

Search quality in practice

Alexander Sibiryakov, ex-Yandex engineer, data scientist at Avast!
sixty-one@yandex.ru

Agenda

- What is search quality?
- Examples of search quality problems.
- Evaluating search quality. Methods.
- Signals is the key.
- Producing good snippets.

Agenda

- **What is search quality?**
- Examples of search quality problems.
- Evaluating search quality. Methods.
- Signals is the key.
- Producing good snippets.

- *Search quality* - abstract term, includes relevance, user experience, and reveals overall effectiveness of search by humans.
- *Relevance* - in search, is the measure of conformity of user information need to document found.

Relevance is subjective

A user takes relevance in a very subjective way:

- The context of the problem, he is trying to solve
- awareness about the problem,
- user interface
 - document annotations,
 - presentation form,
 - order,
 - previous experience with this search system.

AAA AUTO 7 991 Kč
 800 110 800
 40 000 Kč v hotovosti na ruku! V nabídce přes 7000 aut se zárukou 24 měsíců.
www.aaaAuto.cz/Ojeta-Auta

Auto ESA | Váš autobazar
 Ojeté vozy všech značek. Pořídte si nyní **auto** na splátky bez navýšení!
www.AutoEsa.cz

Reklama Sklik

Speciální nabídka Citroën
 Skladové vozy s akční výbavou. Technopaket za 1 Kč. Více zde.
akce.citroen.cz

Auto.cz - vše o autech na jednom místě už 16 let
 ... server na českém internetu. Denně přináší žhavé novinky ze světa aut, aktuální testy, videa a reportáže.
auto.cz

Jaké auto?
 Najděte si články o autě, které vás zajímá? Ať u je to Dacia Lodgy, Škoda Octavia III, Volkswagen ...
jake.auto.cz

Moje.Auto.Cz
 ... uživatelé AUTO.CZ. Zaregistrujte se. Jste-li registrovaný uživatel AUTO.CZ a chcete hodnotit auta,
moje.auto.cz

Seznam.cz, new search UI with big screenshots

← → ↻ 🏠 images.yandex.ru/yandsearch?text=счастье&uinfo=ww-1274-wh-786-fw-1049-fh-580-pd-1 ⚙️ ☆ ☰

наверх счастье 📏 📄 Папки

Файл
JPEG
PNG
GIF

Свежие картинки

На сайте
URL сайта

Искать в других поисковых системах

Waiting time could affect overall search experience, including user decision regarding relevance.

images.yandex.ru - image search from yandex.ru

Search systems behavior could be learned by users

- Seznam.cz has very good document base on Czech internet, bigger than Google, but has less powerful ranking and very sensitive to query formulation.
- Yandex is very bad on software development queries, because of lack of documents or bad ranking.

