

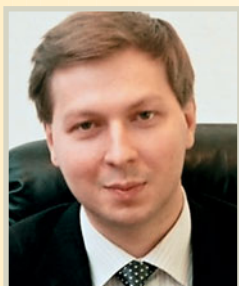
# Гости редакции PC Magazine/RE

Поиск информации — одна из главных проблем эпохи информационных технологий. Постоянно растущие массивы данных, непрерывно пополняемый Интернет, «залежи» документов на жестких дисках рабочих машин... Все чаще встает проблема поиска нетекстовой информации, такой, как видео или содержимое аудиозаписей. Мы попросили разработчиков поисковых технологий прокомментировать сегодняшнее состояние и перспективы развития этого сегмента.

**1** Что бы вы сочли прорывом в технологиях поиска текстовой информации и насколько он возможен?

**2** Какими вы видите перспективы развития поисковых средств для нетекстовых данных — аудио, видео, других мультимедиа-объектов?

**3** Насколько востребованы на рынке развитые технологии поиска, с поддержкой морфологии, лингвистическим анализом и т. д.? Ведь известна статистика, что разделами «расширенный поиск» в поисковых системах пользуется исчезающе малая доля посетителей.



ДМИТРИЙ ГРИШИН

ГЕНЕРАЛЬНЫЙ ДИРЕКТОР  
MAIL.RU  
WWW.MAIL.RU

**1** Прорывом я бы счел создание поисковой машины, способной отвечать, именно отвечать, на вопросы пользователя, а не выдавать ему большое количество ссылок с ключевыми словами.

**2** Делать прогнозы дело неблагоприятное, но, думаю, никаких существенных подвижек здесь в ближайшее время не произойдет. Будут слабые попытки анализа визуальной части (цветовой гаммы, частоты смены кадров и т. п.), но основной упор будет сделан на использование текстовой информации, связанной с мультимедийным объектом, или технической текстовой информации, закодированной в самом объекте (например, поиск Google по видеоконтенту основан именно на этом принципе).

**3** Развитая технология поиска — это не та, которая имеет много настроек и «чек-боксов», а, напротив, та, где от пользователя требуется минимум действий и специальных знаний для получения ответа на свой вопрос. Идеалом, как я уже сказал, станет поисковая машина, которой можно задать вопрос на естественном языке, а она будет выдавать конкретный ответ, а не 10 страниц ссылок на один и тот же анекдот с этой фразой.

Вообще, поисковый сервис — это самая востребованная после коммуникационных сервисов функция Интернета, и с ростом объема информации, содержащейся в сети, с одной стороны, и увеличением количества технически неграмотных пользователей — с другой — необходимость появления качественно новых поисковых технологий только возрастает.



СТАНИСЛАВ ЗАДОРЖНЫЙ

РУКОВОДИТЕЛЬ ГРУППЫ  
ПРОЕКТИРОВАНИЯ И АНАЛИТИКИ  
«МЕДИАЛОГИЯ»  
WWW.MEDIALOGIA.RU

**1** Прорывом можно было бы считать получение от поисковой системы только тех результатов, которые нужны конкретному пользователю в данный конкретный момент. В общем-то эта суперцель всегда стоит перед разработчиками, но пока не достигнута. Если посмотреть поисковые технологии с практических позиций, то можно было бы считать прорывом (хотя я бы скорее это называл эволюцией) массовое промышленное использование актуальных идей из области теории информационного поиска, компьютерной или прикладной лингвистики, прикладной статистики, социологии и психологии. И эти процессы идут; научные идеи и разработки, которые не могли быть использованы в массовых решениях (неважно, по каким причинам — технологическим, из-за отсутствия массового спроса или же данных для алгоритмов), сейчас активно внедряются в поисковые и близкие к ним системы. В качестве примера можно назвать такие задачи, как выделение сущностей, автоматическая классификация, анализ предпочтений пользователей и использование их поведения в качестве обратной связи для моделей информационного поиска. Да и собственно новые модели информационного поиска (невекторные или теоретико-множественные). Особенно такая эволюция заметна в поисковых системах, ориентированных на крупный корпоративный рынок.

