

На правах рукописи

Силич Василий Викторович

**АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ
РЕГИОНАЛЬНОЙ СИСТЕМЫ КОНТЕКСТНОЙ РЕКЛАМЫ
В СРЕДЕ ИНТЕРНЕТ**

Специальность 05.13.11 – Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Томск – 2008

Работа выполнена в Томском государственном университете систем управления и радиоэлектроники (ТУСУР)

Научный руководитель: доктор технических наук, профессор
Ехлаков Юрий Поликарпович

Официальные оппоненты: доктор технических наук, профессор
Цапко Геннадий Павлович

доктор технических наук, профессор
Тарасенко Владимир Феликсович

Ведущая организация: Иркутский государственный технический
университет, г. Иркутск.

Защита состоится «24» декабря 2008 г. в 14.30 на заседании совета по защите докторских и кандидатских диссертаций Д 212.269.06 при Томском политехническом университете по адресу: 634034, г. Томск, ул. Советская, 84.

С диссертацией можно ознакомиться в библиотеке Томского политехнического университета

Автореферат разослан «19» ноября 2008 г.

Ученый секретарь
совета по защите докторских
и кандидатских диссертаций
к.т.н., доцент

М.А. Сонькин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Системы контекстной рекламы (СКР) в настоящее время широко используются как в зарубежном, так и в российском сегментах Интернета. Свою популярность они заработали за счёт того, что позволяют эффективно продвигать рекламодателю сайты в Интернете даже при сравнительно небольшом бюджете. При этом рекламодатель получает гарантированное количество целевых посетителей на свой сайт, а владельцы сайтов и поисковых систем, на которых размещены рекламные объявления – стабильный доход. Для пользователя СКР является источником дополнительной информации, соответствующей контексту просматриваемой им страницы.

Системы контекстной рекламы, как правило, создаются при поисковых системах ввиду их широкой аудитории, наличие явно сформулированного интереса пользователя, выраженного в поисковом запросе, высокому уровню доверия к поисковым системам. При этом СКР интегрируются не только с крупными Интернет-порталами, но и с региональными поисковыми системами (РПС), причем рекламодатели в большинстве случаев заинтересованы именно в последних, т.к. их целевой аудиторией являются преимущественно региональные посетители. Пользователи, в свою очередь, также все больше предпочитают РПС по той причине, что они позволяют искать информацию, актуальную именно для данного региона. В пользу ограничения поиска только региональными ресурсами выступает и наличие в ряде регионов бесплатного доступа к ним. Таким образом, создание системы контекстной рекламы, интегрированной с региональной поисковой системой, является актуальной задачей.

При организации систем контекстной рекламы разработчики используют методы, относящиеся к таким областям, как информационный поиск (Information Retrieval), машинное обучение (Machine Learning), интеллектуальный анализ текстов (Text Mining) и извлечение знаний из Internet (Web mining). Основными задачами при проектировании СКР являются: выбор рекламных объявлений, соответствующих контексту действий пользователя; выделение ключевых фраз из текстов страниц; рекомендация и сравнение ключевых фраз. При решении этих задач, как правило, применяются подходы, основанные на различных обучающих (Murdock V., Zhang W. и др.) и классификационных (Broder A., Josifovski V., Anagnostopoulos A. и др.) алгоритмах. Помимо этого широко распространены методы байесовской сети (Ribeiro-Neto B., Cristo M., Golgher P. и др.), генетического программирования (Lacerda A., Cristo M., Ribeiro-Neto B. и др.), а также метод анализа текстов TF-IDF (Baeza-Yates R., Ribeiro-Neto B., Yih W., Goodman J. и др.). Необходимо отметить, что отечественных разработок, посвященных организации СКР, крайне мало, однако ведутся исследования в смежных областях, таких как информационный поиск и компьютерная лингвистика. Наиболее авторитетными источниками информации в данных областях являются материалы международных конференций ДИАЛОГ и РОМИП.

Использование существующих методов применительно к развивающимся региональным СКР зачастую вызывает трудности. Дело в том, что большинство подходов основывается на различных видах статистического анализа и машин-

ного обучения. При этом необходимо предварительно обучить модель на некотором объеме исходных данных. Такие обучающие выборки отсутствуют как таковые в открытом доступе, поэтому эти данные необходимо накапливать в самой системе контекстной рекламы или поисковой системе. А поскольку на этапе разработки и в первое время существования СКР статистических данных в необходимых объемах просто нет, то соответственно эффективность применения упомянутых методов крайне мала.

Ряд методов основан на принципах полного перебора: каждое объявление сравнивается с содержанием страницы или поисковым запросом. Учитывая, что количество объявлений может достигать тысяч, а запросов и текстов страниц – сотен тысяч, использование подобных методов в режиме реального времени невозможно из-за высоких требований к вычислительным мощностям.

Одним из наиболее важных является вопрос о том, какие факторы и с какой степенью должны учитываться при выборе и ранжировании объявлений. Подавляющее большинство систем контекстной рекламы при выборе руководствуется, прежде всего, стоимостью объявлений, стремясь показать наиболее дорогие объявления. При этом такие факторы, как интересы пользователя или, например, степень соответствия ключевых фраз объявления контексту страницы играют второстепенную роль либо не учитываются вовсе. В условиях большого количества пользователей и рекламных объявлений такой подход вполне оправдан, однако он оказывается не эффективным для развивающихся региональных СКР, конкуренция рекламодателей в которых мала, а количество пользователей невелико. Дело в том, что прибыль зависит не только от ставки за переход по объявлению, но и от количества переходов, а значит, от релевантности показываемых объявлений. Таким образом, более перспективным для региональной СКР представляется подход, ориентированный на выбор рекламных объявлений, наиболее релевантных интересам пользователей. При этом предлагается учитывать множество различных факторов, характеризующих как интересы пользователей и контекст страницы, так и само объявление. Вопросы одновременного использования в рамках систем контекстной рекламы нескольких факторов, влияющих на ранжирование объявлений, недостаточно проработаны в научных исследованиях.

Ввиду отсутствия точных количественных оценок релевантности¹ рекламных объявлений представляется целесообразным при их выборе и ранжировании использовать аппарат нечетких множеств. При этом определение степени релевантности объявлений как по отдельным факторам, так и по их совокупности не должно опираться на методы обработки больших объемов статистических данных и машинного обучения, а также на методы полного перебора.

Цель работы. Разработка алгоритмов и программных средств организации системы контекстной рекламы, интегрированной с региональной поисковой системой, обеспечивающих высокую релевантность рекламных объявлений информационным потребностям пользователя.

¹ Релевантность – субъективное понятие, под которым в системах текстового поиска понимается соответствие ответов системы информационным потребностям пользователя. Концептуально степень релевантности можно измерять вещественным числом от 0 до 1. Используется также термин «пертинентность».

Задачи для достижения поставленной цели

1. Анализ и сравнение существующих систем контекстной рекламы, а также методов их организации.
2. Разработка алгоритмов, не использующих большой объем накопленных данных, для выбора релевантных объявлений по отдельным факторам (поисковому запросу; тегам, характеризующим страницу; навигационной истории и истории поисковых запросов пользователя; качеству и эффективности объявлений) и по совокупности факторов, а также алгоритма выделения тегов из текста страницы.
3. Проведение экспериментальных исследований разработанных алгоритмов.
4. Создание программного комплекса, реализующего разработанные алгоритмы, включающего СКР и региональную поисковую систему.

Методы исследования. В ходе диссертационного исследования были использованы модели и методы теории нечетких множеств, статистического и морфологического анализа, а также методы, относящиеся к областям знаний Information Retrieval, Text Mining и Web Mining. При реализации программного комплекса был применен ряд методов объектно-ориентированного проектирования и программирования.

