Исследование методов трансформации запросов в первом туре Кубка Яндекса[♣]

Добров Б.В.[1,3] Лукашевич Н.В.[1,3]
Добров Г.Б.[4] Резников Я.[2] Штернов С.В.[1,3]
[1] Научно-исследовательский вычислительный центр
МГУ им.М.В.Ломоносова;
[2] Механико-математический факультет
МГУ им.М.В.Ломоносова;
[3] АНО Центр информационных исследований
[4] Лицей г.Троицка Московской области
{dobroff, louk}@mail.cir.ru, dobrov@ttk.ru,
yanreznikov@yandex.ru, sergey@shternov.ru

Аннотация

Представлен отчет о работе, посвященной исследованию действий игроков в первом туре третьего розыгрыша Кубка Яндекса. Проведен анализ различных параметров действий игроков: временные характеристики, использование языка запросов, стратегий добавления или удаления слов запроса, использование синонимов или связанных по смыслу терминов.

Предложены рекомендации для игроков, пожелания организаторам Кубка, рекомендации по функциональности специального программного обеспечения, «помогающего» играть в данную игру.

♣ Работа поддержана компанией Яндекс, грант № 103039

1. Введение

Для работы глобальных поисковых машин характерно обслуживание миллионов пользовательских запросов в день по широкому кругу запросов. Это предъявляет жесткие требования к выбору технических решений: усложнение обработки запросов может приводить к увеличению нагрузки на поисковую машину и к задержке в обслуживании других пользователей.

В результате сложилось некоторое неформальное «соглашение» между пользователями и поисковыми машинами:

- с одной стороны, поисковые машины достаточно успешно справляются с абсолютным большинством простых запросов пользователей;
- с другой стороны, пользователи не ожидают от поисковой машины чуда и стараются обращаться к ней с достаточно простыми вопросами.

Отметим, что в реальной жизни, прежде всего в сфере профессиональной деятельности, часто встречаются ситуации, когда пользователю требуется нетривиальный ответ. Например, на вопросы подобные такому – «*Наша организация занимается рыбоводством. Какую ставку НДС - 10 или 20 процентов - мы должны применять, реализуя рыбопосадочный материал: сеголетков, годовиков, личинок, мальков, а также икру карпа, толстолобика, белого амура и карася? ("Главбух", Отраслевое приложение "Учет в сельском хозяйстве", N 2, II квартал 2002 г.)*». Кроме того, часто пользователь затрудняется выразить свою информационную потребность нужными словами.

В случае, если в Интернет нет «готового» ответа, пользователь либо должен применять специальные стратегии использования поисковых машин – раскладывая свой поиск информации на составные части, либо обращается за советом к человеку-специалисту в специализированные форумы и т.п.

Разработчики поисковых машин понимают такую ситуацию, ведутся обширные исследования по разработке эффективных методов обработки более сложных запросов пользователей. Важным этапом является реализация такого класса задач, как поиск ответов на

достаточно простые «фактографические вопросы» типа «Кто жена Била Клинтона?». Согласно исследованию [9] фактографические запросы в системе Altavista в 2001 году в среднем задавались в 5-12% случаев, при том что тематический поиск осуществлялся в 23-30% случаев, примерно в 25% случаев осуществлялся поиск товаров и сервисов.

В центре внимания настоящей работы находится исследование тактики пользователей при ответе на фактографические вопросы в первом туре розыгрыша Кубка Яндекса по поиску в Интернет (http://kubok.yandex.ru). Игроки (мотивированные самим духом соревнования, наличием призов) могут принять участие в одной или нескольких (всего шесть) *«игр»*. В течение каждой игры игроку предлагается ответить на 20 фактографических вопросов типа *«Из какого материала был сделан первый саксофон?»*, *«Какому городу принадлежит звание "Самый благоустроенный город России" за 2001 год?»* и т.п. Игрок должен сообщить правильный ответ и URL страницы, на которой его сможет найти проверяющий действия игрока оргкомитет. Игрок вправе пользоваться любым способом поиска информации, в частности искать ответ в любой поисковой машине.

Существенной особенностью Кубка Яндекса является трехминутное ограничение на поиск ответа на один вопрос. Отметим, что примерно столько времени требуется в среднем для внимательного прочтения одной страницы обычного книжного текста. То есть существует жесткое ограничение на количество страниц, которые может просмотреть игрок при поиске ответа на вопрос.

Авторам не известны публикации, посвященные исследованию Кубка Яндекса. Периодически проводятся исследования поведения пользователей поисковых машин Интернет [5], в рамках которых выясняется влияние тех или иных параметров. В работе [4] исследуется насколько необходимо использовать операторы языка запросов, в [6] поставлен эксперимент для определения отличия в поисковых стратегиях «новичков» и «экспертов» - специалистов в предметной области и в поиске в Интернет.

Существует обширная библиография, посвященная разработке программного обеспечения для автоматического ответа на вопросы фактографического типа. Алгоритм действия таких программ следующий: исходный вопрос трансформируется [8, 9] в один или

несколько запросов к поисковой машине. Из выдачи поисковой машины берутся документы (обычно 100), имеющие максимальный ранг. После чего с использованием специальных процедур производится выбор наиболее подходящих (в дорожке Question Answering конференции TREC – до пяти документов).

Исследование стратегий автоматической трансформации запросов в случае поисковой машины Яндекс для некоторых типов фактографических вопросов выполнено в [1]. Лучшим оказался метод, когда нарастающим итогом суммируются результаты всех возможных простых преобразований исходного вопроса.

2. Идея исследования

Интерес авторов состоит в исследовании стратегий пользователей при поиске ответов на сложные реальные вопросы, результаты предполагается использовать при создании специализированных поисковых машин для помощи в аналитической работе. Исследование стратегий игроков Кубка Яндекса при ответе на фактографические вопросы является одним из этапов данной работы. Определенный интерес представляет создание специализированного программного обеспечения:

- помогающего игроку играть в Кубок Яндекса;
- автоматически «играющем» в Кубок Яндекса.

