Оценка разнообразия результатов веб-поиска с помощью материалов Wikipedia

© А.В. Стрелковский, И.Е. Кураленок

С.-Петербургский государственный университет

thinkingwhat@yandex.ru, solar@yandex-team.ru

Аннотация

Существует большое количество многозначных поисковых запросов. Целью данной работы является оценка качества результатов, выдаваемых поисковыми системами по таким запросам, рассмотренного с точки зрения количества представленных в них тем, а также исследование метода увеличения количества тем в ответе поисковой системы, основанного на кластеризации результатов поиска. В качестве источника тем и неоднозначных терминов используются материалы Википедии.

1 Введение

Есть два подхода к решению проблемы ответа поисковой системы по неоднозначным запросам: генерализация выдачи и персонализация выдачи. Суть персонализации – подбор дополнительной информации на запрос каждого конкретного пользователя (например, подсказки). Проблема здесь заключается в том, что поисковая система, как правило, владеет довольно скромным количеством информации о большинстве пользователей и не в состоянии обеспечить качественную персонализацию. Поэтому стоит обратить внимание на генерализацию – именно она и будет далее обсуждаться.

Рассмотрим пример неоднозначного запроса: пусть это будет запрос «ягуар». Под ним можно иметь в виду автомобильную компанию, напиток, животное, производителя металлических дверей и т. д. Пользователь, отправивший такой запрос, мог иметь в виду животное, и поэтому ответ из 10 ссылок на сайты по продаже автомобилей его не удовлетворит, в то время как человек, ищущий что-то про автомобили, не будет в восторге, увидев 10 ссылок про двери, напиток или кошек. Возникает проблема обеспечения разнообразия результатов на первой странице ответа поисковой системы, так как нашей целью является удовлетворение нужд всех пользователей (генерализация).

Некоторым может показаться, что большая часть пользователей не будет задавать такие односложные и неоднозначные запросы, что логичнее добавить какие-то конкретизирующие слова, но статистические данные wordstat.yandex.ru говорят об обратном. Рассмотрим тот же «ягуар»: по данным Wordstat за последний месяц (относительно момен-та последней правки статьи) поисковой системе Яндекс было задано 71416 запросов «ягуар», в то время как запросов вида «ягуар + конкретизирующие слова» было задано на порядок меньше. Например, второе место по количеству запросов, содержащих слово «ягуар», – запрос «ягуар напиток» – это всего лишь 7073 показа, третье место – запрос «ягуар машина» – 4653 показа.

Тема разнообразия в результатах поиска рассматривается также в некоторых других работах. Например, в [2] авторы предлагают метод оценки качества поиска, принимающий во внимание разнообразие и объем новой информации, привносимой каждым документом в ответ системы. В [2] также упоминается ряд других работ, связанных с неоднозначностью запросов. В [10] описан обучающийся алгоритм ранжирования, обеспечивающий наличие как минимум одного релевантного документа на первых позициях ответа для любого пользователя, и, таким образом, обеспечивающий разнообразие. В данной же работе с помощью классификатора оценивается количество разных тем, представленных среди документов, возвращенных поисковой системой, и рассматриваются возможности кластеризации в задаче увеличения этого количества.

Есть такая крупная онлайн-энциклопедия: Wikipedia[1]. В ней, очевидно, есть та же самая проблема: многозначные термины. Решается она там с помощью так называемых "disambiguation pages" – страниц с перечислением и кратким описанием всех возможных тем по неоднозначному понятию. Если термин, введенный пользователем , имеет несколько значений, то он сначала попадает на такую страницу и на ней выбирает ту тему, которая его интересует. В данной работе для оценки разнообразия выдачи поисковиков используются данные disambi-guation-страниц Википедии.

2 Метод оценки разнообразия

2.1 Метод оценки разнообразия (классификация документов)

