Statistical linguistics in web search

Igor Andreev

Dialog 2015

@mail.ru®

Mail.Ru Group

Internet communications and entertainment services in Russia and globally

Our products

Email and portal	Search and e-commerce	Social networks	Instant messaging	Games
100 mln active accounts	**26,8 mln** users of Search Mail.Ru	**42,6 mln** users of Odnoklassniki	**18,8 mln** users of Agent Mail.Ru in the world	**175,6 mln** downloads of the games in social networks

@mail.ru

Search@Mail.Ru

http://go.mail.ru

Интернет Картинки Видео Новости Обсуждения Ответы

поиск@mail.ru

поиск@mail.ru конденсат Бозе-Эйнштейна

Интернет Картинки Видео Новости Обсуждения Ответы

Конденсат Бозе — Эйнштейна — Википедия
ru.wikipedia.org/wiki/Конденсат_...
Конденса́т Бо́зе — Эйнште́йна (бо́зе-эйнште́йновский **конденса́т**, **бо́зе-конденса́т**) — агрегатное состояние вещества, основу которого составляют бозоны, охлаждённые до температур, близких к абсолютному нулю (меньше
охлаждённом состоянии...
Теория Модель Эйнштейна

поиск@mail.ru let it be

Интернет Картинки Видео Новости Обсуждения Ответы

The Beatles - Let It Be

00:00 / 00:00

Yellow submarine	John lennon - imagine	While my guitar gentl...
Hey jude	Because	All you need is love
Help!	Something	She loves you
Yesterday	Michelle	Oh! darling

Let It Be — Википедия
ru.wikipedia.org/wiki/Let_It_Be
Let It Be («Пусть будет так» —) — тринадцатый и последний студийный альбом группы The Beatles. Выпущен в мае 1970 года, большинство песен было записано в январе 1969, ещё до записи альбома Abbey Road.

@mail.ru

~ 27 mln users monthly

~ 30 mln queries daily

~ 8 bln documents in index

One of the problems is low-freq queries

“Low-freq query” is a query which has no (or very little) information on user’s behaviour