Научная новизна

1. Впервые предложен алгоритм выбора рекламных объявлений, основанный на формировании нечеткого множества релевантных объявлений по множеству факторов, характеризующих как сами рекламные объявления и контекст страницы, так и личные предпочтения пользователей.
2. Разработан новый алгоритм выбора рекламных объявлений в соответствии с поисковым запросом пользователя, позволяющий учитывать как степень совпадения, так и порядок слов запроса и ключевой фразы объявления с использованием коэффициента Джаккарда и наибольшей общей подпоследовательности фраз.
3. Предложен новый алгоритм выбора рекламных объявлений по тегам текущей страницы, учитывающий релевантность тегов по отношению к тексту страницы и степень их совпадения с ключевыми фразами объявлений. Для выделения тегов из текстов страниц и определения их релевантности разработан алгоритм, модифицирующий метод «ко-появлений» для выделения термов.
4. Разработаны новые алгоритмы выбора рекламных объявлений в соответствии с поисковой и навигационной историями пользователя. Первый учитывает не только степень релевантности ключевых фраз объявления наиболее популярным поисковым запросам пользователя, но и частоту встречаемости запросов в истории, второй – частоту встречаемости категорий, к которым принадлежит сайт объявления, в истории посещенных пользователем сайтов.
5. Предложены новые алгоритмы выбора объявлений в соответствии с их эффективностью и качеством. Первый, основанный на расчете показателя CTR (Click-Through Rate), отличается тем, что учитывает лишь «гарантированные»

просмотры объявления пользователями. Второй, основанный на расчете показателя качества объявления, учитывает не только количество значимых слов объявления, находящихся в тексте целевой страницы, но и их наличие в различных html-тегах.

Основные положения, выносимые на защиту

1. Разработанные алгоритмы выбора релевантных рекламных объявлений, позволяющие учитывать контекст страницы, характеризуемый текущим поисковым запросом или тегами страницы, прошлые интересы пользователя, характеризующиеся поисковой и навигационной историей пользователя, а также характеристики объявлений.

2. Результаты экспериментов, показавшие превосходство разработанных алгоритмов по сравнению с аналогами, а также высокую степень соответствия результатов их работы оценкам экспертов.

3. Созданная система «Поисколог», реализующая предложенные алгоритмы, позволяющая пользователям осуществлять различные виды поиска по региональным ресурсам и получать в качестве дополнительной релевантной информации рекламные объявления, а рекламодателям – создавать и настраивать объявления.

Степень достоверности результатов работы. Достоверность результатов работы обеспечивается корректным применением методов рассматриваемой предметной области и теории нечетких множеств, согласованностью сформулированных выводов с результатами экспериментов. Кроме того, достоверность подтверждается внедрением разработанных алгоритмов в рамках системы контекстной рекламы, интегрированной с региональной поисковой системой.

Теоретическая значимость работы. Предложенные автором алгоритмы, базирующиеся на аппарате нечетких множеств, развивают существующие подходы к организации систем контекстной рекламы в направлении обеспечения высокого уровня релевантности объявлений информационным потребностям пользователей. Разработанные алгоритмы выделения тегов, сравнения запросов и ключевых фраз, сопоставления текстов объявлений и целевых страниц вносят вклад в развитие методов информационного поиска и анализа текста.

Реализация результатов работы. Разработанные алгоритмы использованы при создании СКР в рамках интегрированной системы «Поисколог» (свидетельство о регистрации в "Отраслевом фонде алгоритмов и программ" №10058 от 21.02.2008г. Номер госрегистрации №50200800460 от 28.02.2008 г.).

Практическая значимость работы. Разработанная в рамках диссертационной работы интегрированная поисковая система «Поисколог» используется для поиска информации в томском сегменте Интернета. С помощью данной системы возможен поиск как текстовой, так и мультимедиа-информации. Использование технологии разделения поиска на отдельные вертикали позволяет пользователям производить поиск необходимой информации лишь в интере-

сующей их области, не просматривая множество нерелевантных результатов других тематик.

Разработанная СКР, интегрированная с поисковой системой, позволяет рекламодателям продвигать свои сайты посредством показа объявлений в результатах поиска и на информационных страницах ПС. При этом рекламные объявления, показываемые пользователю, максимально соответствуют контексту страницы и его личным предпочтениям. Кроме того, созданная технология даёт возможность использовать СКР в отрыве от поисковой системы для показа рекламных объявлений на сторонних сайтах.

Практическая ценность работы подтверждается актами внедрения в ряде коммерческих фирм г. Томска.

Личный вклад автора. Все результаты, составляющие основное содержание диссертации, получены автором самостоятельно.

Апробация работы. Основные результаты диссертационной работы докладывались на следующих конференциях: всероссийской конференции студентов, аспирантов и молодых ученых «Энергия молодых – экономике России» (Томск, 2003); всероссийских научно-технических конференциях аспирантов и молодых ученых «Научная сессия ТУСУР» (Томск, 2004, 2005, 2006, 2007); всероссийских научно-технических конференциях студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии» (Томск, 2004, 2006, 2007); международных научно-практических конференциях «Средства и системы автоматизации» (Томск, 2004, 2007); международной научно-практической конференции «Современные техника и технологии» (Томск, 2006).

Публикации. По теме диссертационной работы опубликовано 15 научных работ, в том числе 4 статьи (все в изданиях, рекомендованных ВАК для опубликования результатов диссертаций). Список публикаций приведен в конце автореферата.

Структура и объем работы. Диссертационная работа включает: введение, четыре главы, заключение, список литературы из 198 наименований, 7 приложений. Общий объем диссертации составляет 244 страницы машинописного текста. Работа содержит 40 рисунков, 18 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность работы, приводятся цель, основные положения, выносимые на защиту, задачи и методы исследования, формулируется научная новизна, теоретическая и практическая значимость полученных результатов, приводятся результаты внедрения, а также сведения о публикациях, апробации работы, объеме и структуре диссертации.

Первая глава посвящена сравнению контекстной рекламы с другими видами продвижения, а также обзору современных СКР и методов их организации.

В настоящее время системы контекстной рекламы являются одним из наиболее эффективных средств продвижения сайта в Интернете на всех стадиях его существования. Этот единственный способ, который, в отличие от других видов рекламы, может дать сайту в сжатые сроки гарантированную и при этом целевую аудиторию. Именно по этой причине контекстная реклама уверенно лидирует среди баннерной, e-mail и прочих видов рекламы.

Проведенный анализ как мирового, так и российского рынка контекстной рекламы показал, что, несмотря на ее высокую удельную долю в общем объеме Интернет-рекламы, в данном сегменте существует всего лишь несколько крупных игроков, занимающих более 95% рынка. Лидирующие позиции занимают компании, имеющие в распоряжении крупные поисковые системы и предоставляющие наиболее инновационные на сегодняшний момент достижения в области функциональности и интерфейса системы.

На основе анализа лидирующих СКР был сформирован набор ключевых факторов, используемых ими для подбора объявлений. Большинство систем в качестве основного используют фактор стоимости перехода по объявлению (ставку), а также широко используется значение CTR объявления. Факторы, основанные на качестве целевой страницы и объявления, имеют совсем небольшой вес при вычислении релевантности объявлений и используются, как правило, лишь для новых объявлений, CTR которых ещё не определен или нестабилен. При выборе по фактору соответствия ключевых фраз объявления поисковому запросу никак не учитывается количество совпавших слов, их порядок и словоформы. При вычислении качества целевой страницы не учитывается различная важность слов, которая зависит от их нахождения в заголовке или тексте объявления и от их местоположения в различных html-тегах целевой страницы. Также недостаточное распространение имеют поведенческие технологии, особенно учитывающие историю посещенных пользователем сайтов.