**2** Перспективность технологий мультимедиа-поиска связана в основном со сложностью формулировки поисковой задачи, особенно если мы гово-

рим о массовых применениях. Как таковые же специализированные системы поиска нетекстовой информации прекрасно развиваются. Я имею в виду прежде всего обработку графических данных, например дактилоскопической информации или «фотороботов». Но на массовом рынке (читай: в Интернете) более вероятное направление развития — использование текстовой информации, которая так или иначе связана с мультимедийной. Но еще раз скажу — главная проблема заключается в отсутствии массового спроса на реально решаемые задачи и сложности формулировки пользовательской потребности.

3 Самое неприятное для Интернет-поисковиков — привычка большинства пользователей формулировать запросы из двух-трех слов, да к тому же часто на основании своего, понятного им и совершенно неизвестного поисковой машине, контекста. Любая конкурентоспособная поисковая система должна обеспечивать некоторый лингвистический анализ, но дело не в том, что пользователь напрямую должен задействовать эти средства. Применение лингвистических технологий на уровне базовых алгоритмов поисковой системы позволяет повышать качество поиска или предоставлять дополнительные услуги и информационные продукты.



ЛЕВ МАТВЕЕВ

ДИРЕКТОР «СОФТИНФОРМ»  
WWW.SOFTINFORM.COM

1 Обычный фразовый поиск, реализованный во всех существующих системах (как настольных и корпоративных, так и в Интернете), справляется со своей основной задачей, но далеко не идеально. Сказываются временные затраты на подбор ключевых слов и просмотр ненужных документов, полученных в списке результатов по не слишком корректному поисковому запросу. Мы же ориентируемся на технологии поиска документов, похожих по содержанию. Этот метод позволяет существенно сократить время поиска нужной информации и дает хорошие результаты в реальных задачах.

Главное его преимущество — сокращение времени при индексировании информационных массивов, с одной стороны, и трудозатрат пользователя, особенно при сложном поиске, — с другой. Скажем, в традиционных системах для поиска документов, содержащих информацию о каких-либо

событиях, вероятнее всего, придется сделать несколько «проходов», последовательно уточняя критерии поиска. В случае же «поиска похожих», автоматически сформированного поисковой машиной, можно сразу выбрать интересующий документ и найти «похожие по содержанию». На мой взгляд, это и есть определенный прорыв в области текстового поиска.

2 Интересный вопрос. Тут надо для начала понять, по каким критериям мы хотим искать. Если это, скажем, видео, то следует понимать, что ищем не кадры-картинки, а, к примеру, фильм с каким-то сюжетом. Если мы берем музыку, то тут ведь тоже важны не только слова, но и мелодии. Это необходимо как-то формализовать; на первый взгляд перспективной представляется такая схема: распознавание речи (слова в песнях, тексты в фильмах) с последующим индексированием полученных текстов на базе существующих технологий или более или менее интеллектуальное разбиение музыки на фрагменты с присвоением каждому определенного веса в зависимости от ритма и т. д. (с видео — аналогично). И на основании этого критерия формировать некий интегральный коэффициент схожести текстов и содержимого мультимедиа, выдавая пользователю наиболее релевантные с точки зрения этого критерия результаты. Хотя пока это, безусловно, фантастика.

3 Должен отметить, что поиском в Интернете мир поисковых технологий отнюдь не исчерпывается. Причина низкой популярности всех этих расширенных режимов очевидна, они не позволяют добиться радикального увеличения точности поиска, не обеспечивают хорошей фильтрации «информационного шума». Именно поэтому пользователь тудя и не идет. Но здесь следует помнить, что действительно эффективного поиска можно добиться только при многоуровневой организации поисковой системы. Формирование списка документов, сходных по какому-либо критерию, — это первый уровень, а при дальнейшей работе с ними наличие морфологии и лингвистики, безусловно, позволит повысить качество.

Но безусловно, поисковые технологии востребованы. У нас есть интересные проекты, реализованные в ряде корпоративных систем. Например, система, созданная по заказу компании Alpha Lowyers. Эта компания предоставляет услуги юридического консалтинга по телефону, причем операторы должны максимально оперативно извлекать из общего информационного массива документы, соответствующие вопросу. Клиент оплачивает время разговора, поэтому, естественно, скорость и качество поиска здесь приобретают особое значение, в отличие от поисковых машин для Интернета, где требования к качеству и точности поиска гораздо мягче.



