What Users Ask a Search Engine: Analyzing One Billion Question Queries

Michael Völske¹  Pavel Braslavski²,³  Matthias Hagen¹
Galina Lezina²,⁴  Benno Stein¹

¹Bauhaus-Universität Weimar
<firstname>.<lastname>@uni-weimar.de
www.webis.de

²Ural Federal University
³pbras@yandex.ru
⁴galina.lezina@gmail.com
Question Queries
Question Queries

lose weight
lose weight

how much should I exercise to lose ten pounds
Question Queries
Relevance

"how to"

[Google Trends, 2015]
Question Queries

Relevance

- Increasing prevalence
  - < 1% in the late 90s [Spink & Ozmutlu, Inform. Process. Manag.’02]
  - 2% in 2010 [Pang & Kumar, ACL’11]
  - 3-4% in our dataset from 2012
Question Queries

Relevance

- Increasing prevalence
  - < 1% in the late 90s [Spink & Ozmutlu, Inform. Process. Manag.’02]
  - 2% in 2010 [Pang & Kumar, ACL’11]
  - 3-4% in our dataset from 2012

- Poorer retrieval performance than keywords
  [Bendersky & Croft, WSCD’09] [Aula et al., CHI’10]
Question Queries

Relevance

- Increasing prevalence
  - < 1% in the late 90s [Spink & Ozmutlu, Inform. Process. Manag.’02]
  - 2% in 2010 [Pang & Kumar, ACL’11]
  - 3-4% in our dataset from 2012

- Poorer retrieval performance than keywords
  [Bendersky & Croft, WSCD’09] [Aula et al., CHI’10]

- Topical query classification benefits
  - General search [Bailey et al., ACM TWEB’10]
  - Query disambiguation [Li et al., SIGIR’08]
  - Search advertising [Broder et al., SIGIR’07]
What Users Ask a Search Engine

About this Talk

- Large dataset of ~1 billion question queries from Yandex

- Question query classification using CQA data as training set

- Three classification pipelines: Retrieval, BoW, Topic models

- Insights into asker behavior
What Users Ask a Search Engine

Our Approach

- **Classification task**: given unlabeled question query, predict category
  - Click information not helpful: QQ are rare
What Users Ask a Search Engine

Our Approach

- **Classification task**: given unlabeled question query, predict category
  - Click information not helpful: QQ are rare

- Community question answering (CQA) data as training set
  - CQA users manually select appropriate category for their question
What Users Ask a Search Engine

Our Approach

- **Classification task**: given unlabeled question query, predict category
  - Click information not helpful: QQ are rare

- Community question answering (CQA) data as training set
  - CQA users manually select appropriate category for their question

- Train a classifier that correctly categorizes CQA, then transfer to QQ
## Datasets

### Overview

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Queries</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Queries (yandex.ru)</td>
<td>1 980 million</td>
<td>unlabeled</td>
</tr>
</tbody>
</table>

Cleaning to remove:

- Spam & bots
- Repeated submissions
- Mis-categorized CQA questions
Datasets
Overview

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Queries</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Queries (yandex.ru)</td>
<td>1 980 million</td>
<td>unlabeled</td>
</tr>
<tr>
<td>- after cleaning</td>
<td>900 million</td>
<td>unlabeled</td>
</tr>
</tbody>
</table>

Cleaning to remove:

- Spam & bots
- Repeated submissions
Datasets
Overview

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Queries</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Queries (yandex.ru)</td>
<td>1,980 million</td>
<td>unlabeled</td>
</tr>
<tr>
<td>- after cleaning</td>
<td>900 million</td>
<td>unlabeled</td>
</tr>
<tr>
<td>CQA Questions (<a href="mailto:Otvety@Mail.ru">Otvety@Mail.ru</a>)</td>
<td>11 million</td>
<td>hierarchical (189)</td>
</tr>
</tbody>
</table>

Cleaning to remove:

- Spam & bots
- Repeated submissions
## Datasets

### Overview

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Queries</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Queries (yandex.ru)</td>
<td>1,980 million</td>
<td>unlabeled</td>
</tr>
<tr>
<td>- after cleaning</td>
<td>900 million</td>
<td>unlabeled</td>
</tr>
<tr>
<td>CQA Questions (<a href="mailto:Otvety@Mail.ru">Otvety@Mail.ru</a>)</td>
<td>11 million</td>
<td>hierarchical (189)</td>
</tr>
<tr>
<td>- after cleaning</td>
<td>6 million</td>
<td>flat (14)</td>
</tr>
</tbody>
</table>

Cleaning to remove:

- Spam & bots
- Repeated submissions
- Mis-categorized CQA questions
Datasets

Train and Test Set

- **CQA data**
  - 14 classes derived from CQA taxonomy

  - Society & Culture
  - Computers & Internet
  - Family & Relationships
  - Adult
  - Games & Recreation
  - Education
  - Home & Garden
  - Entertainment & Music
  - Cars & Transportation
  - Health
  - Consumer Electronics
  - Beauty & Style
  - Sports
  - Business & Finance
Datasets

