Obtenção de Fórmulas Matemáticas da Base de Artigos arXiv

Beatriz de O. Rodrigues[1], **Flavio B. Gonzaga**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Alfenas (UNIFAL-MG)
CEP 37.133-840 – Alfenas – MG – Brasil

Abstract. *Information Retrieval has become a major point of information access and a good example of the democratization of knowledge. Great tools today utilize this technology and provide great support for academic research and even ordinary users. Since BRI is based on the concept of manipulating large numbers of data, repositories that provide the same stand out widely. This is the example of arXiv, a repository for electronic preprints of scientific articles in various fields with about* 1.5 *million articles. With this in mind, this work aims to obtain and treat data from this repository in order to obtain mathematical formulas, aiming at future indexing of these data by mathematical content search tools.*

Resumo. *A Busca e Recuperação de Informações (BRI) tem se tornado um grande ponto de acesso à informação e um bom exemplo de democratização do conhecimento. Grandes ferramentas atualmente utilizam esta tecnologia e fornecem grande apoio à pesquisas acadêmicas e até mesmo a usuários comuns. Uma vez que a BRI baseia-se no conceito de manipulação de um grande número de dados, repositórios que fornecem o mesmo destacam-se amplamente. Este é o exemplo do arXiv, um repositório para preprints eletrônicos de artigos científicos nos mais diversos campos contando com cerca de* $1,5$ *milhões de artigos. Pensando nisso, este trabalho tem como objetivo obter e tratar dados provenientes deste repositório a fim de obter fórmulas matemáticas, visando futuramente a indexação desses dados por ferramentas de busca de conteúdo matemático.*

1. Introdução

A BRI atrelada ao reconhecimento de expressões matemáticas possui grande relevância no meio acadêmico, em virtude de oferecer a busca de um tipo de conteúdo ainda não indexado por ferramentas de busca textuais. Em um artigo de Vannevar Bush, cientista do MIT e chefe do departamento científico americano durante a Segunda Guerra Mundial, Bush aborda a questão do crescimento escalar de informações, ele ressalta que a publicação tem tomado espaços maiores do que nossa própria habilidade atual de fazer uso de tal. E que o somatório da experiência humana se expande a uma taxa de grande relevância.[1] Demonstrando que o caso do grande contingente de informações arrasta-se através dos anos e que não existe promessa de desaceleração.

Com o passar do tempo, formas de organização destas informações foram surgindo e dando criação a repositórios, arquivos e bibliotecas virtuais. Esses tipos de fontes se mostram extremamente proveitosas de serem manipuladas a fim de aumentar de modo significativo a esfera de buscas de pesquisas, além de proporcionar democratização do conhecimento, reafirmando a ideia de que o conhecimento deveria ser livre e acessível.

Como exemplo bastante atual e de grande magnitude, o arXiv, repositório de pré-publicações eletrônicas (chamados *preprints*) dos mais diversos campos científicos mantido e operado pela Cornell University, apresenta-se ao meio científico seguindo o conceito de *self-archiving* que corresponde ao ato de disponibilizar um *preprint* de algum documento para proporcionar acesso livre ao mesmo.

Ainda em seus primórdios, em setembro de 1999, o arXiv recebeu um total de 2.502 submissões de artigos. Dez anos depois, em setembro de 2009, este número já estava em 5.696. Como podemos observar na Figura 1, ao analisar o mês de janeiro, desde o ano 1 do arquivo ao ano atual, a taxa mensal de submissões manteve-se em uma crescente, mostrando-se uma ótima fonte de informação científica.