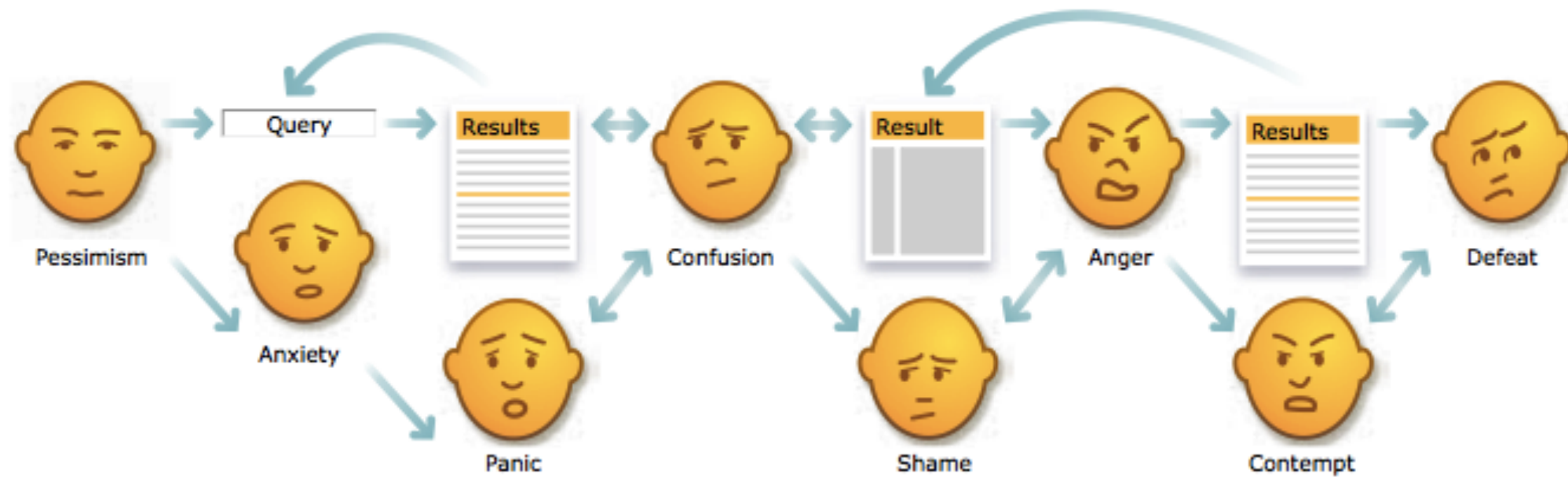
Problems

- No definitive formulation. Considerable uncertainty. Complex interdependencies.
- We, developers, aren't prepared to tackle search. We can't manage high-tech, step-changing, cross-functional, user-centered challenge.
- The role of search in user experience is underestimated. Therefore, nobody measure and knows how good it is.

From «Search Patterns» P. Morville & J. Callender, O'Reilly, 2010

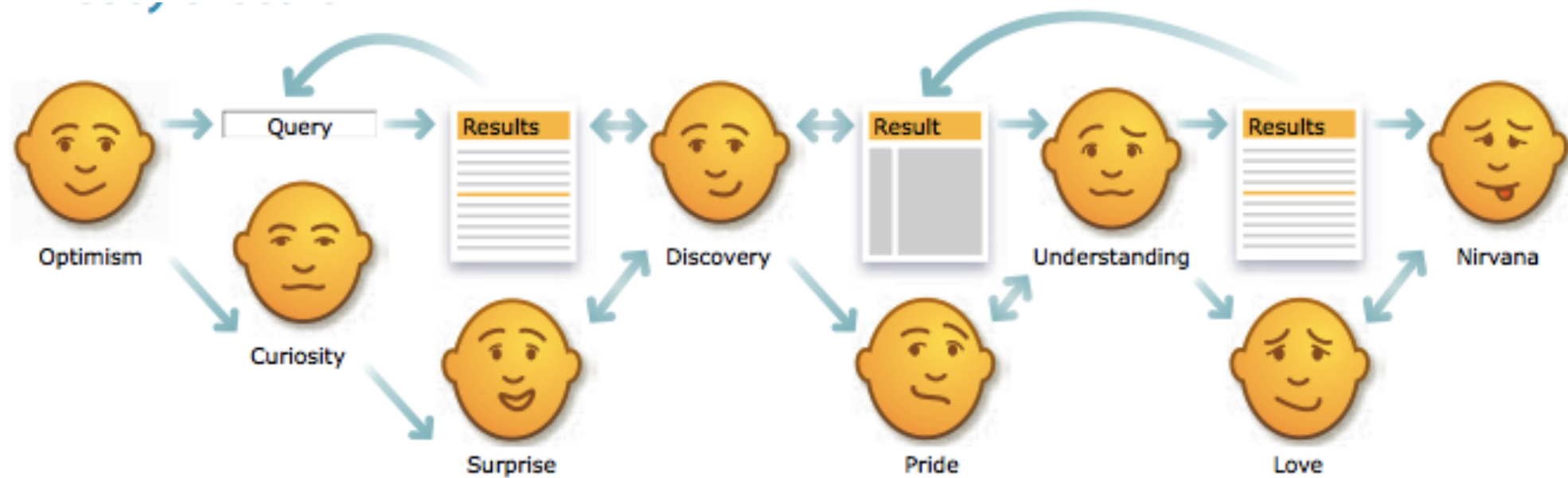


From «Search Patterns» P. Morville & J. Callender, O'Reilly, 2010



Poor search is bad for business and sad
for society

From «Search Patterns» P. Morville & J. Callender, O'Reilly, 2010



Search can be a source of information
and inspiration

From «Search Patterns» P. Morville & J. Callender, O'Reilly, 2010

Agenda

- What is search quality?
- **Examples of search quality problems.**
- Evaluating search quality. Methods.
- Signals is the key.
- Producing good snippets.