Train and Test Set

- **CQA data**
  - 14 classes derived from CQA taxonomy
  - Training/validation set: 70/30 split
Datasets
Train and Test Set

- **CQA data**
  - 14 classes derived from CQA taxonomy
  - Training/validation set: 70/30 split

- **Question queries**
  - Test set: 1000 instances hand-labeled
  - 834 with majority agreement

---

[Diagram showing distribution of classes in train/validation and test sets]
Question Query Classification Pipelines

Pipeline 1: CQA Retrieval
Pipeline 2: Bag-of-Words Classifier
Pipeline 3: Topic Models
Question Query Classification Pipelines

- Pipeline 1: CQA Retrieval
- Pipeline 2: Bag-of-Words Classifier
- Pipeline 3: Topic Models
Question Query Classification Pipelines

Pipeline 1: CQA Retrieval

<table>
<thead>
<tr>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA Training Set</td>
<td>BM25 Scoring + Indexing</td>
<td>CQA Index</td>
</tr>
<tr>
<td>Unlabeled QQ</td>
<td></td>
<td>QQ with Categories</td>
</tr>
</tbody>
</table>

©www.webis.de 2013-15
Question Query Classification Pipelines

Pipeline 1: CQA Retrieval

**Input**
- CQA Training Set
- Unlabeled QQ

**Processing**
- BM25 Scoring + Indexing
- Index lookup
- Top-10 Majority Vote

**Output**
- CQA Index
- QQ with Categories
Question Query Classification Pipelines

- Pipeline 1: CQA Retrieval
- Pipeline 2: Bag-of-Words Classifier
- Pipeline 3: Topic Models
Pipeline 2: **Bag-of-Words Classifier**

<table>
<thead>
<tr>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA Training Set</td>
<td>Term Frequency Vectors</td>
<td>Naive Bayes Training</td>
</tr>
<tr>
<td>Unlabeled QQ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Question Query Classification Pipelines

Pipeline 2: Bag-of-Words Classifier

Input

CQA Training Set

Unlabeled QQ

Processing

Term Frequency Vectors

Naive Bayes Training

Output

QQ with Categories

Naive Bayes Model

Naive Bayes Prediction
Question Query Classification Pipelines

- Pipeline 1: CQA Retrieval
- Pipeline 2: Bag-of-Words Classifier
- Pipeline 3: Topic Models
Question Query Classification Pipelines

Pipeline 3: **Topic Models**

<table>
<thead>
<tr>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlabeled QQ</td>
<td>Topic Model Fitting</td>
<td>k-Topic Model</td>
</tr>
</tbody>
</table>

- **Model type**:
  - \{LDA, BTM\}
- **Topic count (k)**:
  - \{50..500\}

**Latent Dirichlet Allocation**: [Blei et al., JMLR’03]

**Biterm Topic Model**: [Yan et al., WWW’11]
Question Query Classification Pipelines

Pipeline 3: **Topic Models**

<table>
<thead>
<tr>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlabeled QQ</td>
<td></td>
<td>k-Topic Model</td>
</tr>
</tbody>
</table>

**Topic Model Fitting**

- **Model type**: \{LDA, BTM\}
- **Topic count (k)**: {50..500}

**Naive Bayes Training**

**QQ with Categories**

Modeling:

- Latent Dirichlet Allocation: [Blei et al., JMLR'03]
- Biterm Topic Model: [Yan et al., WWW'11]
Question Query Classification Pipelines

Pipeline 3: **Topic Models**

<table>
<thead>
<tr>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlabeled QQ</td>
<td>Topic Model Fitting</td>
<td>k-Topic Model</td>
</tr>
<tr>
<td>Model type ({\text{LDA, BTM}})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic count (k) ({50..500})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CQA Training Set</td>
<td>Topic Distribution Inference</td>
<td>Naive Bayes Training</td>
</tr>
<tr>
<td>CQA Val. Set</td>
<td></td>
<td>Naive Bayes Model</td>
</tr>
</tbody>
</table>

Latent Dirichlet Allocation: [Blei et al., JMLR’03]  
Biterm Topic Model: [Yan et al., WWW’11]
Pipeline 3: **Topic Models**

**Input**
- Model type: \{LDA, BTM\}
- Topic count (k): \{50..500\}

**Unlabeled QQ**

**Processing**
- Topic Model Fitting

**Output**
- k-Topic Model

**CQA Training Set**
- Topic Distribution Inference
- Naive Bayes Training
- Model Selection
- Naive Bayes Prediction

**CQA Val. Set**
- QQ with Categories

Latent Dirichlet Allocation: [Blei et al., JMLR’03]  
Biterm Topic Model: [Yan et al., WWW’11]
## Results

