Санкт-Петербургский Государственный Политехнический Университет
Институт прикладной математики и механики
Кафедра прикладной математики

Диссертация допущена к защите
Зав. кафедрой

_____
"___"_____

ДИССЕРТАЦИЯ
на соискание степени МАГИСТРА

Тема: ***метод ранжирования разнородных результатов поиска***

Направление: 01.04.02 - Прикладная математика и информатика
Магистерская программа: системное программирование

Выполнил студент гр. 63601/2 _____ Толмачев А.С.

Руководитель, к.ф.-м.н., доц. _____ Иванков А.А.

Консультанты:

по вопросам информационного поиска, к.ф.-м.н. _____ Кураленок И.Е.

по вопросам охраны труда, к.т.н., доц. _____ Монашков В.В.

Санкт-Петербург
2015

Содержание

Введение

С развитием информационных систем и ростом их популярности растет и количество информации, производимой с их помощью. Так, по данным аналитической компании IDC (International Data Corporation) общий объем цифровой информации в мире составил на 2013 год примерно 4.4 зеттабайт[1], он увеличивается каждый год примерно на 40% и к 2020 году составит приблизительно 44 зеттабайт [1]. Существенная доля этой информации – информация, размещенная во всемирной сети Интернет. Она большей частью неструктурирована и очень разнообразна. Несомненно, без помощи поисковых систем ориентироваться в этом огромном информационном пространстве не представляется возможным. Поэтому системы веб-поиска на сегодняшний день играют очень важную роль в нашей жизни.

С момента своего возникновения веб-поисковые системы активно развиваются. Одно из современных направлений их развития касается смешивания в результатах поиска разнотипной информации. Первые системы поиска в интернете в ответ на запрос выдавали список ссылок на веб-страницы (рис. 1). Такой вид результатов поиска для того времени был естественным, поскольку представляемая в них информация была достаточно однородной. Но по мере того, как развивались интернет-технологии и увеличивалось число интернет-пользователей, информация, размещаемая в сети, становилось все более разнообразной. На сегодняшний день это разнообразие огромно: в интернете можно найти тексты книг, музыку, фильмы, новости, научные статьи, программные приложения, кулинарные рецепты, технические характеристики товаров и отзывы о них и т. д. – все это различные типы информации. В связи с этим получили развитие системы, предназначенные для агрегации и поиска информации определенного типа. К таковым относятся, например, системы поиска изображений, видео-записей,

[1] 1 зеттабайт (ЗБ) = 1 триллион гигабайт

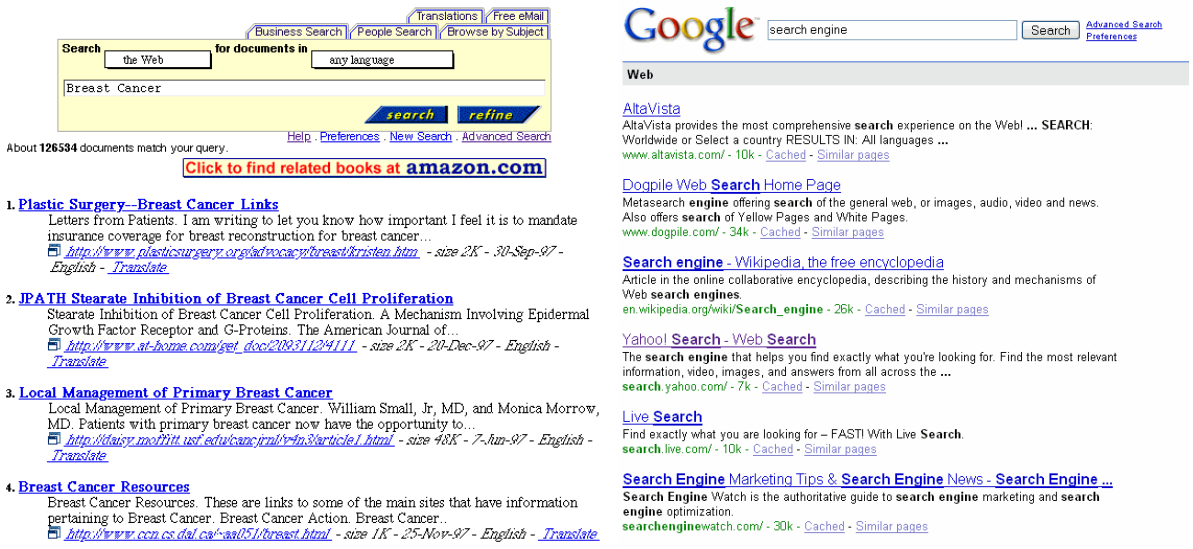

Рис. 1: Страницы результатов поиска системы AltaVista и одной из первых версий системы Google.

