

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

Диссертация допущена к защите
Зав. кафедрой

_____ В.Е.Клавдиев
"___"

**ДИССЕРТАЦИЯ
на соискание степени МАГИСТРА**

Тема: метод ранжирования разнородных результатов поиска

Направление: 010400 - Прикладная математика и информатика
Магистерская программа: системное программирование

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

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

Консультанты:

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

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

Санкт-Петербург
2015

Содержание

Введение	2
1 Обзор литературы	8
2 Постановка задачи	9
Заключение	10
Список литературы	11

Введение

Системы информационного поиска на сегодняшний день играют важную роль в нашей жизни. 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 зеттабайт¹, 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. Эта информация большей частью неструктурирована и очень разнообразна. Несомненно, без помощи поисковых систем ориентироваться в этом огромном информационном пространстве не представляется возможным.

Система информационного поиска или *поисковая система* – это компьютерная система, предназначенная для поиска информации, соответствующей информационной потребности пользователя, в больших массивах неструктурированных данных [2]. Принято выделять три типа поисковых систем:

- *системы веб-поиска (web search engines)* – системы, предназначенные для поиска информации в сети Интернет;
- *предметно-ориентированные поисковые системы (domain-specific search engines)* – системы, ориентированные на поиск информации в определенной предметной области (например, поиск публикаций в электронной библиотеке, поиск патентов в патентной базе или поиск документов во внутренней сети организации);
- *системы персонального поиска (personal information retrieval systems)* – системы

¹1 зеттабайт (ЗБ) = 1 триллион гигабайт

для поиска по персональной информации пользователя (например, поиск файлов на персональном компьютере или поиск электронных писем в почтовом ящике).

Веб-поиск – одна из активно развивающихся областей информационного поиска. Системы поиска в интернете в последнее время стали неотъемлемой частью нашей жизни. Когда возникает необходимость найти какую-то информацию, выбрать товар или услугу, либо найти интернет-ресурс, мы все чаще обращаемся за решением задачи к поисковым системам вместо энциклопедий, справочников, словарей, телефонных книг, газет и т. д. А с развитием и ростом популярности мобильных устройств и мобильного интернета мы получили возможность прибегать к помощи поисковых систем не только дома или на работе, но практически в любом месте, где бы ни находились.

Пользователь взаимодействует с поисковой системой, формулируя свою информационную потребность в виде *поискового запроса*, задавая его системе и получая от нее *результаты поиска*. Поисковый запрос обычно представляет собой набор ключевых слов или короткую фразу. Результаты поиска – это те информационные объекты, поиск которых осуществляется системой. Например, в случае веб-поиска это могут быть страницы веб-сайтов, в случае поиска по электронной библиотеке – книги, журналы и статьи, а при поиске по электронному почтовому ящику – электронные письма. Совокупность результатов поиска, выдаваемых поисковой системой в ответ на запрос, называется *поисковой выдачей* (+представление?). Каждый из найденных результатов может в большей или меньшей степени соответствовать по смыслу заданному поисковому запросу. Эта характеристика поискового результата – мера семантического соответствия поисковому запросу – называется *релевантностью* [3, 4].

Одной из важнейших задач при построении поисковой системы является задача *ранжирования* результатов поиска. Ранжирование – это упорядочение найденных результатов по их релевантности [4, 6] (+определение релевантности?). То, как располагаются найденный результаты в поисковой выдаче, непосредственно влияет на качество работы поисковой системы. Цель поисковой системы – максимально точно и быстро давать ответы на вопросы пользователя, то есть выдавать такие результаты, чтобы пользователь смог найти интересующую его информацию, потратив как можно меньше усилий и времени. Поэтому результаты требуется расположить так, чтобы наи-

более релевантные из них были доступны пользователю в первую очередь. Также чем более релевантны результаты в целом, выдаваемые поисковой системой, тем успешнее пользователь сможет удовлетворить свою информационную потребность в принципе.