Илья СЕГАЛОВИЧ

ДИРЕКТОР ПО ТЕХНОЛОГИЯМ  
«ЯНДЕКС»  
WWW.YANDEX.RU

**1** Будущий интерфейс поиска, равно как и современный, — это строка запроса из нескольких слов. Уже сегодня практически на любой запрос в Web представлено много информации высокого качества. Но поисковым системам есть куда улучшаться. Вектор движения известен, и известны технологии, развитие которых должно здесь помочь.

В первую очередь это технологии, «воспринимающие» и пользователя, и документы как элементы социальной сети. При таком подходе запрос трактуется в контексте данных, полученных с помощью анализа истории поисковой деятельности всех других пользователей, когда-либо задававших этот запрос, с учетом географии, местоположения запрашивающего; неотрывно от истории его собственного поискового поведения. И так далее.

Документы и сайты в социальной сети также подчиняются ее законам: они имеют владельцев, связываются ссылками, копируются, составляются из фрагментов и распадаются на них. Они содержат упоминания людей и фактов, географических объектов и потребительских продуктов. Извлечение этих фактов помогает при анализе сети во время подготовки внутренних информационных структур поисковой системы. Кстати, фраза про «технологии поиска текстовой информации» выдает отчаянно устаревший и неполный взгляд на Web-математику. Поиск — это давным-давно не функция «идеального сопоставления текста и запроса», поиск — это многогранная и сложная обработка всей совокупности Web-данных. Что касается «прорыва»: одна революция недавно уже произошла, когда к тексту перестали относиться только как к тексту. Про новые революции пока ничего неизвестно.

**2** Сегодня пользователи с успехом находят практически любые мультимедийные объекты, пользуясь «традиционным» способом составления текстовых запросов. Достаточно назвать наш сервис поиска изображений, <http://images.yandex.ru>, с месячной аудиторией под 3 млн. пользователей. При этом, действительно, многие ресурсоемкие и несовершенные технологии (например, распознавание образов и речи), которые связаны не с обработкой текстовых аннотаций, а с непосредственно цифровым содержанием мультимедийных файлов, пока не нашли широкого применения и имеют скорее

экспериментальный статус. Однако мне кажется, что первый по-настоящему массовый опыт применения этих способов не за горами.

**3** Следует делать различия между «развитыми технологиями» поиска и «развитыми пользовательскими инструментами по управлению поиском», такими, например, как «расширенный поиск» или «язык поисковых запросов». Это совершенно разные вещи, и степень развитости одного никак не связана со степенью развитости другого.



Олег СЕРЕБРЕННИКОВ

ПРЕЗИДЕНТ «МЕДИАЛИНГВА»  
WWW.MEDIALINGUA.RU

**1** Думаю, что речь может идти о создании технологий, реализующих некоторые механизмы человеческого сознания, и прежде всего технологий, реализующих механизмы ассоциативной памяти. Весьма интересной и уместной для упоминания здесь мне представляется книга «Психозекология» И. В. Смирнова, посвященная исследованию сознания человека и механизмов его работы. Поиск текстовой информации, это чаще поиск информации, соответствующей жизненному опыту конкретного пользователя в контексте его ситуации. Поэтому разработка новых технологий в этой области, по моему мнению, будет симбиозом компьютерной науки и психологии. Вместе с тем для учета «жизненного опыта» пользователя, очевидно, придется сосредоточиться на анализе истории его работы с компьютерной системой. Мы занимаемся этим вопросом, «МедиаЛингва» рассматривает создание ассоциативного хранилища текстовых данных как ближайшую перспективу развития собственных технологий поиска семейства «Следопыт».