Был проведен обзор методов, используемых для решения таких задач, как максимизация прибыли СКР, выбор релевантных объявлений, организация таргетинга, выделение и рекомендация ключевых слов и др. Анализ методов показывает, что наибольшее внимание исследователей уделяется, как правило, вопросам максимизации прибыли системы при сохранении должного уровня релевантности рекламных объявлений. Большая часть используемых методов основывается на принципах машинного обучения или статистической обработки больших объемов накопленной информации.

Во второй главе предлагается новый подход к организации системы контекстной рекламы, интегрированной с РПС, включающий ряд новых алгоритмов.

Региональная поисковая система представляет собой набор поисковых модулей, каждый из которых осуществляет поиск в определенных источниках или по определённом типу информации. В качестве региона, сайты которого будут донорами информации, в данной работе рассматривается Томск и Томская область. Основными модулями поиска являются (рис. 1): «Мета-поиск по Web», «Поиск по новостям», «Поиск по каталогу сайтов», а также модули поиска по различным видам мультимедиа-информации.

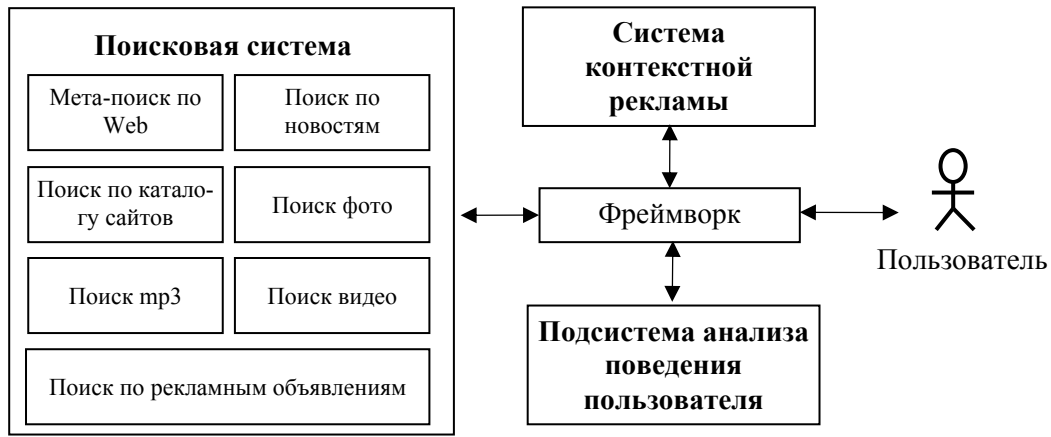


Рис. 1. Общая схема интегрированной РПС

Использование СКР при поисковой системе позволяет не только обеспечить пользователя контекстной информацией, но и быть успешным маркетинговым инструментом для рекламодателей и владельцев ПС. Кроме того, интеграция СКР и ПС позволяет максимально увеличить эффективность обеих систем, а также предоставить поиск по рекламным объявлениям. Для использования поведенческих технологий в составе интегрированной РПС необходим также модуль анализа поведения пользователя, который сохраняет данные обо всех действиях пользователя в его виртуальном профиле. Координирует работу всей интегрированной системы фреймворк (каркас программной системы), управляющий всеми запросами к системе и выводом информации пользователю.

Для выбора рекламных объявлений, показываемых в СКР, *разработан алгоритм выбора объявлений по множеству факторов.*

Обозначим исходное множество рекламных объявлений системы через $O = \{o_i\}$. Каждое объявление характеризуется рядом параметров:

$$o_i = \langle A(o_i), F(o_i), R(o_i), K(o_i) \rangle,$$

где $A(o_i)$ – множество видимых параметров, включающее заголовок объявления $ah(o_i)$, его текст $at(o_i)$, ссылку (url-адрес) $al(o_i)$, видимый url-адрес $av(o_i)$; $F(o_i)$ – множество параметров фокусировки, таких как ключевые фразы $\{fk_j(o_i)\}$, стоп-слова $\{ws_m(o_i)\}$, категории объявления $\{ca_n(o_i)\}$; $R(o_i)$ – множество параметров, задающих ограничения на показ объявления, в том числе $rp(o_i)$ – стоимость объявления, $rd(o_i)$ – дата окончания показа объявления, $rg(o_i)$, $rt(o_i)$, $rv(o_i)$ – параметры, соответствующие ограничениям географического, временного и частотного таргетинга; $K(o_i)$ – множество параметров, включающее показатель эффективности объявления $eCTR(o_i)$ и показатель качества объявления $aq(o_i)$.

Искомое множество Op релевантных рекламных объявлений можно определить как нечеткое:

$$Op = \{ \langle o, \mu_{Op}(o) \rangle \}.$$

Из этого множества необходимо выбрать некоторое количество «наиболее релевантных» объявлений On , которые и будут показаны пользователю. Для этого определим On как подмножество множества Op α -уровня:

$$On \subseteq Op_\alpha \subset Op \subset O,$$

где $Op_\alpha = \{o \in O \mid \mu_{Op}(o) > \alpha\}$ ($\alpha \in (0,1]$) выбирается в соответствии с опытными данными). При этом мощность множества On не может быть больше максимального количества одновременно показываемых на странице объявлений N_{ob} ($|On| \leq N_{ob}$, $N_{ob} \in [7,10]$), что объясняется удобством интерфейса пользователя.

Выбор объявления может происходить в соответствии с интересами пользователя, либо в соответствии с качеством и эффективностью самого объявления. При этом помимо текущих интересов пользователя могут также учитываться интересы, которые были у пользователя в прошлом. Текущие интересы пользователя определяются или по соответствию запроса ключевым фразам объявления (при наличии поискового запроса) или по соответствию тематики текущей страницы тематике объявления. Прошлые интересы пользователя определяются в соответствии с историей поисковых запросов или историей посещенных сайтов. Таким образом, можно выделить следующие факторы, влияющие на выбор объявления: $g_1 = \langle \text{текущий запрос} \rangle$; $g_2 = \langle \text{текущая тематика} \rangle$; $g_3 = \langle \text{история запросов} \rangle$; $g_4 = \langle \text{история тематик} \rangle$; $g_5 = \langle \text{эффективность объявления} \rangle$; $g_6 = \langle \text{качество объявления} \rangle$.

По каждому фактору g_k , $k = \overline{1, m}$ определяется степень соответствия объявлений множеству Op , т.е. формируется свое нечеткое множество релевантных объявлений Og_k с функцией принадлежности $\mu_{Og_k}(o)$. Принадлежность множеству релевантных объявлений по всем факторам будем определять по формуле выпуклой комбинации нечетких множеств:

$$\mu_{Op}(o) = \sum_{k=1}^m w_k \cdot \mu_{Og_k}(o), \quad \sum_{k=1}^m w_k = 1,$$

где w_k – вес k -го фактора.

Выбор факторов, по которым будет осуществляться подбор объявлений, зависит от текущей ситуации, а именно от того, с каким модулем системы в настоящий момент взаимодействует пользователь и от выполняемых им действий (поиск или просмотр информации). При этом ввиду того, что модули поиска «Поиск фото», «Поиск mp3», «Поиск видео» очень схожи, их можно условно объединить под одним названием – «Поиск по мультимедиа». Учитывая, что при взаимодействии с некоторыми модулями пользователь может осуществлять только поиск, можно выделить следующие ситуации: $s_1 = \langle \text{Мета-поиск/поиск} \rangle$, $s_2 = \langle \text{Каталог сайтов/поиск} \rangle$, $s_3 = \langle \text{Каталог сайтов/просмотр} \rangle$, $s_4 = \langle \text{Новости/поиск} \rangle$, $s_5 = \langle \text{Новости/просмотр} \rangle$, $s_6 = \langle \text{Мультимедиа/поиск} \rangle$, $s_7 = \langle \text{Объявления/поиск} \rangle$. Состав и важность используемых факторов выбора объявлений для различных ситуаций будут различными. Так, например, в ситуациях s_1 , s_2 и s_4 используются множества факторов $G_1 = G_2 = G_4 = \{g_1, g_3, g_4, g_5, g_6\}$ с соответствующими векторами весов $W_1 = W_2 = W_4 = \{0.7, 0.12, 0.08, 0.04, 0.06\}$. Веса для каждого из факторов определялись экспериментально на основе метода последовательного сравнения Черчмена-Акоффа.