2.1. Термины и определения

Введем некоторые определения:

- *«Вопрос Кубка»* - вопрос на который, за время не большее чем 3 минуты, должны ответить участники 1го тура Кубка Яндекса, указав URL страницы, содержащей ответ, и лексическую формулу ответа;
- *«Запрос участника Кубка»* - лексическое выражение, направляемая участником в поисковую машину;
- *«Успешный ответ»* - ответ, признанный правильным в соответствии с правилами изложенными в http://kubok.yandex.ru/rules.xhtml;
- *«Успешная серия запросов»* - серия запросов игрока в поисковую машину Яндекс, в результате которой игрок послал успешный ответ;

- *«Неудачная серия запросов»* - серия запросов игрока в Яндекс, в результате которой игрок либо сообщил системе неверный ответ, либо не сообщил никакого, в том числе из-за опоздания;
- *«Успешный запрос» (последний успешный запрос)* - будем называть таким запросом последний запрос игрока в успешной серии запросов.

2.2. Общий первоначальный план исследования

Для достижения поставленных целей были запланированы следующие исследования:

- провести исследование и классифицирование методов трансформации запросов в 1м туре Кубка Яндекса путем проведения анализа тактики участников, которые пользовались для ответов поисковой машиной Яндекс;
- для измерения меры смысловой близости замен исследовать возможность применения тезауруса русского языка РуТез [2] (45 тысяч понятий, 110 тысяч текстовых входов, 360 тысяч прямых отношений между понятиями и 1,900 тысяч наследуемых отношений), рассматривая синонимические замены, замены по иерархии отношений;
- выработать рекомендации по функциональности программного обеспечения, «помогающего» играть в Кубок Яндекса.

3. Описание методов, алгоритмов и экспериментов

3.1. Входные данные и первичная обработка

Для выполнения намеченного плана исследований минимально необходимы следующие исходные данные:

- формулировки вопросов Кубка;
- журнал запросов игроков к поисковой машине;
- ответы игроков, с классификацией на правильные и неправильные.

Компанией Яндекс для выполнения настоящего исследования был любезно предоставлен набор данных «Кубок Яндекса», содержащий в текстовых файлах минимально необходимый комплект данных для

шести игр третьего розыгрыша Кубка Яндекса (http://kubok.yandex.ru/3/).

Требует комментария предоставление в наборе исходных данных «лемматизированного образа» запроса игроков (*«нормализованный запрос: упорядоченный по алфавиту список всех лемм запроса (для слов-агнонимов строятся гипотетические леммы), без снятия омонимии»*). Лемматизированный образ запроса, судя по всему, не представляет собой результата преобразования запроса для выполнения поиска Яндексом (хотя в него не включаются стоп-слова – предлоги и вопросительные слова), так как лемматизируются и элементы запроса, заключенные в кавычки (условие на поиск по точной словоформе).

3.1.1. Конвертирование и очистка данных

Текстовые файлы исходных данных были конвертированы в таблицы Paradox, дальнейшая работа с данными велась с использованием SQL. Помимо нормализации исходных данных каждому запросу была проставлена пометка использования языка запросов Яндекса.

Яндекс при подготовке исходных данных произвел синхронизацию по времени IP адресов игроков, посылающих ответ, с IP-адресами пользователей поисковой машины. Однако, с одного IP-адреса помимо игроков в Кубок Яндексом одновременно могли пользоваться и другие люди, например, использующие тот же прокси-сервер. Мы применили достаточную простую процедуру очистки, состоящую в том, что оставались только запросы, которые имеют пересечение хотя бы по одному слову либо с формулировкой вопроса, либо с одним из ответов, либо с последними запросами игроков, ответивших правильно (служебные слова игнорировались). Процедуру планировалось применять итерационно, расширяя множество допустимых строк, но было оценено, что желаемые результаты были достигнуты уже после первой итерации. В результате, далее мы анализировали 63136 запросов (в исходном файле 75221 запрос), выполненных в Яндекс 897 игроками.

3.1.2. Группирование игроков

Согласно правилам проведения Кубка Яндекса итоговым для игрока признается максимальный результат, достигнутый игроком в любой

из игр. В следующий тур Кубка прошли игроки, набравшие 12 и более очков.

Одним из направлений нашего исследования было попытаться выяснить особенности тактики игры лучших игроков. Поэтому мы произвели деление всех игроков на пять групп в соответствии с величиной достигнутого рекорда (Таблица 1). Кроме того мы переупорядочили игроков по убыванию рекорда и проценту правильных ответов на сделанные запросы, так что игроки с номерами 1-52 представляли группу А, 53-135 группу Б и т.д.

Группа	Рекорд	Кол-во игроков	Номера игроков
А	14-18	52	1-52
Б	12-13	83	53-135
В	10-11	89	136-224
Г	7-9	185	225-409
Д	0-6	388	410-897
Всего		897	

Таблица 1. Распределение игроков по группам

3.2. Общие замечания

3.2.1. Комментарий о поисковой машине Яндекс

В Таблице 2 сведены данные о результатах, достигнутых игроками, использовавшими или не использовавших Яндекс. Как нетрудно видеть, подавляющее большинство ответов получено с использованием Яндекс. При этом другими поисковыми машинами скорее пользовались игроки групп А и Б чем остальных. В целом 842 игрока из 897 использовали Яндекс (небольшое количество использовали и Яндекс и другие машины поиска).

Отметим некоторые особенности использования поисковой машины Яндекс при ответах на вопросы Кубка. Важно, что ответ на запрос формируется Яндексом очень быстро, так что игрок получает ответ в течение 2-4 секунд. На первой странице результатов выдается 10 сниппетов – ссылок на релевантные по мнению программы документы, снабженные контекстно-зависимыми аннотациями, в которых слова запроса подсвечиваются. Имеется возможность перехода на копию документа (в этом нет уверенности, возможно

документ подгружается на лету) из базы Яндекса с подсвеченными найденными словами.