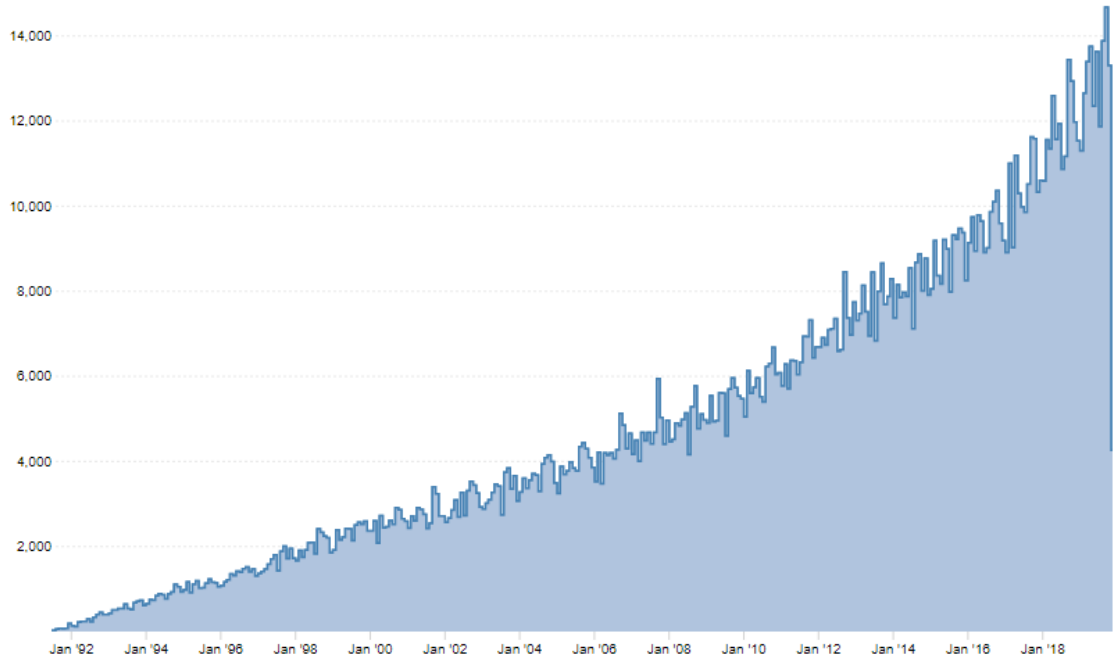

Figura 1. Taxa mensal de submissões ao arXiv em números de artigos enviados. (Fonte: arXiv)

A grande quantidade de informação armazenada nestes arquivos torna a utilização da busca e recuperação aplicada ao reconhecimento de expressões matemáticas por ferramentas de cunho específico muito mais abundante.[2] Não há relatos entretanto de ferramentas que façam a busca e recuperação de fórmulas contidas no arXiv, ainda que muitas façam uso de diversos outros tipos de domínios e bibliotecas.

Este projeto, então, propõe e executa a tarefa de compreender o arquivo em questão, desenvolver algoritmos que sejam capazes de organizar o extenso material obtido, identificar expressões matemáticas e armazená-las de forma a serem referenciadas posteriormente.

O artigo encontra-se organizado de forma a apresentar um referencial teórico na seção 2, seguido pela seção 3 com uma discussão acerca da metodologia utilizada ao desenvolver o projeto, onde encontram-se informações acerca do arquivo arXiv, e também acerca da implementação realizada para a extração dos arquivos e fórmulas. Na seção 4, temos os resultados obtidos ao fim do projeto. E por fim são apresentadas conclusões

acerca destes resultados, além de trabalhos futuros.

2. Referencial Teórico

O termo Busca e Recuperação de Informações (BRI) foi definido de forma pioneira por Calvin Mooers como sendo o nome dado a um processo no qual um usuário de informação é capaz de converter uma necessidade de informação em uma lista real de citações.[3] Com o passar do tempo, o conceito foi tornando-se mais presente e diversos elementos que proporcionam massa de informações foram surgindo. A quantidade de informações existente é uma grande fonte a ser manipulada para um determinado fim, sendo considerada objetivo de estudo de áreas como a Ciência da Informação.[4]

Tefko Saracevic, Professor da Escola de Comunicação e Informação da Rutgers, Universidade Estadual de Nova Jersey, nos Estados Unidos, define a Ciência da Informação como possuindo uma natureza interdisciplinar e ressalta que a evolução da mesma está longe de acabar.[5] Logo, a tecnologia de um sistema de BRI é de benefício geral à área científica como um todo.[6] Áreas dos mais diversos campos científicos têm se adaptado à tecnologia e trazido profundidade à busca de informações, principalmente a área matemática.

Apesar do conceito de Ciência da Informação estar consolidado há algum tempo, concedido durante a Segunda Guerra Mundial, a utilização do mesmo para buscas e recuperação de informações na forma de expressões matemáticas, ao que tudo indica, data de um tempo não tão longo assim. Em um artigo de 2004, Asperti et al [7] já propunha um mecanismo de buscas de cunho matemático porém de domínio de busca muito limitado.

Outro exemplo encontrado na literatura, é uma ferramenta que se insere no conceito de busca e recuperação de fórmulas. Nomeada de MIaS, foi apresentada no ano de 2018 por estudantes do curso de informática da República Tcheca da Masaryk University. O artigo ainda ressalta a importância da fórmula em contraste ao texto para o entendimento da matemática.[8]

3. Metodologia

Este trabalho propõe a obtenção de fórmulas matemáticas de artigos submetidos ao arXiv, seguindo uma abordagem de acesso de dados em massa permitida através da Amazon Simple Storage Service (Amazon S3), um serviço de armazenamento de objetos fornecido pela Amazon Web Service (AWS).