Факторы выбора объявлений можно разделить на два класса – основные (G^{os}), характеризующие интересы пользователя (g_1, g_2, g_3, g_4), и дополнительные (G^{dop}), характеризующие само объявление (g_5, g_6). Дополнительные факторы

играют вспомогательную роль, их целесообразно использовать для выбора не из всего исходного множества объявлений, а для дополнительного отбора из множества объявлений, выбранных с помощью основных факторов. Это позволяет существенно сократить время и вычислительные затраты.

При выборе объявлений для показа необходимо также учитывать ограничения, в соответствии с которыми те или иные объявления могут быть запрещены к показу. Ограничения могут определяться самими рекламодателями либо в целях оптимизации проводимой рекламной кампании (ограничения на общий бюджет и длительность кампании), либо для оптимальной настройки показов объявлений с помощью механизмов таргетинга.

Таким образом, алгоритм в обобщенном виде выглядит следующим образом. Сначала для каждого из основных факторов, соответствующих текущей ситуации, формируются нечеткие множества объявлений с учетом ограничений. Затем из множества выбранных объявлений аналогичным образом формируются нечеткие множества по дополнительным факторам. Далее строится нечеткое множество по всем факторам с учётом их весов, из которого затем выбираются объявления с наибольшими значениями функции принадлежности.

Рассмотрим алгоритмы, осуществляющие формирование нечетких множеств по отдельным факторам.

Алгоритм выбора объявлений по поисковому запросу позволяет определить для каждого объявления значение $\mu_{Og_1}(o)$ функции принадлежности множеству релевантных рекламных объявлений по фактору g_1 <текущий запрос>. Исходной информацией является:

1. Поисковый запрос пользователя fz , представляющий собой последовательность слов запроса: $fz = \langle wz_1, wz_2, \dots, wz_{N_z} \rangle$.

2. Совокупность ключевых фраз $\{fk_j\}$ объявления, $fk_j = \langle wk_{1j}, wk_{2j}, \dots, wk_{N_{kj}} \rangle$. Для каждой фразы рекламодатель задает один из следующих типов ее соответствия поисковому запросу: точное соответствие – фраза должна совпадать с запросом ($fk_j = fz$); точное морфологическое соответствие – множество слов фразы в базовой форме должно совпадать с множеством слов запроса в базовой форме ($\{wk_{nj}^b\} = \{wz_l^b\}$); фразовое соответствие – все слова из базовой формы фразы должны содержаться в базовой форме запроса ($\{wk_{nj}^b\} \subseteq \{wz_l^b\}$); широкое соответствие – хотя бы одно из слов базовой формы фразы должно содержаться в базовой форме запроса ($\{wk_{nj}^b\} \cap \{wz_l^b\} \neq \emptyset$).

3. Совокупность стоп-слов $\{ws_m\}$, заданных для объявления.

Принадлежность объявления o множеству Og_1 определяется по всем ключевым фразам. Обозначим функцию принадлежности объявления данному множеству, определяемую по ключевой фразе fk_j , как $\mu_{Og_1}(o, fk_j)$. Значение данной функции, характеризующее степень сходства запроса и ключевой фразы, будем определять на основе вычисления «наибольшей общей подпоследовательности» слов и коэффициента Джаккарда, что позволит учитывать порядок слов в запросе и степень совпадения его слов со словами ключевой фразы и, таким образом, более точно определять степень релевантности объявлений.

$$\mu_{Og_1}(o, fk_j) = \left(\frac{LCS(fz^b, fk_j^b) \cdot K}{Nz} + 1 - K \right) \cdot JAC(z, k),$$

где Nz – количество слов в запросе fz ; $LCS(fz^b, fk_j^b)$ – наибольшая общая подпоследовательность слов в базовых формах запроса fz^b и ключевой фразы fk_j^b ; K – коэффициент, определяющий вес параметра LCS в формуле и равный 0.02; $JAC(z, k)$ – коэффициент Джаккарда, рассчитанный для векторов запроса и ключевой фразы. Векторы соответствия слов запроса и слов ключевой фразы словам из векторного пространства $\{wp_m\} = \{wz_l^b\} \cup \{wk_{nj}^b\}$ предлагается рассчитывать по формулам:

$$z_m = \begin{cases} 1, & \text{если } wp_m \in \{wz_l^b\} \\ 0, & \text{иначе} \end{cases}, \quad k_m = \begin{cases} 1, & \text{если } \exists wk_{nj}^b \mid wk_{nj}^b = wp_m, wk_{nj}^b \in \{wz_l^b\} \\ 0.9, & \text{если } \exists wk_{nj}^b \mid wk_{nj}^b = wp_m, wk_{nj}^b \in \{wz_l^b\} \\ 0.5, & \text{если } wp_m \in \{wk_{nj}^b\} \\ 0, & \text{иначе} \end{cases}.$$

Функция принадлежности объявления по всем ключевым фразам определяется с помощью операции $\max$: $\mu_{Og_1}(o) = \max_j \mu_{Og_1}(o, fk_j)$.

Обобщенно алгоритм выглядит следующим образом. Выбираются объявления, ключевые фразы которых содержат слова запроса, и происходит проверка на отсутствие стоп-слов объявления в поисковом запросе. Далее для каждой ключевой фразы объявления производится проверка на выполнение условий, соответствующих её типу, и рассчитывается значение функции принадлежности объявления по данной ключевой фразе. После этого рассчитывается значение функции принадлежности по всем ключевым фразам.

Алгоритм выбора объявлений по тегам страницы позволяет определить для каждого объявления значение $\mu_{Og_2}(o)$ функции принадлежности множеству релевантных рекламных объявлений по фактору g_2 <текущая тематика>. В целях снижения вычислительных затрат для выбора объявлений в соответствии с содержанием страницы предлагается использовать не весь её текст, а только теги – ключевые слова, характеризующие содержание текста.

Исходной информацией является:

1. Множество тегов $\{ft_k\}$, характеризующих текущую страницу, $ft_k = \langle wt_{1k}, wt_{2k}, \dots, wt_{Nt k} \rangle$. Каждому тегу сопоставляется вес v , отражающий важность тега.
2. Совокупность ключевых фраз $\{fk_j\}$ объявления.
3. Совокупность стоп-слов $\{ws_m\}$, принадлежащих объявлению.

Принадлежность объявления o множеству Og_2 определяется по всем ключевым фразам объявления и по всем тегам страницы. Обозначим функцию принадлежности объявления данному множеству, определяемую по ключевой фразе fk_j и по тегу ft_k как $\mu_{Og_2}(o, ft_k, fk_j)$. Значение данной функции, характеризующее степень соответствия тега страницы ключевой фразе объявления, будем определять на основе количества совпавших слов и веса тега, что даст возможность учесть не только степень совпадения тега и фразы, но и важность данного тега по отношению к тексту страницы и, в конечном счете, повысить адекватность расчета релевантности объявлений:

$$\mu_{Og_2}(o, ft_k, fk_j) = v(ft_k) \cdot \frac{Ntk_{kj}}{Nk_j},$$

где $v(ft_k)$ – предварительно рассчитанный относительный вес k -го тега для данной страницы, Ntk_{kj} – количество найденных слов wt_{lk} тега ft_k в ключевой фразе fk_j ; Nk_j – количество слов в ключевой фразе fk_j .

