**Literature Review**

Keywords: Information Quality, Wikipedia, Machine Learning, Quality Metrics

**Objective**

Examine state of the art of existing metrics to measure Wikipedia article quality, and the usage of machine learning to predict it.

**Research Questions**

R1: What are the most common metrics used to evaluate article quality in Wikipedia?

R2: To what extent are those metrics currently used, and how effective are they?

R3: Which machine learning approaches can be applied to predict article quality?

**Search Queries (todo: also include URL, code format queries?):**

* Google Scholar: Wikipedia (measuring OR measure) article quality
* ACM Digital Library: [Title: wikipedia assessing quality] AND [[Abstract: metric] OR [Abstract: measure]] AND [Full Text: wikipedia] AND [Full Text: assessing] AND [Full Text: quality] AND [Publication Date: (01/01/2000 TO 12/31/2021)]
* Web of Science: wikipedia assessing quality (metric OR measure)

All searches were restricted to a date range of 2000-2021.

**Inclusion Criteria**

I1: Papers discussing possible metrics to assess data quality in a collaborative encyclopedia.

I2: Papers discussing their findings and evaluating the metrics’ effectiveness.

I3: Papers discussing different machine learning approaches to predict article quality.

**Exclusion Criteria**

E1: Papers discussing vandalism and quality of user edits instead of the articles themselves.

E2: Papers discussing manual approaches to assess article quality, as opposed to automatic ones

E3: Papers in a language other than English

**Process**

P1: Title is assessed for possible relevance within the research area, in order to be screened.

P2: Look for the existence of duplicate publications.

P3: Scan paper abstracts, bringing focus to the research questions, verifying which studies relate to the field.

P4: Analyze full text, starting with motivation, objectives and then results. Focus on inclusion and exclusion criteria: Publications meeting no inclusion criteria or meeting any exclusion criteria are discarded

P5: Order paper quality by the criteria defined in the following section.

**Quality Score**

As described above, each publication will be assessed and assigned a quality score, based on 5 questions. The final score for a paper will consist of the sum of the results of each question. The questions and their score ranges are defined below.

Q1: Does the paper relate to measuring article quality in Wikipedia? [0-2]

Q2: Are the proposed solutions (metrics/approaches) well described? [0-2]

Q3: Are the benefits and limitations well described? [0-2]

Q4: Does the paper avoid restricting to a certain area/field (e.g., Health) [0-1]

Q5: Does the paper focus on more than one article language [0-1]

TODO: Check Q4/Q5 for possible relevance

**Storage**

The references for the papers included at the end of each process phase were stored in a .xml file. The quality scores calculated during P5 were recorded in a wiki.

**Paper Collection**

In order to have a consistent publication search, queries to the different article databases were designed to be as similar as possible. Unfortunately, Google Scholar’s lack of advanced options prevented the search from being refined enough, so around 18000 results were fetched. As it would be impractical and unnecessary to analyze all of them (most results were irrelevant to the research), only the first 300 results were viewed. As for the other databases, the searches returned much shorter lists of publications, therefore all the results could be properly inspected.

Google Scholar: 18300 (300) -> 50 [<https://scholar.google.com/scholar?q=Wikipedia+%28measuring+or+measure%29+AROUND%283%29+article+AROUND+%283%29+quality&hl=en&as_sdt=0%2C5&as_ylo=2000&as_yhi=2021>]

ACM: 99 -> 24 [https://dl.acm.org/action/doSearch?fillQuickSearch=false&target=advanced&expand=dl&AfterMonth=1&AfterYear=2000&BeforeMonth=12&BeforeYear=2021&AllField=Title%3A%28wikipedia+assessing+quality%29+AND+Abstract%3A%28metric+OR+measure%29+AND+Fulltext%3A%28wikipedia+AND+assessing+AND+quality%29]

Web of Science: 121 -> 17

[[**https://www.webofscience.com/wos/woscc/summary/9c126beb-2b28-4fd7-9f5b-a6862c992b25-17f692c7/relevance/1**](https://www.webofscience.com/wos/woscc/summary/9c126beb-2b28-4fd7-9f5b-a6862c992b25-17f692c7/relevance/1)**]**

**AFTER REMOVING ALL THE DUPLICATES 🡪 69**

**Full text assessment (P4)**

Abstract, inclusion, exclusion.

First, the paper structure was analyzed, identifying the sections which would most likely contain information regarding the measurement of article quality. After that, the publication was skimmed through, to exclude the results which were not so relevant to the research questions, or that only discussed them superficially. Finally, the entire text of the article was studied, taking into special account the sections concerning the explanation of article quality metrics and machine learning strategies.

**Data Collection**

Metrics, approaches, limitations, goods, stored in wiki

Vandalismo 🡪 motivaçao

Que algoritmos usaram?

Posso usar outros artigos dps

Livros de machine learning: Ja deste

Pag. 8 -> tabela de números

Pag. 9 -> Medir qualidade de publicação

Usar o VSCode / Wiki para as notas e incluir os scores de qualidade nisso, mas dps também se podem fazer tabelas a partir disso. Por exemplo para os quality scores. E a wiki pode ser a

fonte de referencia para os ids.

Gap analysis?

**Possible useful references:**

https://en.wikipedia.org/wiki/Data\_quality

**Useful**

https://library.acg.edu/how-to-guides/google-scholar/advanced-searching

Extensive literature review example: <https://www.mdpi.com/2073-8994/10/10/470>

Important Tables from it:Table

Description automatically generated

Diagram

Description automatically generated