группа	способ	Игры						всего
		1	2	3	4	5	6	
А(14)	Яндекс	3	2	3	2	4	0	12
	Всего	4	2	3	3	5	0	14
Б(12)	Яндекс	33	34	19	35	39	31	36
	Всего	41	35	23	47	45	38	52
В(10)	Яндекс	58	52	48	75	76	80	201
	Всего	71	60	58	83	86	8	224
Г(7)	Яндекс	137	119	107	147	127	136	370
	Всего	160	127	119	161	142	148	409

Таблица 2. Использование поисковых машин во время 3го Кубка

Известно, что Яндекс, получая запрос, производит его трансформацию и только после этого его выполняет. Однако, существует крайне ограниченное количество точных данных о том, как же собственно Яндекс преобразует запрос. Согласно [3], при вычислении запроса не требуется наличие в возвращаемом документе всех слов запроса (достаточно чтобы количество найденных слов превышало некоторое число – кворум, нелинейно растущее от длины запроса). Более высоко оцениваются документы, где найденные слова запроса находятся близко друг к другу. Также известно, что применяются специфические трансформации запросов фактографического типа «Что такое А?» в «А это». Кроме того, используется отсечение стоп-слов (однако неясно, в любом контексте или нет), анализируя пары запрос – «нормализованный запрос» можно примерно определить список таких стоп-слов для дальнейшей оценки различия лексически разных запросов с точки зрения Яндекса.

3.2.2. Временные характеристики действий игроков

Для лучшего представления предмета исследования приведем наиболее характерные временные характеристики о действиях игроков Кубка.

На рисунках 1 и 2 представлены типичные временные данные двух игроков. На Рис.1 игра лучшего игрока (рекорд 18 очков), в которой данный игрок набрал 15 очков. На Рис.2 лучшая игра игрока с №87 (рекорд 13 очков). По горизонтальной оси отложено время с начала

игры (каждые 180 секунд соответствуют одному вопросу). Ответ обозначен тонкой линией с символом «+» на верху. По вертикальной оси отложены номера запросов для ответа на один вопрос. Более широкие и темные линии обозначают запросы успешных серий. Более тонкие – запросы неудачных серий – после которых последовал либо неправильный ответ, либо ответа не последовало. Ромбиками над соответствующими запросами обозначено использование игроком языка запросов в успешной серии, аналогично треугольниками обозначено использование языка запросов в неудачных сериях.

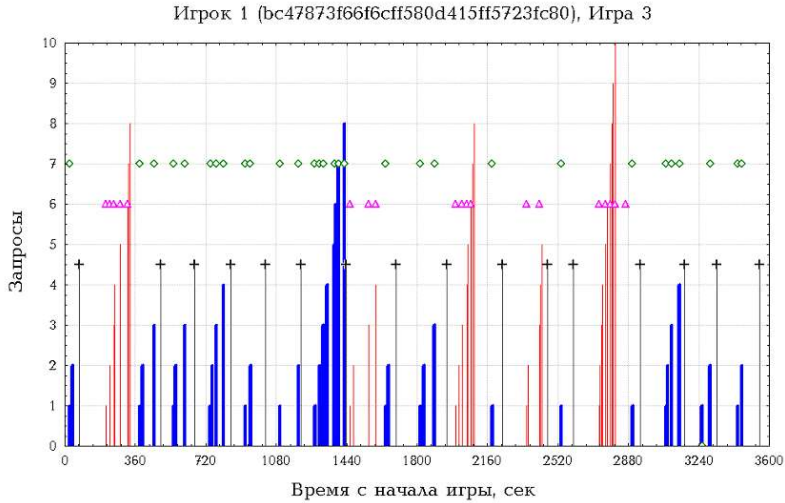

Рис.1. Временные характеристики игры лучшего игрока 3го Кубка (рекорд = 18, в данной игре набрано 15 очков)

Отметим значительное большее число запросов в неудачных сериях, что понятно, так как игрок старается подобрать хороший запрос. Паузы между некоторыми запросами и между последним запросом и ответом могут быть объяснены переходом игроков на страницы документов для поиска окончательного ответа на запрос. Малое время между временем последнего запроса и временем ответа может объясняться нахождением игроком ответа среди сниппетов.

Временные характеристики времени выполнения запросов отображены на рисунке 3. Можно видеть, что первый запрос направляется в поисковую машину в среднем за 20-40 секунд, при

этом есть пауза около 10 секунд, по-видимому, время, требующееся для осознания сущности вопроса. Следующие запросы формулируются в среднем значительно быстрее.

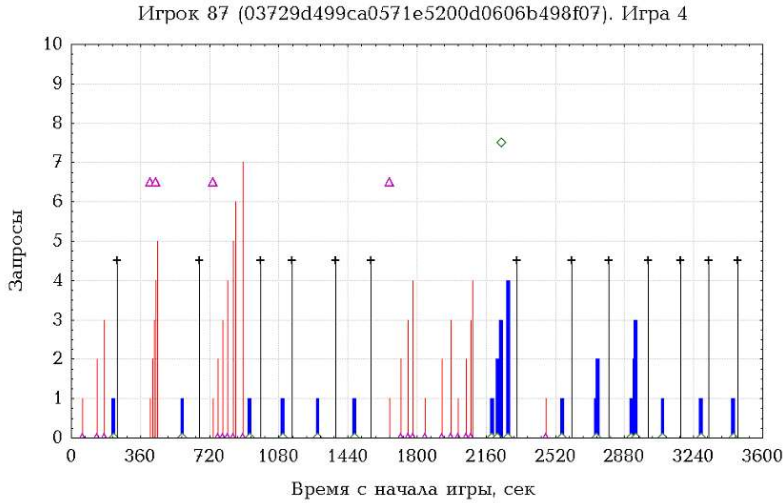

Рис.2. Типичный временной сценарий игры
(лучшая игра игрока с рекордом 13 очков)

На рисунке 4 отражена разница во времени между последним запросом серии и временем посылки ответа игроком. Планируя свои исследования, мы ожидали возможное выделение двух локальных максимумов, соответственно для ответов по сниппетам без перехода на просмотр страницы и для ответов с переходом на страницу. Однако двух пиковой структуры не наблюдается.