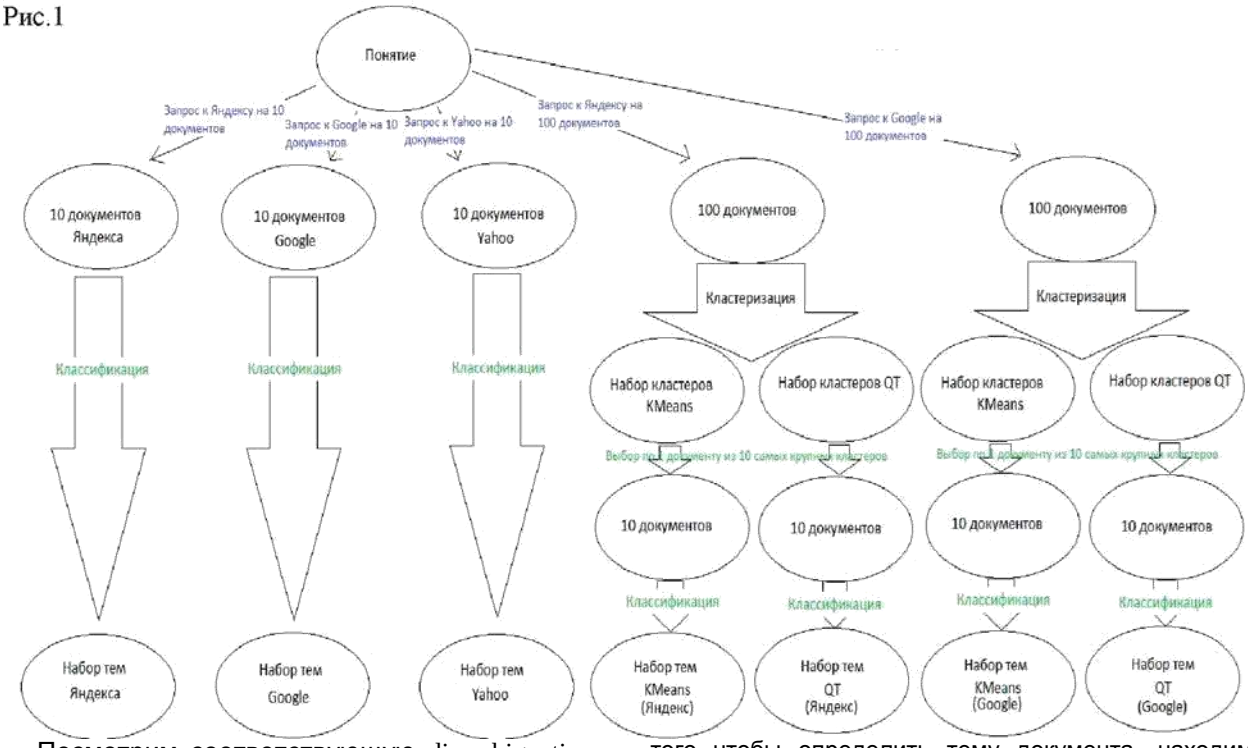
Рассмотрим пример многозначного запроса:

Пусть это снова будет слово «ягуар». Под ним можно подразумевать, например, следующие вещи (темы первых 10 запросов, содержащих данное сло-во, по статистике Wordstat):

- ягуар — напиток,
- ягуар — машина,
- ягуар — животное,
- ягуар — фильм

[1] http://wikipedia.org

Рис.1



Посмотрим соответствующую disambiguation-страницу Википедии. Каждая тема на ней описыва-ется словосочетанием либо одним – двумя предло-жениями. Темы для рассматриваемого термина с этой страницы:

- ягуар — хищник семейства кошачьих,
- Jaguar — английская автомобильная компания,
- Jaguar — слабоалкогольный напиток,
- Ягуар — название нескольких художественных фильмов: фильм Себастьяна Аларкона (1986), фильм Френсиса Вебера (1996), фильм Эрнеста Пинтоффа (1979),
- и еще 5 тем.

Как мы видим, темы, обозначенные нами здесь, присутствуют, помимо этого изложено еще некото-рое количество других тем.

Конечно, таким образом невозможно получить абсолютно все темы по любому понятию, но, так как Википедия является очень крупным ресурсом (3,288,067 статей на английском и 533,423 статьи на русском языках на момент написания работы), ко-торый наполнялся содержанием годами (запущена в январе 2001 года), ожидается, что оценки, получен-ные с ее использованием, будут достаточно полны. Также в 2007 году Wikipedia являлась первым по посещаемости сайтом в мире, посещенным после главной страницы Google. Википедия сейчас явля-ется самым крупным и наиболее популярным спра-вочником в интернете[2] . По объёму сведений и тема-тическому охвату она считается самой полной эн-циклопедией из когда-либо создававшихся за всю историю человечества.

В работе описание каждой темы рассматривается как «мешок слов». По каждому описанию строится tf-idf вектор с tf по форме bm25 (см. [5]). Далее, для того чтобы определить тему документа, находим ближайший к вектору документа (построенному тоже по bm25) вектор темы. Так как рассматрива-ются tf-idf векторы, то в качестве расстояния между векторами x и y бралось $\rho = 1 - \cos(x,y)$. Если рас-стояние от документа до найденного вектора темы меньше, чем некоторая заданная константа ε[3] , то документ относится к теме, соответствующей вы-бранному вектору, иначе – к теме "unclassified".

Автор понимает, что такой способ классифика-ции не всегда дает идеальные результаты, но плю-сом этого метода является то, что он не смещен в сторону какой-либо поисковой системы, так как используются данные из Википедии, априори яв-ляющиеся нейтральными относительно поисковых систем.

