ВВЕДЕНИЕ

Накопленные к настоящему времени колоссальные объёмы информации в совокупности с непрерывно увеличивающимися темпами её роста определяют актуальность и значимость исследований в области информационного поиска. Бурное развитие сетевых технологий, в том числе и Интернета, способствуют значительному увеличению доступных информационных ресурсов и объёмов передаваемой информации. Зачастую это разнородная, слабо структурированная и избыточная информация, обладающая высокой динамикой обновления.

При сегодняшних объёмах доступной информации решение задач информационного поиска является приоритетным для обеспечения своевременного доступа к интересующим данным в рамках *информационной среды* (ИСр).

Концепция информационной среды впервые была предложена Ю.А. Шрейдером [83], который рассматривает информационную среду не только как проводника информации, но и как активное начало, воздействующее на её участников. *Информационная среда* – совокупность технических и программных средств хранения, обработки и передачи информации, а также социально-экономических и культурных условий реализации процессов информатизации.

В настоящее время работает ряд авторитетных международных конференций, посвящённых обсуждению вопросов информационного поиска [24], например, таких как:

- TREC (Text Retrieval Conference) – цикл конференций организованный под эгидой NIST (National Institute for Standards and Technology) – одного из авторитетных органов стандартизации информационных технологий в США [110,111];
- SIGIR (Special Interest Group on Information Retrieval) – цикл конференций проводимых ACM SIGIR (ACM – Association of Computing Machinery) –международной группой специалистов по информационному поиску;

- WWW (World Wide Web) Conference – специально организованная конференция для решения задач, связанных с Интернет [107, 111, 114, 115, 117].

Высокий авторитет конференций TREC, SIGIR, WWW и участие в них ведущих исследовательских коллективов и разработчиков технологий информационного поиска во многом определяет приоритетные направления исследований и задает общие принципы развития поисковых систем.

Из отечественных конференций, посвященных вопросам информационного поиска, нужно отметить ежегодную всероссийскую конференцию «Электронные библиотеки» (RCDL) и семинар по компьютерной лингвистике «Диалог».

Также необходимо отметить ряд отечественных научных школ.

- SPBU IR Group – исследовательская группа в области информационного поиска (Санкт-Петербургский Государственный Университет).
- Исследовательский центр ИИ ИПС РАН.
- Центр информационных исследований (НИВЦ МГУ).

Кроме того, существуют коммерческие организации, занимающиеся не только вопросами исследований, но и вопросами внедрения информационных технологий. Это такие известные организации как Яндекс, Рамблер, Апорт, НейрОК, Гарант-Парк-Интернет, Галактика-Зум, ABBYY-FTR, АОТ и др.

Ряд авторитетных исследователей внесли своими научными трудами значительный вклад в развитие информационно-поисковых систем: И.С. Некрестьянов, И.Е. Кураленок, В.Ю. Добрынин, А.Г. Дубинский, А.Е. Ермаков, М.Р. Когаловский, А.В. Сокирко, G. Salton, A. Singhal, M. Mitra, S. Lawrence, P. Foltz, E. Fox, J. Cho, R. Baeza-Yates, K. Tajima, C. Van Rijsbergen, L. Gravano, J. Kleinberg, J. Sparck, D. Carmel, S. Brin, L. Page, A. Singhal., T. Haveliwala.

Существует широкий спектр предлагаемых решений и перспективных направлений исследований в области информационного поиска, начиная от построения глобальных распределенных информационных структур и поисковых систем, заканчивая элементарными на первый взгляд вопросами анализа документов при помощи латентно семантического анализа [94, 96, 97]. Все они, без-

условно, важны и полезны при решении своих специфических задач. Тем не менее, именно от методов анализа во многом зависит эффективность существующих поисковых систем, так как эти методы являются основой любой поисковой системы и во многом определяют возможности и ограничения таких систем [89].

Современные информационно-поисковые системы, в основе которых по большей степени лежит полнотекстовый поиск, позволили добиться высокой степени классической *релевантности* – соответствия запроса пользователя результатам выдачи поиска. Однако качество информационного поиска характеризуется не только релевантностью, но и *пертинентностью* – соответствием результатов поиска информационной потребности пользователя. Результаты работы поисковой системы часто не удовлетворяют требованиям высокой пертинентности. Это связано с такими свойствами естественного языка как синонимия, полисемия, омонимия и другие [99, 100, 102, 104, 88].

Представленные на сегодняшний день в большинстве популярных поисковых систем способы организации полнотекстового поиска не учитывают в достаточной мере семантику. В то же время, именно *семантическое сходство* непосредственно обусловливает высокую степень пертинентности. Далеко не всегда пользователь информационно-поисковой системы может чётко и однозначно сформулировать именно тот набор ключевых слов, который и приведет его к искомому результату. Зачастую низкая пертинентность обусловлена сложностью формирования информационных запросов для полнотекстового поиска. Эти сложности вызваны следующими причинами:

- незнанием набора ключевых слов, однозначно определяющих семантику искомых документов;
- отсутствием достаточного опыта и квалификации формирования поисковых запросов;
- отсутствием принятой и устоявшейся терминологии в интересующей предметной области.