Up to 30% of queries are low-freq

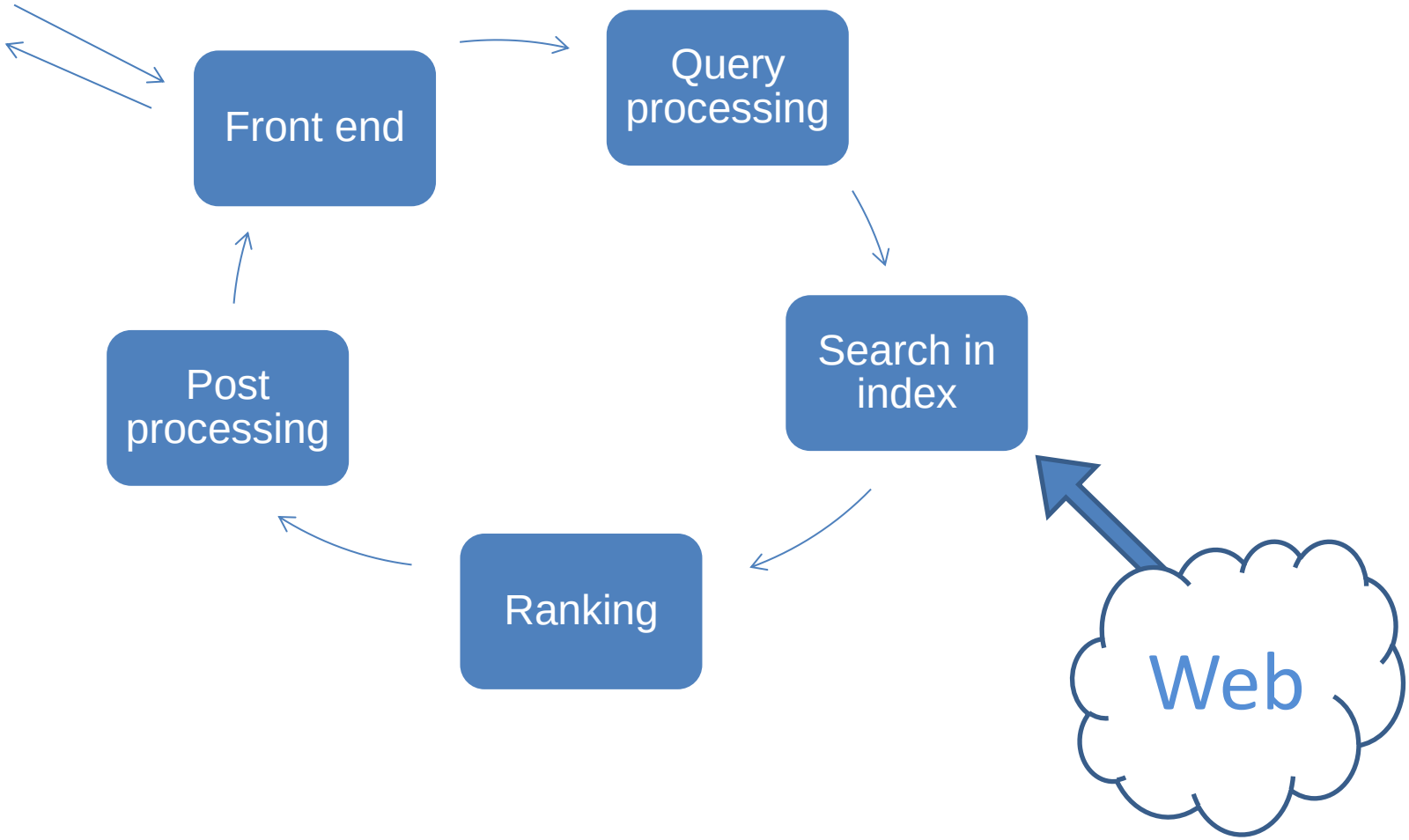
Examples of low-freq queries

- купить елис в минске
- как сделать так чтобы в LEGO гарри поттере 5-7 не было бесконечной загрузки
- к какому виду относится сказка Волк и лиса .Анализ сказки
- акт о выявлении дефектов кровли
- бадьи для зимнего бетонирования в автокад
- видео про контер страйк с лололошкой
- благотворительный фонд хаверим в израиле
- лагер дварец маладежа в Казакстане
- дебилизатор тв смотреть онлайн бесплатно все каналы без регистрации
- как узнать что тебе идет короткая стрижка
- как сплести пупырку без станка на вилках

Q: Why is this a problem?

A: Search engine ranking highly relies on user’s behaviour!

So, we should try to somehow approximate it.


Front end
Query processing
Search in index
Ranking
Post processing
Web

Index

Document 1	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

Document 2	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

Document 3	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

@mail.ru

поиск@mail.ru

query

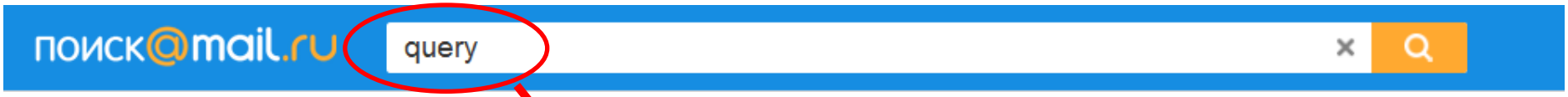
поиск@mail.ru query

1	Document1
2	Document2
3	Document3

поиск@mail.ru query

1	Document1
2	Document2
3	Document3

CLICK!



