Saint-Petersburg State University
Mathematics and Mechanics Faculty

Vladimir Nazarenko

# Incorporating the data quality for information retrieval

Term paper

Scientific advisor:
Professor Novikov B. A.

Saint-Petersburg
2015

# 1 Introduction

Nowadays there is a need to process large amounts of data. In some cases we can accept slightly inaccurate answer in exchange of reduction of computational resources required to calculate this answer. This brings us to the idea of approximate query answering.

Naturally, every query corresponds to some abstract query and we would like to operate with it instead of query in the specific programming language. We use the notion of q-set to achieve this.

**Definition 1.** *Q-set is an abstraction of both a query and the query result. It is a triple, consisting of a query, set of objects and scoring function for the set of objects which takes values from 0 to 1.*

Additional motivation for such definition is present in [1]. Note that we treat the scoring function primarily as objects relevance for query.

For each query we consider its evaluation plan, consisting of elementary operations, such as set-theoretic, filtering and joins. We treat this plan as follows: initial set of objects we represent as a q-set with empty query (or, equally, query which returns all of the objects) and scoring function taking the value 1 for each object. Also we represent each operation in the query evaluation plan as a q-set. In these assumptions we can define quality of a query as an aggregate of qualities of elementary operations.

As we can see the q-set represents the relevance of the object. In homogeneous systems our goal is to provide the most relevant objects. However, in heterogeneous systems we can also be interested in the quality of the retrieved objects if it is supplied. Evidently, we do not need the irrelevant albeit high quality objects, nonetheless we can try to use the quality of the objects to improve the performance of the retrieval.

# 2 Problem statement

Suppose that you have some q-set and some abstract information retrieval engine with an ability to calculate the result of the q-set approximately and hence get the approximate result. Apparently the approximate result may differ from the precise one and the difference is the error of evaluation. Our goal is to get a quantitative measure for the outlined error.

There is also one closely related task. Consider the quality score for each object of the q-set. This quality measure can appear in heterogeneous systems. The natural desire is not just getting the relevant to a query objects, but getting the relevant objects with the highest quality. So our next goal is to add this desire to our measure for a relevance error.

## 3 Related work

Dolmatova et. al. [2] propose the cost models for approximate operations, but authors do not consider supplied quality assessment.

Nevertheless our work lies at the junction of data quality and approximate querying, which are intensively studied areas. In the papers [3][4][5][6] authors discuss the basic ideas of the data quality and propose the tools for building quality metrics. Authors of [7][6] went further and defined more specific quality metrics for Internet pages.

Authors of [8] developed a resource allocation algorithm for approximate query processing, but this algorithm lacks quality measure both for approximate operation and for data. Therefore the mentioned paper can be considered as motivation for our work.

## 4 Operation quality

As was mentioned above we look at the elementary operations and treat them as q-sets.

We have two ways to execute the operation: approximately and precisely. As was mentioned in [4], even the result of the precisely evaluated query may be inaccurate because query is user-specific, but we do not focus on it and suppose that query directly corresponds the informational need of a user. Thus we treat the result of precisely evaluated operation as ideal one.

So what quality metric do we need? First of all, for the sake of convenience this assessment should be normalized which would allow us to easily compare two assessments. Secondly, the assessment should be non-decreasing. This feature brings us some order on the approximate q-sets so we can, for instance, compare two approximate algorithms.

Dolmatova et. al. [2] treat number of correctly ranked results as the quality measure for operation. We would like to go further.

The simple idea is to use the precision and recall, but these metrics are too inflexible. So we propose using the following analogue of nDCG:

Assume

$q^e$ - sorted list of relevance scores from q-set, returned by precisely evaluated query

$q^a$ - sorted accordingly to $q^e$ list of relevance scores from q-set, returned by approximate query evaluation algorithm

$$max\ error(k) = |q_k^e - 0.5| + 0.5 \tag{1}$$

$$Max\ error_p = \sum_{k=1}^{p} \frac{2^{max\ error(k)} - 1}{\log_2(k+1)} \tag{2}$$

$$Error_p = \sum_{k=1}^{p} \frac{2^{|q_k^a - q_k^e|} - 1}{\log_2(k+1)} \tag{3}$$

$$nQuality_p = 1 - Error_p / Max\ error_p \tag{4}$$

Properties of (4):