Нередко человек, осуществляющий поиск, имеет самое приблизительное представление об интересующей его тематике. Все это обусловливает актуальность и значимость исследований, направленных на решение одной из ключевых проблем информаци-

онного поиска – проблемы адекватного отображения информационных потребностей пользователей, и, как следствие, повышения пертинентности поиска.

Одним из вариантов решения проблемы низкой пертинентности в настоящее время является динамично развивающаяся технология Semantic Web [21, 22]. В основе актуальности этой технологии лежит уже осознанная человечеством необходимость представления информационных ресурсов не просто как единиц хранения информации, но как *носителей знаний*. То есть документы, отчёты, статьи, банки данных интересуют специалиста, главным образом, своей семантической составляющей. Семантический подход к глобальным информационным ресурсам, предполагающий не только их индексацию, но и сопровождение семантическим описанием, было предложено реализовать в технологии «Semantic Web», разработанную W3C-консорциумом, занимающимся разработкой и внедрением Web-технологий. В наиболее завершенной форме требования к описанию и стандартизация описания знаний в этой технологии были предложены в 2004 г.

Как правило, авторы программных средств, предназначенных для формального описания знаний в Semantic Web, не претендуют на завершенность своей разработки и отсутствие возможности создания более эффективных теоретических концепций и версий программных систем этого назначения. В то же время нужно согласиться с тем, что принятие единого стандарта в описании знаний – важнейший фактор реальной работы Semantic Web. Исходя из этого, новые формализмы представления знаний целесообразно разрабатывать на принципах совместимости с существующими средствами, такими, как Resource Description Framework (RDF) и Web Ontology Language(OWL DL). RDF – это разработанная консорциумом Всемирной паутины модель для представления данных, OWL – язык описания онтологий для семантической паутины.

Целью данной книги является разработка и исследование способа повышения показателей пертинентности информационного поиска, основанного на концепции интерфейсной поисковой системы (ИнтПС), осуществляющей объединение и переупорядочивание откликов на запросы пользователей популярных поисковых систем сети Интернет. Для достижения поставленной цели решаются следующие задачи:

- формализация описаний факторов ранжирования поисковых систем;
- модификация существующих факторов ранжирования, слабо защищенных от искусственного влияния структуры информационной среды;
- создание методологии оценки пертинентности информационного поиска на основе экспертных оценок;
- разработка концепции поисковой системы многоальтернативного поиска и адаптивного переранжирования.

Совокупность полученных теоретических и практических результатов может использоваться для построения метапоисковых и интерфейсных информационно-поисковых систем, позволяющих повысить эффективность информационной поддержки профессиональной целенаправленной деятельности сотрудников малых и средних предприятий и организаций, для которых гипотеза о тематической однородности запросов наиболее правдоподобна.

Для практического воплощения концепции интерфейсной поисковой системы созданы два программных продукта (AltoSearch/АльтПоиск и SearchAnalizer/ПоискАнализатор), позволяющие автоматически формировать общую выдачу – обобщённый набор документов, получаемых от нескольких поисковых систем сети Интернет в ответ на запрос пользователя и расчёт показателей контентной эквивалентности. Создан макет интерфейсной поисковой системы, опытная эксплуатация которого в рабочем процессе ООО «Мегапром» показала повышение подекадного среднего значения подлинной пертинентности на 10 – 18 % по сравнению с популярными поисковыми системами. Разработанные программные продукты имеют свидетельства об официальной регистрации программных систем и баз данных в Российском агентстве по патентам и товарным знакам (РОСПАТЕНТ) [73, 74]

В первой главе определяются основные цели и задачи разработки информационно-поисковых систем, проблемы современных поисковых систем, приводятся основные направления исследований данной проблемы и обзор работ. Описаны принципы работы документальных поисковых систем, интеллектуальных поисковых систем. Более подробно рассмотрены компоненты поисковых систем сети Интернет.

Вторая глава содержит классификацию поисковых запросов сети Интернет. Описаны основные факторы ранжирования поисковых систем, влияющих на позиции документов в выдаче в ответ на запрос пользователя. Определены методика оценки пертинентности на основе экспертных оценок, понятие информационной единицы, формализованы показатели контентной эквивалентности, применяемые для оценки пертинентности информационного поиска на основе соображений экспертов о предметной области.

В третьей главе дано определение показателя авторитетности страницы PageRank, описаны различные методы вычисления PageRank, их особенности и недостатки. Предложен собственный функциональный способ расчёта PageRank на основе системы линейных алгебраических уравнений. Разработан Solid PageRank, семантически представляющий собой нижнюю оценку авторитетности страницы, позволяющий избежать искусственной накрутки собственного значения за счёт организации топологии фрагмента сети Интернет. Реализован необходимый инструментарий для вычисления значений показателей авторитетности.

В четвертой главе сформулирована концепция интерфейсной поисковой системы, реализующая в себе элементы персонифицированного поиска от Google, социального поиска на основе социальных закладок. Описана концепция многоальтернативного поиска и последующего адаптивного переранжирования при помощи прогнозирования оценок пертинентности. Разработаны теоретические основы для работы с временными рядами оценок пертинентности. Предложен рекурсивный алгоритм определения структуры произвольного фрагмента сети Интернет.

В заключении приводится обобщение основных результатов.