новостей, товаров, музыки. Ясно, что такая специализированная система может быть более удобной и полезной для решения поисковой задачи из соответствующей области, чем система общего назначения. Действительно, если пользователь, к примеру, ищет фотографии Дворцовой площади, то, очевидно, ему будет гораздо удобнее, если в качестве результатов поиска он будет видеть именно фотографии, а не ссылки на веб-страницы – ему не нужно будет переходить по этим ссылкам и самостоятельно исследовать страницы в поисках фотографий. Но выбирать каждый раз, к какой из многочисленных специализированных систем обратиться, неудобно. К тому же пользователь может не знать о существовании той или иной специализированной системы, а для каких-то поисковых задач такой системы может и не быть. Поэтому возникает естественное желание, чтобы система веб-поиска сама “понимала” запрос пользователя, и выдавала в ответ информацию нужного типа. Однако ограничивать ответ на запрос каким-то одним типом информации также неоправданно – для решения своей поисковой задачи пользователю может быть полезна информация сразу нескольких типов. Так, например, при поиске информации о музыкальном исполнителе может быть полезна и его биография, и фотографии с ним, и аудио-записи исполняемой им музыки, и видео-сюжеты о нем. Или же поисковый запрос может быть многозначным – например, по запросу “политика” пользователя может интересовать как информация, касающаяся самого термина, так и политические новости. Стандартным на сегодняшний день решением является смешивание результатов поиска от разных специализированных си-

стем и представление их вместе с традиционными результатами веб-поиска – списком ссылок на интернет-страницы. Такое смешивание мы можем наблюдать, пользуясь современными поисковыми системами. Например, в результатах поиска систем Яндекс и Google можно увидеть разнообразные специализированные результаты: на поисковый запрос о картинах – результаты поиска по изображениям, на запрос об адресе в городе – интерактивную карту с отмеченным адресом, на запрос о новостях – результаты поиска по новостям, а на запрос о кафе – специализированный ответ с найденными заведениями и информацией о них (рис. 2). Таким образом, результаты поиска могут быть *разнородными*, поскольку могут содержать информацию разных типов.

Процесс обработки поискового запроса можно условно разделить на два этапа: поиск информации, соответствующей запросу, и представление найденной информации. На первом этапе из всего множества объектов, известных системе, выбираются те, которые по тем или иным критериям соответствуют заданному запросу. На втором