Отметим наличие незначительного количества запросов, которые продолжали посылаться в поисковую машину уже после посылки ответа, что требует определенного критического отношения к сделанному нами выбору понимания успешного запроса, как последний запрос успешной серии запросов В рамках настоящей работы мы пренебрегаем этим эффектом – можно было бы выбрать таковым последний запрос перед посылкой ответа, однако, утверждать, что этот подход будет значительно точнее, невозможно.

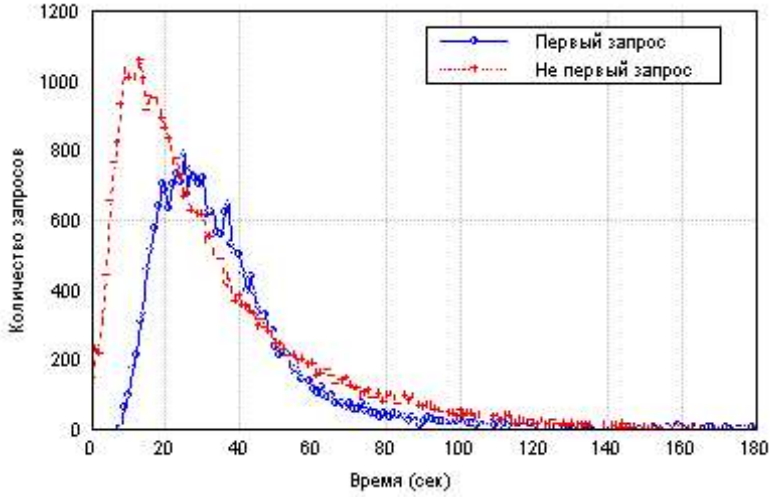

Рис.3. Распределение времени при выполнении запросов

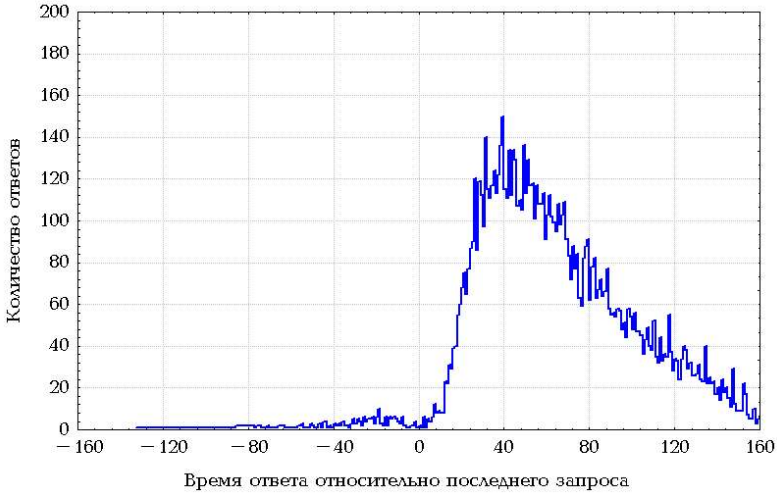

Рис.4. Распределение времени ответа после последнего запроса

3.3. Личные качества игроков

Прежде всего рассмотрим наличие зависимости между более удачливыми игроками и остальными, так как возможна точка зрения, что все игроки ищут одинаково, но некоторые просто ищут значительно быстрее других.

В таблице 3 приведены усредненные временные показатели игроков. Можно сделать вывод, что более успешные игроки являются и в среднем более быстрыми – успевают задавать больше количество запросов при ответе на вопрос. Более успешные игроки и более «трудоспособные» - пытались ответить на большее количество вопросов. Таблица 3 также подтверждает закономерность, что в неуспешных сериях игроки задают в полтора раза больше запросов чем в успешных.

группа	Всего		Успешные серии		Неуспешные серии	
	серий на игрока	запро-сов в серии	серий на игрока	запро-сов в серии	серий на игрока	запро-сов в серии
А(14)	60,2	2,92	37,1	2,30	23,1	3,92
Б(12)	61,6	2,71	32,1	2,08	29,5	3,39
В(10)	44,5	2,63	20,0	1,98	24,5	3,16
Г(7)	32,5	2,43	11,5	1,79	21,0	2,78
Д	18,8	2,08	3,8	1,52	15,0	2,22

Таблица 3. Статистика по игрокам разных групп по сериям и количеству запросов в серии

При проведении любого конкурса организаторы заботятся о двух моментах: с одной стороны, задания должны быть достаточно сложными, чтобы выявить действительно достойных победителей, с другой стороны, необходимо создать «кворум» участников, чтобы было из кого выбирать и выбор был более обоснованным. Нам неизвестна точная мотивация оргкомитета проведения Кубка Яндекса, однако, общий принцип проявляется в наличии вопросов разного уровня сложности (Рис.5).

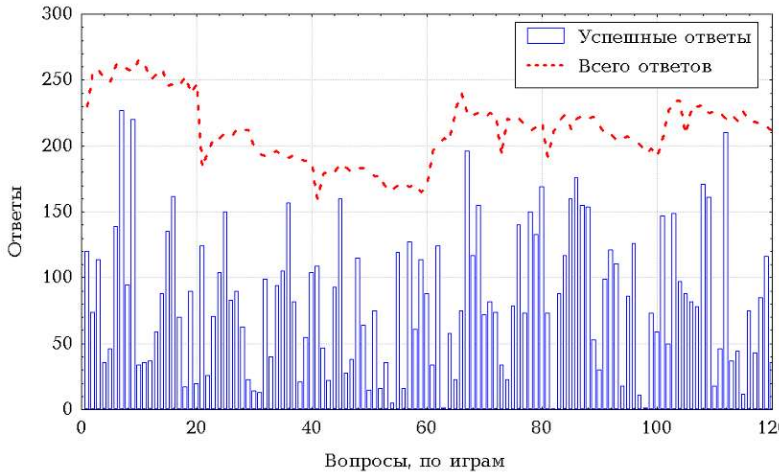

Рис.5. Распределение правильных ответов по вопросам

В Таблицу 4 сведены данные о вопросах, которые оказались самыми трудными для игроков. Как нам представляется, такими вопросами оказались прежде всего те, для которых:

- заранее трудно предсказать ответ (иногда игрок знает/предугадывает ответ заранее и затем ищет только соответствующую страницу, уточняющую ответ);
- даже получив нужную страницу трудно, вычленить правильный ответ в специальном тексте в условиях ограниченного времени;
- вопрос является составным - для ответа на вопрос требуется сначала найти ответ на часть вопроса, затем подставить найденное для поиска ответа на вторую часть (и весь вопрос), например, вопрос про «хока».

В таблице 5 приведены данные по группам игроков для выделенных «трудных» вопросов. В предположении, что более успешные игроки (группы А и Б) представляют собой больших экспертов в пользовании Интернет, чем остальные, результаты Таблицы 5 согласуются с выводами работы [6].

Иг-ра	№ вопроса	Ответили, %	Формулировка вопроса
5	2	0,0	Как в 17 веке в России назывались профессиональные изготовители водки в аптеках?
4	3	0,4	Какие группы крови отца и матери дают максимальную вероятность несовместимости крови матери и плода (в цифровой записи)?
5	18	0,5	В какой телепередаче на телевидении прошел первый телесеанс Кашпировского?
3	14	2,9	Какая трасса "Формулы-1" в 2001 году состояла из наименьшего числа кругов?
6	15	5,3	Сколько метров от бортика до бортика было в разрушенном бассейне Москва?
5	17	5,4	Сколько тонн весил камень, водруженный в начале семидесятых любителями русской поэзии на месте сгоревшей усадьбы Блока?
2	11	6,7	Сколько миллилитров в 10 "американских столовых ложках"?
1	18	6,7	Что находится между 26 и 32 на предмете, который до появления в Европе назывался "хока"?
2	10	7,0	Какая профессия была у победителя первой гонки Тур де Франс?
6	10	7,9	Какой телефон у генерального директора ОАО "Красноярская ГЭС"?
1	20	8,0	Как называется самый длинный туннель на транскавказской автомобильной магистрали?
3	10	8,3	Для чего при игре на оргАне требуется зеркало заднего вида?
5	14	8,7	Откуда родом был рабочий, написавший письмо, опубликованное в газете "Правда" к 27 съезду КПСС, в котором впервые был поднят вопрос о социальном неравенстве и привилегиях руководителям?
3	12	9,0	Как зовут Петрушку в Индии?
3	16	9,3	Как называется объединение старых университетов Америки, возникшее на основе спортивной игры?

Таблица 4. Наиболее трудные вопросы третьего Кубка

Отметим, интересный факт – мера ответственности более успешных игроков «выше» - они реже сообщают неправильный ответ.

Последний столбец показывает долю использования поисковой машины Яндекс при ответе на сложные вопросы.

группа	всего ответов	правильных ответов	процент прав.	серий запросов в Яндекс	успешных серий	процент успешных	исп. Ya, %
А (14)	145	83	57,2	381	70	18,4	84,3
Б(12)	134	53	39,6	602	48	8,0	90,6
В(10)	116	45	38,8	496	32	6,5	71,1
Г(7)	122	24	19,7	737	16	2,2	66,7
Д	149	13	8,7	842	11	1,3	84,6

Таблица 5. Ответы игроков на 15 трудных вопросов

Таким образом, имеется корреляция между личными качествами игроков (опыт работы в Интернет, общая эрудированность) и результатами, достигнутыми в Кубке. Однако, вклад такого рода факторов не велик – имеется достаточно много сравнительно простых вопросов, правильно ответить на которые может любой игрок.

3.4. Исследование пословных трансформаций запросов

3.4.1. Простейшая стратегия

Очевидной простейшей стратегией игроков является прямое повторение формулировки вопроса в качестве запроса к поисковой системе Яндекс.

Существуют обоснования такого подхода:

- зная, что Яндекс производит автоматически усечение стоп-слов не надо тратить время на их ручное удаление;
- это соответствует общей рекомендации сначала выполнить наиболее специфический запрос – с максимальным количеством слов, а затем расширять запрос, анализируя контекст результатов;
- не исключено, что Яндекс автоматически использует шаблоны замещения для вопросительных конструкций.

На рисунке 6 показана степень успешности данной стратегии. Для получения рисунка запросы были объединены по лемматизированным образам, то есть если игрок самостоятельно удалял в запросе по сравнению с формулировкой вопроса только стоп-слова, то запросы считались эквивалентными. По нашим данным в 75 вопросах из 120 игрокам удавалось находить правильный ответ, направив в Яндекс формулировку исходного вопроса. При этом в 20 вопросах такой стратегии придерживалось более 10 процентов игроков (для 4 вопросов более 40% игроков).

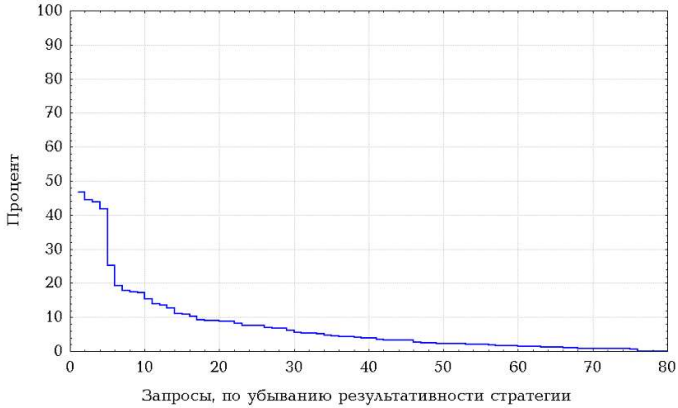

Рис.6. Стратегия повторения вопроса в запросе

Отметим важное обстоятельство, определяемое, по нашему мнению, ограниченностью временем, отводимого для выполнения задания. Несколько игроков в течение выполнения задания направляют в систему одинаковый запрос (предполагаем, что они получают и одинаковый ответ системы). Однако только часть игроков находит правильный ответ.

