Wiki Evolution dataset: English Wikipedia revision articles represented by quality attributes

Ana Luiza Sanches[1], **Sinval de Deus Vieira Júnior**[1], **Daniel Hasan Dalip**[1], **Bárbara Gabrielle C. O. Lopes** [2]

[1]Departamento de Computação
Centro Federal de Educação tecnológica de Minas Gerais (CEFET-MG)
Belo Horizonte, MG – Brasil

[2]Departamento de Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

analuizatrz@gmail.com, sinvalvieirajunior@gmail.com,
hasan@cefetmg.br, barbaragcol@dcc.ufmg.br

Abstract. *This paper presents the creation and publishing of the Wikipedia article's evolution dataset. This dataset is a set of revisions of articles, represented by quality attributes and quality classification. This dataset can be used for studies regarding automatic quality classification that consider the article revision history as well as understanding how the content and quality of articles evolve over time in this collaborative platform.*

Resumo. *Este artigo descreve a criação e disponibilização da base de dados de evolução de artigos da Wikipédia. A base é caracterizada por atributos de qualidades e a classe de qualidade dos artigos em determinada data, sendo cada instância entendida como revisão. Esta base pode ser utilizada para estudos relacionados com classificação automática de qualidade que considerem o histórico de revisão do artigo e entendimento de como o conteúdo e qualidade dos artigos evoluem ao longo do tempo nessa plataforma colaborativa.*

1. Introdução

A Internet permite o amplo compartilhamento de conteúdo por parte de qualquer usuário que tenha acesso à rede. Isso somado ao aumento no número de usuários da Internet possibilitou a existência de um volume enorme e crescente de informação. Um exemplo de plataforma de conteúdo da rede é a Wikipédia, que possui mais de 5 milhões de artigos em inglês escritos por meio de um esforço colaborativo envolvendo 37,6 milhões de usuários registrados e um número indefinido de usuários anônimos [Wikipedia 2019c].

Considerada um dos maiores repositórios de conhecimento humano, a Wikipédia recebeu muita atenção e a avaliação de qualidade de seus artigos se tornou uma preocupação importante durante a década de 2000 [Dang and Ignat 2016]. Essa preocupação com a qualidade dos artigos se deve principalmente à discussão de confiabilidade dos textos colaborativos, pois, como consequência da sua estrutura aberta, a Wikipédia "não pode garantir, de maneira nenhuma, a validade das informações de seu conteúdo" [Wikipedia 2019a]. Sendo assim, é considerado imprescindível a referência a fontes externas e citação "inline" para a verificação das informações contidas nos artigos. Atualmente a classificação

de artigos da Wikipédia envolve avaliadores voluntários responsáveis por classificar os artigos em sete classes de qualidade [Wikipedia 2019b].

Para identificar a qualidade dos artigos, entretanto, os especialistas humanos não são suficientes, uma vez que a alta velocidade de mudança nos artigos torna impossível a execução dessa tarefa de forma manual [Dang and Ignat 2016]. Por esse motivo, vários trabalhos têm explorado a classificação automática de artigos levando em conta diversos critérios como a contribuição colaborativa, identificação de vandalismo, identificação de controvérsia, *feedback* do usuário, entre outros [Jhandir et al. 2017].

Em todos esses trabalhos a coleta e estruturação de dados sobre os artigos é uma etapa essencial. Os dados coletados podem conter vários problemas, como valores ausentes, dados falsos, dados duplicados e falta de padronização [Batista et al. 2018]. Tais problemas geralmente são resolvidos na etapa de pré-processamento de dados, que requer cerca de 80% do tempo de cientistas de dados [Tyagi et al. 2010].

Com isso, o presente trabalho visou a criação e disponibilização de uma base de dados que possa ser utilizada para obter métodos automáticos de predição da qualidade de artigos da Wikipédia. O algoritmo de coleta utilizado na criação desta base de dados também está sendo disponibilizado.