3 Суть работы и результаты

3.1 Суть работы

Для оценки разнообразия выдачи с помощью Википедии было набрано 576 неоднозначных рус-ских понятий и 705 неоднозначных английских по-нятий. Из каждого из этих наборов случайным обра-зом было выбрано по 300 понятий и проведена сле-дующая процедура. По каждому понятию:

- запрос 10 результатов первой страницы от-вета у систем Яндекс, Google, Yahoo;
- запрос 100 результатов первой страницы ответа у системы Яндекс и кластеризация этих ре-зультатов с помощью алгоритмов K-Means (k = 75) (см. [6] и [7]) и Quality Threshold[4] (см. [1]);
- из набора кластеров, полученных с помо-

[2] согласно Alexa® (http://www.alexa.com)

[3] в работе полагается $\varepsilon = 0.9$

[4] в работе использовалась вариация qt-алгоритма, описан-ная в [1]

щью QT (Яндекс), выбиралось 10 самых крупных; далее из каждого из них с равной вероятностью выбиралось по одному документу, таким способом был получен четвертый набор из 10 документов;

- то же самое делалось и для кластеров, по-лученных с помощью K-Means (Яндекс) – был по-лучен пятый набор из 10 документов;
- запрос 100 результатов первой страницы ответа у системы Google и кластеризация этих ре-зультатов с помощью алгоритмов K-Means (k = 75) и Quality Threshold[4];
- из набора кластеров, полученных с помощью QT (Google), выбиралось 10 самых крупных; далее из каждого из них с равной вероятностью вы-биралось по одному документу, был получен шес-той набор из 10 документов;
- то же самое делалось и для кластеров, по-лученных с помощью K-Means (Google) – был по-лучен седьмой набор из 10 документов;
- к полученным 7 наборам применялась классификация по способу, описанному ранее; в итоге были образованы группы тем для каждого набора.

На рис. 1 представлена схема, поясняющая дан-ную процедуру.

В процессе кластеризации использовались заго-ловки и аннотации («сниппеты») документов, так как это как раз то, что видит пользователь в ответе системы, и, соответственно, это то, на основе чего он решает, к какой тематике относится тот или иной документ. Сниппеты и заголовки при этом рассмат-ривались как «мешки слов», по которым строились tf-idf векторы.

Выбор документов из кластеров производился случайным образом с целью сгладить их неодно-родность (так как при выборе, например, медианы, наличие «плохого» документа в кластере, т. е. дале-кого по смыслу от остальных документов, может сильно сдвинуть эту самую медиану). В то же время вероятность выбора таким способом «плохого» до-кумента равна отношению количества этих «пло-хих» документов в кластере к размеру самого кла-стера.

3.2 Спектральный анализ

Число кластеров k = 75 для алгоритма K-Means было подобрано на основе спектрального анализа ([3], [4], [9]) результатов поиска по некоторому на-бору запросов. (K-Means был рассмотрен именно из-за своей линейной по количеству кластеризуе-мых объектов вычислительной сложности, поэтому было решено не проводить спектральный анализ для результатов по каждому из рассматриваемых неоднозначных запросов, а взять среднее число по уже полученной ранее статистике).

Суть анализа заключалась в рассмотрении мно-жества документов в выдаче поисковой системы в виде графа, вершинами которого являются доку-менты, а веса ребер положены равными $\cos(x,y)$, где x и y – векторы, построенные по соответствующим документам (т. е. вершинам графа). Далее строился лапласиан графа. Рассматривались как нормирован-ный, так и ненормированный лапласианы, метод их построения изложен в [3].

Далее строились графики собственных чисел ла-пласиана.

Пример. Запрос: «война и мир», 100 документов; система: Яндекс.

График собственных чисел ненормированного лапласиана (упорядочены по возрастанию, первое собственное число не отображено на графике, так как оно равно нулю для любого лапласиана) пред-ставлен на рис. 2.

Рис. 2 Полагалось, что количество кластеров в выдаче должно быть равно номеру того собственного чис-ла, после которого начинается резкий рост собст-венных чисел (В примере можно положить количе-ство кластеров равным 80). Данный эвристический подход более подробно описан в [3]. Вообще такой способ определения количества кластеров можно обосновать по-разному, например, методами спек-тральной теории графов (см. [9]) либо при помощи теории возмущений (в идеальном случае имеем k несвязанных кластеров, при этом первые k собст-венных чисел лапласиана будут равны нулю, что доказывается достаточно тривиально).

3.3 Результаты

Приведем некоторую общую статистику. По английским понятиям:

Система	Общее количество тем по 300 понятиям
Яндекс	853
QT (Яндекс)	1057
K-Means (Яндекс)	974
Google	1170
QT (Google)	1257
K-Means (Google)	1193
Yahoo	1329

По русским понятиям:

Система	Общее количество тем по 300 понятиям
Яндекс	857
QT (Яндекс)	790

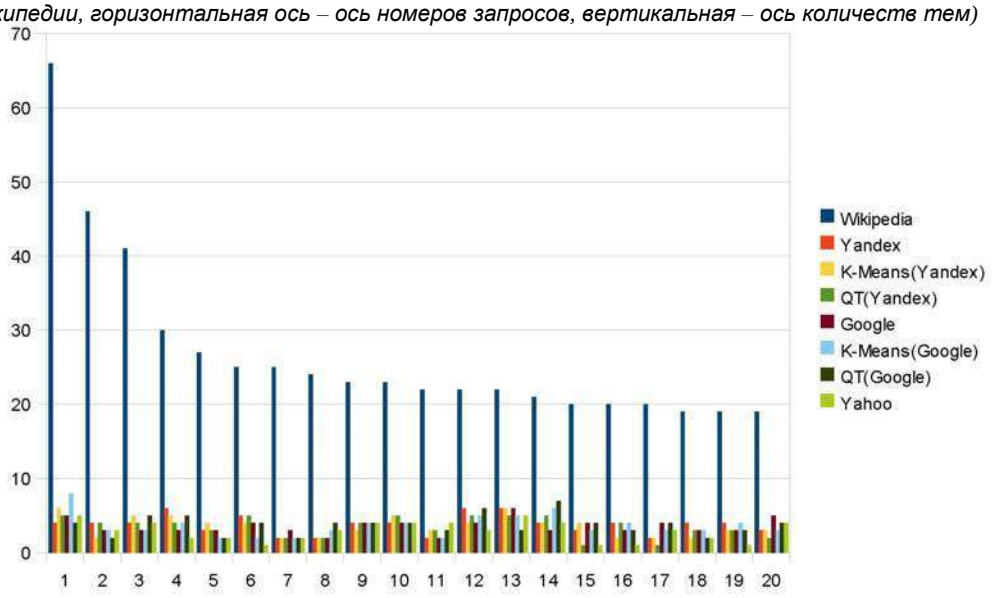
K-Means (Яндекс)	769
Google	905
QT (Google)	805
K-Means (Google)	787
Yahoo	641

Как видно из результатов, для английских понятий алгоритм QT позволил очень существенно улучшить разнообразие выдачи Яндекса (примерно на 23.9%). Более быстрый алгоритм K-Means (лине-ен по k и n) также обеспечил улучшение, но не столь значительное (около 14.2%). Для ответов поисковой системы Google обнаружилось более скромное улучшение разнообразия (около 7.4% для QT и около 2% для K-Means). В то же время, выяснилось, что поисковые системы Google и Yahoo в аспекте разнообразия результатов существенно обходят Яндекс.

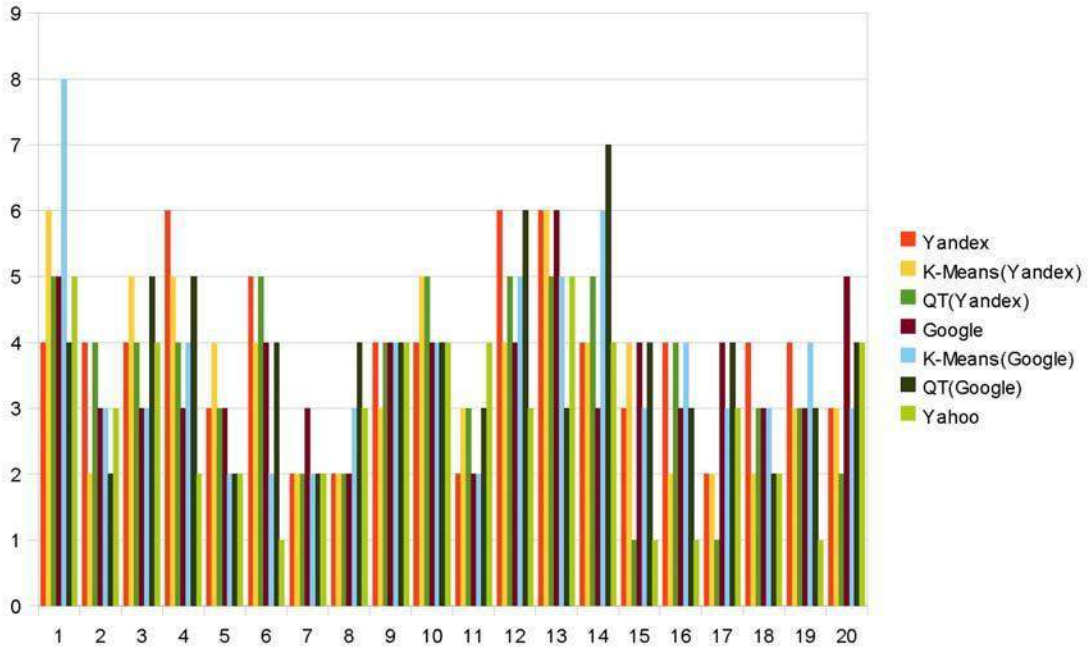
Что касается русских понятий, то тут немного иная картина: лидирует Google, за ним Яндекс, а последним идет Yahoo. В данном случае кластери-зация результат не улучшила ни для Яндекса, ни для Google.

Ниже представлены диаграммы, иллюстрирую-щие соотношение количеств тем по различным за-просам в ответах рассматриваемых поисковых сис-тем (а также в выборках, полученных в результате кластеризации).