Index

Document 1	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

Document 2	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

Document 3	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

Index

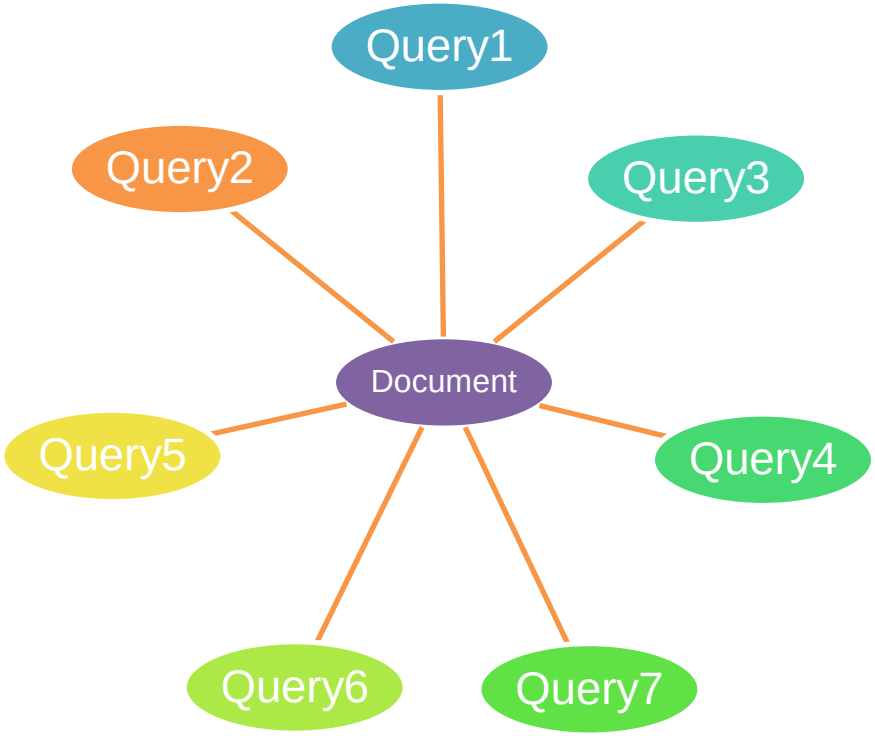
Document 1	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

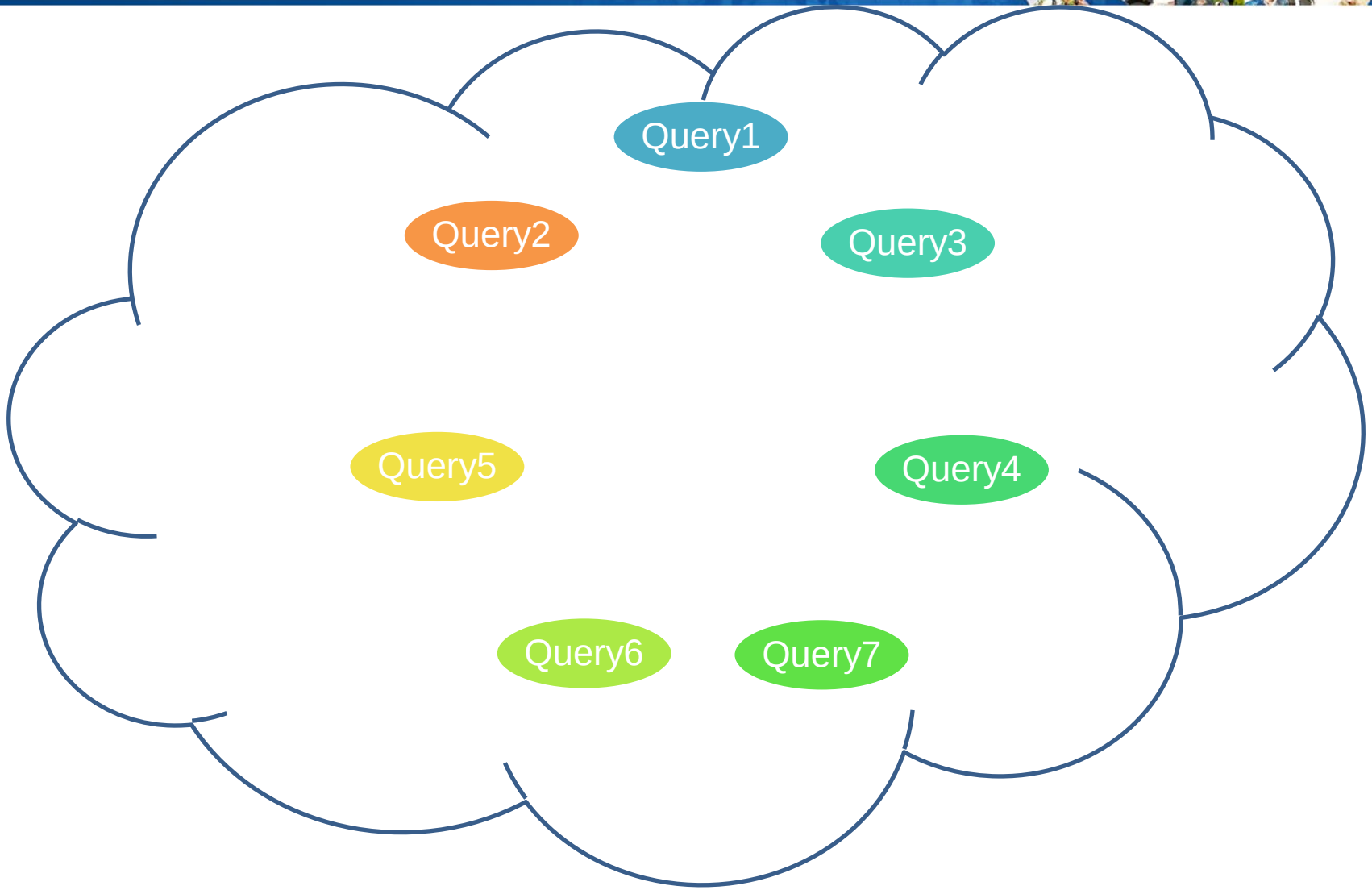
Document 2	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…
Q_links	query