Examples of search quality problems

- Search of model no. or article
[6167 8362823] → [61 67 8 362 823]
(telescopic noozle), proper tokenization
- Detection and correction of typing errors
[drzak myla] → [drzak mydla]
(soap holder), lexical ambiguity
- Question search
[how to buy a used xperia] → [... smartphone]
[how to buy a used fiat] → [... car]
wrong weighting of important words.

Agenda

- What is search quality?
- Examples of search quality problems.
- **Evaluating search quality. Methods.**
- Signals is the key.
- Producing good snippets.

Evaluation of search

- Baseline for improvement of search system,
- as usual, there is no ideal measure,
- use multiple measures,
- keep in mind properties of each measure, when making a decision.

Evaluation of search: methods

- Query-by-query comparison of two systems,
- classic Cleverdon's Cranfield evaluation,
- Pairwise evaluation with Swiss system.

Query-by-query comparison

- Take random queries from the stream, for example 100.
- query each system and evaluate *the whole SERP* of topN results with scale:
 - ++ (very good)
 - + (good)
 - (bad)
 - (very bad)
- Count judgements of each type.

Query-by-query comparison: example

- Comparing Google and Bing

[berlin buzzwords] - G++, B+

[java byteoutputstream] - G+, B-

Google: ++ - 1, + - 1

Bing: + - 1, - - 1



Cyril Cleverdone, born Bristol UK,
1914-1997

British librarian, best known for his work on the
evaluation of information retrieval systems

Cleverdon's Cranfield evaluation

- Components:
 - Document collection,
 - set of queries,
 - set of relevance judgements.
- Measures (per query):
 - **Precision** - fraction of retrieved documents that are *relevant*.
 - **Recall** - percent of *all* relevant documents returned by the search system.

Cleverdon's Cranfield evaluation: example

- [berlin buzzwords]

No.	URL	Judgement
1	berlinbuzzwords.de/	R
2	https://www.facebook.com/berlinbuzzwords	R
3	https://twitter.com/berlinbuzzwords	R
4	www.youtube.com/playlist?list=PLq-odUc2x7i8Qg4j2fix-QN6bjup	NR
5	https://developers.soundcloud.com/blog/buzzwords-contest	R
6	www.retresco.de/the-berlin-buzzwords-over-and-out/	NR
7	planetcassandra.org/events/berlin-de-berlin-buzzwords-2014/	R

$$Pr = C_{Rel} / C = 5 / 7 = 0,71$$

$$Re = C_{Rel} / C_{RelOverall}$$

Cleverdon's Cranfield evaluation: averaging

- Macro-average:

$$PR_{MaA} = (Pr_1 + Pr_2 + \dots + Pr_N) / N$$

- Micro-average:

$$PR_{MiA} = (C_{Rel1} + C_{Rel2} + \dots + C_{RelN}) / (C_1 + C_2 + \dots + C_N)$$

N - count of judged SERP's

- Variations:

Pr1, Pr5, Pr10 - counting only top 1, 5, 10 results.

Normalized Discounted Cumulative Gain (NDCG)

- Measures usefulness, or *gain*, of document based on its position in the result list.
- The gain is accumulated from the top of the result list to the bottom with the gain of each result discounted at lower ranks.

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad NDCG_p = \frac{DCG_p}{IDCG_p}$$

rel_i - graded relevance of the result at position i ,
 DCG_p - discounted cumulative gain for p positions.

From http://en.wikipedia.org/wiki/Discounted_cumulative_gain

Pairwise evaluation with Swiss system (experimental)

- Judgement of **document pairs**,
- «Which document is more relevant to the query X?»
 - answers are:
Left, right, equal.
- Chosen document is getting one point, in case of «equal», both are getting by one point.
- Pairs preparation using Swiss tournament system:
 - First pass. All documents are ordered randomly or by default ranking. Then take first document from first half, and first from second (1-st with 5-th, 2-nd with 6-th, and so on) to get pair.
 - In the next pass, only winners of previous pass are judged. The same way, taking documents from first and second halves starting from top to create pairs for judgement.