Este artigo está estruturado da seguinte forma: A seção 2 apresenta a forma como a Wikipédia disponibiliza seus dados e como eles foram utilizados para gerar o dataset, a seção 3 descreve o algoritmo de coleta desenvolvido, especificando parâmetros e resultados de cada etapa. Em seguida, a seção 4 apresenta a coleta de duas amostras de bases de dados que estão disponíveis para uso. Por fim, a seção 5 apresenta alguns possíveis usos para a base de dados e possibilidades de melhoria deste trabalho.

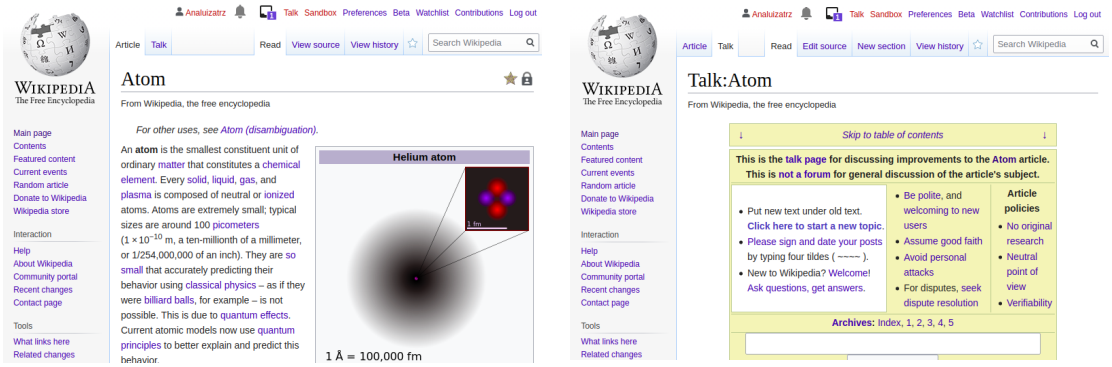
2. Organização da Wikipédia e acesso aos dados

O processamento de dados da Web requer o prévio conhecimento das fontes disponíveis. Com tal entendimento, é possível realizar um planejamento da coleta e definir estratégias que podem melhorar sua eficiência e cobertura [Batista et al. 2018]. Portanto os parágrafos seguintes apresentarão a organização da Wikipédia e como os dados são disponibilizados para consumo.

Os artigos da Wikipédia possuem quatro páginas que são do interesse deste trabalho, visto que apresentam os dados necessários para criação da base de dados: a página de artigo, página de discussão, histórico de revisão da página de artigo e histórico de revisão da página de discussão.

A página de artigo, Figura 1a, apresenta o conteúdo do artigo sendo a mais conhecida por usuários, principalmente os que são apenas leitores da Wikipédia. Além desta, cada artigo possui uma página de discussão (Figura 1b) por meio da qual os editores e avaliadores discutem sobre as modificações feitas no conteúdo do artigo. Essa página também contém a classe de qualidade em que o artigo se encontra, o que é uma informação relevante para este trabalho.

As páginas de revisão do artigo apresentam as versões passadas de um determinado artigo. A partir delas, é possível acessar o conteúdo de um título em determinada data. Neste trabalho, revisão é definida como uma versão do artigo em uma data. Sendo



(a) Página do artigo Atom **(b) Página de discussão do artigo Atom**

Figura 1. Páginas de artigo e de discussão do artigo Atom

assim, a página de artigo exibe a revisão mais atual do artigo, utilizada por leitores, enquanto a página de revisão exibe o histórico.

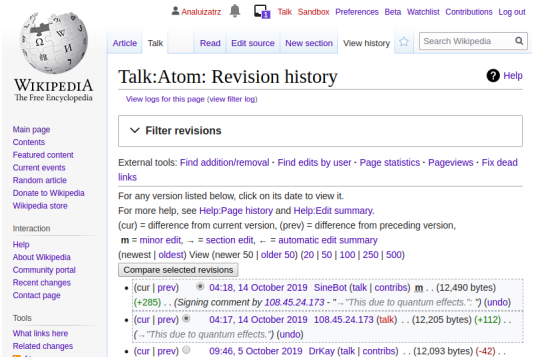


Figura 2. Página do histórico de revisões da página de discussão do artigo Atom

Assim como a página do artigo, a página de discussão também possui um histórico de revisão, possibilitando consultar discussões passadas. A página de discussão apresenta, dentre outras informações, a classe de qualidade, possibilitando a obtenção da qualidade das revisões anteriores de um artigo por meio da página do histórico de revisões do artigo (Figura 2).