Os dados são obtidos através da Amazon S3 por um custo condizente com as solicitações de dados efetuadas. Esses custos são baseados no tipo de solicitações e cobrados de acordo com a quantidade. Os dados fornecidos pelo arXiv são agrupados em arquivos compactados do tipo TAR e possuem um padrão de 500MB por arquivo, tamanho considerado adequado pela plataforma.

De acordo com dados obtidos através das solicitações realizadas a Amazon S3, o conjunto completo de arquivos de código-fonte do repositório, chamados *source*, possuía quantidade de arquivos em torno de 250GB. O projeto possui como foco esses tipos de arquivos para obtenção das fórmulas e através da plataforma os arquivos a serem manipulados puderam ser obtidos.

Os códigos-fonte dos arquivos, necessários para identificação das expressões, são escritos em um sistema de tipografia de nome TEX, sistema muito conhecido e que tornou-se uma alternativa para a digitação de fórmulas complexas e é apontado como um dos sistemas mais sofisticados do mundo, segundo Yannis Haralambous. [9]

3.1. ArXiv

O arXiv, em seu mais recente relatório anual referente a 2018, divulgado em Janeiro de 2019, contava com um acervo de, aproximadamente, 1.5 milhões de artigos nos mais variados campos. Áreas como Física, Matemática, Ciência da Computação e Economia, encontram-se entre as áreas contempladas.

Dados cedidos pelo próprio repositório de *preprints* relatam um aumento de 17% nas submissões de 2018 em contraste à submissões referentes ao ano de 2017, um aumento de exatas 140.616 submissões. Das submissões, segundo o mesmo relatório, áreas de Ciência da Computação e Matemática, representaram, respectivamente, cerca de 26% e 24% do total de envios, o que expressa uma grande vantagem na manipulação destas informações para fins matemáticos.

Os artigos submetidos ao arXiv podem ser obtidos através da Amazon S3. Assim, ao se baixar um bloco de arquivos à partir da Amazon S3, o seu conteúdo será alguns artigos submetidos em determinado mês e ano. Cada arquivo compactado possui somente artigos do mesmo mês e ano. É possível encontrar ainda no diretório do arXiv na Amazon S3 um arquivo chamado *arXiv_src_manifest.xml*. Este carrega informações como nome do arquivo, primeiro item, último item, número de itens contidos na respectiva compressão, tamanho do arquivo e data de submissão. A data de submissão mostrou-se muito importante, pois através dela pode-se analisar a quantidade de artigos submetidos à plataforma ao decorrer do tempo, nos levando a conclusão de que o crescimento da relevância da plataforma é indiscutível. A Figura 2 mostra a quantidade de arquivos compactados de tamanho até 500MB gerados a cada ano desde o início do arXiv.

3.2. Implementação

Após a fase de obtenção dos arquivos através da Amazon S3, os mesmos precisaram passar por uma etapa de extração, uma vez que são disponibilizados em forma de arquivos compactados; e por uma etapa de extração de fórmulas. As etapas são realizadas através de um algoritmo desenvolvido em linguagem JAVA e segue a modelagem de pacotes conforme a Figura 3.

Os algoritmos referentes às duas etapas do projeto possuem a classe *Stack*, pertencente ao pacote *struct*, uma vez que a estrutura de pilha mostrou-se primordial para que fosse possível simular uma árvore de diretórios, tarefa necessária tanto pra criação de diretórios na primeira fase do projeto, quanto na navegação destes mesmos diretórios para indexação na tabela do banco dados na fase posterior. Com o auxílio desta estrutura foi possível tratar os diretórios de maneira a atingir seu mais profundo nível para então dar-se início a extração.

Na Figura 4, podemos observar um fluxograma que representa a lógica referente a extração dos arquivos. Esta etapa resulta em arquivos dos mais variados tipos, tais como pdf, imagens e arquivos do tipo TEX. Não há um padrão quanto ao tipo de arquivo a ser encontrado, dado que depende da maneira como o autor do artigo estruturou o mesmo.