Which document is more relevant to the query [berlin buzzwords] ?

Berlin Buzzwords - Kollwitzkiez - Foursquare

foursquare.com › Professional & Other Places › Convention Center ▾

See 6 photos from 43 visitors to **Berlin Buzzwords**. ... Photo taken at **Berlin Buzzwords** by Robert R. on 6/4/2013; Photo taken at **Berlin Buzzwords** by Robert R.

Berlin Buzzwords 2014

berlinbuzzwords.de/ ▾

This year we've conducted interviews with our **Berlin Buzzwords** Speakers to get to know them better. How did they get started developing software? What will ...

Left

Equal

Right

Pairwise evaluation with Swiss system

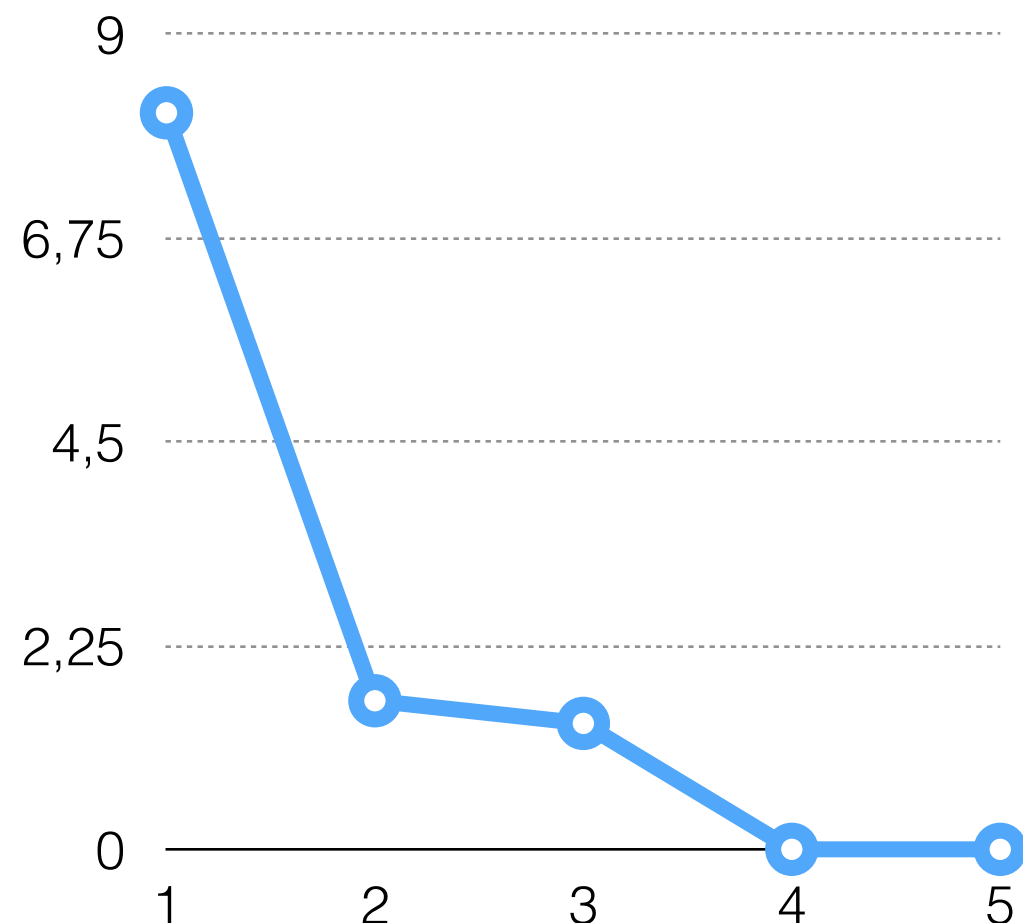
- About 19 judgements is needed for 10 documents retrieved for 1 query.
- After judgement is finished, the ranking is built by gathered points.
- According to position the weights are assigned to the documents.
- Using weights, the machine-learned model can be trained.

Pairwise evaluation with Swiss system: weights assignment

- For example, we can use exponential weight decrement:

$$W = P * EXP (1/pos)$$

1. 8,13 (3)
2. 1,64 (1)
3. 1,39 (1)
4. 0 (0)
5. 0 (0)



Agenda

- What is search quality?
- Examples of search quality problems.
- Evaluating search quality. Methods.
- **Signals is the key.**
- Producing good snippets.