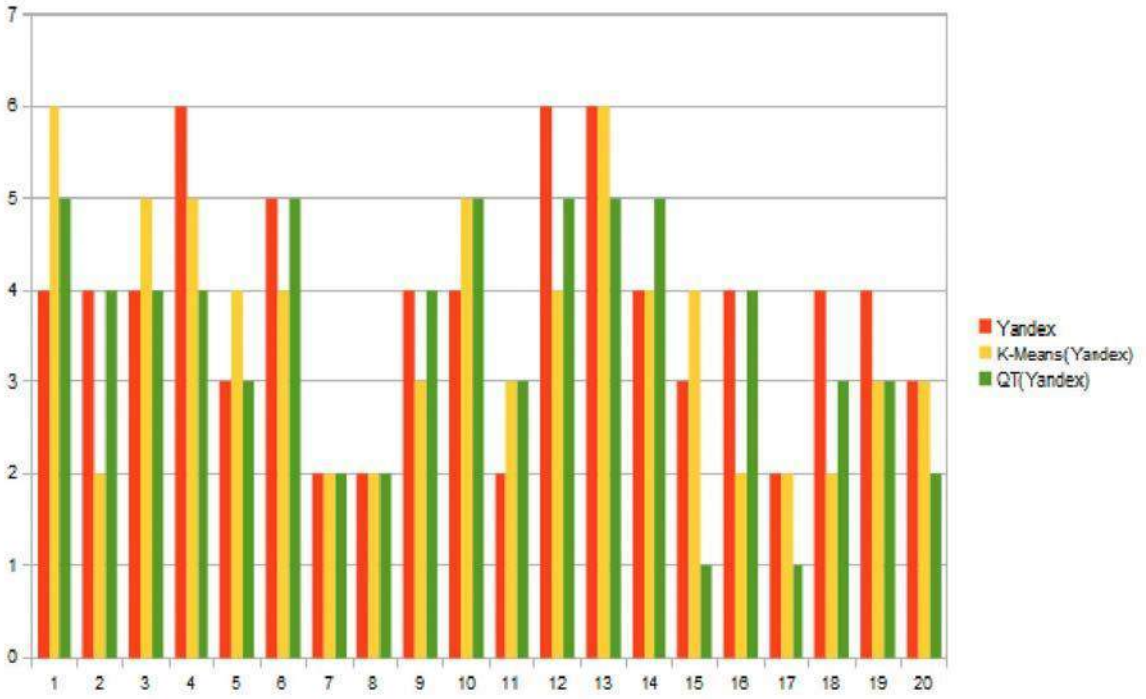
Соотношение по двадцати русским запросам (запросы упорядочены по убыванию количества тем в Википедии, горизонтальная ось – ось номеров запросов, вертикальная – ось количеств тем)



То же самое, но без столбцов по Википедии

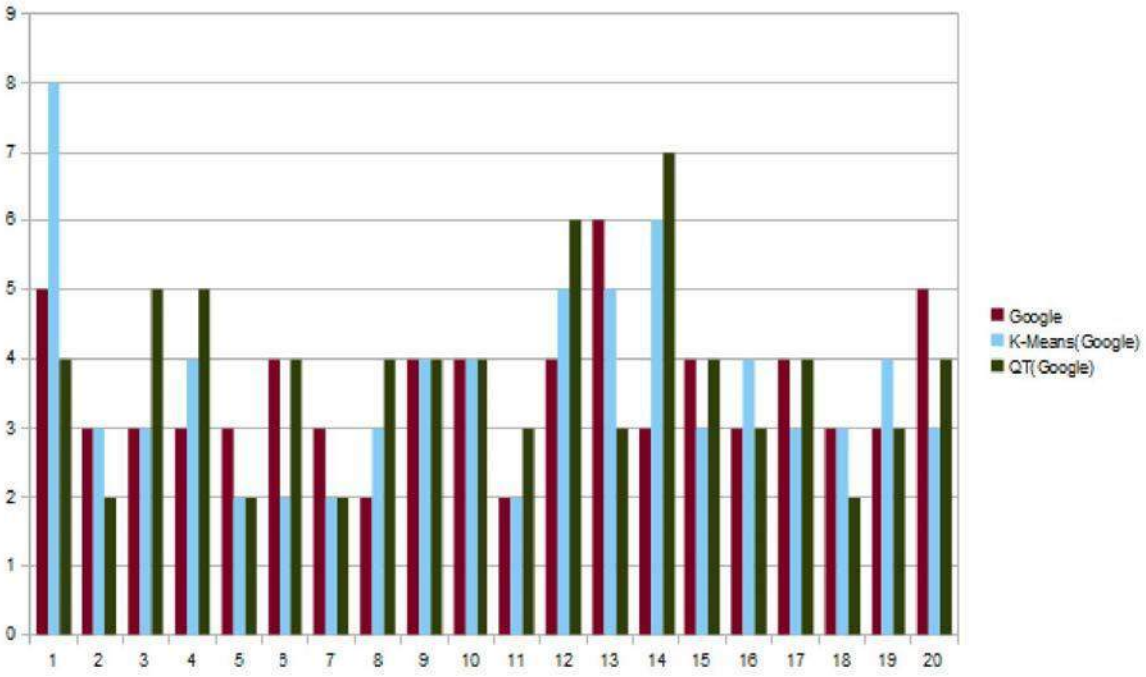


Только Яндекс (русскоязычные запросы)