Исторически веб-поиск зарождался как поиск интернет-страниц. Первые прообразы поисковых систем позволяли искать файлы, размещенные на серверах, – сначала по названию, а затем и по содержащемуся тексту [5]. Интернет-страница является основным типом поисковых результатов и в современных поисковых системах (рис. 1). Однако на сегодняшний день в интернете содержится информация различного вида, и помимо веб-страниц пользователя также могут интересовать и другие информационные объекты – изображения, видеозаписи, аудио-файлы, программные приложения и т. д. Также размещаемая в интернете информация очень разнообразна по своей семантике. Так, например, текст может быть текстом книги, новостью в интернет-газете, пояснением термина в словаре, инструкцией по применению лекарства, описанием характеристик товара или отзывом об этом товаре. В связи с этим одна из тенденций развития современных поисковых систем – встраивание в поисковую выдачу результатов *вертикальных поисков*. Вертикальный поиск – это поиск информации определенного типа (например, поиск изображений или видеозаписей) или информации, посвященной определенной тематике (например, поиск новостей, товаров, авиабилетов). К приме-

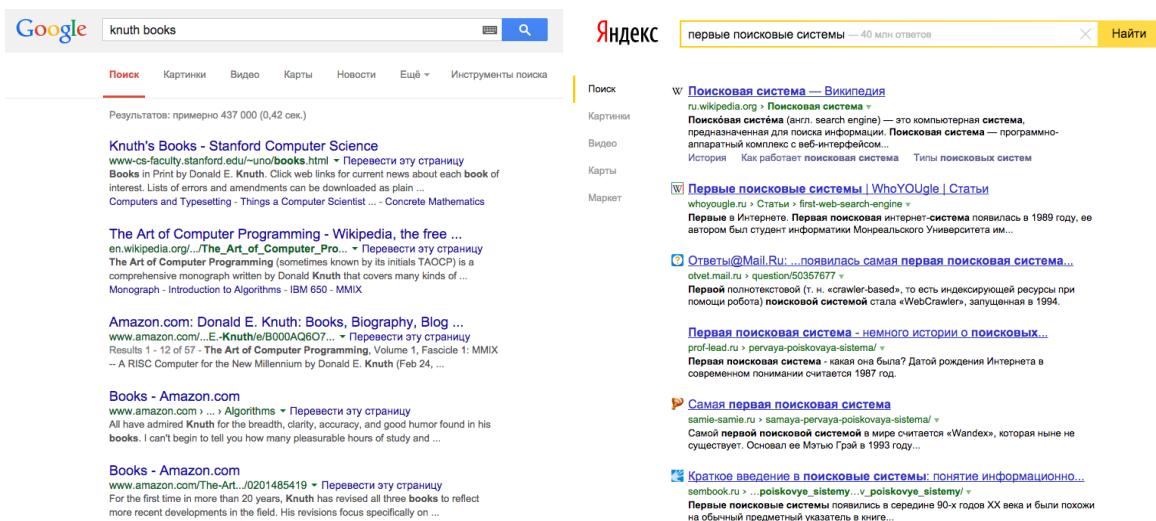


Рис. 1: Примеры поисковых выдач систем Google и Яндекс, состоящих только из ссылок на интернет-страницы.

ру, в поисковой выдаче таких поисковых систем как 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). Такие специализированные результаты так же могут быть в большей или меньшей степени релевантны поисковому запросу, и их так же требуется располагать в поисковой выдаче в соответствии с их релевантностью. Таким образом, возникает задача ранжирования *разнородных* результатов поиска.

Наиболее ранние исследования в области ранжирования разнородных результатов

Яндекс | [картины маринистов — 143 тыс. ответов](#) | [Найти](#)

Поиск | Картины | Видео | Карты | Маркет | Ещё

[Море, картины художников.маринисты...](#)
atap.ru > painting2.htm
Для любителей морских пейзажей – картины художников-маринистов для рабочего стола компьютера, в большом разрешении 1024x768.