Signals is the key: agenda

- Production system: what data is available?
- Text relevance: approaches, no silver bullet.
- Social signals.
- How to mix signals: manual linear model, gradient boosted decision trees.

Production system: what data is available?

- Documents:
 - CTR of the document,
 - absolute number of clicks,
 - count of times, when document was clicked first in SERP,
 - the same, but last
 - count of clicks on the same SERP before/after the document was clicked.
- Displays (shows):
 - Count of times document was displayed on SERP,
 - count of unique queries, where document was displayed,
 - document position: max, min, average, median, etc.

Production system: what data is available?

- Queries:
 - Absolute click count on query,
 - Abandonment rate,
 - CTR of the query,
 - Time spent on SERP,
 - Time spent till first/last click,
 - Query frequency,
 - Count of words in query,
 - IDF of words of query: min/max/average/median, etc.,
 - Count of query reformulations: min/max/average/median.
 - CTR of reformulations.

Text relevance: use cases

- Phrase search,
- search of named entities (cities, names, etc.)
- search of codes, articles, telephone numbers,
- search of questions,
- search of set expressions (e.g. «to get cold»)
- ...

Text relevance: signals

- BM25F zoned version: meta-description, meta-keywords, title, body of the document,
- calculate BM25 on query expansions: word forms, thesaurus based, abbreviations, translit, fragments,
- min/max/average/median of count of subsequent query words found in the document,
- the same, but query order,
- the same, but with distance +/- 1,2,3 words,
- min/max of IDF of query words found,
- to build language model of document and use it for ranking,
- language model of queries, of different words count, use probabilities as a signals.

Text relevance: example model

ScoreTR =

$a * \text{BM25} +$

$b * \text{BM25F}_{\text{Title}} +$

$c * \text{BM25F}_{\text{Descr}} +$

$\text{MAX}(\text{SubseqQWords})^d,$

a, b, c, d - can be estimated manually, or using stochastic gradient descent.

Social signals

- Count of readers/commenters of content,
- count of comments published during some time period (velocity),
- time since last comment,
- speed of likes growth,
- time since last like,
- absolute count of likes,
- etc.

How to mix signals: learning-to-rank

Learning to rank or **machine-learned ranking (MLR)** is the application of machine learning, typically supervised, semi-supervised or reinforcement learning, in the construction of ranking models for information retrieval systems.

From Wikipedia, M. Mohri, et al. Foundations of Machine Learning, The MIT Press, 2012

How to mix signals: full-scale process

- The training set preparation:
 - Documents,
 - Queries,
 - Relevance judgements.
- Framework:
 - Querying of search and dump of feature vectors (incl. assigning relevance judgements),
 - learning model,
 - evaluation of model,
 - adoption of model in production system,
 - repeat after some time.

How to mix signals: DIY way

- Choose manually some set of features, which you think are good predictors,
- create a simple linear model from these predictors,
- fit coefficients manually by selecting few (10) representative queries.

ScoreTR =

$$\begin{aligned} & a * \text{BM25} + \\ & \text{MAX}(\text{SubseqQWords})^b + \\ & c * \text{CTR} + \\ & d * \text{Likes} + \\ & e * \text{QLength}; \end{aligned}$$