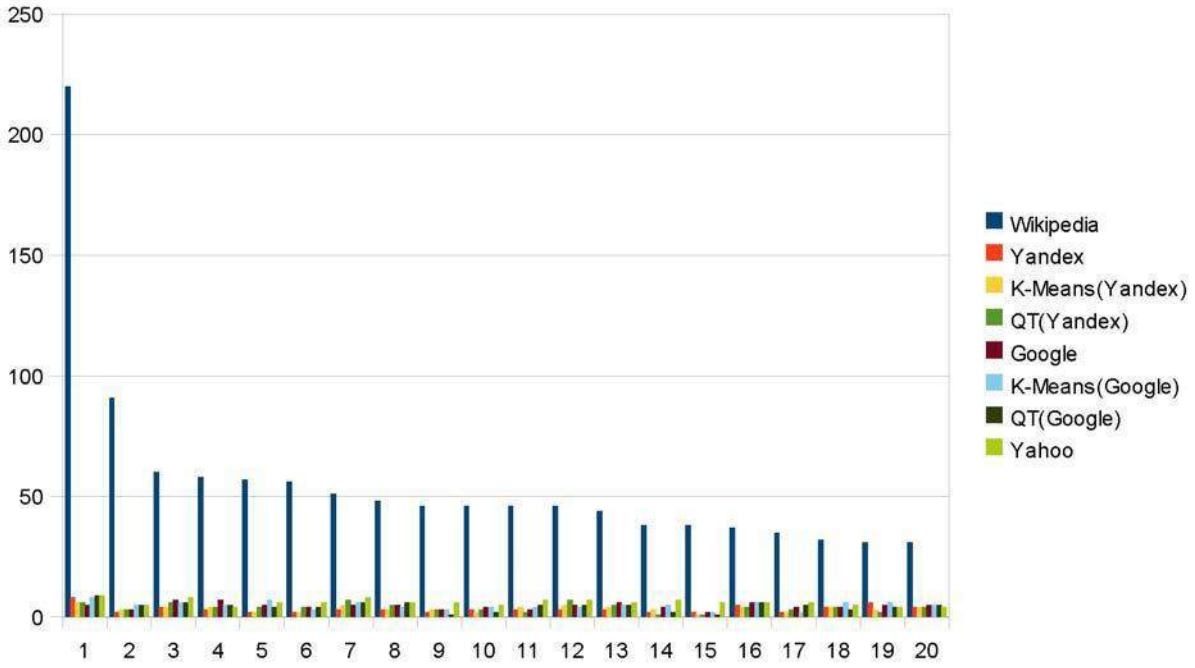
Здесь заметно, что столбцы по обычной выдаче Яндекса выше, чем столбцы по методам кластеризации.

Только Google (русскоязычные запросы)

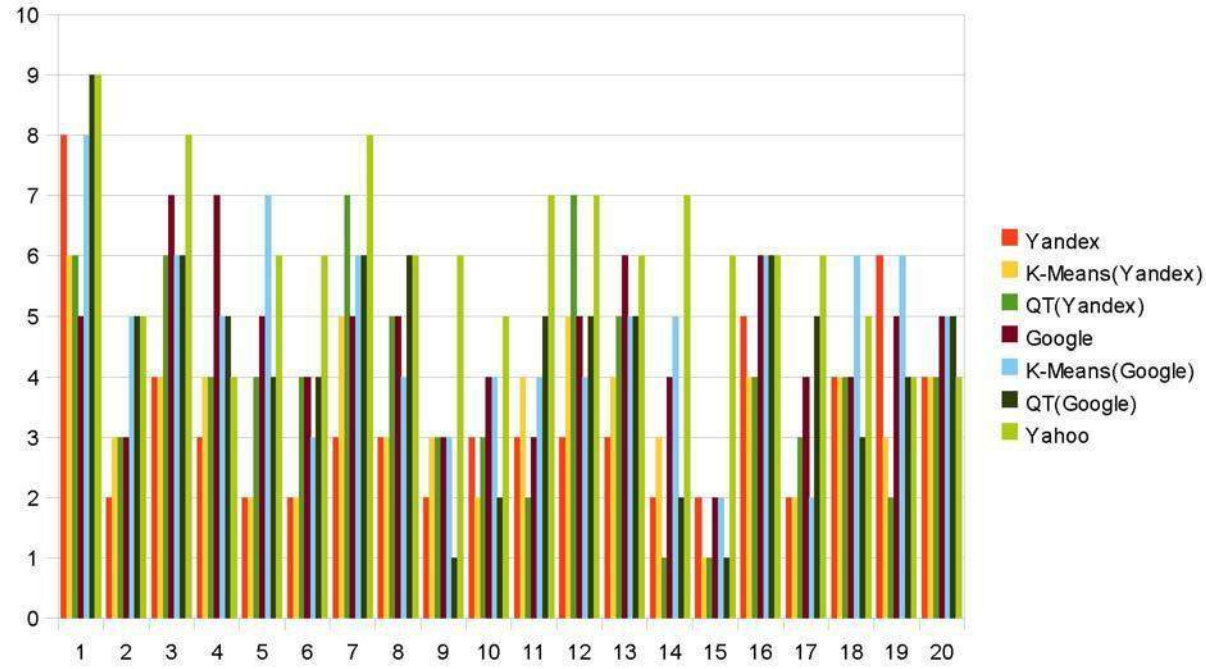


Здесь примерное равенство между обычным ответом Google и результатом работы обоих алгоритмов кластеризации. Заметна нестабильность K-Means (скачки).