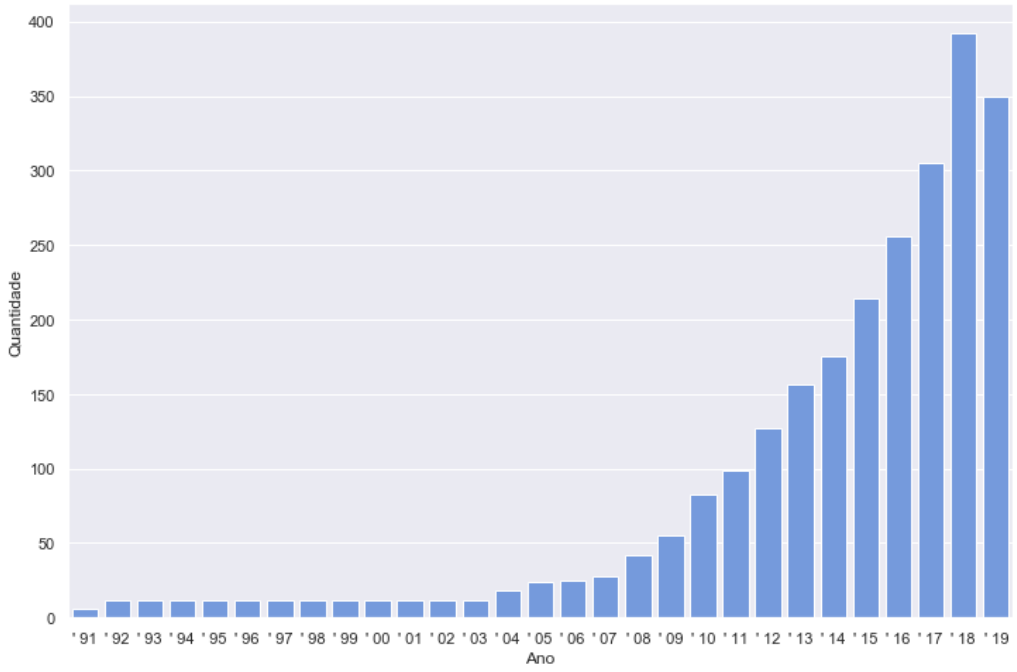

Figura 2. Histórico da quantidade de arquivos compactados (∼500MB) gerados.

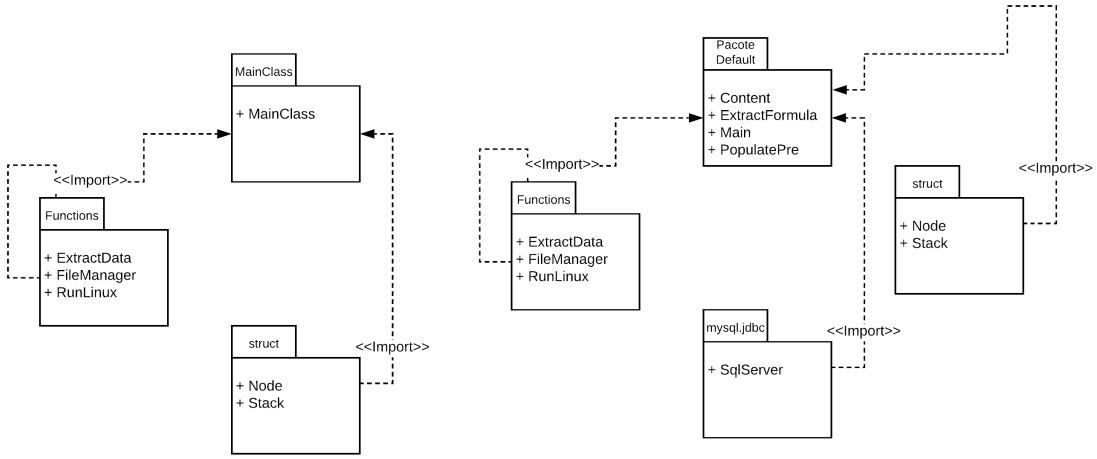

Figura 3. Diagramas de Pacotes dos Algoritmos de extração dos arquivos arXiv e extração de fórmulas.

Ainda no escopo de itens presentes nas submissões, é comum que um artigo submetido seja composto por múltiplos arquivos TEX. Por exemplo, muitos autores dividem os seus textos, criando um arquivo TEX para cada seção do artigo. Além disso, arquivos com extensão *bib*, relacionados à referências bibliográficas, bem como outros relacionados à formatação de estilo também se fazem presente.

Logo, na lógica de extração de fórmulas, foi primordial a junção de todas as seções do artigo que podem ou não estar divididas entre arquivos diferentes, ao menos os com conteúdos relevantes à busca de expressões matemáticas.

