Making Web Search More User-Centric

Yandex

State of Play and Way Ahead
A Yandex Overview

1997
Yandex.ru was launched

№5
Search engine in the world * (# of queries)

180 mln
Search Queries a Day

Offices
> Moscow
> 6 Offices in Russia
> 3 Offices in Ukraine
> Istanbul (Turkey)
> Zurich (Switzerland)
> Palo Alto (CA, USA)

* Source: Comscore 2009
Variety of Markets

16 countries  77 regions in Russia

Outline of the talk

> Challenges in User-Centric (personalized) Search

> Long term (user-based) Personalization
  > Search Topic Preferences
  > Demography
  > Foreign language knowledge

> Short term (session-based) Personalization
  > Query Completion
  > Session-based ranking

> Yandex and IR Research Community
User-Centric Search Challenges
Challenges of User-Centric Search

> Personalized interface and user feedback
> Evaluation and target function
> Over-optimization vs. acceptance test
> Multiple incarnations of a user
> Real-time Infrastructure
User! Be aware!
These results are personalized!
FOR YOU!

Information - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Information
Information, in its most restricted technical sense, is a sequence of symbols that can be interpreted as a message. Information can be recorded as signs, or transmitted as signals.

Information | Define Information at Dictionary.com
dictionary.reference.com/browse/information
2. Information, knowledge, wisdom are terms for human acquirements through reading, study, and practical experience. Information applies to facts told, read...

Images 10 thousand for «information»

What is information? definition and meaning
businessdictionary.com/definition/information.html
Definition: Data that (1) has been verified to be accurate and timely, (2) is specific and organized for a... Click to read more about information. ... information.

information - definition of information by the Free Online Dictionary
thefreedictionary.com/information
In this expression, line often means an anticipated tidbit of information: If you want to get a line on how she feels, she gave me a letter to give you ...
Dear Yandex (Google/Bing/Baidu/Ask)! You (always) have a very wrong assumption about who I am

> I am complicated, you know

> I want the best possible, universal answers, not worse that the ones that my neighbors, friends and wife get

> And you are too sticky!

Social in the SERPs

"How do you feel about social results mixed in with search results?"

- Want: 62%
- Do Not Want: 19%
- No Preference: 19%
User Interface and feedback

> Users
  > are not happy when they know that their results are personalized
  > are “happy” (judging by click metrics) when the results are personalized and they don’t “care” about it

> Hard to convey a message in a consumable and understandable way that
  > the personalization occurred
  > the way it occurred

> Hard to get a explicit feedback on the successful/unsuccessful personalization
Evaluation and Target Function

> Cranfield approach doesn’t work
  > all judgments must be collected in a user context

> We have to rely on user metrics
  > TDI is promising but not yet well studied

> Target function is not well defined
  > could be the “last satisfied click” position
Over-optimization vs. acceptance test

> As a result of the user-behavior centric approach
  > Strong correlation of the ranking function with click features

> Acceptance test tools can be easily “broken”
  > We have to treat click features in a smart way
  > Or even avoid them completely
Multiple incarnations of a user

> A user have several “incarnations”
  > I am at work
  > I am at the cinema (or movie theatre)
  > I am at home

> Each of them requires different rankings
  > One way is to define these environmental settings
  > What else?
Real Time Infrastructure

> Real time session-based algorithms
  > are challenging
  > require support from appropriate infrastructure

> Typical large web-search systems
  > highly heterogenic
  > highly distributed

> Yandex “short history” data delivery
  > for 7 seconds we can collect “short history” context for 95% of queries
Long Term Topic Preferences
Long Term Topic Preferences

> Continuously build user language models
  > by analyzing clicks and skips

> About 20 User Related Features, based on
  > Preferences to specific hosts
  > LMs of clicked/skipped titles/snippets
  > LMs of previous queries
  > User-specific features: e.g. Click Intensity

> Optimize for the “last satisfied” click positions
Foreign Language Knowledge

Krokhalev, Yandex, CIKM’2011
When users don't expect foreign results

> “nokia” (nokia.ru is better)
> “facebook” (facebook has Russian intro page)
> “radio c” (Russian, Brasil or Hungarian?)
> “American pie 2”, “volkswagen tuareg”, “world of tanks”, “harry potter” and many more ...