Соотношение по двадцати английским запросам (запросы упорядочены по убыванию количества тем в Википедии; горизонтальная ось – ось номеров запросов, вертикальная – ось количеств тем)

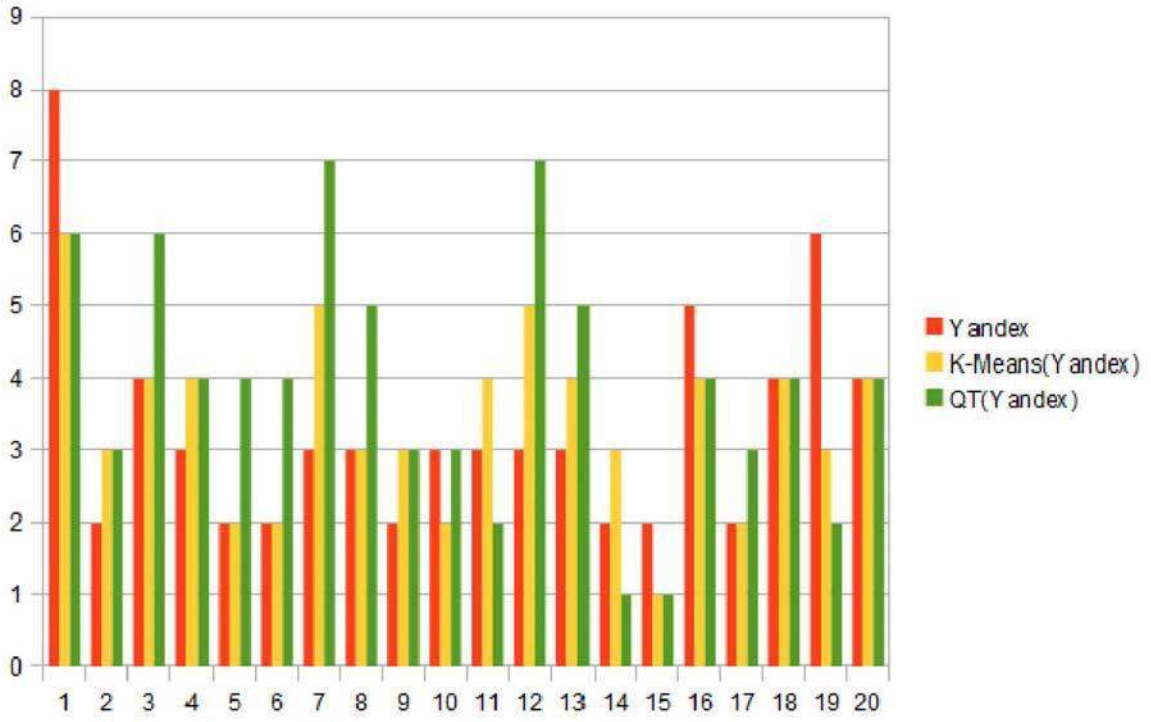


То же самое, но без столбцов по Википедии



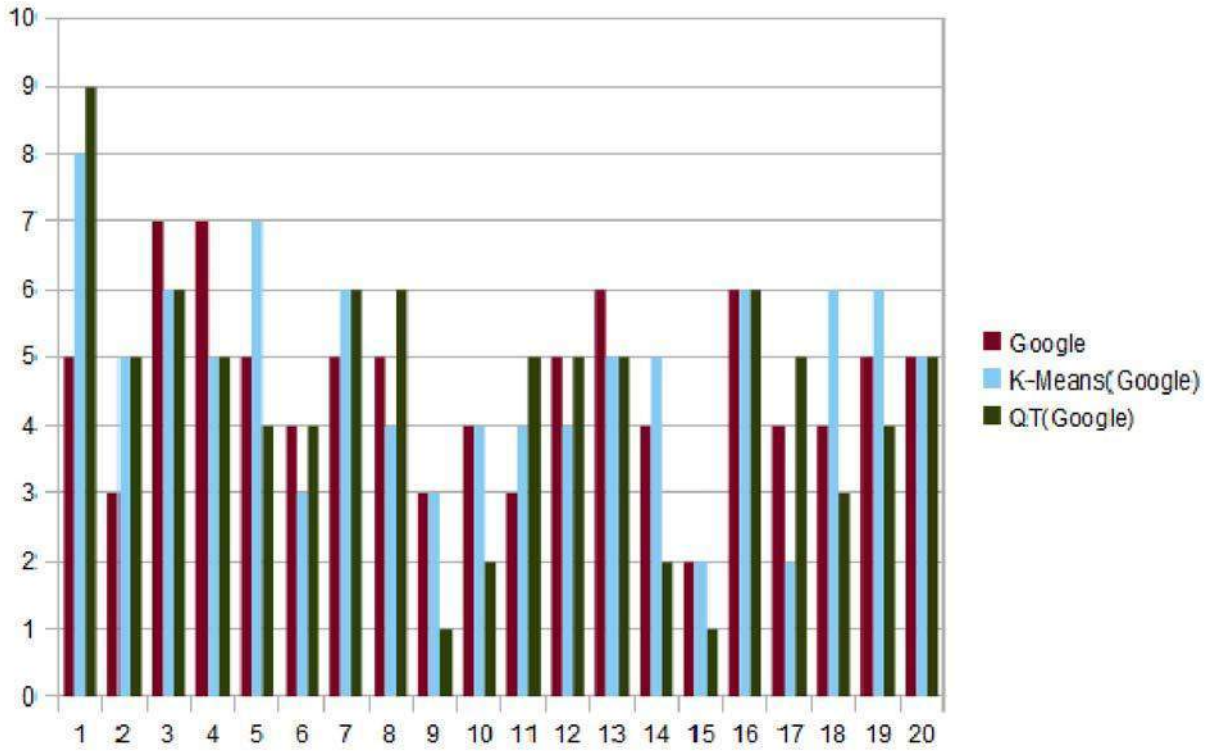
На диаграмме видно, что столбцы, соответствующие методам кластеризации, в основном выше, чем столбцы, соответствующие обыкновенным ответам. Также можно заметить, что столбцы, соответствующие QT, тоже в среднем немного выше, чем столбцы K-Means.