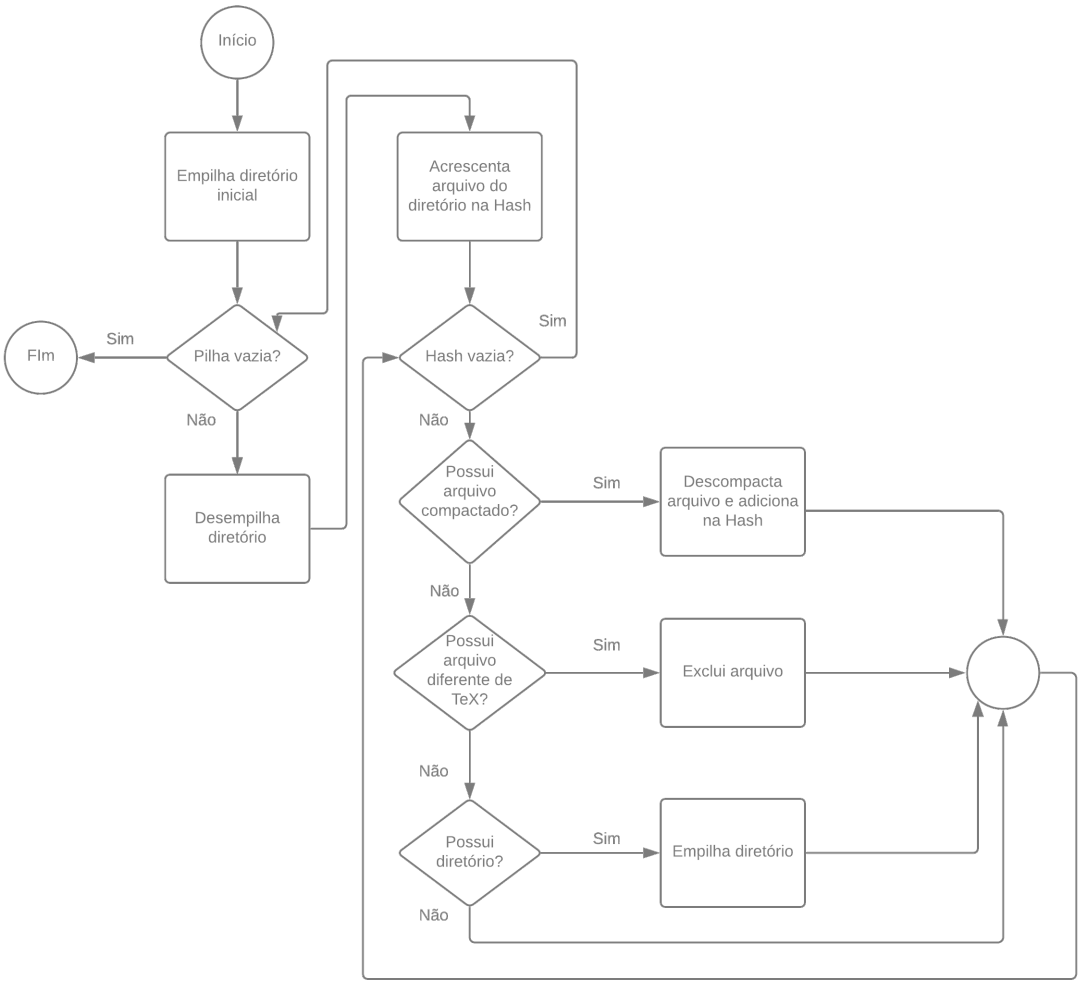

Figura 4. Representação do algoritmo de extração dos arquivos.

Para se identificar o tipo do arquivo, foi utilizado internamente nos algoritmos desenvolvidos o comando *file* do Linux. Assim, arquivos que não sejam do tipo TEX foram excluídos. O comando *file* no entanto, retorna o tipo TEX ainda para arquivos com extensões como: *.cls*, *.sty*, *.aux* e *.bib*, extensões que se referem a códigos de estilo, bibliografia, entre outras coisas não relevantes para a identificação de expressões, sendo estes também excluídos.

Uma vez que a extração dos arquivos foi feita visando a necessidade de futura referenciação ao endereço do artigo no repositório da Web, uma análise inicial da disposição das submissões foi imprescindível.

Segundo informações cedidas na página do repositório arXiv, toda submissão à partir de abril de 2007 possui o padrão de identificador "arXiv:YYMM.number". Onde YY refere-se ao ano da submissão, MM ao mês e *number* ao ID do artigo em si. Já o endereço dos artigos hospedados no repositório possuem formatação "arxiv.org/abs/YYMM.number".

