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 [23]. 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) [24]. Такой вид результатов поиска для того времени был естественным, поскольку представляемая в них информация была однотипной. 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 музыку, фильмы, новости, тексты книг, 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, ново-

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

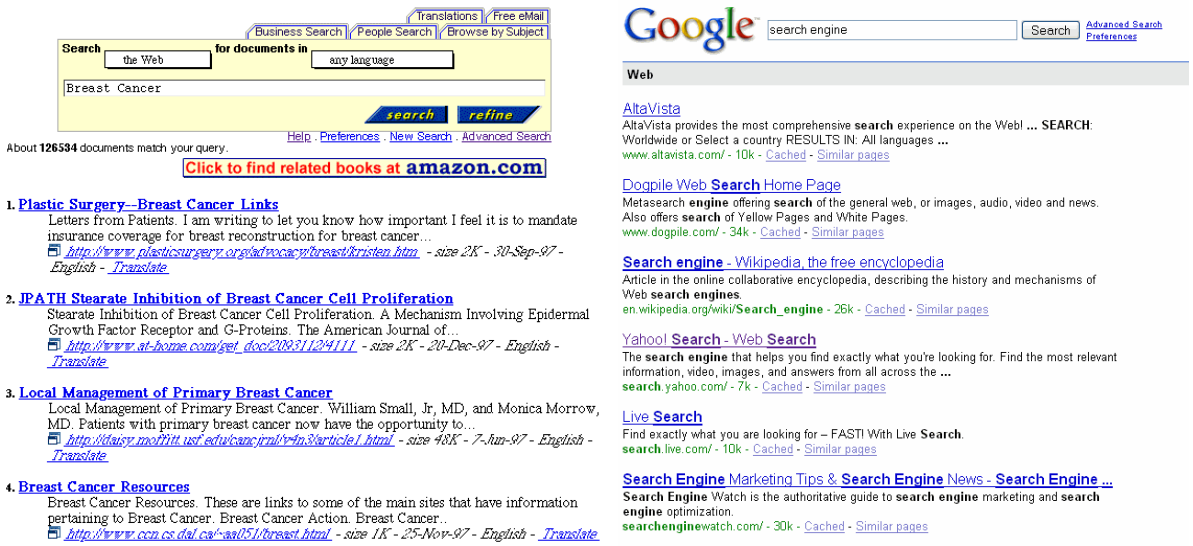

Рис. 1: Страницы результатов поиска системы AltaVista и одной из первых версий системы Google.

стей, 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 систем 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