Только Яндекс (англоязычные запросы)



Здесь видно то улучшение, которое обеспечивает QT.

Только Google (англ. Запросы)



Здесь улучшение не столь очевидно, но, всё же, для большинства запросов выигрывает один из алгорит-мов кластеризации.

4 Заключение

В работе проведено исследование разнообразия результатов веб-поиска и возможности его улучшения с помощью кластеризации. Было выяснено, что кластеризация позволяет повысить количество тем в ответах по англоязычным запросам рассмотренных поисковых систем. При этом сравнивались два алгоритма кластеризации: Quality Threshold и K-Means, причем количество кластеров для алгоритма K-Means подбиралось на основе данных, полученных с помощью спектрального анализа результатов по-иска.

Было также обнаружено, что алгоритм QT действительно помогает существенно повысить разнообразие результатов, в то время как повышение разнообразия от применения K-Means было менее заметно (но, всё же, это было улучшение). В случае же русскоязычных запросов кластеризация не дала положительных результатов ни для Яндекса, ни для Google. Также было проведено сравнение качества выдачи поисковых систем Яндекс, Google и Yahoo с точки зрения разнообразия и выявлено, что по английским запросам лидирует Yahoo, за ним идет Google, а последний – Яндекс, а по русским – лидирует Google, за ним с небольшим разрывом – Яндекс, последний – Yahoo.

В дальнейшем планируется увеличить количест-во рассматриваемых поисковых систем и методов кластеризации, а также повысить качество класси-фикации документов.

Литература

[1] Heyer L.J. et al. Exploring expression data: identification and analysis of coexpressed genes// Genome Research. – 2009. – V. 9. – P. 1106-1115.

[2] Clarke C.L.A. Kolla C.M., Cormack G.V., Vechtomova O., Ashkan A., Büttcher S., MacKinnon I. Novelty and diversity in information retrieval evaluation. – University of Waterloo.

[3] von Luxburg U. A tutorial on spectral clustering// Statistics and Computing. – 2007. – V. 17, No 4.

[4] Ng A.Y., Jordan M.I., Weiss Y. On spectral clustering: analysis and an algorithm//NIPS 14, 2001.

[5] Robertson S.E., Walker S., Hancock-Beaulieu M. Okapi at TREC-7//Proc. of the Seventh Text REtrieval Conf., Gaithersburg, USA, November 1998.

[6] Aloise D., Deshpande A., Hansen P., Popat P. NP-hardness of Euclidean sum-of-squares clustering// Machine Learning. – 2009. – V. 75. – P. 245-249.

[7] Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R., Wu A.Y. An efficient *k*-means clustering algorithm: analysis and implemen-tation// IEEE Trans. Pattern Analysis and Machine Intelligence. – 2002. – V. 24. – P. 881-892.

[8] Hartigan J.A. Clustering algorithms. – Wiley, 1975.

[9] Chung F. Spectral graph theory// CBMS Regional Conf. Series in Math., Conference Board of the Mathematical Sciences, Washington, 1997. – V. 92.

[10] Radlinski F., Kleinberg R., Joachims T. Learning diverse rankings with multi-armed bandits. – 2008.

Evaluation of the diversity of web-search results with the help of Wikipedia materials

A.V. Strelkovskiy, I.E. Kuralenok

There are a lot of ambiguous search requests, which can be sent to a search engine by a user. The aim of this work is the evaluation of the quality of the results provided by search engines for such requests, with respect to the amount of topics presented in them. This work also covers the research of a method of improving the diversity of web-search results, which is based on clusterization. The source of the ambiguous concepts and topics used in this work is Wikipedia.