Функция принадлежности объявления по всем ключевым фразам и по всем тегам определяется с помощью операции максимума:

$$\mu_{Og_2}(o) = \max_k \max_j \mu_{Og_2}(o, ft_k, fk_j).$$

Таким образом, для каждого тега выбираются объявления, ключевые фразы которых содержат слова тега, и происходит проверка на отсутствие стоп-слов в тегах. Для каждой ключевой фразы объявления производится проверка на выполнение условий, соответствующих её типу, и рассчитывается значение функции принадлежности по данной ключевой фразе и тегу. Затем определяется значение функции принадлежности по всем ключевым фразам. Процедура выполняется для всех тегов, после чего рассчитывается значение функции принадлежности по всем тегам.

Алгоритм выделения тегов из текста страницы предназначен для выбора тегов, характеризующих содержание текущей страницы, и определения их весов на основе анализа только одного текста.

Исходной информацией является заголовок NH и текст NT страницы.

Используется модифицированный метод «ко-появлений». В соответствии с ним из множества предложений заголовка страницы NH и множества предложений текста страницы NT выделяются одно-, двух- и трехсложных термы, преобразованные к базовой форме: $P = \{p_m\} = P^h \cup P^t$ (P^h , P^t – множества термов, выделенных из заголовка и текста страницы). При этом в словосочетания, составляющие термы, попадают только слова, идущие подряд в предложении и относящиеся к «значимым» частям речи (исключая предлоги, союзы и т.д.). Для каждого терма p_m определяется количество $f_q(p_m)$ его появлений в предложениях из множеств NH и NT . Затем формируется подмножество, так называемых, «частотных» термов $FR = \{p_n^{fr}\}$, включающее термы, у которых количество появлений превышает единицу ($FR = \{p_n \mid f_q(p_n) > 1\}$), а также его подмножество $G = \{p_n^g\}$ «часто встречающихся термов» с наибольшими значениями частоты ($|G| \leq 0.3 \cdot |FR|$). Для каждого p^{fr} рассчитывается величина $\chi^2(p^{fr})$ по формуле:

$$\chi^2(p^{fr}) = \sum_{p^g \in G} \frac{(f_{qf}(p^{fr}, p^g) - nps(p^{fr}) \cdot nps(p^g) / npt)^2}{nps(p^{fr}) \cdot nps(p^g) / npt},$$

где $f_{qf}(p^{fr}, p^g)$ – частота совместного появления термов p^{fr} и p^g в одном предложении, $nps(p^{fr})$ – количество термов в предложениях, в которых появляется терм p^{fr} , $nps(p^g)$ – количество термов в предложениях, в которых появляется терм p^g , npt – общее количество термов в тексте. Для того чтобы больший вес получали составные термы, рассчитывается величина $\chi'^2(p^{fr})$:

$$\chi'^2(p^{fr}) = \chi^2(p^{fr}) - \max_{p^{fr} \in G} \{\chi^2(p^{fr})\}.$$

Автором предлагается модификация данного алгоритма к задаче выделения тегов для СКР. Для этого, прежде всего, предлагается использовать для слов заголовка повышающий коэффициент $k(p^{fr})$, т.к. эти слова должны иметь больший вес при выделении тегов:

$$\chi_h'^2(p^{fr}) = k(p^{fr}) \cdot \chi'^2(p^{fr}), \quad k(p^{fr}) = \begin{cases} 1,5, & \text{если } p^{fr} \in P^h \\ 1, & \text{если } p^{fr} \in P^t \end{cases}.$$

Затем предлагается исключить из рассмотрения избыточные термы, входящие в состав более сложных термов, которые имеют меньшее значение $\chi_h'^2$:

$$\chi_{hf}'^2(p_n^{fr}) = \begin{cases} 0, & \text{если } p_n^{fr} \text{ такой, что } p_n^{fr} \subset p_m^{fr} \text{ и } \chi_h'^2(p_n^{fr}) \leq \chi_h'^2(p_m^{fr}) \\ \chi_h'^2(p_n^{fr}), & \text{иначе} \end{cases}.$$

Определение весов термов на основе полученных значений $\chi_{hf}'^2$ предлагается осуществлять путем их нормализации. В качестве нормирующего коэффициента выбирается максимальное значение $\chi_{hf \max}'^2 = \chi_{hf}'^2(p_1^{fr})$, однако в случаях, когда оно существенно превышает ближайшее к нему значение ($\chi_{hf}'^2(p_2^{fr})$), предлагается занижать этот максимум. Т.е. если $\chi_{hf}'^2(p_1^{fr}) \geq 1,5 \cdot \chi_{hf}'^2(p_2^{fr})$, то $\chi_{hf}'^2(p_1^{fr}) = \chi_{hf \max}'^2 = 1,5 \cdot \chi_{hf}'^2(p_2^{fr})$. Вес терма рассчитывается по формуле:

$$v(p_n^{fr}) = \frac{\chi_{hf}'^2(p_n^{fr})}{\chi_{hf \max}'^2}.$$

В искомое множество тегов $\{ft_k\}$ входят лишь термы, вес которых превышает пороговую величину: $v(p_n^{fr}) \geq 0.2$.

Так как метод «ко-появлений» не эффективен на малых текстах, то в этом случае предлагается выделять теги на основе метода TF. Процедура выделения тегов и расчета весов в данном случае аналогична шагам предложенного алгоритма, однако в данном случае используются непосредственно значения частоты термов $fq(p_m)$.

Алгоритм выбора объявлений в соответствии с историей запросов пользователя позволяет определить для каждого объявления значение $\mu_{og_3}(o)$ функции принадлежности множеству релевантных рекламных объявлений по фактору g_3 <история запросов>. Исходной информацией является:

1. История запросов пользователя $Iz = \langle fz_1, fz_2, \dots, fz_{Niz} \rangle$, включающая не только запросы, которые пользователь вводил в системе, но также и запросы, по которым пользователь переходил в систему с других поисковых систем.
2. Совокупность ключевых фраз $\{fk_j\}$, характеризующих объявление.
3. Совокупность стоп-слов $\{ws_m\}$, принадлежащих объявлению.

В основу данного алгоритма положен алгоритм выбора объявлений по фактору g_1 , т.е. по соответствию поискового запроса ключевым фразам объявлений. Фактически алгоритм выбора объявлений по запросу расширяется до случая нескольких поисковых запросов, введенных пользователем ранее.

Из списка запросов Iz выделяется список запросов Iz^l , введенных пользователем за последний месяц. Данное ограничение обусловлено тем, что целесообразно учитывать лишь краткосрочные интересы пользователя, т.к. чем свежее эта информация, тем более точными будут предположения об информационных потребностях пользователя. Выделяется множество Iz^{lu} запросов, встречающихся в списке Iz^l , и для каждого запроса fz_k этого множества определяется вес:

$$v(fz_k) = \frac{\varphi(fz_k)}{\max_k \varphi(fz_k)},$$

где $\varphi(fz_k)$ – количество вхождений k -го запроса в список Iz^l .

Из множества Iz^{lu} выделяются запросы с максимальным весом, и для каждого k -го запроса формируется нечеткое множество Og_3^k (аналогично тому, как формируется множество Og_1). Обозначим функцию принадлежности объявления данному множеству, определяемую в соответствии с запросом fz_k , через $\mu_{Og_3}(o, fz_k)$. Тогда функция принадлежности множеству Og_3 по всем запросам будет определяться по формуле: $\mu_{Og_3}(o) = \max_k v(fz_k) \cdot \mu_{Og_3}(o, fz_k)$. Это позволяет дать большее значение функции принадлежности объявлениям, выбранным по запросам, которые чаще встречались в истории запросов пользователя.

Алгоритм выбора объявлений в соответствии с историей посещенных пользователем сайтов позволяет определить для каждого объявления значение $\mu_{Og_4}(o)$ функции принадлежности множеству релевантных рекламных объявлений по фактору g_4 <история тематик>. Исходные данные:

1. Список категорий $Ca = \langle ca_1, ca_2, \dots, ca_{Na} \rangle$, к которым принадлежит сайт объявления.