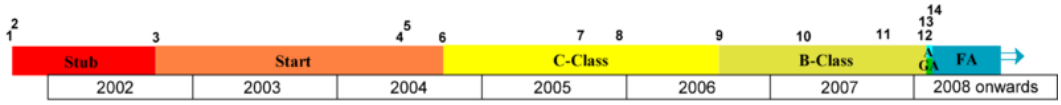


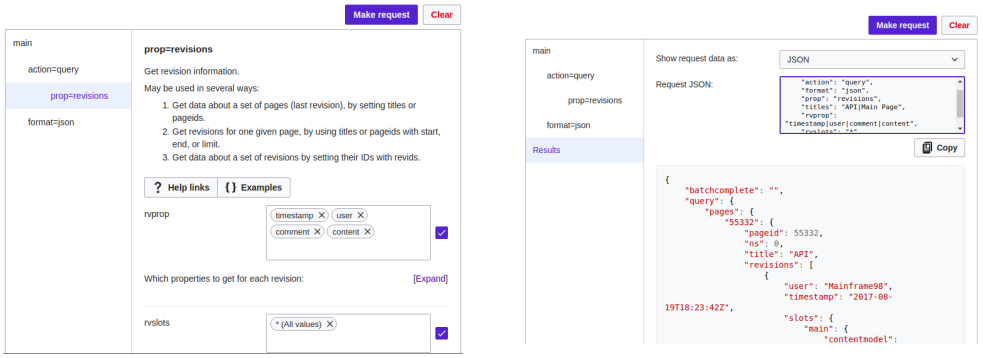
Figura 3. Evolução de qualidade do artigo Atom
Fonte: [Wikipedia 2019b]

As páginas de discussão tornam possível acessar versões passadas dos conteúdos e estudar a evolução de artigos ao longo do tempo. A evolução se caracteriza pelas alterações de conteúdo na página de artigo e reavaliações pela comunidade na página de discussão. A Figura 3 mostra a evolução do artigo "Atom" ao longo do tempo, por meio das classes de qualidade da Wikipédia. Os artigos são avaliados em sete classes de qualidade, sendo elas, em ordem decrescente de qualidade: *Feature Article* (FA), A, *Good Article* (GA), B, C, Start e Stub.

2.1. Consumo via API

Há duas formas principais disponíveis para consumo de dados da Wikipédia, via API[1] e por meio de coleta de páginas Web. Optamos pelo consumo de dados via APIs, visto que essas ferramentas apresentam a vantagem de um formato bem definido e um esquema claro que geralmente são bem documentados para que seus usuários possam as utilizar [Batista et al. 2018].

Foram consideradas as APIs da Wikipedia Special Export[2] e MediaWiki[3]. A MediaWiki (Figura 4a), foi selecionada por possuir possuir mais parâmetros que a *Special Export*, possibilitando consumir informações de forma mais específica e personalizada. [WikiMedia 2019].



(a) Interface de acesso à API MediaWiki utilizando a ação consulta (na imagem action=query)

(b) Resposta e parâmetros para a requisição da Figura 4a

Figura 4. Requisição e resposta da API MediaWiki
Fonte `https://en.wikipedia.org/wiki/Special:ApiSandbox`

A Figura 4b mostra a resposta da requisição mostrada na Figura 4a. A resposta é exibida na parte inferior, e consiste em uma estrutura de dados em formato JSON. Além disso, é possível visualizar a requisição em formato JSON na parte superior.

2.2. Representação dos dados

Mesmo possuindo API bem definida, muitos sistemas não atendem especificamente à demanda de usuários. Isso significa que as APIs podem disponibilizar dados em formatos diferentes daquele que é necessário para o uso, sendo necessário implementar uma aplicação específica para consumir os dados e estruturá-los no formato desejado [Batista et al. 2018].