[Картины маринистов - корабли \(1/133\) | Форумы BalanceForums](#)
airbase.ru » Картины маринистов » Balance.Ru » форумы » старые » Форумы Авиабазы » Флот » Морской » Картины маринистов - корабли. Картины маринистов - корабли.

[Восхитительные картины мариниста Г.Дмитриева....](#)
vdohnovenie2.ru > voskhitelnye-kartiny-marinistov...
Восхитительные картины мариниста Г.Дмитриева. Опубликовано в рубрике живопись, Современные художники | Ноябрь 30th, 2011.

[Смотрите картинки по запросу «картины маринистов»](#)
 Ещё картинки >
images.yandex.ru > картинки маринистов

[Маринисты, картины Арт Холстер, картины, картинная...](#)
Art-hoister.ru » Маринисты » В разделе «Маринисты» нашей галереи Вы можете купить более 60 картин, постеров и репродукций в 5 направлениях, живопись и портреты, предметы интерьера.

Яндекс | [nevsky prospect 20 — 2 млн ответов](#) | [Найти](#)

Поиск | Картины | Видео | Карты | Маркет | Новости | Ещё

[Невский проспект, 20 на карте Санкт-Петербурга](#)

В этом доме: Biblioteca Food and the City, Библиотечно-информационный и культурный центр искусства и музыки, Даники, все организации
Ближайшее метро: • Адмиралтейская • Невский Проспект
Гостиниц Двор
Как добраться на машине, транспортом

[...по адресу Санкт-Петербург, проспект Невский, 20...](#)
MaxiKarta.ru > Справочник > ...проспект_Nevskij_20_0_0...
Санкт-Петербург, проспект Невский, 20. Список компаний в этом здании ... проспект Большохинокий, 25. - 260м до ближайшего отделения связи

[Невский проспект д. 20 на карте Санкт-Петербурга](#)
SpbMap.ru > streets/nevsky-prospekt20.html
Как добраться до проспекта Невский проспект дом 20. Индекс дома 20 по проспекту Невский проспект.

Google | [новости спорта](#) | [Найти](#)

Поиск | Новости | Видео | Картины | Карты | Ещё | Инструменты поиска

Результатов: примерно 23 700 000 (0,39 сек.)

[Новости спорта, Спортивная аналитика, Видео](#)
news.sportbox.ru
Гран-при с Алексеем Поповым. Актуальные новости и последние слухи · Сергей Ковалев - боксер года по версии премии «Звезда бокса» · Роналдо: С...
Футбол - Сиатл - Хоккей - Результаты

[В новостиах](#)

Lenta.ru - 8 ч. назад
Во время смотреть Дома паралимпийского спорта, Фан Ван Туван совместно с ...
Самые подозрительные рекорды в истории спорта
Чемпионат.com - 3 ч. назад
«Чемпионат»: телевизионные частоты «России 2» перейдут к «НТВ-Плюс» - Все виды спорта
Eurosport - 20 ч. назад
Другие новости по запросу новости спорта

[Чемпионат.com: новости спорта - Чемпионат](#)
www.championat.com
Чемпионат — все самые свежие новости спорта, видео, фото. Чемпионаты мира, Европы. Чемпионаты по футболу, хоккею, баскетболу и др. видам ...

Google | [кафе в санкт петербурге](#) | [Найти](#)

Поиск | Карты | Картины | Новости | Видео | Ещё | Инструменты поиска

Результатов: примерно 806 000 (0,51 сек.)

[Кафе, санкт петербурге \(поблизости\)](#)

Zoom	Рейтинг	Количество отзывов	Адрес
	4.6 ★★★★★	123 отзыва	Кафе Городская ул.
	4.6 ★★★★★	85 отзывов	Ресторан, столовая или кафе Большая Морская ул.
	4.2 ★★★★★	35 отзывов	Ресторан, столовая или кафе Гатчинская ул.