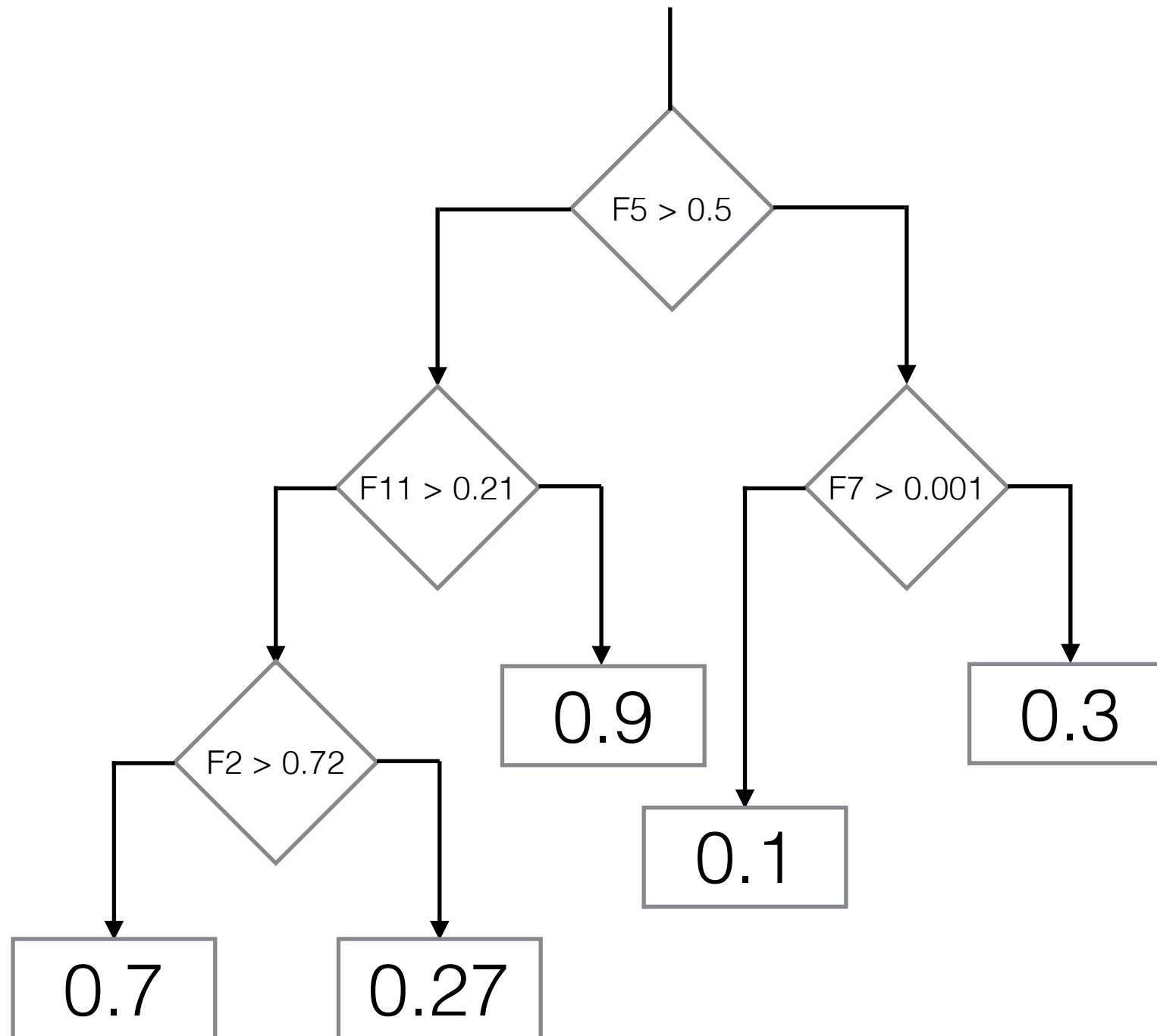
a, b, c, d, e - needs to be fit.