### Classifier Performance

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F$_1$-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA Retrieval</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>6 million</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bag-of-Words</td>
<td>0.61</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>137,032</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA Topics</td>
<td>0.40</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biterm Topics</td>
<td>0.49</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>450</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Simple BoW classifier performs similarly to CQA retrieval
- Biterm topic model outperforms LDA
Results
Classifier Performance

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F$_1$-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA Retrieval</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>6 million</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bag-of-Words</td>
<td>0.61</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>137,032</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDA Topics</td>
<td>0.40</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biterm Topics</td>
<td>0.49</td>
<td>0.53</td>
<td>0.51</td>
</tr>
<tr>
<td>450</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Classification</td>
</tr>
<tr>
<td>CQA Retrieval</td>
<td>medium</td>
</tr>
<tr>
<td>Bag-of-Words</td>
<td>low</td>
</tr>
<tr>
<td>Topic Models</td>
<td>high</td>
</tr>
</tbody>
</table>

- Simple BoW classifier performs similarly to CQA retrieval
- Biterm topic model outperforms LDA
- Topic models less accurate but faster at classification time
Results

Classifier Performance

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA Retrieval</td>
<td>0.67</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>Bag-of-Words</td>
<td>0.61</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>LDA Topics</td>
<td>0.40</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>Biterm Topics</td>
<td>0.49</td>
<td>0.53</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CQA Retrieval</td>
<td>medium high</td>
</tr>
<tr>
<td>Bag-of-Words</td>
<td>low medium</td>
</tr>
<tr>
<td>Topic Models</td>
<td>high low</td>
</tr>
</tbody>
</table>

- Simple BoW classifier performs similarly to CQA retrieval.
- Biterm topic model outperforms LDA.
- Topic models less accurate but faster at classification time.
- We use the Bag-of-Words classifier to analyze the question queries dataset.
Results

Evolution of Categories over Time
Results

Evolution of Categories over Time

- Business & Finance
- Sports
- Beauty & Style
- Consumer Electronics
- Health
- Cars & Transportation
- Entertainment & Music
- Home & Garden
- Education
- Society & Culture
- Computers & Internet
- Games & Recreation
- Adult
- Family & Relationships
- Games & Recreation
- Adult
- Family & Relationships
- Computers & Internet
- Society & Culture
- Business & Finance

Month

01/2012 02/2012 03/2012 04/2012 05/2012 06/2012 07/2012 08/2012 09/2012 10/2012 11/2012
Results

Evolution of Categories Over Time: An Example

![Chart showing the evolution of categories over time with a line graph.

- Two categories are represented:
  - Cars & Transportation
  - Education

- The x-axis represents the months from January 2012 to November 2012.
- The y-axis represents the share of monthly volume.

- The chart indicates a significant peak in Cars & Transportation during July 2012, followed by a decrease.
- Education shows fluctuations, with a notable increase in late 2012.]
Results
Prefixes and Suffixes Across Categories

How-to Questions
Results
Prefixes and Suffixes Across Categories
Summary

1. Analysis of question queries at unprecedented scale
2. First QQ study for Russian language
3. Categorization scheme using CQA data
4. Askers behavior across categories
Summary

1. Analysis of question queries at unprecedented scale
2. First QQ study for Russian language
3. Categorization scheme using CQA data
4. Asker behavior across categories

Future Work

1. Cross-language comparison
2. Deeper insights into asker behavior
3. More advanced classification schemes for short texts
4. Causes of increase in question query prevalence

Thank you :-)
Summary

1. Analysis of question queries at unprecedented scale
2. First QQ study for Russian language
3. Categorization scheme using CQA data
4. Asker behavior across categories

Future Work

1. Cross-language comparison
2. Deeper insights into asker behavior
3. More advanced classification schemes for short texts
4. Causes of increase in question query prevalence

Thank you :-)
Acknowledgements

- Yandex (Alexey Gorodilov, Pavel Serdyukov, Alexander Sadowski)

- Mail.Ru (Andrey Oleynik)
Appendices

Question Queries are Short

Question length (words)

Cumulative percentage of questions

- Question Queries
- CQA Questions
Appendices

Question Queries are Unique

![Graph showing cumulative percentage of question queries against question frequency. The graph compares Question Queries and CQA Questions. The cumulative percentage for Question Queries increases with frequency, reaching a plateau around 0.95. CQA Questions show a similar trend but at a slightly lower cumulative percentage.]
Appendices

CQA Classification Performance

Validation Set (n ≈ 2 million)

![Graph showing F1-Score for different methods vs number of topics]

- **Unigrams**
- **BTM**
- **LDA**

The graph illustrates the performance of different methods (Unigrams, BTM, LDA) across varying numbers of topics, with the F1-Score on the y-axis and the number of topics on the x-axis.
Appendices
Confusion Matrix for Unigram Classifier

Left: CQA Validation set; Right: QQ test set
Appendices

References


Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of CHI 2010*, pages 35–44.


