**OUTDATED INFORMATION: CHECK DISSERTATION OR WIKI**

**Literature Review**

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

**Objective**

Examine state of the art of existing features to automatically measure Wikipedia article quality, as well as the usage of machine learning algorithms for the same purpose.

**Research Questions**

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

R2: To what extent are those features 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 features or metrics to assess data quality in Wikipedia.

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

I3: Papers discussing 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 not meeting all 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. Papers that do not relate to measuring article quality in Wikipedia will not be scored.

Q1: Does the paper describe and compare possible metrics with detail? [0-3]

Q2: Does the paper describe and compare ML approaches with detail? [0-3]

Q3: Are the results, benefits and limitations of their solution well described? [0-3]

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

Q4: Does the paper bring focus to an article language that’s not English? [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** 🡪 70

**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**

We collected: Metrics, approaches, results, pros / limitations, all of these 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?

Deixar TODOs só no literature notes,? Se separarmos o literature notes, talvez deixar no full-text assessment? Ou ent se no lit notes tivermos uma lista deixar la os todos. Ou usar índice??

Mencionar o facto que fomos mais leniente na parte do abstract scanning para garantir q n se perde nada. Isto e no abstract scanning quisemos tudo o que se relacionasse com medir qualidade na wikipedia.

A ideia é pegar num conjunto de metricas tanto simples (numero de palavras, numero de seccoes, imagens, citacoes, etc.) como mais complexas (reputações dos contribuidores, histórico de revisões, etc.) e, baseando-nos na escala de qualidade da wikipedia (featured article, good article sao exemplos de bons artigos mas há poucos artigos categorizados) tentar arranjar um modelo que tente correlacionar essas metricas com a qualidade do artigo. Claro que nao se preve muito classificar a veracidade da informacao, mas provavelmente vai haver alguma correlação tambem entre a qualidade e a veracidade de artigos

**Possible useful references:**

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

<https://en.wikipedia.org/wiki/Recurrent_neural_network>

What if we showed all consulted references were good quality? :o :D

**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