[Ещё результаты по запросу "кафе"](#)

[Лучшие рестораны и кафе Санкт-Петербурга, рейтинг ...](#)
spb.zoon.ru/restaurants/
Рейтинг лучших ресторанов Санкт-Петербурга по отзывам посетителей на Zoon.ru. Удобный поиск ресторанов и кафе на карте Санкт-Петербурга по ...

[Кафе Санкт-Петербурга - RestoClub.RU](#)
wwwレストoclub.ru/search/?cats=2
Гранд-кафе «21-я верстка», что на пересечении Московского проспекта и 7-й

Рис. 2: Специализированные результаты в выдаче поисковых систем Яндекс и Google.

поиска касаются встраивания одного специализированного результата конкретного вертикального источника на первое место в списке результатов поиска (**(TODO: ссылки)**) и встраивания одного из нескольких специализированных результатов так же на первое место (**(TODO: ссылки)**). 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)

Также следует отметить, что список, в котором элементы упорядочены по релевантности, – 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.

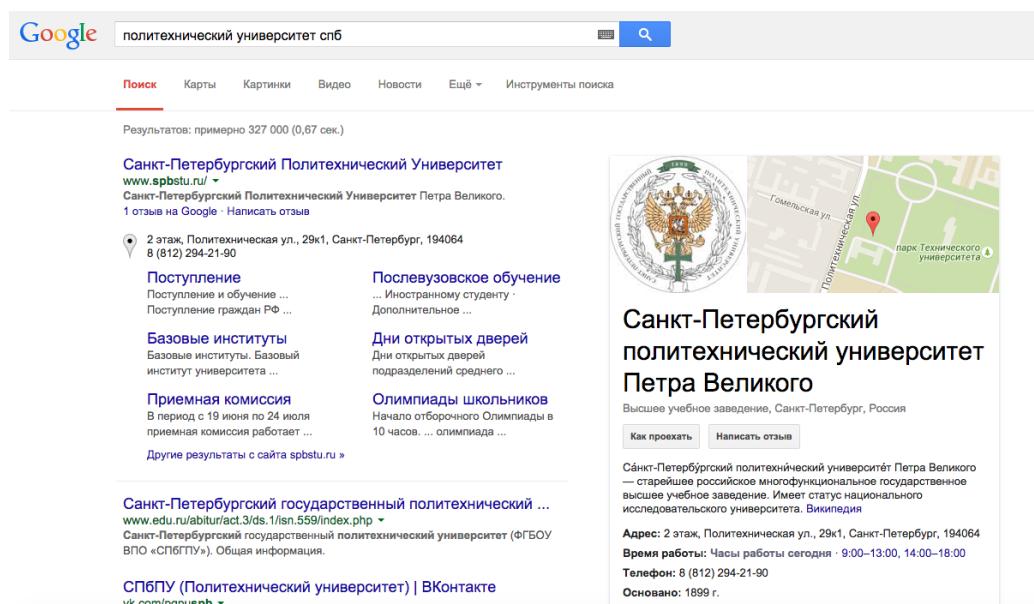


Рис. 3: Поисковая выдача системы Google со специализированным результатом в отдельной колонке.