Document 3	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…

поиск@mail.ru query

1	Document2
2	Document1
3	Document3


Query1
Query2
Query3
Document
Query5
Query4
Query6
Query7


Query1
Query2
Query3
Query5
Query4
Query6
Query7

Source	Target
Query1	Query3
Query2	Query5
Query6	Query7

Statistical machine translation

Alignment	
Parallel	corpus

Training stage	
Phrase table (PhT)	Language model (LM)

Translation stage	
Source sentence	Target sentence

тушеная свинина с овощами
как приготовить свинину с овощами
простой рецепт готовим дома рецепт
свинина тушеная
свинина с овощами рецепты
свинина с овощами тушеная
свинина тушенная с овощами
свинина тушеная с овощами картошка
свинина тушеная на сковороде
свинина тушеная с картошкой
свинина с овощами на сковороде
тушёная свинина с овощами
картошка со свининой на сковороде
свинина с овощами рецепт с фото тушеная свинины
свинина тушеная с овощами
рецепт свинины с овощами
как приготовить свинину на сковороде
рецепты свинины с овощами
рецепт приготовления свинины с овощами
как приготовить мясо с овощами
как можно приготовить свинину
рецепт свинины тушеной с овощами
свинина с овощами рецепт
свинина с овощами

«Parallel» corpus

…	…
тушеная свинина с овощами	свинина с овощами тушеная
как приготовить свинину с овощами	свинина тушенная с овощами
свинина с овощами рецепты	свинина тушеная с овощами картошка
свинина тушеная с овощами	свинина тушеная на сковороде
…	…

Currently ~150 mln pairs

...

как приготовить свинину с овощами ||| приготовление свинины ||| 0.0769231

как приготовить свинину с овощами в ||| приготовление свинины в ||| 0.166667

как приготовить свинину с овощами в горшочке ||| приготовление свинины в горшочках ||| 0.333333

как приготовить свинину с овощами на сковороде ||| свинина жареная на сковороде с овощами ||| 0.333333

как приготовить свинину с овощами по китайски ||| свинина с овощами на пару ||| 0.25

как приготовить тушеную свинину с овощами ||| приготовить мясо с овощами ||| 0.0833333

как тушить мясо свинину ||| тушить мясо с овощами ||| 0.25

...

мясо с овощами в духовке в фольге ||| свинина с грибами в фольге ||| 0.166667

мясо с овощами в кисло ||| свинина в кисло ||| 0.0169492

мясо с овощами в кисло сладком ||| свинина в кисло сладком ||| 0.0192308

мясо с овощами в кисло сладком соусе ||| свинина в кисло сладком соусе фото ||| 0.25

мясо с овощами в сливочном соусе ||| свинина с овощами в мультиварке ||| 0.125

мясо с овощами в фольге ||| запеченая свинина в фольге ||| 0.333333

мясо с овощами в фольге в ||| запеченая свинина в фольге ||| 0.333333

...