A partir disso, foi fundamental que a criação automática de diretórios e extrações

pudessem seguir um padrão que proporcionasse a restauração dos endereços dos artigos posteriormente. Uma vez que a nomenclatura dos arquivos eram obtidos no formato desejado, nenhum tratamento especial foi necessário.

Após a extração dos arquivos, a segunda parte do projeto baseou-se na extração de fórmulas dos arquivos resultantes da extração da fase anterior. Seguindo o mesmo conceito que o algoritmo anterior, o desta fase percorre os diretórios em busca de arquivos para manipulação. Ao encontrar um arquivo, este é enviado como parâmetro para a classe *PopulatePre*, que se encontra no *Pacote Default* do projeto referente a extração de fórmulas. Caso haja presença de expressões matemáticas a database passa a referenciar o arquivo e a extração de fórmula é feita. A base lógica da fase de extração de fórmulas é representada pelo fluxograma da Figura 5.

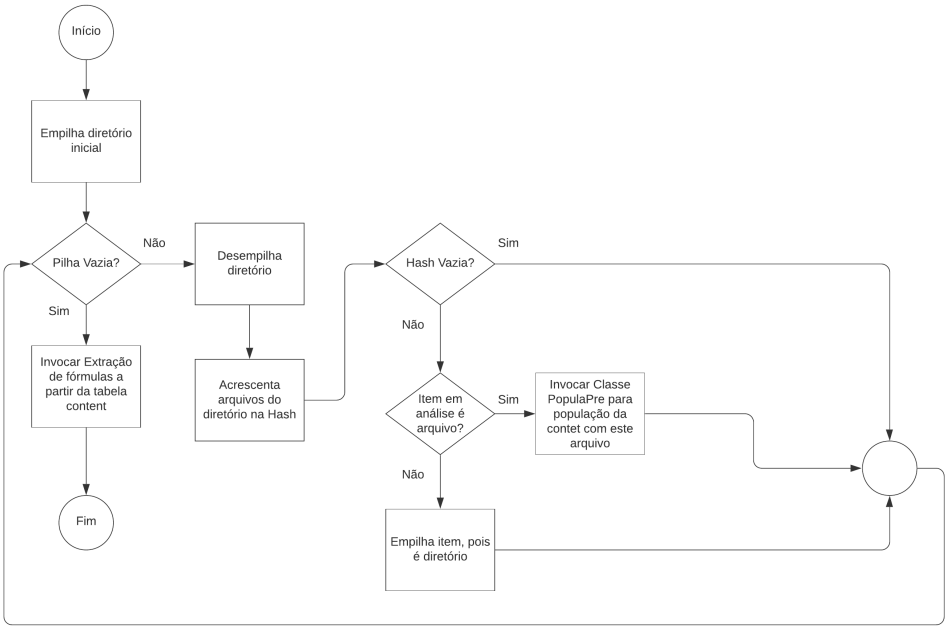

Figura 5. Representação do algoritmo de extração de fórmulas.

Mais adiante apresentam-se diagramas referente ao funcionamento das classes *PopulatePre* e *ExtractFormula*, duas das classes mais importantes de todo o projeto, uma vez que são as que de fato realizam a extração de fórmulas dos artigos.

A extração de fórmulas começa de forma semelhante a etapa de extração de arquivo. Isto por que, para poder caminhar em meio aos diretórios obtendo conteúdo dos arquivos, utiliza-se da mesma lógica utilizada na etapa anterior do projeto.

Uma vez dentro de um diretório, e obtido um arquivo do mesmo, a indexação do artigo no banco de dados acontece por meio da classe *PopulatePre*. Na Figura 6, podemos observar sua lógica e logo após, na Figura 7, podemos observar a organização do banco de dados de nome *pre_search_arxiv*.