**2** Вопросами поиска нетекстовой информации давно и небезуспешно занимаются многие компании. Я имею в виду «истинную» технологию работы с медиа-данными, а не ту, что использует для поиска текстовые деривативы, такие, как надписи под изображениями, и другую, связанную с ними, текстовую метаинформацию. Это вполне реальная задача, еще в середине 1990-х мы разработали технологию распознавания стереоизображения лица и голоса человека, но интерес к биометрическим технологиям начал расти лишь в последние годы. Думаю, что в значительной степени двигателем развития технологий поиска медиа-данных и сейчас

является интерес служб безопасности и правительств к вопросам идентификации. В то же время если не спрос, то интерес конечных пользователей к таким системам также увеличивается. В частности, лицензия на технологию биометрической идентификации «МедиаЛингвы» в 2004 г. была приобретена одним из крупнейших мировых производителей сложного бытового и электронного оборудования. Некоторые технологии биометрической идентификации уже появились в компьютерах и автомобилях, на рынке устройств связи и рынках другого оборудования. Вопрос об адекватности предлагаемых технологий поиска медиа-данных задачам, которые они решают, сложный, и его лучше оставить рынку. Однако ясно, что существующие технологии пока не воспроизводят способностей даже простейших организмов.

③ Разумеется, лингвистический анализ нужен. Им часто «не пользуются» именно потому, что поиск производится в массиве данных, предварительно обработанных с помощью таких технологий еще на этапе индексации. Лингвистический анализ текстовой информации, в частности, необходим для снижения объема хранимых данных и управления «шумом» при их поиске. Лингвистический анализ в этом аналогичен распознаванию медиа-данных в смысле их предобработки или «нормализации». Это позволяет упростить и ускорить технологию поиска «распознанных» форм. В идеале, наверно, можно представить себе универсальную технологию хранения «нормализованных» данных и их поиска, которая использует специализированные препроцессоры исходных данных для их «нормализации». Подытоживая, я бы сказал, что потребность в технологиях анализа и препроцессинга вообще не вызывает сомнений и дело лишь в степени их необходимого применения в разных случаях.



КОНСТАНТИН ЧУБИНИДЗЕ

ДИРЕКТОР ПО НАУКЕ И РАЗВИТИЮ  
CONVERA  
WWW.CONVERA.RU

① Безусловно, прорыв возможен, но революций не будет. Скорее следует говорить о эволюции, в результате которой поисковые технологии выйдут на качественно новый уровень. Направлений развития на сегодня несколько. Прежде всего, это реализация семантического поиска, увеличение его точности и производительности, в основном благодаря развитию и накоплению лингвистических

ресурсов. При этом произойдет переход от большого общелексического словаря к набору согласованных словарей по достаточно узким проблемным областям. В таких семантических словарях ожидается доминирование многословных лексических единиц, а не отдельных слов. Будут реализованы системы распознавания и идентификации сущностей, интеллектуальный анализ запроса с интерфейсом, который «вынудит» пользователя произвести необходимые уточнения для снятия полисемии и определения корректного набора слов-расширений. И конечно, продолжится слияние функциональных возможностей поисковых механизмов с аналитическими системами, системами поддержки коммерции и т. д.

② Действительно, мало кто имеет дело с реальными решениями в этой области. Это вполне объяснимо, особенно для тех, кто сталкивался с проектированием подобных систем. В распоряжении нашей компании есть инструменты для проектирования такого рода решений, например Visual Retrieval Ware (средство разработки систем поиска изображений) и Screening Room (для индексации и поиска видео). Главная причина, по которой такие разработки не получили большого распространения, заключается в специфике самой проблемы индексации и поиска такого типа данных. Здесь требуется, например, «нормализация» поисковых образов, т. е. они должны быть по крайней мере приведены к одному разрешению и размеру. Другое, не менее важное, но сложновыполнимое условие — предварительная обработка изображений. В общем, проблем здесь довольно много. Такие системы развиваются и будут развиваться, но в обозримом будущем мы не ожидаем их популяризации.

③ Да, это так. Мы обнаружили даже более интересную закономерность. Большинство опытных экспертов в области поиска текстовой информации тоже не используют «расширенный» с точки зрения лингвистики поиск. Они составляют весьма громоздкие логические запросы. А обычные пользователи, которые не используют раздел «расширенный поиск» и, несмотря на это, находят необходимую информацию, достигают успеха именно благодаря лингвистической обработке текстов.

Лингвистический анализ в поисковых системах должен облегчить, а не усложнить работу пользователей. Он необходим для того, чтобы система могла автоматически определить семантические значения, которые присутствуют в тексте документа и в запросе пользователя, для корректного их сопоставления при поиске.

А споры о целесообразности таких решений, на мой взгляд, объясняются несовершенством лингвистических алгоритмов. Эту задачу необходимо решать, и наша компания активно работает в этой области.