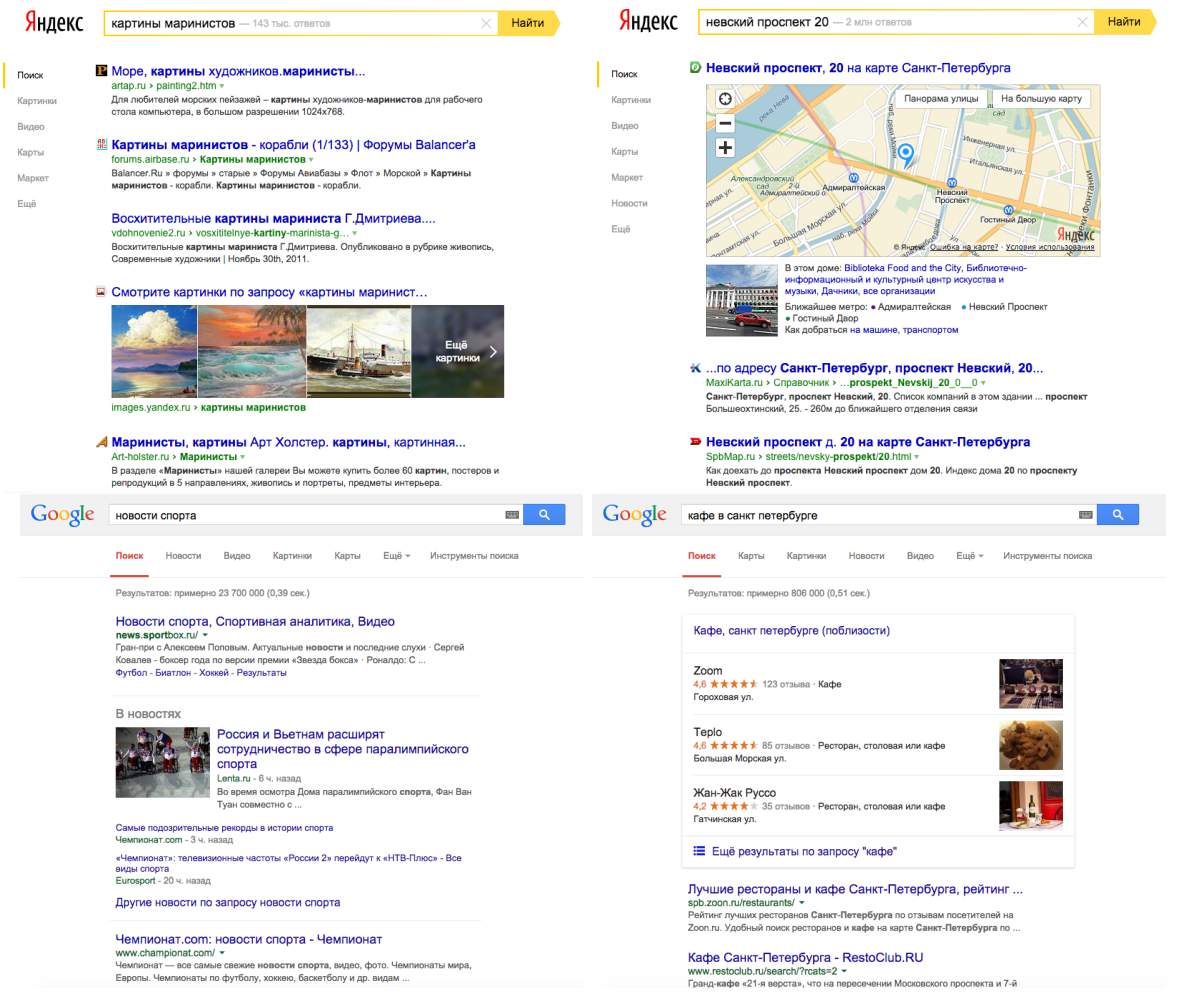

Рис. 2: Специализированные ответы в результатах поиска систем Яндекс и Google.

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 [19, ?]. 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. Каждый из специализированных ответов, встраиваемых в результаты поиска, также может быть более или менее полезен пользователю в сравнении с другими результатами при решении своей поисковой задачи. Кроме этого, какие-то из специализированных ответов могут быть в принципе не уместны для заданного поискового запроса. Таким образом, возникает задача выбора специализированных результатов для показа по запросу и ранжирования их в поисковой выдаче.

Задача ранжирования, возникающая при смешивании специализированных ответов с традиционными результатами веб-поиска, существенно отличается от классической задачи ранжирования [?]. Во-первых, классическая задача формулируется для однотипных объектов, в то время как в рассматриваемой задаче ранжируемые результаты разнородны. Во-вторых, специализированные ответы встраиваются только на первую страницу результатов поиска, поэтому количество ранжируемых объектов при этом невелико. Эти отличия, в особенности первое их них, не позволяют применить для решения рассматриваемой задачи существующие методы решения классической задачи. Соответственно, требуется либо их обобщение, либо разработка принципиально новых подходов.