2. Список сайтов $Is = \langle is_1, is_2, \dots, is_{Ns} \rangle$ из истории посещенных пользователем сайтов. При этом учитываются не только сайты, на которые пользователь переходил из результатов поиска или рекламных объявлений системы, но также и сайты, с которых он переходил в систему. Каждому is_k соответствует список категорий $Ci_k = \langle ci_{k1}, ci_{k2}, \dots, ci_{kNi} \rangle$, определенных в соответствии с региональным рубрикатором сайтов (Каталогом сайтов), в котором проклассифицированы все сайты, известные системе.

Из списка Is выделяется список Is^l сайтов, посещенных пользователем за последние два месяца, т.к. целесообразно учитывать лишь краткосрочные интересы пользователя, чтобы точнее определять информационные потребности пользователя. Далее строится общий список Ci категорий всех сайтов из истории Is^l : $Ci = Ci_1 \cdot Ci_2 \cdot \dots \cdot Ci_{Ns}$. Из списка Ci выделим множество категорий Ci^u и для каждой его категории ci_m определим её относительный вес:

$$v(ci_m) = \frac{\varphi(ci_m)}{\max_m \varphi(ci_m)},$$

где $\varphi(ci_m)$ – количество вхождений категории ci_m в список Ci .

Из множества Ci^u выделяются категории с максимальным весом, и для каждой j -й категории формируется нечеткое множество Og_4^j . Обозначим функ-

цию принадлежности объявления данному множеству, определяемую для категории ca_j , через $\mu_{og_4}(o, ca_j)$. Ее значение будем определять в соответствии с весом категории, т.е. с тем, насколько часто пользователь интересовался соответствующей тематикой:

$$\mu_{og_4}(o, ca_j) = \begin{cases} 0, & \text{если } \forall ci_m \in Ci'' \quad ca_j \neq ci_m \\ v(ci_m), & \text{если } \exists ci_m \in Ci'' \quad ca_j = ci_m \end{cases}.$$

Функция принадлежности объявления по всем его категориям будет определяться с помощью операции $\max$: $\mu_{og_4}(o) = \max_j \mu_{og_4}(o, ca_j)$.

Алгоритм выбора объявлений в соответствии с их эффективностью позволяет определить для каждого объявления значение $\mu_{og_5}(o)$ функции принадлежности множеству релевантных рекламных объявлений по фактору $g_5 = \langle \text{эффективность объявления} \rangle$. Исходными данными являются:

1. Множество $AC = \{ac_m\}$ всех переходов по объявлению. Каждый переход ac_m сохранен в системе в виде отдельной записи в базе данных.

2. Множество $AV = \{av_k\}$ всех показов объявления. Здесь av_k – количество показов объявления в k -тый день.

Основным показателем эффективности интернет-рекламы является CTR (Click-Through Rate), равный отношению количества переходов пользователей по рекламному объявлению к количеству его показов. Однако значение CTR зачастую не отражает реальную статистику просмотров объявлений, т.к. показ объявления не гарантирует того, что пользователь его увидел. Чтобы более адекватно отразить эффективность объявления, предлагается учитывать лишь, так называемые, гарантированные просмотры объявления пользователем, т.е. только те показы объявления, которые были осуществлены в блоке рекламных объявлений, в котором в последствии был осуществлен переход хотя бы по одному из них. CTR, рассчитанный соответствующим образом, будем называть эффективным CTR и обозначать как eCTR. Также предлагается значение eCTR рассчитывать не за всё время существования объявления, а лишь за последний месяц, чтобы учитывалась лишь текущая актуальная его эффективность.

Таким образом, из множеств AC и AV выделяются подмножества AC^l и AV^l соответственно переходов и показов за рассчитываемый срок. При этом в AV^l включаются только гарантированные показы. Значение eCTR объявления определяется по следующей формуле:

$$eCTR(o) = \frac{Nc}{Nv},$$

где $Nc = |AC^l|$ – количество всех переходов по объявлению за рассчитываемый срок, $Nv = \sum_{k=1}^{30} av_k$ ($av_k \in AV^l$) – количество гарантированных просмотров объявления за рассчитываемый срок. При $Nv = 0$ будем считать, что $eCTR = 0$.

Значение eCTR для любого объявления, как правило, не превышает максимальную отметку в 0,2, поэтому эффективность объявлений с $eCTR > 0,2$ можно считать максимальной. При малых значениях просмотров объявлений

($Nv \leq 20$), значение eCTR не может адекватно отражать его эффективность, поэтому для оценки эффективности таких объявлений предлагается использовать среднее значение $eCTR^{avg}$ для объявлений с аналогичными категориями.

Таким образом, значение функции принадлежности будем рассчитывать на основе eCTR или $eCTR^{avg}$:

$$\mu_{O_{g_5}}(o) = \begin{cases} \min(1, eCTR(o) \cdot \frac{1}{0,2}), & \text{если } Nv > 20 \\ eCTR^{avg}(o), & \text{если } Nv \leq 20 \end{cases}.$$

Алгоритм выбора объявлений в соответствии с их качеством позволяет определить для каждого объявления значение $\mu_{O_{g_6}}(o)$ функции принадлежности множеству релевантных рекламных объявлений по фактору $g_6 = \langle \text{качество объявления} \rangle$. Исходными данными являются: заголовок ah объявления; текст at объявления; html-текст lpt целевой страницы объявления.

Формируются множества значимых слов заголовка и текста объявления (в базовой форме) путем удаления всех дубликатов слов и стоп-слов (предлогов, союзов, частиц и т.д.): $\{wa_n^{bu}\} = \{wah_m^{bu}\} \cup \{wat_l^{bu}\}$. Аналогично формируется множество значимых слов целевой страницы и его подмножества слов, содержащихся в html-тегах «title», «h1», «h2», «em»/«i», «strong»/«b»: lp^t , lp^{h1} , lp^{h2} , lp^{te} , lp^{ts} . Множество остальных слов текста обозначим как lp^t .

Значение функции принадлежности $\mu_{O_{g_6}}(o)$ будем рассчитывать на основе показателя качества объявления $AQ(o)$, отражающего степень соответствия объявления тексту целевой страницы, следующим образом:

$$\mu_{O_{g_6}}(o) = AQ(o) = 0.5 \cdot \left(\frac{\sum_{m=1}^{Nah} k_m}{Nah} + \frac{\sum_{l=1}^{Nat} k_l}{Nat} \right),$$

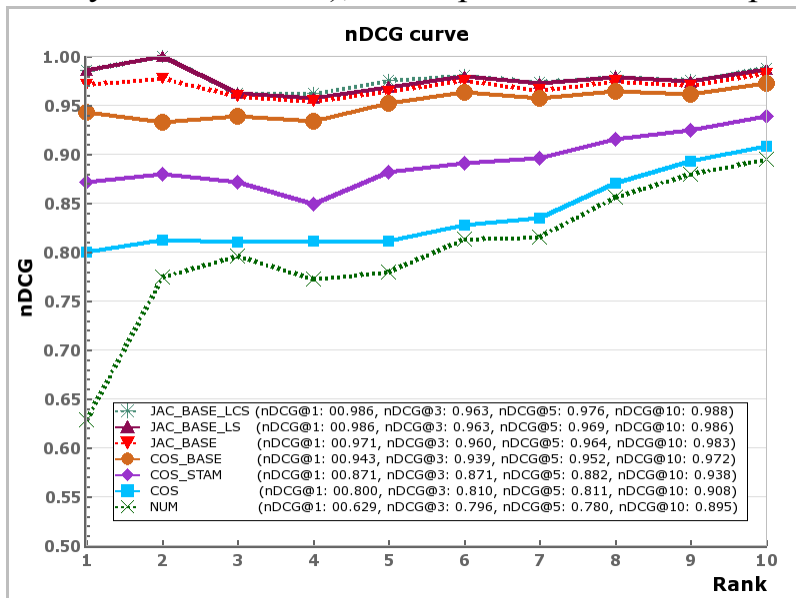
где Nah – общее количество значимых слов заголовка объявления, Nat – общее количество значимых слов текста объявления, k_m , k_l – показатели, характеризующие наличие слов из множеств $\{wah_m^{bu}\}$ и $\{wat_l^{bu}\}$ в различных подмножествах слов целевой страницы. Показатель k_i равен 1, если i -е слово содержится во множестве lp^{ti} , 0.9 – если в lp^{h1} , 0.8 – если в lp^{h2} , 0.7 – если в lp^{te} , 0.6 – если в lp^{ts} , 0.5 – если в lp^t и 0 – иначе.