A partir da API MediaWiki, por exemplo, é possível consumir o histórico de revisão da página de artigo e também da página de discussão, isoladamente. Porém ainda é necessário fazer um processamento dos dados para que seja direcionado para fins específicos, como o uso em treinamento de modelos de aprendizado de máquina. Para este trabalho o conteúdo de cada revisão é utilizado para a geração dos atributos de qualidade relacionados à tamanho, estilo, estrutura e legibilidade [Hasan Dalip et al. 2009]. A página de discussão da revisão é utilizada para a extração da classe de qualidade que

[1]APIs, do inglês "Interface de Programação de Aplicativos" são interfaces bem definidas para consumo de dados na Web

[2]`https://en.wikipedia.org/wiki/Special:Export`

[3]`https://en.wikipedia.org/wiki/Special:ApiSandbox`

categoriza o conteúdo da revisão. E, por fim ainda, é necessário unificar essas duas informações a fim de obter os atributos e a classificação de cada revisão. A base de dados final é composta por 44 atributos de qualidade, descritos em [Dalip 2015], e a classe de qualidade.

3. O Algoritmo de coleta

A coleta tem como objetivo a criação da base de dados de evolução. O algoritmo recebe como parâmetro uma lista de títulos e um período de tempo e retorna como resultado uma base de evolução em forma de arquivo CSV. O período de tempo é determinado por um uma data de início e uma data de fim. O algoritmo encontra-se disponível em `https://github.com/analuizatrz/wiki-crawler`.

A Figura 5 apresenta as etapas de coleta, que utilizam a WikiMedia API por meio de requisições HTTP, e as etapas de tratamento de dados, que extraem ou transformam a resposta das requisições. Os retângulos representam coletas enquanto os círculos representam tratamento de dados.

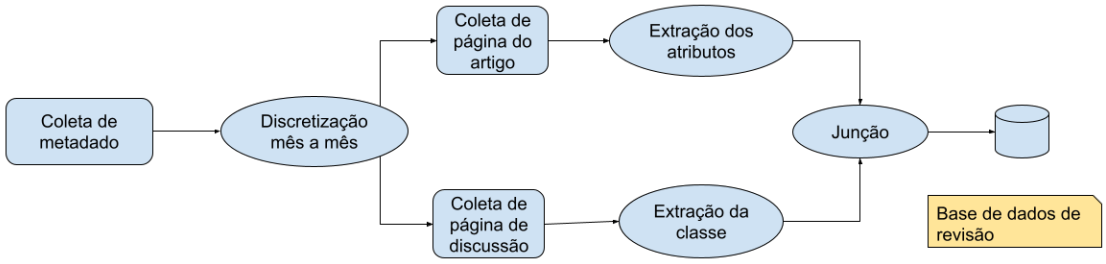


Figura 5. Diagrama da coleta e tratamento de dados para criação da base de dados de revisão

Fonte: Os autores

A primeira coleta realizada foi a dos metadados. Por meio dela, foram obtidas as informações contidas na Tabela 1. Depois de coletados, os metadados foram filtrados de forma a selecionar apenas uma instância por mês do período definido. Os metadados selecionados são das últimas revisões anteriores ao primeiro dia de cada mês do período de tempo estabelecido. Por exemplo, se há duas revisões contíguas, uma do dia 29 de setembro e a próxima é dia 2 de outubro, a revisão do dia 29 de setembro será a selecionada. O motivo de filtrar os metadados dessa forma é que mudança de classe de qualidade de um artigo não ocorre com tanta frequência.

Tabela 1. Metadados de uma instância da base de Revisão

Metadado	Descrição
ID	identificador da revisão
ID do pai	id da revisão pai, ou seja, da revisão anterior sem as modificações atuais
comentário	descrição da modificação inserida na revisão
timestamp	data de criação da revisão
acesso	data fictícia de acesso ao artigo
usuário	editor da revisão

A partir dos metadados discretizados mês a mês, são realizadas outras duas etapas de coleta, a primeira para obtenção de página de artigo e a segunda para obtenção de página de discussão. Das páginas de artigo são extraídas os atributos de qualidade e das páginas de discussão são extraídas as classes de qualidade. Cada metadado é a representação de uma revisão, caracterizada principalmente por título e data, que auxilia nas próximas coletas. Após as etapas de coleta e extração, cada revisão é caracterizada, além de título e data, por um vetor de atributos de qualidade de 44 dimensões e uma classe de qualidade.