How to mix signals: more work

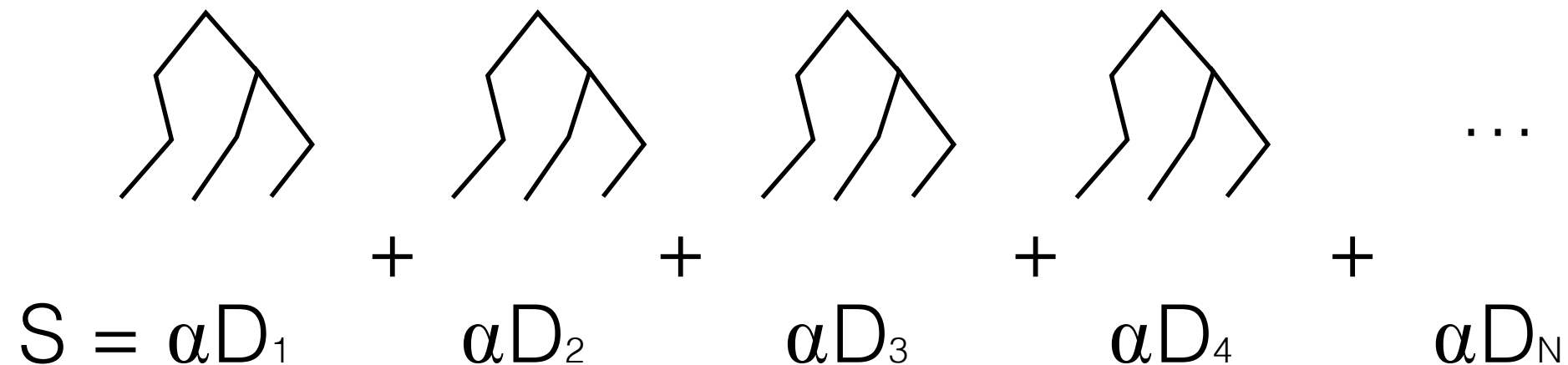
- Get some relevance judgements:
 - pairwise evaluation,
 - classic Cranfield way,
 - using some good signal, sacrificing it *
- Learn a more complex model: Ranking-SVM, or Gradient Boosted Decision Trees (GBDT).

* - make sure there are no big correlations with other signals.

Decision tree



Gradient boosted decision trees



The diagram illustrates the equation $S = \alpha D_1 + \alpha D_2 + \alpha D_3 + \alpha D_4 + \dots + \alpha D_N$. Above the equation, there are four identical decision tree symbols, each consisting of a root node with two children, and a third node branching from the right child. These are followed by an ellipsis and then another decision tree symbol. Plus signs are placed between each tree symbol and between the ellipsis and the final tree symbol. Below the trees, the equation is written as $S = \alpha D_1 + \alpha D_2 + \alpha D_3 + \alpha D_4 + \alpha D_N$.

α - step,

D_i - result of each weak predictor (tree),

N - count of weak predictors

Each weak predictor is learned on subsample from the whole training set.

Yahoo! Learning to rank challenge, 2011

Table 5: Performance of the 3 baselines methods on the validation and test sets of SET 1: BM25F-SD is a text match feature, RankSVM is linear pairwise learning to rank method and GBDT is a non-linear regression technique.

	Validation		Test	
	ERR	NDCG	ERR	NDCG
BM25F-SD	0.42598	0.73231	0.42853	0.73214
RankSVM	0.43109	0.75156	0.43680	0.75924
GBDT	0.45625	0.78608	0.46201	0.79013

Table 2: Statistics of the two datasets released for the challenge.

	SET 1			SET 2		
	Train	Valid.	Test	Train	Valid.	Test
Queries	19,944	2,994	6,983	1,266	1,266	3,798
Dococuments	473,134	71,083	165,660	34,815	34,881	103,174
Features		519			596	

Agenda

- What is search quality?
- Examples of search quality problems.
- Evaluating search quality. Methods.
- Signals is the key.
- **Producing good snippets.**

Producing good snippets: text summarization

The problem is to generate a summary from original document taking into account

1. Query words,

2. length,

3. style.

[mardi gras fat tuesday]

What is the origin of **Fat Tuesday / Mardi Gras**?



www.gotquestions.org/Mardi-Gras-Fat-Tuesday.html ▼

by S. Michael Houdmann - in 528 Google+ circles

Answer: **Mardi Gras**, which is French for "Fat Tuesday," is the last day of a season called ... **Mardi Gras** is the culmination of festivities and features parades, ...

Producing good snippets: types

1. *Static* - generated once, their content will not change when query changes, may not have query words at all.
2. *Dynamic* - generated individually for each query, usually contain query words.

Almost all modern search systems use dynamic generation of snippets or combination.

Producing good snippets: algorithm

1. Generate presentation of the document as a set of paragraphs, sentences and words.
2. Generate candidates for snippet for given query.
3. For each candidate generate signals and rank candidates with machine learned model.
4. Selection of most suitable candidate(s) fitting requirements.

Producing good snippets: example signals

- Length of candidate text,
- amount of query words in candidate text,
- BM25,
- IDF of query words in candidate text,
- is there beginning/ending of sentence ?
- conformity of query words order,
- conformity of word forms between query and text,
- etc.

Thank you.

Alexander Sibiryakov, sixty-one@yandex.ru