Таким образом, в алгоритме учитывается не только количество слов заголовка и текста объявления, находящихся на целевой странице, но и их наличие в различных html-тегах, что позволяет придать больший вес объявлениям, содержание которых наиболее выразительно представлено на странице.

Третья глава посвящена исследованию эффективности разработанных алгоритмов и прототипа системы в условиях малого количества данных, основываясь лишь на тестовых выборках. Для проведения тестирования был спроектирован и реализован соответствующий программный комплекс, организующий обработку данных, полученных от экспертов, а также реализующий различные стратегии, выдвинутые для сравнения с оцениваемыми алгоритмами.

В ходе *тестирования алгоритма выбора рекламных объявлений по поисковому запросу* результаты его работы сравнивались с результатами реализации других алгоритмических стратегий, которые могут использоваться для сопоставления ключевых фраз объявления и поисковых запросов, а также с результатами экспертизы. Стратегии отличались способами учета встречаемости слов, сравнения слов, учёта порядка слов. В ходе проведения экспертизы экспертами были проранжированы и оценены предложенные им ключевые фразы по степени соответствия различным поисковым запросам коммерческих тематик. На той же выборке было проведено ранжирование ключевых фраз с помощью программного комплекса, реализующего выбранные стратегии.

Для оценки результатов использовались метрики DCG («обесцениваемая совокупная выгода»), её нормализованная версия nDCG. Результаты анализа



значений рассчитанных метрик, а также графиков DCG/nDCG для каждой из стратегий (рис. 2) показали, что стратегии, основанные на векторном представлении слов, учёте морфологии и порядка слов в ключевой фразе, дают наилучшие результаты по сравнению с другими. Предпочтительной стратегией на всех уровнях ответов является стратегия, реализованная в

Рис. 2. Графики значений nDCG для различных стратегий тестируемом алгоритме.

Аналогично, с использованием метрик DCG/nDCG было проведено *тестирование алгоритма выбора объявлений в соответствии с их качеством* в сравнении с результатами экспертизы, а также различными стратегиями сопоставления текстов объявлений и их целевых страниц. Наиболее эффективной по результатам тестирования была признана стратегия, реализованная в тестируемом алгоритме, использующая морфологию и учитывающая положение слов в разных частях объявления и html-тегах целевой страницы.

В ходе *тестирования работы алгоритма выделения тегов из текста страницы* для сравнения был предложен ряд стратегий выделения ключевых фраз, а также результаты ранжирования экспертами предложенных им фраз по степени соответствия их тематике страницы. Были выбраны бинарные метрики на последовательностях (MAP, MRR и Precision) и 11-точечный график полноты/точности. Наиболее эффективной по результатам тестирования была признана стратегия, реализованная в тестируемом алгоритме, основанная на подсчёте статистики совместной встречаемости ключевых фраз, учёте морфологии, выборе слов только определенных частей речи и использовании более высокого приоритета для фраз, выделенных из заголовка.

Основной целью *тестирования эффективности алгоритма выбора объявлений по множеству факторов* было выявление того, как влияет на релевантность выбора объявлений использование тех или иных факторов. Для сравнения были выбраны стратегии, отличающиеся набором учитываемых факторов. Из полученных значений оценок и DCG/nDCG был сделан вывод о том, что использование всех шести рассматриваемых факторов в качестве финальной стратегии в ряде случаев позволяет существенно улучшить результаты.

Также была произведена *оценка эффективности выбора объявлений в СКР «Поисколог» в сравнении с другими системами контекстной рекламы*. В качестве сравниваемых систем были выбраны две наиболее крупные СКР российского сегмента Интернета – «Яндекс.Директ» и «Бегун». В ходе эксперимента был выбран ряд объявлений, присутствующих одновременно по одному и тому же поисковому запросу в выдаче систем «Яндекс.Директ» и «Бегун». Выбранные объявления были занесены также в систему «Поисколог». Экспертами были проранжированы данные объявления в соответствии с поисковыми запросами, а также с качеством и эффективностью объявлений.

По результатам тестирования можно отметить достаточно высокий уровень значений рассматриваемых метрик DCG/nDCG (рис. 3) для всех оцениваемых систем на первых позициях. А также можно сделать вывод, что разработанная система в данном тестировании показала наилучший результат.

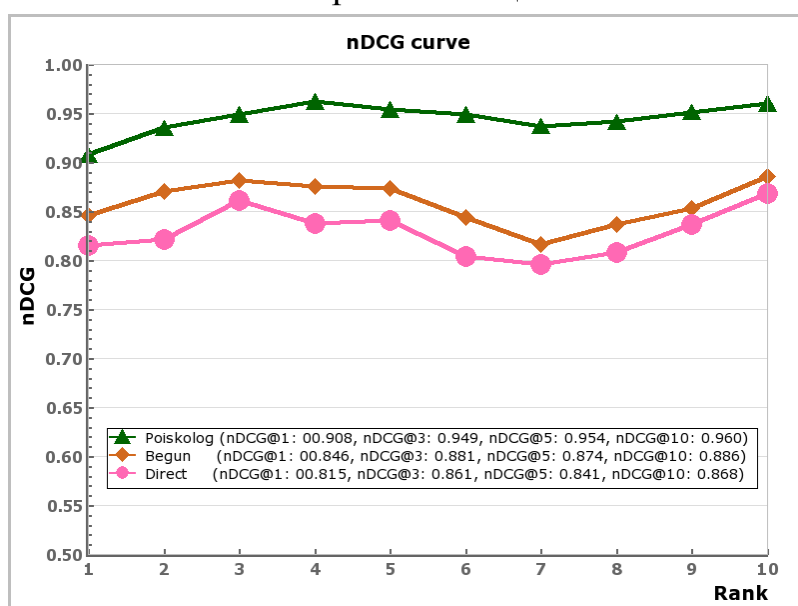


Рис. 3. Графики значений nDCG для различных СКР

В **четвертой главе** рассматривается программная реализация разработанных алгоритмов организации системы контекстной рекламы, интегрированной с региональной поисковой системой.

Основным средством реализации системы был выбран язык программирования PHP, который позволил быстро, в сжатые сроки реализовать систему. Кроме того, на выбор данной технологии оказало влияние наличие ряда развитых библиотек для поиска и морфологического анализа.

Основными компонентами разработанной интегрированной системы (ИС) «Поисколог» являются: база данных; поисковая система, содержащая модули вертикального поиска; система контекстной рекламы, включающая модули вы-

ботанная система в данном тестировании показала наилучший результат.

При проведении всех тестирований процессорные затраты и объем выделяемой оперативной памяти для различных стратегий отличались незначительно. Время, затрачиваемое на расчеты, по восходящей увеличивалось для более сложных стратегий, однако также находилось в заданных пределах.

бора рекламных объявлений и интерфейс рекламодателя; модуль анализа поведения пользователя. Каркас системы составляет MVC-фреймворк CodeIgniter, через который проходят все запросы и вызываются другие компоненты ИС.