Algoritmo 1: Coleta de artigo que resulta na base de dados de revisão

Data: titulos, dataInicio, dataFim
Result: R
$R \leftarrow \emptyset$;
M $\leftarrow$ coletaMetadados(titulos, dataInicio, dataFim);
foreach $\boldsymbol{m}_t^p \in M$ **do**
  $W_p \leftarrow$ coletaPaginaArtigo($\mathbf{m}_t^p$);
  $\mathbf{a}_t^p \leftarrow$ extraiAtributos(W_p);
  $W_d \leftarrow$ coletaPaginaDiscussao($\mathbf{m}_t^p$);
  $c_t^p \leftarrow$ extraiClasse(W_d);
  $r_t^p \leftarrow [\mathbf{m}_t^p, \mathbf{a}_t^p, c_t^p]$;
  $R \leftarrow R \cup \{r_t^p\}$;

O Algoritmo 1 mostra o processo em alto nível. A partir dos metadados coletados, são obtidos os atributos e a classe, que compõem a instância da base de revisão. O desenvolvimento dos algoritmos de cada etapa da coleta, coletaMetadados(), coletaPaginaArtigo() e coletaPaginaDiscussao() foi iterativo. Isso significa que uma coleta amostral foi realizada enquanto os algoritmos foram sendo aprimorados. Esta base de dados amostral será apresentada seção 5.

O algoritmo evoluiu para incorporar mecanismos de recuperação de erro bem como lidar com desambiguação e redirecionamentos. Para a recuperação de erro foi criado um registro das páginas já coletadas e as páginas que não foram coletadas dada algum erro. Quando a qualquer etapa de coleta se inicia, os artigos já coletados não são colocados na fila de coleta novamente. Também é possível estudar os artigos que não foram coletados e entender o porque o erro ocorreu. O redirecionamento ocorre quando a página acessada é uma página de redirecionamento[4], que envia automaticamente o visitante para outra página. As páginas de redirecionamento são úteis para referenciar o mesmo artigo por outro título (e.g. Einstein[5] e Albert Einsten[6] são redirecionados para o mesmo conteúdo. A desambiguação, por sua vez, ocorre quando um título é ambíguo e pode se referir a dois ou mais artigos distintos. As páginas de desambiguação [7] são úteis para ajudar o visitante a encontrar o artigo que o interessa ao exibir os artigos relaciona-

[4] `https://en.wikipedia.org/wiki/Wikipedia:Redirect`. Acesso em 15 de novembro de 2019.

[5] `https://en.wikipedia.org/wiki/Einstein`). Acesso em 15 de novembro de 2019.

[6] `https://en.wikipedia.org/wiki/Albert_Einstein`. Acesso em 15 de novembro de 2019.

[7] `https://en.wikipedia.org/wiki/Wikipedia:Disambiguation`. Acesso em 15 de novembro de 2019.

dos ao título ambíguo (e.g. há três artigos ligados ao título *"Mercury"*[8], o elemento[9], o planeta[10] e mitologia[11]).

3.1. Coleta dos metadados

A coleta dos metadados recebe como parâmetros os títulos a serem coletados em um período, especificado por meio de datas inicial e final, em que as revisões foram feitas. As informações obtidas por este coletor são id da revisão, id da página, data da revisão, usuário que realizou a modificação e comentário da revisão. Os metadados são salvos em arquivo CSV para posteriormente serem utilizados pelas próximas coletas.

3.2. Coleta da página de artigo

A coleta da página de artigo consiste na coleta de conteúdo das página de artigo de cada revisão. O coletor possui como parâmetros os pares título e data da revisão, advindo de um metadado.

Depois de coletados o conteúdo das páginas de artigo de cada revisão, foi executada a extração de atributos. Para a extração dos atributos de estrutura, em especial, foi necessário converter o formato das páginas principais. O formato original das páginas é um formato próprio da Wikipédia denominado Wikitext [12]. As páginas foram então convertidas para HTML e em sequência foram extraídos os atributos de qualidade. Para a geração dos atributos de qualidade foi utilizada a biblioteca web quality [Pinto et al. 2020].