Foreign Language Knowledge
> User classification by foreign language knowledge
  > 80.0%: users with undefined language preferences
  > 12.5%: users with who never read/click English pages ("Anglophobes")
  > 7.5%: users who do not mind English pages ("Anglophiles")

> The results can be adjusted accordingly
  > maximize for diversity-aware measure (e.g. ERR-IA), or
  > downgrade English pages, or
  > maximize for ERR without any language adjustments
Foreign Language Knowledge

> Independent ranking for “Anglophiles”

> Abandonment -1.9%
> CTR of 1,2,3 positions +2.1%
> CTR of non-Russian results +7%
Demography
Demography-aware ranking

> Our goal is to optimize for the position of the “last satisfied” click
> Define **user demography vector** (UDV) as a list of values for his/her probability to belong to a specific
  > gender
  > age group
  > income class
> Where UDV itself is obtained from
  > training set of users explicitly self-identified
  > lots of features inside (proprietary) classification engine (**Crypta**)  
> Adjust Dynamic Bayesian Network Model (Chapelle. et. al, 2009)
  > Predict click and satisfaction probabilities by using UDV via  
  > Training separate logistic regression functions to estimate each variable
> MRR improvement (%) for the “last satisfied” click, for queries with different level of click entropy:

<table>
<thead>
<tr>
<th></th>
<th>Full, %</th>
<th>Gender, %</th>
<th>Age, %</th>
<th>Income, %</th>
<th>share</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H \geq 2$</td>
<td>+1.1</td>
<td>+0.6</td>
<td>+1.1</td>
<td>+0.6</td>
<td>42%</td>
</tr>
<tr>
<td>$H \geq 1$</td>
<td>+1.0</td>
<td>+0.4</td>
<td>+0.6</td>
<td>+0.7</td>
<td>68%</td>
</tr>
<tr>
<td>$H \geq 0$</td>
<td>+0.4</td>
<td>+0.1</td>
<td>+0.01</td>
<td>+0.3</td>
<td>100%</td>
</tr>
</tbody>
</table>
Session-based Query Completion
[harry potter]
Session-based Query Completion

- Lightweight approach
  - Previous query supplied via HTTP Referrer field
- Candidate suggestions collected through 3 previous months
- Find frequent pairs among candidates
- Integrate these “rules” into suggestions ranking mechanism
- It took one week from the idea to the production
Session-based Reranking

single previous query
Session-based Ranking

> Throw away unconnected pairs of consecutive queries
  > based on classifier telling if two queries have the same information need

> Compute features based on the previous query $q_0$
  > User behavior
  > Text relevance
  > Query relatedness

> Rerank and improve target click metrics
  > mean click position MCP
  > mean first click position FCP
  > mean reciprocal rank of clicks MRR.
Yandex and IR Research Community
Challenges organized by Yandex

> Previous **Relevance Prediction Challenge:**
> > A large piece of anonymized click data, and...
> > A set of query-URL pairs with relevance labels
> > Predict relevance labels for a test set
> > Submit at [http://imat-relpred.yandex.ru](http://imat-relpred.yandex.ru)
> > 500 teams registered / 100 participated
> > Competed from 22 Oct to 22 Dec 2011
> > 1400 submissions
> > Best solutions presented at Web Search Click Data workshop at WSDM 2012
Next **Search Action Prediction Challenge:**

- A large piece of anonymized click data, including...
- User IDs (so, please, personalize)
- Predict next search action in a session:
  - Query, Click, EndOfSession or SearchEngineSwitch

> Will be a part of Web Search Click Data workshop at WSDM 2013

> **Dates:** 15 October – 15 December, or earlier...
EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL
Moscow, 24-27 March 2013
Workshop/tutorial submission: **16 September**
Abstracts submission: **1 October**
Paper submission: **8 October**
Posters/Demo submission: **22 October**

<table>
<thead>
<tr>
<th>24.03 Sun</th>
<th>25.03 Mon</th>
<th>26.03 Tue</th>
<th>27.03 Wed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workshops &amp; Tutorials</td>
<td>Day 1</td>
<td>Day 2</td>
<td>Day 3</td>
</tr>
<tr>
<td>Reception</td>
<td>Posters &amp; reception</td>
<td>Conference banquet</td>
<td>Industry day</td>
</tr>
</tbody>
</table>
See you in Moscow!

http://ecir2013.org
iseg@yandex-team.ru

See you in Moscow at ECIR 2013
organized by Yandex and HSE