Q: первое блюдо на косточках свинных	
суп на свиных косточках	-18.8081
плов на свиных косточках	-19.080498
блюда на свиных косточках	-20.160805
суп на свиных косточек	-20.285103
плов на свиных косточек	-20.5575

Let’s apply (1)

Document	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…
Q_links	query

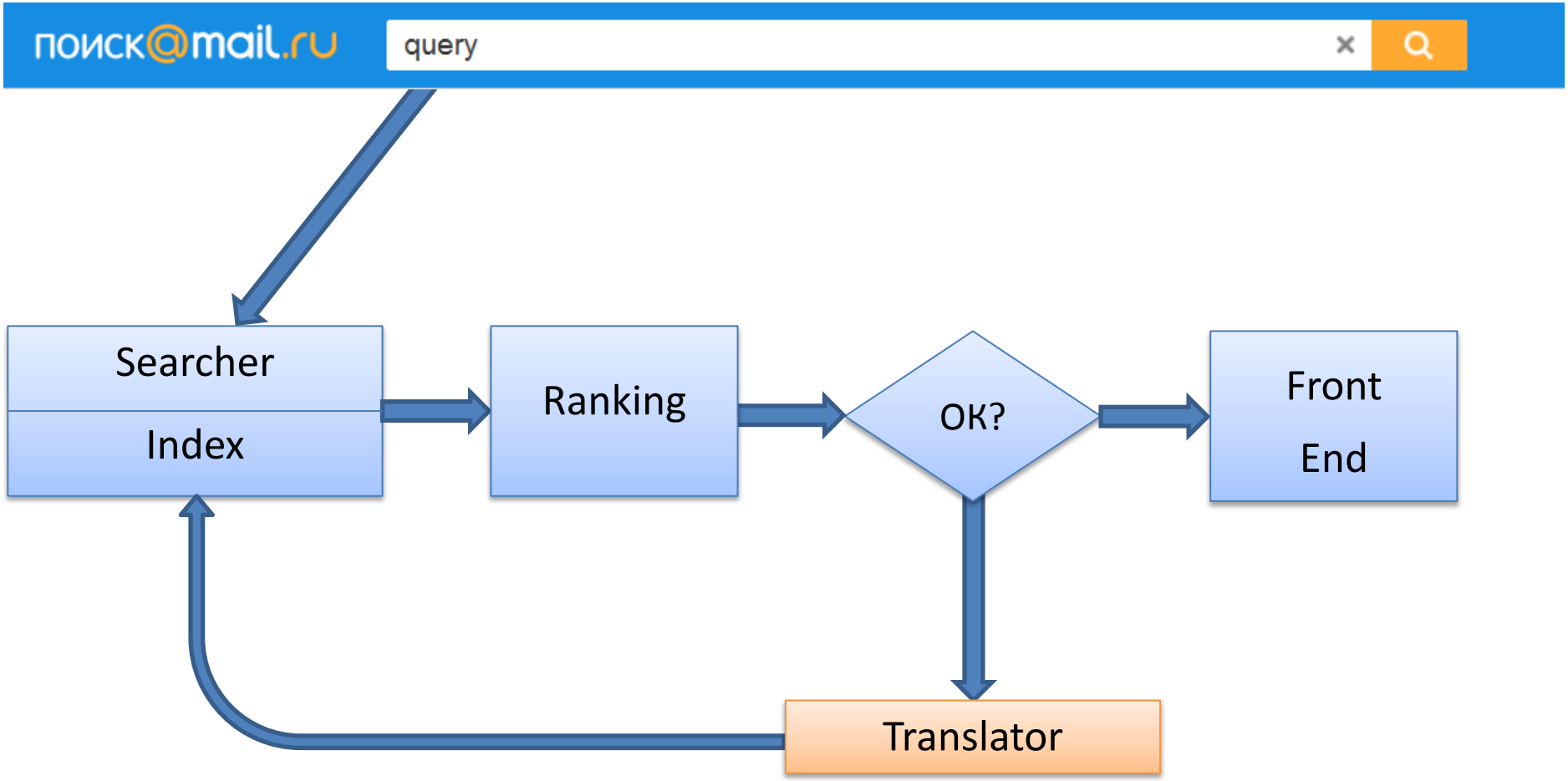
query

Translator