3.3. Coleta da página de discussão

Para associar uma revisão da página de discussão à uma revisão da página de artigo foram considerados o título e a data, de forma que a revisão de discussão associada à uma revisão de artigo é aquela que possui a maior data menor que a data da revisão de artigo.

A coleta do conteúdo da página de discussão possui como parâmetros os pares título e data da revisão. As revisões da página de artigo não são diretamente relacionadas com as revisões da página de discussão. Neste trabalho foi considerado que dado que a qualidade é avaliada sobre uma revisão já escrita, a página de discussão considerada é aquela cuja data é a posterior mais próxima à data de revisão. Por exemplo, dado uma revisão da página de artigo de 5 de abril e duas revisões de página de discussão dos dias 4 e 6 de abril, a qualidade considerada será extraída da revisão do dia 6 abril. Estes parâmetros são obtidos dos metadados discretizados mês a mês.

A partir do conteúdo da página de discussão, é extraída a classe de qualidade da revisão. Como as páginas da Wikipédia são de um formato próprio denominado Wikitext, foi necessário implementar um tratamento para identificação da classe, a partir de uma expressão regular capaz de identificar o padrão "{{project—class=x—propriedade_1=?—(...)}}"[13],

[8] `https://en.wikipedia.org/wiki/Mercury`. Acesso em 15 de novembro de 2019.

[9] `https://en.wikipedia.org/wiki/Mercury_(element)`. Acesso em 15 de novembro de 2019.

[10] `https://en.wikipedia.org/wiki/Mercury_(planet)`. Acesso em 15 de novembro de 2019.

[11] `https://en.wikipedia.org/wiki/Mercury_(mythology)`. Acesso em 15 de novembro de 2019.

[12] `https://en.wikipedia.org/wiki/Help:Wikitext`

[13] `https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Avalia%C3%A7%C3%A3o_autom%C3%A1tica`

sendo x a classe que se quer obter. A expressão regular para detectar esse trecho foi ”class( )*=( )*([a-z]+)”. A Figura 6 ilustra esse processo com um exemplo da página de discussão do artigo *Binary Search*.

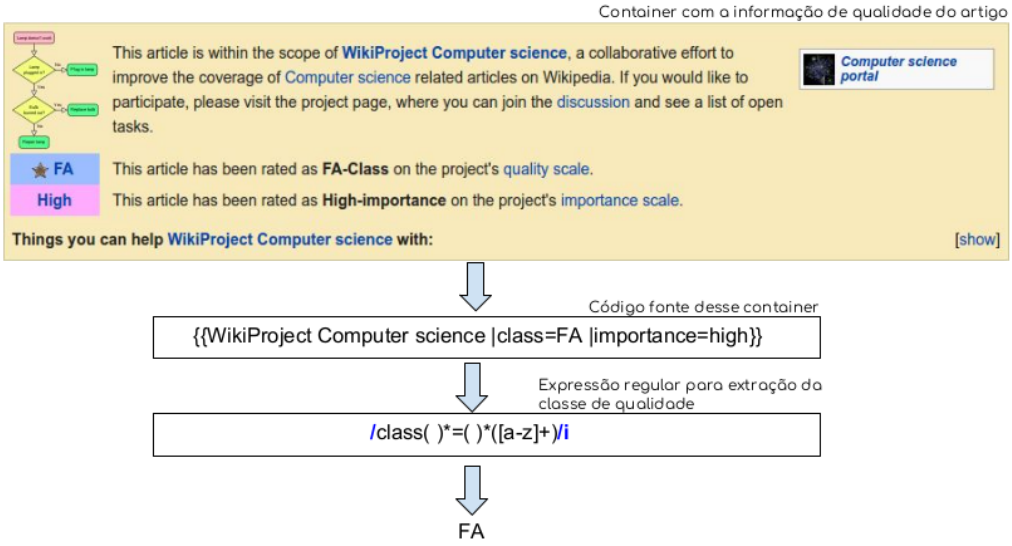


Figura 6. Demonstração de extração de classe de qualidade a partir de uma página de discussão do artigo *Binary Search*.
Fonte: batista:2018