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

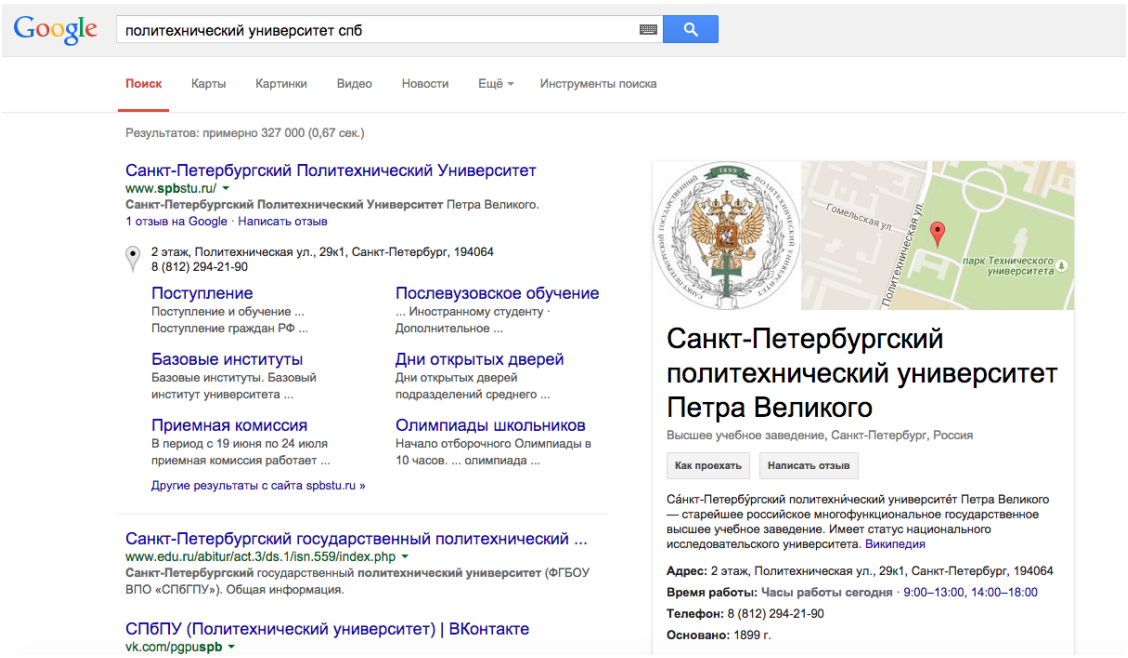

Рис. 3: Поисковая выдача системы Google со специализированным результатом в отдельной колонке.

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 таких случаях задача ранжирования усложняется и превращается в задачу расположения результатов поиска в соответствии с заданной моделью поисковой выдачи. Это также требует обобщения существующих методов ранжирования.

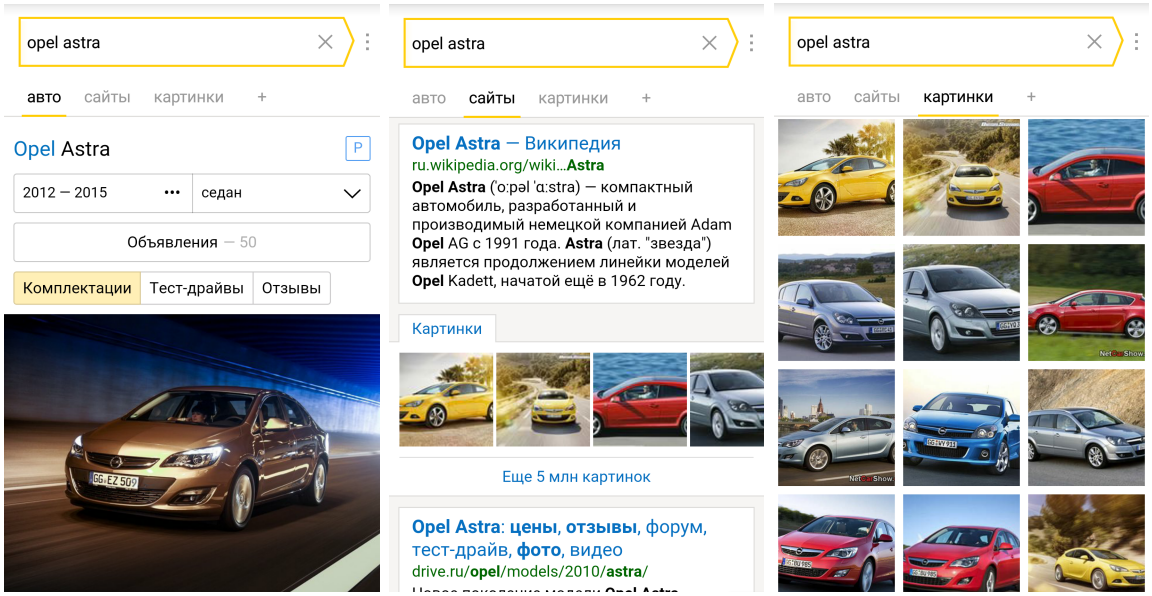

Рис. 4: Выдача мобильного приложения поисковой системы Яндекс с результатами поиска на отдельных страницах.

В данной работе рассматривается задача ранжирования разнородных результатов поиска и предлагается универсальный метод ее решения. Этот метод может быть применен для ранжирования результатов произвольного вида и для разных моделей поисковой выдачи. Также он не требует экспертных оценок, поскольку основывается на поведении пользователей. (TODO: Описать, что в каком разделе).

Глава 1

Обзор литературы

Задача ранжирования занимает важное место в информационном поиске, и на эту тему имеется достаточно большое количество публикаций. Однако подавляющее большинство из них относится к задаче ранжирования однотипных объектов (классической задаче ранжирования) [15]. При встраивании же специализированных ответов в результаты поиска требуется ранжировать разнородные объекты. Это не дает возможности применить имеющиеся методы решения классической задачи – главным образом потому, что разнородные объекты нельзя описать одним набором признаков. В связи с этим рассматриваемая задача требует отдельных исследований, которых на данный момент пока не так много.