На рисунке 7 данный тезис иллюстрируется для стратегии повторения исходного вопроса в запрос на материале первой игры Кубка. Темными широкими прямоугольниками обозначено количество успешных запросов, совпадающих с формулировкой вопроса. Более узкими прямоугольниками обозначено количество запросов из успешных серий, совпадающих с исследуемым (включает предыдущие). Отличие в высоте этих прямоугольников от предыдущих говорит о том, что игрок использовал такой запрос, но затем выбрал для окончательного ответа другой. Самыми узкими и

светлыми прямоугольниками обозначено общее количество запросов в розыгрыше вопроса, совпадающих с анализируемым. Где данный прямоугольник выше – значит игрок спрашивал данный запрос, но не нашел правильного ответа.

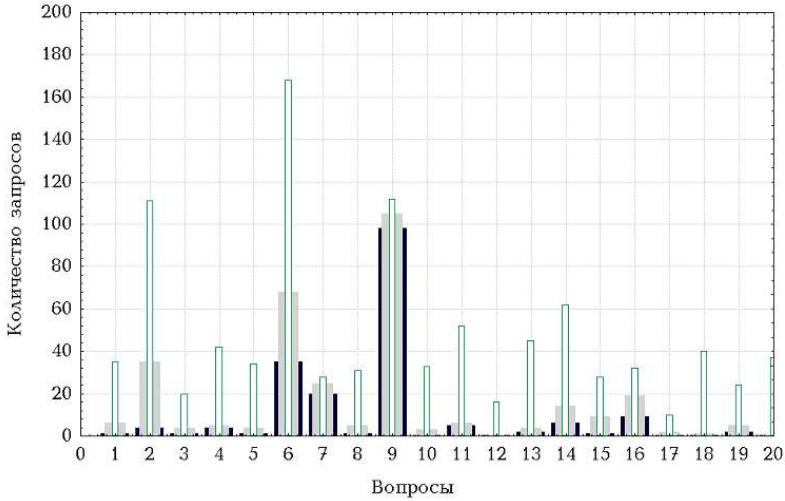

Рис.7. Стратегия повторения формулировки вопроса в запросе

3.4.2. Удаление и добавление слов

Представим результаты суммарного анализа состава слов, удаляемых игроками для получения успешного запроса (Таблица 6, слова откинутые только в 1 запросе и предлоги не рассматривались).

Как можно видеть, практически любое слово исходного вопроса может быть отброшено. Однако наиболее часто игроками отбрасывались:

- вопросительные слова (возможно, что они автоматически отбрасываются и поисковой машиной);
- формы глаголов-связок (*быть*, *являться* и т.п.), модальные глаголы (*должен*);
- глаголы, которые могут сочетаться с большим количеством существительных (*произойти*, *сделать*, и т.п.).

В какой день недели произошло Ледовое побоище?				*Из какого материала был сделан первый саксофон?*			
1	4	какой	28	1	13	какого	54
1	4	день	26	1	13	материала	54
1	4	недели	26	1	13	был	31
1	4	произошло	25	1	13	сделан	24
1	4	ледовое	4	1	13	первый	12
1	4	побоище	4	1	13	саксофон	7
…				…			

Таблица 6. Примеры отброшенных игроками слов для некоторых вопросов

В какой день недели произошло Ледовое побоище?				*Из какого материала был сделан первый саксофон?*			
1	4	суббота	10	1	13	сакс	7
1	4	дата	6	1	13	металла	4
1	4	апреля	5	1	13	изготовлен	3
1	4	5	4	1	13	82	2
1	4	ледового	4	1	13	адольф	2
1	4	побоища	4	1	13	адольфом	2
1	4	вторник	3	1	13	история	2
1	4	понедельник	3	1	13	саксофона	2
1	4	пятница	3	1	13	чего	2
1	4	среда	3	…			
1	4	четверг	3				
1	4	воскресенье	2				
…							

Таблица 7. Примеры добавленных игроками слов для некоторых вопросов

Дальнейшие стратегии отбрасывания слов, по нашему мнению, связаны с представлением игроков о сочетаемости слов вопроса в контексте ожидаемого ответа. Например, трудно предположить предложение текста *«…саксофон сделан из … материала …»*, скорее будет *«…саксофон сделан из X»* (что не исключает повторения вопроса в каком-нибудь заголовке). Отметим, что для собственных имен чаще наблюдается стратегия сохранения фамилии, отбрасывая имя.

В Таблице 7 представлены наиболее частые новые слова в успешных запросах по сравнению с формулировкой вопроса.

Игроками широко используются морфологические варианты (по-видимому, игроки перенабирают запрос руками, выбирая при этом нормализованную форму), словообразовательные варианты (*весить – вес*, *бейсбол – бейсбольный*).

Любопытна стратегия «подбора» правильного ответа, которая может состоять:

- в угадывании лексической последовательности в ответе (последовательность может учитываться поисковой машиной). Например, *«саксофон сделан из»*;
- в переборе вариантов конечного списка вариантов ответа. Например, замечательный запрос *«ледовое побоище (понедельник | вторник | среда | четверг | пятница | суббота | воскресенье)»*.

Некоторое количество запросов свидетельствует, что игрок сначала догадывается об ответе (получает его на предыдущих запросах), но затем продолжает искать, по-видимому, либо подтверждение ответа, либо более удачные страницы (иногда ссылки с выдачи поисковой машины ведут на «мертвые» страницы).

Отметим, что особенность Яндекса выдавать документы, содержащие не все слова запроса позволяют стать успешными запросам с ложной гипотезой, например, «*ледовое побоище вторник»* (правильный ответ – «суббота»).