Após a extração dos atributos e da classe de qualidade de cada revisão, os vetores metadados $\mathbf{m}_t^p$, atributos $\mathbf{a}_t^p$ e classe c_t^p, de cada revisão, são concatenados formando uma instância $\mathbf{r}_t^p$ da base de dados. Como a revisão pode ser identificada por um página (título/artigo) p e uma tempo (data) t, uma instância da base de dados de revisão R pode ser representado conforme a Equação 1.

$$\mathbf{r}_t^p = [\mathbf{m}_t^p, \mathbf{a}_t^p, c_t^p] \forall p \in P, \forall t \in T \tag{1}$$

4. Bases de dados coletadas

O algoritmo descrito na seção 2 foi desenvolvido de forma iterativa, sendo usada como teste a coleta de um conjunto de títulos da amostra. A versão final do algoritmo foi então executada com um número maior de títulos e um período maior, visando a criação de uma base de dados mais abrangente. A coleta da amostra inicial foi bem documentada com relação a tempo de execução e armazenamento de documentos intermediários. Já a amostra final foi coletada de forma fracionada em um esforço que durou meses, e não registramos o tempo exato de execução de cada etapa.

4.1. Coleta da amostra inicial

Para a construção da base da amostra inicial foram utilizados como parâmetros 3.294 títulos presentes na base de dados de [Hasan Dalip et al. 2009], e o período definido de 2007 à 2009. A Tabela 2 mostra detalhes das etapas da coleta e em quantos títulos erros ocorreram.

No coleta da página de artigo foram obtidas 53.023 revisões de 3.242 páginas, o que consome 1,5 GB de armazenamento. O tempo de execução do algoritmo para a coleta desses dados totalizou aproximadamente 5 horas e 17 minutos.

Tabela 2. Estatística da coleta da amostra inicial

Coleta	# Artigos	Erros	Total
Metadado	3.246	48	3.294
Página do Artigo	3.242	4	3.246
Classe de qualidade	2.999	247	3.246

Depois de coletados o conteúdo das páginas de artigo de cada revisão foi executada a extração de atributos. Uma das etapas da extração de atributos é a conversão do formato wiki para HTML, o que resultou na persistência de aproximadamente 53 mil páginas que consomem um armazenamento de 2,2 GB. O tempo para essa conversão foi de aproximadamente 3 horas. Foram coletados ao todo páginas de discussão de revisões de 2.998 artigos, que compõe uma base de dados de 58.024 instâncias de 70 atributos.

4.2. Coleta da amostra final

Primeiramente foram coletados mais de 6.000.000 de títulos de artigos. A Tabela 3 apresenta, entre outras informações, a distribuição dos títulos por classe de qualidade, na data de março de 2020. Como mostra essa tabela, algumas classe possuem uma representatividade muito pequena, como a classe A que é responsável por menos de 0,02% dos títulos.

Tabela 3. Quantidade de artigos por classe de qualidade

classe	# Artigos total (títulos)	# Artigos da amostra	# Revisões da amostra
FA	5.705	5.750	609.502
A	1.072	1.072	7.053
GA	11.689	5.750	348.257
B	100.410	5.750	450.237
C	328.210	5.750	359.277
Start	1.691.461	5.750	225.451
Stub	4.085.974	5.750	111.979
Total	6.224.521	35.572	2.175.236

Visando tornar a base final mais balanceada alguns títulos foram desconsiderados, em um processo de subamostragem. Para isso, foram selecionados, aleatoriamente, a partir dos títulos coletados, 5750 artigos de cada classe. Para a classe A isso não foi possível, pois haviam apenas 1072 títulos disponíveis, então foram considerados todos.

O algoritmo de coleta foi executado para os títulos da subamostragem, resultando em uma base de dados de mais de 2.000.000 de revisões. O período que marca as datas de início e fim das revisões escolhido foi de janeiro de 2003 até março de 2020, que configura todo o período de dados disponível quando a coleta se iniciou. A distribuição dessas revisões por classe se encontra na última coluna da tabela Tabela 3 e também na forma de gráfico na Figura˜7. As classes com maior representatividade são B, FA e Start.