Após a etapa de preenchimento da tabela referente ao conteúdo dos arquivos *(tb_content)*, vem a etapa de extração das fórmulas em si. Isto acontece a partir do conteúdo fonte do artigo (em formato LaTeX), já extraído e inserido no campo *co_source* da

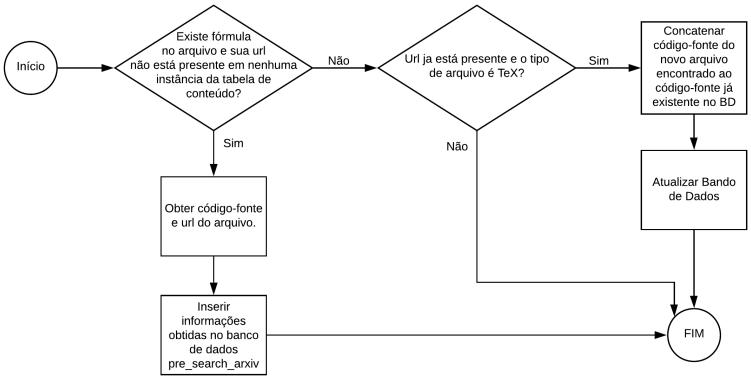

Figura 6. Representação da classe de população do banco de dados

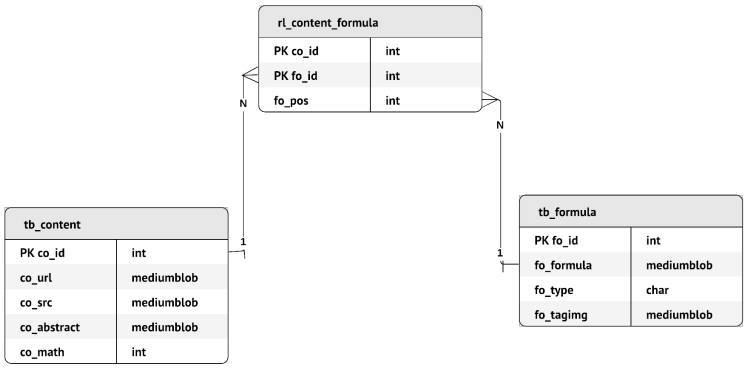

Figura 7. Modelo Relacional do Banco de dados pre_search_arxiv

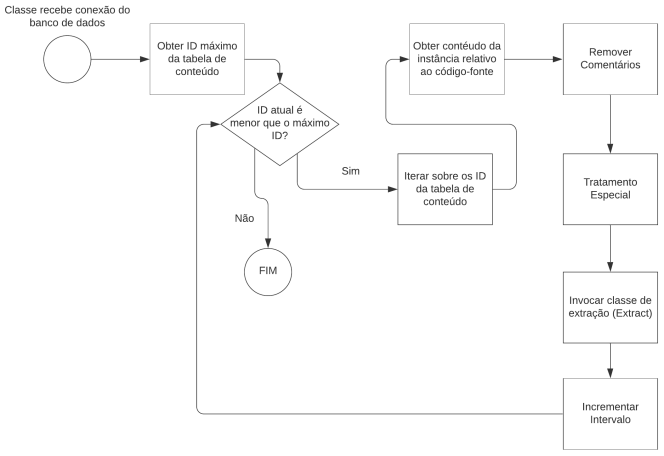

Figura 8. Representação da classe para população da tabela de fórmulas

tb_content. Na Figura 8, vemos a representação do funcionamento da classe para popular

a tabela de fórmulas.

Uma vez obtido o conteúdo do campo *co_source*, já com todas as seções existentes reunidas, o método *Extract* da classe *ExtractData*, do *Pacote Default* realiza a identificação das fórmulas através da análise de códigos em linguagem TEX referentes a expressões matemáticas. Caso exista a presença de uma fórmula, ela é indexada na tabela de nome *tb_formula* e relacionada com o artigo original através de seu ID na tabela de nome *rl_content_formula*.

4. Resultados

Os artigos com submissão de Outubro de 2018 a Setembro de 2019 totalizavam 224GB, já extraídos, contudo, após a execução do código de extração para captura de seus conteúdos e exclusão de arquivos avaliados como irrelevantes, esse tamanho foi reduzido a 13GB. Esta esta etapa referente a extração dos arquivos, durou cerca de 3 dias para a quantidade de arquivos apresentada.

Após a extração dos arquivos compactados, temos a etapa de popular o banco de dados. A tabela *tb_content*, ao final da etapa referente à mesma, totalizou um conjunto de 136.064 artigos científicos contendo fórmulas. À partir desses artigos, foram extraídas $22.346.623$ fórmulas diferentes. Cerca de 4 dias foram necessários para que a execução chegasse ao seu fim.

À partir da tabela de fórmulas, foi possível analisar a quantidade de fórmulas submetidas ao repositório arXiv mês a mês. Para esse gráfico, contou-se quantas fórmulas cada artigo submetido possuía, desconsiderando-se repetições para um mesmo artigo. Contudo, se uma mesma fórmula aparece, por exemplo, em dois artigos diferentes, ela será contada duas vezes, uma para cada artigo. A Figura 9 exibe esses resultados. É possível observar que a arXiv recebe a cada mês algo entre $2,5$ e $3,0$ milhões de fórmulas. Contudo, conforme já mencionando, quando se desconsidera repetições, o valor ao final de um ano é de $22.346.623$ de fórmulas.