Разработанная СКР, основанная на алгоритмах, предложенных в данной работе, предоставляет следующие возможности: с минимальными трудозатратами создавать и настраивать рекламные объявления без предварительной модерации; контролировать общий бюджет и сроки проведения рекламной кампании; использовать различные виды таргетинга с целью фокусировки кампании на целевую аудиторию; отслеживать статистику переходов по объявлениям в режиме реального времени. К достоинствам программной реализации СКР относится гибкость в управлении факторами выбора объявлений, а также возможность в дальнейшем легко расширить алгоритм выбора объявлений по множеству факторов за счёт добавления новых факторов.

В ходе создания поисковой системы были разработаны и программно реализованы алгоритмы: работы мета-поисковых систем (агрегации, фильтрации, ранжирования и кеширования результатов поиска); работы систем поиска по новостям и каталогу сайтов (агрегации, индексации, поиска информации); поиска мультимедиа-ресурсов (агрегации и извлечения информации из html-содержимого). Реализованные алгоритмы позволяют повысить эффективность работы поисковых модулей, увеличив их быстродействие и улучшив качество выдаваемых результатов.

Создание модуля анализа поведения пользователя включало в себя разработку и программную реализацию алгоритма анализа http-заголовка “referer” браузера пользователя, а также алгоритмов сохранения информации о действиях пользователя. Внедрение модуля позволяет анализировать сетевую активность пользователей, фиксируя не только их действия в системе, но и частично – за её пределами. Полученная информация используется СКР для реализации поведенческих технологий при выборе объявлений.

Исследование внедренной системы показывает высокую релевантность выдаваемых пользователям объявлений, выбираемых в соответствии с контекстом просматриваемых страниц. Средний CTR объявлений в ходе тестирования системы превысил отметку 6,7%, что доказывает эффективность разработанных алгоритмов даже в условиях отсутствия большого объема накопленной статистики о показах объявлений и истории сетевой активности пользователей.

Основные результаты работы:

1. Разработан ряд новых алгоритмов работы системы контекстной рекламы:
 - алгоритм выбора рекламных объявлений по множеству факторов;
 - алгоритм выбора объявлений в соответствии с поисковым запросом пользователя;
 - алгоритм выбора объявлений по тегам страницы;
 - алгоритм выделения тегов из текстов страниц;
 - алгоритм выбора объявлений в соответствии с историей запросов пользователя;

- алгоритм выбора объявлений в соответствии с историей посещенных пользователем сайтов;
- алгоритм выбора объявлений в соответствии с их эффективностью;
- алгоритм оценки качества объявления.

2. Проведено тестирование эффективности основных разработанных алгоритмов, показавшее превосходство алгоритмов по сравнению с альтернативными стратегиями и высокую степень соответствия результатов работы предлагаемых алгоритмов оценкам экспертов. Тестирование разработанного прототипа СКР «Поисколог» показало высокую релевантность выдаваемых объявлений в сравнении с другими коммерческими СКР.

3. Разработана интегрированная система «Поисколог», включающая систему контекстной рекламы, поисковую систему и модуль анализа поведения пользователя. ИС реализует предложенные в работе алгоритмы выбора релевантных объявлений, а также ряд поисковых алгоритмов. Система «Поисколог» позволяет пользователям осуществлять различные виды поиска по региональным ресурсам и получать в качестве дополнительной релевантной информации рекламные объявления, а рекламодателям – создавать, настраивать и контролировать рекламные кампании и объявления.

4. Внедрение системы «Поисколог» показало высокую релевантность выдаваемых объявлений даже при отсутствии большого объема накопленной статистики о показах объявлений и длительной истории сетевой активности пользователей. Следовательно, достигается цель диссертационной работы – обеспечить высокую релевантность рекламных объявлений информационным потребностям пользователя.

Таким образом, по результатам выполненных теоретических и экспериментальных исследований разработано алгоритмическое и программное обеспечение системы контекстной рекламы, интегрированной с региональной поисковой системой. Решена задача, имеющая существенное значение для области информационного поиска в среде Интернет.

Публикации по теме диссертации:

1. Силич В.В. Методы организации системы поиска и индексации новостей в сети Интернет и её взаимодействие с системой контекстной рекламы / В.В. Силич // Вестник компьютерных и информационных технологий. – 2008. – №9. – С. 50-55.
2. Силич В.В. Система автоматизации поиска информации в сети Интернет / В.В. Силич // Приборы. – 2008. – №3. – С. 55-60.
3. Силич В.В. Поисковый Интернет-портал и система контекстной рекламы "Поисколог" / В.В. Силич // Компьютерные учебные программы и инновации – М: ГОСКООРЦЕНТР. – 2008. – №6. – С. 156.
4. Силич В.В. Модель выбора рекламных объявлений веб-сайтов в системе поисковой рекламы / В.В. Силич // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2007. - №2(16). – С. 229-235.

5. Силич В.В. Метод организации системы поисковой рекламы в сети Интернет / В.В. Силич // Известия Томского политехнического университета. – 2006. – №8. – С. 140-143.
6. Силич В.В. Организация системы поисковой рекламы в сети Интернет на основе нечетких множеств / В.В. Силич // Научная сессия ТУСУР – 2007: Мат-лы докладов Всерос. науч.-техн. конф. студентов, аспирантов и молодых ученых. – Томск: Изд-во «В-Спектр», 2007. – Ч. 1. – С. 332-335.
7. Силич В.В. Технология выборки оптимальных информационных блоков в системе поисковой рекламы в сети Интернет / В.В. Силич // Молодежь и современные информационные технологии: сб. трудов V Всерос. науч.-практ. конф. студентов, аспирантов и молодых ученых. – Томск: Изд-во ТПУ, 2007. – С. 260-261.
8. Силич В.В. Использование систем мета-поиска для нахождения информации в сети Интернет / В.В. Силич // Современная техника и технологии : сб. трудов XII междунар. науч.-практ. конф. студентов, аспирантов и молодых ученых / ТПУ. – Томск, 2006. – Т.2. – С. 159-162.
9. Силич В.В. Развитие технологий поиска в среде Интернет за счет применения метапоисковых машин / В.В. Силич, Ю.П. Ехлаков // Молодежь и современные информационные технологии : сб. трудов IV-ой Всерос. науч.-практ. конф. студентов, аспирантов и молодых ученых / ТПУ. – Томск, 2006. – С. 364-365.
10. Силич В.В. Технологии интерактивного взаимодействия веб-приложений в среде Интернет / В.В. Силич // Научная сессия ТУСУР-2006: Мат-лы докладов Всерос. науч.-техн. конф. аспирантов и молодых ученых. – Томск, 2006. – С. 35-37.
11. Силич В.В. Методы обработки запросов к веб-сайту / В.В. Силич // Научная сессия ТУСУР-2005: Мат-лы Всерос. науч.-техн. конф. аспирантов и молодых специалистов. – Томск, 2005. – С. 189-192.
12. Силич В.В. Технологии разделения данных и их оформления в web-программировании / В.В. Силич // Научная сессия ТУСУР – 2004: Мат-лы Всерос. науч.-техн. конф. – Томск, 2004. – С. 100-103.
13. Силич В.В. Использование единого механизма управления динамическим web-сайтом / В.В. Силич, В.Д. Автайкин // Средства и системы автоматизации: мат-лы 5-й науч.-практ. конф. – Томск, 2004. – с.123-124.
14. Силич В.В. Использование шаблонов при создании сайтов в сети Интернет / В.В. Силич, Ю.П. Ехлаков // Молодежь и современные информационные технологии: сб. тр. 1-й Всерос. науч.-практ. конф. – Томск, 2004. – С.164-165.
15. Силич В.В. Электронная торговля в России / В.В. Силич, М.В. Сподобаев // Энергия молодых – экономике России: тез. докл. 4-й Всерос. конф. студентов, аспирантов и молодых ученых. – Томск, 2003. – Т.1. – С. 165-166.