3.5. Использование языка запросов

В практически любой поисковой машине имеется специальный язык запросов, при анализе использования которого при Интернет-поиске обычно рассматривают основные логические операции, операцию, задающую требование поиска на точное совпадение словоформ фразы. В случае Яндекса можно задавать допустимое расстояние между словами запроса, выбор режима поиска – в рамках предложения (по умолчанию) или по документу, поиск по зонам документа и т.п. (подробнее см. www.yandex.ru/info/syntax.html).

Существуют разные ответы [4] на вопрос «Помогает ли использование операторов при поиске информации в поисковой машине при поиске в Интернет?». Посмотрим, что можно сказать для случая игроков Кубка Яндекса (Таблица 8).

- язык запросов в среднем использовался достаточно равномерно вне зависимости от успехов игроков;
- процент использования языка запросов был больше, чем описано в литературе [4, 5] при поиске в Интернет;
- применение языка запросов не давало преимущества по сравнению с неиспользованием языка запросов;
- лучшие игроки применяют язык запросов с большей эффективностью, чем игроки в среднем.

Группа	Процент запросов с ЯЗ	Процент успешных запросов с ЯЗ
А(14)	20,1	16,9
Б(12)	17,6	15,2
В(10)	18,0	14,2
Г(7)	17,3	14,6
Д	18,0	11,8

Таблица 8. Использование языка запросов (ЯЗ)

3.6. Оценка использования семантического преобразования запросов

Авторы работы имеют возможность автоматически определить в текстах вопросов и запросов текстовые входы и соответствующие им понятия тезауруса русского языка РуТез [2]. Для этого были образованы текстовые файлы, включающие формулировку вопроса и формулировку успешного запроса. Результатом сопоставления с РуТез является список понятий с указанием текстовых входов, найденных в формулировке вопроса и запросе, а также список связанных понятий с учетом иерархии тезауруса.

3.6.1. Использование синонимов

Представление об использовании синонимов в успешных запросах дает содержимое Таблицы 9. Как можно видеть, использование синонимических замен достаточно распространено, всего примерно в полтора раза реже, чем использование языка запросов, при этом имеет аналогичную зависимость в среднем от успехов игрока.

Группа	Успешные запросы, в которых использовались синонимические замены, %
А(14)	11,3
Б(12)	10,9
В(10)	11,5
Г(7)	8,1
Д	7,1

Таблица 9. Использование синонимических замен в успешных запросах

В Таблице 10 приведены примеры использованных игроками синонимических замен.

Текстовый вход вопроса	Текстовый вход запроса
ТЭТЧЕР	*МАРГАРЕТ ТЕТЧЕР*
МИФОЛОГИЯ	*МИФОЛОГИЧЕСКИЙ*
ИМПЕРАТОР ЯПОНИИ	*МИКАДО*
ПРИДНЕСТРОВСКИЙ	*ПРИДНЕСТРОВЬЕ*
САМОХОДКА	*САМОХОДНЫЙ*
ЛАТИНСКИЙ	*ЛАТЫНЬ*
ВРУЧАТЬ	*ВРУЧЕНИЕ*

Таблица 10. Примеры синонимических замен

3.6.2. Трансформация запросов по смыслу

В Таблице 11 приведены данные об использовании игроками в запросах понятий, связанных отношениями в тезаурусе РуТез. Отношения, содержащиеся в самом вопросе, не учитывались. Проведена классификация по типам использованных отношений. Пример приведен в Таблице 12.

Смысловые замены использовались игроками всех групп. Анализ показывает, что большинство замен представляли собой простые лексические замены, что определялось как лексически связанные термины. То есть замена «*шариковая ручка*» на «*ручка*», определялось как пара *ШАРИКОВАЯ РУЧКА выше РУЧКА (ПРИНАДЛЕЖНОСТЬ ДЛЯ ПИСЬМА)*.

группа	Всего успешных запросов, исп. замены по смыслу, %	из них сужение запроса, %	из них расширение запроса, %	из них отношение ассоциации, %
А	22,3	43,1	55,3	7,7
Б	21,8	52,7	43,4	11,2
В	19,1	36,2	59,4	12,9
Г	18,6	36,3	57,7	12,3
Д	15,1	34,5	59,5	12,7
	20,9	39,6	56,2	11,1

Таблица 11. Использование в запросах замен, связанных с понятиями вопроса разными типами отношений

Запрос	Понятие вопроса	Понятие запроса
ручка изобретатель	*ЧЕЛОВЕК*	*ИЗОБРЕТАТЕЛЬ*
ручка изобретатель	*ШАРИКОВАЯ РУЧКА*	*РУЧКА (ПРИНАДЛЕЖНОСТЬ ДЛЯ ПИСЬМА)*

Таблица 12. Преобразование вопроса
«Как звали человека, сконструировавшего шариковую ручку?»

4. Выводы и обсуждение результатов

Большая часть работы была направлена на исследование различных параметров действий игроков Кубка Яндекса, выделения из них наиболее важных.

1) Проведено исследование использования трансформаций запросов различных типов по группам игроков в зависимости от достигнутого результата в Кубке:

- использование языка запросов;
- использование синонимических замен;
- использование смысловых замен.

В целом, лучшие игроки успешнее используют указанные трансформации (на 20% по сравнению со средним уровнем), но это не является определяющим для выигрыша в соревновании.

2) Самым важной характеристикой розыгрыша Кубка Яндекса является, по нашему мнению, трехминутное ограничение времени

ответа на вопрос. Часто встречается ситуация, когда игроки направляют одинаковые запросы в поисковую машину, однако только часть игроков замечает правильный ответ. По-видимому, для успеха в Кубке помимо хороших навыков поиска в Интернет большую роль играют такие личные качества, как хорошая общая эрудированность, скорость реакции, внимательность.

3) Большинство вопросов, предложенных игрокам, допускает простые стратегии формирования успешных запросов. Для 75 вопросов (из 120) были игроки, которые находили правильный ответ, по сути, повторяя в запросе формулировку ответа.