- Normalized
- Nondecreasing
- Less relevant documents have less influence on the result of the quality
- May be applied at the different levels by changing parameter p
- It is easy to incorporate the quality of an object: we can take the harmonic mean of relevance and quality instead of just relevance
- Authors of [9] analytically proved that the nDCG has the property of consistent distinguishability, hence our formulas inherit it

# 5 Information quality

It is hard to define the quality of an abstract object so that it is useful in practice, though we can state some common principles. Usually in this case they say about the dimensions of the data quality. Here are some dimensions from the [5]:

- Accessibility
- Appropriate Amount of Data
- Believability
- Completeness
- Understandability

These dimensions are quite rational but still quite abstract, so we do not have a set of metrics to access the data quality uniformly, but if we restrict the area, we can achieve better results.

The world wide web is a bright example of heterogeneous system and we decided to use it as an example for this work. In the meantime we face two problems trying to define the quality of web pages:

1. The lack of content verification leads to the huge amount of the low quality pages. Also SEOs generate large amount of the spam.
2. The uncertainty of the informational need. Strong and Wang suggest [4] that quality of data cannot be assessed independently of the people who use data. Taking into account the number of users of the Internet and the difference of their interests we may say that defining the data quality still seems to be not a trivial task.

Authors of [6] provided a detailed comparison of information quality frameworks in context of World Wide Web. According to this paper the dimensions of quality can be divided into to classes: subjective, i.e. user-specific and objective – user-independent.

For now we are interested only in objective metrics for data quality. Eppler et al. suggest [7] a convenient and concise set of objective criteria to assess the quality of the page which can be obtained either by computing or by user surveys.

# 6 Web page quality as ranking signal

As an example of the task we try to solve, let us consider the problem of ranking web pages in search engine results. As was stated earlier, query is usually user-dependant, so natural desire for user is to get the web pages of reasonably good quality. Suppose that the relevance judgement by experts is a precise solution and the relevance estimated from document features like tf-idf or bm25 is an approximate one. Additionally, we can extract some quality features from documents, as described in [7]. So here we may want to optimize $nQuality_p$ with supplied quality for each object as proposed in the properties of (4) instead of $nDCG_p$ or any other usual metric for assessing the ranking quality.

Researchers showed [10][11] that solving this problem leads to better ranking results both for homogeneous collection and for heterogeneous ones. They extracted quality features such as depth of the url, fraction of table conent, entropy of the page text, fraction of stopwords in the page text from web-collections and used them as ranking signals. This resulted in significant ranking improvement.

# 7 Conclusions and further research

We discussed the information quality generally and the quality of web pages. We proposed the metric to assess the quality of an approximate operation. We showed the practical example of using our results. The directions of the further investigation are following:

- Define quality for another data
- Conduct computational experiments on web data using features from [7] to show that proposed formula (4) is empirically correct

# References


[1] Boris Novikov, Natalia Vassilieva, and Anna Yarygina. Querying big data. In *Proceedings of the 13th International Conference on Computer Systems and Technologies*, pages 1–10. ACM, 2012.

[2] Oxana Dolmatova, Anna Yarygina, and Boris Novikov. Cost models for approximate query evaluation algorithms. In *DB&Local Proceedings*, pages 20–28. Citeseer, 2012.

[3] Stuart E Madnick, Richard Y Wang, Yang W Lee, and Hongwei Zhu. Overview and framework for data and information quality research. *Journal of Data and Information Quality (JDIQ)*, 1(1):2, 2009.

[4] Diane M. Strong, Yang W. Lee, and Richard Y. Wang. Data quality in context. *Commun. ACM*, 40(5):103–110, May 1997.

[5] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[6] Shirlee-ann Knight and Janice M Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science: International Journal of an Emerging Transdiscipline*, 8(5):159–172, 2005.

[7] Martin J Eppler and Peter Muenzenmayer. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In *IQ*, pages 187–196, 2002.

[8] Anna Yarygina and Boris Novikov. Optimizing resource allocation for approximate real-time query processing. *Computer Science and Information Systems*, (00):63–63, 2014.

[9] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT 2013)*, 2013.

[10] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 288–295, New York, NY, USA, 2000. ACM.

[11] Michael Bendersky, W Bruce Croft, and Yanlei Diao. Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 95–104. ACM, 2011.