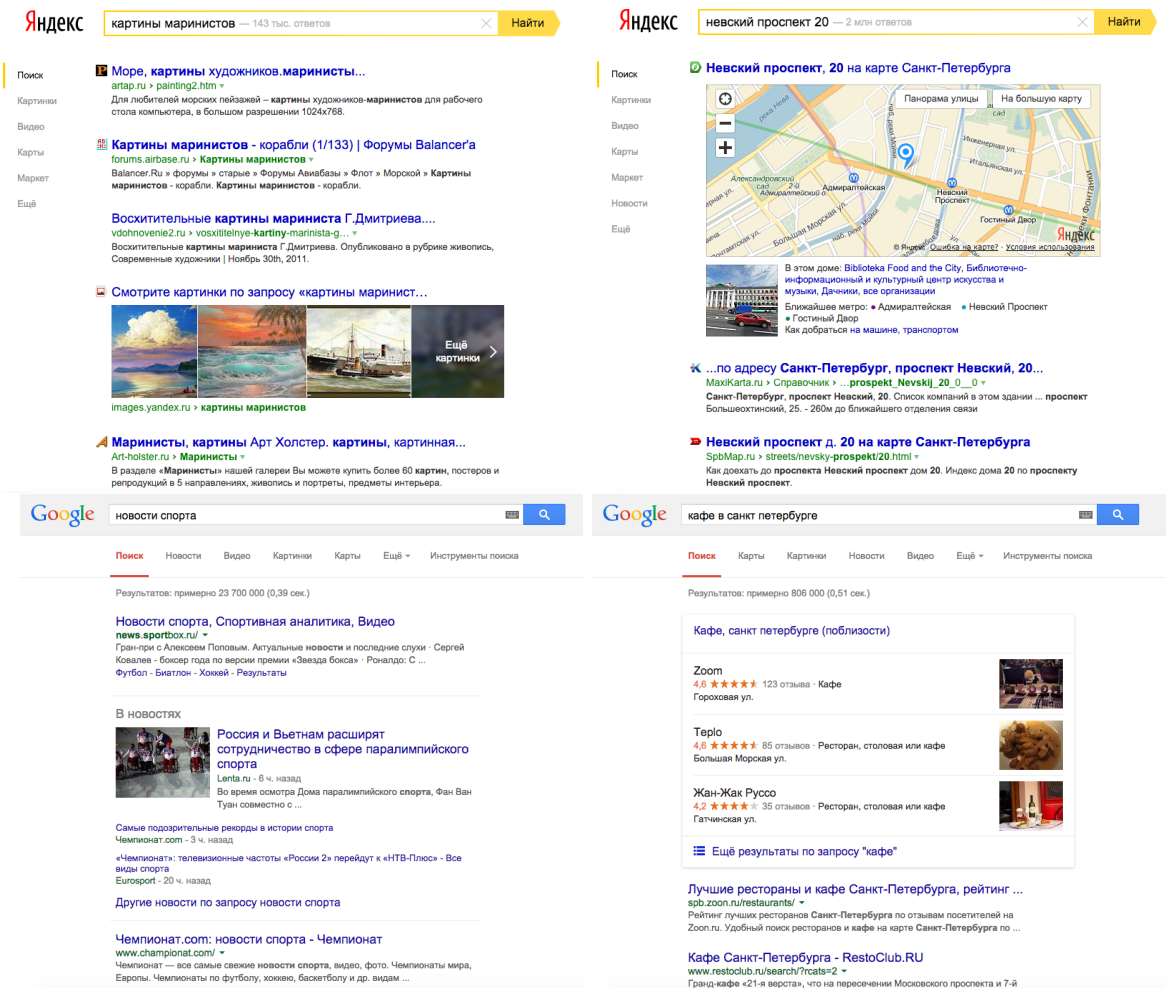

Рис. 2: Специализированные ответы в результатах поиска систем Яндекс и Google.

этапе множество найденных объектов представляется некоторым образом и выдается пользователю. Задача *ранжирования* относится ко второму из этих этапов. Ранжирование – это упорядочение результатов поиска в соответствии с некоторым принципом [4, 6]. От того, как упорядочены результаты, во многом зависит то, насколько успешно пользователь сможет решить свою поисковую задачу. Так, например, если пользователь задал запрос “скалолазание википедия”, то, вероятно, он ищет статью о скалолазании из интернет-энциклопедии Википедия. Если среди найденных результатов эта статья присутствует, то разумно расположить ее первой в списке результатов поиска, чтобы пользователь смог сразу ею воспользоваться. В противном случае ему будет сложнее найти этот результат среди остальных, а если расположить его за пределами первых десяти результатов, то он и вовсе может решить, что эта статья не была найдена.

(TODO: Что-то еще о ранжировании?)

Задача ранжирования, возникающая при смешивании результатов поиска от разных специализированных систем, отличается от классической задачи ранжирования (TODO: ссылка). Во-первых, ... следующими особенностями:

Возникающая при смешивании ... задача ранжирования отличается

Классическая задача ранжирования формулируется для однотипных объектов. (TODO: О классической задаче ранжирования и особенностях ранжирования разнородных результатов)

———

Также следует отметить, что список, – не единственный способ представления результатов поиска. Модели поисковой выдачи могут быть различными. Например, поисковая выдача системы Google для настольных компьютеров имеет две колонки, в левой из которых располагается список результатов, а в правой могут располагаться специализированные ответы (рис. 3). А выдача мобильного приложения поисковой системы Яндекс состоит из страниц, каждая из которых может содержать список результатов или специализированные ответы. Набор этих страниц и их порядок зависит от запроса (рис. 4). В таком случае задача ранжирования усложняется и превращается в задачу расположения результатов поиска в соответствии с заданной моделью поисковой выдачи. Это также требует обобщения методов ранжирования.

В данной работе рассматривается задача ранжирования разнородных результатов поиска и предлагается универсальный метод ее решения. ...