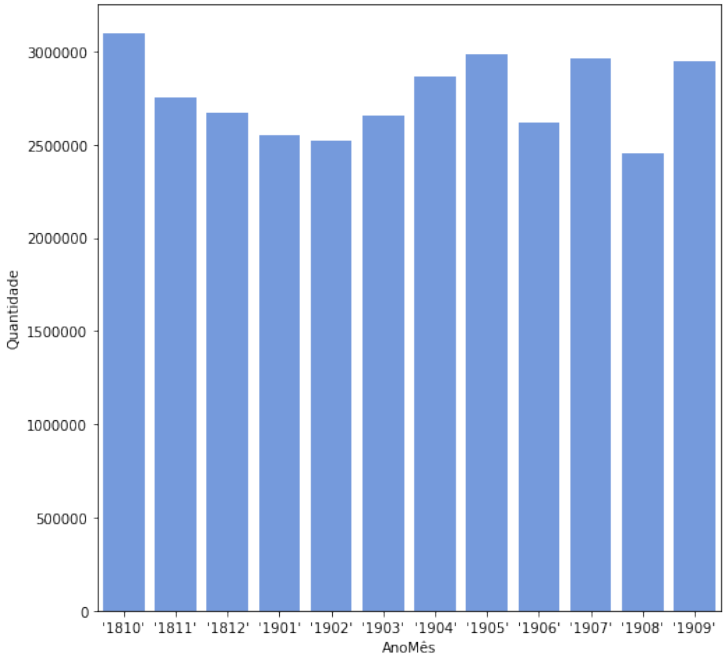

Figura 9. Gráfico de quantidade de fórmulas por Mês e Ano

5. Conclusão e Trabalhos Futuros

Em virtude do crescimento de informações a disposição, encontrar formas de organizar o número elevado de dados disponíveis é primordial. E em se tratando da manipulação de um arquivo com as dimensões do arXiv, em quesito de submissões e visibilidade, o ganho para o meio científico se apresenta ainda mais relevante.

Outro fator de relevância, é a contribuição à acessibilidade das buscas de fórmulas, especialmente por estudantes e profissionais de diversas áreas do meio científico, pelo fato de sites de buscas, agora, poderem indexar uma maior variedade de expressões matemáticas. A vantagem proporcionada pelo arquivo eletrônico arXiv provêm do fato de artigos serem enviados para a plataforma antes mesmo de serem publicados em anais de evento, periódicos ou revistas.

Espera-se, futuramente, executar todo o processo para a base de artigos da plataforma partindo desde seu ano um até os dias atuais. Tornando a cobertura de fórmulas indexadas ainda maior.

Este trabalho forneceu dados para o projeto "SearchOnMath for Research" aprovado preliminarmente para a segunda fase do Programa Centelha, executado pela FAPEMIG, que visa estimular a criação de empreendimentos inovadores.

Referências

[1] Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[2] Ilza Leite Lopes. Estratégia de busca na recuperação da informação: revisão da literatura. 2002.

[3] Calvin Mooers. Zatocoding applied to mechanical organization of knowlegde. *American Documentation*, 2(1):20–32, 1951.

[4] Richard Zanibbi and Dorothea Blostein. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(4):331–357, 2012.

[5] Tefko Saracevic. A natureza interdisciplinar da ciência da informação. *Ciência da Informação*, 24(1), 1995.

[6] Silvana Drumond Monteiro, Rogério Paulo Muller Fernandes, Gian Carlo DeCarli, and Gustavo Lunardelli Trevisan. Sistemas de recuperação da informação e o conceito de relevância nos mecanismos de busca: semântica e significação. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 22(50):161–175, 2017.

[7] Andrea Asperti, Ferruccio Guidi, Claudio Sacerdoti Coen, Enrico Tassi, and Stefano Zacchiroli. A content based mathematical search engine: Whelp. In *International Workshop on Types for Proofs and Programs*, pages 17–32. Springer, 2004.

[8] Petr Sojka, Michal Růžička, and Vít Novotný. Mias: Math-aware retrieval in digital mathematical libraries. 2018.

[9] Yannis Haralambous. *Fonts & encodings*. "O'Reilly Media, Inc.", 2007.