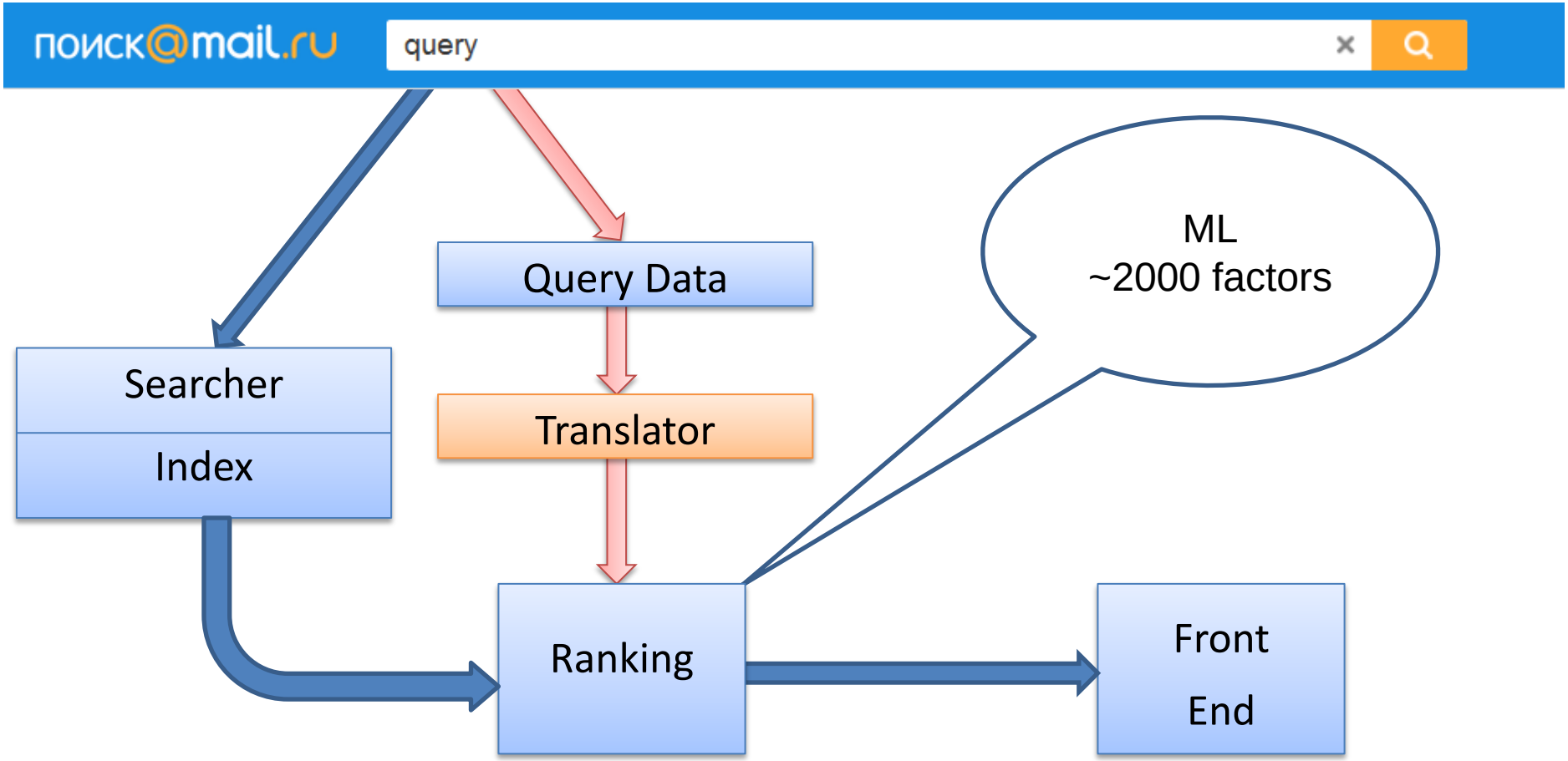
query1, query2, …

Document	
Title	w1, w2, w3…
Body	w1, w2, w3…
Ext_links	w1, w2, w3…
Q_links	query
T_links	query1, query2…

Let’s apply (2)


поиск@mail.ru
query
Searcher
Index
Ranking
OK?
Front
End
Translator

Let’s apply (3)


поиск@mail.ru
query
Query Data
Translator
Searcher
Index
ML
~2000 factors
Ranking
Front
End

