 251-265.
17. *Shan GAO and Bo XU and Taiyi HUANG and Chengqing ZHONG.* An Adaptive Information Retrieval Approach Based on Fuzzy Set. // ISCSLP 2000. :[http://www.isca-speech.org/archive_open/archive_papers/iscslp2000/PSA2/076.pdf]
18. *Jana Kludas, Eric Bruno, and Stephane Marchand-Maillet.* Adaptive multimedial retrieval: Retrieval, user, and semantics. // Information Fusion in Multimedia Information Retrieval. –Berlin, Heidelberg: Springer-Verlag, 2008. –P. 147-159.
19. *Mieczyslaw M. Kokar, Jerzy A. Tomasik, and Jerzy Weyman.* Formalizing classes of information fusion systems. // Information Fusion. –2004. –Vol. 5, No.3. –P. 189-202.
20. *D. Lillis, F. Toolan, R. W. Collier, and J. Dunnion.* Probfuse: a probabilistic approach to data fusion. // E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Jarvelin, editors, SIGIR, 2006. –ACM, 2006. –P. 139-146.
21. *D. Lillis, F. Toolan, A. Mur, L. Peng, R. Collier, and J. Dunnion.* Probability based fusion of information retrieval result sets. // Artificial Intelligence Review. –2006. –Vol. 25. –P. 179-191.
22. *J. W. S. Liu and J. M. Milner.* Probabilistic models of inverted file information retrieval systems. // 1976 ACM SIGMETRICS conference on Computer performance modeling

measurement and evaluation, SIGMETRICS '76, New York, NY, USA, 1976, Proceedings. –ACM, 1976. –P. 25-37.

23. *M. Melucci.* A basis for information retrieval in context. // ACM Trans. Inf. Syst. –June 2008. –Vol. 26, No.14. –P. 1-41.
24. *Olivas, Jose.* Fuzzy Sets and Web Meta-search Engines. // Bustince, Humberto and Herrera, Francisco and Montero, Javier, editors, Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, volume 220 of Studies in Fuzziness and Soft Computing. –Berlin/Heidelberg: Springer, 2008. –P. 537-552.
25. *J. Pound, I. F. Ilyas, and G. Weddell.* Expressive and flexible access to web extracted data: a keyword-based structured query language. // 2010 international conference on Management of data, SIGMOD '10, 2010, Proceedings. –New York, NY, USA: ACM, 2010. –P. 423-434.
26. *N. Rubens.* The application of fuzzy logic to the construction of the ranking function of information retrieval systems. // CoRR. –2006. –Vol. abs/cs/0610039.
27. *X. Shen and C. X. Zhai.* Exploiting query history for document ranking in interactive information retrieval. // 26th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '03, 2003, Proceedings. –New York, NY, USA: ACM, 2003. –P. 377-378.
28. *S. Shi, B. Lu, Y. Ma, and J.-R. Wen.* Nonlinear static-rank computation. // D. W.- L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, CIKM. –ACM, 2009. –P. 807-816.
29. *K. H. Stirling.* On the limitations of document ranking algorithms in information retrieval. // SIGIR Forum. –May 1981. –Vol. 16. –P. 63-65.
30. *M. Stricker and A. Dimai.* Spectral covariance and fuzzy regions for image indexing. // Mach. Vision Appl. –1997. –Vol. 10, No. 2. –P. 66-73.
31. *L. Valet, G. Mauris, and Ph. Bolon.* A statistical overview of recent literature in information fusion. // 3rd International Conference on Information Fusion, MOC3, 2000, Proceedings. –P. 22-29.
32. *J.-N. Vittaut and P. Gallinari.* Machine learning ranking for structured information retrieval. // M. Lalmas, A. MacFarlane, S. M. Ruger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, ECIR, volume 3936 of Lecture Notes in Computer Science. –Springer, 2006. –P. 338-349.
33. *Voorhees, E. M., N. K. Gupta, and B. Johnson-Laird.* Learning collection fusion strategies. // 18th annual international ACM SIGIR conference on Research and development

in information retrieval, SIGIR ’95, 1995, Proceedings. –New York, NY, USA: ACM Press, 1995. –P. 172–179.

34. *Wei Wang, Amelie Marian, and Thu D. Nguyen.* Unified structure and content search for personal information management systems. // 14th International Conference on Extending Database Technology (EDBT/ICDT '11), Anastasia Ailamaki, Sihem Amer-Yahia, Jignesh Pate, Tore Risch, Pierre Senellart, and Julia Stoyanovich (Eds.), 2011, Proceedings. –New York, NY, USA: ACM, 2011.
35. *M. Wechsler and P. Schauble.* A new ranking principle for multimedia information retrieval. // 4th ACM conference on Digital libraries, DL '99, 1999, Proceedings. –New York, NY, USA: ACM, 1999. –P. 146-151.
36. *Wiguna, Wiratna and Fernndez-Ibar, Juan and Garcна-Serrano, Ana.* Using a Fuzzy Model for Combining Search Results from Different Information Sources to Build a Metasearch Engine. // Reusch, Bernd, editor, Computational Intelligence, Theory and Applications.–Berlin Heidelberg: Springer, 2006. –P. 325-334.
37. *Wu, S. and F. Crestani:* Shadow document methods of results merging // 2004 ACM Symposium on Applied Computing, SAC ’04, 2004, Proceedings. –New York, NY, USA: ACM Press. –P. 1067-1072.
38. *Shengli Wu and Sally McClean.* Performance prediction of data fusion for information retrieval. // Inf. Process. Manage. –July 2006. –Vol. 42. –P. 899-915.
39. *R.-J. Yamashita, T. Ito, and H.-H. Yao.* Essql: an enhanced semi-structured query language for composite document retrievals. // 16th annual international conference on Computer documentation, SIGDOC '98, 1998, Proceedings. –New York, NY, USA: ACM, 1998. –P. 120-126.
40. *Yeh, Lin-Ju and Yao, Hsiu-Hsen and Chen, Yuan-Kuo*. SSQL: a semi-structured query language for SGML document retrievals. // 14th annual international conference on Systems documentation: Marshaling new technological forces: building a corporate, academic, and user-oriented triangle, SIGDOC '96, Research Triangle Park, North Carolina, United States, 1996, Proceedings. –New York, NY, USA: ACM, 1996. –P. 221-228.
41. *L. A. Zadeh.* Fuzzy sets. // Information and Control. –Vol. 8. –P. 338-353.
42. *Zadrozny, Slawomir and Nowacka, Katarzyna.* Fuzzy information retrieval model revisited. // Fuzzy Sets Syst. –Vol. 160. –Amsterdam, The Netherlands, The Netherlands, Elsevier North-Holland, Inc., 2009. –P. 2173-2191.