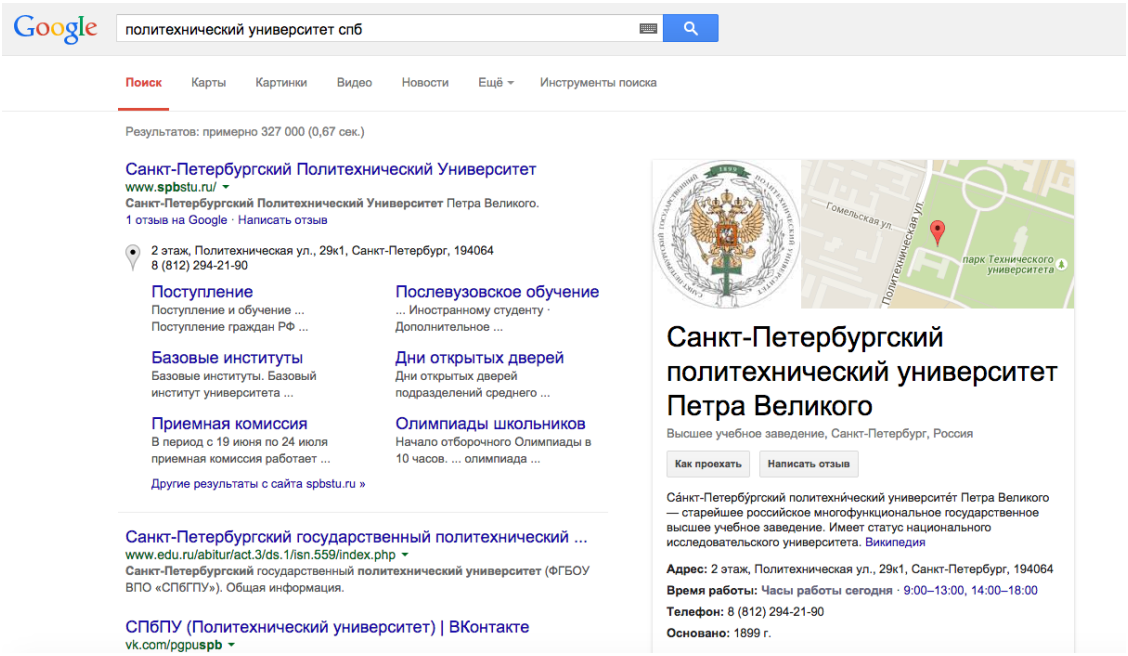

Рис. 3: Поисковая выдача системы Google со специализированным результатом в отдельной колонке.

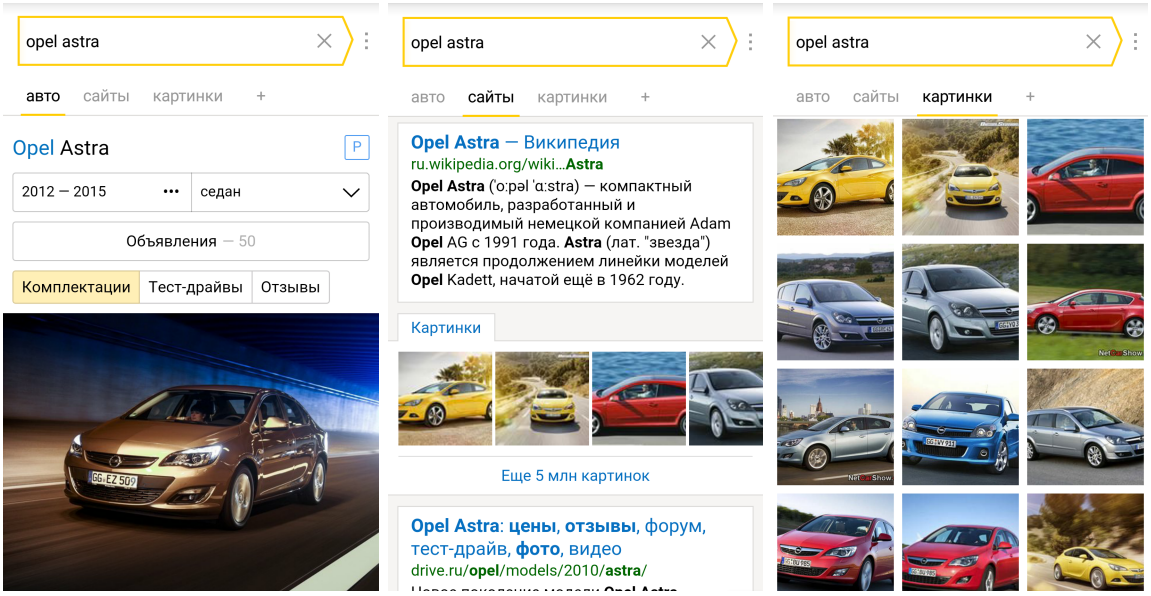

Рис. 4: Выдача мобильного приложения поисковой системы Яндекс с результатами поиска на отдельных страницах.