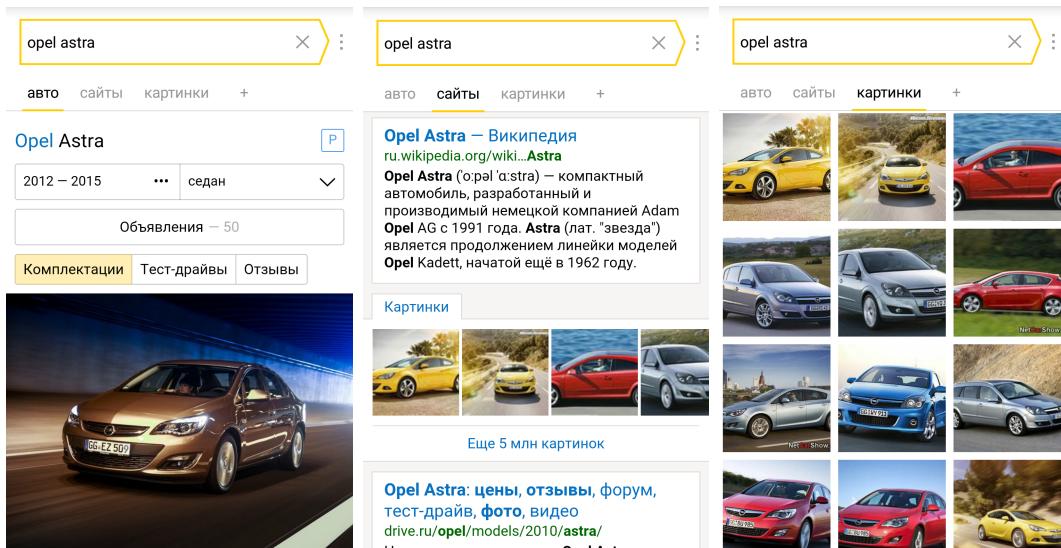


Рис. 4: Выдача мобильного приложения поисковой системы Яндекс с результатами поиска на отдельных страницах.

3), а выдача мобильного приложения поисковой системы Яндекс состоит из страниц, каждая из которых может содержать список результатов или специализированные ответы и результаты вертикальных поисков, и которые располагаются в соответствии с релевантностью содержащихся результатов (рис. 4). В таком случае задача ранжирования усложняется и превращается в задачу расположения поисковых результатов в соответствии с заданной моделью поисковой выдачи. Это также требует обобщения методов ранжирования результатов поиска.

Глава 1

Обзор литературы

План:

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

Глава 2

Постановка задачи

Заключение

В данной работе предложен новый метод ранжирования разнородных результатов поиска. Главная отличительная особенность метода состоит в том, что результаты поиска располагаются исходя из соображений максимизации релевантности всей поисковой выдачи в целом, а не в соответствии с релевантностями отдельных результатов. Благодаря этому метод является универсальным – он может быть применен к разнообразным видам поисковых результатов и для разных моделей поисковой выдачи. Также переход от рассмотрения поисковых результатов по отдельности к рассмотрению выдачи в целом позволяет естественным образом учитывать зависимости и отношения между разными типами результатов. Кроме этого предложенный метод не требует асессорских оценок – он основывается на поведении пользователей на поисковой выдаче.

(Еще преимущества?)

Предложенный метод был реализован и применен для встраивания 31 типа специализированных результатов вертикальных поисковых источников в мобильную поисковую выдачу системы Яндекс. Использовались специализированные результаты поиска по картинкам, видео, мобильным приложениям, поиска товаров, новостей, погоды, результаты гео-поиска и других сервисов компании Яндекс. Была проведена оценка качества работы метода [ref] и сравнение с текущим используемым методом встраивания специализированных результатов [ref] по метрикам, основанным на асессорских оценках: по точности и полноте показа специализированных результатов и метрике *pfound* [ref] (+ online-метрики?). Сравнение показало 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)

В ходе реализации метода и встраивания его в поисковую систему также была

решена задача эффективного нахождения аргумента максимизации функции, представляющей собой ансамбль решающих деревьев специального вида (*oblivious decision trees*), и нахождения заданного числа кандидатов в аргументы максимизации при наличии частично вычисленного вектора признаков [ref]. Решение этой задачи позволяет избежать задания запросов к тем поисковым источникам, результаты которых будут заведомо нерелевантны заданному поисковому запросу. Также следует отметить, что решение данной задачи имеет самостоятельную ценность, и может быть применено не только для реализации предложенного метода ранжирования разнородных результатов поиска, но 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 Information Retrieval, vol. 3, no. 3, pp. 225–331, 2009.