В этой главе сначала даются краткие сведения о классической задаче ранжирования и основных методах решения. Затем проводится обзор имеющихся на данный момент публикаций, посвященных ранжированию разнородных результатов.

1.1 Классическая задача ранжирования

1.1.1 Формулировка задачи

(TODO:)

1.1.2 Обзор методов решения

(TODO:)

1.2 Ранжирование специализированных ответов

(TODO:)

1.2.1 Формулировка задачи

1.2.2 Обзор методов решения

Наиболее ранние исследования в области смешивания разнородных результатов поиска касаются встраивания одного конкретного специализированного ответа на первое место в поисковой выдаче [4, 12, 14] и встраивания одного из нескольких специализированных ответов так же на первое место [1, 2, 5]. Однако подобные методы подходят лишь в тех случаях, когда для заданного запроса уместен только один специализированный результат, и этот результат наиболее подходящий из всех остальных. В более поздних исследованиях рассматривается возможность встраивания специализированных ответов на различные позиции в поисковой выдаче [13, 16, 3, 8]. Однако методы, предлагаемые в работах [13, 16, 3] не достаточно универсальны – они предназначены для встраивания какого-то конкретного специализированного ответа, или существенно зависят от специфики используемых специализированных ответов. Также методы [16, 3] используют экспертные оценки. Это делает их применение дорогостоящим и затрудняет регулярное обновление используемых машинно-обученных моделей, необходимое для учета изменений в потоке поисковых запросов.

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] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. // SIGIR 2009, pp. 315–322. ACM, 2009.

[2] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. // SIGIR 2010, pp. 691–698. ACM, 2010.

[3] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. // CIKM 2011, pp. 201–210. ACM, 2011.

[4] F. Diaz. Integration of news content into web results. // WSDM 2009, pp. 182–191. ACM, 2009.

[5] F. Diaz and J. Arguello. Adaptation of offline vertical selection predictions in the presence of user feedback. // SIGIR 2009, pp. 323–330. ACM, 2009.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais, T. White. Evaluating Implicit Measures to Improve Web Search. // ACM TOIS, 23(2), pp. 147-168, 2005.

[7] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2nd edition. // Springer, 2009.

[8] L. Jie, S. Lamkhede, R.t Sapra, E. Hsu, H. Song, Y. Chang. A Unified Search Federation System Based on Online User Feedback. // Proceedings of KDD 2013. ACM, 2013.

[9] T. Joachims. Optimizing Search Engines using Clickthrough Data. // Proceedings of KDD 2002. ACM, 2002.

[10] Y. Kim, A. Hassan, R. White, I. Zitouni. Modeling Dwell Time to Predict Click-level Satisfaction. // Proceedings of WSDM 2014.

[11] R. Kohavi. Ch. Li. Oblivious Decision Trees, Graphs, and Top-Down Pruning. // IJCAI, pp. 1071-1079, 1995.

[12] A. C. König, M. Gamon, and Q. Wu. Click-through prediction for news queries. // SIGIR 2009, pp. 347–354. ACM, 2009.

[13] D. Lefortier, P. Serdyukov, F. Romanenko, M. de Rijke. Blending Vertical and Web results: A Case Study using Video Intent. // ECIR 2014, pp. 184-196.

[14] X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs. // SIGIR 2008, pp. 339–346. ACM, 2008.

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

[16] A. K. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach, and T. Kanungo. On composition of a federated web search result page: Using online users to provide pairwise preference for heterogeneous verticals. // WSDM 2011, pp. 715–724. ACM, 2011.

[17] M. Sokolova, G. Lapalme. A systematic analysis of performance measures for classification tasks. // Information Processing and Management 45, p. 427–437. Elsevier, 2009.

[18] R. White, J. Huang. Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs. // Proceedings of SIGIR 2010. ACM 2010.

[19] И. Ашманов, А. Иванов. Оптимизация и продвижение сайтов в поисковых системах, 3-е издание. // Спб.: Питер, 2011.

[20] А. Гулин, П. Карпович, Д. Расковалов, И. Сегалович. Оптимизация алгоритмов ранжирования методами машинного обучения. // РОМИП 2009.

[21] А. Фонарев, А. Дьяконов. Обзор алгоритмов бустинга. // МГУ, 2012.

[22] Алгоритм машинного обучения MatrixNet. // URL: `https://company.yandex.ru/technologies/matrixnet/`. Дата обращения: 3.06.2015.

[23] 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).

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

Словарь терминов

Поисковая система?

Поисковый запрос

Поисковая выдача

Поисковый источник

Специализированный ответ (специализированный результат)

Ранжирование

Тип информации