В результате нашего анализа выработана достаточно забавная рекомендации игрокам – обнаружилось, что есть небольшое время (около 8-10 секунд), которое тратится всеми игроками на понимание вопроса. В то же время имеются определенные шансы получить правильный ответ, как можно быстрее копируя вопрос Кубка в качестве запроса к Яндекс, а уже после начинать разбираться с сутью вопроса. В случае использования Яндекса, возможно (для уверенного вывода данных недостаточно), лучше не удалять из запроса служебные слова – либо Яндекс их отфильтрует самостоятельно, либо автоматически применит специальную трансформацию запроса.

4) Основной выявленной нетривиальной стратегией игроков является предугадывание возможного ответа:

- отбрасывание тех слов запроса, которые несовместимы со словами, наличие которых в ответе обязательно;
- угадывание ответа непосредственно, в том числе перебирая элементы из возможного ограниченного списка.

5) Наибольшие трудности у игроков вызывали следующие типы вопросов:

- специальные вопросы, когда игроку трудно в сжатое время разобраться является ли найденный по контексту ответ правильным или нет (*«Какие группы крови отца и матери дают максимальную вероятность несовместимости крови матери и плода (в цифровой записи)?»*);
- составные вопросы, которые требовали для получения правильного ответа сначала выяснить часть ответа, а затем

использовать вновь полученное знание для получения окончательного ответа;
- трудности вызывали и вопросы, когда в Интернет невозможно было найти ответ, имеющий лексическое совпадение с одним из элементом множества, получаемого из исходного вопроса набором простых трансформаций.

6) Рекомендации по функциональности специального программного обеспечения, «помогающего» играть в Кубок Яндекса:

- программа должна сокращать время анализа игроком результатов поиска, что подразумевает суммирование результатов запроса, в том числе автоматический анализ страниц, на которые ссылается поисковая машина. Важным является стараться получать контекстно-зависимые аннотации не только в контексте запроса, но и в контексте исходного вопроса;
- программа должна параллельно посылать в поисковую систему несколько запросов, полученных путем простых трансформаций, и выдать игроку суммированные результаты;
- для фактографических вопросов необходимо предусмотреть использование шаблонов замен и связанных с ними алгоритмов поиска наиболее подходящих фрагментов документов (собственные имена, группы времени, локативы и т.п.), ссылки на которые возвращает поисковая машина.

7) Пожелания организаторам:

- в настоящее время нет полной уверенности, что по имеющимся данным удается точно определять момент получения игроком правильного ответа. Для более аккуратного исследования поведения игроков необходимы данные о содержимом страниц выдачи поисковой системы (сниппетах), а также о переходах с этих страниц. Наиболее удобным для проведения анализа могло бы стать журналирование действий игрока на его же компьютере (мотивируя игроков каким-либо способом);
- имеет смысл рассмотреть вопрос о проведении, возможно другого, «соревнования» по несколько другим правилам – когда игроки должны ответить в течение часа на те же 20 вопросов, но без 3х минутного ограничения на один вопрос.

Конечно, при этом теряется определенная зрелищность, увеличивается риск нечестной групповой игры, однако такой регламент представляется более подходящим для моделирования реального поиска ответов на фактографические вопросы;

- желательно опубликовать сведения об используемых в поисковой машине шаблонах трансформации запросов – по сути это дополнительные операторы языка запросов. Иначе трудно сформулировать для игроков уверенные рекомендации относительно удаления тех или иных служебных слов.

5. Литература

1. Бояндин И., Некрестьянов И., Преобразование запросов для предметно-независимого фактографического поиска в Интернет // "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. Труды пятой всероссийской научной конференции – СПб, НИИ Химии СПбГУ, 2003, С.159-168.

2. Лукашевич Н.В, Добров Б.В., Взаимодействие лексики и терминологии в общезначимой сфере языка // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2004 («Верхневолжский», 2-7 июня 2004г.) / Под ред. И.М.Кобозевой, А.С.Нариньяни, В.П.Селегея. – М.:Наука, 2004. – С.172-178.

3. Сегалович И., Маслов М., Яндекс на РОМИП-2004. Некоторые аспекты полнотекстового поиска и ранжирования в Яндекс // Труды второго российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова - Санкт-Петербург: НИИ Химии СПбГУ, 2004, 214 с

4. Eastman C.M., Jansen B.J., Coverage, Relevance, and Ranking: The Impact of Query Operators on Web Search Engine Results – ACM Transactions on Information Systems, V.21, No.4, 2003, p.383-411.

5. Jansen B.J., Spink A., An Analysis of We Searching by European AlltheWeb.com Users. – Information Processing ans Management, 41, 2005, p.361-381.

6. Holscher C., Strube G., Web Search Behavior of Internet Experts and Newbies // Proceedings of the 9th International World Wide Web Conference. – 2000, p.81.

7. Kwok C.C.T., Etzioni O., Weld S.W., Scaling Question Answering to the Web // // Proceedings of the 10th International World Wide Web Conference. – 2001.

8. Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R., Rus V., LASSO: A Tool for Surfing the Answer Net // NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC-8) / Editors: E.M.Voorhees and D.K.Harman - NIST, 1999, pp.175-184.

9. Rose D.E., Levinson D., Understanding User Goals in Web Search // Proceedings of the 13th World Wide Web Conference. – New York, 2003, p.13-19.

Study of Query Transformations in the First Round of Yandex's Cup

Dobrov Boris, Loukachevitch Natalia,
Dobrov Gregory, Reznikov Yan, and Shternov Sergey

In the report we study activities of players in the first round of the third game of Yandex's cup. The goal of the study was to formulate recommendations for development of special software, helping to play the game, and, further, automatically playing this game. We analyse various parameters of players actions: temporal characteristics, use of query language, addings and deletions of query words, use of synonyms and sense-related terms. We propose some recommendations for players, Cup organizers, guidelines for development of special software.