Глава 1

Обзор литературы

План:

- Традиционная задача ранжирования, обзор методов, способов оценки
- Ранжирование разнородных результатов, обзор методов, способов оценки

Наиболее ранние исследования в области ранжирования разнородных результатов поиска касаются встраивания одного конкретного специализированного результата на первое место в списке результатов поиска (TODO: ссылки) и встраивания одного из нескольких специализированных результатов так же на первое место (TODO: ссылки). Однако встраивание только одного специализированного результата на самую верхнюю позицию подходит лишь для тех случаев, когда поисковый запрос выражено относится к какой-то вертикали, и рассматриваемый специализированный результат более релевантен, чем все остальные. Однако специализированный результат может быть более или менее релевантен по сравнению с другими результатами, а также для запроса могут быть уместны одновременно несколько специализированных результатов. Поэтому более поздние исследования нацелены на встраивание специализированных результатов на различные позиции в поисковой выдаче (TODO: ссылки). (TODO: Дописать еще)

1.1 Классическая задача ранжирования

1.1.1 Формулировка задачи

1.1.2 Обзор методов решения

1.2 Задача ранжирования разнородных результатов поиска

1.2.1 Формулировка задачи

1.2.2 Обзор методов решения

Глава 2

Описание метода

2.1 Основные идеи

2.2 Статистический критерий полезности поисковой выдачи

2.3 Формальная постановка задачи

2.4 Модель оценки полезности поисковой выдачи

2.5 Алгоритм ранжирования

2.5.1 Базовый алгоритм

2.5.2 “Жадный” вариант алгоритма

Глава 3

Программная реализация

3.1 Схема системы ранжирования

3.2 Уменьшение числа обращений к поисковым источникам

3.3 Используемые технологии

Глава 4

Оценка качества работы метода

4.1 Методы оценки качества поиска

4.1.1 Методы, основанные на экспертных оценках

4.1.2 Методы, основанные на поведении пользователей

4.2 Выбор данных

4.3 Описание результатов

Глава 5

Вопросы охраны труда

5.1 Общая характеристика санитарно-гигиенических условий труда

5.2 Эргономические требования

5.3 Микроклиматические условия

5.4 Уровень шума

5.5 Системы освещения

5.6 Излучения

5.7 Электробезопасность

5.8 Инженерно-технические мероприятия по созданию благоприятных условий труда

5.9 Методика и приборы контроля параметров среды

Заключение

В данной работе предложен новый метод ранжирования разнородных результатов поиска. Его отличительные особенности состоят в следующем:

- ранжируемые результаты рассматриваются в совокупности, а не по отдельности;
- результаты располагаются в соответствии с критерием полезности, основанным на действиях пользователей на поисковой выдаче.

Благодаря этим особенностям метод обладает рядом преимуществ. Во-первых, он универсален: он может быть применен для ранжирования результатов произвольного вида и для разных моделей поисковой выдачи. Во-вторых, он позволяет естественным образом учитывать взаимосвязи между результатами. И в-третьих, он не требует экспертных оценок для обучения.

Предложенный метод был реализован и применен для встраивания 32-х видов специализированных результатов в поисковую выдачу системы Яндекс для мобильных устройств. Встраивались результаты поиска по изображениям, видео, мобильным приложениям, товарам, новостям, результаты гео-поиска и других сервисов компании Яндекс.

Было оценено качество работы метода по поисковым метрикам, основанным на экспертных оценках и на поведении пользователей. В сравнении с текущим используемым методом было получено улучшение точности показа специализированных результатов на 21.22% при снижении полноты на 29.21% и прирост качества по метрике *pfound* на 0.27%. (TODO: уточнить результаты) (TODO: + online-метрики)

В процессе реализации метода также была решена задача нахождения заданного числа кандидатов в аргументы максимизации значения функции, представляющей собой ансамбль “забывчивых” деревьев решений (oblivious decision trees), по частично

вычисленному вектору признаков. Решение этой задачи позволяет избежать обращения к тем поисковым источникам, результаты которых заведомо нерелевантны заданному поисковому запросу. Разработанное решение имеет самостоятельную ценность и может быть применено и в других задачах.

Список литературы

[1] International Data Corporation (IDC). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. // EMC website, URL: `http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm` (дата обращения: 7.05.2015).

[2] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval. // Cambridge University Press. 2008.

[3] Дородницын А. А. и др. Словарь по кибернетике. 2-е издание, под ред. Михалевича В.С. // Гл. ред. УСЭ им. М. П. Бажана, 1989.

[4] Ашманов (TODO)

[5] A Brief History of Search Engines. // Webreference website, URL: `http://www.webreference.com/authoring/search_history` (дата обращения: 19.05.2015).

[6] Tie-Yan Liu. Learning to rank for information retrieval // Foundations and Trends in Informaton Retrieval, vol. 3, no. 3, pp. 225–331, 2009.

Словарь терминов

Поисковая система?

Поисковый запрос

Поисковая выдача

Поисковый источник

Специализированный ответ (специализированный результат)

Ранжирование