Figura 7. Distribuição de revisões por classe
Fonte: Os autores

A Tabela 3 apresenta a quantidade de revisões por classe de qualidade da amostra. Essa informação também é representada na Figura 7. A base de dados gerada está disponibilizada publicamente em `https://figshare.com/articles/dataset/FinalWikiEvolutionSample_csv/20154434` e a descrição das colunas encontra-se em `https://github.com/analuizatrz/wiki-crawler/tree/master/FinalWikiEvolutionSample`.

5. Aplicações

O dataset criado pode servir de base de dados para diversos estudos. Dentre eles, os estudos que buscam analisar a evolução da escrita humana, principalmente durante o século 21, poderão utilizar a metodologia proposta para gerar uma amostra abrangente que permita representar a evolução de conteúdo da Internet. Assim, será possível analisar quais aspectos da escrita foram mais desenvolvidos, além de permitir que se estude quais desses aspectos mais influenciam na qualidade geral de um artigo, e, dessa forma, criar ferramentas que determinam sua qualidade de forma automática e façam sugestões de melhoria relevantes para o escritor.

Uma possível ferramenta é a criação de um modelo sequencial de Aprendizado de Máquina que determine a qualidade final de um artigo depois de ter recebido várias alterações. Ou seja, dado um artigo no instante T com uma classe de qualidade C, qual sua classe final Y em um instante T + X, em que X é a quantidade de representações futuras do artigo.

Este trabalho possui algumas limitações. Com relação à discretização mês a mês, houveram revisões que foram agrupadas. Isso se deve a escolha de priorizar um dataset com revisões por um longo período de tempo ao invés de todas as revisões de um período menor, o que poderia afetar algumas aplicações do dataset. Outra limitação é a representação de qualidade baseada apenas no conteúdo da revisão. Uma possível melhoria nesse sentido seria enriquecer a base de dados com outros atributos, como por exemplo de autoria dos artigos, que podem ser coletados com base nos metadados de autoria das revisões e incorporados na representação dos artigos.

Referências

Batista, N. A., Brandão, M. A., Pinheiro, M. B., Dalip, D. H., and Moro, M. M. (2018). Dados de múltiplas fontes da web: coleta, integração e pré-processamento. In de Computação – SBC, S. B., editor, *Anais do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web: Minicursos*, chapter 5, pages 153–192. Sociedade Brasileira de Computação – SBC.

Dalip, D. H. (2015). *Uma Abordagem Multi-Visão para a Estimativa Automática da Qualidade de Conteúdo Colaborativo na Web 2.0*. PhD thesis, UFMG.

Dang, Q. V. and Ignat, C.-L. (2016). Quality assessment of wikipedia articles without feature engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, pages 27–30, New York, NY, USA. ACM.

Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. (2009). Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 295–304, New York, NY, USA. ACM.

Jhandir, M. Z., Tenvir, A., On, B.-W., Lee, I., and Choi, G. S. (2017). Controversy detection in wikipedia using semantic dissimilarity. *Inf. Sci.*, 418(C):581–600.

Pinto, A. C., Silva, B. S., Carmo, P. R. M., Lima, R. L. A., Amorim, L. S. P., Viana, R. T. C., Dalip, D. H., and Oliveira, P. A. C. (2020). Webfeatures: A web tool to extract features from collaborative content. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 103–106, Porto Alegre, RS, Brasil. SBC.

Tyagi, N., Solanki, A., and Tyagi, S. (2010). An algorithmic approach to data preprocessing in web usage mining. *International Journal of Information Technology and Knowledge Management*, 2.

WikiMedia (2019). Mediawiki api help. Disponível em `https://wiki.f-si.org/api.php`. Acesso em 30 de setembro de 2019.

Wikipedia (2019a). Wikipédia: :aviso geral. Disponível em `https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Aviso_geral`. Acesso em 21 de mar de 2019.

Wikipedia (2019b). Wikipédia:content assessment. Disponível em `https://en.wikipedia.org/wiki/Wikipedia:Content_assessment`. Acesso em 20 de jun de 2019.

Wikipedia (2019c). Wikipedia:size of wikipedia. Disponível em `https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`. Acesso em 12 de nov de 2019.