Results

Now we can reformulate low-freq queries

Correct pairs' selection improves «adequacy» of translation

Random pairs	~75%
Frequency-alingned pairs	~90%

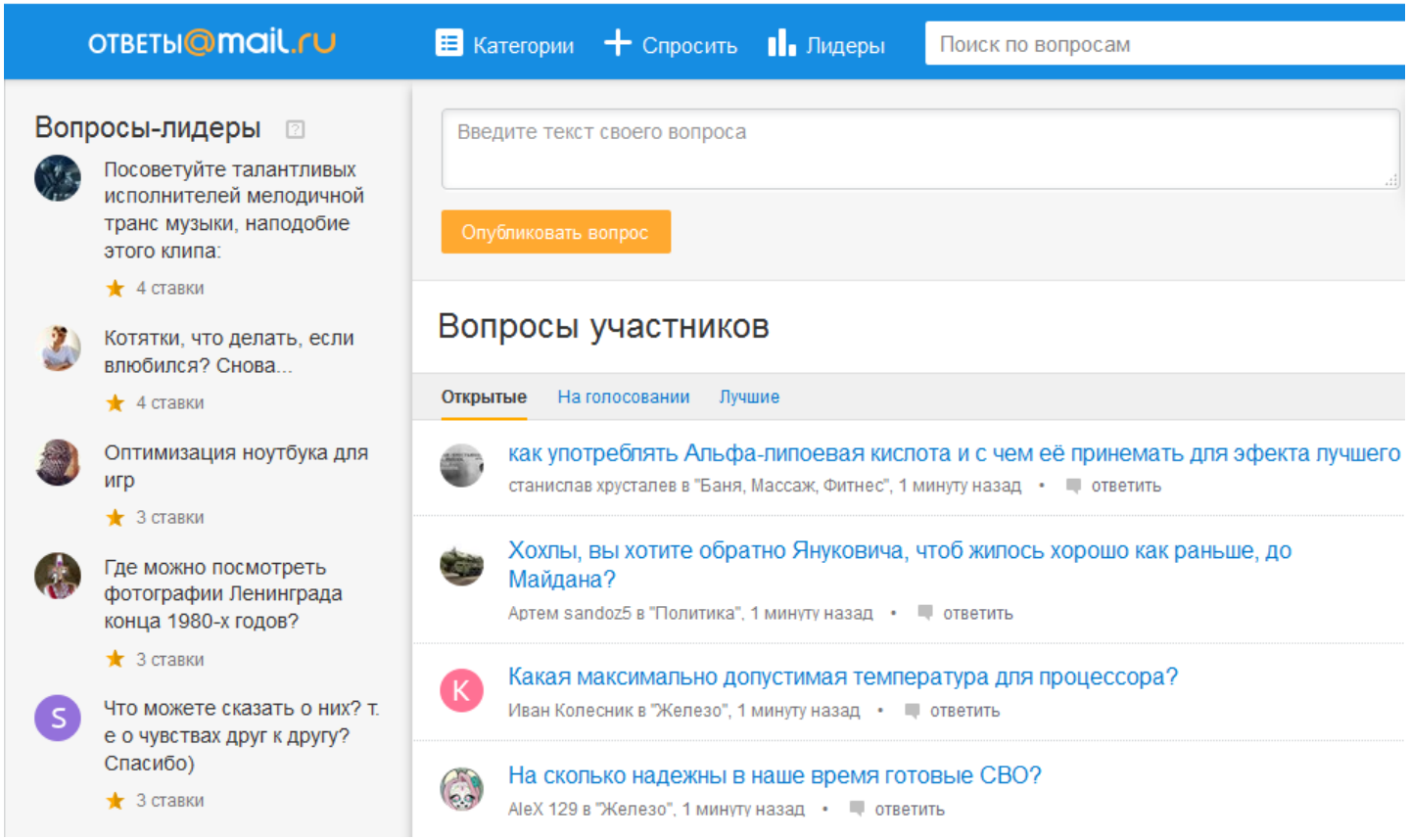
Less «traumatical» implementation is #3

- Translator works on ~3% of queries
- Quality
  - +5% on asessors' marks
  - +1% on interleaving

Technical issues

- Based on Moses
  - http://www.statmt.org/moses/
- Monolingual alignment
  - Some special techniques
- Perfomance issues
  - Multithreading
  - Memory sharing
- Language modelling

On other (possible) applications



What to read

1. Stefan Riezler and Yi Liu. Query Rewriting Using Monolingual Statistical Machine Translation.
2. Philipp Koehn. Statistical Machine Translation.

THAT’S ALL!

Igor Andreev
Search@Mail.ru
i.andreev@corp.mail.ru

@mail.ru®

www.mail.ru